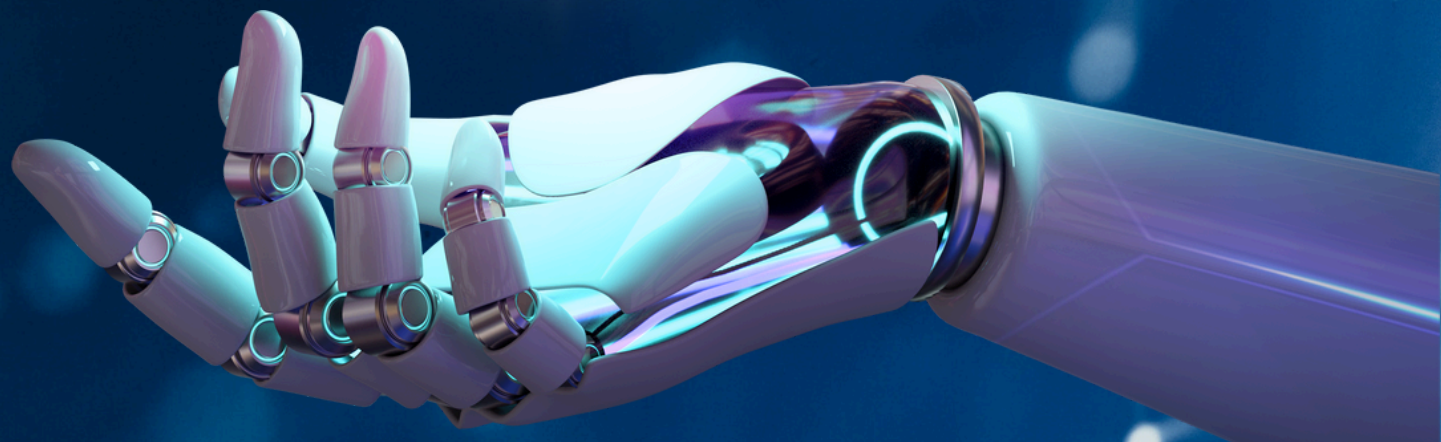
ISSN: 3023 - 8609

VOLUME 3, ISSUE 1, JANUARY 2026

AN INTERDISCIPLINARY JOURNAL OF MEDICAL TECHNOLOGIES

COMPUTERS AND ELECTRONICS IN MEDICINE



ADBA

https://journals.adbascientific.com/cem

ISSN: 3023-8609

Computers and Electronics in Medicine
Volume: 3 – Issue No: 1 (January 2026)

EDITORIAL BOARD

ISSN: 3023-8609

Computers and Electronics in Medicine
Volume: 3 – Issue No: 1 (January 2026)

CONTENTS

https://journals.adbascientific.com/cem/issue/view/19

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.1-10
https://doi.org/10.69882/adba.cem.2026011

Evaluating the Performance Disparity and the Role of Gender-Aware Approaches in Machine Learning Based Disease Detection

Önder Çoban *,[1] **and Ayşe Kartal** *,[2]
*Department of Computer Engineering, Faculty of Engineering, Atatürk University, 25240, Erzurum, Türkiye.

ABSTRACT Machine Learning (ML) is gaining attraction in medical research due to its ability to identify unnoticeable patterns by the human eye. However, concerns about fairness in ML models, particularly performance differences across groups, are growing. This study, therefore, focuses on evaluating the performance disparity and the role of gender-aware approaches in ML-based disease detection. It uses the gender-aware approach and introduces its two new variants by testing them on nine different disease datasets. Intensive experimental evaluations reveal that the detection performance can increase up to an F1-score of 1.0, depending on the nature of the dataset at hand. On the other hand, the gender-aware approach is successful in mitigating the performance disparity only in three out of nine cases. The variants relying on a crossing-over fashion can capture the relationships and different patterns in some cases, but often fall behind the gender-aware approach. This research distinguishes itself through the use of a significant number of datasets and implemented pipelines, of which two are employed for mitigating performance disparity in disease detection for the first time in the literature. The findings of this study, therefore, make important contributions to the field of disease detection in terms of the aforementioned aspects.

KEYWORDS
Gender bias
Performance disparity
Disease Detection
Machine learning

INTRODUCTION

Diseases are continuing to be the top global causes of human deaths. According to a report by WHO (World Health Organization), seven of the ten leading causes of death were noncommunicable diseases, accounting for 38% of all deaths, or 68% of the top ten causes at a global level in 2021 (W.H.O. 2024). This reveals that there is a vital need for effective diagnosis of various diseases globally (Ahsan *et al.* 2022). However, the complexity of the different disease mechanisms and underlying symptoms of the patient population presents solid challenges in the early diagnosis phase and in providing effective treatments. This is because many indications and symptoms are ambiguous and can only be diagnosed by trained health experts who are often prone to error (Ahsan *et al.* 2022). For instance, the symptoms become worse and almost unmanageable as the Alzheimer's disease progresses (Negi *et al.* 2025), and this makes it hard for health workers to diagnose it.

In this context, a report by the National Academies of Science, Engineering, and Medicine revealed that the majority of people encountered at least one diagnostic mistake during their lifespan (Ball *et al.* 2015). The misdiagnosis may be influenced by various factors, like a lack of proper symptoms, which are often unnoticeable for rare diseases, and the disease is mistakenly omitted from the consideration (Ahsan *et al.* 2022). Note that misdiagnosis (or biased outputs) could have severe implications, such as unequal access to diagnosis and treatment (Lozano 2025) in the healthcare context. Contrarily, early diagnosis is highly beneficial in tackling the challenges posed by diseases (Negi *et al.* 2025). Such a task may be achieved more efficiently by predicting results from the data, which is very helpful in making decisions for the medical supervisors. The predicted results may also be useful in medical research where practitioners can get benefits for their medical trials (Sharad *et al.* 2025).

As such, Machine Learning (ML) has attracted the attention of researchers whose recent studies have demonstrated its potential in the medical field with the use of large datasets (Straw and Wu 2022). This is because ML can learn and recognize patterns that may not be apparent to the human eye (Islam and Khanam 2024; Raza *et al.* 2024; Petersen *et al.* 2023) due to the aforementioned challenges in the medical domain. Accordingly, ML techniques can identify trends in medical data and help to develop prediction models which are useful in increasing efficiency of the healthcare

Manuscript received: 9 November 2025,
Revised: 20 December 2025,
Accepted: 25 December 2025.

[1]onder.coban@atauni.edu.tr (**Corresponding author**)
[2]aysekartal62@gmail.com

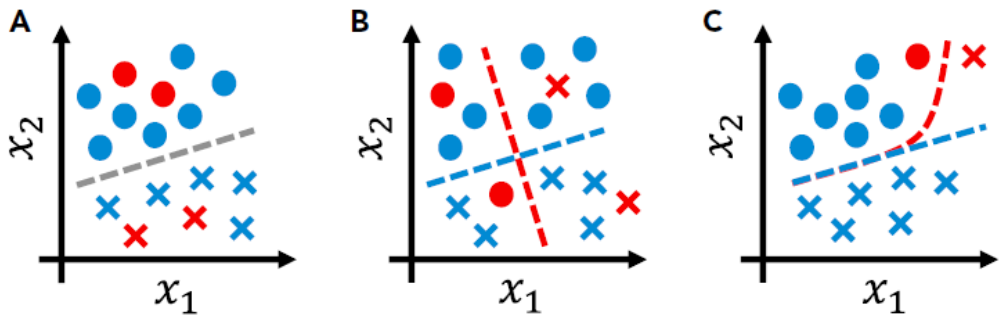


Figure 1 Cases of binary classification for creation of different decision boundaries. Blue circles (majority) and red crosses (minority) represent two patient groups of interest (Petersen *et al.* 2023; Raza *et al.* 2024).

system and managing electronic data in a better way (Sharad *et al.* 2025). Consequently, ML can help with better in-time and correct diagnosis of diseases and offer solutions to difficult medical problems, including the detection of Parkinson's disease (Islam and Khanam 2024), Psychogene Dysphonie (Singhal and Sharma 2024), liver disease (Straw and Wu 2022), Alzheimer's disease (Negi *et al.* 2025), and coronary artery disease (CAD) (Hogo 2020). Despite the promise of ML tools, however, the fairness of ML models has come under increased scrutiny in recent years, with respect to the performance disparities between different groups being one potential source of unfairness (Petersen *et al.* 2023). The discussion has also reached the medical ML community, where the effects of group underrepresentation have received much attention in recent years.

As depicted in Figure 1, ML models are often assumed to find an optimal decision boundary for all subgroups (case A, not problematic). However, existing research efforts often report that there is a performance disparity between subgroups created considering several demographic attributes, including gender (Straw and Wu 2022). In such a situation (case B), optimal decision boundaries differ between groups, and either the model or the input data are not sufficiently expressive to capture the optimal decision boundaries for all groups. Standard ML models or approaches often optimize performance in the majority group (i.e., the blue circles in this case). It is worth noting that the relationship between a group's representation in the training dataset and the model performance for that group is complex. Using similar amounts of data from different groups does not ensure equal model performance across groups, and unfairness exists even with balanced data (Meissen *et al.* 2024; Klingenberg *et al.* 2023). Group underrepresentation, on the other hand, does not necessarily result in poor model performance (Petersen *et al.* 2023). Finally, an expressive model armed with an efficient mitigation technique can learn a decision boundary (red dashed line in Case C) that is optimal for both groups (Petersen *et al.* 2023; Raza *et al.* 2024).

In light of the aforementioned cases, it is clear that ML models should address disparities by properly training the algorithms before implementation in the healthcare sector. This is because any bias in an arbitrary ML model affects its diagnostic accuracy as well as the treatment recommendation given by a medical supervisor (Sharad *et al.* 2025). Nevertheless, bias detection is frequently overlooked, and this is causing to have less-than-ideal results (Kumar and Prabha 2025). The performance disparity often exists due to the estimator (or learner) bias, and using several data-oriented mitigation techniques does not come with guarantees. As a result, improved diagnostics and mitigation remain an open research problem in the medical field (Petersen *et al.* 2023). The data-oriented solutions include using other target variables, bias-robust learners, stratification, more samples, and additional features. On the one hand, narrow algorithmic fairness solutions cannot address all of these issues (Petersen *et al.* 2023).

All of these cases make it clear that fairness and interpretability are as crucial as predictive accuracy when applying ML in healthcare (Lozano 2025), and gender bias is one of the frequently observed complications of ML models in the medical domain. Recent research has demonstrated that ML-based methods are revealing several differences in sex-based health research (Kumar and Prabha 2025). For instance, gender plays a vital role in deciding probable targets of Alzheimer's disease (Negi *et al.* 2025). The performance disparity observed in favour of males (Islam and Khanam 2024; Singhal and Sharma 2024; Kondaka 2024) or females (Hogo 2020; Mushta *et al.* 2024; Klingenberg *et al.* 2023) depending on the nature of the dataset at hand. Not that many studies operate under the assumption that balancing datasets is sufficient to address bias, but biases can still arise at the model level (Lozano 2025).

Building upon these findings, this study aims to inspect the effect of the gender-aware approach in mitigating performance disparity across male and female instances in ML-based disease detection. To the best of our knowledge, it also uniquely presents and utilizes its two variations, never before seen in existing mitigation research. The salient contributions of this study are as follows:

- We implement four different ML pipelines to uncover the real power of the gender-aware approach in automatic disease detection.
- We introduce two variants of the gender-aware approach and comparatively employ them for the first time for mitigating performance disparity in automatic disease detection.
- We use nine different datasets to evaluate the behaviours of pipelines under different circumstances.
- We report our findings by relying on extensive experimental evaluations.

We believe that the contributions given above make this study different from the existing research effort. Please note that this study employs crossing-over pipelines for mitigating the performance disparity in disease detection for the first time in the literature. As such, the findings of this study provide useful insight for researchers studying automatic disease detection.

LITERATURE REVIEW

This section presents our literature review that covers existing studies focusing merely on fairness and bias in ML-based disease detection. The studies that we are aware of are as follows: different ML classifiers are used to detect Parkinson's disease based on MRI (Magnetic Resonance Imaging) data in (Islam and Khanam 2024). The authors analyzed male and female brain scans separately and reported both complete and gender-specific results, which showed that there is a performance disparity in favor of male instances for all brain structures, even when they balanced the instances using a sampling method. Hence, the authors emphasized the need for using possible mitigation techniques to remove the performance disparity. ML is employed to detect Psychogene Dysphonie using voice signals in (Singhal and Sharma 2024). This study performed gender-wise analysis in disease detection by using a hybrid algorithm (Recurrent Neural Network Bidirectional Long Short-Term Memory - RNN_BiLSTM) and revealed that there is a performance disparity between male and female instances. This disparity is observed in favour of males. ML models are used to detect liver disease with the aim of stratified sex analysis in (Straw and Wu 2022). The authors employed their algorithms on both sex-balanced and sex-imbalanced datasets and built their pipelines with and without feature selection. They observed the superiority of Support Vector Machine (SVM) to other ML models and observed that females suffer from a higher false negative rate.

Various ML classifiers are used to diagnose Alzheimer's disease on a single dataset in (Negi *et al.* 2025). The authors experimented with both non-gender-aware and gender-aware approaches and observed that there is a performance disparity across male and female instances. The overall best performance is provided by a modified *k*-Nearest Neighbor (*k*-NN) classifier, which provided higher results on male instances. Another study Hogo (2020) used ML for the diagnosis of CAD. The authors employed a gender-aware approach and observed that the patient's gender affects the structure and performance of the CAD diagnosis system. Unlike the studies (Islam and Khanam 2024; Singhal and Sharma 2024) in which the performance disparity is observed in favour of males, the disparity is observed in favour of females. The use of transcribed speech is explored for automated Alzheimer's disease detection in (Lozano 2025). Unlike the aforementioned studies, this study relies on text data and uses Random Forest (RF), which is fed with linguistic features as well as an LLM (Large Language Model). The author inspected the bias concerning both gender and age attributes of patients and revealed that demographic disparities exist in both models, particularly related to age. Another difference of this study is that it uses Reject Option Classification (ROC) as a mitigation method that significantly improves fairness without substantial reductions in performance. Unstructured text datasets are similarly used in (Kumar and Prabha 2025) for examining gender and sex disparities and suggestions offered for maximizing technology use to improve global health outcomes and reduce inequality.

A framework based on a Multiple Domain Adversarial Neural Network (MDANN) is proposed for mitigating performance disparity in (Li *et al.* 2025). The authors used pre-trained convolutional autoencoders (CAEs) to extract deep representations of brain image data. The findings of this study show that the proposed framework achieves the best balance in terms of accuracy for both sex and handedness in Autism disease diagnosis. Similarly, ML models are used to evaluate bias across race and gender in (Raza *et al.* 2024). The authors detected that there is a performance disparity between male and female instances in favour of females. Convolutional Neural Network (CNN) is used to evaluate potential performance bias for age and sex on MRI data for the diagnosis of Alzheimer's disease (Klingenberg *et al.* 2023). The authors made their evaluation on both balanced and imbalanced data and found that the CNN performed significantly better for women than for men. They concluded that sex differences cannot be attributed to an imbalanced training dataset and therefore point to the importance of examining and reporting classifier performance across population subgroups to increase transparency and algorithmic fairness. Pretrained CNNs (i.e., specifically DenseNet-121 and ResNet-50) are used to evaluate bias and fairness in skin lesion diagnosis in (Kondaka 2024). The findings of this study reveal statistically significant differences in diagnostic performance between genders in favour of males. Additionally, the study found that data augmentation improved accuracy, especially for female skin lesions. ML is used for the diagnosis of PD by exploring the potential imaging biomarkers in (Mushta *et al.* 2024). The authors used Dopamine transporter scan (DATSCAN) images to feed their learners, from which the best model was found to be AdaBoost (AB). The authors also evaluated their pipeline by using the gender-aware approach that separates the dataset by gender to independently evaluate classification performance for male and female participants. The results of this study again show that there is a performance disparity across male and female instances, and the disparity is observed in favour of female instances. Fairness of unsupervised ML models is evaluated on three large-scale publicly available chest X-ray datasets in (Meissen *et al.* 2024). The results of this study revealed that unfairness exists even with balanced data, and it cannot be mitigated by balanced representation alone. On the other hand, male subjects consistently received significantly higher scores across all datasets, even under balanced conditions.

Our review of the literature, briefly provided above, emphasizes that fairness and interpretability are as crucial as predictive accuracy when applying ML in healthcare. As such, a large majority of the studies research bias, especially across gender and age in ML models. On the other hand, studies (Islam and Khanam 2024; Singhal and Sharma 2024; Straw and Wu 2022; Negi *et al.* 2025) focusing on mitigating such a disparity mostly use the gender-aware approach. Some other ones pursue a different purpose and focus on providing fairness metrics (e.g., sAUROC (Meissen *et al.* 2024), ROC (Lozano 2025), and PPGR (Raza *et al.* 2024)) to quantify the fairness of ML models in disease detection. To summarize, existing studies often report bias in ML models across demographic attributes, including gender. They often rely on the gender-aware approach for mitigating the performance disparity and experimenting on a single disease type.

In this study, we therefore employ the gender-aware approach not only for a single disease type but also for nine different disease types to provide a more general view, unlike existing research efforts. In addition, we employ its two variants by following a crossing-over fashion for the first time in the literature for the purpose of mitigating the performance disparity in disease detection. Our extensive results obtained comparatively revealed that the gender-aware approach improves subgroup performance in only three cases. Its two variants rely on crossing-over ML models; on the other hand able to capture the relationships and different patterns in some cases. Building upon these findings, we believe that this study makes an important contribution to the literature by showing the limited efficiency of both the gender-aware approach and its variants, as well as the need for more sophisticated approaches to remove performance disparity, which is not a trivial task.

Table 1 A summarized quantitative description of the underlying datasets

Dataset	# of ...					Distribution of ... instances across targets		B
	INS	F	Targets	Males	Females	male	female	
ALZ	2149	38	2 [0: 1389, 1: 760]	1088	1061	0: 714, 1: 374	0: 675, 1: 386	✗
AND	1421	5	2 [0: 801, 1: 620]	740	681	0: 328, 1: 412	0: 473, 1: 208	✗
ADD	2392	32	2 [0: 2268, 1: 124]	1180	1212	0: 1118, 1: 62	0: 1150, 1: 62	✗
CDD	2206	34	2 [Yes: 1843, No: 363]	1122	1084	Yes: 914, No: 208	Yes: 929, No: 155	✗
CKD	1659	59	2 [0: 135, 1: 1524]	855	804	0: 60, 1: 795	0: 75, 1: 729	✗
HFD	918	19	2 [0: 410, 1: 508]	725	193	0: 267, 1: 458	0: 143, 1: 50	✗
HRD	4240	12	2 [0: 2923, 1: 1317]	1820	2420	0: 1249, 1: 571	0: 1674, 1: 746	✗
LCD	309	28	2 [Yes: 270, No: 39]	162	147	Yes: 145, No: 17	Yes: 125, No: 22	✗
ASD	1054	33	2 [Yes: 728, No: 326]	735	319	Yes: 534, No: 201	Yes: 194, No: 125	✗

INS and F represent the number of instances and features, respectively. The last column B stands for if the corresponding dataset is imbalanced (shown with ✗).

DATASETS

In this study, we use nine different datasets to inspect the effect of the employed methods under different circumstances. The datasets are briefly described as follows:

- Alzheimer's Disease Dataset (ALZ): This dataset (El Kharoua 2024a) includes health information of 2,149 patients. It has been made publicly available in a tabular form and includes demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and a diagnosis of Alzheimer's disease.
- Anemia Dataset (AND): This dataset (Ranjan 2022) includes five features, which are gender, hemoglobin, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). It is created for the purpose of predicting whether a patient is likely to suffer from anemia.
- Asthma Disease Dataset (ADD): This dataset (El Kharoua 2024b) contains health information for 2,392 patients diagnosed with asthma disease. It is composed of several features, including demographic details, lifestyle factors, environmental and allergy factors, medical history, clinical measurements, symptoms, and a diagnosis indicator.
- Celiac Disease Dataset (CDD): It is created by Wageningen University & Research Biotechnology Department to diagnostically predict whether or not a patient has Celiac disease. This dataset (Win 2022) is comprised of several features derived from certain diagnostic measurements.
- Chronic Kidney Disease Dataset (CKD): This dataset (El Kharoua 2024c) contains health information for 1,659 patients diagnosed with chronic kidney disease. It includes features like demographic details, lifestyle factors, medical history, clinical measurements, and medication usage, as well as symptoms, quality of life scores, environmental exposures, and health behaviors.
- Heart Failure Dataset (HFD): This dataset (Soriano 2021) contains 11 features that can be used to predict a possible heart failure, which is a common event caused by cardiovascular diseases. It includes several features like age, sex, serum cholesterol, resting blood pressure, maximum heart rate achieved, and so on.
- Hypertension Risk Dataset (HRD): The dataset (Khan 2023) comprised of both demographic and health-related attributes and was created for predicting the risk of hypertension. It has a total of 13 features, which are gender, age, smoking habits (current smoker and cigarettes per day), medication for high blood pressure (BPMeds), presence of diabetes, total cholesterol levels, systolic and diastolic blood pressure, body mass index (BMI), heart rate, glucose levels, and the corresponding hypertension risk label.
- Lung Cancer Dataset (LCD): This dataset (Bhat 2021) is created to predict cancer risk status based on 16 different attributes. It has a total of 284 instances (or records) and includes features like age, sex, smoking status, existence of chest pain, and so on.
- Autism Screening Data for Toddlers (ASD): This dataset (Fayez 2018) is created with the help of a mobile application called ASDTests to screen autism in toddlers. It has 1,054 records, each of which has values of 17 features.

Note that a quantitative description of the datasets briefly described above is given in Table 1. For more detailed information, the reader is advised to refer to the respective cited references.

METHODS

In this study, we implement four different ML pipelines, which are depicted in Figure 2, showing that all pipelines commonly involve preprocessing, classification, and performance measurement steps. The only difference is arising from the way of cross-validation employment. As seen in Figure 2, the baseline pipeline simply employs cross-validation on overall instances, while the gender-aware pipeline divides the instances into two disjoint subsets that

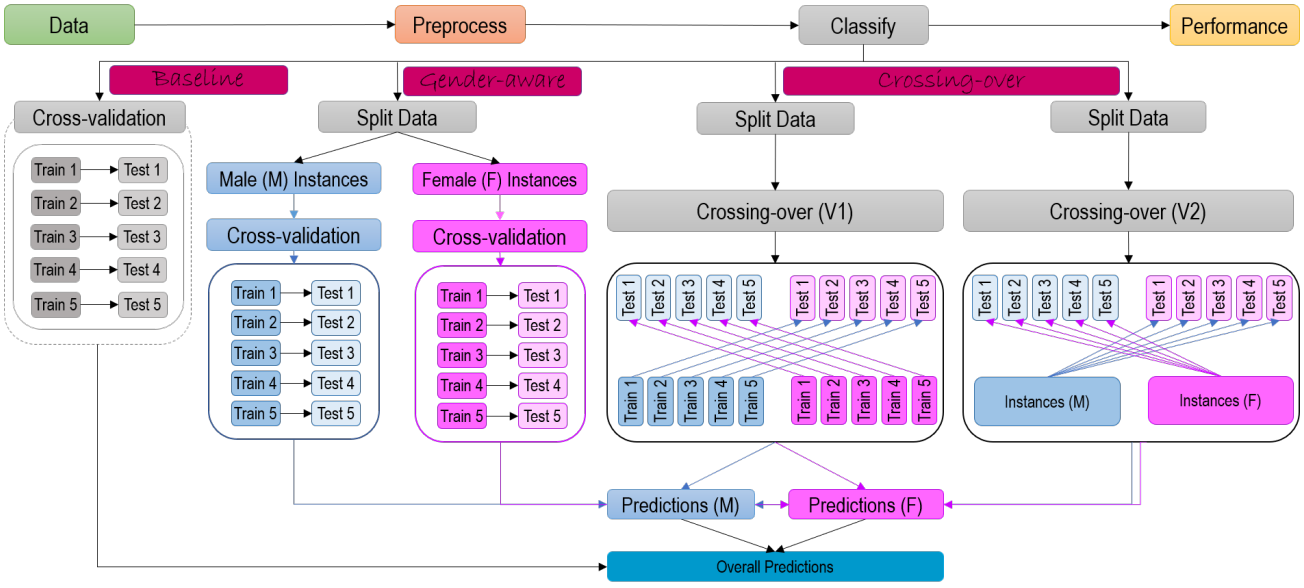


Figure 2 Flowchart of our evaluation that relies on four different pipelines.

include only female and male instances. It then employs cross-validation on these subsets and finally reports both gender-wise and overall performances. On the other hand, the two variants of gender-aware pipeline simply rely on the idea of the crossing-over technique, in which any ML model trained on male instances is then used to predict the labels of female instances, and vice versa. There is also a difference between the crossing-over pipelines, considering the size of the instances used to train the model at hand. The first variant (i.e., V1) simultaneously applies cross-validation on both male and female instances and uses only single training and test sets at each fold. Contrarily, the second variant (i.e., V2) uses the complete sets of male and female instances to make predictions on their subsets again, following the crossing-over fashion. The following subheadings provide the details of methods used to implement the aforementioned pipelines.

Preprocessing

This is the first step employed on our datasets. It involves the following two data cleaning tasks (Freire *et al.* 2025):

- Imputing: Some datasets used in this study include missing values. Hence, missing values are filled using a linear interpolation method implemented in the Pandas (McKinney 2011) library.
- Scaling: The feature values in each dataset are scaled and translated into a range between 0 and 1. This task is achieved by using MinMaxScaler implemented in sklearn (Pedregosa *et al.* 2011) package.

Classification

Upon completion of the preprocessing, the datasets are used to feed ML classifiers for binary classification. We used several classifiers, which are briefly described as follows:

- SVM: This classifier tries to find a linear or non-linear hyperplane that separates classes from each other. Maximizing the margin between the hyperplane and instances on either side corresponds to a lower generalization error (Yang *et al.* 2025; Kotsiantis *et al.* 2007).
- RF: This classifier is also a type of ensemble learning and trains multiple number of decision trees on different subsets of the data at hand. The final decision for any test instance is then given by using majority voting among the trees (Breiman 2001).
- AB: Follows an ensemble fashion in which a meta classifier (often selected to be a tree-based learner) is trained on the data at first. The copies of this classifier are trained on the same data set again, with a difference that they mainly focus on the instances the meta classifier has difficulty in classifying (Zhu *et al.* 2009). The final decision to classify an instance is given by a weighted voting such that the more a classifier provides good performance, the more it has influence on the final decision.
- Gradient Boosting Classifier (GBC): This learner relies on multiple regression trees as an additive model that follows a stage-wise fashion. It aims to minimize the loss (i.e., the difference between the actual and predicted classes of the training data) to make a better classification (Friedman 2001).
- Logistic Regression (LR): It is a statistical algorithm for transforming a linear regression by a sigmoid function and deciding classification by calculating the distance to the decision boundaries it previously built between classes (Yang *et al.* 2025).
- Sthocasting Gradient Descent Classifier (SGD): This is actually a way of training any ML model, like SVM and LR, by optimizing several loss functions. In other words, it tries to minimize the loss by iteratively updating the parameters of the model at hand (Zhang 2004).

Please note that we employ the learners briefly described above by using their implementations in the scikit-learn Python package (Pedregosa *et al.* 2011). The kernel of SVM is selected to be Radial Basis Function (RBF), while the solver and the number of iterations (i.e., max_iteer) of LR are set to be liblinear and 1000, respectively. The remaining ones and all settings of parameters for other classifiers are left untouched. It is important to note that we intentionally left almost all of the parameters at their default values since this strategy is more appropriate than an aggressive hyperparameter optimization when fairness is the key concern. Us-

ing default parameter settings provides a strong and reproducible baseline for comparison. Accordingly, any practitioner ensures that all models and all groups are treated consistently. Otherwise, it may become unclear if observed unfairness is due to data, model design, or tuning choices.

Performance Measurement and Evaluation

The globally accepted way of evaluating the performance of any ML classifier is to use precision (P), recall (R), accuracy, and /or F1-score metrics, which are actually derived from the confusion matrix. For a binary classification task, this matrix stores four values, namely true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In this study, we report our results using True Positive Rate (TPR), True Negative Rate (TNR), and F1-score that is calculated as follows (Chicco and Jurman 2020; Freire *et al.* 2025):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{1}$$

where $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$. Note that R, also known as sensitivity, is a synonym of TPR and stands for the proportion of all actual positives that are classified correctly as positives. The $TNR = \frac{TN}{TN+FP}$ is, on the other hand, also known as specificity, which measures the proportion of all actual negatives that are classified correctly as negative (Monaghan *et al.* 2021). We would also like to note that the performance measurement results are reported by using a stratified cross-validation (Kohavi 1995; Coban 2022) strategy, which is configured to run with five folds in this study. This means that the dataset at hand is divided into train and test subsets five times, and the performance of the learner at hand is measured on five disjoint subsets (i.e., training and test sets). An average value of the five folds is reported so as to ensure that our results are as reliable as possible.

We would like to strongly emphasize that we intentionally rely on well-known performance metrics and report the results only with weighted F1 score (robust across imbalanced datasets), TPR, and TNR values to save space and reduce complexity. Even though there exist several fairness metrics (e.g., sAUROC (Meissen *et al.* 2024), ROC (Lozano 2025), and PPGR (Raza *et al.* 2024)), they rely on different aspects and are not accepted as benchmark methods. Hence, we used classical performance metrics that do not harm the reliability of our evaluation and ease the burden of using additional metrics for possible comparative analysis in the future.

RESULTS AND DISCUSSION

In this section, we provide our experimental results and their discussion. Using the methods introduced in Section Methods, we performed classification experiments on our datasets (see Section Datasets). Please note that we employed all classifiers on all datasets by creating four pipelines, namely baseline, gender-aware, cross-over-v1, and cross-over-v2. This yielded too many results, making it challenging to report all in this study. Hence, we only report the results for the best cases of pipelines for each dataset. We report the results not only with F1-score but also with TPR and TNR to provide a deeper insight into the performance effect of pipelines on different instance groups for each dataset. For example, we experimented with the aforementioned pipelines with all classifiers on the ALZ dataset, but only reported the best cases in which GBC is the best model to reduce complexity and space. The results are provided in Table 2, where the best overall F1 scores for each scenario are in bold typeset. As seen in Table 2, the best pipeline on the ALZ dataset is baseline with the best f1-score of 0.947. On the other hand, the crossing-over approach (with both v1 and v2) outperforms the gender-aware approach, and the cross-over-v2 pipeline achieves the second-best f1-score of 0.942. Inspecting the performance metric values across instance groups also reveals that the performance of pipelines is higher for females than for males in all cases.

On the AND dataset, baseline and gender-aware pipelines show the same behavior with the same best f1-score of 1.0, and they outperform the crossing-over pipelines. Interestingly, crossing-over pipelines show different behaviours. Using v1, scores on males are higher than the results obtained on females. The pipeline v2, on the other hand, reverses this situation and provides a better overall f1-score of 0.764. Baseline and gender-aware pipelines provide perfect classification of both male and female instances. On the ADD dataset, the behaviour of the pipelines is the same, and they provide the same overall f1-score of 0.922. Their performances are also the same for male and female instance groups. Unlike the ALZ dataset, the performance values are higher than the that ones for the females on this dataset, which shows that performance disparity can also be observed in favor of males. On the CDD dataset, the behaviours of the baseline and gender-aware pipelines are the same, with an f1-score of 0.996. Similarly, two versions of the crossing-over pipelines show the same behavior with an f1-score of 0.994. Performance values obtained on female instances are higher than the results obtained on males in all cases. However, baseline and gender-aware pipelines outperform the other two pipelines and provide perfect classification of female instances. On the CKD dataset, the best f1-score of 0.911 is provided by the baseline pipeline. On the other hand, the crossing-over-v2 pipeline provides the second-best f1-score of 0.908 and provides a slightly different improvement on male instances. However, the f1-scores are higher again on female instances compared to male instances in all cases. On the HFD dataset, the best f1-score of 0.867 is provided by the gender-aware pipeline. The crossing-over pipelines fall behind the baseline and gender-aware pipelines, but v2 outperforms v1.

A closer look at the results of crossing-over pipelines shows that these pipelines classify female instances incorrectly. Considering the overall f1-score, it seems that the results are higher on male instances compared to the results obtained on females. The gender-aware pipeline mitigates performance disparity in favour of male instances and also improves the overall f1-score from the baseline pipeline's f1-score of 0.864 to 0.867. On the HRD dataset, the gender-aware pipeline again mitigates performance disparity in favour of female instances, also by improving the overall best f1-score of 0.900. The crossing-over pipelines fall behind the baseline pipeline and do not help to mitigate the performance disparity. Nevertheless, the behaviors of crossing-over pipelines are different even though their overall f1-scores are the same (i.e., 0.892). On the LCD dataset, the best f1-score of 0.915 is provided by the gender-aware pipeline. On the other hand, crossing-over pipelines fall behind the baseline pipeline. There is a performance disparity in favour of male instances in all cases. Interestingly, the crossing-over v2 pipeline provides a slight improvement on the TNR value on male instances, even though it provides a lower overall f1-score (i.e., 0.906) than the respective baseline. Finally, on the ASD dataset, the behaviours of all pipelines are the same with a perfect classification.

These results make it clear that the gender-aware approach improves the overall f1-score effectively in only three instances. The second version (v2) of the crossing-over approach performs as well or better than the first version (v1). However, both crossing-over

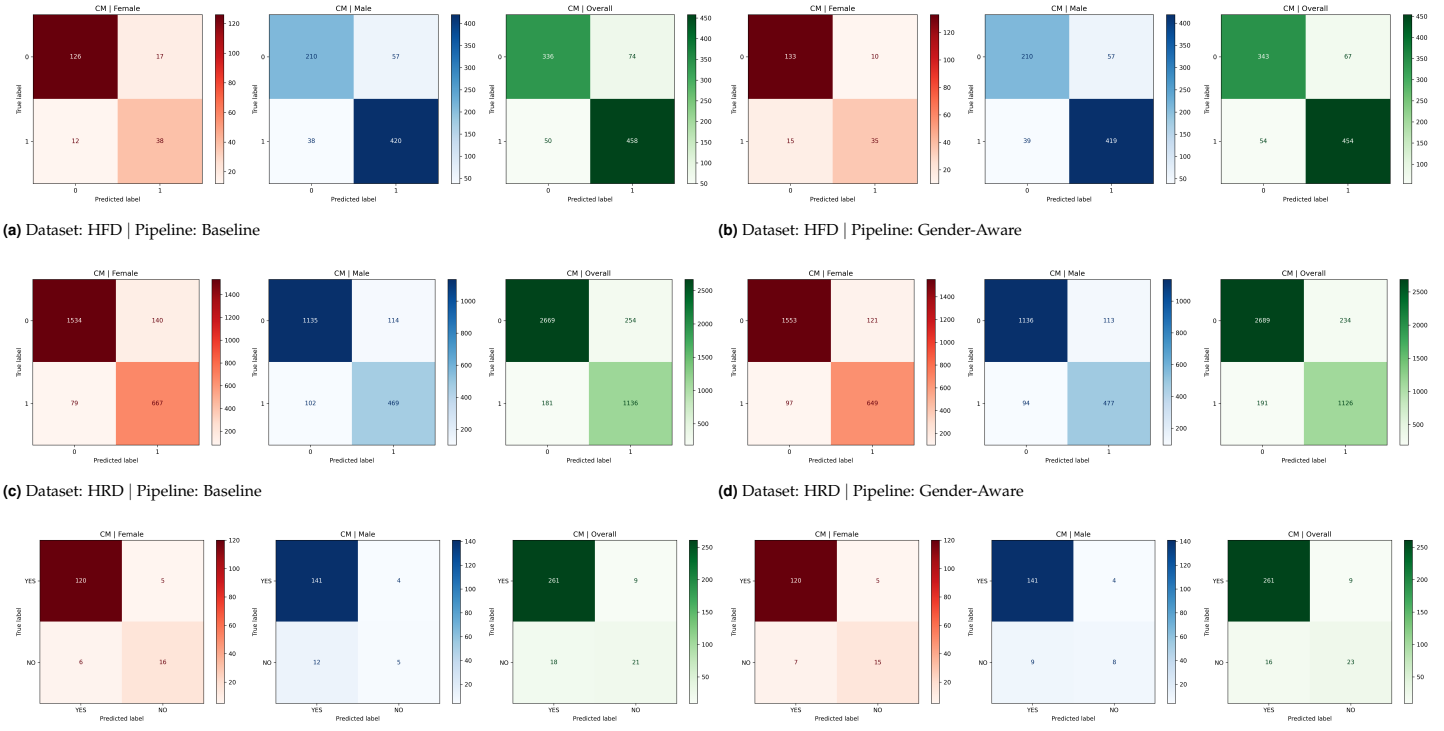
Table 2 Weighted average F1, TPR, and TNR values considering four pipelines on nine datasets across different instance groups

Data	IG	Baseline				Gender-Aware				Cross-Over (V1)				Cross-Over (V2)			
		BC	F1	TPR	TNR	BC	F1	TPR	TNR	BC	F1	TPR	TNR	BC	F1	TPR	TNR
ALZ	M	GBC	0.938	0.938	0.939	GBC	0.921	0.921	0.921	GBC	0.927	0.927	0.928	GBC	0.937	0.937	0.938
	F		0.956	0.956	0.957		0.940	0.940	0.940		0.940	0.940	0.940		0.947	0.947	0.948
	O		**0.947**	0.947	0.948		**0.931**	0.931	0.931		**0.933**	0.933	0.934		**0.942**	0.942	0.943
AND	M	GBC	1.000	1.000	1.000	GBC	1.000	1.000	1.000	SGD	0.783	0.775	0.848	SGD	0.759	0.750	0.840
	F		1.000	1.000	1.000		1.000	1.000	1.000		0.692	0.708	0.861		0.771	0.777	0.866
	O		**1.000**	1.000	1.000		**1.000**	1.000	1.000		**0.738**	0.740	0.747		**0.764**	0.764	0.764
ADD	M	LR	0.923	0.948	1.000	LR	0.923	0.948	1.000	LR	0.923	0.948	1.000	LR	0.923	0.948	1.000
	F		0.921	0.947	1.000		0.921	0.947	1.000		0.921	0.947	1.000		0.921	0.947	1.000
	O		**0.922**	0.948	1.000		**0.922**	0.948	1.000		**0.922**	0.948	1.000		**0.922**	0.948	1.000
CDD	M	GBC	0.992	0.992	0.993	AB	0.992	0.992	0.993	GBC	0.992	0.992	0.993	GBC	0.992	0.992	0.993
	F		1.000	1.000	1.000		1.000	1.000	1.000		0.995	0.995	0.995		0.995	0.995	0.995
	O		**0.996**	0.996	0.996		**0.996**	0.996	0.996		**0.994**	0.994	0.994		**0.994**	0.994	0.994
CKD	M	GBC	0.899	0.919	0.967	AB	0.897	0.907	0.932	AB	0.892	0.911	0.956	GBC	0.901	0.920	0.967
	F		0.922	0.934	0.963		0.915	0.926	0.951		0.906	0.915	0.935		0.915	0.930	0.967
	O		**0.911**	0.927	0.965		**0.906**	0.917	0.942		**0.899**	0.913	0.945		**0.908**	0.925	0.967
HFD	M	RF	0.851	0.849	0.846	LR	0.868	0.870	0.876	LR	0.808	0.797	0.808	LR	0.827	0.818	0.828
	F		0.867	0.868	0.872		0.866	0.867	0.870		0.779	0.775	0.804		0.783	0.779	0.806
	O		**0.864**	0.864	0.867		**0.867**	0.868	0.869		**0.780**	0.779	0.792		**0.787**	0.787	0.799
HRD	M	RF	0.910	0.909	0.908	GBC	0.910	0.909	0.909	RF	0.908	0.906	0.906	RF	0.910	0.908	0.909
	F		0.881	0.881	0.880		0.886	0.886	0.885		0.870	0.871	0.877		0.866	0.868	0.873
	O		**0.898**	0.897	0.896		**0.900**	0.899	0.899		**0.892**	0.891	0.891		**0.892**	0.891	0.890
LCD	M	LR	0.924	0.925	0.926	LR	0.916	0.918	0.922	LR	0.904	0.911	0.932	LR	0.922	0.925	0.931
	F		0.887	0.901	0.934		0.913	0.919	0.935		0.892	0.907	0.945		0.887	0.901	0.934
	O		**0.907**	0.912	0.925		**0.915**	0.919	0.927		**0.898**	0.909	0.937		**0.906**	0.912	0.928
ASD	M	GBC	1.000	1.000	1.000	GBC	1.000	1.000	1.000	GBC	1.000	1.000	1.000	GBC	1.000	1.000	1.000
	F		1.000	1.000	1.000		1.000	1.000	1.000		1.000	1.000	1.000		1.000	1.000	1.000
	O		**1.000**	1.000	1.000		**1.000**	1.000	1.000		**1.000**	1.000	1.000		**1.000**	1.000	1.000

IG and BC stand for instance group and the best classifiers, respectively. Values M, F, and O under column IG represent males, females, and overall (males and females), respectively.

variants often underperform the gender-aware approach in terms of overall f1-score. Despite this, the crossing-over approach shows advantages in specific cases. For example, both crossing-over variants outperform the gender-aware approach on the ALZ dataset. On the CDD dataset, they match the gender-aware approach's performance on male instances. On the LCD dataset, both variants provide better TNR values across all instances compared to the baseline and gender-aware pipelines. These findings emphasize that pipeline performance is dataset-dependent. The gender-aware approach can mitigate performance disparities, while the crossing-over approach, particularly v2, can enhance TNR values. Performance disparities favoring males are seen in three datasets (i.e., ADD, HRD, and LCD), while disparities favoring females are seen in four datasets (i.e., ALZ, CDD, CKD, and HFD). No significant performance disparity between genders is observed only in two datasets (i.e., AND and ASD).

A closer look at the confusion matrices makes it clearer to easily observe the changes in values of both correctly and incorrectly



(a) Dataset: HFD | Pipeline: Baseline

(b) Dataset: HFD | Pipeline: Gender-Aware

(c) Dataset: HRD | Pipeline: Baseline

(d) Dataset: HRD | Pipeline: Gender-Aware

(e) Dataset: LCD | Pipeline: Baseline

(f) Dataset: LCD | Pipeline: Gender-Aware

Figure 3 Confusion matrices of different instance groups (i.e., male, female, and overall) for three cases in which the gender-aware pipeline outperforms the baseline pipeline on HFD, HRD, and LCD datasets (see Table 2).

classified instances. As seen from the confusion matrices depicted in Figure 3, the gender-aware pipeline only improves results in favour of both male and female instances on the HRD dataset. It improves the overall f1-score on HFD and LCD datasets, but causes a slightly different decrease in male and female instances, respectively. Table 3 provides the summarized view of Figure 3 with respect to the total number of correctly classified instances (CCIs) and incorrectly classified instances (ICIs). As seen in Table 3, the total number of ICIs on female, male, and entire instances is reduced by 4 (i.e., from 29 to 25), increased by 1 (i.e., from 95 to 96), and reduced by 3 (i.e., from 124 to 121), respectively, on the HFD dataset by using the gender-aware pipeline. The number of CCIs increased by 4 (i.e., from 164 to 168) on female instances and 3 (i.e., from 194 to 797) on overall instances, while it decreased by 1 (i.e., from 630 to 629) on male instances. On the HRD dataset, the number of CCIs is increased by 1 (i.e., from 2201 to 2202), 9 (i.e., from 1604 to 1613), and 10 (i.e., from 3805 to 3815) for female, male, and entire instances, respectively. This paved the way to decrease the number of ICIs by 1 (i.e., from 219 to 218), 9 (i.e., from 216 to 207), and 10 (i.e., from 435 to 425) on the same instance groups, respectively. Finally, on the LCD dataset, the number of CCIs is increased on male instances by 3 (i.e., from 146 to 149) while decreased by 1 (i.e., from 136 to 135) on female instances. This situation resulted in a total of 2 (i.e., from 282 to 284) increase in the number of CCIs on the entire dataset.

As such, the findings reveal that the dataset-specific performance variations likely stem from differences in data characteristics, feature distributions, and the presence of biases. The gender-aware approach's success in mitigating disparities suggests that it effectively addresses gender-related biases present in the data or model. The crossing-over approach's occasional advantages indicate that it might be capturing different patterns or relationships within the data, potentially related to how gender interacts with other features. In other words, the feature interactions are very close across male and female instances, and this explains why any model trained on males improves TNR values on females. On the other hand, the varying performance of crossing-over variants could be due to differences in feature interactions in different datasets. Complications arise from the potential for overfitting to specific datasets and the difficulty in generalizing results to new data.

All of the aforementioned implications show that the role of gender-aware approaches is limited in mitigating the performance disparity. Unfortunately, using similar amounts of data from different groups does not ensure equal model performance across groups, and unfairness exists even with balanced data. This observation shows that equal representation of subgroups is a necessary but not sufficient condition for fairness. The performance disparity may still arise from physiological differences, feature-label mismatches, model biases, and structural inequities in healthcare data. As such, ML-based disease detection must address how data is modeled and evaluated, not just how much data is collected. Sample size (i.e., for both the overall dataset and subgroups) is another case that may directly affect the performance since the generalization ability of ML models often increases on large datasets. However, the results of both this study and existing research efforts reveal that the performance disparity persists on large datasets as well. This case is also showing that mitigating performance disparity in disease detection cannot be solved solely by using data-oriented approaches but also requires developing fairness-aware modeling strategies.

This analysis does not delve into the specific features or biases driving these results, which limits a deeper understanding. Hence, further investigation is necessary to fully understand the implica-

Table 3 Total number of CCIs and ICIs across different instance groups of three datasets on which the gender-aware pipeline outperforms the baseline pipeline

Pipeline	# of ...	Dataset | Instance Groups (Male →, Female → F, and Overall → O)								
		HFD (see Figures 3a, 3b)			HRD (see Figures 3c, 3d)			LCD (see Figures 3e and 3f)		
		F	M	O	F	M	O	F	M	O
Baseline	CCIs	164	630	794	2201	1604	3805	136	146	282
	ICIs	29	95	124	219	216	435	11	16	27
Gender-Aware	CCIs	168	629	797	2202	1613	3815	135	149	284
	ICIs	25	96	121	218	207	425	12	13	25

CCI and ICI stand for the number of correctly and incorrectly classified instances, respectively. The values are extracted from the confusion matrices of Figure 3.

tions of these findings and develop more effective and equitable ML systems. A more detailed feature analysis or an efficient feature selector may be used to mitigate the performance disparity as much as possible. Note that this is not a trivial task, and a large majority of existing research efforts rely on the gender-aware approach whose success is shown to be limited in this study. This study, therefore, strongly suggests understanding the underlying causes of performance disparities, especially concerning the feature correlations across instance groups, and using robust ML models in mitigating performance disparities.

CONCLUSION

In this study, we studied the problem of automatic disease detection using classical ML techniques. We aim to inspect the effect of gender-aware and crossing-over approaches in the mitigation of performance disparity, mostly observed between male and female instances in disease detection. For this purpose, we intentionally used nine different disease datasets to provide a more general view. One interesting outcome of this study is that the performance disparity can also be observed in favour of male instances in disease detection. On the other hand, the gender-aware approach helps to efficiently mitigate the performance disparity only in three cases, and therefore, its success is limited. Crossing-over approach, on the other hand able to capture different patterns and relationships within data, again with a limited capability. Hence, we conclude that there is still an open room for mitigating the performance disparity, which is not a trivial task in automatic disease detection. Further investigation is also necessary to fully understand the implications of these findings and develop more effective and equitable ML systems.

As future work, we are planning to conduct further research on the mitigation of the performance disparity. For this purpose, we will similarly employ several well-known deep learners to inspect if the performance disparity still exists when traditional learners are replaced with deep learners. Providing an intense effort to run deep learners on 1D patient records will be another future direction of this study. Another direction will be making an effort to find the fairest ML models in general by evaluating several well-known fairness metrics.

Author Contributions

The authors equally contributed to this work. This paper is derived from the second author's master's thesis, supervised by the first author. They all read and approved the final version of the paper.

Funding Information

The authors received no financial support for the research, authorship, and/or publication of this study.

Availability of data and material

Available upon request.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this study.

LITERATURE CITED

Ahsan, M., S. A. Luna, and Z. Siddique, 2022 Machine-learning-based disease diagnosis: A comprehensive review. Healthcare **10**: 541.

Ball, J., B. Miller, and E. Balogh, 2015 *Improving diagnosis in health care*. National Academies Press, Washington.

Bhat, A. M., 2021 Lung cancer. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer, Accessed 4 Aug 2025.

Breiman, L., 2001 Random forests. Machine learning **45**: 5–32.

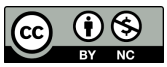
Chicco, D. and G. Jurman, 2020 The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics **21**: 6.

Coban, O., 2022 A new modification and application of item response theory-based feature selection for different machine learning tasks. Concurrency and Computation: Practice and Experience **34**: e7282.

El Kharoua, R., 2024a Alzheimer's disease dataset. https://www.kaggle.com/dsv/8668279, Accessed 4 Aug 2025.

El Kharoua, R., 2024b Asthma disease dataset. https://www.kaggle.com/dsv/8669080, Accessed 4 Aug 2025.
El Kharoua, R., 2024c Chronic kidney disease dataset. https://www.kaggle.com/dsv/8658224, Accessed 4 Aug 2025.
Fayez, F., 2018 Autism screening data for toddlers. https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers, Accessed 4 Aug 2025.
Freire, P., D. Freire, and C. C. Licon, 2025 A comprehensive review of machine learning and its application to dairy products. Critical reviews in food science and nutrition **65**: 1878–1893.
Friedman, J. H., 2001 Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232.
Hogo, M. A., 2020 A proposed gender-based approach for diagnosis of the coronary artery disease. SN Applied Sciences **2**: 1060.
Islam, N. and R. Khanam, 2024 Gender variability in machine learning based subcortical neuroimaging for parkinson's disease diagnosis. Applied Computing and Informatics .
Khan, R., 2023 Exploring predictive factors for hypertension risk prediction. https://www.kaggle.com/datasets/khan1803115/hypertension-risk-model-main, Accessed 4 Aug 2025.
Klingenberg, M., D. Stark, F. Eitel, C. Budding, M. Habes, *et al.*, 2023 Higher performance for women than men in mri-based alzheimer's disease detection. Alzheimer's Research & Therapy **15**: 84.
Kohavi, R., 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pp. 1137–1145, Montreal, American Association for Artificial Intelligence.
Kondaka, A., 2024 Evaluating gender bias and fairness in skin lesion diagnoses using convolutional neural networks. The National High School Journal of Science **2024**: 1–14.
Kotsiantis, S. B., I. Zaharakis, and P. Pintelas, 2007 Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering **160**: 3–24.
Kumar, V. and C. Prabha, 2025 Unlocking gender-based health insights with predictive analytics. In *AI-Based Nutritional Intervention in Polycystic Ovary Syndrome (PCOS)*, edited by A. N. Rakesh K., Meenu G., pp. 141–165, Springer, Singapore, first edition.
Li, B., X. Jiang, K. Zhang, A. Harmanci, B. Malin, *et al.*, 2025 Enhancing fairness in disease prediction by optimizing multiple domain adversarial networks. PLOS Digital Health **4**: e0000830.
Lozano, R. S., 2025 *Assessing Bias in Machine Learning Models for Alzheimer's Disease Detection Across Gender and Age*. Master's thesis, Leiden University, Leiden.
McKinney, W., 2011 Pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing **14**: 1–9.
Meissen, F., S. Breuer, M. Knolle, A. Buyx, R. Muller, *et al.*, 2024 (predictable) performance bias in unsupervised anomaly detection. Ebiomedicine **101**: 1–10.
Monaghan, T. F., S. N. Rahman, C. W. Agudelo, A. J. Wein, J. M. Lazar, *et al.*, 2021 Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina **57**: 503.
Mushta, I., S. Koks, A. Popov, and O. Lysenko, 2024 Exploring the potential imaging biomarkers for parkinson's disease using machine learning approach. Bioengineering **12**: 11.
Negi, H. S., R. Indu, S. C. Dimri, B. Kumar, N. Bisht, *et al.*, 2025 Detecting alzheimer's disease (gender-based) using different machine learning approaches. In *10th International Conference on Signal Processing and Communication (ICSC)*, pp. 357–362, Noida, IEEE.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, *et al.*, 2011 Scikit-learn: Machine learning in python. The Journal of Machine Learning Research **12**: 2825–2830.
Petersen, E., S. Holm, M. Ganz, and A. Feragen, 2023 The path toward equal performance in medical machine learning. Patterns **4**: 1–9.
Ranjan, R. B., 2022 Anemia dataset. https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset, Accessed 4 Aug 2025.
Raza, S., A. Shaban-Nejad, E. Dolatabadi, and H. Mamiya, 2024 Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review. IEEE Access **12**: 180815–180829.
Sharad, C., P. Mrinal, M. Nandita, and G. Meenu, 2025 *AI and Machine Learning in Modern Healthcare*. Transforming Gender-Based Healthcare with AI and Machine Learning, Taylor & Francis, New York.
Singhal, A. and D. K. Sharma, 2024 Comparative analysis of gender-wise disease detection based on voice signal analysis. In *International Conference on Next-Generation Communication and Computing*, edited by S. K. D., S. R., and P. S., pp. 389–401, Ghaziabad.
Soriano, F., 2021 Heart failure prediction dataset. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction, Accessed 4 Aug 2025.
Straw, I. and H. Wu, 2022 Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ health & care informatics **29**: e100457.
W.H.O., 2024 The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accessed 4 Aug 2025.
Win, J., 2022 Celiac disease (coeliac disease). https://www.kaggle.com/datasets/jackwin07/celiac-disease-coeliac-disease, Accessed 4 Aug 2025.
Yang, G., S. Luo, and P. Greer, 2025 Advancements in skin cancer classification: a review of machine learning techniques in clinical image analysis. Multimedia tools and applications **84**: 9837–9864.
Zhang, T., 2004 Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116–116, Banff Alberta, ACM.
Zhu, J., H. Zou, S. Rosset, and T. Hastie, 2009 Multi-class adaboost. Statistics and its Interface **2**: 349–360.

How to cite this article: Coban, O., and Kartal, A. Evaluating the Performance Disparity and the Role of Gender-Aware Approaches in Machine Learning Based Disease Detection. *Computers and Electronics in Medicine*, 3(1), 1-10, 2026.



ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.11-26
https://doi.org/10.69882/adba.cem.2026012

Comparison of Artificial Intelligence Applications of EEG Signals in Neuroscience

Ahmet Hüsrev Akdeniz *,[1] **and Can Bülent Fidan** *,[2]
*Department of Mechatronics Engineering, Faculty of Engineering and Natural Science, Karabuk University, Karabuk, 78050, Türkiye.

ABSTRACT In recent years, there has been a growing interest in the artificial intelligence (AI)-based analysis of electroencephalography (EEG) signals. This surge has made the potential of EEG more evident, both in monitoring cognitive states and in the early diagnosis of neurological disorders. This review systematically evaluates the academic literature from the past decade focusing on the processing of EEG signals through machine learning (ML), deep learning (DL), and other alternative techniques. The study compares personalized ML models (e.g., SVM, Random Forest) with wavelet decomposition–based optimized approaches and further analyzes the performance of Hilbert transform–based Convolutional Neural Network (CNN) architectures, label-free autoencoder frameworks, and multi-architecture DL systems in contemporary brain–computer interface (BCI) applications. In addition, incremental learning models based on multimodal data fusion are reviewed in the context of diagnosing disorders such as Alzheimer's disease and epilepsy. The findings indicate that EEG–AI integration holds substantial potential for both research and clinical applications.

KEYWORDS
Artificial intelligence
Machine learning
Deep learning
Brain-computer interface
Neuroscience
EEG signals

INTRODUCTION

Electroencephalography (EEG) is a non-invasive neuroimaging technique that measures electrical fluctuations on the scalp resulting from action potentials generated during neuronal activity (Edelman *et al.* 2025; Kimmatkar and Babu 2021). The signals obtained through metal electrodes allow for high temporal resolution recording of electrical patterns in the brain. Owing to this feature, EEG stands out as a particularly valuable tool for real-time monitoring of mental and emotional processes.

The history of electroencephalography (EEG) dates back to 1875, when Richard Caton made the first observations on animals; following Hans Berger's successful recording of human EEG signals in 1924, it gained widespread clinical and scientific application (Wan *et al.* 2019). Today, EEG is effectively used in a wide range of areas, from epilepsy diagnosis to the analysis of sleep disorders, from detecting attention deficit, depression, and mood disorders to rehabilitation-oriented systems (Liu *et al.* 2025). This broad spectrum of applications has also facilitated the development of EEG devices with various electrode types, connectivity technologies, and user-friendly designs (Soufineyestani *et al.* 2020). In particular, brain-computer interfaces (BCIs) have attracted significant attention due to their potential to interpret EEG data and translate individuals' mental states into external system commands (Wan *et al.* 2019; Sozer and Fidan 2017; Sözer and Fidan 2019). Over the past five decades, research efforts have enabled brain–computer interface (BCI) systems to evolve from experimental foundations into applicable technologies across various domains, including clinical practice, rehabilitation, and human–machine interaction. This transformation has encompassed not only technical advancements but also sparked multidimensional discussions surrounding ethics, user experience, and societal acceptance (Kawala-Sterniuk *et al.* 2021). For example, motor imagery-based prosthetic control can be provided for paralyzed individuals, while alternative communication mechanisms can be developed for those with limited interaction abilities (Orban *et al.* 2022).

The integration of electroencephalography (EEG) into artificial intelligence (AI)-based systems, particularly for the diagnosis and monitoring of neurodegenerative diseases, represents a new paradigm in clinical decision-making (Mouazen *et al.* 2025). This integration is supported by digital tools that enable the preprocessing and enhancement of EEG signals for analytical purposes. Notably, open-source MATLAB-based platforms such as EEGLAB and Fieldtrip facilitate artifact removal, time–frequency analysis, and feature extraction through standardized modules. Extensions integrated into these platforms, such as MARA, AAR, ADJUST, clean rawdata, and icablinkmetrics, allow for the automatic detection and removal of artifacts caused by eye blinks, muscle activity, and environmental noise. This provides a robust preprocessing infrastructure that enhances the reliability and accuracy of downstream analyzes (Gu *et al.* 2021).

Manuscript received: 5 November 2025,
Revised: 1 January 2026,
Accepted: 2 January 2026.

[1]2428133509@karabuk.edu.tr (**Corresponding author**)
[2]cbfidan@karabuk.edu.tr

With the advancement of AI algorithms, EEG data analysis has undergone a significant transformation. Machine learning (ML) and deep learning (DL) approaches have demonstrated high accuracy in classifying emotional states, cognitive levels, and mental tasks by learning spatiotemporal patterns from EEG signals (Khan *et al.* 2024; Nandakumar *et al.* 2025). In this context, AI-based analytical approaches involving multi-layered processes, such as feature extraction, classification, and even compression of EEG signals, are becoming increasingly comprehensive (Khlief and Idrees 2023). These developments not only enhance clinical diagnostic workflows but also expand the role of EEG in multidisciplinary applications, including human–computer interaction (HCI), driver fatigue detection, and emotion-aware user interfaces.

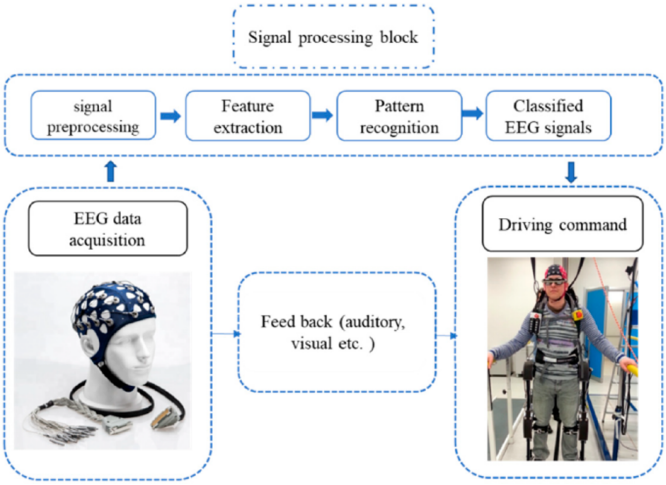


Figure 1 The rule of the EEG signal as the main element in the BCI-EEG rehabilitation system (Orban *et al.* 2022).

In this study, the scientific literature focusing on the analysis of EEG signals through AI-based methods is examined in depth. Within this scope, prominent machine learning and deep learning approaches are systematically categorized and compared, both in the context of individualized models and collective frameworks.

METHODOLOGY OF LITERATURE REVIEW

This review adopts a structured methodology to explore the recent advances in artificial intelligence applications on EEG signals. Academic articles published between 2018 and 2025 were systematically retrieved from reputable databases including PubMed, IEEE Xplore, Scopus, and SpringerLink. The search strategy involved Boolean keyword combinations such as "EEG + Deep Learning", "Brain-Computer Interface + AI", and "EEG + Classification"

Studies were included in the article based on the following criteria:

- Use of real EEG data (clinical or experiment in in most studies, traditional ML methods such as Support Vector Machines (SVM), k-nearest neighbors (kNN), and decision trees have been used for classification purposes due to their simplicity and interpretabilityental)
- Implementation of AI techniques (ML or DL)
- Focus on emotion recognition, mental task classification, BCI systems, or neurological diagnosis

Studies were not included in the article based on the following criteria:

- Simulation-only data
- Hardware-only reviews without algorithmic analysis
- Articles lacking peer-review

After deduplication and abstract screening, 63 articles were included in this review.

RECENT ADVANCES IN AI-DRIVEN EEG APPLICATIONS

Electroencephalography (EEG) signals are currently evaluated in integration with disciplines such as artificial intelligence, bioinformatics, psychology (Elnaggar *et al.* 2025), and neuroscience, and are utilized across a wide variety of application domains. In this context, existing studies span a broad spectrum, ranging from the characterization of meditation to the diagnosis of neurological disorders, and from mental command recognition to human–robot interaction.

Meditation, Mental State, and Mindfulness Classifications

In a study aiming to distinguish techniques such as Vipassana, Isha Shoonya, and Himalayan Yoga, a one-dimensional convolutional neural network (1D-CNN) combined with chi-square–based feature selection achieved an accuracy rate of 60% (Jain *et al.* 2025). The differentiation of meditation states through EEG signals contributes to the objective assessment of individuals' levels of mental awareness.

Studies aimed at classifying cognitive states such as mental relaxation and concentration constitute the fundamental building blocks of BCI systems (Aggarwal and Chugh 2022; You 2021). In addition, in studies focusing on the classification of imagined words, classification was performed using CNN after Hilbert transformation; however, model complexity has been a limiting factor in real-time applications (Agarwal and Kumar 2024). In the field of mental task recognition, autoencoder architectures operating with unlabeled data have been proposed, where "abnormal" states are identified based on reconstruction errors (Dairi *et al.* 2022).

Classification of Emotional States

The classification of emotional states using EEG signals holds a significant place in the literature (Mendivil Sauceda *et al.* 2024). The biologically informed Spiking Neural Network-based BISNN model has improved emotional classification performance and enhanced sensitivity through the biological meaningfulness of synaptic parameters (Sun *et al.* 2025).

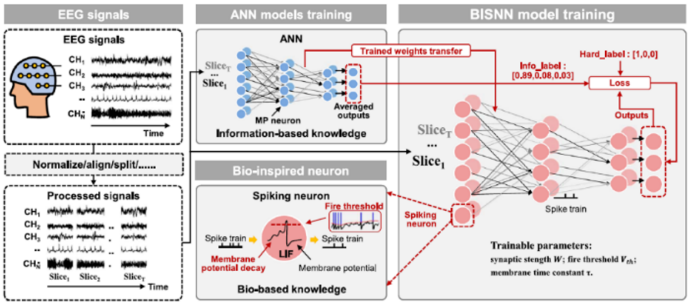


Figure 2 Overview of the training process of the proposed BISNN method for EEG-based emotion recognition. Pre-processed EEG signals are divided into T-slices to reduce temporal feature complexity. Pre-trained weights from a knowledge-based ANN model are transferred to a homogeneous BISNN model composed of biologically inspired neurons (Sun *et al.* 2025).

Similarly, models developed for the classification of emotions in the context of human-robot interaction (HRI) have achieved high accuracy through the use of global optimization algorithms, demonstrating that emotional intelligence can be integrated into HRI systems (Staffa *et al.* 2023)

Studies on the classification of anxiety levels have been conducted using algorithms such as SVM, kNN, and Decision Tree; the use of closed-loop neurofeedback systems for therapeutic purposes has also been evaluated (Chen et al., 2021). Additionally, a personalized machine learning approach was developed using the DEAP dataset, taking into account individual emotional differences (Barrowclough *et al.* 2025)

Additionally, a comprehensive study conducted by Reza et al. proposed a machine learning framework to classify four emotional states, positive, neutral, depressed, and anxiety, using EEG signals. The study utilized a dataset of 300 patients and addressed class imbalance through advanced data augmentation techniques including GAN, SMOTE, and ADASYN. Discrete Wavelet Transform and Shannon entropy were applied for feature extraction, and nine different ML/DL algorithms (MLP, CNN, RF, SVM, etc.) were compared. The highest accuracy of 98.8% was achieved with the MLP model (Sobhani *et al.* 2025).

Recent review studies have also highlighted the progress and limitations of deep learning techniques in EEG-based emotion recognition. In particular, Abgeena and Garg systematically analyzed more than one hundred studies published between 2018 and 2024, emphasizing that models such as BiLSTM, CNN, and hybrid CNN–LSTM architectures achieved the highest accuracy rates across datasets like SEED and DEAP (Abgeena and Garg 2025). Similarly, Li and Chen (2025) proposed a cross-modal alignment and fusion framework that combines EEG and visual features through a hybrid attention mechanism, achieving up to 96.49% accuracy on the SEED dataset.

The analysis of mental stress levels using EEG in virtual reality (VR) environments has also emerged as a prominent topic in recent years. In one study, stress levels were classified using EEG signals recorded in a VR environment, and the performances of algorithms such as SVM and Random Forest were compared (Albayrak-Kutlay and Bengisu 2025; Kamińska *et al.* 2021).

Diagnosis and Follow-up of Neurological Diseases

In the differentiation of neurodegenerative diseases such as Alzheimer's Disease (AD) and Frontotemporal Dementia (FTD), the Coherence-CNN method achieved three-class classification with over 94% accuracy by utilizing functional connectivity measures (Jiang *et al.* 2025). In another recent study, a wavelet-ML framework incorporating OTFL-THFB decomposition was proposed for Alzheimer's detection using EEG signals. With features derived from Hjorth Parameters and Higuchi's Fractal Dimension, the model achieved up to 98.91% accuracy, outperforming existing state-of-the-art classifiers (Puri *et al.* 2025).

Wang *et al.* (2025b) extracted effective connectivity information using the GRU-GC algorithm for pre-surgical focus localization in patients with refractory epilepsy, enabling the classification of epileptic foci through directional connectivity graphs. Pacia (2023) demonstrated that sub-scalp implantable telemetric EEG (SITE) systems enable the identification of diagnostic biomarkers and objective monitoring of treatment responses in neurological and behavioral disorders beyond epilepsy through long-term EEG recording. Tautan *et al.* (2025) systematically reviewed the use of unsupervised and self-supervised machine learning methods on EEG data for epilepsy, highlighting methodological trends and clinical application gaps in the field.

Shafiezadeh *et al.* (2024) evaluated patient-independent AI models for EEG-based epileptic seizure prediction and emphasized the lack of generalizability and methodological validation in most existing studies. Gurmessa and Jimma (2025) conducted a systematic review evaluating interpretable artificial intelligence (XAI) methods for EEG-based epileptic seizure diagnosis, emphasizing the balance between model performance and clinical interpretability. The study provided a comprehensive methodological framework for developing reliable, ethical, and explainable AI systems in epilepsy diagnosis. The study developed by Leela and Helenprabha (2025) stands out with the proposed TMDFILE model, a two-level multimodal data fusion and incremental learning approach aimed at the early diagnosis of Alzheimer's disease. By integrating different data types such as EEG, MRI/PET, speech, and written text, this approach dynamically optimizes inter-modality contribution levels through the use of a gating mechanism.

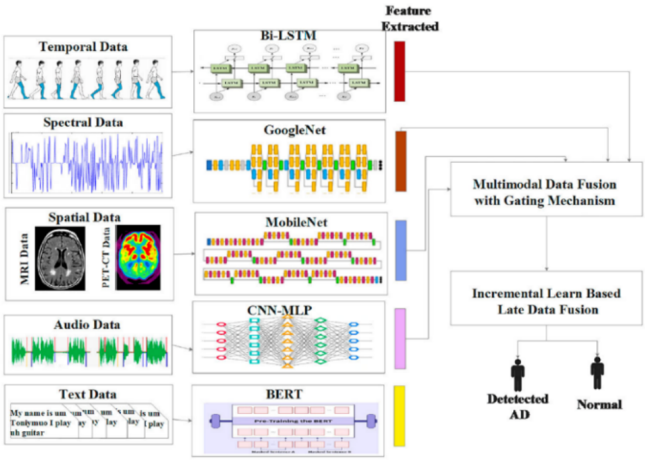


Figure 3 Overview of workflow process (Leela and Helenprabha 2025).

Thanks to its incremental learning structure, the model can adapt to newly incoming data and achieved 94.5% accuracy when tested on five different datasets (ADNI, OASIS, EEG, AUSD, BRATS). Compared to traditional methods such as CNN, SVM, and RF, TMDFILE demonstrated superior performance in terms of both accuracy and generalizability (Leela and Helenprabha 2025; Shang *et al.* 2024; Uyanik *et al.* 2025).

Brain-Computer Interfaces (BCI)

BCI systems are effectively utilized, particularly in the rehabilitation of individuals with limited motor abilities Fig. 4 A review study on these systems evaluated the performance of potentials such as P300 and SSVEP in terms of signal processing (Sözer and Fidan 2018), pattern recognition and control techniques (Orban *et al.* 2022).

In another study, EEG-based BCI applications for elderly and disabled individuals were examined, and the application methods of machine learning algorithms such as ANN, SVM, and LDA to EEG data were elaborated (Wan *et al.* 2019). Endogenous EEG-based BCI systems developed to enable online communication for individuals with complete loss of motor functions rely on paradigms that operate without external stimuli (Turi *et al.* 2021). In the study conducted by Han et al., a classification approach based on Riemannian geometry achieved an accuracy of 87.5% (Han *et al.* 2019; Ma *et al.* 2025; Remsik *et al.* 2022).

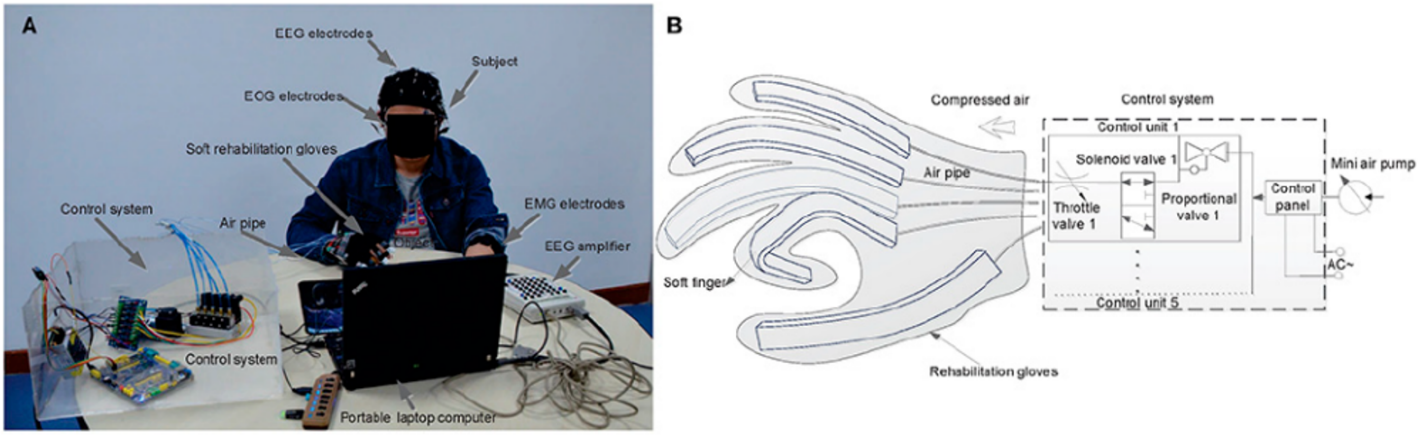


Figure 4 (a) Shows the prototype model BCI and its experimental conditions, (b) is the control scheme of the soft robot hand (Orban *et al.* 2022).

Furthermore, a comprehensive review on EEG-based BCI systems systematically presented the strengths and weaknesses of deep learning architectures such as CNN, RNN, LSTM, GAN, and Transformer models (Alshehri *et al.* 2025; Hossain *et al.* 2023). Another noteworthy study on the classification of motor movements from EEG signals was conducted by Al-Dabag and Ozkurt (2019). In this study, classification was performed using statistical features obtained from EEG signals through artifact removal, wavelet-based frequency band decomposition, and cross-correlation with effective channels (Al-dabag and Ozkurt 2019). In experiments conducted on both the BCI Competition III Dataset IVa and data recorded using the Emotiv device, classification accuracies exceeding 98% were achieved with ANN and SVM algorithms (Al-dabag and Ozkurt 2019). Similar signal processing approaches contribute to the development of BCI systems used not only in clinical settings but also in various application areas such as gaming, education, and marketing (Maiseli *et al.* 2023).

Mental Command and Human-Machine Interaction

Mental command recognition using EEG expands the possibilities of thought-based control in human–machine interaction. Classification of EEG signals preprocessed with EMD using deep neural networks has enabled the differentiation of various mental commands (Agrawal *et al.* 2024). In another study, biometric identification based on individuals' EEG signals was achieved, where a multi-band deep embedding learning network provided high-accuracy identification using features extracted from SSVEP signals (Gu *et al.* 2025). In a study based on a hybrid BCI architecture, motor imagery (MI) and steady-state visual evoked potential (SSVEP) signals were combined to control a quadcopter in 3D space. The system, which switches between two modes using an eye-blinking signal, is capable of controlling flight directions through eight different EEG commands, achieving a classification accuracy of 87.09%. Enhanced with real-time feedback and offline optimization, this architecture offers a high level of control capacity based on mental commands in human–machine interaction (Yan *et al.* 2020).

Sleep, Fatigue, and Mental State Monitoring

In studies focusing on tracking time-dependent cognitive states, fatigue levels were successfully detected using a spatial–temporal CNN and a bidirectional LSTM-based model (Ahn *et al.* 2016; Jeong *et al.* 2019). Building on EEG-derived cognitive state models, a vocal fatigue detection system was developed for air traffic controllers using SVM and Random Forest classifiers. The system performs cognitive state classification based on speech signal features such as MFCCs, fundamental frequency, and energy (Kouba *et al.* 2023).

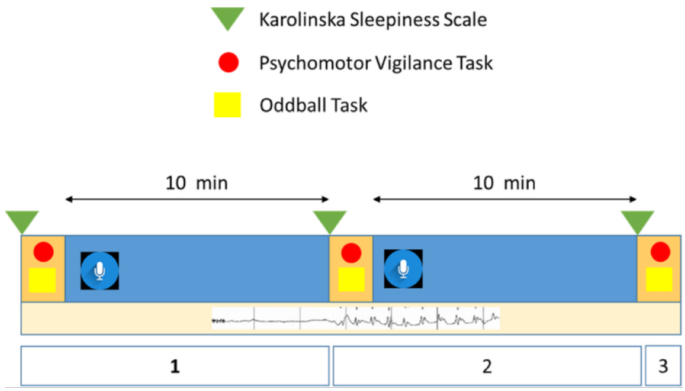


Figure 5 The experimental session representation (Kouba *et al.* 2023).

Additionally, wavelet transform-based multilayer models were employed to classify mental states such as alertness and drowsiness by enhancing time–frequency resolution (Joo *et al.* 2025; Khare *et al.* 2023). In a study on the detection of driver fatigue using EEG signals, the fatigue state was classified with high accuracy through the fusion of four different entropy measures (spectral, approximate, sample, and fuzzy). Using selected classifiers (SVM, BP, RF, and kNN), an accuracy of up to 98.3% was achieved, and effective results were obtained with only four channel regions. These findings highlight the traceability of mental states in human–machine interaction systems and their potential contribution to driving safety (Kouba *et al.* 2023).

In this context, Hassan et al. conducted a comprehensive survey summarizing the current progress and key challenges in electroencephalogram (EEG)-based driver fatigue detection. The review systematically analyzed 87 studies published between 2015 and 2025, focusing on signal preprocessing, feature extraction, and classification techniques. Commonly adopted classifiers such as

Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) networks were found to achieve accuracies ranging from 85% to 99%. The study emphasized that increased power in theta (4–8 Hz) and alpha (8–13 Hz) frequency bands is strongly correlated with fatigue onset. Moreover, the authors highlighted critical limitations, including inter-subject variability, real-time implementation challenges, and a lack of standardized EEG acquisition protocols, suggesting the need for adaptive and wearable EEG systems for practical applications (Hassan *et al.* 2025).

Unique Sensory Experiences and Neurophysiological Representations

The study by Shen *et al.* (2025) investigated the effects of acupuncture sensation on brain functional connectivity using EEG data and analyzed this sensory experience through graph theory metrics. Based on the obtained measures, acupuncture sensation was predicted with a mean absolute error (MAE) of 0.65%. EEG-based investigations conducted on individuals with high autistic traits revealed the neurophysiological foundations of social interaction difficulties and identified a negative correlation between low alpha coherence in the occipital region and AQ scores (Wang *et al.* 2025a).

Music, as a form of expression capable of eliciting strong emotional responses, enables the recognition of music-evoked emotions through EEG signals, contributing to a deeper understanding of the underlying neural mechanisms of such responses. Artificial intelligence plays a crucial role in this process by facilitating the extraction of characteristic frequencies (Su *et al.* 2024) and the identification of novel features, thereby enhancing the development of emotion recognition models.

The studies presented demonstrate that the analysis of EEG signals using AI-assisted approaches plays a significant role in the objective assessment of cognitive, emotional, and neurological states. Deep learning and graph theory-based methods reveal the versatile potential of EEG in clinical, rehabilitation, and human–computer interaction domains. In this context, the development of personalized EEG analyses and real-time systems represents a significant advancement in literature.

ARTIFICIAL INTELLIGENCE APPLICATIONS ON EEG SIGNALS

Several studies identified in the literature that utilize artificial intelligence methods are comparatively examined in the following section.

Machine Learning Method (ML)

Comprehensive reviews are also available in the literature regarding the role of machine learning (ML) algorithms in enhancing the processing, interpretation, accuracy, and effectiveness of EEG signals (Hosseini *et al.* 2021). The methods used in the study conducted by Joseph et al. (2025) are as follows:

Preprocessing: In this section, the researchers explain that the four-channel signals obtained from the MUSE EEG device were converted into microvolts using the BlueMuse and Muse-lsl software and then standardized by resampling them to 200 Hz through a Fourier transform. Each EEG recording, approximately one minute in duration, was segmented using a one-second "sliding window" with a 0.5-second overlap to capture temporal variations. Statistical features (such as skewness and variance) were extracted from these windows and subsequently used for signal analysis and emotional state classification.

The graph in Figure 6 illustrates the segmentation of the EEG signal over time for analysis. In other words, the continuous signal is divided into overlapping small segments (windows). Each segment (for instance, one second in length) serves as an analytical unit from which statistical features are extracted.

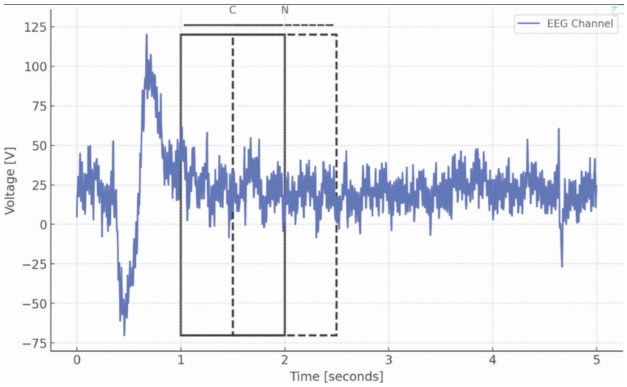


Figure 6 Sliding window approach (Barrowclough *et al.* 2025).

$$\tilde{\mu}_3 = \frac{\sum_i^N (x_i - \bar{x})^3}{(N-1) * \sigma^3} \quad (1)$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (2)$$

Overall, the data processing procedure involves standardizing the raw EEG signals, segmenting them into time windows, and extracting statistical features (such as skewness and variance) from each window. These features were then utilized in the affective classification model.

Feature Extraction: To capture different dimensions of emotions (e.g., valence and arousal), features were extracted from the signals in the time–frequency domain. Among the prominent techniques are methods such as Empirical Mode Decomposition (EMD) and Wavelet Transform.

Classification: For personalized models, the best results were obtained using Support Vector Machines (SVM) and Random Forest algorithms. The researchers also experimented with ensemble learning strategies to improve classification performance. According to the results, personalized models performed significantly better than general models.

Performance Criteria: Classical evaluation metrics such as accuracy, sensitivity, and specificity were used as performance measures. The SVM-based personalized model achieved an accuracy of up to 85%, which outperforms many general models reported in the literature.

Evaluation and Observation: Based on the assumption that each individual's EEG signal structure differs, the researchers investigated whether personalized models outperform generic ones. To this end, six machine learning algorithms were tested on the MUSE and DEAP datasets: K-NN, Decision Tree, Random Forest, SVM, Naïve Bayes, and Stacking Ensemble Classifier. The experiments aimed to enhance prediction performance by accounting for EEG features specific to each individual.

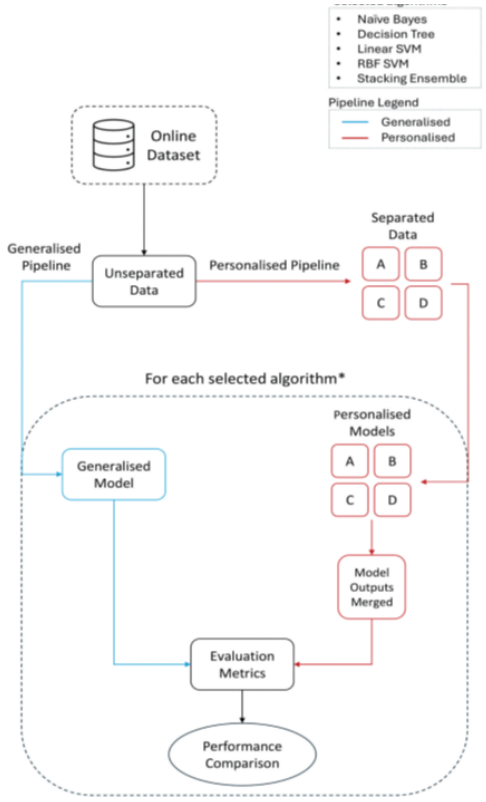


Figure 7 High level method pipeline (Barrowclough *et al.* 2025)

The experimental design of the study is visualized in Fig. 7. This diagram illustrates, step by step, how both the generic and personalised model pipelines are implemented with the machine learning algorithms. The fundamental mathematical foundations of the algorithms (e.g., Minkowski distance, Gini impurity, Bayesian distributions, SVM kernel functions) are also provided; during evaluation, confusion matrices were used and accuracy rates were compared.

Table 1 Summary of Methods and Techniques Used in the Study by Joseph et al.

Feature	Methods Used	Advantages	Disadvantages
Feature Extraction	EMD, Wavelet Transform	High time-frequency resolution	High computational cost
Classifier	SVM, Random Forest, Ensemble Methods	High accuracy with personalized modeling	Slow due to the need for training multiple models
Dataset	DEAP	Widely used, suitable for comparative analysis	Limited number of subjects
Model Structure	Personalized models	Able to capture individual differences	Not suitable for generalization

The methodological contribution of this study lies in demonstrating that a personalized modeling strategy can substantially enhance the performance of EEG-based emotion classification. Such approaches are particularly well-suited for applications such as user-specific neurofeedback systems, mental health monitoring, and human–computer interaction.

The methods used in the study conducted by Khare *et al.* (2023) are as follows:

Dataset and Preprocessing: In this part of the study, a publicly available EEG dataset from the Kaggle platform was utilized. Five participants performed experiments using a train simulator (Amtrak–Philadelphia route). The participants' mental states were categorized into three classes: Focused, Unfocused, and Drowsy. EEG recordings were collected using the Emotiv EPOC EEG system with a sampling frequency of 128 Hz, a bandwidth of 0.2–43 Hz, and a resolution of 0.51 μV. Each 10-minute session was divided into 30-second segments, resulting in 680 EEG segments per class.

The raw EEG signals were processed using an ensemble wavelet decomposition approach to remove noise and artifacts. In this process, three different wavelet-based methods were employed in combination:

- MDWT (Multilevel Discrete Wavelet Transform): It separates low- and high-frequency components.
- TQWT (Tunable Q Wavelet Transform): It performs parameter-based (q, R, B) decomposition without requiring the selection of a main wavelet.
- FAWT (Flexible Analytic Wavelet Transform): It analyzes complex signals such as EEG through two high-pass and one low-pass channels.
- The combination of these methods allows for precise analysis of the signals in both time and frequency domains.

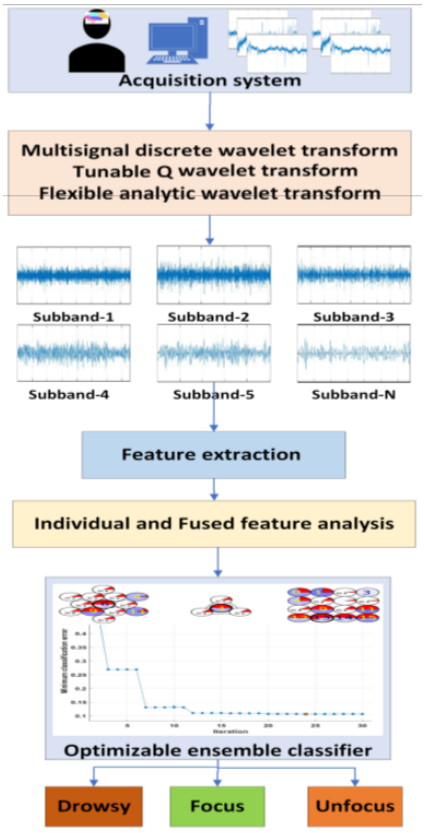


Figure 8 Proposed ensemble model for mental state detection (Khare *et al.* 2023)

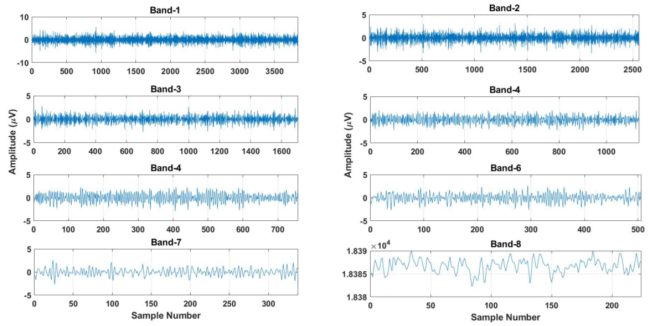


Figure 9 A typical example of SBs generated by the TQWT (Khare *et al.* 2023)

Feature Extraction and Wavelet-Based Decomposition: A total of 27 statistical, fractal, and nonlinear features were extracted from the decomposed EEG signals. These features include indicators such as mean energy, variance, skewness, Hurst exponent, Hjorth mobility, Higuchi fractal dimension, Lyapunov exponent, and zero-crossing rate. The extracted features represent the dynamics of brain activity and were utilized to differentiate between various mental states.

Classification: Optimized ensemble classifiers were employed in the study. The classifiers used included models such as boosted trees, bagged trees, ensemble discriminant, and subspace KNN. Feature fusion was applied to achieve dimensionality reduction and enhance performance.

During the training and testing phases, holdout, 5-fold, and 10-fold cross-validation techniques were implemented.

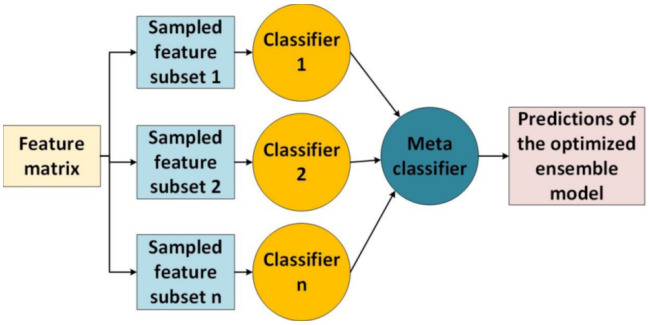


Figure 10 Typical working of ensemble classifier techniques (Khare *et al.* 2023)

Success Performance and Results:

- The best performance was achieved using FAWT, yielding 97.8% accuracy with the Iterative Majority Voting technique.
- In the three-class analysis (F, UF, D), FAWT-based features provided higher separability compared to other methods.
- Feature fusion achieved the highest F1 score (98.18%) particularly for the drowsy class.
- The best-performing subband (SB) levels were as follows:
  - B-1 for MDWT and TQWT
  - SB-7 for FAWT

Evaluation and Observation: This study demonstrates how the combination of wavelet transform techniques and optimized ensemble classification can yield robust results in EEG-based mental state classification. In particular, the high time–frequency resolution of FAWT has proven to be ideal for complex and oscillatory signals such as EEG.

Table 2 Summary of Methods, Features, and Classification Performance in the Study by Khare *et al.* (2023)

Feature	Methods Used	Advantages	Disadvantages
Feature Extraction	MDWT, TQWT, FAWT	Versatile time–frequency decomposition	Parameter tuning is complex
Feature Type	27 statistical/chaotic features	Rich info, high discriminative power	Large features → reduction may be required
Classifier	Ensemble (optimized)	High accuracy and generalization	High computational cost
Accuracy (with IMV)	97.8%	Superior performance in the literature	Limited number of subjects

Compared to previous studies, this work adopts an ensemble wavelet decomposition with feature fusion strategy rather than personalization. While personalized models focus on capturing individual variations, the advantage of the present study lies in providing a more generalizable classification model.

The two studies considered (Barrowclough *et al.* 2025; Khare *et al.* 2023) focus on the machine learning-based classification of mental states from EEG signals and present distinct methodological approaches. Although the objectives of both studies are similar, they demonstrate significant differences in the signal processing and classification strategies employed.

Modeling Approach:

- The first study, "Personalised Affective Classification Through Enhanced EEG Signal Analysis," focused on individual-specific emotion classification and developed personalized machine learning models (Khare *et al.* 2023). Training a separate model for each participant aimed to enhance classification accuracy by accounting for personal variations in EEG signals.
- The second study, "Ensemble Wavelet Decomposition-Based Detection of Mental States," aimed to establish a generalizable model and adopted an ensemble wavelet decomposition combined with an optimizable ensemble classifier approach rather than personalization (Ma *et al.* 2025). In this study, mental states such as focused, unfocused, and drowsy were classified into three distinct classes.

The first study emphasizes personalized models, whereas the second focuses on generalized models. While the former highlights individual adaptation, the latter underscores the power of signal processing and algorithm optimization.

Feature Extraction and Signal Processing Techniques:

- In the first study, features were extracted from EEG signals using conventional time–frequency decomposition methods, such as Empirical Mode Decomposition (EMD) and Wavelet Transform. These extracted features were used to model affective dimensions, such as valence and arousal.
- In the second study, a more sophisticated approach was adopted. Multilevel Discrete Wavelet Transform (MDWT),

Tunable Q Wavelet Transform (TQWT), and Flexible Analytic Wavelet Transform (FAWT) were used in combination to perform ensemble decomposition. Each of these techniques excels at decomposing different frequency bands and signal characteristics. A total of 27 statistical and chaotic features were extracted, and dimensionality reduction was achieved using a feature fusion method.

In the second study, the feature space is considerably richer and more diverse. In particular, the use of FAWT provides a significant advantage in analyzing the temporal complexity of EEG signals. The methods employed in the first study are comparatively more basic in nature.

Table 3 Machine learning based Comparative Performance and Methodological Analysis of Barrowclough *et al.* (2025) and Khare *et al.* (2023)

Feature	Study 1: Personalized Model	Study 2: Ensemble Wavelet Model
Modeling Approach	Individual-specific	Generalizable + optimized
Feature Extraction	EMD, Wavelet	MDWT, TQWT, FAWT
Classifier	SVM, RF	Ensemble + IMV
Accuracy	~85%	97.8%
Signal Type	DEAP dataset	Kaggle EEG
Strengths	Captures individual variations	Advanced decomposition + fusion
Limitations	Limited generalization	Small number of participants

Classification Methods:

- In the first study, Support Vector Machines (SVM) and Random Forest classifiers were primarily employed, and personalization increased the accuracy to approximately 85% (Khare *et al.* 2023).
- In the second study, optimized ensemble classifiers were preferred. A combination of methods, including bagged trees, boosted trees, and subspace KNN, was applied, and the final decision mechanism was implemented using Iterative Majority Voting (IMV). This framework increased the accuracy rate to 97.8%. The classification approach in the second study demonstrated superior performance due to its more advanced framework and inclusion of hyperparameter optimization. Despite the use of personalized models in the first study, the overall accuracy remained lower.

Deep Learning Method (DL)

In the literature review section, three key studies employing deep learning (DL) approaches for EEG-based classification problems were analyzed (Agarwal and Kumar 2024), and their methodological differences and performances were compared in detail. Prabhakar et al. in 2024, the methods employed are as follows:

Preprocessing:

- The preprocessing stage aimed both to enhance signal quality and to provide a suitable foundation for feature extraction.
- The signals were divided into segments of equal length to enable meaningful analysis.
- Each segment was time-normalized to ensure consistent learning by the model.
- The preprocessing stage aimed both to enhance signal quality and to provide a suitable basis for feature extraction.

Feature Extraction:

- The Hilbert Transform was applied to the EEG signals to obtain the envelopes of each signal.
- This transformation provided both instantaneous amplitude and phase information, offering a richer representation compared to classical time–frequency analysis methods.
- The resulting envelopes were converted into a suitable matrix format to serve as input for the classification model.
- In this way, both the neural and semantic components of the signals were successfully represented

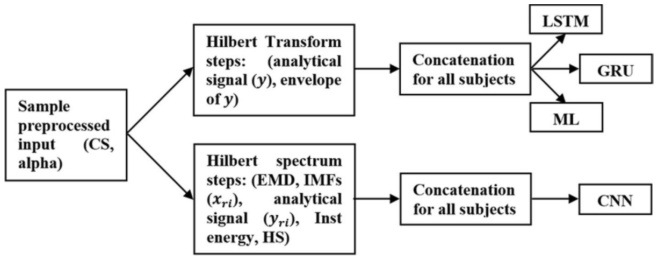


Figure 11 Detailed steps for feature extraction (Agarwal and Kumar 2024)

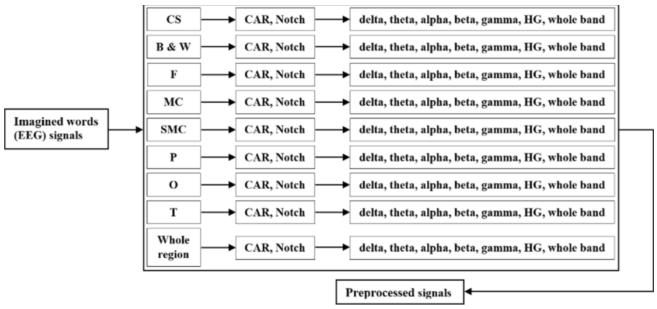


Figure 12 Detailed steps for preprocessing (Agarwal and Kumar 2024).

Classification:

- A deep neural network (DNN) based on a multilayer perception (MLP) was employed for classification.
- The model was trained under a supervised learning paradigm and optimized using the backpropagation algorithm.
- The network architecture enabled the deep processing of features, allowing for successful discrimination of imagined word tasks.

Performance Criteria:

- The model's performance was primarily evaluated in terms of accuracy. Additionally, its cross-subject generalizability was tested to assess the ability for participant-independent classification.

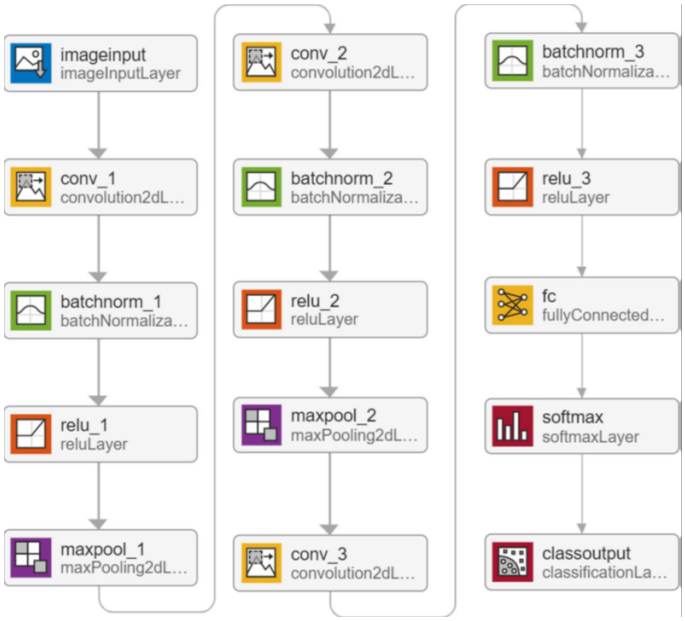


Figure 13 Proposed CNN architecture (Agarwal and Kumar 2024).

Table 4 Summary of Preprocessing, Feature Extraction, and Classification Methods Used by Remsik *et al.* (2022)

Feature	Methods Used	Advantages	Disadvantages
Preprocessing	Band-pass filtering, segmentation, time normalization	Signals cleaned from noise and artifacts; suitable for analysis	Filter parameters may not be universal due to individual variations
Feature Extraction	Envelope extraction via Hilbert Transform	Rich representation containing both phase and amplitude info	Computational load of Hilbert Transform can be high
Classification	MLP-based Deep Neural Network, backpropagation	Strong classification capability; parametric flexibility	Deep learning models require large datasets; risk of overfitting
Performance Evaluation	Accuracy, cross-subject generalizability tests	Effective performance across different individuals	Despite high accuracy, not yet tested in real-time applications

- The study demonstrated high performance with similar accuracy rates across different individuals. However, the applicability of the model in real-time systems has not yet been experimentally tested.
- The F1 score is not directly reported as a metric in the article content and tables; the study focuses more on accuracy and Network Prediction Time (NPT).
- Accuracy Percentage:
  - Maximum Accuracy: 94.29% was achieved.
  - Average Accuracy: An average accuracy of 71.75% was achieved throughout the study.

In the study conducted by (Dairi *et al.* 2022), the methods employed are as follows:

Preprocessing:

- In this section of the study, the Multi-Channel Wiener Filter (MWF) algorithm was applied for artifact removal. MWF has been shown to perform effectively on both real and hybrid EEG data, offering superior performance compared to conventional techniques.
- The fundamental principle of the algorithm involves modifying the artifact covariance matrix using a low-rank approximation and processing it through generalized eigenvalue decomposition. This approach provides a generalizable and robust solution for the removal of various types of EEG artifacts.

Feature Extraction:

- In this study, conventional manual feature extraction was not performed. Instead, the task of feature extraction was directly handled by the encoder layers of the model.
- In particular, the convolutional layers (CNN) acted as a structure that automatically learned the spatial and temporal patterns of the EEG signals to extract features. This approach aligns with the deep learning capability to reduce the need for manual feature engineering.
- Consequently, the data was fed directly into the neural network, allowing the model itself to generate meaningful representations without relying on hand-crafted features.

Classification:

- The novelty of the study lies in employing an anomaly detection approach instead of conventional classifiers. In this context, the developed system is based on an autoencoder architecture. The encoder–decoder model aims to summarize and reconstruct the input signal.
- During training, the model was allowed to learn only "normal" mental tasks. In the testing phase, reconstruction errors were measured to identify "anomalous" patterns. This approach is particularly effective for high-variance data such as EEG signals, providing both the ability to work with unlabeled data and robustness to individual differences.

Performance Criteria:

- The success of the DBN-iF model proposed in this study has proven its superiority compared to other results in the literature:
  - DBN-iF (This study): 98.5%
  - LS-SVM (Closest competitor): 97.56%
  - KNN: 92.8%
  - CNN-SAE: 90.0%
- The model's performance was evaluated primarily based on reconstruction error rather than conventional accuracy metrics. Anomalous events were identified when this error exceeded a predefined threshold, enabling functional output even on unlabeled samples.
- Experimental results demonstrate the system's high capacity both for discriminating between different tasks and for adapting to inter-subject variations.However, the absence of

Table 5 Summary of Preprocessing, Feature Extraction, and Classification Methods Used by (Dairi *et al.* 2022)

Feature	Methods Used	Advantages	Disadvantages
Preprocessing	Multi-Channel Wiener Filter (MWF); low-rank covariance; eigenvalue decomposition	Superior denoising on real/hybrid EEG; robust to multiple artifacts	More complex than classical filtering; requires expertise
Feature Extraction	Automatic extraction via CNN layers; no manual features	Eliminates manual engineering; learns spatial-temporal patterns	Requires large data; less interpretable than handcrafted features
Classification / Detection	Autoencoder-based anomaly detection; reconstruction error	Learning from unlabeled data; robust to inter-subject variability	Lack of softmax makes direct comparison difficult; threshold selection issues
Performance Evaluation	Reconstruction error and anomaly thresholding	Enables detection without labels; shows generalization	No classical metrics (accuracy, etc.) reported; limited comparability

comparison using traditional metrics such as class-wise accuracy limits the direct comparability of these results with conventional systems.

- According to the findings of this article, combining QTFD (Quadrificial Time-Frequency Distribution) feature extraction with DBN-based Anomaly Detector (Isolation Forest) can achieve higher accuracy in recognizing mental tasks compared to traditional supervised learning models (such as CNN or SVM).

In the study conducted by Khondoker Murad et al. (Hossain *et al.* 2023), the techniques employed are as follows:

Preprocessing: The raw EEG signals inherently contain noise, artifacts, and low signal-to-noise ratios, which complicate their direct use in BCI applications. Therefore, among the prominent preprocessing methods highlighted in the studies reviewed, the following are commonly employed:

- Frequency filtering (particularly band-pass filtering between 0.5–45 Hz).
- Artifact removal (especially signals arising from eye blinks and muscle movements).
- Manual cleaning and methods such as Independent Component Analysis (ICA) and Discrete Wavelet Transform (DWT).

Feature Extraction. One of the main advantages of deep learning is its ability to eliminate the need for manually defined feature extraction. In the studies reviewed, feature extraction was primarily performed automatically by the initial layers of CNN, RNN, LSTM, or even Transformer architectures.

However, some hybrid approaches employed time–frequency representations to support model learning, applying transformations such as STFT, CWT, or Morlet wavelets as a pre-processing step. Notably, in tasks such as motor imagery or emotion recognition, these rich representations were observed to significantly enhance classification performance.

Classification: In the review, various deep learning architectures employed for EEG-based BCI applications are presented in detail. The most frequently used classifiers include:

- Convolutional Neural Networks (CNN) – effective for capturing spatial features.
- Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) – suitable for temporal signals such as EEG.
- Generative Adversarial Networks (GAN) – employed for data augmentation and balancing.
- Transformer architectures – offer potential for modeling long-range dependencies.

Performance Criteria: In this study, deep learning models demonstrated different success levels depending on the application type:

- Driver Fatigue Detection: Studies in this area have resulted in high accuracy rates between 83% and 98%.
- Epileptic Seizure Detection: Accuracy rates of over 99% have been reported using CNN and RNN models.
- Emotion Recognition: Most researchers have achieved accuracy rates above 90% in datasets such as DEAP and SEED.
- Motor Imagery (MI): While classical machine learning methods struggle in this area (between 72% and 86%), deep learning (especially LSTM and CNN) has significantly improved performance.

Table 6 Summary of Preprocessing, Feature Extraction, and Classification Methods Used by (Hossain *et al.* 2023)

Feature	Methods Employed	Advantages	Limitations
Preprocessing	Frequency filtering, artifact removal (ICA, DWT), manual visual inspection	Noise reduction and improved signal quality	Standardization is challenging due to methodological diversity
Feature Extraction	CNN, LSTM, Transformer architectures; time–frequency transformations (CWT, STFT)	Automatic learning instead of manual extraction; rich representations	Loss of interpretability and high computational cost in some models
Classification	CNN, LSTM, GAN, Transformer; hybrid architectures (e.g., CNN+LSTM)	High accuracy with task-specific selection; captures patterns	Model selection depends on task type; no universal architecture
Performance Evaluation	Accuracy, F1-score, sensitivity, specificity, cross-validation	Multi-dimensional analysis; improved generalization via transfer learning	Direct comparison is difficult due to differences in datasets

Additionally, some approaches employed cross-validation strategies to assess generalizability. It has also been highlighted that increasing the amount of data, as well as applying domain adaptation and transfer learning techniques, has a significant impact on accuracy. However, the lack of methodological consistency among the datasets used has led to inconsistencies in comparative results.

Comparison of Deep Learning-Based Studies

These three studies collectively illustrate the multidimensional utilization of deep learning techniques in EEG signal analysis. The first study introduces a CNN-based framework focusing on a specific task (imagined words), whereas the second emphasizes an anomaly detection paradigm rather than direct classification. The third article provides a higher-level synthesis encompassing multiple studies, underscoring that the selection of an appropriate methodological approach should be determined contextually, depending on the experimental design and target application.

Explanations of Technical Terms:

- Hilbert Transform: An analytical signal approach used to extract the temporal amplitude and phase components of EEG signals. It enhances temporal resolution.
- Autoencoder: An unsupervised learning architecture that identifies anomalies by learning low-dimensional representations of data and minimizing reconstruction error.
- STFT / CWT / Wavelet Transforms: Fundamental time-frequency transformation techniques used in EEG signal analysis. They facilitate the generation of more meaningful inputs for deep learning models.
- Domain Adaptation: A subfield of transfer learning employed to maintain model consistency across EEG data collected from different individuals.

Similarities and Differences Between Methods:

- CNN architecture plays a significant role in all three studies. In the first study, it is employed directly for classification purposes, whereas in the second study, it is incorporated within an autoencoder structure for reconstruction. In the third study, CNN is discussed within the context of a literature review as a commonly used method in EEG applications.
- Noticeable differences emerge in terms of pre-processing techniques. The first study enriches the data using advanced signal processing methods such as the Hilbert transform, whereas the second study employs simpler, more basic preprocessing procedures. This distinction clearly demonstrates the impact of data preparation methods on the resulting model performance.
- In terms of learning paradigms, the first study is based on supervised learning and utilizes labeled data. In contrast, the second study employs an unsupervised learning approach applied to unlabeled datasets. The third study compares these two methods and discusses which approach may be more appropriate depending on the type of application.
- Advanced techniques such as generalizability and transfer learning are discussed in detail particularly in the third study. Since EEG signals exhibit substantial inter-individual variability, these strategies are crucial for ensuring that models remain effective across different subjects. In contrast, the other two studies address this topic in a more limited manner.

Results: Although these three studies serve different objectives, they collectively demonstrate that deep learning–based analyses of EEG signals encompass a broad methodological spectrum. The common use of CNN-based architecture indicates that such models can effectively capture and represent spatial patterns within EEG data. However, the adaptation of these architectures to dataset-specific characteristics, such as preprocessing strategies, labeling schemes, and levels of architectural complexity, constitutes a critical factor influencing application performance. As a consequence of this methodological diversity, models such as autoencoders tend to offer advantages in scenarios emphasizing unsupervised learning, whereas CNN-based models enriched with information-dense inputs yield more efficient outcomes in contexts requiring linguistic imagery or complex classification.

Studies Utilizing Alternative Modeling Approaches

The methods and techniques used in the studies of Al-Dabag and Ozkurt (Al-dabag and Ozkurt 2019),

Preprocessing:

- Artifact Removal (EEG Subtraction): Noise was suppressed by subtracting the resting-state (motionless) EEG signal from the EEG signal containing movement. This procedure was applied to enhance signal clarity and make it suitable for classification.
- Channel Selection: EEG channels located near the motor cortex were pre-identified and selected. In particular, the F3 and F4 channels were utilized as "effective channels.

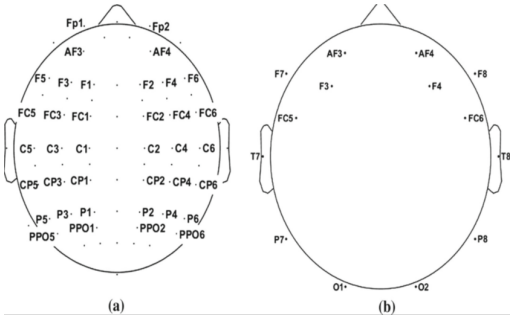


Figure 14 Selected EEG channels in both datasets. (A) Selected channels of BCI datasets, (b) selected channels of Emotiv datasets (Al-dabag and Ozkurt 2019).

Feature Extraction:

- Discrete Wavelet Transform (DWT): EEG signals were decomposed into beta and gamma bands using the DWT method. This decomposition allowed the extraction of frequency components related to movement.
- Cross-Correlation: The selected effective channels (F3, F4) were subjected to cross-correlation with other channels in the right and left hemispheres, thereby extracting information on similarity/synchronization between two channels.
- Statistical Features: From the cross-correlation outputs, 10 normalized statistical features were calculated, including minimum, maximum, mean, median, mode, standard deviation, variance, entropy, and the first and third quartiles.

Classification:

- Artificial Neural Network (ANN): A multilayer ANN was implemented. The number of nodes in the hidden layer was optimized through multiple trials (e.g., 14 nodes yielded the best performance).

Table 7 Comparative Deep Learning-Based EEG Applications

Criterion / Study	Imagined Words Classification (Hilbert Transform + CNN)	Mental Task Recognition (Autoencoder-Based Anomaly Detection)	Comprehensive Literature Review (Multiple Deep Learning Architectures)
Application Domain	Recognition of mentally imagined words (*Imagined Speech BCI*)	Identification of different cognitive tasks (*Mental Task Recognition*)	General methodological categorization for EEG-based BCI systems
Preprocessing Approach	Extraction of instantaneous amplitude and phase components using the Hilbert Transform	Z-norm–based standardization and basic spectral filtering	Time–frequency transformations (STFT, CWT, Wavelet Decompositions)
Modeling Architecture	Deep Convolutional Neural Network (Deep CNN)	Encoder–decoder–based Autoencoder structure (CNN embedded)	Advanced DL architectures such as CNN, LSTM, GAN, and Transformer
Methodological Contribution / Innovation	Enriched input representation via Hilbert Transform for meaningful signal encoding	Mental state differentiation through unsupervised detection of anomalous patterns	Architecture–application mapping and architecture recommendations based on task type
Data Labeling Requirement	Moderate (requires labeled data for supervised learning)	Low (suitable for unsupervised or anomaly detection tasks)	High (requires extensively labeled data or transfer learning scenarios)
Generalizability / Subject Independence	Cross-participant generalizability tested	Proposed structure tolerant to participant variability	Discussion of domain adaptation and transfer learning strategies
Highest Accuracy	%94.29	%98.50	99.00%+ (Epilepsy)
Average / Other Results	Average: 71.75%	LS-SVM: %97.56, KNN: %92.8, CNN-SAE: %90.0	Driver Fatigue: 83–98%, Emotion Recognition: >90%, Motor Imagery: 72–86%

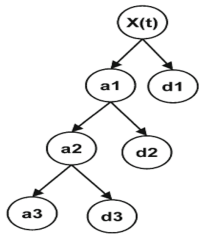


Figure 15 Three-level DWT decomposition (Al-dabag and Ozkurt 2019)

- Support Vector Machine (SVM): An SVM with a Radial Basis Function (RBF) kernel was employed. The kernel scale was determined automatically.
- Datasets: Classification was performed using both the BCI Competition III Dataset IVa (imagined movements) and Emotiv Epoc+ data (actual movements).

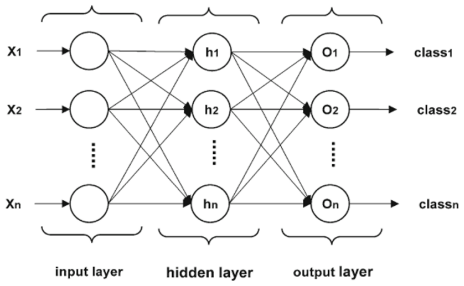


Figure 16 ANN architecture (Al-dabag and Ozkurt 2019)

Performance Criteria:

- Accuracy: An accuracy of 99.33% was achieved with the ANN, and 99.69% with the SVM (BCI dataset).
- Cross-Validation (10-Fold): A 10-fold cross-validation was performed for both classifiers to assess their validity.
- User Independence: Both subject-specific (patient-based) and pooled-subject (movement-based) classifications were tested. In movement-based classification, accuracy did not drop below 92%.

The methods and techniques used in the studies of Leela and colleagues (Leela and Helenprabha 2025),

Preprocessing: Since different data modalities (temporal, spectral, spatial, audio, and text data) were used for Alzheimer's disease prediction, specific preprocessing procedures were applied for each data type:

- Temporal (gait) data: Erroneous and missing data were removed using statistical methods, and outliers were eliminated based on standard deviation thresholds.
- Spectral (EEG) data: Signals were filtered within the 1–45 Hz band and subsequently prepared for time-frequency transformation.
- Spatial (MRI/PET/CT) data: Imaging data underwent size standardization and resampling procedures.
- Audio data: Noise reduction and normalization were performed.
- Text data: After converting audio data to text, tokenization, stop-word removal, and vectorization were carried out.

Feature Extraction: The study employed deep learning models that extract distinctive representations from multimodal datasets:

- Temporal (gait) data: A BiLSTM (Bidirectional Long Short-Term Memory) network was used for the time series data.

Table 8 Summary of Preprocessing, Feature Extraction, and Classification Methods Used by (Al-dabag and Ozkurt 2019)

Feature	Methods Used	Advantages	Disadvantages
Preprocessing	Resting-state EEG subtraction, channel selection	Noise was suppressed, resulting in cleaner signals. Focus on motor cortex regions was enhanced.	Manual channel selection may require expert knowledge. Performance may depend on recording conditions.
Feature Extraction	Frequency decomposition using DWT, cross-correlation, statistical parameter extraction	Provided meaningful and effective features with low computational cost. A comprehensive feature set was created.	Features were intuitively extracted from the signal; additional optimization may be required for universal generalization.
Classification	ANN (Levenberg-Marquardt), SVM (RBF kernel)	Tested with two different classifiers, achieving high accuracy. Comparison between ANN and SVM was conducted.	ANN results show run-to-run variability; parameter tuning may require careful adjustment.
Performance Evaluation	98–99+% accuracy, 10-fold cross-validation, two datasets (BCI and Emotiv), subject-specific and pooled classification	Method generalizability was tested, achieving success across different individuals and devices. High performance achieved with a simple approach.	Slight drop in accuracy observed in pooled classification scenarios (∼92%); individual differences may have a stronger effect.

- Spectral (EEG) data: Frequency-dynamic features were extracted from EEG images using the GoogLeNet convolutional neural network.
- Spatial data: Imaging data were analyzed using the MobileNet architecture, achieving high-accuracy representations with low computational cost.
- Audio data: Audio files were analyzed using a CNN-MLP (convolutional neural network combined with a multilayer perceptron).
- Text data: Semantic information was captured using the BERT (Bidirectional Encoder Representations from Transformers) model.

A two-level data fusion approach was implemented:

- Level 1: Temporal, spatial, and spectral data were combined through a gating mechanism to generate an intermediate representation.
- Level 2: This intermediate representation was then integrated with audio and text representations to obtain the final fusion vector.

Classification: The final feature vector was given to the Incremental Learner Ensemble Classifier (TMDFILE) model, which was specifically developed for Alzheimer's prediction. The main characteristics of this model are as follows:

- Ensemble structure: It combines different classifiers (e.g., decision trees, SVM) to enable more robust decision-making.
- Incremental learning: The model can be updated with newly incoming data; in other words, it is not static but has an adaptive learning process.
- Gating mechanism: It assigns weights by determining which modality is more effective in the classification process.

Performance Metrics: ADNI, OASIS, EEG Emotion, Aberystwyth Dementia, and BRATS. The obtained performance metrics:

- Accuracy: %94.5
- Precision: %93.5
- Recall: %95.1
- F1 Score: %94.1

Table 9 Summary of Preprocessing, Feature Extraction, and Classification Methods Used by (Leela and Helenprabha 2025)

Feature	Methods Used	Advantages	Disadvantages
Data Types	Temporal, spectral, spatial, audio and text data	Integrated analysis of disease dimensions	Complex data collection and synchronization
Preprocessing	Band-pass filtering, outlier removal, resampling	Modality-specific cleaning improves quality	Long processing time; high technical expertise
Feature Extraction	BiLSTM, GoogLeNet, MobileNet, CNN-MLP, BERT	Optimized DL architectures for each modality	High hardware requirements; long training time
Data Fusion Approach	Two-level: hierarchical fusion with gating	Optimized modality contribution	Complex structure; limited explainability
Classification Method	Incremental Learner Ensemble (TMDFILE)	Updatable, adaptive system; high accuracy	Needs additional management for real-time
Performance Metric	Acc: 94.5%, Prec: 93.5%, Recall: 95.1%, F1: 94.1%	High success rates across datasets	High system complexity for integration
Clinical App.	Multimodal neurological imaging system	Evaluates cognitive and motor symptoms	Access to all data sources may not be feasible

Comparative Analysis in Terms of Methods and Techniques Al-Dabag and Ozkurt (2018) proposed a simple and low-cost method that can be integrated into online BCI systems for motor movement classification based on EEG data. By utilizing DWT, cross-correlation, and statistical features, high accuracy was achieved without the need for complex optimization processes, and classical models such as ANN and SVM were employed for classification. On the other hand, Leela et al. (2025) introduced an incremental learning-based system called TMDFILE, which integrates EEG, MRI, speech, text, and gait data for Alzheimer's diagnosis. This approach is grounded in multidisciplinary data fusion. The two studies differ significantly in terms of data diversity, scale, and intended application.

Table 10 Comparative Analysis of EEG-Based Studies by (Al-dabag and Ozkurt 2019) and (Leela and Helenprabha 2025)

Comparison	Al-Dabag & Ozkurt (2018)	Leela et al. (2025)
Application Area	EEG-based motor movement recognition, BCI systems	Alzheimer's disease diagnosis, multi-modal biomedical data
Data Type	EEG only (BCI Dataset & Emotive)	EEG, MRI/PET, gait, audio, and text data
Preprocessing	EEG subtraction, channel selection	Modality-specific preprocessing: filtering, normalization, resampling, etc.
Feature Extraction	DWT, cross-correlation, 10 statistical features	BiLSTM, GoogLeNet, MobileNet, CNN-MLP, BERT, two-level data fusion
Classification Algorithms	Artificial Neural Network (ANN), Support Vector Machine (SVM)	TMDFILE (Ensemble & Incremental Learning Classifier)
AI Level	Traditional ML techniques	Advanced deep learning and incremental learning
Fusion Usage	None	Two-level data fusion (EEG + MRI → + audio + text)
Accuracy	99.33% (ANN), 99.69% (SVM)	94.5% on average (across five datasets)
Advantages	Simple, low computational cost, suitable for online BCI applications	Comprehensive modality analysis, adaptive learning, clinical generalizability
Disadvantages	EEG-only usage → limited information representation, sensitivity to individual data	High computational and data requirements, system complexity
Generalizability	Limited (restricted to BCI systems)	High (broad clinical application with data from multiple sources)

CONCLUSION

This review systematically classified the literature from the past decade on AI-based EEG analysis and provided comprehensive methodological comparisons. The findings highlight that personalized machine learning models, supported by advanced signal representation techniques such as Hilbert-based analytic signal extraction and time–frequency decompositions using STFT, CWT, and wavelet transforms, significantly improve sensitivity to individual neural dynamics. In parallel, representation learning approaches based on autoencoders contribute to robust feature extraction and dimensionality reduction in high-dimensional EEG data. Moreover, advanced multichannel signal enhancement techniques, particularly the Multi-Channel Wiener Filter (MWF), play a critical role in artifact suppression, signal-to-noise ratio improvement, and preservation of physiologically meaningful neural information, thereby strengthening the robustness of downstream machine learning and deep learning models. Deep learning architectures such as CNN, LSTM, and Transformer networks demonstrate strong capability in capturing both spatial and temporal dependencies inherent in multichannel EEG recordings.

Beyond conventional clinical diagnosis and monitoring, EEG-based AI systems are increasingly applied in interdisciplinary domains, including emotional state recognition, human–computer interaction, and driving safety. In this context, incremental learning frameworks combined with multimodal data fusion show promising potential for the early detection of complex neurological disorders such as Alzheimer's disease. Looking forward, the more systematic integration of domain adaptation and transfer learning strategies that explicitly address inter-subject and cross-session variability is strongly recommended. Additionally, the development of low-latency, optimized software–hardware co-design solutions remains essential for real-time EEG applications. Finally, the adoption of ethical data practices and explainable AI frameworks is critical to ensure the safe, transparent, and clinically reliable translation of EEG–AI technologies into real-world healthcare systems.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Availability of data and material

Available upon request.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Abgeena, A. and S. Garg, 2025 Unravelling emotions: exploring deep learning approaches for eeg-based emotion recognition with current challenges and future recommendations. Cognitive Neurodynamics **19**.

Agarwal, P. and S. Kumar, 2024 Eeg-based imagined words classification using hilbert transform and deep networks. Multimedia Tools and Applications **83**: 2725–2748.

Aggarwal, S. and N. Chugh, 2022 Review of machine learning techniques for eeg based brain computer interface. Archives of Computational Methods in Engineering **29**: 3001–3020.

Agrawal, R., C. Dhule, G. Shukla, S. Singh, U. Agrawal, *et al.*, 2024 Design of eeg based thought identification system using emd & deep neural network. Scientific Reports **14**: 26621.

Ahn, S., T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun, 2016 Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous eeg, ecg, and fnirs data. Frontiers in human neuroscience **10**: 219.

Al-dabag, M. L. and N. Ozkurt, 2019 Eeg motor movement classification based on cross-correlation with effective channel. Signal, Image and Video Processing **13**: 567–573.

Albayrak-Kutlay, Y. and M. Bengisu, 2025 Exploring vr and neuroscience methodologies in interior design: A systematic review. Human Behavior and Emerging Technologies **2025**.

Alshehri, H., A. Al-Nafjan, and M. Aldayel, 2025 Decoding pain: A comprehensive review of computational intelligence meth-

ods in electroencephalography-based brain-computer interfaces. Diagnostics **15**.

Barrowclough, J., N. Nnamoko, and I. Korkontzelos, 2025 Personalised affective classification through enhanced eeg signal analysis. Applied Artificial Intelligence **39**: 2450568.

Dairi, A., N. Zerrouki, F. Harrou, and Y. Sun, 2022 Eeg-based mental tasks recognition via a deep learning-driven anomaly detector. Diagnostics **12**: 2984.

Edelman, B., S. Zhang, G. Schalk, P. Brunner, G. Müller-Putz, *et al.*, 2025 Non-invasive brain-computer interfaces: State of the art and trends. IEEE Reviews in Biomedical Engineering **18**: 26–49.

Elnaggar, K., M. El-Gayar, and M. Elmogy, 2025 Depression detection and diagnosis based on electroencephalogram (eeg) analysis: A systematic review. Diagnostics **15**.

Gu, C., X. Jin, L. Zhu, H. Yi, H. Liu, *et al.*, 2025 Cross-session ssvep brainprint recognition using attentive multi-sub-band depth identity embedding learning network. Cognitive Neurodynamics **19**: 15.

Gu, X., Z. Cao, A. Jolfaei, P. Xu, D. Wu, *et al.*, 2021 Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. IEEE/ACM transactions on computational biology and bioinformatics **18**: 1645–1666.

Gurmessa, D. and W. Jimma, 2025 A comprehensive evaluation of interpretable artificial intelligence for epileptic seizure diagnosis using an electroencephalogram: A systematic review. Digital Health **11**.

Han, C.-H., Y.-W. Kim, D. Y. Kim, S. H. Kim, Z. Nenadic, *et al.*, 2019 Electroencephalography-based endogenous brain–computer interface for online communication with a completely locked-in patient. Journal of neuroengineering and rehabilitation **16**: 1–13.

Hassan, J., S. Naziullah, M. Rashid, T. Islam, M. Islam, *et al.*, 2025 Current status and challenges in electroencephalography (eeg)-based driver fatigue detection: a comprehensive survey. Cognitive Neurodynamics **19**.

Hossain, K. M., M. A. Islam, S. Hossain, A. Nijholt, and M. A. R. Ahad, 2023 Status of deep learning for eeg-based brain–computer interface applications. Frontiers in computational neuroscience **16**: 1006763.

Hosseini, M., A. Hosseini, and K. Ahi, 2021 A review on machine learning for eeg signal processing in bioengineering. IEEE Reviews in Biomedical Engineering **14**: 204–218.

Jain, A., R. Raja, M. Kumar, and P. K. Verma, 2025 A novel classification of meditation techniques via optimised chi-squared 1d-cnn method based on complexity, continuity and connectivity features. Connection Science **37**: 2467387.

Jeong, J.-H., B.-W. Yu, D.-H. Lee, and S.-W. Lee, 2019 Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional lstm network using electroencephalography signals. Brain sciences **9**: 348.

Jiang, R., X. Zheng, J. Sun, L. Chen, G. Xu, *et al.*, 2025 Classification for alzheimer's disease and frontotemporal dementia via resting-state electroencephalography-based coherence and convolutional neural network. Cognitive Neurodynamics **19**: 46.

Joo, E., H. Altier, C. Selai, M. Gratton, A. Kim-Dahl, *et al.*, 2025 Neurobiological mechanisms of sleep state misperception in insomnia disorder: A theoretical review. Sleep Medicine Reviews **81**.

Kamińska, D., K. Smółka, and G. Zwoliński, 2021 Detection of mental stress through eeg signal in virtual reality environment. Electronics **10**: 2840.

Kawala-Sterniuk, A., N. Browarska, A. Al-Bakri, M. Pelc, J. Zygarlicki, *et al.*, 2021 Summary of over fifty years with brain-computer interfaces-a review. Brain Sciences **11**.

Khan, S., S. M. Umar Saeed, J. Frnda, A. Arsalan, R. Amin, *et al.*, 2024 A machine learning based depression screening framework using temporal domain features of the electroencephalography signals. Plos one **19**: e0299127.

Khare, S. K., V. Bajaj, N. B. Gaikwad, and G. R. Sinha, 2023 Ensemble wavelet decomposition-based detection of mental states using electroencephalography signals. Sensors **23**: 7860.

Khlief, M. and A. Idrees, 2023 A comprehensive review of electroencephalography data analytics. International Journal of Computer Applications in Technology **71**: 78–88.

Kimmatkar, N. V. and B. V. Babu, 2021 Novel approach for emotion detection and stabilizing mental state by using machine learning techniques. Computers **10**: 37.

Kouba, P., M. Smotek, T. Tichy, and J. Koprivová, 2023 Detection of air traffic controllers? fatigue using voice analysis - an eeg validation study. International Journal of Industrial Ergonomics **95**.

Leela, M. and K. Helenprabha, 2025 Incremental learning based two-level multimodal data fusion model for alzheimer disease prediction on different data modalities. Connection Science **37**: 2458501.

Li, L. and W. Chen, 2025 Cross-modal alignment and fusion of eeg-visual based on mixed attention mechanism for emotion recognition. Cognitive Neurodynamics **19**.

Liu, X., W. Wang, M. Liu, M. Chen, T. Pereira, *et al.*, 2025 Recent applications of eeg-based brain-computer-interface in the medical field. Military Medical Research **12**.

Ma, Y., K. Karako, P. Song, X. Hu, and Y. Xia, 2025 Integrative neurorehabilitation using brain-computer interface: From motor function to mental health after stroke. Bioscience Trends .

Maiseli, B., A. Abdalla, L. Massawe, M. Mbise, K. Mkocha, *et al.*, 2023 Brain-computer interface: trend, challenges, and threats. Brain Informatics **10**.

Mendivil Sauceda, J. A., B. Y. Marquez, and J. J. Esqueda Elizondo, 2024 Emotion classification from electroencephalographic signals using machine learning. Brain Sciences **14**: 1211.

Mouazen, B., A. Bendaouia, E. Abdelwahed, and G. De Marco, 2025 Machine learning and clinical eeg data for multiple sclerosis: A systematic review. Artificial Intelligence in Medicine **166**.

Nandakumar, R., S. Deivanayagi, S. A. Kirubha, and R. Prabu, 2025 Recognizing emotions from physiological data in a eeg signals using a novel deep learning technique. Circuits, Systems, and Signal Processing pp. 1–24.

Orban, M., M. Elsamanty, K. Guo, S. Zhang, and H. Yang, 2022 A review of brain activity and eeg-based brain–computer interfaces for rehabilitation application. Bioengineering **9**: 768.

Pacia, S., 2023 Sub-scalp implantable telemetric eeg (site) for the management of neurological and behavioral disorders beyond epilepsy. Brain Sciences **13**.

Puri, D. V., J. P. Gawande, P. H. Kachare, and I. Al-Shourbaji, 2025 Optimal time-frequency localized wavelet filters for identification of alzheimer's disease from eeg signals. Cognitive Neurodynamics **19**: 12.

Remsik, A., P. van Kan, S. Gloe, K. Gjini, L. Williams, *et al.*, 2022 Bci-fes with multimodal feedback for motor recovery poststroke. Frontiers in Human Neuroscience **16**.

Shafiezadeh, S., G. Duma, M. Pozza, and A. Testolin, 2024 A systematic review of cross-patient approaches for eeg epileptic seizure prediction. Journal of Neural Engineering **21**.

Shang, S., Y. Shi, Y. Zhang, M. Liu, H. Zhang, *et al.*, 2024 Arti-

ficial intelligence for brain disease diagnosis using electroencephalogram signals. Journal of Zhejiang University-Science B **25**: 914–940.
Shen, D., B. Yang, J. Li, J. Zhang, Y. Li, *et al.*, 2025 The potential associations between acupuncture sensation and brain functional network: a eeg study. Cognitive Neurodynamics **19**: 1–14.
Sobhani, M., S. Srestha, M. Shomoy, A. Reza, and N. Siddique, 2025 A machine learning-based eeg signal analysis framework to enhance emotional state detection. Cognitive Neurodynamics **19**.
Soufineyestani, M., D. Dowling, and A. Khan, 2020 Electroencephalography (eeg) technology applications and available devices. Applied Sciences **10**.
Sozer, A. and C. Fidan, 2017 Novel detection features for ssvep based bci: Coefficient of variation and variation speed. BrainBroad Research in Artificial Intelligence and Neuroscience **8**: 144–150.
Staffa, M., L. D'Errico, S. Sansalone, and M. Alimardani, 2023 Classifying human emotions in hri: applying global optimization model to eeg brain signals. Frontiers in Neurorobotics **17**: 1191127.
Su, Y., Y. Liu, Y. Xiao, J. Ma, and D. Li, 2024 A review of artificial intelligence methods enabled music-evoked eeg emotion recognition and their applications. Frontiers in Neuroscience **18**.
Sun, H., S. Mao, W. Cai, Y. Cui, D. Chen, *et al.*, 2025 Bisnn: bioinformation-fused spiking neural networks for enhanced eeg-based emotion recognition. Cognitive Neurodynamics **19**: 1–12.
Sözer, A. and C. Fidan, 2018 Novel spatial filter for ssvep-based bci: A generated reference filter approach. Computers in Biology and Medicine **96**: 98–105.
Sözer, A. T. and C. B. Fidan, 2019 Emotiv epoc ile durağan hal görsel uyarılmış potansiyel temelli beyin bilgisayar arayüzü uygulaması. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi **8**: 158–166.
Tautan, A., A. Andrei, C. Smeralda, G. Vatti, S. Rossi, *et al.*, 2025 Unsupervised learning from eeg data for epilepsy: A systematic literature review. Artificial Intelligence in Medicine **162**.
Turi, F., M. Clerc, and T. Papadopoulo, 2021 Long multi-stage training for a motor-impaired user in a bci competition. Frontiers in human neuroscience **15**: 647908.
Uyanik, H., A. Sengur, M. Salvi, R. Tan, J. Tan, *et al.*, 2025 Automated detection of neurological and mental health disorders using eeg signals and artificial intelligence: A systematic review. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery **15**.
Wan, X., K. Zhang, S. Ramkumar, J. Deny, G. Emayavaramban, *et al.*, 2019 A review on electroencephalogram based brain computer interface for elderly disabled. IEEE Access **7**: 36380–36387.
Wang, J., L. Zhang, S. Chen, H. Xue, M. Du, *et al.*, 2025a Individuals with high autistic traits exhibit altered interhemispheric brain functional connectivity patterns. Cognitive Neurodynamics **19**: 9.
Wang, X., D. Wu, and C. Yang, 2025b Localization of epileptic foci from intracranial eeg using the gru-gc algorithm. Brain Informatics **12**: 6.
Yan, N., C. Wang, Y. Tao, J. Li, K. Zhang, *et al.*, 2020 Quadcopter control system using a hybrid bci based on off-line optimization and enhanced human-machine interaction. ieee access 8, 1160–1172 (2020).
You, S., 2021 Classification of relaxation and concentration mental states with eeg. information, 12 (5), 187.

How to cite this article: Akdeniz, A. H., and Fidan, C. B. Comparison of Artificial Intelligence Applications of EEG Signals in Neuroscience. *Computers and Electronics in Medicine*, 3(1), 11-26, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.27-35
https://doi.org/10.69882/adba.cem.2026013

Determinants of Viewer Engagement in Health and Sports Videos: A Quantile Regression Forest Machine Learning Approach Applied to Reformer Pilates Content

Gülçin Aydoğdu[*,1], **Sibel Yıldırım**[α,2], **Serhat Hayme**[β,3] **and Emre Demir**[*,4]

[*]Hitit University, Faculty of Medicine, Department of Biostatistics, Çorum, Türkiye, [α]Hitit University, Faculty of Sport Sciences, Department of Coach Training, Çorum, Türkiye, [β]Erzincan Binali Yıldırım University, Faculty of Medicine, Department of Biostatistics and Health Informatics, Erzincan, Türkiye.

ABSTRACT This study assessed the quality, reliability, and popularity of Reformer Pilates videos on YouTube, evaluating educational value, credibility, and viewer engagement using standardized scoring while exploring factors affecting popularity. On November 1, 2024, a YouTube search for "Reformer Pilates" identified the top 70 most-viewed videos. Videos were excluded if irrelevant, promotional, or under 25 seconds. Video quality and reliability were assessed using the Global Quality Scale (GQS), Modified DISCERN (mDISCERN), and JAMA benchmarks. Viewer engagement and popularity were measured with the Video Power Index (VPI). Factors associated with engagement were analyzed using Quantile Regression Forest modeling. Most videos (87.1%) were uploaded by individual channels, 12.9% by Pilates studios. Average GQS (3.06), mDISCERN (3.09), and JAMA (2.61) scores indicated moderate quality and reliability. Videos claiming instructor expertise had lower mDISCERN scores ($p = 0.005$). Studio-recorded videos had higher GQS scores than home-recorded videos, though not statistically significant. Quantile Regression Forest analysis indicated that mentioning safety information, shorter video duration, and higher GQS scores were among the factors associated with higher viewer engagement (VPI). Older videos tended to exhibit lower engagement levels, reflecting a preference for newer content. Findings highlight the importance of high-quality, concise, and safety-focused Reformer Pilates videos, as these characteristics are associated with higher engagement and popularity. Content creators may benefit from prioritizing these factors to attract and retain audiences, underscoring YouTube's role in delivering potentially reliable Pilates content.

KEYWORDS
Machine learning
Reformer pilates
Video power index
YouTube
Viewer engagement

INTRODUCTION

Reformer Pilates is an effective and modern exercise system that supports the foundation of the musculoskeletal system, enhances the movement capacity and quality of the human body, and enables the body to move as a whole by correcting muscle weaknesses and imbalances (Lee 2023; Pereira *et al.* 2022). In Reformer Pilates, individuals engage in exercises in a controlled and focused manner, discovering their movement capacities (Lim and Hyun 2021). This makes the exercise more beneficial and improves the quality of individuals' daily life activities. Moreover, Reformer Pilates has become a commonly preferred method in the treatment processes of orthopedic problems in the muscles and joints (Rangabprai *et al.* 2024). In athletes, it is frequently chosen to enhance performance and accelerate the recovery process after injuries (Kaner and Ayer 2022). As Reformer is a functional and complex piece of equipment, ensuring multifaceted physical development and maintenance is prioritized. Therefore, it is essential to perform exercises under the guidance of expert trainers and with proper technique to achieve effective results and avoid injuries (Sim *et al.* 2022).

In today's digital age, YouTube has become more than just a social media and entertainment platform; it has transformed into a tool for beneficial content creation and a large-scale information sharing and interaction medium (Aharul 2023; Di Virgilio and Das 2023). In this context, YouTube has become a highly preferred tool for reaching a wide audience in exercise and healthy living fields, particularly for popular exercise techniques like Reformer Pilates. The global reach of YouTube plays a significant role in the growing popularity of Reformer Pilates (McDonough *et al.* 2022). Content created by both professional and amateur trainers has raised awareness of this exercise technique by making the system and principles of Reformer Pilates easy to understand. However, the quality, reliability, and academic accuracy of these created contents have become a part of the general discussions about exercise content on YouTube (Ratwatte and Mattacola 2021; Sari *et al.* 2025).

The purpose of this study is to evaluate the accuracy and reliability of Reformer Pilates videos shared on YouTube, in addition to analyzing key viewer engagement metrics. By examining the

Manuscript received: 10 November 2025,
Revised: 30 December 2025,
Accepted: 30 December 2025.

[1]gulcinaydogdu06@gmail.com (**Corresponding author**)
[2]sibelyildirim@hitit.edu.tr
[3]serhathayme@gmail.com
[4]emredemir82@gmail.com

visibility of Reformer Pilates on the platform, analyzing the diversity of content, the expertise and quality of the content creator, its popularity, and user interaction, the study seeks to reveal the impact of these videos on users and determine the role of YouTube in the widespread adoption of Reformer Pilates.

MATERIALS AND METHODS

Data Collection

On November 1, 2024, data for this analysis were sourced from YouTube, a widely used platform for video sharing and social networking. A search was performed using the term "Reformer Pilates" as the only keyword, and the resulting videos were sorted by their view counts, with the highest-ranked videos appearing first. The top 70 videos, all demonstrating Pilates exercises, were identified according to specific inclusion criteria. Videos were excluded if they were irrelevant, categorized as shorts, or were promotional in nature. Additionally, videos containing personal testimonials, those in languages other than English, those without sound, or those shorter than 25 seconds were also removed. The final dataset included videos with view counts ranging from over 2 million for the most watched to over 32 thousand for the least watched. Data from these videos were analyzed statistically. An instructor, accredited by the Turkish Gymnastics Federation, conducted the evaluation of each video. This evaluation process was based on standardized instruments, namely the Modified DISCERN (mDISCERN) scale, the Global Quality Scale (GQS), and the Journal of the American Medical Association (JAMA) scoring method Furthermore, metrics such as the view-to-like ratio and the Video Power Index (VPI) were computed from the available video data.

Evaluation Criteria

In this cross-sectional investigation, YouTube video reliability, quality, and popularity were assessed through various evaluation methods. The mDISCERN, GQS, and JAMA benchmarks were specifically used to ascertain the credibility and educational merit of the videos. Viewer engagement and popularity levels were gauged using the VPI.

The GQS, established by Bernard *et al.* (2007), is a five-point scale designed for evaluating the quality of medical video content intended to support patient education. It evaluates core elements such as scientific accuracy, clarity and effectiveness of communication, comprehensiveness, educational value, and the overall potential of the content to benefit viewers. Given the growing role of Reformer Pilates in rehabilitation, posture correction, pain management, and overall wellness, this scale was used to evaluate how effectively the videos present accurate and understandable information about Pilates exercises and their potential health benefits. Scores range from 1 (poor quality, useless/limited use to viewers) to 5 (excellent quality, highly beneficial/useful for viewers) (Bernard *et al.* 2007).

To assess the reliability and accuracy of content in the Reformer Pilates videos, the mDISCERN scale was utilized. This instrument, adapted by Singh *et al.* (2012) from the original DISCERN tool by Charnock *et al.* (1999), comprises five items. It evaluates dependability, clarity, effective presentation, and the capacity to deliver accurate health-related information. It was applied to examine whether the Reformer Pilates videos clearly define their purpose, provide balanced and evidence-based explanations of Reformer Pilates techniques, and address any benefits or limitations relevant to health outcomes. Each item receives a score of 1 (yes) or 0 (no), yielding a total score between 0 and 5, where higher scores denote superior informational quality (Singh *et al.* 2012; Charnock *et al.* 1999).

Evaluation of the Reformer Pilates videos' reliability and quality was conducted using the JAMA scale, a tool developed by Silberg *et al.* (1997). This scale examines essential components such as the accuracy of the information, scientific validity, clarity of expression, and the usefulness of the content for viewers.The four evaluation points—authorship (is it clear who is responsible for the content and their qualifications?), attribution (are sources for claims clearly cited?), disclosure (are conflicts of interest or sponsorships declared?), and currency (is the information up-to-date?)—were examined in the context of Reformer Pilates videos. Fulfillment of each criterion contributes one point to the score, leading to a possible range of 0 to 4, with higher scores signifying greater trustworthiness and reliability (Silberg *et al.* 1997).

Finally, the VPI, proposed by Erdem and Karaca (2018), was calculated to quantify viewer engagement and the popularity of Reformer Pilates videos. The VPI is used to determine the effectiveness of video content on social media platforms and to evaluate the extent of viewer interest it generates Given the growing interest in Reformer Pilates as both a fitness and rehabilitative practice, VPI quantified public interaction by incorporating metrics such as the likes and dislikes counts, total views, and the duration of time passed since each video was published (Erdem and Karaca 2018; Çoşkun and Demir 2024, 2025).

To perform the calculations, the following formulas were applied (Erdem and Karaca 2018; Çoşkun and Demir 2024).

$$VPI = \frac{\text{Like ratio} \times \text{View ratio}}{100}$$

$$\text{Like ratio} = \frac{100 \times \text{Like count}}{\text{Like count} + \text{Dislike count}}$$

$$\text{View ratio} = \frac{\text{View count}}{\text{Days since initial upload}}$$

Statistical Analyses

Analysis of the data was performed using SPSS software (Version 22.0, SPSS Inc., Chicago, IL, USA, License: Hitit University). For categorical variables, frequencies (n) and percentages (%) were used for description. Continuous data were summarized as mean $\pm$ standard deviation for those with a normal distribution, and as median (min–max) for non-normally distributed data. The assessment of data normality was carried out through the Shapiro–Wilk test along with graphical techniques. Since the data were not normally distributed, the Mann–Whitney U test was utilized for comparisons between two independent groups. The Kruskal–Wallis test was applied for comparisons across more than two independent groups when the assumptions of parametric tests were not satisfied. In cases of significant differences, pairwise post-hoc comparisons were conducted using the Dunn–Bonferroni test. Depending on the distributional features of the continuous variables, Spearman's correlation coefficient was employed to assess their associations. Statistical significance was set at $p < 0.05$.

Because the Video Power Index (VPI) is inherently skewed and heavy-tailed, mean-based regression models and variance-explained performance measures (e.g., R^2) were considered suboptimal. Therefore, a Quantile Regression Forest (QRF) approach was employed to estimate conditional quantiles of VPI and to identify predictors associated with high and very high viewer engagement. QRF is a non-parametric, tree-based machine learning method that extends random forests and is robust to nonlinear relationships, mixed predictor types, and non-normal error structures. The

outcome variable was VPI (continuous). Predictors included quality and reliability indicators (Global Quality Scale, modified DISCERN, and JAMA scores), video characteristics (video length and time since upload), and categorical descriptors related to content features and source characteristics (safety mention, publisher type, application location, instructor expertise, accessory use, machine settings demonstration, and region). The dataset was randomly divided into a training set (70%) and an independent test set (30%). All model selection and tuning procedures were conducted on the training set, with internal validation performed using 5-fold cross-validation. Model hyperparameters were optimized via randomized search using cross-validated pinball loss as the optimization criterion for $\tau = 0.75$ and $\tau = 0.90$. Model performance was evaluated using pinball loss, the standard metric for quantile regression models, with lower values indicating improved predictive accuracy. To enhance interpretability, quantile-specific permutation feature importance was computed by assessing changes in pinball loss following permutation of each predictor in the test set. All machine learning analyses were conducted using Python.

RESULTS

The first 70 videos featuring Reformer Pilates exercises were uploaded by individual YouTube channels (87.1%, n=61) and Pilates studio YouTube channels (12.9%, n=9). Of these videos, 55.7% (n=39) were filmed at home, while 44.3% (n=31) were filmed in a studio. It was found that 81.4% of the Pilates exercise videos did not specify the trainer's expertise, while 18.6% did indicate the trainer's expertise. The exercises targeted the following areas: 72.9% of the videos focused on the arms, 91.4% on the legs, 70% on the abdomen, 54.3% on the back, 77.1% on the hips, and 52.9% on the neck. The average video length was 1943.44 ± 1001.1 seconds, with video lengths ranging from 105 to 4234 seconds. The period since video upload ranged from 102 to 5507 days, averaging 1318 ± 1035.8 days. The number of views per video averaged 203335.4 ± 299235.8 (ranging from 32062 to 2003528), and the average number of likes was 2397.9 ± 2077.4 (ranging from 159 to 8828). The average Video Power Index (VPI) was 173.5 ± 193.1 (ranging from 22 to 1214.7). Additional descriptive statistics concerning the videos are detailed in Table 1.

The mean scores calculated for each evaluation tool were as follows: GQS, 3.06 ± 0.899; mDISCERN, 3.09 ± 1.073; and JAMA, 2.61 ± 1.09. The GQS assessment categorized 28.6% of the 70 analyzed videos as low quality, 45.7% as moderate quality, and 25.7% as high quality. Six videos achieved the maximum possible GQS score of 5. Eight videos attained a perfect score on the mDISCERN tool, and eighteen videos received a score of 4 on the JAMA criteria. Notably, a single video, which had garnered 77,000 views and was uploaded in 2020 by Homebody Pilates, achieved maximum scores across all three evaluation instruments; this video presented a studio-based Pilates routine targeting multiple body regions. The video with the highest view count (2003528 views), uploaded by VivaPilatesStudio, received scores of 4 for both GQS and mDISCERN, and 3 for JAMA. The second most viewed video (984,980 views) received scores of 4 for GQS, 3 for mDISCERN, and 3 for the JAMA criteria. The third and fourth most viewed videos (867220 and 683466 views) had scale scores of 3, 3, 4 and 3, 2, 2, respectively. The fifth most-viewed video (658421 views) scored 4 on GQS, 4 on mDISCERN, and 3 on JAMA.

Videos featuring expert trainers exhibited shorter durations, fewer views, lower VPI, and lower JAMA scores relative to those by non-expert trainers; however, these observed differences did not reach statistical significance (P-values: 0.862, 0.294, 0.886, and 0.324, respectively; see Table 2). Conversely, expert trainer videos showed higher GQS scores than those without expert trainers, though this difference also lacked statistical significance (P-value: 0.809; Table 2). Videos by expert trainers had statistically significantly lower mDISCERN scores than those by non-expert trainers (P-value: 0.005; see Table 2). A jittered boxplot illustrating the distribution of GQS, mDISCERN, JAMA, and VPI scores across different instructor expertise levels is presented in Figure 1A.

Videos recorded in studios tended to have higher views, VPI, GQS, mDISCERN, and JAMA scores than home-filmed videos, yet none of these differences were statistically significant (P-values: 0.274, 0.411, 0.098, 0.820, and 0.257, respectively; see Table 2). The length of studio-filmed videos was shorter than that of home-filmed videos, but this difference was not statistically significant (P-value: 0.290; see Table 2). A jittered boxplot depicting the distribution of GQS, mDISCERN, JAMA, and VPI scores across different pilates location categories is presented in Figure 1B.

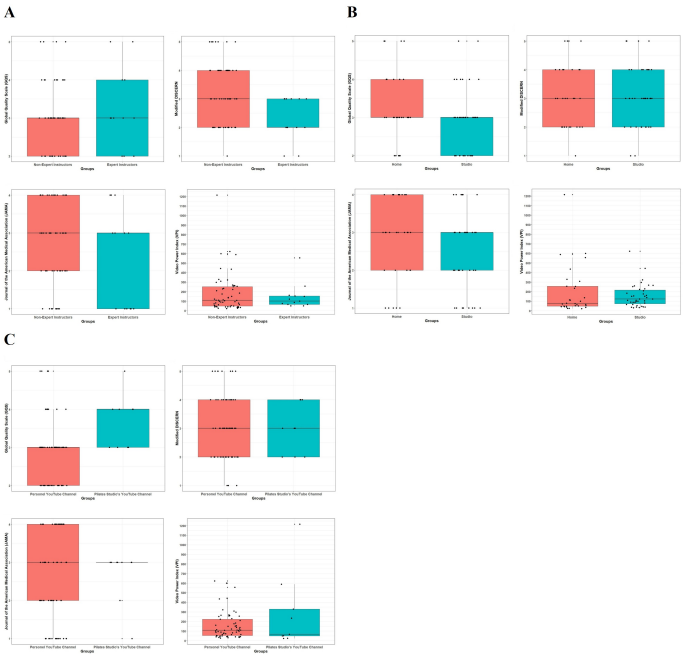


Figure 1 Distribution of quality, reliability, and engagement metrics across video characteristics. (A) Boxplot with jitter illustrating the distribution of Global Quality Scale (GQS), modified DISCERN (mDISCERN), Journal of the American Medical Association (JAMA), and Video Power Index (VPI) scores across different instructor expertise. (B) Boxplot with jitter illustrating the distribution of Global Quality Scale (GQS), modified DISCERN (mDISCERN), Journal of the American Medical Association (JAMA), and Video Power Index (VPI) scores across different pilates locations. (C) Boxplot with jitter illustrating the distribution of Global Quality Scale (GQS), modified DISCERN (mDISCERN), Journal of the American Medical Association (JAMA), and Video Power Index (VPI) scores among different video publishers.

Content from Pilates studio YouTube channels demonstrated higher view counts and JAMA scores when compared to videos from personal YouTube channels, though these distinctions were not statistically significant (P-values: 0.101, and 0.971, respectively; see Table 3). Additionally, Pilates studio YouTube channel videos had lower VPI and mDISCERN scores compared to personal YouTube channel videos, but these differences were not

Table 1 Descriptive statistical findings regarding the characteristics of the analysed YouTube videos (n=70)

	n	%		Mean ± SD	Median (min–max)
Video publisher			**Video features**		
Personal YouTube Channel	61	87.1	Video length (seconds)	1943.44 ± 1001.1	2077.5 (105–4234)
Pilates studio's YouTube channel	9	12.9	Time since upload (days)	1318 ± 1035.8	1077.5 (102–5507)
Pilates application location			Number of views	203335.4 ± 299235.8	94415.5 (32062–2003528)
Home	39	55.7	Number of likes	2397.9 ± 2077.4	1591 (159–8828)
Studio	31	44.3	Number of dislikes	46.81 ± 76.99	19.5 (0–510)
Instructor expertise			Comments	107.8 ± 143.4	76.5 (1–1018)
No	57	81.4	View ratio	177.8 ± 200.2	108.5 (23.3–1283.5)
Yes	13	18.6	Like ratio	97.9 ± 2.52	98.6 (87.8–100)
Activated regions			VPI	173.5 ± 193.1	106.2 (22–1214.7)
Arm			**Scales**		
No	19	27.1	GQS	3.06 ± 0.899	3 (2–5)
Yes	51	72.9	Modified DISCERN	3.09 ± 1.073	3 (1–5)
Leg			JAMA	2.61 ± 1.09	3 (1–4)
No	6	8.6			
Yes	64	91.4			
Abdomen					
No	21	30.0			
Yes	49	70.0			
Back					
No	32	45.7			
Yes	38	54.3			
Hips					
No	16	22.9			
Yes	54	77.1			
Neck					
No	33	47.1			
Yes	37	52.9			
Region					
Single region	5	7.1			
Multiple regions	65	92.9			
Is the activated region specified?					
No	25	35.7			
Yes	45	64.3			
Are the machine settings demonstrated?					
No	23	32.9			
Yes	47	67.1			
Is there use of additional accessories?					
No	42	60.0			
Yes	28	40.0			
Is safety mentioned?					
No	54	77.1			
Yes	16	22.9			
Content type					
Theoretical knowledge	5	7.1			
Application	18	25.7			
Mixed	47	67.1			

VPI: Video Power Index; JAMA: Journal of the American Medical Association; GQS: global quality scale

Table 2 Statistical findings comparing video length, number of views, VPI, GQS, Modified DISCERN, and JAMA scores among groups categorized by instructor expertise and reformer Pilates application location.

	Instructional Expertise			Location		
	No (n=57)	**Yes (n=13)**	***P* values**	**Studio (n=31)**	**Home (n=39)**	***P* values**
Video length (seconds)	2082 (105–3425) 1948.6±980.5	1812 (148–4234) 1920.5±1129.47	0.862	1812 (105–4234) 1793.6±1151.2	2198 (148–3425) 2062.5±860.7	0.290
Number of views	78787 (32062–2003528) 211461.5±325237.8	123615 (37141–575970) 167705.5±139785.3	0.294	110952 (32916–2003528) 264331.8±402417.6	88503 (32062–867220) 154851.1±171156.4	0.274
VPI	108.9 (22–1214.7) 179.8±204.3	100.4 (50.1–555.8) 145.9±136.2	0.886	74.8 (22–1214.7) 196.02±256.8	122 (30.9–622.6) 155.6±122.1	0.411
GQS	3 (2–5) 3.04±0.865	3 (2–5) 3.15±1.068	0.809	3 (2–5) 3.26±0.930	3 (2–5) 2.9±0.852	0.098
Modified DISCERN	3 (1–5) 3.26±1.061	2 (1–3) 2.31±0.751	**0.005**	3 (1–5) 3.13±1.118	3 (1–5) 3.05±1.05	0.820
JAMA	3 (1–4) 2.68±1.038	3 (1–4) 2.31±1.316	0.324	3 (1–4) 2.77±1.117	3 (1–4) 2.49±1.073	0.257

Mann Whitney U test with median (min–max) and mean ± SD
VPI: Video Power Index, GQS: Global Quality Scale, JAMA: Journal of the American Medical Association

statistically significant (P-values: 0.951, and 0.848, respectively; see Table 3). However, Pilates studio YouTube channel videos had statistically significantly higher GQS scores compared to personal YouTube channel videos (P-value: 0.016; see Table 3). Videos from Pilates studio channels were statistically significantly shorter in length than those from personal channels (P-value: 0.043; see Table 3). A jittered boxplot illustrating the distribution of GQS, mDISCERN, JAMA, and VPI scores across different video publishers is presented in Figure 1C.

Correlation analysis revealed no statistically significant relationships between GQS, Modified DISCERN, or JAMA scores and time since video upload, video length, number of views, number of likes, or VPI (all $P > 0.05$; see Table 4). Nevertheless, a weak positive correlation was observed between time since upload and number of views ($r = 0.466$, $P < 0.001$), whereas weak negative correlations were identified between time since upload and VPI ($r = -0.298$, $P = 0.012$) and between video length and number of views ($r = -0.308$, $P = 0.009$). No other statistically significant relationships were detected (all $P > 0.05$; Table 4).

Given the heavy-tailed and highly skewed distribution of the Video Power Index (VPI), Quantile Regression Forest (QRF) analysis was performed to model high and very high levels of viewer engagement. Model performance was evaluated using the pinball loss function, which provides quantile-specific predictive accuracy without reliance on variance-explained measures. Using a 70/30 train–test split with internal 5-fold cross-validation on the training set, the QRF model demonstrated stable and consistent performance across quantiles. For high engagement ($\tau = 0.75$), the mean pinball loss during cross-validation was 66.96 ± 36.95, while the corresponding pinball loss in the independent test set was 53.22. For very high engagement ($\tau = 0.90$), model performance improved, with a cross-validated pinball loss of 54.06 ± 35.73 and a lower pinball loss of 35.80 in the test set (Table 5).

Quantile-specific permutation feature importance analysis was conducted to improve interpretability of the model. Feature importance profiles differed across quantiles, indicating heterogeneity in the factors associated with viewer engagement at different levels of VPI. At the high engagement level ($\tau = 0.75$), time since upload was the most influential predictor, followed by overall content quality, as measured by the Global Quality Scale (GQS). Video duration also contributed notably, with longer videos associated with lower engagement, while explicit mention of safety information and publisher type showed additional influence (Figure 2).

At the very high engagement level ($\tau = 0.90$), time since upload remained the dominant predictor; however, indicators related to professional credibility, particularly publisher type, gained relative importance. GQS continued to contribute meaningfully, and video duration retained a negative association with engagement, even among the most highly viewed videos (Figure 3). Across both upper quantiles, time since upload, GQS, video duration, safety mention, and publisher type were consistently ranked among the most influential predictors of viewer engagement.

Table 3 Statistical findings comparing video length, number of views, VPI, GQS, Modified DISCERN, and JAMA scores among groups created based on video publisher

	Video Publisher		
	Personal YouTube Channel (n=61)	**Pilates Studio's YouTube Channel (n=9)**	***P* values**
Video length (seconds)	2132 (105–4234) 2039.1±971.09	1210 (192–2920) 1294.6±1014.4	**0.043**
Number of views	88503 (32062–867220) 154882.5±168284.6	211989 (36628–2003528) 531738.4±648592.4	0.101
VPI	108.9 (26.10–622.60) 156.7±140.5	65.8 (22–1214.7) 287.05±396.1	0.951
GQS	3 (2–5) 2.97±0.894	4 (3–5) 3.67±0.707	**0.016**
Modified DISCERN	3 (1–5) 3.1±1.106	3 (2–4) 3±0.866	0.848
JAMA	3 (1–4) 2.61±1.144	3 (1–3) 2.67±0.707	0.971

Mann Whitney U test with median (min–max) and mean ± SD
VPI: Video Power Index, GQS: Global Quality Scale, JAMA: Journal of the American Medical Association

Table 4 Correlation analysis findings determining the relationships between video metrics and GQS, Modified DISCERN, and JAMA scale scores, along with the relationships between video characteristics and video metrics (n=70)

		GQS	**Modified DISCERN**	**JAMA**
Time since upload (days)	*r*	0.114	-0.209	0.055
	P	0.349	0.083	0.649
Video length (seconds)	*r*	0.078	-0.094	0.019
	P	0.521	0.437	0.875
Number of views	*r*	0.195	-0.029	-0.036
	P	0.105	0.814	0.767
Number of likes	*r*	0.047	0.038	-0.008
	P	0.701	0.753	0.949
VPI	*r*	0.105	0.185	-0.078
	P	0.386	0.126	0.520
Correlation between video characteristics and video metrics				
		Time since upload (days)	**Video length (seconds)**	
Number of views	*r*	**0.466****	**-0.308****	
	P	< 0.001	0.009	
Number of likes	*r*	0.169	-0.017	
	P	0.161	0.889	
VPI	*r*	**-0.298***	-0.128	
	P	0.012	0.289	

GQS: Global Quality Scale; JAMA: Journal of the American Medical Association; VPI: Video Power Index

DISCUSSION

The findings of this study reveal several key factors influencing the popularity, perceived quality, and characteristics of Reformer Pilates videos on YouTube. In particular, Quantile Regression Forest analysis focusing on high and very high engagement levels

Table 5 Quantile Regression Forest performance for predicting Video Power Index (VPI)

Quantile (τ)	Evaluation Strategy	Pinball Loss (Mean $\pm$ SD)
0.75	5-fold CV (training set, 70%)	66.96 ± 36.95
0.75	Hold-out Test Set (30%)	53.22
0.90	5-fold CV (training set, 70%)	54.06 ± 35.73
0.90	Hold-out Test Set (30%)	35.80

Abbreviations: Pinball loss: quantile-specific loss function (lower values indicate better predictive accuracy); VPI: Video Power Index; CV: cross-validation.

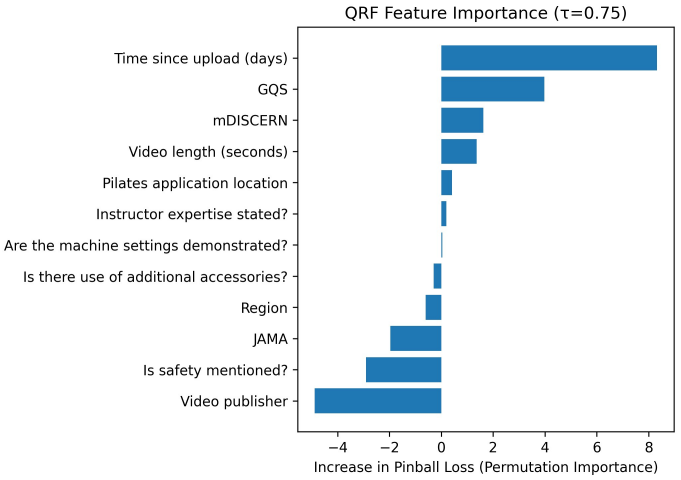


Figure 2 Feature importance at high engagement level ($\tau = 0.75$). *Footnote:* Permutation-based feature importance for the Quantile Regression Forest model. Importance values represent the increase in pinball loss following permutation of each predictor.

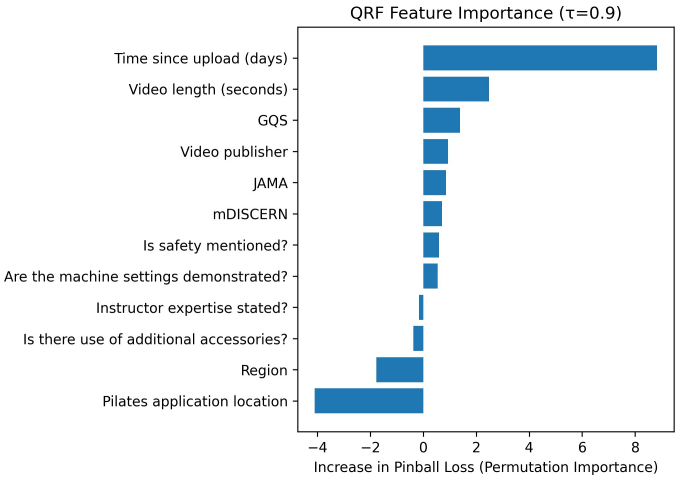


Figure 3 Feature importance at very high engagement level ($\tau = 0.90$). *Footnote:* Permutation-based feature importance for the Quantile Regression Forest model at the very high engagement quantile ($\tau = 0.90$). Importance values represent the increase in pinball loss following permutation of each predictor.

indicated that specific content-related features are associated with higher viewer engagement, with explicit mention of safety information emerging as one of the influential predictors of the Video Power Index (VPI). This indicates that viewers, likely including a significant proportion of beginners or those less familiar with the specialized equipment, prioritize safety when engaging with exercises like Reformer Pilates, which inherently involve technical skill and potential injury risk. The explicit emphasis on safety appears to foster trust and drive engagement. This result is in line with the study by Ratwatte and Mattacola (2021), which suggested that safety warnings in fitness content enhance viewer trust and positively affect engagement. The importance of clear, actionable safety information in online exercise content cannot be overstated, especially given the autonomous nature of at-home exercise guided by such videos Gronwald *et al.* (2022), while focusing on hamstring injuries in football, highlight the general need for demand-specific risk reduction, a principle applicable to online exercise guidance).

Time since upload was identified as the most influential predictor of viewer engagement in the Quantile Regression Forest analysis at both high and very high engagement levels. This finding reflects the "content decay" phenomenon, in which older videos gradually lose visibility and relative popularity as platforms and viewers prioritize newer content. This can be explained by social media algorithms prioritizing new content and viewers' tendency to engage with new videos. Supporting this trend, Di Virgilio and Das (2023) noted that social media platforms highlight continuously updated content, causing older videos to lose visibility. While older videos accumulate more total views over time (as shown by the weak positive correlation between time since upload and views), their power to engage relative to newer content diminishes. This necessitates a strategy of regular content creation or strategic re-promotion of older, high-quality content for creators aiming for sustained impact on platforms driven by novelty and recency.

Video duration was identified as an influential predictor of viewer engagement in the Quantile Regression Forest analysis at both high and very high engagement levels. Longer videos were consistently associated with lower engagement, indicating that shorter and more concise content is more effective in achieving higher Video Power Index (VPI) values. This finding is consistent with McDonough *et al.* (2022), who found that shorter, goal-oriented content in exercise programs on YouTube receives higher engagement. The average video length in our sample was substantial (1943.44 ± 1001.1 seconds), suggesting that many videos may exceed the optimal duration for viewer attention and engagement in the current digital landscape, which often favors "snackable" content. Interestingly, our finding that videos from Pilates studio channels were statistically significantly shorter in length than those from personal channels might indicate that professional studios are more attuned to this preference for brevity or are structuring content in more digestible segments.

Overall content quality, as measured by the Global Quality Scale (GQS), consistently ranked among the influential predictors of viewer engagement in the Quantile Regression Forest analysis at both high and very high engagement levels. This finding suggests that higher-quality instructional content is more strongly represented among videos achieving elevated engagement levels. The GQS scale, developed by Bernard *et al.* (2007), is a recognized tool for assessing video quality. Our results align with Erdem and Karaca (2018), who found that channels producing high-quality content on YouTube received more engagement in their study of kyphosis exercise videos.

Furthermore, videos published by Pilates studio channels exhibited higher GQS scores compared with personal channels, suggesting that professionally produced content may be associated with higher perceived quality. Given that GQS consistently ranked among the influential predictors of engagement in the Quantile Regression Forest analysis, this finding supports the relevance of professional production standards in achieving higher viewer engagement. The average GQS score for the analyzed videos was 3.06 ± 0.899, with 45.7% rated as medium quality and only 25.7% as high quality, indicating substantial room for improvement in the overall quality of Reformer Pilates content on YouTube. The potential for video analysis to assess and improve movement quality, as demonstrated in various contexts from running gait (Vergeer *et al.* 2023) to specific exercises like back squats (Peres *et al.* 2024), underscores the value of high-quality visual presentation in exercise videos.

An unexpected finding of this study was that videos featuring trainers who explicitly stated their expertise tended to receive lower Modified DISCERN scores compared with videos presented by trainers without stated expertise. The Modified DISCERN tool assesses reliability and content accuracy, particularly the clarity and effective presentation of health-related information (Singh *et al.* 2012; Charnock *et al.* 1999). This counterintuitive result does not necessarily imply that expert trainers provide less reliable information, but rather, as Singh *et al.* (2012) highlighted, the presentation of content, particularly the excessive use of technical terms, can reduce viewer comprehension. With 81.4% of videos not specifying trainer expertise, those that did might have fallen into the trap of using overly technical language, thereby diminishing clarity and accessibility for a general audience, which is reflected in lower DISCERN scores. This underscores the critical importance for expert trainers to employ clear, simple, and understandable language when conveying information to a broad audience on platforms like YouTube, ensuring their expertise translates into accessible and truly informative content. While methods like 2D video analysis are increasingly used to evaluate exercise form and ROM (Tanioka *et al.* 2022), the verbal communication accompanying these visuals is equally crucial for the viewer's understanding and perceived quality.

Regarding filming location, videos filmed in studios had descriptively higher views, VPI, GQS, mDISCERN, and JAMA scores compared to those filmed at home. However, these differences were not statistically significant, although studio-filmed videos tended to exhibit higher GQS scores. This suggests that while professional settings can contribute to higher production quality, high-quality content can also be produced in home settings with appropriate equipment and planning. The lack of statistical significance may also be due to the relatively small number of studio channel videos (n=9) in our sample. The ability of platforms like YouTube to host content from diverse creators, regardless of access to professional studios, is a key aspect of its democratizing influence on information dissemination.

Overall, these findings provide valuable insights into how Reformer Pilates videos are perceived by their audience and which content features enhance engagement. The average quality scores (GQS 3.06, Modified DISCERN 3.09, JAMA 2.61) indicate a general landscape of moderate quality, highlighting a significant opportunity for content creators to improve. Creators can increase their success on digital platforms by producing content that emphasizes safety, is concise and up-to-date, is of high overall quality, and communicates expertise in an accessible manner.

Limitations

This study encountered some methodological limitations while evaluating Reformer Pilates content on YouTube. Due to the study design, the analysis was limited to the 70 most-watched English-language videos at a specific point in time (November 1, 2024). This limitation may affect the generalizability of the findings, as the dynamic nature of the platform and algorithmic changes can influence content visibility over time. Despite these limitations, the study provides a robust methodological framework for evaluating digital exercise content and lays an important foundation for future research. In particular, it paves the way for multicenter and longitudinal studies that encompass content produced across different cultures and languages. Furthermore, the findings offer practical value by providing concrete recommendations for content creators to enhance the quality and engagement potential of their videos.

CONCLUSION

The results of this study reveal factors influencing the popularity of Reformer Pilates videos, offering important insights for digital exercise content creators. The findings suggest that videos that include safety information, are of high quality, short in duration, and up to date are more likely to be associated with higher viewer engagement. The observed association between video quality, content reliability, and viewer interaction highlights the importance for content creators to consider professionalism and clarity when developing educational materials. Furthermore, the observed decline in engagement of older videos over time underscores the critical need for consistent production of updated content to maintain visibility on digital platforms. It has been observed that technical language in content created by expert trainers can sometimes reduce comprehensibility, negatively affecting the informative quality. Therefore, it is recommended that content be prepared to maintain scientific accuracy while using clear and accessible language. Based on the data obtained, effective presentation of exercise content such as Reformer Pilates on digital platforms is more likely when videos emphasize safety, are high quality, concise, and up to date. This approach will enhance both viewer satisfaction and the level of information provided. The findings of this study highlight the need to evaluate Reformer Pilates content on YouTube not only based on view counts but also in terms of content quality and relevance.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Aharul, J. A., 2023 Social media influenced lexicons: A child's vocabulary production in talk-in interactions. International Journal of Learning, Teaching and Educational Research **22**: 541–556.

Bernard, A., M. Langille, S. Hughes, C. Rose, D. Leddin, *et al.*, 2007 A systematic review of patient inflammatory bowel disease information resources on the world wide web. American Journal of Gastroenterology **102**: 2070–2077.

Çoşkun, N. and E. Demir, 2024 Analysis of youtube videos on circumcision: evaluating reliability and quality for patients and parents. European Journal of Therapeutics **30**: 626–637.

Çoşkun, N. and E. Demir, 2025 Analysing youtube as a health resource: quality and reliability of videos on pediatric appendicitis. BMC Medical Education **25**: 1229.

Charnock, D., S. Shepperd, G. Needham, and R. Gann, 1999 Discern: an instrument for judging the quality of written consumer health information on treatment choices. Journal of Epidemiology and Community Health **53**: 105–111.

Di Virgilio, F. and S. Das, 2023 Digital solutions for social media: role of digital evolution on business enterprises. In *Digital Entertainment as Next Evolution in Service Sector: Emerging Digital Solutions in Reshaping Different Industries*, pp. 127–146, Springer Nature Singapore, Singapore.

Erdem, M. N. and S. Karaca, 2018 Evaluating the accuracy and quality of the information in kyphosis videos shared on youtube. Spine **43**: 1334–1339.

Gronwald, T., C. Klein, T. Hoenig, M. Pietzonka, H. Bloch, *et al.*, 2022 Infographic. video analysis of match hamstring injury patterns in professional male football (soccer) teaches us about the need for demand-specific multicomponent exercise-based risk reduction programmes. British Journal of Sports Medicine **56**: 1194–1195.

Kaner, G. and c. Ayer, 2022 The effect of a weight-loss diet in women doing reformer pilates: a 12-week evaluation. Journal of Contemporary Medicine **12**: 19–26.

Lee, K., 2023 Motion analysis of core stabilization exercise in women: kinematics and electromyographic analysis. Sports **11**: 66.

Lim, E. J. and E. J. Hyun, 2021 The impacts of pilates and yoga on health-promoting behaviors and subjective health status. International Journal of Environmental Research and Public Health **18**: 3802.

McDonough, D. J., M. A. Helgeson, W. Liu, and Z. Gao, 2022 Effects of a remote, youtube-delivered exercise intervention on young adults' physical activity, sedentary behavior, and sleep during the covid-19 pandemic: Randomized controlled trial. Journal of Sport and Health Science **11**: 145–156.

Pereira, M. J., R. Mendes, R. S. Mendes, F. Martins, R. Gomes, *et al.*, 2022 Benefits of pilates in the elderly population: a systematic review and meta-analysis. European Journal of Investigation in Health, Psychology and Education **12**: 236–268.

Peres, A. B., A. Sancassani, E. A. Castro, T. A. F. Almeida, D. A. Massini, *et al.*, 2024 Comparing video analysis to computerized detection of limb position for the diagnosis of movement control during back squat exercise with overload. Sensors **24**: 1910.

Rangabprai, Y., W. Mitranun, and W. Mitarnun, 2024 Effect of 60-min single bout of resistance exercise, reformer pilates, on vascular function parameters in older adults: a randomized crossover study. Gerontology **70**: 764–775.

Ratwatte, P. and E. Mattacola, 2021 An exploration of 'fitspiration' content on youtube and its impacts on consumers. Journal of Health Psychology **26**: 935–946.

Sari, F., Z. Bazancir Apaydin, and S. Sari, 2025 Assessment of reliability and quality of youtube exercise videos in people with rheumatoid arthritis. Physiotherapy Theory and Practice **41**: 362–369.

Silberg, W. M., G. D. Lundberg, and R. A. Musacchio, 1997 Assessing, controlling, and assuring the quality of medical information on the internet: caveant lector et viewor—let the reader and viewer beware. JAMA **277**: 1244–1245.

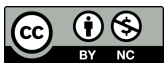
Sim, G., D. Kim, and H. Jeon, 2022 Effects of pilates reformer core and mat core exercises on standing posture alignment. Physical Therapy Korea **29**: 282–288.

Singh, A. G., S. Singh, and P. P. Singh, 2012 Youtube for information on rheumatoid arthritis—a wake-up call? The Journal of Rheumatology **39**: 899–903.

Tanioka, R., H. Ito, K. Takase, Y. Kai, K. Sugawara, *et al.*, 2022 Usefulness of 2d video analysis for evaluation of shoulder range of motion during upper limb exercise in patients with psychiatric disorders. The Journal of Medical Investigation **69**: 70–79.

Vergeer, R., H. Bloo, F. Backx, M. Scheltinga, and E. Bakker, 2023 Reliability of 2d video analysis assessing running kinematic variables in patients with exercise-related leg pain in a primary care practice. Gait and Posture **105**: 117–124.

How to cite this article: Aydoğdu, G., Yıldırım, S., Hayme, S., and Demir, E. Determinants of Viewer Engagement in Health and Sports Videos: A Quantile Regression Forest Machine Learning Approach Applied to Reformer Pilates Content. *Computers and Electronics in Medicine*, 3(1), 27-35, 2026.



ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.36-41
https://doi.org/10.69882/adba.cem.2026014

Exploring the Chemical Space of BACE-1 Inhibitors: Structure-Based Prediction with Deep Learning and Machine Learning

Duygu Selen Yılmazcan [*,1] **and Muhammed Ali Pala** [α,β,2]

*Department of Biomedical Engineering, Sakarya University of Applied Sciences, Sakarya 54050, Türkiye, αDepartment of Electrical and Electronics Engineering, Faculty of Technology, Sakarya University of Applied Sciences, Sakarya 54050, Türkiye, βBiomedical Technologies Application and Research Center (BIYOTAM), Sakarya University of Applied Sciences, Sakarya 54050, Türkiye.

ABSTRACT The identification of effective inhibitors targeting β-site amyloid precursor protein cleaving enzyme-1 (BACE-1) is crucial for developing therapeutic strategies for Alzheimer's disease. This study developed a structure-based computational framework for predicting BACE-1 inhibitory activity using both deep learning and conventional machine learning techniques. A publicly available BACE-1 dataset with chemical structures defined in SMILES (Simplified Molecular Input Line Entry System) format was subjected to feature extraction using the RDKit program. Global molecular characteristics and substructural information were captured using both molecular fingerprint representations and physicochemical descriptors. Circular (Morgan/ECFP4) fingerprints, RDKit fingerprints, and MACCS keys were used to encode molecular substructures into binary vectors. Subsequently, Support Vector Machines (SVM), k-Nearest Neighbors (kNN), deep neural networks (DNN), and enhanced deep neural networks were trained and validated using these features under the same experimental conditions. Confusion-matrix analysis and standard classification metrics (accuracy, precision, recall, and F1-score) were used to assess the model's performance. Deep learning models outperformed traditional machine learning techniques in capturing intricate nonlinear structure–activity correlations, according to comparison research. The proposed enhanced DNN demonstrated balanced precision and recall across both classes and achieved an accuracy of 0.99 on a test set of 303 molecules, including 138 active inhibitors and 165 inactive non-inhibitors. All things considered, these results imply that deep learning models, in conjunction with molecular fingerprints, offer a robust and reliable approach to BACE-1 inhibitor prediction and could accelerate early-stage virtual screening. All experiments were conducted using a fixed random seed and a held-out random split to ensure reproducibility.

KEYWORDS
BACE-1 inhibitors
Deep learning
Machine learning
Molecular fingerprints
Structure activity relationship

INTRODUCTION

The most common cause of dementia globally, Alzheimer's disease, is a progressive neurological condition that primarily affects older persons (Knopman et al. 2021). Memory and other cognitive abilities gradually deteriorate because of the illness, which eventually affects independence and day-to-day activities. Alzheimer's disease has a significant macroeconomic impact on healthcare systems in addition to its dramatic effects on patients and caregivers. The need for more potent disease-modifying therapies is highlighted by the fact that, despite symptomatic therapies, current therapeutic options remain limited in their ability to prevent or reverse disease progression.

The build-up of β-amyloid (β) peptides is one of the main pathogenic characteristics of Alzheimer's disease at the molecular level (Hampel et al. 2023). The amyloidogenic processing of the amyloid precursor protein, involving multiple cleavage events, results in the production of β (Hampel et al. 2023). The first cleavage step that commits amyloid precursor protein to β synthesis is catalysed by β-site amyloid precursor protein cleaving enzyme-1 (BACE-1, β-secretase). As a result, pharmacological inhibition of BACE-1 has been extensively studied as a tactic to lower β formation and impede downstream aggregation-related processes, making BACE-1 a significant therapeutic target and an ongoing focus of drug development research (Coimbra et al. 2024; Ghosh 2024).

Simultaneously, the availability of vast biochemical datasets and the expansion of computational resources have made machine learning, especially deep learning, increasingly crucial in early-stage drug discovery (Zhang and Saravanan 2024; Pala 2025a). By learning quantitative structure–activity relationships (QSAR) from molecular representations, machine-learning-based screening can save time and money compared to conventional trial-and-error methods (Pala 2025b). Previous studies have reported the effectiveness of deep neural network–based approaches for drug interaction and bioactivity-related prediction tasks (Pala 2025c). DNNs are particularly well-suited to capturing nonlinear patterns in chemical data (Pala 2025a; Qian et al. 2023). Deep learning-based

Manuscript received: 21 November 2025,
Revised: 18 January 2026,
Accepted: 21 January 2026.

[1]25500305002@subu.edu.tr (**Corresponding author**)
[2]pala@subu.edu.tr

methods have also been shown to outperform traditional machine learning baselines for BACE-1 inhibitor prediction.

Inspired by these advancements, this study proposes a structure-based computational framework that utilises molecular information derived from SMILES strings to classify BACE-1 inhibitors. We extract a feature set comprising physicochemical descriptors generated with RDKit tools and fingerprint-based representations that encode substructural patterns. To facilitate a fair and direct comparison, we integrate these molecular representations into both deep learning architectures and traditional machine learning models, such as SVM and KNN, within a single experimental setup. All models are trained and evaluated on the same dataset and shared feature space, providing a controlled, systematic comparison of DNN-based and traditional ML approaches for differentiating BACE-1 inhibitors from non-inhibitors, in contrast to prior studies that primarily focus on a single modelling paradigm.

MATERIALS AND METHODS

Dataset

The dataset used in this study was obtained from an open-access BACE-1 inhibitor classification dataset widely used in the literature and contains a total of 1513 compounds with experimentally reported BACE-1 inhibitory activity (Wu et al. 2018). Each compound is represented in SMILES format and annotated with binary class labels (active/inactive), framing the prediction task as a binary classification problem (Weininger 1988; Lenselink et al. 2017).

The raw data were preprocessed to remove any erroneous or illegible molecular structures. The RDKit chemical computing library was then used to calculate molecular descriptors that quantitatively represent the compounds' structural and physicochemical properties. The calculated descriptors include molecular weight (MolWt), octanol/water partition coefficient (LogP), numbers of hydrogen bond donors and acceptors (HBD, HBA), topological polar surface area (TPSA), numbers of rotatable bonds, rings, and heavy atoms, and the carbon sp^3 ratio (FractionCSP3) (Todeschini and Consonni 2009). These descriptors are widely used in structure–activity relationship (SAR) modelling and exhibit strong discriminatory power for predicting inhibitory activity.

In addition to global physicochemical descriptors, molecular structures were also encoded using fingerprint-based representations to capture local substructural information. Circular and key-based molecular fingerprints were generated for each compound using the RDKit library. Circular (Morgan/ECFP4) fingerprints, RDKit fingerprints, and MACCS keys were used to encode molecular substructures (Rogers and Hahn 2010; Durant et al. 2002; Yang et al. 2019). Fingerprints encode molecular substructure patterns as binary vectors, enabling machine learning and deep learning models to learn structural similarities and recurring motifs among compounds effectively. This representation facilitates the incorporation of local chemical features that are not fully captured by traditional descriptor-based approaches. After feature generation, the dataset was split into training and test sets using a stratified random split with an 80/20 ratio, ensuring that the class distributions of active and inactive compounds were preserved across both subsets.

Table 1 summarises the various fingerprint types, computational factors, and basic characteristics used in this study. In this study, predicting BACE-1 inhibitor activity is treated as a binary classification problem. Both standard machine learning techniques and deep learning models were trained on a high-dimensional feature space comprising chemical IDs and fingerprint images. All models attempt to distinguish between inhibitory and non-inhibitory substances.

Machine Learning Models

Support Vector Machine (SVM) Support vector machines are supervised learning algorithms that try to maximise class separation by defining an ideal decision boundary (Cortes and Vapnik 1995). They produce effective results, particularly on high-dimensional and nonlinear datasets. In this study, kernel functions were used to create the SVM model and capture intricate interactions between chemical characteristics. SVM's fundamental purpose is to identify a hyperplane that maximises the margin between classes. The optimisation problem presetnted in Equation 1./2. is solved by minimising a regularised loss function, where x_i represents the molecular feature vector, y_i the class label, w the weight vector, b the bias term, and ξ_i the slack variables. The regularisation parameter C controls the trade-off between maximising the margin and allowing classification errors.

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{1}$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \tag{2}$$

Equation 1./2. The Support Vector Machine (SVM) optimisation issue is formulated mathematically, with the goal of maximising the margin between classes while permitting controlled misclassification using slack variables and the regularisation parameter C.

k-Nearest Neighbors (kNN) The k-nearest neighbour (kNN) algorithm is a sample-based classification method that accounts for similarities among samples (Cover and Hart 1967). A test sample's class is determined by the class labels of its k nearest neighbours in the feature space. When predicting activity based on molecular similarities, the kNN algorithm produces intuitive and effective results.

Equation 3 illustrates the class prediction rule of the kNN classifier, where $N_k(x)$ represents the set of k nearest neighbours of the test sample x, and $I(\cdot)$ denotes the indicator function. Distance between samples is measured using the Euclidean distance.

$$\hat{y} = \arg\max_{c} \sum_{i \in N_k(x)} \mathbb{I}(y_i = c) \tag{3}$$

Equation 3 decision rule of the k-Nearest Neighbors (kNN) classifier, where the predicted class is determined by majority voting among the k nearest neighbors in the feature space.

Deep Learning Models

Deep Neural Network (DNN) For binary classification, a feed-forward deep neural network architecture was used. In addition to batch normalisation, dropout regularisation, and early stopping to enhance generalisation and avoid overfitting, the model was trained with class weighting to mitigate potential class imbalance. Based on validation performance, hyperparameters were selected empirically. Multilayer artificial neural networks have been successfully applied to model complex nonlinear structure–activity relationships in various molecular property and drug interaction prediction tasks (Pala 2025a,b). These networks provide a hierarchical learning process that begins at the input layer and propagates through one or more hidden layers to the output layer (LeCun et al. 2015).

The following equation defines the output of a neuron in a fully connected neural network layer:

Table 1 Summarises the descriptive features, parameter settings, vector types, and dimensionalities of the chemical fingerprint representations used in this investigation.

Fingerprint Type	Description	Parameters	Vector Type	Bit Length
Morgan (ECFP4)	Circular fingerprints encoding atomic neighbourhoods up to a radius of two bonds	Radius = 2	Binary	1024
RDKit Topological (RDKFingerprint)	Path-based fingerprint encoding linear atom paths and molecular bonding patterns	Default	Binary	2048
MACCS Keys	Predefined chemical substructure keys representing common chemical patterns	166 keys	Binary	166

$$\mathbf{h}^{(l)} = f\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{4}$$

Equation 4. Forward propagation in a fully connected layer, where $\mathbf{h}^{(l)}$ denotes the activation vector of layer l, $\mathbf{h}^{(l-1)}$ the activations from the previous layer, $\mathbf{W}^{(l)}$ the weight matrix, $\mathbf{b}^{(l)}$ the bias vector, and $f(\cdot)$ the activation function.

Here, $\mathbf{W}^{(l)}$ represents the weight matrix, $\mathbf{b}^{(l)}$ the bias vector, and $f(\cdot)$ the activation function. For the binary classification problem, the sigmoid activation function is used in the output layer:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{5}$$

Equation 5. Sigmoid activation function for binary classification, where the function maps real-valued inputs to the interval $(0, 1)$.

A binary cross-entropy loss function was selected to reduce the discrepancy between the expected outputs and the actual class labels during model training:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{6}$$

Equation 6. Binary cross-entropy loss function used for training the deep neural network, where y_i denotes the true class label, $\hat{y}_i$ the predicted probability, and N the total number of samples.

Improved Deep Neural Network Architecture A binary cross-entropy loss function was selected to reduce the discrepancy between expected outputs and actual class labels during model training. Several regularisation strategies were incorporated into the baseline DNN architecture to improve model performance and prevent overfitting. During training, a proportion of the network's neurons was randomly deactivated using dropout layers, which helps reduce co-adaptation of neurons and enhances generalisation (Srivastava et al. 2014). To address potential class imbalance in the dataset, class weights were incorporated into the training process, following strategies adopted in previous deep learning studies (Pala 2025a). Furthermore, an early stopping mechanism was employed to monitor the model's validation loss and terminate training once performance ceased to improve, thereby preventing unnecessary overtraining (Prechelt 1998). The hyperparameter configuration of the improved DNN model is summarized in Table 2.

EXPERIMENTAL RESULTS

This section presents and analyses the experimental findings from using deep learning and classical machine learning models on the BACE-1 dataset. Standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrices, were used to assess the model's performance (Sokolova and Lapalme 2009). These criteria provide a thorough evaluation of each model's ability to differentiate between BACE-1 inhibitors and non-inhibitors accurately. To guarantee the validity and applicability of the presented findings, all experiment was carried out on a separate test set. The efficacy of the deep neural network model in capturing intricate nonlinear correlations between chemical fingerprints and BACE-1 inhibitory activity was evaluated.Table 3 summarises the classification performance of the DNN model on the test dataset.

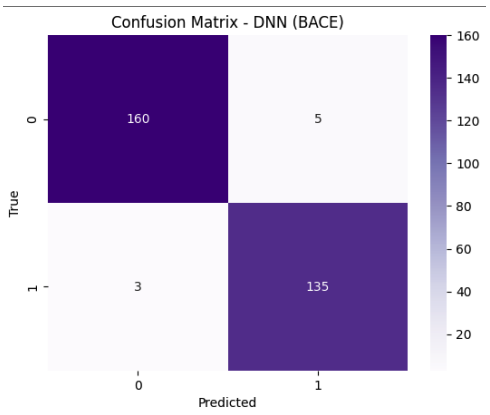


Figure 1 The confusion matrix of the DNN model illustrating its ability to correctly classify both inhibitory and non-inhibitory compounds, while revealing a limited number of misclassifications.

The performance of the proposed deep neural network (DNN) model was compared with that of traditional machine learning techniques, including Support Vector Machines (SVM) and k-Nearest Neighbours (kNN), to further evaluate the effectiveness of the proposed deep learning methodology. To ensure a fair and unbiased comparison, all models were trained and evaluated using identical molecular feature representations and the same held-out test dataset. Standard classification metrics were employed for the comparative analysis, enabling an objective assessment

Table 2 Hyperparameter configuration of the improved deep neural network (DNN) model used for BACE-1 inhibitor classification.

Hyperparameter	Value	Description
Neurons per layer	512–256–128	Decreasing layer sizes (funnel-shaped architecture)
Activation function	ReLU	Nonlinear activation for hidden layers
Output activation	Sigmoid	Binary classification output
Loss function	Binary cross-entropy	Optimization objective for binary labels
Optimizer	Adam	Adaptive gradient-based optimizer
Learning rate	0.001	Initial learning rate for Adam
Batch size	32	Number of samples per gradient update
Max epochs	100	Upper bound; training may stop earlier with early stopping
Dropout rate	0.30	Regularization to reduce overfitting
Validation split	0.20	Fraction of training data reserved for validation

Table 3 Classification performance on the BACE-1 inhibitor prediction problem.

Class	Precision	Recall	F1-Score	Support
0	0.98	0.97	0.98	165
1	0.96	0.98	0.97	138
Accuracy			**0.97**	303
Macro Avg	0.97	0.97	0.97	303
Weighted Avg	0.97	0.97	0.97	303

of each model's capability to distinguish BACE-1 inhibitors from non-inhibitors.

Table 4 Standard classification criteria used to compare the performance of deep neural network designs and classical machine learning models on the BACE-1 inhibitor prediction problem.

Model	Accuracy	Precision	Recall	F1-score
SVM	0.97	0.97	0.97	0.97
kNN	0.82	0.82	0.82	0.82
DNN	0.97	0.97	0.97	0.97
Improved DNN	**0.99**	**0.99**	**0.99**	**0.99**

As shown in Table 4, the improved DNN model achieved the highest overall performance across all evaluation metrics, outperforming both classical machine learning methods and the baseline DNN architecture. While the SVM and standard DNN models demonstrated strong and comparable predictive performance, with an accuracy of 0.97, the kNN model exhibited substantially lower performance, indicating a limited ability to capture complex structure–activity relationships. In contrast, the improved DNN architecture achieved an accuracy of 0.99, along with precision, recall, and F1-score values of 0.99, highlighting its superior capability to model nonlinear relationships between molecular fingerprints and BACE-1 inhibitory activity. These results demonstrate the effectiveness of deep learning–based approaches, particularly optimised DNN architectures, for reliable inhibitor classification.

In addition to threshold-dependent performance metrics, receiver operating characteristic (ROC) analysis was conducted to evaluate the models' discriminative capability across varying decision thresholds. Unlike accuracy-based measures, ROC curves provide a threshold-independent assessment of classification performance, making them particularly suitable for imbalanced datasets such as BACE-1. The area under the ROC curve (AUC) reflects the overall ability of the models to distinguish between inhibitors and non-inhibitors. ROC curves are reported for classical machine learning models (SVM and KNN) to highlight their threshold-independent discrimination. In contrast, deep learning models were primarily evaluated using confusion matrices and standard classification metrics.

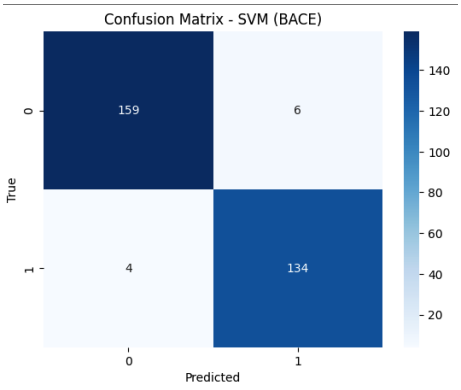


Figure 2 (a)

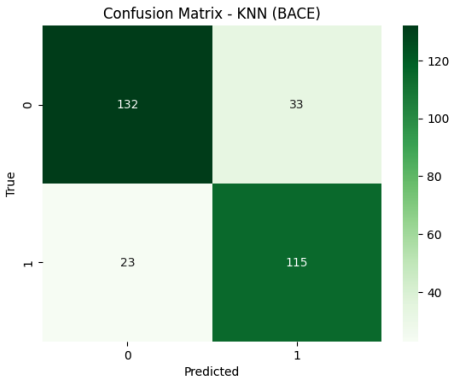


Figure 2 (b)

Figure 2 The confusion matrices of SVM and kNN highlight the differences in their classification behavior, with SVM exhibiting fewer misclassifications compared to kNN.

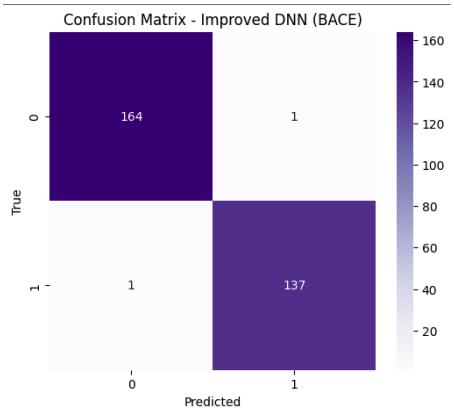


Figure 3 Compared to the baseline DNN, the improved DNN confusion matrix demonstrates a further reduction in classification errors, indicating enhanced generalisation capability.

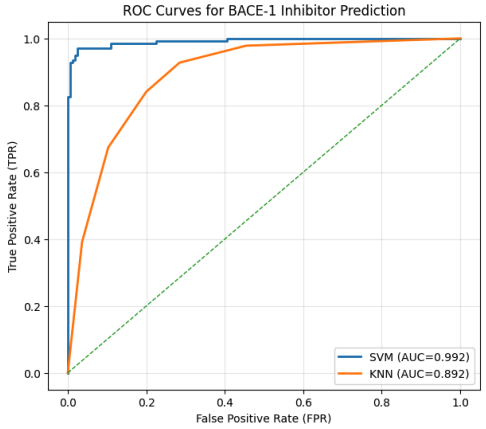


Figure 4 ROC curve analysis indicating that classical machine learning models, particularly SVM, exhibit strong discriminative capability for BACE-1 inhibitor prediction.

CONCLUSION

This study integrated chemical descriptors and fingerprint-based representations to develop a structure-based machine learning framework for predicting BACE-1 inhibitory activity. To assess how well deep learning techniques and classical machine learning models, such as Support Vector Machines and k-Nearest Neighbours, model the structure–activity relationships in the BACE-1 dataset. The experimental findings showed that deep neural network models outperform traditional techniques for extracting intricate nonlinear patterns from chemical fingerprint data. With an overall accuracy of 0.99, the enhanced DNN architecture outperformed SVM across prediction accuracy, precision, recall, and F1-score. These results show that for BACE-1 inhibitor classification tasks, deeper architectures in conjunction with optimised feature representations offer a substantial advantage. All things considered, the suggested method emphasises the promise of deep learning-based models as dependable and effective instruments for early-stage drug discovery, especially in the identification of prospective BACE-1 inhibitors. To improve predictive performance, future research might focus on extending this framework to regression-based activity prediction, external validation on separate datasets, and incorporating sophisticated representation learning methods, such as graph neural networks.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

All datasets analysed in this manuscript are publicly available. The MoleculeNet datasets can be accessed at the following link: https://moleculenet.org/datasets-1.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Knopman, D.S., Amieva, H., Petersen, R.C. et al. Alzheimer disease. Nat Rev Dis Primers 7, 33 (2021). https://doi.org/10.1038/s41572-021-00269-y

Harald Hampel, Giuseppe Caruso, Robert Nisticò, Gaia Piccioni, Nicola B. Mercuri, Filippo Sean Giorgi, Fabio Ferrarelli, Pablo Lemercier, Filippo Caraci, Simone Lista, Andrea Vergallo, Neurodegeneration Precision Medicine Initiative (NPMI), Biological Mechanism-based Neurology and Psychiatry: A BACE1/2 andDownstream Pathway Model, Current Neuropharmacology; Volume 21, Issue 1, Year 2023, e011221198450. DOI: 10.2174/1570159X19666211201095701

Coimbra JRM, Resende R, Custódio JBA, Salvador JAR, Santos AE. BACE1 Inhibitors for Alzheimer's Disease: Current Challenges and Future Perspectives. Journal of Alzheimer's Disease. 2024;101(s1):S53-S78. doi:10.3233/JAD-240146

Arun K. Ghosh, BACE1 inhibitor drugs for the treatment of Alzheimer's disease: Lessons learned, challenges to overcome, and future prospects†, Global Health & Medicine, 2024, Volume 6, Issue 3, Pages 164-169, Released on J-STAGE July 02, 2024, Advance online publication June 06, 2024, Online ISSN 2434-9194, Print ISSN 2434-9186, https://doi.org/10.35772/ghm.2024.01033.

Haiping Zhang, Konda Mani Saravanan, Advances in Deep Learning Assisted Drug Discovery Methods: A Self-review, Current Bioinformatics; Volume 19, Issue 10, Year 2024, e290124226290. DOI: 10.2174/0115748936285690240101041704

Pala, M. A. (2025). XP-GCN: Extreme learning machines and parallel graph convolutional networks for high-throughput prediction of blood-brain barrier penetration based on feature fusion. Computational Biology and Chemistry, 120, 108755. https://doi.org/10.1016/j.compbiolchem.2025.108755

Pala, M. A. (2025). Graph-Aware AURALSTM: An attentive unified representation architecture with BiLSTM for enhanced molecular property prediction. Molecular Diversity. https://doi.org/10.1007/s11030-025-11197-4

Pala, M. A. (2025). DeepInsulin-Net: A deep learning model for identifying drug interactions leading to specific insulin-related adverse events. Sakarya University Journal of Computer and Information Sciences, 8(2), 245–259. https://doi.org/10.35377/saucis...1646658

Qian, C., Tang, H., Yang, Z., Liang, H., & Liu, Y. (2023). Can Large Language Models Empower Molecular Property Prediction? http://arxiv.org/abs/2307.07443

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical Science, 9(2), 513–530. https://doi.org/10.1039/C7SC02664A

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, 28(1), 31–36. https://doi.org/10.1021/ci00057a005

Lenselink, E. B., ten Dijke, N., Bongers, B., Papadatos, G., van Vlijmen, H. W. T., Kowalczyk, W., IJzerman, A. P., & van Westen, G. J. P. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of Cheminformatics, 9(1), 45. https://doi.org/10.1186/s13321-017-0232-0

Todeschini, R., & Consonni, V. (2009). Molecular Descriptors for Chemoinformatics. Wiley. https://doi.org/10.1002/9783527628766

Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling, 50(5), 742–754. https://doi.org/10.1021/ci100050t

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. Journal of Chemical Information and Computer Sciences, 42(6), 1273–1280. https://doi.org/10.1021/ci010132r

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing Learned Molecular Representations for Property Prediction. Journal of Chemical Information and Modeling, 59(8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Prechelt, L. (1998). Early stopping—but when? In G. B. Orr & K.-R. Müller (Eds.), Neural networks: Tricks of the trade (pp. 55–69). Springer.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

How to cite this article: Yılmazcan, D. S., and Pala, M. A. Exploring the Chemical Space of BACE-1 Inhibitors: Structure-Based Prediction with Deep Learning and Machine Learning. *Computers and Electronics in Medicine*, 3(1), 36-41, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.42-47
https://doi.org/10.69882/adba.cem.2026015

Benchmarking State-of-the-Art Vision Transformer Architectures for the Automated Classification of Pigmented Skin Lesions

Md Saiful Islam [*,1], **André Chéagé Chamgoué** [β,2] **and Gurvinder Pal Dubb** [α,3]

*School of Engineering, Deakin University, Waurn Ponds Campus, Victoria, Australia, [β]Ngaoundere University, School of Geology and Mining Engineering, Meiganga, Cameroon, [α]Military Technological College, Department of Systems Engineering, Muscat, Sultanate of Oman.

ABSTRACT Skin cancer represents an escalating global public health challenge where early detection is paramount, potentially increasing five-year survival rates to 99%. While dermoscopy improves diagnostic sensitivity, its effectiveness often depends on clinician experience and is subject to inter-observer variability. To address these limitations, this study presents a rigorous comparative analysis of four state-of-the-art Vision Transformer (ViT) architectures, DeiT III-Base, Swin-Base, ViT-Base, and PiT-B, for the automated classification of pigmented skin lesions. We utilized the HAM10000 dataset (n=10,011) and implemented a stratified 70-15-15 split to ensure balanced training, validation, and testing phases. Images were resized to 224×224 pixels and normalized using ImageNet parameters, while transfer learning was employed to stabilize training and enhance generalization. Experimental results indicate that DeiT III-Base achieved superior diagnostic efficacy, reaching an accuracy of 92.04% and an F1-score of 85.44%. Furthermore, computational evaluation revealed that DeiT III-Base and ViT-Base offered highly efficient clinical throughput with sub-millisecond inference times (0.5674 ms and 0.5459 ms, respectively), whereas PiT-B exhibited the lowest computational workload (21.1067 GFLOPs). These findings underscore the viability of attention-based paradigms as robust real-time Computer-Aided Diagnosis (CAD) tools. Future research will explore the integration of multi-modal patient data and Explainable AI (XAI) to foster transparency and clinical trust.

KEYWORDS
Vision transformers (ViTs)
Skin cancer classification
HAM10000 Dataset
Computer-aided diagnosis (CAD)
DeiT III-Base

INTRODUCTION

Skin cancer constitutes a significant and escalating public health challenge globally, accounting for a substantial proportion of all diagnosed malignancies. With incidence rates rising steadily due to factors such as increased exposure to ultraviolet (UV) radiation, environmental changes, and aging populations, the burden on healthcare systems continues to grow. Among the various types, melanoma represents the most aggressive and lethal form, characterized by a high potential for metastasis (Gloster Jr and Neal 2006; Armstrong and Kricker 1995; Madan *et al.* 2010). However, prognosis is strongly correlated with the stage at diagnosis; early detection can increase the five-year survival rate to nearly 99%, whereas delayed diagnosis significantly diminishes treatment success and patient survival. Consequently, the development of rapid, accessible, and precise diagnostic mechanisms is not merely a clinical preference but a vital necessity to reduce mortality rates and improve patient outcomes (Gloster Jr and Brodland 1996).

Traditionally, the diagnosis of pigmented skin lesions relies on visual examination followed by dermoscopy, a non-invasive imaging technique that visualizes subsurface skin structures. While dermoscopy significantly enhances diagnostic sensitivity compared to naked-eye examination, its efficacy remains heavily dependent on the clinician's experience and training (Siegel *et al.* 2024; Jerant *et al.* 2000). The visual similarity between benign lesions (e.g., nevi, benign keratosis) and malignant tumors (e.g., melanoma, basal cell carcinoma) often leads to diagnostic ambiguity, resulting in either unnecessary biopsies or missed malignancies. Factors such as clinician fatigue, inter-observer variability, and the sheer volume of patients further complicate the manual diagnostic process. These limitations have necessitated the integration of Computer-Aided Diagnosis (CAD) systems to provide objective, consistent, and "second opinion" support for dermatologists.

In the last decade, the landscape of medical image analysis has been revolutionized by Deep Learning (DL), particularly Con-

Manuscript received: 5 September 2025,
Revised: 18 November 2025,
Accepted: 20 November 2025.

[1]Saiful.Islam@deakin.edu.au
[2]acchamgoue@gmail.com
[3]dubbgurvinderpal@gmail.com (**Corresponding author**)

volutional Neural Networks (CNNs) (ThangaPurni and Braveen 2025; Ozdemir and Pacal 2025; Çakmak and Pacal 2025; Cakmak and Maman 2025). Architectures such as ResNet, DenseNet, and EfficientNet have established themselves as the gold standard in skin lesion classification, demonstrating the ability to learn hierarchical feature representations directly from raw images (Karthik *et al.* 2024; Cakmak and Pacal 2025; Pacal and Cakmak 2025a,b). Despite their success, traditional CNNs primarily focus on local features due to their receptive field limitations, potentially missing long-range dependencies and global contextual information crucial for differentiating complex lesions. This limitation has paved the way for the adoption of Vision Transformers (ViTs) and hybrid architecture. Unlike CNNs, ViTs utilize self-attention mechanisms to capture global relationships across the entire image, offering a robust alternative for capturing fine-grained morphological details.

However, the rapid evolution of DL has precipitated a shift from pure convolutional networks to sophisticated attention-based paradigms. Consequently, there remains a critical need to rigorously evaluate how these distinct architectural strategies perform within the specific domain of skin cancer classification. In this study, we propose a comprehensive comparative analysis of advanced ViT architectures to identify the most effective mechanisms for skin lesion diagnosis. Rather than employing a broad spectrum of legacy models, our experimental framework rigorously benchmarks four distinct transformer-based methodologies that represent the state-of-the-art in attention mechanisms: the Data-efficient Image Transformer (DeiT III-Base), the standard ViT-Base, the hierarchical Swin Transformer (Swin-Base), and the Pooling-based Vision Transformer (PiT-B). Through this targeted evaluation, we aim to assess these architectures under unified conditions, providing critical insights into their generalization capabilities and clinical applicability for early skin cancer detection.

RELATED WORK

To overcome the locality bias of CNNs, researchers have increasingly adopted ViTs to model global context within dermoscopic images. Aruk *et al.* (2026) conducted a comprehensive comparative study evaluating 15 different CNNs against 15 ViT variants, including Swin and BeiT architectures, under identical training conditions. Their extensive analysis revealed that ViT models, particularly the Swin Transformer, consistently outperformed CNNs in classification accuracy, albeit at the cost of higher parameter counts and computational demands. Addressing the data-hungry nature of transformers, novel training paradigms have been introduced to improve robustness. Chaurasia *et al.* (2025) developed a multi-resolution model utilizing the DINOv2 self-supervised learning method to classify skin cancer subtypes from whole slide images (WSIs). By training on histological patches at various magnifications (10x to 400x), their model effectively captured multi-scale features, achieving an F1-score of 0.898 on external validation datasets.

Furthermore, specific architectural modifications have been proposed to tailor ViTs for the nuances of skin lesion analysis. Manju *et al.* (2025) proposed a preprocessing-optimized ViT model that integrates contrast enhancement and lesion segmentation directly into the workflow to remove artifacts before tokenization. This attention-enhanced model achieved an AUC-ROC score of 0.97, proving that feeding cleaner, segmented data into self-attention mechanisms significantly boosts diagnostic performance. Finally, the challenge of class imbalance in transformer training has been rigorously addressed. Sakib *et al.* (2025) introduced "LEVit," a framework that combines a hybrid ViT with extensive data augmentation and oversampling techniques to ensure uniform class distribution. Their approach not only achieved an F1 score of 98.11% on the ISIC 2019 dataset but also integrated Grad-CAM to generate class-specific heatmaps, ensuring the model's decisions were interpretable.

Building on the theme of architectural refinement for dermatological assessment, the literature has further evolved to address the synergy between localized features and global spatial reasoning. Pacal *et al.* (2024) advanced this structural capacity by proposing a Swin Transformer model that incorporates hybrid shifted window-based multi-head self-attention. Their methodology utilizes SwiGLU-based MLP layers to more effectively synchronize localized texture orientations with global spatial cues, thereby enhancing the network's sensitivity to minute malignant patterns that often elude standard convolutional filters. Beyond individual architectural optimizations, the focus has shifted toward improving decision robustness through multi-model integration. Bruno *et al.* (2025) expanded upon these gains by introducing a multi-scale attention and ensemble framework designed to aggregate features from diverse transformer-based learners. Their research demonstrates that combining multiple attention mechanisms effectively overcomes the inherent biases of single-architecture systems, leading to superior classification stability even in the presence of noisy or heterogeneous dermatoscopic data.

Finally, ensuring that these sophisticated models are both trustworthy and clinically viable has become a primary objective. Dagnaw *et al.* (2024) addressed this by integrating ViTs with Explainable Artificial Intelligence (XAI) to bridge the gap between high-performance computing and clinical transparency. By providing dermatologists with interpretable saliency maps, their framework ensures that the diagnostic logic of the transformer is both visible and verifiable, fostering the necessary confidence for the adoption of automated screening tools in high-stakes medical environments.

MATERIALS AND METHODS

Dataset and Data Preprocessing

In this research, we worked with the HAM10000 dataset, which is essentially the gold standard collection used in the ISIC 2018 challenge (HAM 2025). It contains 10,011 dermoscopic photos covering seven different types of skin conditions, ranging from harmless moles to dangerous cancers like melanoma. If you look at Figure 1, you can see why this is such a difficult task: many of these lesions look incredibly similar to the untrained eye, making manual diagnosis a real challenge.

When setting up our experiments, we didn't just split the data randomly. We used a stratified split to make sure the balance of diseases remained consistent across our training and testing sets. We settled on a 70-15-15 distribution, which gave us 7,005 images to train our models, 1,498 to fine-tune them, and a final 1,508 to test how well they actually perform on "unseen" cases. You can see the exact breakdown of these numbers in Table 1.

To get the images ready for the AI, we resized everything to 224×224 pixels. We also normalized the colors based on ImageNet standards. This step is vital because it helps the model ignore background "noise" like lighting differences and instead focus strictly on the textures and patterns that actually matter for a correct diagnosis.

Vision Transformers (ViTs)

To address the inherent limitations of local receptive fields in convolutional networks, this study leverages ViTs to model long-range

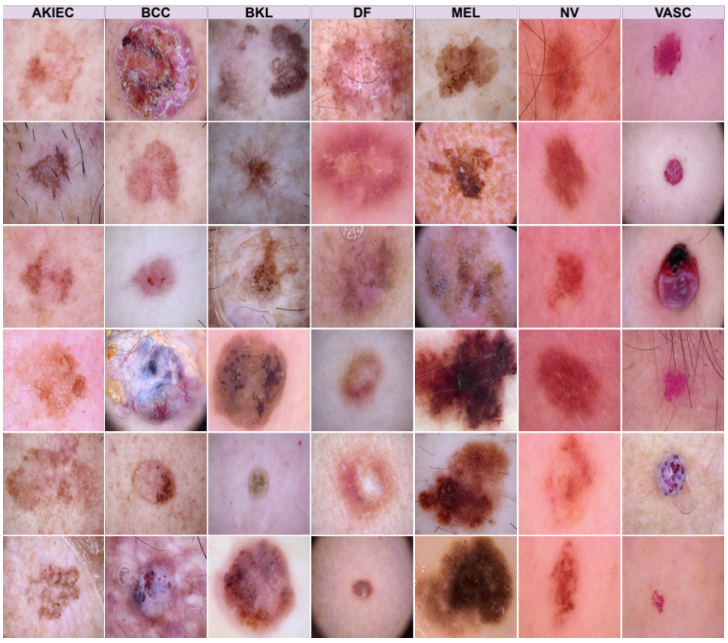


Figure 1 Representative dermoscopic samples from the HAM10000 dataset illustrating the seven skin lesion classes used in this study (AKIEC, BCC, BKL, DF, MEL, NV, and VASC).

Table 1 Distribution of the HAM10000 dataset across the seven diagnostic categories for training, validation, and testing subsets.

Class Name	Total	Train	Val	Test
BKL	1099	769	164	166
DF	115	80	17	18
VASC	142	99	21	22
AKIEC	323	226	48	49
MEL	1113	779	166	168
BCC	514	359	77	78
NV	6705	4693	1005	1007
Grand Total	10011	7005	1498	1508

dependencies and global semantic context, which are essential for distinguishing the subtle morphological variations in skin lesions. Our methodological framework rigorously benchmarks four distinct attention-based paradigms to evaluate their efficacy in dermoscopic analysis: the standard ViT-Base (Dosovitskiy *et al.* 2020), serving as a pure self-attention anchor; the DeiT III-Base (Touvron *et al.* 2022), selected for its superior data efficiency and refined training strategies suitable for medical datasets; the Swin Transformer (Swin-Base) (Liu *et al.* 2021), which introduces a hierarchical architecture with shifted windows to capture multi-scale features; and the Pooling-based Vision Transformer (PiT-B) (Ren *et al.* 2024), which incorporates spatial dimension reduction to bridge the gap between transformer flexibility and the structural abstraction of CNNs. By deploying this diverse set of architectures, we aim to dissect how specific structural innovations, ranging from hierarchical attention to pooling mechanisms, impact diagnostic precision in skin cancer classification.

Transfer Learning

Training high-capacity DL architectures, particularly Vision Transformers, from scratch necessitates massive volumes of annotated data to overcome the lack of inherent inductive biases found in convolutional networks. Given the intrinsic scarcity of labeled medical imaging datasets compared to general-domain repositories, initializing models with random weights often leads to poor convergence and significant overfitting. To mitigate this challenge, we adopted a transfer learning strategy by leveraging models pre-trained on the ImageNet-1K dataset. This approach allows our network to inherit robust hierarchical feature representations, ranging from low-level edge detection to high-level semantic abstractions, learned from millions of natural images. By transferring this prior knowledge to the dermatological domain, we effectively bypass the computational burden of learning fundamental visual patterns de novo, thereby accelerating training stability and enhancing generalization performance on dermoscopic images (Ren *et al.* 2022).

In the implementation phase, the architectural adaptation involved replacing the original classification head, designed for the 1,000 distinct classes of ImageNet, with a task-specific Multi-Layer Perceptron (MLP) tailored to our seven diagnostic categories. We employed an end-to-end fine-tuning protocol rather than merely using the backbone as a fixed feature extractor. This allowed the pre-trained weights to be subtly adjusted via backpropagation, facilitating the domain adaptation process. By doing so, the models could recalibrate their attention mechanisms to focus on the specific, fine-grained morphological details pertinent to skin lesions, such as pigment networks and vascular structures, while retaining the robust generalization capabilities acquired during the pre-training phase (Bengio 2012).

Experimental Design and Training Protocol

To ensure the reproducibility and rigor of our comparative analysis, all DL architectures were implemented using the PyTorch framework on a high-performance workstation equipped with an NVIDIA GPU. The dataset was partitioned using a stratified sampling strategy to maintain the inherent class distribution across subsets, resulting in a split of approximately 70% for training (n=7,005), 15% for validation (n=1,498), and 15% for testing (n=1,508). Prior to feeding the networks, all images underwent standardization, including resizing to uniform dimensions compatible with the pre-trained transformer backbones (typically 224×224) and normalization using ImageNet mean and standard deviation parameters. This rigorous preprocessing pipeline was essential to stabilize the training dynamics and ensure that the attention mechanisms could effectively attend to lesion-specific features without being swayed by lighting or resolution inconsistencies.

The training phase was conducted using the Cross-Entropy Loss function to penalize classification errors across the seven diagnostic categories. To optimize the network weights, we employed the AdamW optimizer, widely recognized for its efficacy in training transformer models, initialized with an empirically tuned learning rate and weight decay to prevent overfitting. We utilized a dynamic learning rate scheduler (Cosine Annealing) to progressively reduce the learning rate, allowing the model to settle into sharper minima as training converged. The best-performing model weights were saved based on the validation loss metric to avoid the pitfalls of overfitting during extended epochs. Finally, the quantitative evaluation was performed on the unseen test set using standard metrics, including Accuracy, Precision, Recall, and F1-Score, to provide a holistic view of each model's diagnostic capability.

Performance Evaluation Metrics

To truly understand how these models hold up, we didn't just look at their accuracy. We used a multi-layered approach to evaluate them, looking at how well they handle different skin diseases and how they would actually run in a real-world clinic. We tracked standard scores like Accuracy, Precision, Recall, and the F1-Score to see how reliable the diagnoses are. However, for a model to be useful in a hospital, it also needs to be efficient. That's why we also measured "Params" to see how much memory the model takes up, "GFLOPs" to calculate the raw processing power required, and "Inference Time" to see how many milliseconds it takes for a doctor to get a result. You can see how all these factors compare for each model in Table 2. The formulas we used to calculate these results are shown above. Accuracy (Eq. 1) gives us the big picture of correct guesses, while Precision (Eq. 2) tells us how often the model is right when it flags a lesion as concerning. Recall (Eq. 3) is perhaps the most important for patients because it measures how many actual cancer cases the model caught without missing any. Finally, the F1-Score (Eq. 4) helps us find the sweet spot between being precise and being thorough, which is vital since some types of skin cancer in our dataset are much rarer than others.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

RESULTS

Quantitative Performance and Comparative Analysis

The empirical evaluation of the four benchmarked ViT architectures reveals distinct trade-offs between diagnostic precision and computational efficiency. As summarized in Table 2, the DeiT III-Base model achieved the highest overall performance, reaching an accuracy of 92.04% and an F1-score of 85.44%. This superior performance suggests that the data-efficient training strategies and refined attention mechanisms of DeiT III are particularly effective for capturing the fine-grained morphological features necessary for skin lesion classification. Swin-Base followed closely with an accuracy of 91.64%, demonstrating the benefit of hierarchical feature extraction in dermoscopic analysis.

In terms of computational complexity, PiT-B exhibited the highest efficiency regarding memory footprint and processing workload, utilizing only 72.75M parameters and 21.1067 GFLOPs. However, this reduction in GFLOPs did not translate to the fastest execution; PiT-B recorded the highest Inference Time (1.5418 ms), likely due to its unique pooling-based architecture which may hinder parallelization compared to the standard transformer blocks. Conversely, ViT-Base and DeiT III-Base provided the most rapid predictions with inference times of 0.5459 ms and 0.5674 ms, respectively, making them the most suitable candidates for high-throughput clinical screening.

Analysis of Model Predictions and Error Patterns

To dissect the model's decision-making process, we analyzed the confusion matrix for the top-performing DeiT III-Base model, as shown in Figure 2. The diagonal elements indicate high sensitivity for the most prevalent class, Melanocytic Nevi (NV), with 977 correct predictions. However, notable confusion exists between Actinic Keratoses (AKIEC) and Benign Keratosis (BKL), as well as between Melanoma (MEL) and NV. This reflects the inherent "visual mimicry" described in clinical literature, where benign and malignant lesions share overlapping pigment patterns.

A qualitative assessment of these results is provided in Figure 3, which visualizes both successful and erroneous predictions alongside their confidence scores. True predictions often correspond to lesions with clear, well-defined diagnostic structures, such as the distinct vascular patterns in VASC or the characteristic symmetry in NV. In contrast, misclassified cases, such as a MEL predicted as AKIEC with 49.3% confidence, often involve ambiguous textures or peripheral artifacts that challenge the self-attention mechanism.

Table 2 Performance comparison of the benchmarked ViT architectures on the HAM10000 test set, evaluated by predictive metrics and computational efficiency.

Models	Accuracy	Precision	Recall	F1 Score	Params (M)	Gflops	Inference Time (Ms)
DeiT III-Base	0.9204	0.8794	0.8348	0.8544	85.82	33.6955	0.5674
ViT-Base	0.8952	0.8508	0.7704	0.8060	85.80	33.6955	0.5459
Swin-Base	0.9164	0.8703	0.8356	0.8508	86.75	30.3375	0.7927
PiT-B	0.9072	0.8677	0.8599	0.8606	72.75	21.1067	1.5418

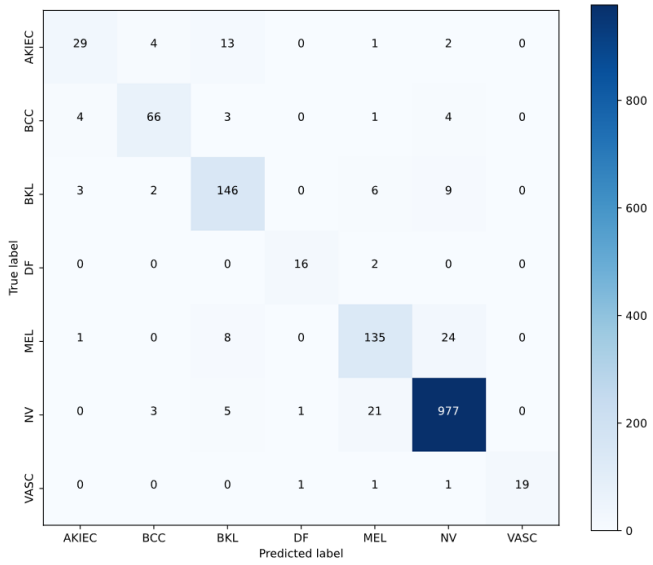


Figure 2 Confusion matrix for the top-performing DeiT III-Base model, illustrating class-specific diagnostic performance and inter-class misclassification patterns.

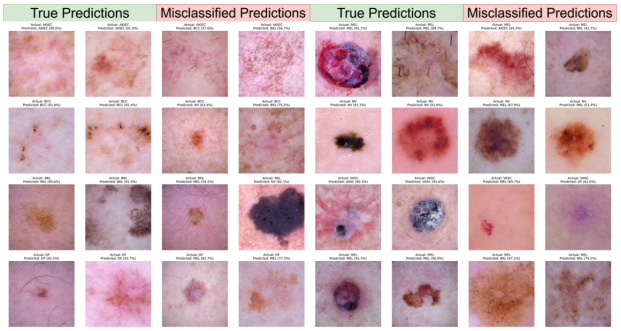


Figure 3 Qualitative analysis of model predictions showing representative examples of successful classifications and erroneous predictions with their corresponding confidence scores.

DISCUSSION

Interpretation of Key Findings

The results of this study underscore the effectiveness of ViTs in overcoming the locality bias of traditional CNNs. The top performance of DeiT III-Base and Swin-Base confirms that modeling global contextual relationships is vital for differentiating complex lesions. Specifically, the self-attention mechanism allows these models to attend to subtle, long-range dependencies across the lesion surface, which are often missed by the localized receptive fields of standard convolutional filters. Furthermore, the use of Transfer Learning from ImageNet-1K was essential in stabilizing training and achieving high accuracy despite the limited size of medical datasets compared to natural image repositories.

Clinical Implications, Limitations, and Future Directions

From a clinical perspective, the high accuracy and sub-millisecond inference times of models like DeiT III-Base suggest their viability as real-time CAD tools. Such systems could provide a critical "second opinion," potentially reducing the rate of unnecessary biopsies for benign lesions while ensuring that early-stage melanomas are not overlooked. However, this study has limitations. The HAM10000 dataset is significantly imbalanced, with NV accounting for over 60% of the samples. As seen in the confusion matrix (Figure 2), this imbalance can bias models toward predicting the majority class. Future research should focus on incorporating multi-modal data, such as patient age, anatomical location, and clinical history, to further refine diagnostic precision. Additionally, exploring Explainable AI (XAI) techniques, beyond the visual samples in Figure 3, will be crucial for building clinician trust and ensuring the transparency of DL models in high-stakes medical environments.

CONCLUSION

This study demonstrates the efficacy of advanced ViT architectures in the automated classification of pigmented skin lesions using the HAM10000 dataset. By leveraging global self-attention mechanisms to model long-range dependencies, our framework successfully overcame the locality limitations of traditional CNNs, with the DeiT III-Base architecture emerging as the superior model, achieving a peak accuracy of 92.04% and an F1-score of 85.44%. The comparative analysis revealed that while hierarchical structures like Swin-Base offer competitive diagnostic precision, the data-efficient strategies of DeiT III provide an optimal balance between predictive sensitivity and computational throughput, characterized by sub-millisecond inference times suitable for real-time clinical screening. Furthermore, the integration of transfer learning from large-scale natural image repositories was essential for stabilizing training and achieving high performance despite the inherent class imbalances within the dermoscopic data. Ultimately, these findings underscore the potential of transformer-based paradigms as robust CAD tools in dermatology. Future research will focus on the integration of multi-modal data, such as patient history and anatomical location, alongside XAI techniques to ensure the transparency and clinical reliability of DL deployments in high-stakes healthcare environments.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The dataset analyzed for this study is the public dataset, which is available on Kaggle: https://www.kaggle.com/datasets/surajghuwalewala/ham1000-segmentation-and-classification?select=GroundTruth.csv

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

2025 Skin cancer: Ham10000 dataset.

Armstrong, B. K. and A. Kricker, 1995 Skin cancer. Dermatologic Clinics **13**: 583–594.

Aruk, I., I. Pacal, and A. N. Toprak, 2026 A comprehensive comparison of convolutional neural network and visual transformer models on skin cancer classification. Computational Biology and Chemistry **120**.

Bengio, Y., 2012 Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the 29th International Conference on Machine Learning*, volume 27, pp. 17–36.

Bruno, A., A. Artesani, P. L. Mazzeo, F. Janan, G. Yang, *et al.*, 2025 Boosting skin cancer classification: A multi-scale attention and ensemble approach with vision transformers. Sensors **25**: 2479.

Cakmak, Y. and A. Maman, 2025 Deep learning for early diagnosis of lung cancer. Computational Systems and Artificial Intelligence **1**: 20–25.

Cakmak, Y. and I. Pacal, 2025 Comparative analysis of transformer architectures for brain tumor classification. Exploratory Medicine **6**.

Çakmak, Y. and N. Pacal, 2025 Deep learning for automated breast cancer detection in ultrasound: A comparative study of four cnn architectures. Artificial Intelligence in Applied Sciences **1**: 13–19.

Chaurasia, A. K., P. W. Toohey, H. C. Harris, and A. W. Hewitt, 2025 Multi-resolution vision transformer model for histopathological skin cancer subtype classification using whole slide images. Computers in Biology and Medicine **196**.

Dagnaw, G. H., M. El Mouhtadi, and M. Mustapha, 2024 Skin cancer classification using vision transformers and explainable artificial intelligence. Journal of Medical Artificial Intelligence **7**.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, *et al.*, 2020 An image is worth 16x16 words: Transformers for image recognition at scale.

Gloster Jr, H. M. and D. G. Brodland, 1996 The epidemiology of skin cancer. Dermatologic Surgery **22**: 217–226.

Gloster Jr, H. M. and K. Neal, 2006 Skin cancer in skin of color. Journal of the American Academy of Dermatology **55**: 741–760.

Jerant, A. F., J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, 2000 Early detection and treatment of skin cancer. American family physician **62**: 357–368.

Karthik, R., R. Menaka, S. Atre, J. Cho, and S. V. Easwaramoorthy, 2024 A hybrid deep learning approach for skin cancer classification using swin transformer and dense group shuffle non-local attention network. IEEE Access **12**: 158040–158051.

Liu, Z., Y. Lin, Y. Cao, *et al.*, 2021 Swin transformer: Hierarchical vision transformer using shifted windows.

Madan, V., J. T. Lear, and R.-M. Szeimies, 2010 Non-melanoma skin cancer. The lancet **375**: 673–685.

Manju, V. N., D. S. Dayana, N. Patwari, K. P. B. Madavi, and K. K. Sowjanya, 2025 Attention-enhanced vision transformer model for precise skin cancer detection. In *Proceedings of the 2025 International Conference on Emerging Technologies in Computing and Communication (ETCC)*, IEEE.

Ozdemir, B. and I. Pacal, 2025 An innovative deep learning framework for skin cancer detection employing convnextv2 and focal self-attention mechanisms. Results in Engineering **25**: 103692.

Pacal, I., M. Alaftekin, and F. D. Zengul, 2024 Enhancing skin cancer diagnosis using swin transformer with hybrid shifted window-based multi-head self-attention and swiglu-based mlp. Journal of Imaging Informatics in Medicine **37**: 3174–3192.

Pacal, I. and Y. Cakmak, 2025a A comparative analysis of u-net-based architectures for robust segmentation of bladder cancer lesions in magnetic resonance imaging. Eurasian Journal of Medicine and Oncology **9**: 268–283.

Pacal, I. and Y. Cakmak, 2025b *Diagnostic Analysis of Various Cancer Types with Artificial Intelligence*. Duvar Yayınları.

Ren, H., J. Guo, S. Cheng, and Y. Li, 2024 Pooling-based visual transformer with low complexity attention hashing for image retrieval. Expert Systems with Applications **241**: 122745.

Ren, Z., H. Zhang, T. Huang, *et al.*, 2022 Deep learning (cnn) and transfer learning: A review. Journal of Physics: Conference Series **2273**: 012029.

Sakib, A. H., M. I. H. Siddiqui, S. Akter, A. Al Sakib, and M. R. Mahmud, 2025 Levit-skin: A balanced and interpretable transformer-cnn model for multi-class skin cancer diagnosis. International Journal of Science and Research Archive **15**: 1860–1873.

Siegel, R. L., A. N. Giaquinto, and A. Jemal, 2024 Cancer statistics, 2024. CA: a cancer journal for clinicians **74**: 12–49.

ThangaPurni, J. and M. Braveen, 2025 Unified arp-vit-cnn system: Hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions. Array p. 100515.

Touvron, H., M. Cord, and H. Jégou, 2022 Deit iii: Revenge of the vit.

How to cite this article: Islam, M. S., Chamgoué, A. C. and Dubb, G. P. Benchmarking State-of-the-Art Vision Transformer Architectures for the Automated Classification of Pigmented Skin Lesions. *Computers and Electronics in Medicine*, 3(1), 42-47, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.48-53
https://doi.org/10.69882/adba.cem.2026016

Design and Development of a Low-Cost EMG-Controlled Prosthetic Hand

Fakeraldeen Mohamed Abdalla Ali *,[1], **Youssouf Fadile Raye Bowou** *,[2], **Ghassan Ali Mohammed Al-Shafali** *,[3] and **Kamal Abdulrahman Adam Wady** *,[4]

*Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Hitit University, Çorum, Türkiye.

ABSTRACT This study focuses on the design, development, and enhancement of a cost-effective myoelectric prosthetic arm intended for daily functional use. The initial prototype was fabricated using Fused Deposition Modeling (FDM) 3D printing technology with acrylonitrile butadiene styrene (ABS) as the base material. This approach enabled the creation of a lightweight, portable prosthetic arm with a human-like appearance and six degrees of freedom, allowing for the execution of essential daily activities. The actuation mechanism is based on an artificial tendon-driven system inspired by prior works such as the Vanderbilt Hand and the prosthetic hand developed at Hitit University. The tendon-driven structure allows for coordinated finger movement while preserving mechanical simplicity. Actuation is achieved using standard servo motors, controlled by surface electromyography (sEMG) signals acquired from the user's forearm muscles. The initial version of the device was constructed with a material cost of approximately 250, achieving a grip force of around 3 N per finger and a complete actuation cycle time of approximately 0.4 s. Despite demonstrating satisfactory functional performance, early user evaluations revealed challenges related to control intuitiveness and system integration, as reflected in user feedback surveys. To address these limitations and enhance usability, several technological upgrades were implemented. The original microcontroller was replaced with an Arduino Mega, and an ESP8266-07 Wi-Fi module was integrated to enable wireless communication. These enhancements significantly improved data transmission, real-time signal processing, and remote monitoring capabilities.

KEYWORDS
Electromyography (EMG)
Low-cost prosthetics
Myoelectric prosthetic arm
3D printing
Tendon-driven actuation

INTRODUCTION

The loss of an arm can be one of the most devastating events in a person's life, deeply affecting their dependence on others, mental health, and work capacity. Although advanced myoelectric prostheses provide very comfortable control for the user, their very high price, often over $20,000, makes them inaccessible for most people in the world (Belter et al. 2013). This financial obstacle causes a strong necessity for cheap but still functional alternatives.

Step by step, additive manufacturing technology has made it possible for individuals to create complex mechanical structures by themselves. This is evident in the example of the open-source InMoov humanoid project (Langevin 2014). At the same time, the growing availability of open-source microcontroller and sensor platforms has made it easier to develop complicated electronic systems for control. The current work took advantage of the combination of these two trends to build a fully functional myoelectric prosthetic arm at a low cost from scratch. The work was divided into two major stages: 1) initial design and demonstration of a fully functional 3D-printed arm, and 2) a major upgrade with a focus on the modernization of the control system through wireless communication. Mechanical design, component testing, and the application of a wireless control system were parts of the whole development process detailed in the present paper, which also discusses the advantages of the system and offers a platform for further studies in the field of advanced bio-mechatronics (Brooker 2012).

The primary objective of this study was to design and fabricate a fully functional, portable myoelectric prosthetic arm from scratch, while maintaining a total material cost below 500. This goal was set in contrast to commercially available devices such as the Bebionic 3 (RSL Steeper 2014) and i-Limb (Touch Bionics 2014), which are significantly more expensive. The developed prototype was evaluated in terms of grip strength, actuation speed, weight, durability, and control reliability. The performance metrics obtained were benchmarked against both commercial and research-grade prosthetic systems, based on values reported in the literature (Belter et al. 2013; exrx.net 2023). Furthermore, key limitations of the initial prototype were identified particularly those related to control intuitiveness and system expandability both of which are commonly reported challenges in myoelectric prosthetic systems (Pylatiuk et al. 2007; Lake 2010). In response, a systematic upgrade process was implemented, which included replacing the original microcontroller with an Arduino-based platform and integrating a Wi-Fi module (ESP8266-07). These improvements enabled wireless control and data telemetry, laying the groundwork for advanced,

Manuscript received: 21 December 2025,
Revised: 20 January 2026,
Accepted: 23 January 2026.

[1]fkhri.mohammed.6@gmail.com (**Corresponding author**)
[2]rayebowou@gmail.com
[3]ghassanalshafali@gmail.com
[4]wadyk44@gmail.com

cloud-based signal processing techniques beyond conventional EMG algorithms (Hudgins et al. 1993).

MATERIALS AND METHODS

The methodology for this particular project proceeded through two comprehensive phases: first, the initial construction of a functional prosthetic arm that was low-cost, and then a massive upgrade to the control technology. This encompassed mechanical drawing, 3D printing, electronic parts integration, firmware development, and system validation.

Data Acquisition and Signal Processing

The research utilized two disposable Ag/AgCl electrodes (Kendall™ MediTrace™ ECG electrodes), which were arranged in a bipolar montage to record surface electromyography (EMG) signals. The first electrode pair was positioned on the biceps brachii muscle belly, while the second pair extended across the forearm flexor group, with their midpoint located at one-third the distance between the elbow and wrist. A common ground electrode was attached to the olecranon process to reduce common-mode interference. The inter-electrode distance was maintained at 20 mm (center-to-center). The Muscle Sensor V3 modules from Advancer Technologies served as raw myoelectric signal conditioning devices, providing on-board amplification (approximately 1000 gain), band-pass filtering (20–450 Hz using 2nd-order Butterworth), and full-wave rectification with a smoothing RC circuit ($\tau = 50$ ms).

The PIC18F25K22 microcontroller sampled the conditioned analog signals (ranging from 0 to 3 V) using its built-in 10-bit Analog-to-Digital Converter (ADC) at a constant sampling frequency of 1 kHz. The firmware applied a moving-average filter with a window length of 5 samples to reduce high-frequency noise, while a software-based 50 Hz notch filter eliminated mains interference. The system was powered by a 2-cell Lithium Polymer battery (7.4 V, 1600 mAh), and two low-dropout linear regulators (AP1117 series) provided stable 5 V and 3.3 V outputs to protect the sensitive analog front-end from switching noise and voltage fluctuations caused by servo motor loads. The complete acquisition system operated under typical daily conditions, including motion artifacts and electromagnetic interference, to ensure signal integrity.

Design and Fabrication of the Core Prosthetic Arm

Mechanical Design and Additive Manufacturing: The mechanical design was determined by the low cost, human-like shape, and basic functionality aims. A modular architecture was formed of a hand/wrist module and a separable forearm/elbow module that allowed for different levels of amputation and application (Lake 2010). The prosthetic hand offers six degrees of freedom (DOF): independent bending of the thumb, index, and middle fingers; combined bending of the ring and pinky fingers; 180° wrist rotation; and 110° elbow bending. It uses five TowerPro MG996R standard hobby servo motors, each providing a stall torque of 10 kg-cm at 6 V. Finger bending is powered by a tendon-based mechanism, which uses high-strength, low-stretch braided polyethylene fishing line (approx. 50 lb test). This line runs through internal channels in the 3D-printed phalanges and is secured at the tips. The tendons come together at a narrow opening in the wrist. This design allows for palm rotation but causes some cable overlap during extreme rotation movements.

Tendon tension comes from the servos located in the forearm. Custom 3D-printed servo horns serve as winches. To maintain the strength of the printed ABS parts especially in stressed areas like the wrist pivot and finger joints the firmware limits the servo command range. By restricting the pulse width modulation (PWM) signal for each servo (usually between 1000–2000 µs), the movement of each servo horn is capped. This limits the maximum force the tendons can apply. This software-based mechanical stop ensures that the grip strength at each fingertip remains around 3 N (approx. 300 g). This safe limit was determined through testing to prevent permanent damage or breakage.

Hand and Finger Assembly: A tendon-driven actuation mechanism was chosen because of its benefits in joint congruity and easy mechanical design compared to complicated linkage systems (Wiste et al. 2009; Melchiorri et al. 2013). Each finger was designed in SolidWorks with three phalanges, closely resembling human anatomy (ElKoura & Singh 2003). 3D printing was done using a fused deposition modeling (FDM) printer and ABS plastic filament. Finger segments were fastened with 3 mm polypropylene pin joints. The high-strength, non-stretch braided fishing line was passed through internal channels and secured at the distal phalanx, creating the artificial tendon. Each joint was provided with a passive extension spring that would return the finger to its open position once relaxation of the tendon occurred.

Actuation and Drive Train: The pathways of the tendons were set through the palm and wrist, and then stopped at the palm of the hand where custom 3D-printed servo horns were located. Within the forearm were housed standard hobbyist servo motors (TowerPro MG996R). The servos were assigned for the thumb, index, and middle fingers while the ring and pinky fingers were linked to minimize actuator count.

Wrist and Elbow Joints: The wrist joint allowed nearly 180° of rotation by placing the palm structure directly on the servo output shaft. The elbow joint was made to carry the weight of the forearm. A bespoke 2.1:1 reduction gear train was created and 3D-printed to increase the servo motor's output torque, resulting in around 110° of flexion.

Post-Processing: All ABS parts that were 3D printed were treated with acetone vapor as a post-processing step in order to achieve better bonding of the layers and to improve the strength of the structure overall (i.materialise 2014).

Electrical System and Initial Control Logic: The first electrical layout was created to make myoelectric operation portable and useable.

Power Management: The system was powered by one 7.4V, 1600mAh Lithium-Polymer (LiPo) battery. The voltage was stabilized by several 5V and 3.3V low-dropout regulators, which fed the servo motors and the microcontroller, respectively, thus ensuring smooth and stable operation.

Signal Acquisition: Myoelectric control was achieved by two commercial single-channel EMG sensor kits (e.g., Muscle Sensor V3). These modules did the on-board amplification, rectification, and smoothing of the raw bioelectric signals that were picked up by the surface electrodes placed on the user's residual limb (Day 2010; Cotton et al. 2014).

Microcontroller and Basic Firmware: The PIC18F25K22 microcontroller was the main processor. Firmware was written that would read the analogue EMG signals through its analogue-to-digital converter (ADC). A basic control algorithm was put in place where one EMG signal would switch between preset device "states" (e.g., hand open/close mode, wrist rotation mode) and the second EMG signal would move the state that was selected. This finite-state machine method is a typical simple strategy for multifunctional control (Hudgins et al. 1993).

System Control Logic: The PIC18F25K22 microcontroller was the main processor. Firmware was written that would read the

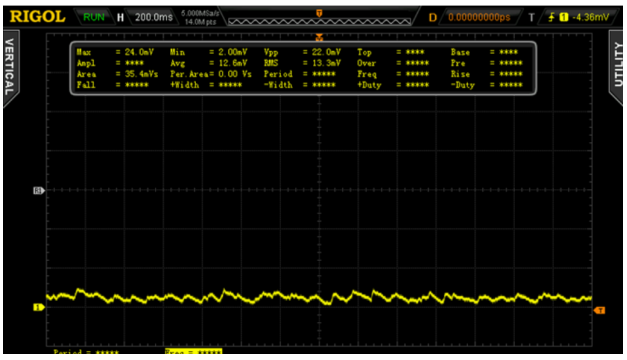


Figure 1 Light flex output (200ms/div, 50mV/div) showing EMG signal acquisition for low-intensity muscle contraction

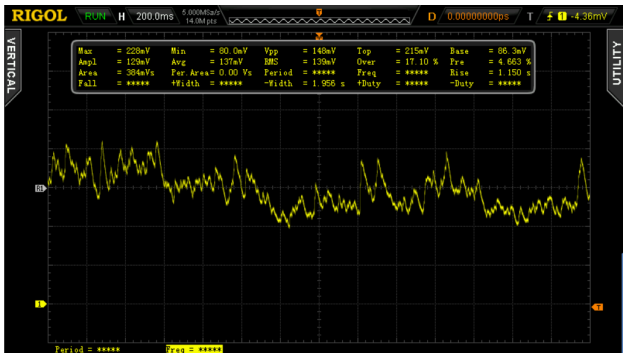


Figure 2 Strong flex output (200ms/div, 50mV/div) showing EMG signal acquisition for high-intensity muscle contraction

analogue EMG signals through its analogue-to-digital converter (ADC). A basic control algorithm was put in place where one EMG signal would switch between preset device "states" (e.g., hand open/close mode, wrist rotation mode) and the second EMG signal would move the state that was selected. This finite-state machine method is a typical simple strategy for multifunctional control (Hudgins et al. 1993).

The device uses a hybrid control design that merges a finite-state machine (FSM) for mode selection with proportional, amplitude-based control for actuation within each mode. It continuously monitors surface EMG signals from two muscle sites, usually the biceps and forearm extensors. The signal from the main muscle site is rectified and averaged over a 200 ms window. When this processed amplitude goes above a set threshold (determined during a user-specific training session at about 20% of maximum voluntary contraction), the system moves to the next grip state in order: Rest, Precision Grip (e.g., pinch), Power Grip, Wrist Rotation, Rest. This setup enables access to multiple functions using just one control signal.

When a specific state is active, proportional control is activated with the continuous amplitude of the secondary EMG signal. This signal is normalized and mapped linearly to the servo command. The relationship is defined by:

$$W_{\mathrm{PWM}}(t) = a + k \cdot \mathrm{EMG}_{\mathrm{norm}}(t)$$

In this equation, W_{PWM} represents the pulse width sent to the servo (in microseconds), a is the baseline pulse width for the open-hand position (typically 1000 µs), k is a gain constant that scales the normalized EMG amplitude ($\mathrm{EMG}_{\mathrm{norm}} \in [0, 1]$) to a practical pulse width range (e.g., 1000–2000 µs), and t indicates time. This mapping allows users to easily control both the speed of finger closure and the grip force by adjusting muscle contraction intensity. For example, a gentle flexion leads to slow, light closure for delicate items, while a strong contraction results in fast, strong grasping. The combination of state-based selection and proportional control offers a practical and intuitive way to perform sequential tasks with different force needs.

The EMG signal acquisition quality, as demonstrated in Figures 1 and 2, shows effective signal capture for both light and strong muscle contractions. The voltage levels recorded (22.0 mV peak-to-peak for light flex and 14.6 mV for strong flex) fall within the expected range for surface EMG signals (Day 2010). The ADC sampling mechanism illustrated in Figure 3 effectively converts analog EMG signals to digital control signals for servo positioning.

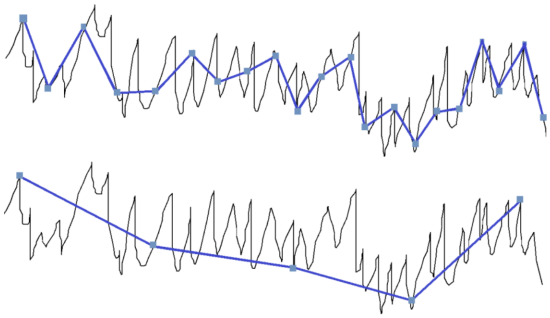


Figure 3 EMG signals (black) with ADC samples (blue). The voltage level of these samples controls servo positions, demonstrating the proportional control mechanism.

Wireless Control System Upgrade and Integration

In order to improve the limitations of control flexibility and system isolation, strong upgrades were undertaken, aiming at modern and accessible hardware with wireless connectivity.

Hardware Modernization: The core electronic backbone was replaced for more versatile and community-supported components.

Controller Replacement: Replacing the PIC microcontroller was an Arduino Mega 2560 development board. This platform was chosen due to its large I/O capability, easy C++ programming environment, and enormously rich ecosystem of libraries to speed the development and prototyping cycle.

Wireless Communication Module: An ESP8266-07 Wi-Fi transceiver was incorporated with a serial UART interface connection to the Arduino. This module permits the prosthesis to connect to the local wireless network for data transfer.

Sensors and Actuators Interface: The old EMG sensor boards, servo motors, and power system nicely interfaced with the new Arduino controller, confirming backward compatibility of the upgrade path.

Firmware Development and Network Architecture: New firmware was developed for the Arduino to set up a bidirectional data pipeline.

Sensor Firmware On the Device: The Arduino was programmed to continuously sample the analogue EMG signals packetize this data along with system status parameters (e.g., battery voltage), and send it to a specified network address via the ESP8266 module.

External processing and control server: To this end, a companion software application was developed in Python, which runs on

a standard desktop computer or at a cloud server. This application: (1) listened to and received the incoming wireless data stream from the prosthesis; (2) allowed implementing advanced signal processing techniques, filtering or machine learning algorithms on the received EMG data (Shedeed et al. 2013; Hudgins et al. 1993); and (3) correspondingly sent actuator control commands back to the Arduino via the Wi-Fi link.

Control Interface Diversification: It decoupled control logic from the physical hardware. Different control interfaces could be created, for example, a simple graphics application on a paired smartphone that allows touch-based control or visualization of system sensor data in real-time.

System Integration and Functional Validation

Integration of the total systems occurred once both portions of the development had been completed, followed by functional validation.

Mechanical Assembly and Tuning: The assembly of 3D-printed parts, tendon tensioning, and calibration of tuning ranges for the servos were performed. Structural integrity of main joints, particularly wrist joints, was tested under actual working loads.

Electrical Integration and Testing: All electronic subsystems were coupled together and tested for power delivery, signal integrity, and for alignment with the defined communication protocol. Wireless characteristics of latency and reliability were tested.

Holistic Functionality Tests: Final integral prototypes (original and upgraded) were activated and made to perform commanded gestures and grasping in sequences. Performance metrics were noted and qualitatively evaluated, including response latency, smoothness of motion, and wireless control stability, confirming the operational success of both designs.

The entire work used an iterative design-build-test methodology. Solidworks was used to design the initial concepts; 3D printing was then used to test the design. After the initial test, mechanical and electrical subsystems were designed. After this, system integration and performance testing were done. This methodology is in line with most other prosthetic device development (Weir 2003). Upgrading to wireless control also used the same iterative methodology; this was done with a focus on firmware and network integration.

Data Collection Methods

Data was collected via direct measurement and controlled testing:

- **Grip Force:** This was measured with prosthetic fingers fully closed and the scales were used as a functional measure. This is practical as the closing force is realistic.
- **Actuation Speed:** A crucial performance measure for user acceptance (Lake 2010), this was measured by timing using a high-speed camera to record the flexion and extension cycles of the fingers.
- **System Weight:** The system was weighed on a precision scale. Weight is an essential factor when considering the comfort of the socket and the suspension (Brooker 2012).
- **Battery Life:** This was estimated using typical operating cycles and measured as average current draw and the mAh of the battery.
- **Control Accuracy & Latency:** For the wireless system, signal transmission delay was measured from EMG input to actuator response.

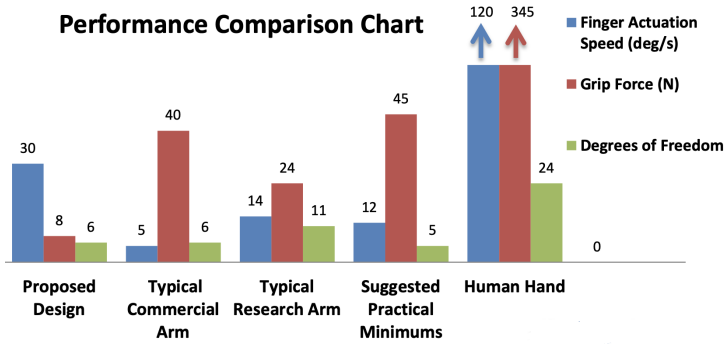


Figure 4 Performance comparison of the proposed prosthetic arm with commercial and research arms, along with the human hand and suggested practical minimums. Values taken from Belter et al. (2013) and exrx.net (2023). The chart shows finger actuation speed (deg/s), grip force (N), and degrees of freedom for different categories.

FINDINGS AND RESULTS

The work proves that it is possible to create and construct an operational myoelectric prosthetic arm with multiple degrees of freedom at a very low cost compared to commercial devices. The performance, although not equal to that of the high-end commercial products, confirms the core design principles and offers a considerable baseline.

As shown in Figure 4, the proposed design achieves a balanced compromise between cost and functionality. The grip force of approximately 3 N per finger, while lower than commercial devices, is sufficient for many activities of daily living. The actuation speed of 0.4 s compares favorably with both commercial and research prosthetics, addressing user concerns regarding responsiveness (Lake 2010).

Quantitative performance evaluation focused on two main functional metrics: actuation speed and reliable gripping capability. The wrist rotation mechanism, powered by a dedicated servo, completes a full 180° arc (from full pronation to full supination) in 0.45 seconds. This results in an average rotational speed of 400°/s, which exceeds many commercial myoelectric prostheses and approaches the reflexive movement of a biological wrist. We characterized finger actuation performance through systematic grasping tests with standardized objects, such as cylinders, blocks, and various everyday items. When all fingers engaged in a full power grip, the hand showed a static lifting capacity of about 600 g. This capacity comes from the combined strength of each finger (around 300 g per fingertip), limited by the programmed servo travel range to protect the integrity of the 3D-printed joints.

While the actuation speed of the system is a clear advantage, its current open-loop control has a drawback: when encountering an object, the servos keep applying tension until they hit their pre-programmed endpoint. This can cause high internal stresses, servo stall, and long-term component wear. To address this, future versions are set to include fingertip pressure sensors, like force-sensitive resistors or piezoresistive films. This feedback will create a closed-loop grip-force control system, where servo motion can be dynamically stopped or adjusted upon reaching a specific contact pressure. This improvement would protect the mechanical structure and save battery power while giving the user better control over grip strength, enabling safe handling of fragile or compliant objects. This is a vital step toward more natural and adaptive

prosthetic manipulation.

Table 1 System Specifications and Performance Parameters

Parameter	Value	Description
Electronics		
Microcontroller	PIC18F25K22 Arduino Mega	8-bit / 16MHz processing core
ADC Resolution	10-bit	1024 discrete sensitivity levels
Sampling Frequency	500 Hz	EMG data capture rate
Communication	Wi-Fi/Serial	Telemetry and logic interface
Mechanical		
Degrees of Freedom	6 DOFs	5 fingers + wrist rotation
Actuator Type	TowerPro MG996R	Metal gear servos
Stall Torque	10 kg-cm	Maximum rotational force
Wrist Rotation	180° in 0.45s	Travel speed (400°/s)
Performance		
Grip Force (Max)	600g	Combined power grasp force
Total Weight	480g	With motors and 3D frame
Material	ABS Thermoplastic	FDM 3D printed

Table 2 Quantitative Performance Metrics Comparison

Parameter	Measured	Benchmark	Ref.
Material Cost	$250	$20,000+	Belter et al. (2013)
Grip Force (per finger)	3 N	10 N to 15 N	exrx.net (2023)
Actuation Speed	0.4 s	0.5 s to 0.8 s	Lake (2010)
System Weight	950 g	800 g to 1200 g	Belter et al. (2013)
Degrees of Freedom	6	6–24	Belter et al. (2013)

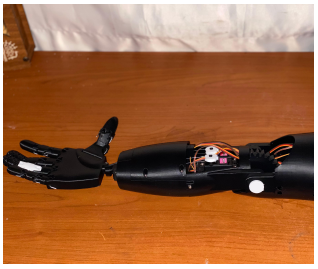
Table 1 and Table 2 provide a detailed comparison of the prototype's performance against commercial benchmarks. The cost reduction by a factor of 100 represents a significant achievement in making prosthetic technology more accessible.

The prosthetic hand demonstrated versatile grasping capabilities as shown in Figure 5. The four grasping patterns pinch grip, two variations of power grip, and handle grip enable the user to perform a wide range of daily activities. The pinch grip (Figure 5a) allows for precise manipulation of small objects such as pens, coins, or keys. The power grips (Figures 5c and 5b) provide stable holding of variously sized objects like water bottles, books, or tools. The handle grip (Figure 5d) is particularly useful for cylindrical objects such as cups, mugs, or door handles.

(a) Pinch Grip

(b) Power Grip

(c) Before the final wiring



(d) Project Final Form

Figure 5 Demonstration of various grasping patterns achieved by the prosthetic hand: (a) Pinch Grip for precision tasks, (b) Power Grip for medium objects, (c) the Project before connecting the Battery and finishing the wiring for larger objects, and (d) Represents The project's Final Form . These patterns demonstrate the hand's versatility in activities of daily living.

CONCLUSION

The successful integration of an Arduino-based controller with an ESP8266 Wi-Fi module represents a significant evolutionary milestone in the development of the proposed prosthetic system, effectively transforming the device into an open and connected platform. This architectural shift addresses several inherent limitations of traditional standalone myoelectric prostheses by enabling remote monitoring, improved computational offloading, and the development of personalized user interfaces. Through wireless connectivity, the system gains flexibility, scalability, and adaptability, which are essential characteristics for modern assistive technologies.

Building upon this foundation, future research efforts will focus on the implementation of machine learning algorithms for real-time gesture classification using streaming EMG data, thereby enhancing control accuracy and user intuitiveness. In addition, the integration of fingertip pressure sensors is planned to enable closed-loop grip force regulation, which is expected to improve object manipulation safety and reliability (Dhillon & Horch 2005). Further work will also include the design and fabrication of a more robust, custom-printed circuit board (PCB) to improve system durability, compactness, and long-term operational stability. This study provides a comprehensive and easily replicable blueprint for the development of low-cost myoelectric prosthetic systems, while also offering a forward-looking framework that supports future enhancements in intelligence, connectivity, and user-centered design. By bridging affordability with advanced control and communication capabilities, the proposed approach contributes meaningfully to the ongoing advancement of biomechatronics and supports the development of the next generation of economical, smart, and user-friendly assistive devices (Brooker 2012).

Acknowledgments

This work was supported by Hitit University Scientific Research Projects Unit. The authors would like to thank the Department of Electrical and Electronics Engineering for providing the necessary

facilities and support.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Belter, J. T., Segil, J. L., Dollar, A. M., & Weir, R. F. (2013). Mechanical design and performance specifications of anthropomorphic prosthetic hands: A review. *Journal of Rehabilitation Research and Development*, 50(5), 597-618.

Brooker, G. (2012). *Introduction to Biomechatronics*. Scitech Publishing.

Cotton, D. P. J., Cranny, A., White, N. M., & Chappell, P. H. (2014). Control strategies for a multiple degree of freedom prosthetic hand. *Measurement and Control*, 47(2), 49-54.

Day, S. (2010). Important factors in surface EMG measurement. *Bortec Biomedical Ltd. Technical Report*, 1-12.

Dhillon, G. S., & Horch, K. W. (2005). Direct neural sensory feedback and control of a prosthetic arm. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(4), 468-472.

ElKoura, G., & Singh, K. (2003). Handrix: Animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (pp. 110-119).

exrx.net. (2023). Body segment data. Retrieved from https://exrx.net/Kinesiology

Hudgins, B., Parker, P., & Scott, R. N. (1993). A new strategy for multifunctional myoelectric control. *IEEE Transactions on Biomedical Engineering*, 40(1), 82-94.

i.materialise. (2014). ABS 3D printing design guide. Retrieved from https://i.materialise.com/blog/abs-3d-printing-design-guide

Lake, C. (2010). Partial hand amputation: Prosthetic management. In *Atlas of Amputations and Limb Deficiencies* (4th ed., pp. 189-198). American Academy of Orthopaedic Surgeons.

Langevin, G. (2014). InMoov: Open source 3D printed life-size robot. Retrieved from https://inmoov.fr/

Melchiorri, C., Palli, G., Berselli, G., & Vassura, G. (2013). Development of the UB hand IV: Overview of design solutions and enabling technologies. *IEEE/ASME Transactions on Mechatronics*, 18(4), 1441-1451.

Pylatiuk, C., Schulz, S., & Doderlein, L. (2007). Results of an internet survey of myoelectric prosthetic hand users. *Prosthetics and Orthotics International*, 31(4), 362-370.

RSL Steeper. (2014). *Bebionic 3 Technical Information*. RSL Steeper Ltd.

Shedeed, H. A., Issa, M. F., & El-sayed, S. M. (2013). Brain EEG signal processing for controlling a robotic arm. In *2013 8th International Conference on Computer Engineering & Systems (ICCES)* (pp. 152-157). IEEE.

Touch Bionics. (2014). *i-limb digits clinician user manual*. Touch Bionics Ltd.

Weir, R. F. (2003). Design of artificial arms and hands for prosthetic applications. In M. Kutz (Ed.), *Standard Handbook of Biomedical Engineering and Design* (pp. 32.1-32.39). McGraw-Hill.

Wiste, T. E., Dalley, S. A., Withrow, T. J., & Goldfarb, M. (2009). Design of a multifunctional anthropomorphic prosthetic hand with extrinsic actuation. In *2009 IEEE International Conference on Rehabilitation Robotics* (pp. 675-681). IEEE.

How to cite this article: Ali, F. M. A., Bowou, Y. F. R., Al-Shafali, G. A. M., and Wady, K. A. A. Design and Development of a Low-Cost EMG-Controlled Prosthetic Hand. *Computers and Electronics in Medicine*, 3(1), 48-53, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.54-59
https://doi.org/10.69882/adba.cem.2026017

Towards Robust CAD Systems for Digital Pathology: Evaluating Transformer-Based Backbones for Breast Cancer Classification

Seref Koyuncu *,1, **Yigitcan Cakmak** *,2 **and Ishak Pacal** *,β,3

*Department of Computer Engineering, Faculty of Engineering, Igdir Universty, 76000, Igdir, Türkiye, βDepartment of Electronics and Information Technologies, Faculty of Architecture and Engineering, Nakhchivan State University, AZ 7012, Nakhchivan, Azerbaijan.

ABSTRACT Breast cancer is one of the greatest global health burdens today and demands accurate diagnosis because of the vast histological variety. CNN-based systems had been the dominant technology in Digital Pathology, but with their inability to create a global representation has allowed other technologies such as Vision Transformers to compete. This paper evaluate the performance of three different transformer-based backbone architectures (DeiT Base, Swin Base, and ViT Base) for classifying breast histopathological images into eight granular classes using the BreaKHis database. To facilitate this comparison, we utilize transfer learning and distinct data augmentation methods. Each architecture was fine-tuned to classify four benign and four malignant subtypes with a minimum reported accuracy of 94%, with Swin Base performing more optimally than either of the other two approaches, obtaining highest reported accuracy of 0.9511 and an F1 score of 0.9434. The unique design and shifted windowing processes of Swin Base have allowed this architecture to capture detailed nuclear information as well as the larger context regarding breast cancers, to an extent greater than the other two architectures. Additionally, we provide an in-depth study of confusion matrices in conjunction with high classification accuracy, even when dealing with minor morphological overlap, to further support their claim regarding the ability of Swin Base and the remaining transformer architectures to successfully differentiate between histologically similar classes.

KEYWORDS
Breast cancer
Histopathology images
Multi-class classification
Computer-aided diagnosis
Digital pathology

INTRODUCTION

Breast cancer remains one of the most formidable challenges in global healthcare, consistently ranking as a leading cause of oncological mortality among women worldwide (Getu *et al.* 2025; Chikkala *et al.* 2025). The complexity of the disease, characterized by its high histological heterogeneity, demands diagnostic precision that is both rapid and highly accurate. In recent years, the field of pathology has undergone a significant paradigm shift with the advent of Digital Pathology (DP) and Whole Slide Imaging (WSI) (Karthiga *et al.* 2025; Hayat *et al.* 2024). This transition from traditional glass slides to high-resolution digital patches has paved the way for Computer-Aided Diagnosis (CAD) systems to assist pathologists in navigating the immense workload and reducing the inherent subjectivity of manual assessments (Murphy and Singh 2024; Logu and Thangaraj 2024; Asha 2025).

Deep learning and convolutional neural networks (CNNs) are transforming the way Computer-aided design (CAD) systems operate (Singh and Kaswan 2024; Alshehri 2025). Early CAD systems relied very heavily on the handcrafted features created by DLT and modified by the CAD software (Abdulaal *et al.* 2024b; Behzadpour *et al.* 2024). CNV models are designed specifically to recognize and represent locality patterns at a small scale (in a local-neighborhood context) within an image, constraining their ability to relate and connect long-range or distant regional features to each other and to map across an image with respect to the overall context (Ramamoorthy *et al.* 2024; Ukwuoma *et al.* 2025; Jackson *et al.* 2025). This limitation has led researchers to the Vision Transform is (ViT) architecture, in which all patches of an image processed sequentially to incorporate their information into a complete image representation (Kansal *et al.* 2025; Chitta *et al.* 2025; Jakkaladiki and Maly 2024). However, before ViTs can be routinely used in clinical practice, the full measure of robustness, computational efficiency, and classification accuracy across the full spectrum of breast cancer subtypes must be assessed and validated (Akshaya *et al.* 2024; Abdulaal *et al.* 2024a; Simonyan *et al.* 2024).

The primary objective of this study is to evaluate the efficacy of transformer-based backbones, specifically DeiT Base, Swin Base, and ViT Base, in the multi-class classification of breast histopathology images. Unlike binary classification tasks that merely distinguish between benign and malignant tissue, our work tackles an 8-class challenge involving specific subtypes: Adenosis, Ductal Carcinoma, Fibroadenoma, Lobular Carcinoma, Mucinous Carcinoma, Papillary Carcinoma, Phyllodes Tumor, and Tubular Adenoma. This granular approach is essential for providing clinicians

Manuscript received: 1 September 2025,
Revised: 28 November 2025,
Accepted: 30 November 2025.

[1]serefkoyuncu77@gmail.com
[2]ygtcncakmak@gmail.com(**Corresponding author**)
[3]ishakpacal@igdir.edu.tr

with the detailed diagnostic insights necessary for personalized treatment planning.

RELATED WORKS

The clinical management of breast cancer has been fundamentally reshaped by the transition to DP and the subsequent integration of CAD systems. Historically, pathologists relied on manual microscopic examination, a process prone to inter-observer variability and high cognitive load. Recent advancements in DL have mitigated these challenges by enabling the automated extraction of high-level features that capture the complex morphological heterogeneity of breast tissue. Current research in this domain focuses on enhancing diagnostic reliability through hybrid architectures, ensemble strategies, and novel feature descriptors, particularly when dealing with the granular multi-class classification of histological subtypes.

Ogundokun *et al.* (2024) proposed a robust hybrid DL framework that integrates CNN with Artificial Neural Networks (ANN) specifically for 400x magnification images. By utilizing a dual-pathway approach for feature processing, their model achieves exceptional diagnostic precision, reaching a near-perfect accuracy of 99.94%, which significantly reduces the margin for diagnostic error in high-resolution histopathology. Gul (2025) introduced an innovative textural analysis method termed Quad Star Local Binary Pattern (QS-LBP) paired with a customized 20-layer CNN architecture. This combination effectively encodes fine-grained tissue textures, allowing the system to outperform existing methodologies in distinguishing subtle morphological differences between benign and malignant growths, boasting a peak accuracy of 98.27

Rajaram *et al.* (2024) conducted a systematic evaluation of various ResNet architectures to determine their effectiveness in classifying BreaKHis dataset samples. Their comparative study highlighted that deeper residual networks, when combined with transfer learning, provide a highly scalable solution for multi-class classification, successfully capturing the intrinsic hierarchical patterns of cancerous cell structures. Balasubramanian *et al.* (2024) developed an ensemble learning strategy that fuses the outputs of VGG16, ResNet34, and ResNet50 models to tackle both cancer subtyping and invasiveness. This multi-model approach increases the overall stability of the CAD system and minimizes the risk of misclassification by leveraging a diversified feature set from multiple architectural families.

This work on the performance benchmarking of the BreaKHis dataset through the application of SVM on the textural features as proposed by Thakur *et al.* (2025) sets the standard for the BreaKHis dataset. This work on the benchmarking of performance through the experimental work on patients according to the patient-centric benchmarking methodology of this work on the rigorous benchmarking of patients helps to prevent data leakage. Aldakhil *et al.* (2025) provided a comprehensive review and performance assessment of transfer learning-based architectures, specifically ResNet18, Inception-V3, and ShuffleNet. Their findings emphasize that fine-tuned pre-trained models offer a computationally efficient path toward autonomous breast cancer detection, achieving high accuracy rates while significantly reducing the training time and data requirements typically associated with training from scratch.

MATERIALS AND METHODS

Dataset and Data Preprocessing

The BreaKHis (Breast Cancer Histopathological Image Classification) dataset (Dataset 2025), which is widely used as a benchmark in the field of digital pathology (DP), was used to conduct experiments and evaluate the proposed transformer-based architectures. This dataset contains a complete set of all types of Normal Breast Tissue (eight types), and there are two distinct classes of breast tissue; four (four types) benign, and four (four types) malignant. The dataset consists of 1995 high resolution images. The training and test sets were created by utilizing a structured data split, where 70% of the datasets is used to train the models (1393 samples), 15% is used for validation (295 samples), and 15% is used for independent testing (307 samples) in order to ensure that each model is trained appropriately and evaluated fairly. The exact breakdown of the samples across the eight types of breast tissue is provided in Table 1 and depicts the inherent imbalances found in clinical datasets.

The visual complexity and histological diversity of the samples are demonstrated in Figure 1, where some sample images of the eight histopathological classes are shown. These sample images demonstrate the level of cellular architecture similarity between the different classes, an issue that has often been known to make the task of pathological evaluation challenging when performed on a slide by slide basis. To eventually preprocess the visual data in a form conducive to the transformers chosen in this work as their base models, namely DeiT, Swin, and ViT, some image processing operations such as resizing and normalization have been performed on the sample images. Moreover, in order to help the transformers attain the required level of generalizability and avoid the problems of a small number of samples in classes such as Adenosis, a data augmentation process has been followed.

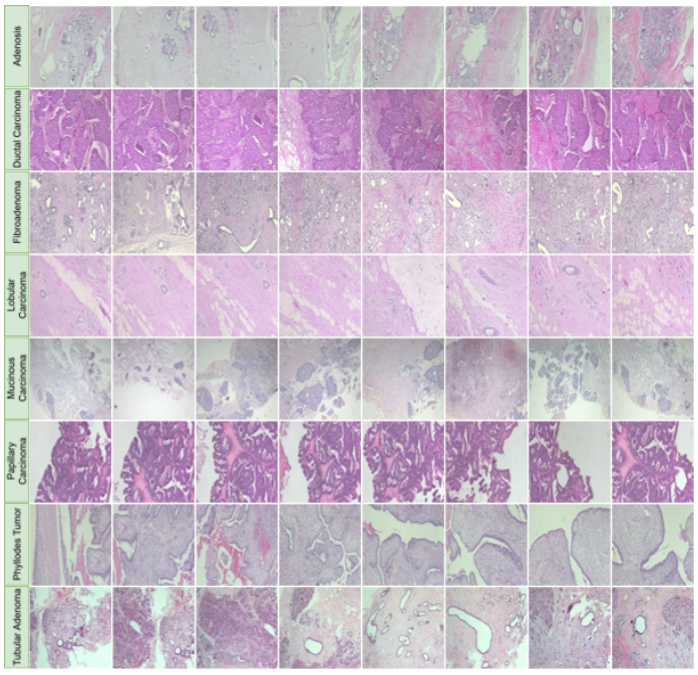


Figure 1 Sample images from the breast histopathology classes in the dataset.

Vision Transformers (ViTs)

The advent of ViTs marks the beginning of an important shift in the paradigm of medical image processing, diverging from the local receptive field requirement imposed by conventional CNNs. Although CNNs excel in local texture description, their efficacy in describing global dependencies in the image is inherently

Table 1 Dataset Distribution by Class.

Class	Original Dataset	Train Set (70%)	Validation Set (15%)	Test Set (15%)
Adenosis (ADE)	114	79	17	18
Ductal Carcinoma (DC)	864	604	129	131
Fibroadenoma (FIB)	253	177	37	39
Lobular Carcinoma (LC)	156	109	23	24
Mucinous Carcinoma (MC)	205	143	30	32
Papillary Carcinoma (PC)	145	101	21	23
Phyllodes Tumor (PT)	109	76	16	17
Tubular Adenoma (TA)	149	104	22	23
Total	1995	1393	295	307

limited, making it difficult to capture the essence of complicated histopathological patterns without considering the whole image in the process. This is where ViTs accentuate the advantage with their patch-based division of the image into a series of discrete segments and self-attention techniques, which help the model extract correlations even in the far-off regions of cells right from the beginning, opposed to CNNs, which tend to concentrate on local cell morphologies without emphasis on global architecture in digital pathology, where the difference in tissue might be quite subtle and relied on entirely on the global pattern.

In this study, we evaluate three distinct transformer-based architectures, ViT Base (Dosovitskiy *et al.* 2025), DeiT Base (Touvron *et al.* 2025), and Swin Base (Liu *et al.* 2025), to determine their effectiveness in classifying eight specific breast cancer subtypes. While the standard ViT Base provides a robust foundation for learning global representations, the Data-efficient Image Transformer (DeiT Base) is specifically designed to perform effectively on smaller datasets through a teacher-student distillation strategy, making it highly relevant for specialized medical imaging tasks where data can be scarce. Furthermore, we investigate the Swin Transformer (Swin Base), which introduces a hierarchical structure and shifted windowing scheme. This approach allows the model to process images at multiple scales, effectively capturing both fine-grained nuclear details and broader tissue patterns, which is vital for the granular 8-class classification challenge involving diverse benign and malignant subtypes.

Transfer Learning and Data Augmentation Strategy

The size and variability of the BreaKHis dataset and the depth of the deep transformer models make deep learning converge inefficiently. To counter these issues, we employed the concept of transfer learning. As ViTs are huge and demand the availability of vast amounts of data to train and converge properly, doing so on the relatively small number of images in the histopathology domain would be inefficient and would indulge the model into the problem of overfitting. To overcome this problem, we made use of the PyTorch Image Models library to load the pre-trained models of DeiT Base, Swin Base, and ViT Base pre-trained models on the ImageNet-1k dataset. The use of these models enables them to focus and work on the high-level visual representations right from the start, which would be further tuned to focus on the complexities of the eight sub-types of breast cancer.

As a way of compensating for the natural imbalance between classes in the dataset and to ultimately strengthen our model's performance, we created a highly effective method of augmenting our images through the use of a complete set of augmentation strategies provided by the timm library. Examples of these strategies include random resizing, random cropping, and random horizontal mirroring. The application of these augmentation techniques in the field of digital pathology represents a novel approach because they replicate the variations that occur naturally in a tissue's orientation and staining intensity during slide processing. By exposing our transformer models to the diverse range of visual challenges that arise from these augmentation methods during training, our models were trained to learn the underlying pattern of abnormality rather than simply learning specific signs of pathology through visual memorization (Wang *et al.* 2024; Mumuni *et al.* 2024).

Performance Evaluation Metrics

In this paper, we have conducted a rigorous quantification of the diagnostic efficacy of the transformer-based backbones that were studied using a comprehensive set of performance evaluation metrics derived from the confusion matrix. Overall, classification performance was primarily conducted with Accuracy, representing the ratio of correctly identified benign and malignant samples out of all total predictions defined in Equation (1). Given the clinical need to minimize both false positives and false negatives in breast cancer screening, Precision and Recall (Sensitivity) were calculated. Precision, defined in Equation (2), reflects the model's ability not to label a negative sample as positive. On the other hand, Recall, defined in Equation (3), characterizes the ability to detect all positive examples within the dataset. Since there are severe class imbalances within our histological subtypes, we present the F1-Score as the single, balanced metric that considers the possible trade-offs between those two measures, which is the harmonic mean of Precision and Recall, and defined in Equation (4). The virtues of such multi-faceted performance metrics ensure a robust evaluation of the models' reliability in high-stakes computer-aided diagnosis.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Table 2 Comparison of Performance Metrics and Computational Complexity of the Models.

Models	Accuracy	Precision	Recall	F1-score	Params	GFLOPs
DeiT Base	0.9446	0.9466	0.9192	0.9308	85.80M	336.955
Swin Base	0.9511	0.9548	0.9332	0.9434	86.75M	303.375
ViT Base	0.9414	0.9353	0.9319	0.9323	85.80M	33.6955

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

RESULTS AND DISCUSSION

The experimental results obtained in this study underscore the remarkable potential of ViTs in navigating the complex histological landscape of breast cancer. Our comprehensive evaluation across eight distinct tissue subtypes reveals that transformer-based architectures can effectively learn the subtle, high-dimensional features required for precise computer-aided diagnosis. As summarized in Table 2, all three evaluated models, DeiT Base, Swin Base, and ViT Base, exhibited robust performance, with accuracy scores consistently exceeding 94%. This high baseline suggests that the self-attention mechanism is inherently well-suited for capturing the structural variations present in BreaKHis histopathology images.

The Swin Base model outperformed all other backbones tested in this experiment, reaching a peak accuracy of 0.9511 and an F1-score of 0.9434. The Swin architecture has a hierarchical design along with a shifted windowing method; this allows the Swin model to handle tissue slides at multiple scales. The ability to work at various scales is critical when detecting small nuclear atypia and more general architectural distortions. The second and third most successful backbones in accuracy were the DeiT Base model with an accuracy of 0.9446 and the ViT Base model with an accuracy of 0.9414. Despite being the least computationally intensive backbone (GFLOPs of 33.69), the ViT Base model still demonstrated a high recall value of 0.9319, indicating that standard transformer-based models can also have excellent detection sensitivity to malignant cells. However, the Swin Base model's high levels of precision (0.9548) and recall (0.9332) make it the most suitable choice for clinical use, where the prevention of false positives and false negatives is critical.

The diagnostic reliability of our leading machine learning (ML) model can be evaluated using the confusion matrix in Figure 2. Based on the confusion matrix, the Swin Base model demonstrated considerable accuracy for high-frequency immortalized (cancerous) classes, such as Ductal Carcinoma (DC), which correctly classified 128 out of 131 cases of Ductal Carcinoma (DC). Additionally, the Swin Base model exhibited strong predictive ability for Fibroadenomas (FIB); there was only one case in which Fibroadenomas (FIB) were incorrectly classified. The Swin Base model encountered minor confusion regarding histologically similar subtypes: four out of four Lobular Carcinomas (LC) were misclassified as Ductal Carcinomas (DC). As can be expected from a biological standpoint, Ductal Carcinomas (DC) and Lobular Carcinomas (LC) are both malignant carcinomas that may show overlap in morphology, depending on their magnification level. Additionally, the Swin Base model had slightly more difficulty with Tubular Adenomas (TA) than with Fibroadenomas (Fib), and there were occasions in which Tubular Adenomas (TA) were misclassified as Fibroadenomas (Fib). Nevertheless, the relatively high diagonal values for all classes in the Swin Base model (Figure 2) indicate that it has successfully generalized to the considerable histologic heterogeneity present in the dataset.

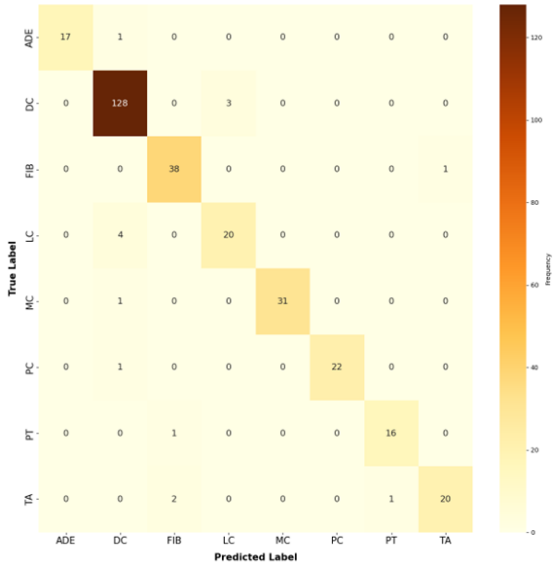


Figure 2 Confusion Matrix of the Swin Base Model.

The findings of this study indicate that transitioning away from utilizing convolutional filters to employing hierarchical transformer systems can enhance the overall experience one has with viewing digital slides. In particular, the excellent performance of the Swin Transformer shows the value of being able to extract features at multiple scales in digital pathology. Thus, capturing long-range spatial relationships between various features in conjunction with maintaining local texture analysis allows for the development of a strong base for future CAD systems that are designed to help pathologists throughout their highly complicated workflows related to diagnosing patients.

CONCLUSION

This study has systematically assessed how effective ViTs are as a new model for multi-class classification of breast cancer histopathology. Results indicate that the ability of the self-attention mechanism to manage multiple structural complexities of digital slides allows for a broader understanding of tissue architecture than the limited local receptive field of Classic CNNs. Of the models tested, the Swin Base was determined to provide the most

reliable performance, achieving an ideal balance of precision and recall, while also successfully addressing the inherent imbalances among the eight class categories present in the BreaKHis dataset. Transfer learning combined with extensive data augmentation was critical to mitigate the impact of infrequent impressions typically seen in this type of medical imaging task and ensure the models created general representations of pathology and did not rely on representations of image artifacts. Mismatches did occur within histologically similar subtypes (e.g., Lobular and Ductal Carcinoma), but the very high diagonal values achieved within the performance evaluation indicate the strong capacity of hierarchical transformers to differentiate between the eight tissue categories.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Abdulaal, A. H., M. Valizadeh, M. C. Amirani, and A. F. S. Shah, 2024a A self-learning deep neural network for classification of breast histopathological images. Biomedical Signal Processing and Control **87**: 105418.

Abdulaal, A. H., R. A. Yassin, M. Valizadeh, A. H. Abdulwahhab, A. M. Jasim, *et al.*, 2024b Cutting-edge cnn approaches for breast histopathological classification: The impact of spatial attention mechanisms. ShodhAI: Journal of Artificial Intelligence **1**: 109–130–109–130.

Akshaya, K., A. Bhan, and S. Pathan, 2024 Deep learning based classification approach to improve breast cancer screening using histopathological images. 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science, AMATHE 2024 .

Aldakhil, L. A., H. F. Alhasson, S. S. Alharbi, R. U. Khan, and A. M. Qamar, 2025 Image-based breast cancer histopathology classification and diagnosis using deep learning approaches. Applied Computational Intelligence and Soft Computing **2025**: 7011984.

Alshehri, H. M., 2025 Brenet: Attention-enhanced multi-scale cnn framework for breast cancer classification in histopathological images. IEEE Access .

Asha, N., 2025 Breast cancer data augmentation with detection using cnn model in deep learning. Lecture Notes in Electrical Engineering **1246 LNEE**: 325–339.

Balasubramanian, A. A., S. M. A. Al-Heejawi, A. Singh, A. Breggia, B. Ahmad, *et al.*, 2024 Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology. Cancers **16**.

Behzadpour, M., B. L. Ortiz, E. Azizi, and K. Wu, 2024 Breast tumor classification using efficientnet deep learning model .

Chikkala, R. B., C. Anuradha, P. S. C. Murty, S. Rajeswari, N. Rajeswaran, *et al.*, 2025 Enhancing breast cancer diagnosis with bidirectional recurrent neural networks: A novel approach for histopathological image multi-classification. IEEE Access **13**: 41682–41707.

Chitta, S., S. Sharma, and V. K. Yandrapalli, 2025 Hybrid deep learning model for enhanced breast cancer diagnosis using histopathological images. Procedia Computer Science **260**: 245–251.

Dataset, 2025 Breakhis - breast cancer histopathological dataset.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, *et al.*, 2025 An image is worth 16x16 words: Transformers for image recognition at scale .

Getu, M. A., C. Lu, Y. Liu, A. Mehmood, Z. Iqbal, *et al.*, 2025 Bcnet: A novel deep learning model for enhanced breast cancer classification using histopathological images .

Gul, M., 2025 A novel local binary patterns-based approach and proposed cnn model to diagnose breast cancer by analyzing histopathology images. IEEE Access **13**: 39610–39620.

Hayat, M., N. Ahmad, A. Nasir, and Z. A. Tariq, 2024 Hybrid deep learning efficientnetv2 and vision transformer (effnetv2-vit) model for breast cancer histopathological image classification. IEEE Access **12**: 184119–184131.

Jackson, J., L. E. Jackson, C. C. Ukwuoma, M. D. Kissi, A. Oluwasanmi, *et al.*, 2025 A patch-based deep learning framework with 5-b network for breast cancer multi-classification using histopathological images. Engineering Applications of Artificial Intelligence **148**: 110439.

Jakkaladiki, S. P. and F. Maly, 2024 Integrating hybrid transfer learning with attention-enhanced deep learning models to improve breast cancer diagnosis. PeerJ Computer Science **10**: e1850.

Kansal, K., S. Kumar, and K. Kansal, 2025 Advances in deep learning techniques for breast cancer classification: A comprehensive review. Archives of Computational Methods in Engineering pp. 1–36.

Karthiga, R., K. Narasimhan, N.Raju, and R. Amirtharajan, 2025 Automatic approach for breast cancer detection based on deep belief network using histopathology images. Multimedia Tools and Applications **84**: 4733–4750.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, *et al.*, 2025 Swin transformer: Hierarchical vision transformer using shifted windows .

Logu, K. and S. J. J. Thangaraj, 2024 Progressing breast cancer assessment: Precise tumor categorization through densenet201-based deep learning. TQCEBT 2024 - 2nd IEEE International Conference on Trends in Quantum Computing and Emerging Business Technologies 2024 .

Mumuni, A., F. Mumuni, and N. K. Gerrar, 2024 A survey of synthetic data augmentation methods in machine vision. Machine Intelligence Research 2024 21:5 **21**: 831–869.

Murphy, G. and R. Singh, 2024 Comparative analysis and ensemble enhancement of leading cnn architectures for breast cancer classification .

Ogundokun, R. O., A. R. T. Abdullahi, A. R. Adenike, C. Awoniyi, H. B. Akande, *et al.*, 2024 Hybrid deep learning for breast cancer diagnosis: Evaluating cnn and ann on breakhis_v1_400x. In *International Conference on Science, Engineering and Business for Driving Sustainable Development Goals, SEB4SDG 2024*, Institute of Electrical and Electronics Engineers Inc.

Rajaram, G., V. Rajinikanth, and B. S. Latha, 2024 Classification of breast histology into benign and malignant using resnet-variants: A study with breakhis database. In *2024 9th International Conference on Applying New Technology in Green Buildings, ATiGB 2024*, pp. 479–483, Institute of Electrical and Electronics Engineers Inc.

Ramamoorthy, P., B. R. R. Reddy, S. S. Askar, and M. Abouhawwash, 2024 Histopathology-based breast cancer prediction using deep learning methods for healthcare applications. Frontiers in Oncology **14**: 1300997.

Simonyan, E. O., J. A. Badejo, and J. S. Weijin, 2024 Histopathological breast cancer classification using cnn. Materials Today: Proceedings **105**: 268–275.

Singh, A. and K. S. Kaswan, 2024 Empirical analysis on breast cancer datasets with machine learning. Proceedings - International Conference on Computing, Power, and Communication Technologies, IC2PCT 2024 pp. 223–227.

Thakur, N., S. Shrivastava, S. Shukla, and M. Gyanchandani, 2025 Patient-centric multilabel classification with svm on breakhis dataset. In *2025 3rd International Conference on Data Science and Network Security (ICDSNS)*, pp. 1–7, IEEE.

Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, *et al.*, 2025 Training data-efficient image transformers & distillation through attention .

Ukwuoma, C. C., D. Cai, E. O. Eziefuna, A. Oluwasanmi, S. F. Abdi, *et al.*, 2025 Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable ai – lime & shap. Biomedical Signal Processing and Control **100**: 107014.

Wang, Z., P. Wang, K. Liu, P. Wang, Y. Fu, *et al.*, 2024 A comprehensive survey on data augmentation .

How to cite this article: Koyuncu, S., Cakmak, Y. and Pacal, I. Towards Robust CAD Systems for Digital Pathology: Evaluating Transformer-Based Backbones for Breast Cancer Classification. *Computers and Electronics in Medicine*, 3(1), 54-59, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.60-76
https://doi.org/10.69882/adba.cem.2026018

Enhancing Hospital Inventory Forecasting Accuracy through Hybrid and Ensemble Learning Models

Cem Özkurt[*,1], **Ahmet Kutey Küçükler**[α,2], **Murat Karslıoğlu**[β,3] **and Ruveyda Nur Özdemir**[§,4]

[*]Department of Computer Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye, [α]Department of Computer Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye, [β]Department of Computer Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye, [§]Department of Computer Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye.

ABSTRACT Demand forecasting for medical and consumable supplies in healthcare institutions is a challenging problem due to irregular usage patterns, seasonality, sudden demand spikes, and data sparsity, and inaccurate forecasts may lead to stock-outs, excessive inventory costs, and disruptions in patient care. This study proposes an anomaly-aware, hybrid and ensemble-based forecasting and decision support framework for short-term hospital inventory demand prediction using real-world operational data obtained from a hospital inventory management system. The proposed approach integrates density-based anomaly detection, material-level behavioral feature extraction, supervised time series transformation, and a multi-model ensemble architecture combining linear models, tree-based methods, and boosting-based learners, with model selection and weighting performed via time-series cross-validation. To ensure operational robustness, a multi-layer fallback strategy incorporating classical exponential smoothing and conservative heuristics is employed for data-scarce scenarios, and an interpretable rule-based forecast confidence score together with an integrated ABC–XYZ segmentation scheme is used to directly link forecasts with inventory control policies. Experimental results on real hospital inventory data demonstrate that the proposed framework significantly improves forecasting stability and accuracy compared to single-model approaches, particularly for heterogeneous and irregular consumption patterns, while providing a practical, explainable, and operationally actionable solution for hospital inventory management.

KEYWORDS
Hospital inventory management
Demand forecasting
Ensemble learning
Hybrid models
Time series forecasting
Anomaly detection
Decision support systems

INTRODUCTION

The effective and efficient operation of healthcare delivery systems plays a critical role in today's complex and dynamic healthcare environment. In hospitals, having the right amount of medical and consumable supplies in the right place at the right time is essential for the uninterrupted delivery of patient care services (Joshi *et al.* 2025). Inventory management is an important component not only in terms of operational efficiency, but also in terms of financial performance, the quality of hospital services, and human health. Incorrectly estimating the demand for medical and consumable supplies can have very serious consequences for hospitals. Excessive stockpiling leads to increased storage costs, deterioration of material quality, and significant financial losses, while insufficient stock levels cause disruptions in patient care, reduced treatment quality, and, in some cases, pose a threat to patient safety. Therefore, demand forecast accuracy is a fundamental requirement for optimizing hospital resources and delivering sustainable healthcare services (Umoren *et al.* 2025).

In previous years, traditional statistical methods and simple heuristic rules were used in hospital inventory management. These methods often produced inaccurate forecasts because they did not adequately account for complex factors such as seasonal changes, emergencies, disease outbreaks, and sudden fluctuations in demand (Adedunjoye and Enyejo 2024). Over the past twenty years, the rapid development of machine learning and time series analysis methods has opened up new opportunities in medical demand

Manuscript received: 4 October 2025,
Revised: 11 December 2025,
Accepted: 20 December 2025.

[1]cemozkurt@subu.edu.tr (**Corresponding author**)
[2]23010903094@subu.edu.tr
[3]24010903047@subu.edu.tr
[4]24010903035@subu.edu.tr

forecasting. Algorithms such as tree-based methods (Random Forest, Gradient Boosting, XGBoost, LightGBM), linear regression approaches (Ridge, Lasso), and time series models (ARIMA, Exponential Smoothing) yield good results when used individually (Darshan *et al.* 2025). Academic research and practical applications have shown that a single model does not always deliver optimal performance and may be insufficient in adapting to different data patterns and changing conditions (Punnahitanond *et al.* 2025).

In recent years, hybrid and ensemble model combinations have emerged as innovative solutions to the demand forecasting problem, both in academic settings and industrial applications (Mbonyinshuti *et al.* 2024). Hybrid models combine machine learning and time series methods, leveraging the strengths of each technique, while ensemble approaches systematically combine the forecasts of different models to obtain more stable and reliable results (Vignesh and Vijayalakshmi 2025). Such combination methods capture non-linear relationships that cannot be addressed by individual models, reveal hidden patterns in the data, and significantly reduce prediction errors (Jahin *et al.* 2024). Particularly in the healthcare sector, the use of hybrid and ensemble methods for demand forecasting of hospital supplies has been limited, representing an important area that still needs to be researched (Donkor *et al.* 2024).

The research problem of this study is whether the demand for medical and consumable supplies in hospitals can be predicted more accurately and reliably using hybrid and ensemble model combinations under complex and variable conditions where single prediction models fall short. The main objective of this study is to demonstrate how hybrid and ensemble model combinations can be effectively applied in the demand forecasting of medical and consumable supplies used in hospitals, instead of single models. In this study, historical inventory inflow and outflow data, time series structure, and material usage patterns will be analyzed in detail, and various combinations of tree-based machine learning models, linear regression methods, and time series techniques will be tested. This integrated approach is designed to overcome the limitations of individual models and provide a more reliable, sustainable, and successful forecasting framework for hospital inventory management.

Research Objectives

The main objectives of this research are defined as follows:

- To individually evaluate the success of different machine learning (Random Forest, Gradient Boosting, XGBoost, LightGBM, Ridge, Lasso) and time series models (ARIMA, Exponential Smoothing) in forecasting hospital medical supply demand and to perform a comparative analysis based on performance metrics (MAE, RMSE, MAPE, R^2).
- To design hybrid and ensemble model combinations that combine different machine learning and time series methods in order to overcome the limitations of single model approaches, and to evaluate the prediction success of these combinations.
- Systematically analyze whether the proposed hybrid and ensemble model combinations improve demand forecast accuracy compared to single models, and how they affect model stability and forecast variance.
- To propose a more reliable, economical, and sustainable forecasting framework for inventory management applications in hospitals; to provide recommendations on how this framework can be integrated into practical applications.

Primary Contributions

The main contributions of this study can be summarized as follows:

- Tree-based machine learning models, linear regression methods, and classical time series techniques were comprehensively compared in hospital supply demand forecasting. The performance of each technique was evaluated using standardized metrics.
- Innovative hybrid model combinations integrating machine learning and time series methods have been developed. These hybrid structures overcome the limitations of individual models, providing higher prediction accuracy.
- Various ensemble techniques (bagging, boosting, stacking, weighted averaging) have been systematically applied to optimally combine the predictions of different models. These strategies reduce prediction error and increase system stability.
- Factors specific to the hospital environment, such as seasonal changes, emergencies, and sudden fluctuations in demand, were taken into account in model design. Thus, solutions adapted to the healthcare sector were presented, unlike general industry applications.
- Detailed recommendations regarding model selection criteria, data preprocessing procedures, hyperparameter tuning, and system integration enhance the practical applicability of the study.
- The proposed hybrid and ensemble combinations significantly improve prediction accuracy compared to individual models, thereby contributing to reduced inventory costs and optimized hospital operations.

Structure of the Article

The structure followed in the article begins with the Related Work section, which addresses existing academic studies on demand forecasting techniques, their applications in hospital settings, and the effectiveness of machine learning and ensemble methods. Subsequently, the Materials and Methods section detail the dataset used, data preprocessing steps, and the applied methodology, introducing tree-based models, linear regression methods, time series techniques, and ensemble strategies. In the Hybrid Model Design and Ensemble Strategies sections, combinations of different techniques are systematically created and integrated, and the design principles of each hybrid architecture are compared with the advantages and disadvantages of various ensemble combination methods. In the Results and Discussion section, the performance of all models is compared using standardized metrics (MAE, RMSE, MAPE, etc.), performance improvements of hybrid and ensemble models compared to individual models are presented, detailed comparisons between different model categories are made, and the implications of the findings for hospital operations are outlined by explaining their meaning in the context of the healthcare sector. Finally, the Conclusion section summarizes the main findings of the research, highlights the contributions of the study, provides practical recommendations for hospital managers and operational staff, and outlines guidelines for future research.

RELATED WORK

The healthcare sector faces unique challenges in inventory management and demand forecasting due to its complex structure and critical outcomes. The availability of medical supplies and consumables has a direct impact on the quality and safety of patient care.

Insufficient stock levels can lead to operational disruptions, service interruptions, and even situations that threaten patient safety. Excessive inventory, on the other hand, causes financial problems such as expired shelf life, storage costs, and insufficient capital. In this context, accurately forecasting the demand for medical and consumable supplies in healthcare institutions is of critical strategic importance for cost optimization, operational efficiency, and, most importantly, the provision of uninterrupted, high-quality patient care. In this context, recent literature has been reviewed thematically.

Traditional Time Series Approaches and Their Limitations

Among the classical time series methods commonly used in the literature for forecasting demand for medical and consumable supplies in the healthcare sector are Moving Average (MA), Exponential Smoothing (ES), and Autoregressive Integrated Moving Average (ARIMA) models (Shih and Rajendran 2019; Khalid 2024). While these methods demonstrate a certain degree of success in forecasting future demand based on historical data (Luo *et al.* 2017), they fail to adequately model external factors arising from the dynamic nature of the healthcare sector (Khalid 2024). Particularly in demand series with non-linear relationships and high volatility, the forecasting performance of these models may decline and they may be insufficient in capturing complex patterns (Kolambe 2024; Dachepalli 2025).

In a study predicting hospital admissions during the COVID-19 pandemic, ARIMA and Exponential Smoothing models were used, but limitations in their performance were observed due to the dynamic and non-linear nature of the pandemic (Perone 2021). Similarly, despite the use of models such as ARIMA in forecasting demand in the blood product supply chain, difficulties are encountered in managing the uncertainties arising from the short shelf life and highly variable usage rates of these products (Motamedi *et al.* 2024). Although traditional time series methods have advantages such as usability and low computational cost, especially in short-term forecasting, their inability to fully grasp the complex structure of demand in healthcare services is a significant limitation (Luo *et al.* 2017; Fatima and Rahimi 2024).

These studies demonstrate that classical time series methods have certain advantages in demand forecasting in the healthcare sector; however, they also reveal that nonlinear structures, sudden demand spikes, and material-based heterogeneous usage patterns cannot be adequately captured. In this study, in order to overcome these limitations, a hybrid and ensemble approach integrating machine learning-based models was adopted instead of using classical time series models alone. The goal was to produce more flexible and reliable forecasts for complex demand structures where traditional methods fall short.

The Rise of Machine Learning-Based Approaches

In recent years, machine learning (ML) algorithms have emerged as a powerful alternative for time series forecasting problems and, in particular, demand forecasting in the healthcare sector (Kontopoulou *et al.* 2023). Ensemble learning models such as Random Forest (RF), Gradient Boosting (GBM), XGBoost, and LightGBM stand out for their ability to model complex and non-linear relationships (Qiu *et al.* 2019; Erdebilli and Devrim-Tenba 2022). These models can provide higher prediction accuracy compared to traditional methods and can include a large number of explanatory variables, such as hospital capacity, patient numbers, and seasonal indices, in the analysis (Shern *et al.* 2024).

In predicting peak demand days for cardiovascular diseases, the LightGBM model stood out with a higher AUC value (0.940) compared to other ML models such as logistic regression and SVM (Qiu *et al.* 2019). Similarly, in a study on medical waste prediction, an ensemble consensus regression model using algorithms such as Random Forest, Gradient Boosting, and AdaBoost demonstrated superior performance with a lower RMSE value compared to single models (Erdebilli and Devrim-Tenba 2022). Such tree-based models generally demonstrate strong capabilities in analyzing and predicting complex data structures (Gldoan *et al.* 2023).

Studies using linear models such as Ridge and Lasso also offer limited but valuable contributions, particularly in the context of feature selection and preventing overfitting (Abdul-Rahman *et al.* 2021). The literature emphasizes that machine learning methods are more flexible and adaptable than traditional statistical methods, especially when dealing with complex and high-dimensional data (Fatima and Rahimi 2024). These developments support more accurate decision-making and increased operational efficiency in demand management in the healthcare sector.

Hybrid and Ensemble Models: Potential and Gaps in the Literature

In order to overcome the limitations of single models, hybrid and ensemble modeling approaches have recently gained attention. Combining the strengths of different model families, these approaches generally exhibit superior forecasting performance compared to models used alone (Perone 2021). Ensemble models offer the potential to reduce forecast variance, correct bias, and create more generalizable models by bringing together multiple base learners (Fatima and Rahimi 2024). Such model combinations are particularly promising for demand series in the healthcare sector, which often involve heterogeneous data structures. For example, in a study on energy demand forecasting, hybrid combinations of different time series and machine learning models significantly outperformed individual models.

While much of the existing literature focuses on optimizing the performance of a single model, studies that systematically combine multiple models and comprehensively evaluate their forecasting success are limited (Kontopoulou *et al.* 2023). Particularly in the healthcare sector, there is a need for in-depth analysis of the effectiveness of combinations using different model families in forecasting medical and consumable demand. This points to a significant gap in the literature, as the complex and highly variable demand structures in the healthcare sector suggest that a single model may not always provide the best solution (Qiu *et al.* 2019).

This study aims to fill this gap in the literature by focusing on medical and consumable supply demand forecasting in the healthcare sector. Rather than demonstrating the superiority of a single model, it will compare the forecasting performance of different combinations of classical time series, machine learning, and hybrid approaches. This approach has the potential to contribute to the development of stronger and more consistent forecasting models in healthcare supply chain management, thereby increasing operational efficiency and minimizing costs (Aifuwa *et al.* 2020). Thus, it is assumed that by bringing together complementary information obtained from different model families, more robust and reliable forecasting systems that better adapt to the unique dynamics of the healthcare sector can be created.

Although these machine learning-based studies have made significant progress in capturing non-linear relationships in demand forecasting in the healthcare sector, they mostly focus on individual model performance and address the potential for using different model families together to a limited extent. Unlike the existing

literature, this study aims to provide an integrated forecasting framework by systematically combining tree-based methods with linear models, complementing the strengths and weaknesses of individual models.

METHODOLOGY

This section details the methodological framework of the forecasting and decision support system used in hospital inventory management. The primary objective of the study is to generate short-term (three-month) demand forecasts for the future by utilizing historical consumption records on a material-warehouse basis and to convert these forecasts into outputs that can be directly integrated into operational decision-making processes. In this context, all steps, from raw data processing to the creation of monthly time series, from multi-model forecasting approaches to reliability assessment, and from ABC-XYZ segmentation to optimal stock level calculations, are addressed under a comprehensive methodology. The method followed aims to reduce model fragility in healthcare inventories with different consumption regimes, ensure the physical and operational suitability of forecasts, and make the results suitable for internal auditing and reporting (Silva-Aravena *et al.* 2020; Karamshetty *et al.* 2022; Subramanian 2021).

Problem Definition and Notation

The fundamental problem addressed in this study is the accurate and reliable prediction of future short-term consumption behavior for each material based on hospital warehouses. Specifically, the objective is to estimate consumption (output) quantities for the next three months for each material-warehouse combination and to use these estimates to generate decision support outputs for inventory management. The generated forecasts are not merely numerical predictions; they also form the basis for advanced analyses such as ABC-XYZ segmentation, optimal stock level calculation, and automatic identification of problematic materials. This approach directly links the forecasting process to operational policy generation (Balkhi *et al.* 2022; Feibert *et al.* 2019; Polater and Demirdogen 2018).

The system generates monthly consumption time series from raw material movement records and analyzes these time series using an ensemble forecaster. Only physically meaningful movement types are considered in the forecasting process; in this context, transactions in the raw data set are filtered only as G (input) and C (output). Thus, the model operates on a data structure that represents the actual effect of stock movements on inventory levels.

The mathematical notation used throughout this section is defined as follows. Here, m represents the material ID and corresponds to the `HASTANE_MALZEME_ID` field in the application. Similarly, d represents the warehouse ID and is represented by the `AD` field. The time dimension t is defined in terms of the timestamp corresponding to the beginning of the month. The expression $y_{m,d,t}$ used in Equation (1) indicates the output, i.e., the consumption quantity of material m in warehouse d in month t. The forecast horizon is expressed by the parameter h, which in this study covers three-month forward forecasts with $h = 1, 2, 3$.

Data Source and Transformation

The raw dataset used in this study consists of detailed material movement records obtained from a real-world hospital inventory management system through an institutional collaboration. The dataset was provided under a confidentiality agreement. Due to data security and institutional privacy policies, the identity of the data-providing institution and the raw dataset itself cannot be publicly disclosed. The dataset includes fields such as material ID, receipt date, transaction type (inbound or outbound), transaction quantity, and warehouse information.

In order to obtain consistent and reproducible results in forecasting and decision support processes, the raw data was first subjected to a systematic transformation process. In this process, fields unnecessary for analysis were eliminated, numerical variables were converted to appropriate data types, and date fields were converted to timestamp format to ensure temporal consistency (Merkuryeva *et al.* 2019; Subramanian 2021).

To create monthly consumption time series, the exit transactions for each material-warehouse combination were aggregated on a monthly basis. As shown in Equation (1), the consumption quantity for a given month t for a specific material m and warehouse d is defined as the sum of all exit transactions occurring within that month (Mbonyinshuti *et al.* 2022):

$$y_{m,d,t} = \sum_{i \in I(m,d,t)} \text{MIKTAR}_i \qquad (\text{only ISLEM\_TUR} = C) \tag{1}$$

The time series created using the monthly consumption values obtained in Equation (1) were converted to a fixed monthly frequency to make them suitable for the analysis and forecasting process. During this conversion process, the time axis was reindexed in monthly periods for each material-warehouse combination, and it was checked whether there were any outflow transactions in specific months.

This process was carried out as shown in Equation (2). Accordingly, if a consumption observation exists for the relevant month, the time series value is retained as is; if there is no discharge operation in the relevant month, the consumption quantity is assigned a value of zero:

$$y_t \leftarrow \text{asfreq(MS)}, \qquad y_t = \begin{cases} y_t, & \text{if there is consumption} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In Equation (2), the expression `asfreq(MS)` represents the resampling of the time series to a monthly start frequency. As a result of this process, the time series is converted into a regular monthly structure, and missing months are explicitly identified. Assigning a value of zero when there is no consumption in the relevant month prevents "no usage" situations from being implicitly lost in the data and allows the model to learn irregular or infrequent usage patterns.

Thanks to this approach, continuous, regular, and comparable monthly time series are obtained for each material-warehouse combination; at the same time, zero consumption periods are preserved as a meaningful signal in statistical and behavioral analyses. Thus, the time series created provide a suitable input structure for both classical time series methods and feature-based machine learning models.

Data Cleaning and Outlier Management

This subsection details how data quality issues that could affect the reliability of forecasting and inventory decisions are addressed. Health inventory data, being derived from operational processes, may contain missing records, incorrect date information, and consumption values that are statistically extreme. Since such problems can lead to misleading patterns and unstable model behavior, especially in time series-based forecasting models, a systematic data

cleaning and outlier management process was applied prior to analysis (Karamshetty *et al.* 2022; Gurumurthy *et al.* 2021).

During the preprocessing stage, records with missing information in fields required for analysis were first removed from the dataset. These fields include basic variables such as material ID, warehouse information, transaction date, and transaction amount. Subsequently, quantity fields were converted to appropriate numerical data types, and values with logical inconsistencies were checked. Finally, date fields were parsed into timestamp format, and records with invalid or incorrect date information were eliminated from the dataset, thus ensuring the chronological integrity of the time series.

During the outlier management phase, a top-end trimming (minorize-like) approach was adopted to prevent records with extremely high values, which are rarely observed in consumption amounts, from disproportionately affecting the modeling process. In this approach, the distribution of raw consumption quantities was considered, and the 99.9th percentile value was set as the threshold. As shown in Equation (3), this threshold value is represented by τ, and only observations below this value are retained in the analysis:

$$\tau = Q_{0.999}(x), \qquad x_i^* = \begin{cases} x_i, & x_i \leq \tau \\ \text{remove}, & x_i > \tau \end{cases} \tag{3}$$

In Equation (3), x_i represents the raw consumption (transaction amount) value, while $Q_{0.999}(x)$ represents the upper 99.9th percentile of the entire consumption distribution. According to this definition, observations satisfying the condition $x_i \leq \tau$ are retained in the data set, while extreme outliers above the threshold value are excluded from the analysis. This method aims to prevent the model from developing excessive sensitivity without completely suppressing the effect of outliers, while preserving the overall structure of the distribution.

This outlier cleaning strategy offers a balanced solution that maintains both statistical robustness and operational realism, particularly in healthcare inventory data where high-volume outputs are rarely used but occasionally observed. Thus, the resulting time series achieve a more stable and reliable input structure for both classical statistical analyses and machine learning-based forecasting models (Ingle *et al.* 2021; Kim *et al.* 2023).

Material-Based Statistical and Behavioral Feature Extraction

This subsection explains how statistical and behavioral features that quantitatively summarize the past consumption behavior of each material are extracted. The aim is to improve prediction performance and generalizability by including attributes that reflect the general usage character of the material in the model, rather than using only the past values of the time series. This approach enables the differentiation of materials that have the same lagged values but exhibit different usage patterns (Kim *et al.* 2023; Subramanian 2021).

For each material–storage combination, basic descriptive statistics such as total number of entries and exits, average consumption amount, maximum and minimum values, and standard deviation were calculated. However, the coefficient of variation (CV) was used to evaluate consumption behavior independently of scale. As shown in Equation (4), CV is defined as the ratio of the standard deviation to the mean and measures the relative variability in consumption:

$$CV = \frac{\sigma(y)}{\mu(y)} \qquad (\mu(y) > 0) \tag{4}$$

In Equation (4), σ represents the standard deviation of monthly output quantities, while μ represents the mean value of the same series. The CV metric prevents absolute variance from being misleading in materials with low averages, enabling a comparative assessment of consumption stability.

Seasonality strength is defined to quantitatively express the degree of seasonal fluctuation in consumption throughout the year. In this context, the seasonality indicator is obtained by normalizing the variability of average consumption values by month. As shown in Equation (5), this metric is calculated based on the ratio of the variation between monthly averages to the overall average:

$$S = \frac{\sigma(\bar{y}_{ay})}{\mu(\bar{y}_{ay})} \qquad (\mu > 0) \tag{5}$$

Here, $\bar{y}_{ay}$ represents the average consumption calculated for each month, while $\bar{y}$ represents the overall average obtained throughout the entire period. Higher S values indicate strong seasonal fluctuations and show that the forecasting process is relatively more complex.

The trend slope was obtained by applying linear regression to the monthly total consumption values in order to determine the general upward or downward direction of the consumption series over time. As shown in Equation (6), the slope of the consumption values against the time variable is used as a trend indicator:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \tag{6}$$

In this equation, the coefficient β_1 represents the trend slope, indicating whether consumption shows an increasing or decreasing trend over time.

Finally, the regularity score is defined to measure the degree to which a material is used regularly. As shown in Equation (7), regularity is calculated as the ratio of the months in which consumption occurred to the total number of months:

$$R = \frac{\text{Number of months consumption was observed}}{\text{Total number of months}} \tag{7}$$

This metric plays a critical role in both prediction reliability and inventory policy design by enabling the differentiation of irregular and infrequently used materials.

Time Series Feature Engineering

This subsection details the feature engineering process applied to transform monthly consumption time series into an input space suitable for supervised learning models. In time series forecasting problems, the direct use of past observations is often insufficient; instead, derived features representing temporal dependencies, short-term trends, and calendar effects must be included in the model. Within the scope of this study, a comprehensive feature vector consisting of lagged values, rolling statistics, short window trends, and calendar variables was created for each month (Kim *et al.* 2023; Ingle *et al.* 2021).

The feature vector created for each time step primarily includes calendar components. In this context, the variables "month(t)", "year(t)", and "quarter(t)" are used to enable the model to learn monthly, yearly, and quarterly seasonal effects. Additionally, the variable k, an increasing index starting from the beginning of the time series, is defined to represent long-term trends through a linear time indicator.

To capture temporal dependencies, lagged values of past consumption for each month are added to the feature vector. In this study, consumption amounts from one-, two-, and three-months

prior were defined as y_{t-1}, y_{t-2}, and y_{t-3}, respectively, and the short-term autocorrelation structure was targeted for transfer to the model.

Rolling statistics were calculated to represent the short-term consumption level and volatility. As shown in Equation (8), the three-month rolling average $RM_3(t)$ is defined as the arithmetic mean of the previous three observations, including the relevant month:

$$RM_3(t) = \frac{1}{3}\sum_{i=0}^{2} y_{t-i} \tag{8}$$

In Equation (8), y_{t-i} represents the consumption amount i months prior to time t. This metric represents the short-term consumption level in a smoothed form, reducing the impact of sudden spikes.

Additionally, a three-month rolling standard deviation was calculated to measure short-term consumption volatility. As shown in Equation (9), the $RS_3(t)$ value is defined as the square root of the average of the squares of deviations around the rolling average:

$$RS_3(t) = \sqrt{\frac{1}{3}\sum_{i=0}^{2} (y_{t-i} - RM_3(t))^2} \tag{9}$$

Equation (9) quantitatively expresses the magnitude of short-term consumption fluctuations and enables the model to distinguish between stable and volatile periods.

The short window trend is defined to capture the direction of increase or decrease in the recent past. In this context, the slope value obtained by applying a linear regression on the last three observations is expressed as $\text{Trend}_3(t)$ as shown in Equation (10):

$$\text{Trend}_3(t) = \text{slope}(y_t, y_{t-1}, y_{t-2}) \tag{10}$$

This slope value represents the recent consumption direction independently of long-term trends and ensures that short-term changes are incorporated into the forecasting process.

All features defined in this subsection are explicitly generated within the system and are perfectly aligned with the feature space used during the training phase. The same feature generation scheme is maintained for forecasts covering the next three months; feature vectors for future periods are created using lagged values derived from the latest observations, rolling statistics, and trend components. This ensures structural consistency between the training and forecasting phases, enhancing the model's stability and generalizability.

Multi-Model Forecasting and Ensemble Combination

This subsection details how monthly consumption time series are handled within a multi-model forecasting framework and how models with different learning biases are combined to obtain more stable forecasts. Health inventory consumption data can exhibit heterogeneous behaviors such as high variation, irregular usage, sudden spikes, and regime shifts, making a forecasting approach based on a single model family often insufficient. Therefore, the study adopts an ensemble structure that combines linear, tree-based, and boosting-based models (Chien *et al.* 2023; Sina *et al.* 2023).

The system uses XGBoost and LightGBM models when appropriate libraries are available in the working environment; if these libraries are not available, it constructs an equivalent ensemble space using Random Forest and Gradient Boosting algorithms representing the same model class. Additionally, linear-regularized models such as Ridge and Lasso are included in the ensemble to capture linear components in consumption behavior. This approach aims to achieve more balanced prediction performance across different consumption regimes through model diversity.

Converting Time Series to Regression (Supervised TS Formulation): This subsection explains the process of converting monthly consumption time series into a structure that can be processed by supervised learning algorithms. In time series forecasting problems, it is known that each observation is not solely based on past values; it must also be considered alongside calendar effects, short-term statistics, and local trend information. Therefore, in this study, the time series problem is transferred to a regression framework by defining a comprehensive feature vector for each time step.

For the monthly consumption series y_t, the feature vector created at each time step t is defined as shown in Equation (11):

$$x_t = [\underbrace{\text{month}(t), \text{year}(t), \text{quarter}(t), k}_{\text{calendar + index}}, \underbrace{y_{t-1}, y_{t-2}, y_{t-3}}_{\text{delays}}, \underbrace{\mu_t(3), \sigma_t(3)}_{\text{rolling}}, \underbrace{s_t(3)}_{\text{short trend}}] \tag{11}$$

The variables $\text{month}(t)$, $\text{year}(t)$, and $\text{quarter}(t)$ in Equation (11) ensure that the monthly, annual, and quarterly seasonal effects on consumption are incorporated into the model. The k variable represents the increasing time index from the beginning of the time series and allows long-term trends to be learned through a linear time indicator. The lagged variables y_{t-1}, y_{t-2}, y_{t-3} reflect the short-term autocorrelation structure, ensuring that past consumption information is directly included in the model.

Short-term statistical properties are calculated based on the last three observations. The three-month moving average $\mu_t(3)$ and moving standard deviation $\sigma_t(3)$ are defined as shown in Equation (12):

$$\mu_t(3) = \frac{1}{3}\sum_{i=0}^{2} y_{t-i}, \qquad \sigma_t(3) = \sqrt{\frac{1}{3}\sum_{i=0}^{2} (y_{t-i} - \mu_t(3))^2} \tag{12}$$

These metrics represent the short-term level and volatility of consumption, ensuring that sudden spikes and temporary fluctuations are incorporated into the model in a balanced manner.

The short-window trend is defined by the slope value obtained by applying a linear fit to the last three observations. This process is performed as shown in Equation (13):

$$y_{t-i} \approx ai + b \quad (i = 0, 1, 2), \qquad s_t(3) = a \tag{13}$$

Equation (13), where a represents the trend coefficient reflecting the short-term consumption trend. This value ensures that the recent upward or downward direction is incorporated into the model independently of long-term trends.

In practice, this structure was implemented through the variables `lag_1-3`, `rolling_mean_3`, `rolling_std_3`, and `trend_3`. In addition, attributes such as the coefficient of variation (CV) representing material behavior, seasonality score, long-term trend slope, and regularity were also added numerically to the feature vector. In this way, the consumption context of different materials with the same lag values is clearly presented to the model, enabling the model to learn material-specific behaviors (Kim *et al.* 2023).

Mathematical Foundations of the Models Used

This subsection presents the mathematical foundations of the prediction models used in the study. The models were selected to have different learning biases and positioned within the ensemble structure to complement each other.

In linear regression models, the objective is to estimate the coefficients representing the linear relationship between the input features and the target variable. This fundamental relationship can be expressed as shown in Equation (14):

$$\hat{y} = X^T \beta \tag{14}$$

In this model, β represents the coefficient vector to be estimated. In high-dimensional and correlated feature spaces, the classical least squares approach can lead to overfitting. Ridge and Lasso regularizations have been used to mitigate this problem.

Ridge regression adds a penalty term based on the L_2 norm to limit the magnitude of the coefficients. As shown in Equation (15):

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2 \tag{15}$$

This approach reduces the model's variance by bringing the coefficients closer to zero, thereby producing more stable estimates.

Lasso regression, on the other hand, produces sparse solutions using a penalty term based on the L_1 norm. This indirectly provides a feature selection mechanism by setting the coefficients of unimportant features to zero. Equation (16) illustrates this structure:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1 \tag{16}$$

Ridge and Lasso models form a strong baseline, particularly in feature spaces containing partially linear relationships such as calendar variables, lagged consumption values, and rolling statistics. Using these models in conjunction with tree-based methods allows the balanced capture of linear and nonlinear components in the ensemble structure (Kim *et al.* 2023).

Random Forest, one of the tree-based methods used in this study, has been included in the ensemble structure to obtain stable estimates, especially for high-variance and noisy consumption time series. Since health inventory consumption data can contain sudden usage spikes and irregular usage periods, models based on a single decision tree often face high variance problems. The Random Forest approach aims to mitigate this limitation through bootstrap sampling and averaging mechanisms (Mbonyinshuti *et al.* 2022; Kim *et al.* 2023).

The Random Forest model produces the final output by averaging the predictions of numerous decision trees trained on different subsets of data created using the bootstrap method. This structure is expressed as follows, as shown in Equation (17):

$$\hat{y}_{RF}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{17}$$

Equation (17) shows that $f_t(x)$ represents the prediction generated by the t-th decision tree for the input feature vector x, while T denotes the total number of trees in the ensemble. Each decision tree is trained on a different subset of the original dataset, sampled using the bootstrap method; additionally, randomly selected feature subsets are used in the splitting operations at tree nodes. These two randomness mechanisms significantly reduce the variance of the ensemble average by lowering the correlation between trees.

While a single decision tree typically exhibits low bias but high variance, the Random Forest ensemble suppresses this high variance through averaging. This feature enables more stable and reliable predictions, especially in health inventory time series where consumption amounts fluctuate irregularly, very high values are rarely observed, and the noise level is high. Therefore, the Random Forest model plays a critical role as a variance-reducing component within the ensemble structure.

Gradient Boosting, another tree-based method used in this study, was included in the ensemble structure to gradually correct systematic errors that arise in consumption estimation. While the Random Forest approach focuses on reducing variance, the Gradient Boosting model primarily aims to reduce bias and gradually learns the error components that previous models could not explain. This feature provides a significant advantage, especially in complex consumption patterns where trends, seasonality, and lagged interactions are observed together (Sina *et al.* 2023; Kim *et al.* 2023).

The Gradient Boosting approach builds the model incrementally. In the first step, an initial model that minimizes the loss between observed values and predictions is defined. This process is expressed as shown in Equation (18):

$$L(y_i, c) \tag{18}$$

In Equation (18), $L(y_i, c)$ represents the loss between the actual consumption value y_i and the constant estimate c. Following the initial model, the model is updated incrementally. At each step, a new weak learner is added that attempts to explain the residuals (negative gradients) of the previous model. This update process is shown in Equation (19):

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \tag{19}$$

Here, $h_m(x)$ represents the new weak learner attempting to explain the error component that the previous model could not predict, while ν represents the learning rate. The learning rate reduces the risk of overfitting by controlling the extent to which each new model contributes to the ensemble output. Thanks to this stepwise structure, the Gradient Boosting model gradually captures components that could not be explained in previous steps and systematically improves prediction accuracy.

Due to these characteristics, Gradient Boosting plays an important role as a bias-reducing component within the ensemble structure.

XGBoost and LightGBM are enhanced versions of the Gradient Boosting approach in terms of regularization and computational efficiency. These models are designed to offer stronger generalization performance, particularly in high-dimensional feature spaces and complex interaction structures. In this study, these models were included in the ensemble structure to more effectively capture nonlinear and interactive consumption dynamics (Cao and Gui 2019; Kim *et al.* 2023).

The optimization process in XGBoost and LightGBM is based on jointly minimizing the loss function, which measures data fit, and a regularization term that penalizes model complexity. This objective function is generally defined as shown in Equation (20):

$$L = \sum_{i} L(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{20}$$

In Equation (20), $L(y_i, \hat{y}_i)$ represents the loss between the actual consumption value and the model estimate, while the term $\Omega(f_k)$ represents the model complexity of the k-th tree. A typical regularization function used for tree complexity is given in Equation (21):

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \tag{21}$$

In this expression, T represents the number of leaves in the tree, w represents the weights associated with leaf nodes, and γ and λ represent the regularization coefficients that penalize model complexity. These penalties imposed on the number of leaves and weights prevent the model from learning overly complex structures, thereby improving its generalization ability.

Thanks to this regularization mechanism, XGBoost and LightGBM can effectively model complex consumption dynamics such as interactions between consumption values, seasonal effects, and irregular usage patterns. On the code side, if these libraries are not available in the working environment, the model is preserved by substituting Random Forest or classic Gradient Boosting algorithms representing the same model class.

Time Series Cross-Validation (TS-CV) and Model Selection with MAE

This subsection explains how the performance of prediction models is evaluated in a manner appropriate to the nature of time series. In time series problems, classical shuffled cross-validation approaches can cause information about the future to influence past model parameters, a situation referred to in the literature as "data leakage". Since such leakage can cause model performance to appear more optimistic than it actually is, this study adopts a time-axis-sensitive validation strategy (Shaub 2020; Kim *et al.* 2023).

In this context, the system gradually expands the training window along the time axis using the TimeSeriesSplit approach and positions the validation set chronologically ahead of the training set at each fold. Thus, observations after time t are not used in learning the model's parameters at time t, and the forward prediction scenario is simulated realistically.

Mean Absolute Error (MAE) was chosen as the error metric for evaluating model performance. Equation (22) provides the mathematical definition of MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{22}$$

In Equation (22), y_i represents the observed actual consumption value, while $\hat{y}_i$ represents the estimate generated by the model. The MAE metric provides a more robust measure than MSE against singular jumps in time series that may contain extreme values, such as stock consumption, because it does not involve squaring. This feature makes MAE a suitable performance criterion, especially for healthcare inventory data where both high-volume consumption and zero consumption can occur.

Selecting the Top 3 Models and Weighted Ensemble

As a result of the time series cross-validation process, MAE-based performance scores were obtained for each candidate model. Using these scores, the models with the lowest error values were selected and used to create the ensemble structure. Specifically, as shown in Equation (23), the first three models with the smallest MAE values were determined:

$$\mathcal{B} = \text{Top3}\left(\arg\min_{j} MAE_j\right) \tag{23}$$

In Equation (23), $\mathcal{B}$ represents the set of models that will be included in the ensemble, while MAE_j denotes the error score obtained from the cross-validation of the j-th model.

The predictions of the selected models are combined using a performance-based weighting approach. The weights are defined as inversely proportional to the error, as shown in Equation (24) (Pawlikowski and Chorowska 2020; Shaub 2020):

$$w_j = \frac{1}{MAE_j + \varepsilon} \tag{24}$$

Here, ε represents a small constant value added to ensure numerical stability and prevent division by zero. The obtained weights are normalized, and the final ensemble estimate is calculated as shown in Equation (25):

$$\hat{y}(x) = \frac{\sum_{j\in\mathcal{B}} w_j \hat{y}_j(x)}{\sum_{j\in\mathcal{B}} w_j} \tag{25}$$

The statistical interpretation of this approach is that models with low MAE values have smaller overall error levels and are therefore targeted to reduce the expected absolute error by receiving higher weights in the combination process.

Constraining Negative Predictions (Physical Feasibility)

Since consumption quantities cannot take negative values by nature, a one-sided constraint has been applied to all model predictions to preserve the physical meaningfulness of the prediction outputs. This constraint ensures that predicted values below zero are clipped to zero, as shown in Equation (26):

$$\hat{y} \leftarrow \max(\hat{y}, 0) \tag{26}$$

This process can be interpreted as a non-negativity constraint applied in the output space and is applied both in individual model forecasts and in the ensemble output vector. This ensures that stock and consumption forecasts are consistent with physical reality.

Multi-Step ($H = 3$) Forecast Generation

In this study, the forecast horizon is set to $H = 3$ months. To generate forecasts for future periods, feature vectors for future periods are constructed using lagged values derived from the latest observations, rolling statistics, and trend components. These feature vectors are defined as $x_{t+1}, x_{t+2}, x_{t+3}$, respectively (Cao and Gui 2019).

The selected ensemble model directly produces multi-step forecasts based on these feature vectors. This process can be expressed as shown in Equation (27):

$$[\hat{y}_{t+1}, \hat{y}_{t+2}, \hat{y}_{t+3}] = \text{Ensemble}(x_{t+1}, x_{t+2}, x_{t+3}) \tag{27}$$

The code creates a `future_df` for the next three months and aligns it with the training feature space using `X_test = future_df[X_train.columns]`.

Robust Fallback Strategies

Health inventory consumption series can exhibit heterogeneous behaviors such as sparse usage, sudden spikes, institution/clinic-driven regimen changes, and seasonal fluctuations. Therefore, relying on a single model family increases the risk of "single-model fragility" (the model becoming fragile under certain data regimes). In this study, the prediction engine is designed with a multi-layered fallback architecture (Sina *et al.* 2023; Kim *et al.* 2023).

Primary Layer: Ensemble Learning: When the length of the time series created at monthly frequency is $T \geq 6$ months, the system primarily performs prediction through a feature-based ensemble learning architecture. At this stage, the time series is modeled not directly from raw values but through an explanatory feature space representing the series' temporal structure and short-term dynamics. The feature vector created for each time step includes calendar information (month, year, and quarter), the time index, lagged variables representing past consumption (values from the last one, two, and three months), local statistics such as the rolling average and standard deviation for the last three months, and the local trend slope calculated over a short window. Additionally, statistical measures summarizing the historical usage behavior of the relevant material, such as the coefficient of variation (CV), regularity of usage, and seasonality indicator, are also included in this feature space as numerical attributes. This ensures the model is sensitive not only to the latest values in the time series but also to the material's long-term behavioral characteristics.

Multiple learners are trained in parallel on this expanded feature space. Gradient-boosted tree-based methods such as XGBoost and LightGBM are used in the application environment whenever possible; when these libraries are not available, Random Forest and Gradient Boosting algorithms, which represent the same model family, are used as backups. The fundamental purpose of this model diversity is to capture nonlinear interactions, sudden jumps, or smoother trend structures that may arise in different consumption regimes using models with different biases.

Model selection and performance evaluation are performed using a TimeSeriesSplit-based cross-validation approach to prevent information leakage in time series. In this method, the training window is gradually expanded along the time axis, and each validation step is performed using only historical data. The Mean Absolute Error (MAE) is preferred as the error metric, and the most successful models are determined based on the average validation error obtained for each model. The predictions of the models with the lowest error are combined using weights inversely proportional to the corresponding error values, thus yielding the final ensemble prediction. This weighted combination strategy aims to reduce the impact of systematic errors that a single model may make under certain data regimes and to increase overall generalization capability by allowing more reliable models to contribute more to the prediction.

Secondary Layer: Exponential Smoothing Backup: The ensemble learning layer may fail in practice due to unexpected numerical or structural errors during feature extraction, model training, or prediction generation. In such cases, a complete halt in the prediction process is unacceptable for operational decision support systems. Therefore, the system incorporates the Holt–Winters Exponential Smoothing approach from classical time series methods as a backup mechanism in the second layer. This method tends to produce stable and interpretable results, especially for medium and long-length series, thanks to its fewer parameters and its ability to directly model the level, trend, and, if necessary, seasonality of the time series.

Whether or not the seasonality component is included in the model is determined conditionally. If the time series is of sufficient length (at least $T \geq 24$ months in practice) and the seasonality indicator, which shows a meaningful seasonal fluctuation in the historical behavior of the series, is above a certain threshold value, the seasonal component is added to the model. Otherwise, a simpler structure working only with level and trend components is preferred. The main objective of this approach is to ensure the continuity of the system's forecasting by compensating for the breaks that may occur in machine learning-based methods when the data is of medium or long length with a more classical and stable time series model.

Third Layer: Simple Scaled Average / Last-Value Heuristic: When the time series length is very short, i.e., especially in cases where $T < 6$ months, the risk of both machine learning-based methods and parametric time series models overfitting or producing unreasonably volatile predictions increases significantly. In such data-starved scenarios, the system consciously adopts the principle of "safe heuristics over complex modeling." If at least a few observations are available for the relevant material, the last observed consumption value is scaled by a specific coefficient (approximately 50% in practice) to produce a fixed projection. If the time series is effectively empty or there is no meaningful historical data, the estimates are fixed to zero, and these outputs are also reported with a very low confidence label.

While not numerically perfect, this approach ensures that the system exhibits a more cautious operational behavior by preventing it from producing random or overly confident but unfounded estimates in the absence of data.

When considered together, this three-layered structure presents a robust prediction architecture that enables the system to behave in an adaptable manner to data length and data quality without being dependent on a single model family. Thus, the goal is for the prediction pipeline to operate seamlessly and reliably in healthcare inventories with heterogeneous consumption patterns.

Prediction Confidence Score:

Simply producing point estimates is not sufficient for prediction outputs to be used in real-world inventory management and supply planning processes. It is also necessary to quantitatively express the conditions under which these predictions are made and the degree to which they are based on reliable information. Especially in critical areas with high operational risk, such as healthcare inventory, ignoring forecast uncertainty can lead directly to costly outcomes such as stock-outs or excess inventory. Therefore, this study defines a rule-based and explainable forecast confidence score accompanying the forecasts generated for each material.

This score is not a machine learning output; it is designed as an interpretable quality measure obtained by jointly evaluating a set of statistical indicators that reflect the amount of data, behavioral stability, timeliness, and seasonal complexity. Thus, decision-makers can obtain a quantitative answer not only to the question "How much will we consume?" but also to the question "How much can I trust this forecast?"

Components and Normalization: In order to make the different indicators calculated for each material comparable, the system first reduces all components to the range $[0, 1]$ and then calculates the average of these values to produce the overall confidence score. This approach allows criteria with different scales and meanings to be combined under a single unified quality metric.

The first component of the confidence score is the data quantity factor, which represents the amount of data on which the prediction is based. This factor is defined in Equation (28) below, using T to denote the length of the time series:

$$f_{data} = \min\left(\frac{T}{24}, 1\right) \tag{28}$$

In this formula, T represents the number of monthly observations available for the relevant material. The value of 24 used in

the denominator was chosen as an intuitive reference, assuming that at least two years of observation history is a reasonable "sufficiency threshold" to capture annual seasonal cycles. If the time series length is 24 months or longer, this factor takes the value 1, and the data quantity is considered sufficient. Conversely, for shorter series, this ratio remains below 1, and the confidence score is suppressed downward due to data scarcity.

The second component represents regularity of use. This measure is derived from the "proportion of months with zero consumption" in the monthly consumption series for the relevant material and is used directly as a normalized value between 0 and 1. A high regularity value indicates that the material is consumed regularly in most months and therefore exhibits predictable behavior. Conversely, low regularity indicates that the series has infrequent and irregular usage and that prediction uncertainty should naturally be higher.

The third component is a measure representing the stability of variation in the series and is based on the coefficient of variation (CV). CV is a dimensionless statistic defined as the ratio of the standard deviation to the mean, measuring the relative volatility of the series. In this study, CV was converted into a confidence score using a three-interval interval scoring method to enhance practical interpretability, rather than being used directly as a continuous function. If the CV value is below 0.5, the series is considered relatively stable, and in this case, a score of 0.8 is assigned to this component. If the CV is between 0.5 and 1.0, it is assumed to indicate moderate volatility, and the component score is taken as 0.6. In high volatility regimes where the CV is 1.0 or above, the series is considered highly irregular, and this factor contributes only a lower score of 0.3. This threshold directly reflects the assumption that point estimates are naturally less reliable in high-variance series.

The fourth component represents the recency of the data. The magnitude Δ used here indicates the number of days between the last output process for the relevant material and the current date. If a material has not been used for a very long time, predictions based on its past behavior are less likely to reflect current operational reality. Therefore, materials with a small Δ value, i.e., those used recently, receive a higher score. In practice, this factor is taken as 1.0 for $\Delta \leq 90$ days, reduced to 0.7 for values between 90 and 365 days, to 0.4 for values between 365 and 730 days, and to 0.1 for materials that have not been used for more than two years. This gradual decrease directly integrates the concept of "temporal decay" of information into the confidence score.

The fifth and final component is the indicator representing seasonal complexity, denoted by S. This magnitude is a normalized measure summarizing the relative intensity of seasonal fluctuations in the series. In series with a weak seasonal effect, i.e., $S < 0.3$, this factor takes a high value such as 0.8; in series with moderate seasonality ($0.3 \leq S < 0.6$), it drops to 0.6; and in series with strong seasonal fluctuations ($S \geq 0.6$), it is taken as 0.4. The basic assumption here is that as seasonality increases, the behavior of the series becomes more complex and, consequently, the uncertainty of short-term forecasts rises.

Overall Score and Verb Levels: All of the components defined above are combined to calculate a single composite confidence score for each material. The overall confidence score C is defined as the arithmetic mean of the total K components in Equation (29):

$$C = \frac{1}{K} \sum_{k=1}^{K} f_k \tag{29}$$

This expression defines f_k as the normalized sub-scores corresponding to data quantity, regularity, variation stability, timeliness, and seasonal complexity, respectively, and $K = 5$ was used in this study. The reason for choosing the arithmetic mean is to prevent any single factor from overly dominating the score by giving equal weight to all components and to obtain a more balanced reliability assessment.

This calculated continuous value is then converted into four ordinal confidence levels (e.g., "very high", "high", "medium", "low", "very low") for easier interpretation by decision-makers. This conversion ensures that the system outputs are directly usable not only numerically but also in managerial and operational reporting contexts (Subramanian 2021).

ABC–XYZ Segmentation

The outputs of the forecasting module are not limited to producing total future consumption quantities; they also reveal a rich set of information reflecting the extent to which the demand behavior of each material is predictable. From an operational inventory management perspective, the question "how much of which material will be consumed?" is not sufficient; it is also necessary to answer the questions "which materials have more stable demand, and which are more volatile and uncertain?" In order to address these two dimensions together in this study, the ABC classification, commonly used in classical inventory literature, was combined with the XYZ classification, which is based on demand predictability, to create a combined ABC–XYZ segmentation for each material. The application aims to ensure that the results obtained are transparent and reproducible in internal auditing, reporting, and decision support processes, thanks to the "rule-based" and deterministic definition of classification thresholds.

ABC: For each material in the ABC dimension, the sum of the estimated consumption quantities for the next three months is used as a proxy variable representing the relative "value" or "intensity" of the relevant item in the inventory system. This total value, V_m, is obtained by summing the three-month forecasts generated for material m, and all materials are ranked from largest to smallest according to this magnitude. Subsequently, the cumulative share of each material in the total forecast is calculated, and the classification is performed based on the position of this share within the total.

In this approach, materials in the top group with a cumulative share up to 80% of the total are labeled as Class A because they represent a large portion of the total consumption volume. Materials with a cumulative share between 80% and 95% form the Class B group, which has a medium consumption profile. Materials in the bottom 5% of the total, with a relatively low consumption volume, are classified as Class C. The selection of these thresholds is consistent with the classical ABC approach, which is based on the Pareto principle, providing a prioritization logic that directs resources and attention primarily to Class A items in inventory management.

XYZ: The XYZ dimension aims to measure the stability and predictability of demand behavior rather than the consumption quantity of materials. To this end, the coefficient of variation (CV) calculated for each material and the regularity indicator representing usage regularity are used together. CV measures the relative volatility of the series, while regularity reflects the degree of continuity in consumption over months. Evaluating these two metrics together provides a more balanced classification that accounts not

only for the magnitude of fluctuations but also for the structural continuity of the series.

Within this framework, materials with a CV value below 0.5 and a regularity value above 0.7 are classified as Class X, as they exhibit both low volatility and high usage continuity. This group represents the most predictable demand and the most reliable items for inventory planning. Materials with a CV value below 1.0 but a regularity value above 0.4 are labeled as Class Y, assuming they have medium volatility and partial regularity. This group represents a transition zone with medium uncertainty in terms of predictability. Materials that do not meet these conditions, i.e., those with irregular and complex behavior due to high variation or low regularity, are classified as Class Z and constitute the most problematic group in terms of demand forecasting.

Calculating the Optimal Stock Level

ABC–XYZ segmentation is used in this study not only to classify materials but also to systematically determine the stock management policy to be applied for each segment. In other words, the segment label obtained is treated directly as a "decision parameter," and different safety stock ratios, maximum stock levels, and control frequencies are defined for each segment. This approach ensures the establishment of an adaptable and risk-sensitive structure that takes into account both consumption volume and the predictability of demand behavior, rather than applying a uniform policy in inventory management. In practice, the matching between the segment and the inventory policy is defined through a clear and deterministic dictionary structure, ensuring that the results produced by the system are verifiable and easily integrated into corporate procedures (Goncalves *et al.* 2020).

Demand Representation: The base demand level used in inventory planning calculations for each material is obtained by taking the average of three-month forecasts representing the short-term forecast horizon. This approach aims to mitigate the impact of short-term volatility and produce a more balanced representation of "typical monthly demand" by preventing excessive reliance on a single month's forecast value. Thus, both sudden spikes and temporary declines do not disproportionately affect inventory level decisions; instead, calculations are based on a more stable reference level.

Safety Stock, Maximum Stock and Reorder Point: The safety stock and maximum stock level to be determined for each material are calculated based on policy coefficients defined according to the ABC–XYZ segment to which it directly belongs. In this context, the segment label represents a kind of "risk profile"; for example, higher safety margins are anticipated for segments that are both high-volume and low-predictability, while more limited buffer levels are considered sufficient for low-volume and more stable segments.

The safety stock can be interpreted as a buffer level obtained by applying a segment-based safety factor to the typical monthly demand for the relevant material, while the maximum stock level represents the upper limit that the stock is allowed to reach, obtained by scaling the same reference demand size with a wider multiplier.

The reorder point is defined in the system based on a rule that is practical and operationally easy to implement. In this study, the reorder threshold is considered as a fixed percentage of the maximum stock level, and this percentage is directly applied as a safety percent in practice. This choice aims to balance both the risk of excessive stock accumulation and the risk of stock depletion by automatically triggering an order when the stock level falls below a certain safe margin.

Seasonality Adjustment: The consumption behavior of some materials involves significant seasonal fluctuations. In such cases, stock levels determined by fixed coefficients may be insufficient to meet seasonal demand peaks. Therefore, for materials with a high seasonal effect, the safety stock and maximum stock levels are adjusted upward by an additional multiplier. This adjustment aims to reduce the risk of stock depletion, particularly in clinical usage scenarios where sudden and intense demand spikes occur during specific periods. Thus, the system indirectly reflects not only the average demand level but also the structural fluctuations in the temporal distribution of demand in its stock policies.

Problem Material Identification and Intervention Strategies

A critical step in operational decision support is to automatically flag risky or problematic materials on the same line where forecasts are generated. The system labels a material as "problematic" when any of the following conditions are met **?**:

- Low or very low reliability level
- Forecast quantity is zero
- $CV > 1.5$ (highly variable)
- Last use > 365 days (long period of non-use)

Behavioral Problem Groups: Problem materials are divided into subgroups based on their typical behaviors to facilitate intervention design:

- **OLD_USE**: Materials that have not been used for a very long time.
- **HIGH_VARIATION**: Materials exhibiting very high consumption volatility.
- **IRREGULAR_USE**: Materials with low regularity of usage.
- **INSUFFICIENT_DATA**: Materials with very few historical outputs.
- **COMPLEX_PATTERN**: Materials that do not fit into the above categories and exhibit complex or mixed behavior.

Action Recommendations and Prioritization: This subsection explains how actionable recommendations and priority levels are generated for each problem group, going beyond the identification of problematic materials. The goal is to transform forecasting and classification outputs from merely descriptive reports into a decision support layer that directly supports operational decision-making processes. This approach elevates the system from a "passive reporting" level to an "active and directive decision support" level.

The system determines the problem type by jointly evaluating the forecast results, uncertainty indicators, behavioral characteristics (variation, regularity, trend), and inventory classifications (ABC–XYZ) obtained for each material–warehouse combination. As a result of this comprehensive evaluation, predefined but data-triggered action templates are activated for each problem type.

For example, for materials that exhibit long-term low or zero consumption and whose predicted future consumption is also negligible, removal from stock or evaluation of alternative uses is recommended. Such materials, especially if they are in the C or Z class, can negatively impact warehouse space efficiency and inventory carrying costs. Therefore, the system flags this group as a low operational priority but strategically important problem requiring cleanup.

In contrast, for materials exhibiting high variation and irregular consumption behavior, action recommendations are generated to increase safety stock (buffer levels) or re-adjust reorder points. In such cases, the system proposes a more protective policy aimed at minimizing stock-out risk, taking into account prediction error and uncertainty levels.

When data length is insufficient or consumption patterns cannot be modeled statistically reliably, the system flags these materials as "high uncertainty" and adopts an approach that prioritizes expert evaluation over automatic decision-making. The recommended action plan for such materials is to initiate a manual review process alongside temporary conservative stock policies.

Each generated action recommendation is labeled with a priority score, taking into account the operational impact of the relevant problem, the expected risk level, and the potential cost outcome. This prioritization enables managers to focus limited resources on the most critical materials and systematizes the decision-making process. Thus, the outputs obtained from the forecasting module are transformed into actionable, traceable, and justifiable decision recommendations.

RESULTS AND DISCUSSION

Medical supply demand forecasting in healthcare institutions is a decision problem characterized by high uncertainty, dynamism, and multiple factors. Sudden changes in patient numbers, emergency department workloads, epidemic periods, updates to clinical protocols, and administrative decisions directly and often unpredictably affect consumption patterns. Therefore, healthcare inventory data are often described in the literature as irregular, sparse, highly variable, and prone to anomalies (Merkuryeva *et al.* 2019; Subramanian 2021; Feibert *et al.* 2019; Silva-Aravena *et al.* 2020; Balkhi *et al.* 2022)

The anomaly-aware and ensemble-based demand forecasting architecture developed in this study aims to directly address these structural challenges. The results obtained are interpreted by jointly evaluating anomaly detection outputs, comparisons of parametric and non-parametric methods, time series structural analysis, and operational impacts.

Forecast Model Types Distribution

In this study, the demand forecasting process was evaluated not only based on forecast accuracy but also on which model types can be preferred under which data conditions. Health inventory consumption series exhibit a highly heterogeneous structure in terms of frequency of use, continuity, variability, and clinical dependency. This situation makes it difficult to apply a uniform forecasting approach for all materials and makes data adequacy a critical factor in model selection (Subramanian 2021; Balkhi *et al.* 2022).

The model type distribution results obtained in this context reveal that a significant portion of health inventory data lacks sufficient and continuous data for model training. Particularly for infrequently used, procedure-specific, or seasonally active medical consumables, the short length of the historical observation window limits the ability of forecasting models to learn meaningful patterns. The literature indicates that such irregular and intermittent demand series specific to the health sector pose a significant challenge for both statistical time series methods and machine learning-based approaches (Silva-Aravena *et al.* 2020; Subramanian 2021).

When examining cases where data adequacy is ensured, ensemble approaches are seen to be more prominent than individual models. Linear models (ridge, lasso) offer advantages in capturing low-variance and more regular components; while tree-based methods (gradient boosting, random forest, XGBoost) can more effectively model non-linear relationships, sudden demand spikes, and complex interactions. Bringing these different model families together contributes to balancing prediction errors and obtaining more generalizable results (Shaub 2020; Kim *et al.* 2023; Chien *et al.* 2023).

These findings show that, rather than searching for the "best single model" in healthcare inventory demand forecasting, a flexible, multi-model approach that can adapt to the data structure is more rational. The model type distribution reveals that the ensemble-based strategy adopted in the study is supported not only theoretically but also by the data.

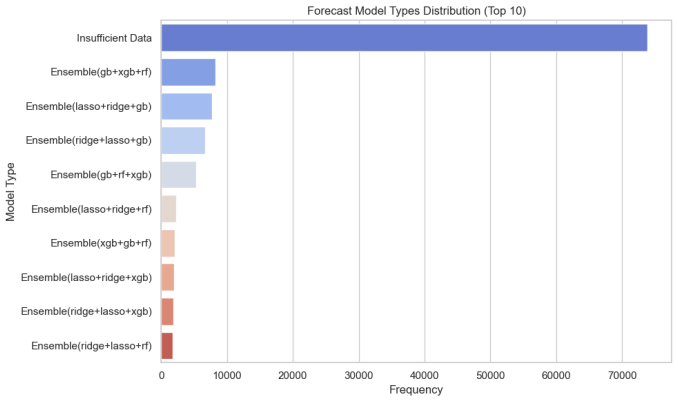


Figure 1 Distribution of the most commonly used prediction model types (Top 10).

As shown in Figure 1, one of the first outputs of the applied estimation process is the model distribution, which indicates which model types stand out in different material–unit pairs. Figure 1 presents the frequencies of the most commonly used estimation model types. The clear dominance of the "Insufficient Data" category in the graph indicates that a significant portion of the health inventory consumption data lacks sufficient and continuous observations for model training. In a hospital setting, some medical consumables are used infrequently, only in connection with specific clinical procedures, while others may be completely out of use during certain periods. This situation leads to short, irregular, and intermittent demand series, limiting the learning capacity of both statistical and machine learning-based models (Subramanian 2021; Balkhi *et al.* 2022).

When examining results outside the Insufficient Data category, it is observed that the most frequently preferred approaches are predominantly ensemble model combinations. This finding indicates that a single model family is insufficient to capture all behavioral patterns in health inventory demand series. Linear models (ridge, lasso) offer advantages in capturing regular and low-variance components, while tree-based methods (gradient boosting, random forest, XGBoost) can more effectively model non-linear relationships, sudden jumps, and interactions. Ensemble approaches, which combine these different strengths, produce more stable and generalizable predictions by balancing errors (Kim *et al.* 2023; Chien *et al.* 2023; Shaub 2020). Therefore, Figure 1 shows that the multi-model and flexible prediction strategy adopted in this study is also supported by the data.

Forecast Reliability Levels

Another factor as important as error metrics in evaluating forecast performance is the reliability level of the generated forecasts. Since forecast results directly influence operational decisions in healthcare inventory management, the question of how reliable the forecasts are is of critical importance. Particularly in demand series with high uncertainty, careful consideration should be given to how forecast outputs can be used for decision support (Goncalves *et al.* 2020). In this study, forecasts were classified under Very Low, Low, Medium, High, and Very High Reliability levels. The distribution obtained indicates that forecast uncertainty is high in a significant portion of healthcare inventory consumption data. Short data windows, high variance, sudden demand spikes, and irregular usage patterns widen the uncertainty intervals of forecast models, thereby reducing their reliability. The literature emphasizes that demand forecasting in the healthcare sector inherently involves high uncertainty and that this uncertainty must be explicitly managed (Subramanian 2021; Balkhi *et al.* 2022).

However, the presence of significant density in the Medium, High, and Very High Reliability categories indicates that some materials exhibit more regular and predictable consumption behavior. It has been observed that prediction reliability increases in series with more pronounced trend and seasonality components, lower coefficient of variation (CV), and higher observation continuity. In such series, ensemble approaches in particular are seen to produce more stable and reliable forecasts (Shaub 2020; Kim *et al.* 2023).This differentiation in forecast reliability levels reveals that a uniform forecasting strategy is not appropriate for all materials. While prediction outputs for materials in the Very Low Reliability group should be used only to a limited extent for decision support purposes, greater weight can be given to prediction-based automatic stock decisions for materials in the High and Very High Reliability groups. This approach is consistent with the literature arguing that prediction results in healthcare inventory management should be differentiated based on reliability (Goncalves *et al.* 2020).

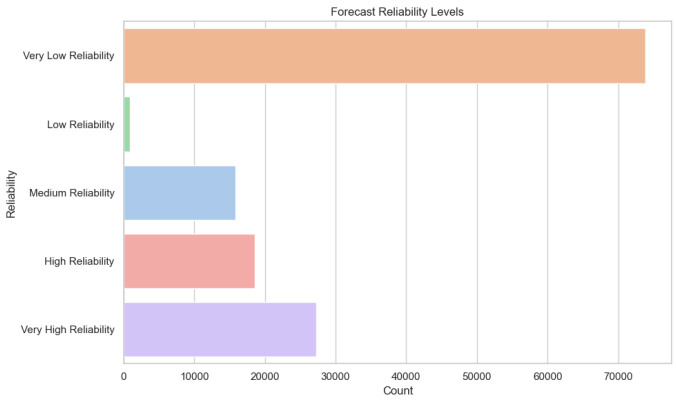


Figure 2 Distribution of prediction reliability levels.

Forecast performance has been evaluated not only through error metrics but also through the reliability levels of the forecasts. Figure 2 shows the distribution of the generated forecasts according to Very Low, Low, Medium, High, and Very High Reliability levels. The dominance of the Very Low Reliability category in the graph reveals that uncertainty is quite high in health inventory consumption series. Short data windows, high variance, sudden demand spikes, and irregular usage patterns widen the uncertainty intervals of prediction models, thereby reducing reliability (Subramanian 2021).Nevertheless, a significant concentration is also observed in the Medium, High, and Very High Reliability categories. This indicates that some materials have more regular consumption behaviour and that prediction models can produce more reliable outputs for these series. Prediction reliability increases in series where trend and seasonality components are more pronounced, the coefficient of variation (CV) is lower, and observation continuity is higher. The fact that ensemble approaches produce more successful results for materials in this group is consistent with the properties of reducing error variance and increasing generalisability emphasised in the literature (Shaub 2020; Kim *et al.* 2023).

One of the most important contributions of Figure 2 is that it shows that a single prediction strategy is not suitable for all materials. It is understood that prediction outputs should be used for decision support purposes only to a limited extent for materials in the Very Low Reliability group, whereas more weight should be given to prediction-based automatic stock decisions for materials in the High and Very High Reliability groups. This finding is consistent with studies arguing that prediction results should be differentiated based on reliability in healthcare inventory management (Goncalves *et al.* 2020; Balkhi *et al.* 2022).

Anomaly Detection Findings and Their Effects on the Health Inventory

The anomaly detection step applied prior to the forecasting process is a critical pre-processing stage in health inventory demand forecasting. The literature clearly states that feeding outliers and irregular observations directly into the model increases estimation errors, reduces model stability, and can lead to incorrect stock decisions (Ingle *et al.* 2021; Kim *et al.* 2023; Subramanian 2021; Merkuryeva *et al.* 2019).In this study, the DBSCAN algorithm is preferred for anomaly detection. Thanks to its density-based structure, DBSCAN can produce effective results in complex and heterogeneous data sets without the need for predefined threshold values or distribution assumptions.

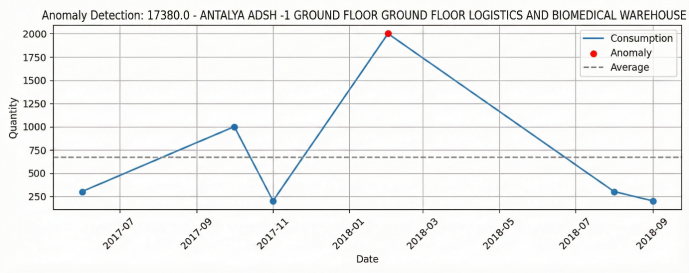


Figure 3 Singular and high-impact anomaly detected in the ADSH -1 Ground Floor Inventory and Biomedical Storage consumption series using the DBSCAN algorithm.

As observed in Figure 3, the DBSCAN algorithm clearly detected a high-volume singular anomaly in the consumption series belonging to the central storage facility. It can be seen that the consumption amount during the relevant period was significantly above the historical average. Such sudden and sharp consumption spikes are generally associated with urgent clinical needs, unexpected patient surges, or operational planning-out usage scenarios in the context of healthcare inventory (Feibert *et al.* 2019; Subramanian 2021).The literature emphasises that such singular but highly impactful anomalies create a disproportionate disruptive effect on prediction models. Particularly when error metrics have

a squared structure (e.g., RMSE), such outliers can significantly degrade model performance (Merkuryeva *et al.* 2019; Ingle *et al.* 2021).Therefore, identifying and separating these anomalies prior to the prediction process facilitates the model's learning of the overall behaviour.

The most significant contribution of the DBSCAN algorithm in this example is its ability to successfully distinguish sudden changes in consumption intensity without relying on distribution assumptions or using fixed threshold values. This clearly demonstrates why density-based approaches are more suitable for heterogeneous and irregular data structures such as health inventories (Balkhi *et al.* 2022).

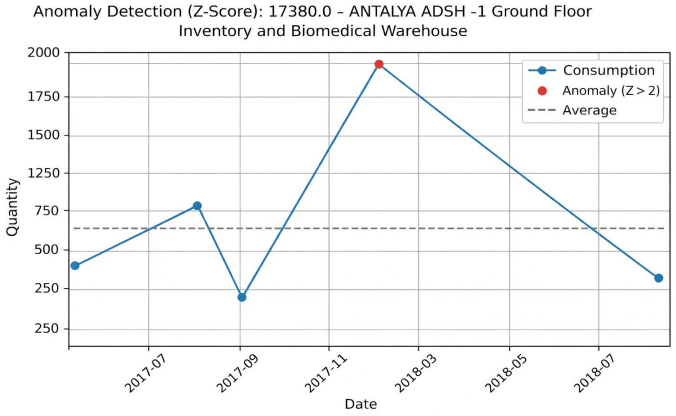


Figure 4 Multiple and irregular anomalies detected in the consumption series of the Güzeloba ADSP Laboratory unit using the DBSCAN algorithm

As shown in Figure 4, the consumption series of the laboratory unit contains multiple and irregular anomalies. Such consumption patterns are frequently observed in healthcare institutions because laboratory services are directly linked to patient profiles and clinical demand. Fluctuations in diagnostic and examination processes, in particular, can cause changes in consumption series intensity (Polater and Demirdogen 2018; Feibert *et al.* 2019).

The most important point to note in Figure 4 is that anomalies appear not only at a single point but at different time intervals and at different density levels. This explains why parametric methods based on fixed thresholds can be inadequate. The DBSCAN algorithm has successfully modelled this complex structure thanks to its ability to distinguish between different density regions. The literature indicates that consumption series for laboratory and imaging units in healthcare supply chains are typically multimodal and irregular (Balkhi *et al.* 2022; Subramanian 2021). In this context, Figure 4 visually supports why the DBSCAN-based approach should be preferred in complex systems such as healthcare inventory.

Comparative Evaluation of Parametric and Non-Parametric Anomaly Approaches

In order to evaluate the effectiveness of the DBSCAN-based approach more comprehensively, a comparative analysis was performed with the Z-Score, a classical parametric method. The Z-Score method is based on the assumption that the data follows an approximate normal distribution and identifies outliers using fixed threshold values.

Figure 5 shows the anomaly detection results obtained using the Z-Score method for the same central storage facility. Although

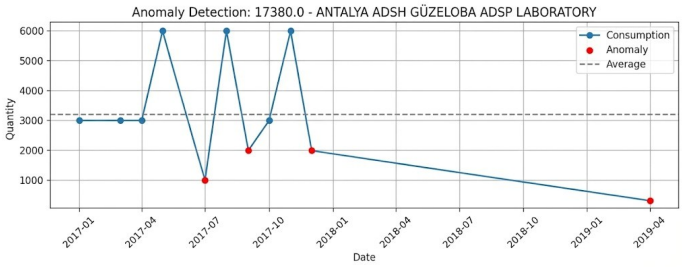


Figure 5 Anomalies detected in the ADSH -1 Ground Floor Inventory and Biomedical Storage consumption series using the Z-Score method.

the Z-Score method was able to detect significant outliers, it could only capture structural fluctuations and intensity changes in the consumption series to a limited extent. This situation stems from the Z-Score's dependence on the distribution assumption. Health inventory consumption data often do not satisfy the normal distribution assumption; instead, they exhibit skewed, multimodal, and seasonally varying structures (Merkuryeva *et al.* 2019; Subramanian 2021).It is frequently emphasised in the literature that methods based on fixed thresholds in such data structures can only detect extreme observations but may overlook more complex anomalies (Ingle *et al.* 2021; Kim *et al.* 2023).

In this context, Figure 5 clearly illustrates the limitations of the Z-Score method in the health inventory context and supports why non-parametric approaches such as DBSCAN offer more flexible solutions (Silva-Aravena *et al.* 2020).

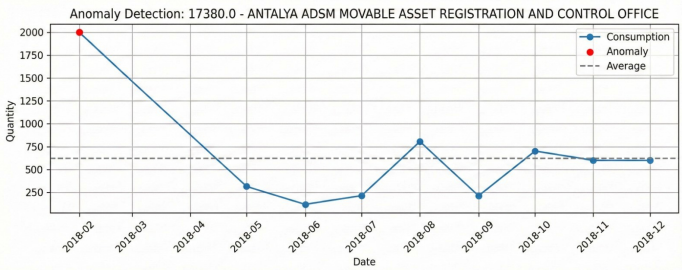


Figure 6 Results of anomaly detection performed on the consumption series of the ADSM A17 X-ray unit using the Z-Score method.

Figure 6 clearly shows that the Z-Score method can completely miss anomalies in some periods. The consumption series for the X-ray unit exhibits structural breaks over time due to changing clinical demands and patient profiles. The performance of methods based on fixed thresholds is significantly reduced in such series.

The literature indicates that consumption series for imaging and diagnostic units typically contain seasonality, trends, and sudden increases in demand (Feibert *et al.* 2019; Balkhi *et al.* 2022). In this context, Figure 4 illustrates why parametric methods are limited in the healthcare inventory context and why approaches based on distribution assumptions do not always produce reliable results. This finding directly aligns with studies explaining why non-parametric methods and density-based approaches are increasingly preferred in the healthcare supply chain literature (Merkuryeva *et al.* 2019; Subramanian 2021).

Following anomaly cleaning, the time series decomposition method was applied to enable a more in-depth analysis of the

fundamental behavioural characteristics of the consumption series. This analysis allows for the separate examination of the trend, seasonality, and residual components.

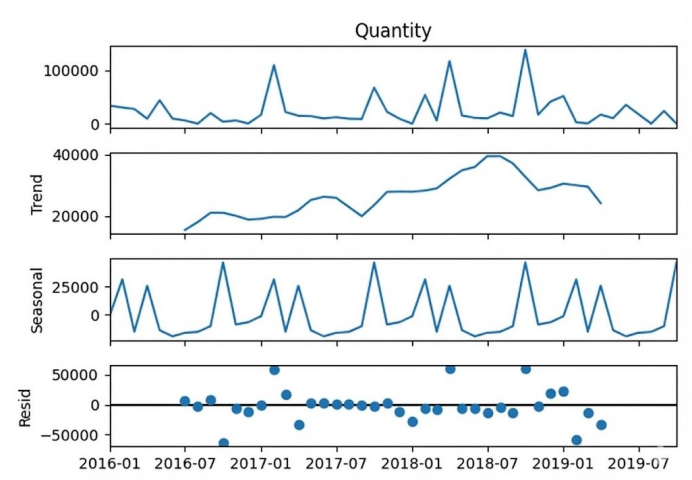


Figure 7 Time series decomposition analysis for material 17380: trend, seasonality and residual components.

Figure 7 shows the time series decomposition results obtained for material 17380. While a long-term increasing structure is observed in the trend component, pronounced fluctuations are noticeable in the seasonal component. This indicates that the consumption of the relevant material exhibits both an increasing trend over time and is influenced by seasonal factors. The irregularities observed in the residual components reveal that this series has a structure susceptible to anomalies. The literature emphasises that single forecasting models are generally inadequate for such complex time series, and that ensemble approaches produce more successful results (Kim *et al.* 2023; Chien *et al.* 2023; Sina *et al.* 2023; Shaub 2020).

In this context, Figure 5 is one of the key figures that visually supports why ensemble and hybrid forecasting approaches are necessary.

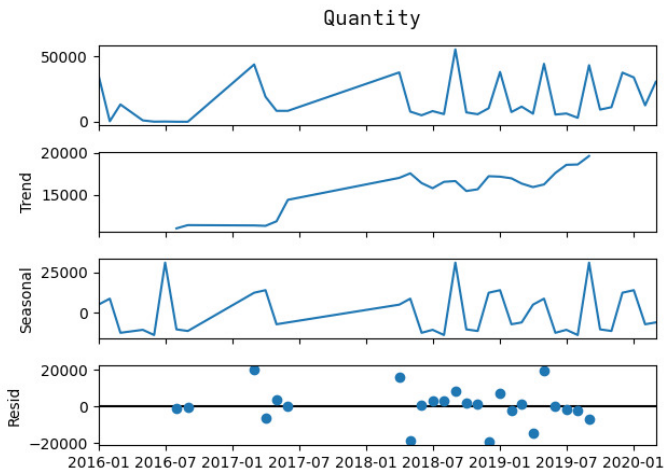


Figure 8 Time series decomposition analysis for material 17305: trend, seasonality and residual components.

Figure 8 presents the decomposition results obtained for a material with a high-volume and complex consumption structure. In addition to the trend and seasonality components, the high variance observed in the residual components indicates that this series is both prone to anomalies and difficult to forecast. The literature indicates that such high-volume and irregular series cannot be effectively predicted using simple statistical models; instead, multiple model combinations provide more stable results (Sharma *et al.* 2021; Sina *et al.* 2023). Therefore, Figure 6 is of critical importance in justifying the proposed ensemble-based approach.

CONCLUSION

The findings obtained within the scope of this study not only present results related to statistical prediction accuracy but also demonstrate that the developed anomaly-aware and ensemble-based prediction architecture produces outputs that can be directly integrated into the operational and strategic decision-making processes of healthcare institutions. As frequently emphasised in the literature, healthcare inventory data has an irregular, sparse, highly variable, and anomaly-prone structure necessitated the design of the developed system to directly address these structural challenges (Merkuryeva *et al.* 2019; Subramanian 2021; Feibert *et al.* 2019; Silva-Aravena *et al.* 2020; Balkhi *et al.* 2022).The anomaly-aware forecasting process enables unexpected demand spikes to be detected at an early stage, contributing to reducing the risk of stock depletion, preventing excessive inventory costs, and ensuring service continuity (Karamshetty *et al.* 2022; Goncalves *et al.* 2020). The generated forecasts are not only reported as point values but are also presented with an explainable reliability score calculated based on indicators such as the amount of data, behavioural stability, timeliness, and seasonal complexity for each material. Thus, decision-makers can obtain a quantitative answer not only to the question 'how much will we consume?' but also to the question 'to what extent can I trust this forecast?' (Subramanian 2021).

Situation indicates that a uniform stock and forecasting policy for all materials will not be effective and necessitates the integrated use of ABC classification, commonly used in classical inventory literature, and XYZ classification, which is based on demand predictability (Aktunc *et al.* 2019; Babai *et al.* 2015; Demiray Kirmizi *et al.* 2024).Thanks to the combined ABC–XYZ segmentation developed within the scope of the study, while stock removal or alternative usage evaluations are recommended for materials exhibiting low and irregular consumption, more protective policies such as increasing safety stock and redefining reorder points are systematically developed for critical materials exhibiting high variation. In cases where the data length is insufficient or the consumption pattern cannot be modelled in a statistically reliable manner, a conservative approach prioritising expert assessment is adopted instead of automatic decision-making. This holistic structure enables the evolution of forecasting outputs into actionable, traceable, and justifiable decision recommendations, demonstrating the practical application of integrated forecasting–classification approaches advocated in the literature for healthcare inventory management (Feibert *et al.* 2019; Subramanian 2021).

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Abdul-Rahman, S., N. H. Zulkifley, I. Ibrahim, and S. Mutalib, 2021 Advanced machine learning algorithms for house price prediction: Case study in Kuala Lumpur. Science and Information Organization .

Adedunjoye, A. S. and J. O. Enyejo, 2024 Leveraging predictive analytics to improve demand forecasting and inventory management in healthcare supply chains. International Journal of Scientific Research in Science Engineering and Technology .

Aifuwa, S. E., T. Oshoba, E. Ogbuefi, P. N. Ike, S. B. Nnabueze, *et al.*, 2020 Predictive analytics models enhancing supply chain demand forecasting accuracy and reducing inventory management inefficiencies. International Journal of Multidisciplinary Research and Growth Evaluation .

Aktunc, E. A., A. Atamturk, and E. Demir, 2019 Demand classification and inventory control in healthcare systems using ABC–XYZ analysis. European Journal of Operational Research **274**: 539–551.

Babai, M. Z. *et al.*, 2015 Demand forecasting and inventory control in healthcare: A review. International Journal of Production Economics **167**: 130–146.

Balkhi, Z. *et al.*, 2022 Healthcare supply chain management: A review of quantitative models. Operations Research for Health Care **33**: 100349.

Cao, Y. and L. Gui, 2019 Multi-step wind power forecasting model based on LightGBM. In *Proceedings of the International Conference on Signal and Artificial Intelligence (ICSAI 2018)*, pp. 192–197.

Chien, C.-F., Y.-J. Chen, and J.-T. Peng, 2023 Hybrid and ensemble forecasting models for complex demand patterns. International Journal of Forecasting **39**: 287–302.

Dachepalli, V., 2025 An ARAS-based evaluation of AI applications for demand forecasting and inventory management in supply chains. International Journal of Cloud Computing and Supply Chain Management .

Darshan, U., G. P. Chinmaya, H. Abhinav, S. V. Kumar, R. Deepak, *et al.*, 2025 Predictive analysis on medicines availability in hospitals using machine learning and deep learning technique. International Journal For Multidisciplinary Research .

Demiray Kirmizi, M. *et al.*, 2024 Integrated inventory classification and forecasting in healthcare supply chains. Journal of Manufacturing Systems **72**: 102–118.

Donkor, A. A., S. A. Dada, A. Korang, and J. Umoren, 2024 The role of artificial intelligence and machine learning in optimizing U.S. healthcare supply chain management. World Journal of Advanced Research and Reviews .

Erdebilli, B. and B. Devrim-Tenba, 2022 Ensemble voting regression based on machine learning for predicting medical waste: A case from Turkey. Multidisciplinary Digital Publishing Institute .

Fatima, S. S. W. and A. Rahimi, 2024 A review of time-series forecasting algorithms for industrial manufacturing systems. Multidisciplinary Digital Publishing Institute .

Feibert, D. C., R. K. Hansen, and P. Jacobsen, 2019 Measuring supply chain resilience in healthcare: An empirical study. Supply Chain Management: An International Journal **24**: 78–96.

Gldoan, E., F. H. Yan, A. Pinar, C. Olak, S. Kadry, *et al.*, 2023 A proposed tree-based explainable artificial intelligence approach for the prediction of angina pectoris. Nature Portfolio .

Goncalves, J. N. *et al.*, 2020 Decision support systems for healthcare inventory management. Computers & Operations Research **113**: 104785.

Gurumurthy, A., V. K. Nair, and S. Vinodh, 2021 Application of a hybrid selective inventory control technique in a hospital. The TQM Journal **33**: 568–595.

Ingle, V., P. Shah, and A. Shukla, 2021 Time series forecasting for healthcare demand using hybrid models. Expert Systems with Applications **184**: 115555.

Jahin, M. A., A. Shahriar, and M. A. Amin, 2024 MCDFN: Supply chain demand forecasting via an explainable multi-channel data fusion network model. Evolutionary Intelligence .

Joshi, S. V., S. B. Patil, S. N. Patil, S. Bhosle, N. Karhadkar, *et al.*, 2025 Machine Learning Applications in Hospital Pharmacy for Predicting Drug Shortages and Supply Chain Optimization. Research Journal of Pharmacy and Technology .

Karamshetty, S., R. Akkiraju, and K. Bhaduri, 2022 Anomaly detection in supply chain demand using density-based approaches. IEEE Access **10**: 41245–41257.

Khalid, O., 2024 Short-term and long-term product demand forecasting with time series models. Journal of Trends in Financial and Economics .

Kim, S., J. Park, and D. Lee, 2023 Ensemble learning-based demand forecasting for healthcare inventory management. Applied Soft Computing **132**: 109840.

Kolambe, S. A. M., 2024 Forecasting the future: A comprehensive review of time series prediction techniques. None .

Kontopoulou, V. I., A. D. Panagopoulos, I. Kakkos, and G. K. Matsopoulos, 2023 A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. Multidisciplinary Digital Publishing Institute .

Luo, L., L. Luo, X. Zhang, and X. He, 2017 Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. BMC Health Services Research .

Mbonyinshuti, F., J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare, 2022 The prediction of essential medicines demand. Processes **10**: 26.

Mbonyinshuti, F., J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare, 2024 Health supply chain forecasting: A comparison of ARIMA and LSTM time series models for demand prediction of medicines. Acta Logistica .

Merkuryeva, G., A. Valberga, and A. Smirnov, 2019 Demand forecasting in healthcare supply chains: A systematic literature review. Procedia Computer Science **149**: 541–550.

Motamedi, M., J. W. Dawson, N. Li, D. G. Down, and N. M. Heddle, 2024 Demand forecasting for platelet usage: From univariate time series to multivariable models. Public Library of Science .

Pawlikowski, M. and A. Chorowska, 2020 Weighted ensemble of statistical models. International Journal of Forecasting **36**: 93–97.

Perone, G., 2021 Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. Springer Science+Business Media .

Polater, A. and G. Demirdogen, 2018 Healthcare logistics and demand uncertainty: A case-based analysis. Journal of Health Systems **4**: 45–59.

Punnahitanond, W., P. Kantavat, N. Prompoon, and T. Ananpiriyakul, 2025 Forecasting drug dispensation using machine learning: A case study on a large public hospital. In *2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)*.

Qiu, H., L. Luo, Z. Su, L. Zhou, L. Wang, *et al.*, 2019 Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. BMC Medical Informatics and Decision Making .

Sharma, A., R. Gupta, and J. Saini, 2021 Machine learning approaches for demand forecasting: A review. Materials Today: Proceedings **46**: 11131–11138.

Shaub, M., 2020 Why ensemble learning improves forecasting accuracy. Journal of Forecasting **39**: 38–55.

Shern, S. J., M. T. Sarker, M. H. S. M. Haram, G. Ramasamy, S. P. Thiagarajah, *et al.*, 2024 Artificial intelligence optimization for user prediction and efficient energy distribution in electric vehicle smart charging systems. Energies .

Shih, H. and S. Rajendran, 2019 Comparison of time series methods and machine learning algorithms for forecasting Taiwan Blood Services Foundation's blood supply. Hindawi Publishing Corporation .

Silva-Aravena, M., R. Morales, and L. Cisternas, 2020 Forecasting healthcare demand using time series and machine learning models. Computers & Industrial Engineering **143**: 106419.

Sina, M., M. Rahimi, and R. Tavakkoli-Moghaddam, 2023 A hybrid machine learning approach for demand forecasting in healthcare supply chains. Computers & Industrial Engineering **176**: 108933.

Subramanian, N., 2021 Healthcare inventory management: A review of demand forecasting and decision support systems. International Journal of Production Economics **231**: 107867.

Umoren, J., T. O. Agbadamasi, T. K. Adukpo, and N. Mensah, 2025 Leveraging artificial intelligence in healthcare supply chains: Strengthening resilience and minimizing waste. EPRA International Journal of Economics, Business and Management Studies .

Vignesh, A. and N. Vijayalakshmi, 2025 A weighted ensemble model combining ARIMA, LSTM, and GBM for robust time series prediction. Scientific and Technical Journal of Information Technologies, Mechanics and Optics .

How to cite this article: Özkurt, C., Küçükler, A. K., Karslıoğlu, M., and Özdemir, R. N. Enhancing Hospital Inventory Forecasting Accuracy through Hybrid and Ensemble Learning Models. *Computers and Electronics in Medicine*, 3(1), 60-76, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.77-85
https://doi.org/10.69882/adba.cem.2026019

Classification of Brain MRI Images using Deep Learning: The DeiT3 Model and the Use of Feature Fusion Methods

Sena Kahraman *,1 **and Mesut Toğaçar** α,2

*Technology and Information Management, Institute of Social Sciences, Fırat Unıversity, Elazığ 23119, Türkiye, αDepartment of Management Information Systems, Faculty of Economics and Administrative Sciences, Fırat Unıversity, Elazığ 23119, Türkiye.

ABSTRACT Brain tumors are among the diseases that can seriously threaten human life and can be fatal. Early diagnosis of brain tumors plays a crucial role in the treatment process of the disease. However, accurately and quickly diagnosing this disease remains one of the significant challenges of modern medical technologies. Currently, advanced imaging techniques such as magnetic resonance imaging (MRI) are generally used for detecting brain tumors. This study proposes an artificial intelligence-based diagnostic approach using MRI images that include brain tumor types and consist of four classes. The proposed approach includes preprocessing, model training, feature fusion, and selection as final steps. In the preprocessing step, Grad-CAM and LBP techniques are applied to the original dataset, resulting in a total of three datasets, including the original one. These datasets are then trained with the DeiT3 model to obtain three separate feature sets (original, Grad-CAM-based, LBP-based). The feature sets are fused using a feature fusion technique, and the performance of the combined sets is evaluated using SVM methods. Feature selection methods (Chi2, Relief) are applied to the best-performing Grad-CAM & LBP-based feature set to highlight the most efficient features. Experimental analysis results show that a success rate of 99.5% was achieved using the SVM method.

KEYWORDS
Brain tumor
Transformer model
Image processing
Feature fusion
Feature selection

INTRODUCTION

A brain tumor is a mass formed by the abnormal growth of cells located in the brain. The brain is the control center of the body, enabling humans to perform their basic life functions. Therefore, the detection and treatment of brain tumors are also very important for human life. It has been observed that brain tumor disease significantly affects mortality rates (Yılmaz 2023). It has been reported that the number of people dying from brain tumor disease in China is between 50,000 and 100,000 per year, and that approximately 80,000 people are diagnosed with brain tumors in the United States each year (Taşdemir and Barışçı 2024). When looking at causes of death, brain tumors rank tenth worldwide and eighth in Türkiye (Erçelik and Hanbay 2023b).

Medical imaging techniques, which are also used in the diagnosis of many diseases, are used to detect brain tumors. Medical imaging techniques include techniques such as computed tomography, magnetic resonance imaging (MRI), mammography, etc. The most preferred imaging technique for detecting brain tumors is MRI. Erçelik et al. stated that it is difficult to detect brain tumors using only MRI images and that this is due to the complex structure of the brain. The difficulty doctors face in diagnosing diseases using such traditional methods and the advancement of technology have led to the emergence of the concept of "Artificial Intelligence in Health." Today, approaches such as image processing and deep learning are used for disease detection (Serttaş and Deniz 2023). There are multiple types of brain tumors. Some studies extract features for classification, while others incorporate deep learning methods (Aslan 2022).

Numerous artificial intelligence-based studies have been conducted on brain tumors in the literature. Examining some recent studies, Aslan (2024) presents the LSTM-ESA model, which combines LSTM (Long Short-Term Memory) and ESA (Convolutional Neural Network Numerous artificial intelligence-based studies have been conducted on brain tumors in the literature. E s). He states that this model achieved a score of 98.1% in brain tumor detection. He notes that this score is higher than the score achieved by the ESA model. Demirel and Soylu (2024) aimed to compare the performance of MobileNet, DenseNet-121, and DenseNet-201 models in terms of brain tumor detection in their study. It was stated that all models achieved excellent accuracy rates during the training phase, but the DenseNet-121 model showed the best performance in terms of generalization. It was stated that an accuracy rate of 98.81% was achieved with the DenseNet-121 model. Erçelik and Hanbay (2023a) aimed to compare the Gaussian Filtering and ResNet50 models and identify the model with the best performance. Noise components in the images were cleaned us-

Manuscript received: 11 November 2025,
Revised: 15 January 2026,
Accepted: 15 January 2026.

[1]sena_kahraman2002@hotmail.com (**Corresponding author**)
[2]mtogacar@firat.edu.tr

ing the Gaussian function. It was determined that cleaning the noise yielded better performance compared to ResNet50. Yenikaya and Oktaysoy (2023) aimed to test the success of ResNet101 and GoogLeNet models in brain tumor detection in their study. At the end of model training, it was found that the ResNet101 model achieved better success than the GoogLeNet model with 91.5%. The ResNet model achieved 87.9% success. Das *et al.* (2025) developed a method using a VGG-16-based deep learning model to accurately classify brain tumors in FLAIR MR images. The model classified these images, which had three different tumor classes, with 99% accuracy. In their study, Yang *et al.* (2025) developed an artificial intelligence model called GMDNet to accurately classify brain tumor MR images. With this model, they successfully classified the relationship between different MR images. The GMDNet model also provided high accuracy in the presence of missing data. To achieve this, a special method called reuse modality was developed.

In this study, data diversity was increased and the model's learning capacity was enhanced by applying Grad-CAM and LBP techniques to the original data set. The model was trained using next-generation technology-based transformers. The features extracted from the model training offer an innovative approach that is more powerful and contributes to performance thanks to the feature fusion technique. Feature selection was applied to the dataset that yielded the best performance as a result of feature fusion, and the most efficient features were obtained. Consequently, time and cost savings were also achieved. This developed approach produces reliable outputs in the brain tumor detection process and makes an important contribution to the early diagnosis of the disease. This will help in rapid detection and determining the correct treatment methods. A brief summary of the other sections of the article is as follows: Information about the data set and model training is provided in the second section. Detailed information about the analysis results is provided in the third section. The fourth section contains the discussion. The final section, the fifth section, contains the conclusions.

MATERIALS AND METHODS

In this study, a hybrid approach incorporating the DeiT3 model and feature fusion is proposed for the automatic classification of brain MRI images according to brain tumor types. The proposed hybrid model consists of data set preparation, preprocessing steps, model training, feature fusion, and feature selection steps.

Dataset

In this study, a new data set was used, created from a combination of the Figshare (Cheng 2024), Sartaj dataset (Bhuvaji 2025), and Br35H (Hamada 2020) data sets, which consist of brain MRI images. This data set was obtained from the open-access Kaggle website (Nickparvar 2021). The dataset contains 7,023 brain MRI images in JPG format, divided into four classes. The class types consist of glioma, meningioma, pituitary diseases, and non-tumor MRI images. The non-tumor class of the Br35H dataset was not included in the combined new dataset. By class type, there are 1,621 glioma, 1,645 meningioma, 1,757 pituitary, and 2,000 non-tumor images. The class types are evenly distributed. The images in the dataset have variable resolution. A sample subset of images from the dataset is shown in Figure 1.

Image processing methods

Gradient-weighted class activation mapping (Grad-CAM) is a technique used to visualize the regions that a convolutional neural network (CNN) model considers most important when making predictions in the form of a heatmap (Senjoba *et al.* 2024). Grad-CAM helps understand which regions influence an image processing model to select a specific class during classification (Livieris *et al.* 2023). This method is also widely used in different fields such as object detection and medical image analysis. It is a general and flexible method that can be applied to different CNN architectures (Ennab and Mcheick 2025). In this study, Grad-CAM was trained using the DarkNet-19 model architecture. This facilitated the detection of diseased regions in images and allowed Grad-CAM to focus on these regions in the images. Sample images from the new dataset obtained by applying the Grad-CAM technique to the original dataset are shown in Figure 2.

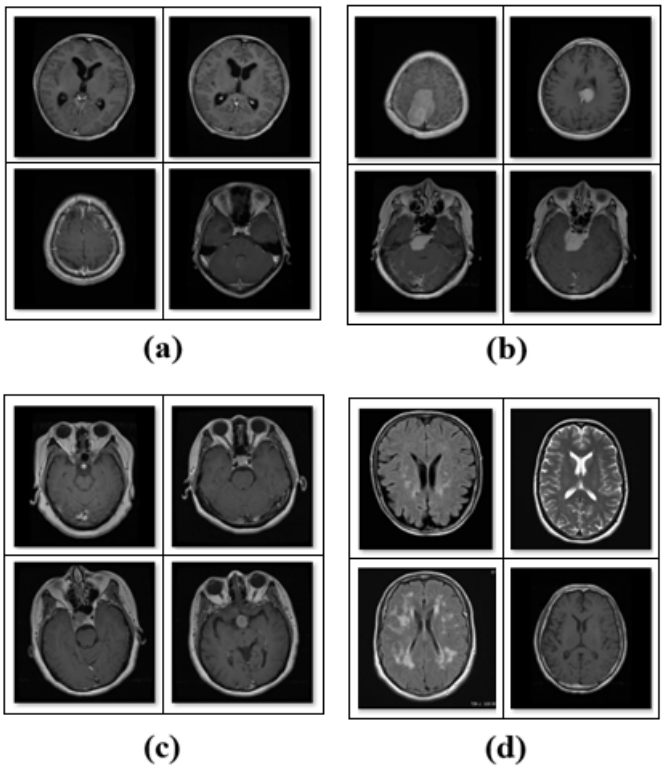


Figure 1 Sample images of the class types in the dataset: (a) glioma, (b) meningioma, (c) pituitary, and (d) non-tumor.

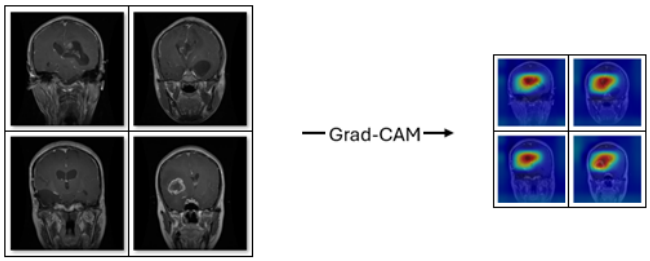


Figure 2 Images from the original dataset and the dataset created using Grad-CAM.

Local Binary Pattern (LBP) is a widely used method in image processing and computer vision for extracting texture features from

images (Almohamade *et al.* 2025). Its computational efficiency and adaptability to different scenarios demonstrate its versatility. It produces a binary code consisting of 0s and 1s by comparing the intensity of each pixel in an image with its neighboring pixels. The resulting binary codes describe the local textural structure of the image. These codes are converted into a histogram, which is a numerical representation summarizing the texture in the image (Attallah 2025). This histogram provides features (Aydemir 2022). The steps of the LBP method are shown in Figure 3. Sample images

Figure 3 Stages of the LBP method.

from the new dataset obtained by applying the LBP technique to the original dataset are shown in Figure 4.

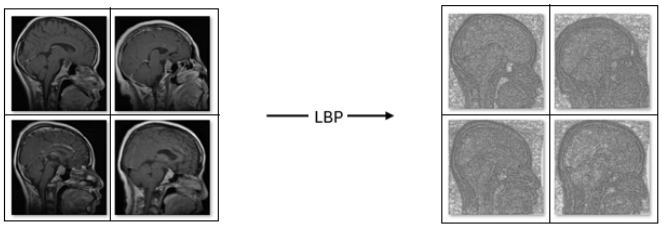


Figure 4 Images from the original dataset and the dataset created by applying LBP.

New generation transformer model: DeiT3

Transformer models typically process data in six stages. Sometimes, additional layers may be included in this basic process due to structural differences in these models. The first of these stages involves dividing the received image into smaller pieces. The divided image pieces are converted into vectors by passing through an embedding layer. Location information is added to each piece to ensure that the model preserves the spatial order. Subsequently, information is processed in layers through encoder blocks, and deep features are extracted. The Multi-Head Self-Attention mechanism identifies relationships between image fragments and enables the model to generate contextual meanings. The information obtained in the final layer is processed using neural networks and transmitted to the output layer for classification (Toğaçar 2025).

Figure 5 Stages of the LBP method.

Data-efficient image transformers (DeiT) are a transformer-based deep learning architecture that aims for high performance with minimal data. DeiT is derived from the vision transformer (ViT) architecture. It reduces data requirements by utilizing the knowledge distillation method. In this architecture, the input image is typically resized to 224x224 pixels. The image is then divided into small 16x16 pixel patches. Location information and distillation tokens are added to each patch to preserve the structural features of the data. These patches are then converted into a vector format that the transformer can process. The relationships between the pieces are analyzed through the attention mechanism and feedforward layers. The encoder block consists of three recurrent self-attention and feedforward neural network layers. In the final stage, the image labels are determined through the classification layer (Sevinc *et al.* 2025).

Feature selection methods

The chi-square (Chi2) test is a statistical method used for feature selection. This test helps determine how one variable affects another. Specifically, it examines the relationship between the target variable and the features. Features that show a stronger relationship with the target variable are selected, while independent feature variables are removed. This ensures that only important features are included in the model (Rahman *et al.* 2023). The Chi2 test works by calculating the difference between expected and observed frequencies. If there is independence between two features, the Chi2 value will be low. A high Chi2 value indicates a stronger relationship with the target variable. This feature selection method is particularly effective in high-dimensional and complex data sets. However, it can lead to erroneous results in low-frequency cells and cause poor performance in data imbalance (Devi *et al.* 2023).

The RelieF method is an effective feature selection method commonly used in classification. This method aims to improve model performance by analyzing the importance level of different features in the data. For each data point, the nearest neighbors are analyzed; the weight score for each feature is determined using both neighbors from the same class and neighbors from different classes. In improving model performance, features with high weight values have a more significant impact than those with low weights. Features with low weights may contain unnecessary or misleading information. The greatest advantage of the RelieF method is its ability to deliver successful results in high-dimensional, complex, and large datasets (Gür *et al.* 2025).

Support vector machines method

Support vector machines (SVM) effectively perform both regression and classification operations on high-dimensional and complex data sets. Data is transformed into a higher-dimensional space using kernel functions. This makes it easier to capture non-linear relationships between features. The model's performance and confusion are adjusted through the regularization parameter and kernel coefficient (gamma) (Haldar *et al.* 2025). The SVM algorithm is a powerful machine learning technique that works successfully with both linear and non-linear data structures, providing high accuracy by making clear distinctions between classes. It achieves high accuracy by separating classes with the maximum margin (Toğaçar 2025). It can classify with high accuracy despite complex and noisy data (Tubog *et al.* 2025). The parameter information and selected values of the SVM method used in the experimental analyses of this study are given in Table 1.

Proposed Approach

The proposed approach consists of artificial intelligence-based models that analyze CT images to detect brain tumors. In this study, the DeiT3 model from the ViT family forms the basis of this system to successfully detect the disease. This approach combines the DeiT3 model with approaches such as Grad-CAM and LBP to achieve more successful disease detection. The proposed approach consists of preprocessing steps, model training, feature fusion, and post-processing steps such as feature selection.

In the preprocessing step, image processing methods are applied. CT images collected in a hospital environment may have

Table 1 Parameters of the SVM method used

Parameter	Selection/Value
Function	Cubic
Box Restriction Level	1
Core Scale Mode	Automatic
Multi-Class Method	One-on-One

different resolutions. In this step, all images are cropped to a fixed resolution of 224 × 224 pixels. This resolution is chosen because ViT-based models process images of this size as input. Then, new image sets are created using image processing methods (Grad-CAM and LBP) on the original data set. DeiT3, which stands out among the new generation image transformers, is used in the model training step. Feature extraction is applied to the data sets trained with the DeiT3 model. In this step, 768 features are obtained from each of the Grad-CAM (A), original (B), and LBP (C) datasets. Feature fusion is then applied to the datasets, and the best fusion set is selected from the resulting fusion sets. The goal of this step is to determine the most efficient feature set for feature selection in the next step.

In the final processing step, the most significant features are selected from the 1536 features obtained from the best combination set (for this study, the combination of B and C is the best combination set) using Chi2 and Relief techniques, respectively. This approach assigns a score to each of the 1536 features according to its own statistical methods and ranks them. The features are ranked from highest to lowest according to their importance. The most meaningful 1000 or 500 features are selected from this ranking. They are then classified using the SVM method.

The design of the proposed model is shown in Figure 6.

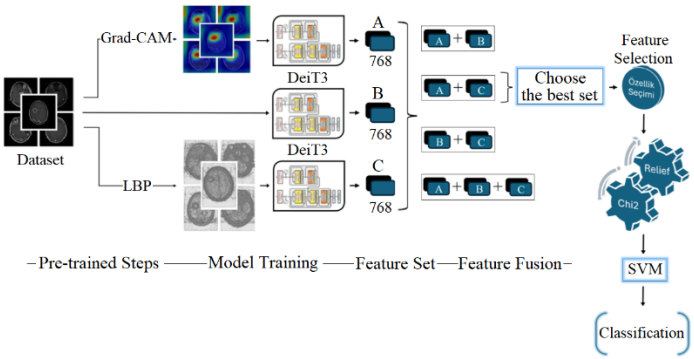


Figure 6 Design of the proposed model.

EXPERIMENTAL RESULTS

During the experimental analyses, MATLAB 2024 software was used to perform feature fusion, feature selection, and classification stages (SVM method). The Python programming language was used for the LBP and Grad-CAM methods, which are preprocessing steps, and for training the DeiT3 model, and these codes were run via the Jupyter Notebook interface. A computer with a 3.40 GHz Intel Core i7 processor, 32 GB RAM, and a 10 GB graphics card was used to perform the experimental analyses. A confusion matrix, which provides the number of correctly and incorrectly classified data in the dataset, was used to evaluate the analysis results (Çelik and Koç 2021). The metrics and formulas for the confusion matrix are given below. There are four metrics: accuracy, f-score, precision, and sensitivity (Kuştaşı and Yağanoğlu 2024). The equations contained in the metrics have four basic classification elements: positive (P), negative (N), true (T), and false (F).

$$\text{Sensitivity (Se)} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision (Pre)} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{F-score (f-scr)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Accuracy (Acc)} = \frac{TP + TN}{FP + FN + TP + TN} \tag{4}$$

The preferred parameters in the recommended approach are given in Table 2. Default values have been selected for parameters other than those listed.

Table 2 Parameters used in the proposed approach and selected values

Model / Method	Parameter	Preference / Value
DeiT3	Classifier	Linear
DeiT3	Epoch	11
DeiT3	Learning rate	10-4
DeiT3	Loss function	CrossEntropyLoss
DeiT3	Mini-batch	32
DeiT3	Optimization	SGD
DeiT3	Training & Test rate	%80 – %20
SVM	Preset	Cubic
SVM	Core scale	Auto
SVM	Box restriction level	1
SVM	Multi-class method	One-vs-One

The experimental analysis consists of three stages. The first stage consists of the preprocessing process. In this step, Grad-CAM and LBP methods were applied to the original data set, resulting in three different image sets. The images belonging to these three data sets are shown in Figure 7.

The second stage involves evaluating the performance of the three datasets obtained in the first stage using a model. In this stage, the new-generation DeiT3 model, developed to increase the efficiency of transformation models in image processing, was used. After the three data sets were trained by the model, the training-test success graphs shown in Figure 8 were obtained. The confusion matrices obtained at the end of the model training are given in Figure 9. The overall success of the metric results obtained from the analyses is shown in Table 3. According to the metric values obtained as a result of model training, the original dataset showed 89.75% overall accuracy, the dataset obtained with the LBP method showed 83.27%, and the dataset obtained with the Grad-CAM method showed 98.22%.

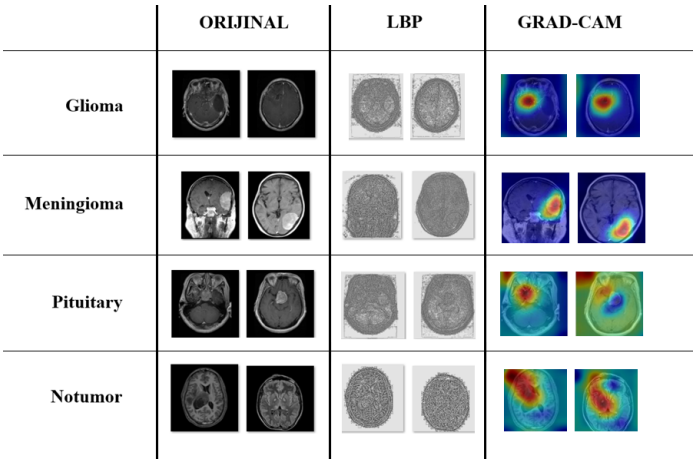


Figure 7 Sample images obtained by applying LBP and Grad-CAM methods to the original dataset.

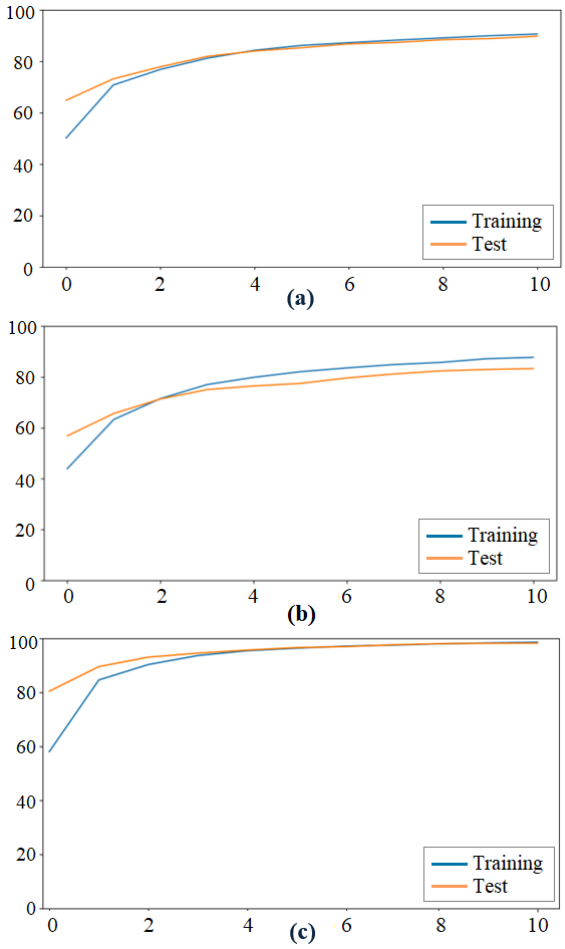


Figure 8 Training-test success graphs of the DeiT3 model; a) Original dataset, b) LBP dataset, c) Grad-CAM dataset.

The third stage involves applying feature fusion to the feature sets extracted from each image set in the previous stage. The goal of this stage is to determine the most efficient combination obtained by applying feature fusion. A feature set of size [image count x 768] was extracted from the final layer of the DeiT3 model. This

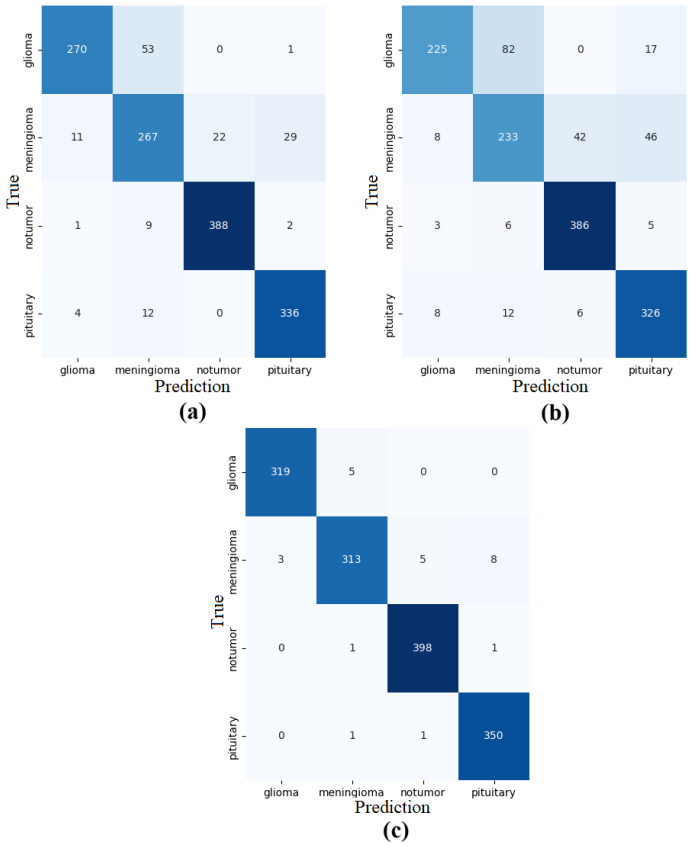


Figure 9 Confusion matrix obtained from training the DeiT3 model; a) Original dataset, b) LBP dataset, c) Grad-CAM dataset.

Table 3 Metric results obtained from the analyses of the DeiT3 model (%)

Dataset	Se	Pre	f-scr	Acc
Original	89.23	89.66	89.34	89.75
LBP	82.34	83.47	82.40	83.27
Grad-CAM	98.13	98.22	98.17	98.22

process was performed individually for all image sets (Original, Grad-CAM, LBP). As a result of feature fusion; [image count x 1536] from the (Grad-CAM & Original) set, [image count x 1536] from the (Original & LBP) set, (Grad-CAM & LBP) set [image count x 1536], (Grad-CAM & Original & LBP) set [image count x 2304]. The SVM method was used to evaluate the performance of the obtained feature sets. At this stage, the training and test ratios were selected in the same proportion as the model training (training data 80%, test data 20%). The confusion matrices obtained from the classification of the feature sets obtained as a result of feature fusion using the SVM method are given in Figure 10. The confusion matrix results are given in Table 4.

Table 5 shows that the (Grad-CAM & LBP) set achieved a general accuracy success rate of 99.42%, while the (Grad-CAM & Original) set achieved a general accuracy success rate of 99.36%. Therefore, it was determined that the (Grad-CAM & LBP) set provides better performance than the (Grad-CAM & Original) set. It was observed that the success of the new sets obtained with the feature

Table 4 Metric results (%) of confusion matrices obtained after feature fusion

Set / Fusion	Class	Se	Pre	F-scr	Acc
GC & LBP	0	97.84	99.69	99.76	
GC & LBP	1	99.69	97.91	98.79	
GC & LBP	2	100	100	100	
GC & LBP	3	99.71	100	99.86	
					99.36
GC & ORJ	0	98.15	99.38	98.76	
GC & ORJ	1	99.39	98.19	98.78	
GC & ORJ	2	100	100	100	
GC & ORJ	3	99.71	99.71	99.71	
GC & ORJ					99.36
LBP & ORJ	0	95.68	99.04	96.91	
LBP & ORJ	1	98.78	94.20	96.32	
LBP & ORJ	2	99.75	99.75	99.75	
LBP & ORJ	3	98.58	99.43	98.97	
LBP & ORJ					98.15
GC & LBP & ORJ	0	97.84	99.37	98.37	
GC & LBP & ORJ	1	99.39	97.90	98.48	
GC & LBP & ORJ	2	99.75	100	99.87	
GC & LBP & ORJ	3	99.71	99.71	99.71	
GC & LBP & ORJ					99.22

Table 5 Metric results (%) of confusion matrices obtained after cross-validation

Set / Fusion	Class	Se	Pre	F-scr	Acc
GC & LBP	0	99.07	99.69	99.38	
GC & LBP	1	99.21	98.85	99.03	
GC & LBP	2	99.80	99.65	99.72	
GC & LBP	3	99.21	99.32	99.26	
					99.42
GC & ORJ	0	99.01	99.63	99.32	
GC & ORJ	1	99.15	98.85	99.00	
GC & ORJ	2	99.85	99.75	99.80	
GC & ORJ	3	99.49	99.32	99.41	
					99.36

fusion approach was higher than that obtained from the DeiT3 model. The proposed approach had a positive impact on overall performance with all steps in the process.

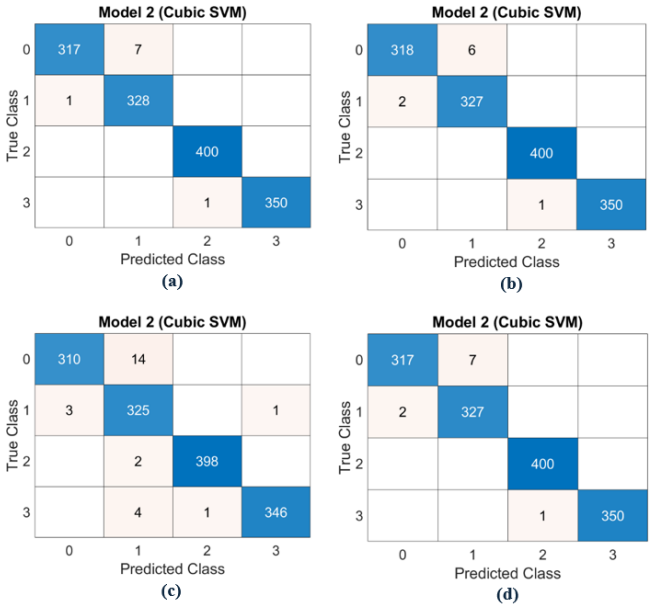


Figure 10 Confusion matrices obtained after feature fusion; a) GC & LBP, b) GC & ORJ, c) LBP & ORJ, d) GC & ORJ & LBP.

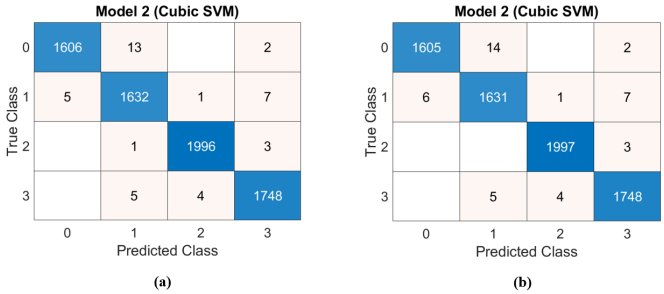


Figure 11 Confusion matrices obtained after cross-validation; a) GC & LBP, b) GC & ORJ

CONCLUSION

In this study, the model's effectiveness was enhanced by using the Grad-CAM and LBP techniques in the preprocessing steps. The regions of interest in the model's decision-making process were identified using the Grad-CAM-based visual interpretability approach. This step made the anatomical structures associated with the tumor more prominent while reducing the impact of the background and clinically insignificant regions. Thus, the deep learning model was directed to more meaningful regions in terms of classification, and the clinical consistency of the learned representations was increased. In addition, tissue features were extracted using the LBP method. LBP is an effective method, particularly for capturing microstructural differences between tumorous and healthy tissue. Thanks to this approach, local tissue variations specific to tumor regions were quantitatively represented and used as a complement to deep learning-based features. These different representations obtained during the preprocessing stage were converted into high-level features via the DeiT3 model. The combination of feature sets obtained from different preprocessing methods (feature fusion) created a richer and more discriminative feature structure compared to representations obtained from a single source. This directly contributed to strengthening the distinction between classes.

The preprocessing steps applied in the study played a critical role in reducing noise, highlighting tumor-related tissue and structural information, and making the model's learning process more targeted. The experimental results obtained demonstrate that these preprocessing strategies improve classification performance and that the proposed approach offers an effective and reliable solution for the brain tumor detection problem. The main reason for choosing the DeiT3 model in this study is that it can offer high generalization performance even without a very large dataset and provides stability during the training process. Although ViT approaches have the ability to effectively learn visual features, they often require a large amount of labeled data and complex training strategies. DeiT3, on the other hand, significantly alleviates these structural constraints through methods that balance the training process and limit the model's tendency to overfit.

The reason for using SVM in the classification stage is based on the method's ability to create effective decision boundaries in high-dimensional feature structures. Particularly when features obtained from deep learning-based models exhibit non-linear distributions, SVM's margin-based optimization approach offers more stable and discriminative classification performance. In addition, SVM has been preferred because it produces more consistent results by reducing the risk of overfitting under limited data conditions. The proposed approach offers several advantages in the analysis of brain tumor MRI images. Combining high-level features derived from the deep learning-based DeiT3 architecture with textural features extracted using the LBP method has created a complementary representation structure that incorporates both global and local information. This multi-feature representation has contributed to strengthening the distinction between classes and supported the improvement of the obtained classification performance. Furthermore, the use of Grad-CAM has enabled the visualization of the regions on which the model focuses during the decision-making process, allowing for a clearer analysis of tumor-related anatomical structures and increasing the clinical interpretability of the method.

However, the proposed method also has limitations. Its multi-stage structure, which includes preprocessing, feature extraction, feature fusion, and classification steps, increases computational cost and requires a more complex workflow compared to end-to-end learning approaches. Furthermore, the method's performance is sensitive to the quality of the preprocessing steps and the selected features, and parameter choices in these stages can influence the results. Finally, while the results are promising, validating the method's generalizability on larger, multi-center datasets is important for increasing its reliability for clinical applications. The proposed approach can be considered as a decision support tool that can assist the clinical diagnosis process through the automatic analysis of brain tumor MRI images. Especially under heavy clinical workload, it can contribute to accelerating the preliminary evaluation phase, allowing specialist physicians to focus on more complex cases. Furthermore, Grad-CAM-based visualizations enhance the clinical interpretability of the findings by making the model's decision-making process more understandable. In this regard, the proposed method can be used in clinical practice as a tool that supports the diagnostic process rather than replacing the physician's final decision.

The experimental results show that the proposed hybrid approach offers an effective method for the accurate and rapid diagnosis of brain tumor disease. The results of the study reveal that new-generation ViT models such as Deit3 are much more effective than traditional diagnostic methods. During the analyses in

Table 6 Comparison of studies using the same dataset

Research	Number of Images	Class	Model/Method	Değer (%)
(Gómez-Guzmán *et al.* 2023)	7.023	4	ResNet50, InceptionV3, InceptionResNetV2, Xception, MobileNetV2 and EfficientNetB0	%97.1
(Bayaral *et al.* 2025)	7.023	4	VGG16, VGG19, ResNet50, MobileNetV2 and SVM / XGBoost	%97.8
This research	7.023	4	DeiT3 model, Grad-CAM and LBP image processing techniques, Cubic SVM method	%99.5

this study, Grad-CAM and LBP image processing techniques were applied to the original dataset. The feature fusion resulted in the best combination set not being the original dataset, but rather the combination of Grad-CAM and LBP, demonstrating that the image processing techniques used provided a meaningful contribution. The use of the DeiT3 model in model training required powerful hardware during training. Therefore, it may be difficult to apply the proposed approach on low-performance systems. A comparison of studies conducted with the same dataset is provided in Table 6.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Almohamade, A., S. Kammoun, and F. Alsolami, 2025 Local binary pattern–cycle generative adversarial network transfer: Transforming image style from day to night. Journal of Imaging **11**: 108.

Aslan, E., 2024 Classification of brain tumors from mr images using the lstm-esa hybrid model. Adıyaman University Journal of Engineering Sciences **11**: 63–81.

Aslan, M., 2022 Deep learning-based automatic brain tumor detection. Fırat University Journal of Engineering Sciences **34**: 399–407.

Attallah, O., 2025 Multi-domain feature incorporation of lightweight convolutional neural networks and handcrafted features for lung and colon cancer diagnosis. Technologies (Basel) **13**: 173.

Aydemir, E., 2022 Speaker recognition and speaker verification by comparing mfcc and lbp methods. Turkish Informatics Foundation Journal of Computer Science and Engineering **15**: 104–109.

Bayaral, S., E. Gül, and D. Avcı, 2025 Classification of brain tumors using artificial intelligence. International Journal of Innovative Engineering Applications **9**: 8–22.

Bhuvaji, S., 2025 Brain tumor classification (mri). Kaggle dataset, Accessed: 10/15/2025.

Çelik, Ö. and B. C. Koç, 2021 Classification of turkish news texts using tf-idf, word2vec, and fasttext vector model methods. Deu Faculty of Engineering Science and Engineering **23**: 121–127.

Cheng, J., 2024 Brain tumor dataset. Figshare dataset, Accessed: 10/15/2025.

Das, D., C. Sarkar, and B. Das, 2025 Real-time detection of meningiomas by image segmentation: A very deep transfer learning convolutional neural network approach. Tomography **11**: 50.

Demirel, C. and E. Soylu, 2024 Comparison of transfer-based deep learning algorithms for tumor detection in mri data. Black Sea Journal of Science **14**: 1322–1339.

Devi, A. G. *et al.*, 2023 An improved chi2 feature selection based on a two-stage prediction of comorbid cancer patient survivability. Revue d'Intelligence Artificielle **37**: 83–92.

Ennab, M. and H. Mcheick, 2025 Advancing ai interpretability in medical imaging: A comparative analysis of pixel-level interpretability and grad-cam models. Machine Learning and Knowledge Extraction **7**: 12.

Erçelik, Ç. and K. Hanbay, 2023a Classification of brain tumors using gaussian filtering and the resnet50 model. Computer Science .

Erçelik, Ç. and K. Hanbay, 2023b Effects of histogram equalization method on some deep learning models in brain tumor classification. Computer Science .

Gómez-Guzmán, M. A. *et al.*, 2023 Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks. Electronics (Basel) **12**: 955.

Gür, Y. E., M. Toğaçar, and B. Solak, 2025 Integration of cnn models and machine learning methods in credit score classification: 2d image transformation and feature extraction. Computational Economics **65**: 2991–3035.

Haldar, B., H. Joardar, A. K. Mondal, N. H. Alrasheedi, R. Khan, *et al.*, 2025 Machine-learning-driven analysis of wear loss and frictional behavior in magnesium hybrid composites. Crystals (Basel) **15**: 452.

Hamada, A., 2020 Br35h: Brain tumor detection. Kaggle dataset, Accessed: 10/15/2025.

Kuştaşı, E. and M. Yağanoğlu, 2024 Classification of variable stars using deep learning and transfer learning methods. Batman University Journal of Life Sciences **14**: 81–97.

Livieris, I. E., E. Pintelas, N. Kiriakidou, and P. Pintelas, 2023 Explainable image similarity: Integrating siamese networks and grad-cam. Journal of Imaging **9**: 224.

Nickparvar, M., 2021 Brain tumor mri dataset. Kaggle dataset, Accessed: 10/15/2025.

Rahman, S., S. N. F. Mursal, M. A. Latif, Z. Mushtaq, M. Irfan,

et al., 2023 Enhancing network intrusion detection using effective stacking of ensemble classifiers with multi-pronged feature selection technique. In *2023 2nd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, pp. 1–6, IEEE.

Senjoba, L., H. Ikeda, H. Toriya, T. Adachi, and Y. Kawamura, 2024 Enhancing interpretability in drill bit wear analysis through explainable artificial intelligence: A grad-cam approach. Applied Sciences **14**: 3621.

Serttaş, S. and E. Deniz, 2023 Disease detection in bean leaves using deep learning. Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering **65**: 115–129.

Sevinc, A., M. Ucan, and B. Kaya, 2025 A distillation approach to transformer-based medical image classification with limited data. Diagnostics **15**: 929.

Taşdemir, B. and N. Barışçı, 2024 Brain tumor segmentation with deep learning. Journal of Information Technologies **17**: 159–174.

Toğaçar, M., 2025 Time-frequency imaging for epilepsy seizure detection: Feature fusion with transformer model. Journal of Engineering Sciences and Research **7**: 93–102.

Tubog, M. V., K. Emsellem, and S. Bouissou, 2025 Detection of agricultural terraces platforms using machine learning from orthophotos and lidar-based digital terrain model: A case study in roya valley of southeast france. Land (Basel) **14**: 962.

Yang, P., R. Zhang, C. Hu, and B. Guo, 2025 Gmdnet: Grouped encoder-mixer-decoder architecture based on the role of modalities for brain tumor mri image segmentation. Electronics (Basel) **14**: 1658.

Yenikaya, M. A. and O. Oktaysoy, 2023 The use of artificial intelligence applications in the healthcare sector: Preliminary diagnosis using deep learning methods. Sakarya University Business Institute Journal **5**: 127–131.

Yılmaz, S., 2023 Design of a decision support system based on the yolov7 algorithm for brain tumor diagnosis. Kocaeli University Journal of Science **6**: 47–56.

How to cite this article: Kahraman, S., and Toğaçar, M. Classification of Brain MRI Images using Deep Learning: The DeiT3 Model and the Use of Feature Fusion Methods. *Computers and Electronics in Medicine*, 3(1), 77-85, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.86-93
https://doi.org/10.69882/adba.cem.20260110

TinyML-Based Machine Learning System for Multi-Class Ear Condition Classification

Serkan Dişlitaş *,1
*Hitit University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, 19030, Çorum, Türkiye.

ABSTRACT This study presents a TinyML-based machine learning system for multi-class ear condition classification on an edge device, designed to process images captured from a digital otoscope camera. Utilizing a dataset comprising five categories, Normal, Acute Otitis Media (AOM), Cerumen Impaction (CI), Chronic Otitis Media (COM), and Myringosclerosis (MYS), the proposed system performs near real-time classification directly on a resource-constrained microcontroller. The model was developed and optimized using the Edge Impulse platform and deployed as a quantized TinyML library. The system architecture incorporates lightweight convolutional neural networks (CNNs) and the EON™ Compiler to ensure efficient memory usage while maintaining high diagnostic performance. Experimental results demonstrate a validation accuracy of 97.5% and a testing accuracy of 96.31%, with a peak RAM footprint of only 240.3K and an inferencing latency of 1482 ms. These findings highlight the potential of TinyML for portable, low-power medical applications, providing a foundation for privacy-preserving, GDPR-compliant, on-device diagnostics in auditory healthcare without reliance on cloud infrastructure.

KEYWORDS
Ear conditions
TinyML
Multi-class classification
Edge impulse
Otoscope camera

INTRODUCTION

Ear diseases represent a significant clinical challenge due to their high prevalence worldwide and the complexity of achieving accurate diagnosis through visual examination. Common otoscopic findings such as Acute Otitis Media (AOM), Chronic Otitis Media (COM), Cerumen Impaction (CI), and Myringosclerosis (MYS) affect diverse populations and may lead to pain, hearing loss, and long-term complications if not diagnosed and treated promptly. Traditional clinical diagnosis of these ear conditions primarily relies on otoscopic examination performed by specialists, which involves visual inspection of the tympanic membrane and ear canal. Although this method is widely adopted in clinical practice, it is inherently subjective and highly dependent on clinician expertise, resulting in variability in diagnostic accuracy, particularly in primary care settings where access to otolaryngology specialists may be limited (Al-Rahim Habib *et al.* 2022; Livingstone and Chau 2020).

The limitations of traditional diagnostic procedures have motivated the exploration of advanced computational methods. In particular, artificial intelligence (AI) and machine learning (ML) approaches have shown promise in automating image based diagnosis of ear disorders. Systematic reviews indicate that AI algorithms, especially convolutional neural networks (CNNs), can achieve high classification accuracy for otoscopic images, often surpassing non expert human performance. For example, AI-based methods have reported classification accuracies of up to 97.6% in multiclass tasks distinguishing normal conditions, acute otitis media (AOM), and otitis media with effusion using otoscopic images (Tatlı 2025; Song *et al.* 2022). Recent studies have begun exploring transformer-inspired feature representations for otoscopic image analysis, aiming to enhance global contextual understanding and feature extraction beyond conventional convolutional architectures (Demircan *et al.* 2025).

However, these AI-driven systems often rely on deep learning models executed on high-performance hardware or cloud services, which introduces latency, dependency on network connectivity, and privacy concerns, limiting their practicality in resource-constrained or portable environments (Diez *et al.* 2024). Furthermore, while high-performance models like YOLOv10 are emerging for real-time mobile endoscopy, they still demand substantial computational overhead compared to ultra-low-power edge solutions (Wang *et al.* 2024).

To overcome these limitations, Tiny Machine Learning (TinyML) has emerged as a key technology for low-power, on-device AI applications on embedded and edge devices. TinyML enables real-time inference directly on microcontrollers and other embedded hardware, reducing latency, preserving privacy, and operating independently from external servers (Heydari and Mahmoud 2025; Schizas *et al.* 2022). Platforms such as Edge Impulse facilitate the design, training, optimization, and deployment of TinyML models for embedded tasks, including image classification on resource-constrained devices. Recent studies and platform evaluations indicate that tools such as the EON Tuner allow these models to maintain clinically relevant accuracy while fitting within the strict memory limits of Cortex-M microcontrollers (Hymel *et al.* 2022).

Manuscript received: 17 November 2025,
Revised: 14 January 2026,
Accepted: 20 January 2026.

[1]serkandislitas@hitit.edu.tr (**Corresponding author**)

Despite these advances, the application of TinyML for multi-class ear condition classification remains underexplored. While prior studies demonstrate the potential of machine learning for diagnosing ear disorders using deep learning, most focus on computationally intensive models that require cloud or Graphics Processing Unit (GPU) resources (Tatlı 2025; Song *et al.* 2022; Bingol 2022). Therefore, there is a clear need for research combining lightweight ML models, edge computing, and embedded platforms for real-time, on-device classification. In this study, we propose a TinyML-based machine learning system for multi-class ear condition classification deployable on edge devices. Using a publicly available dataset covering five categories, Normal, AOM, CI, COM, and MYS this work demonstrates the design, training, optimization, and deployment of TinyML models with Edge Impulse, offering portable, low-power auditory diagnostics for clinical support.

RELATED WORKS

Automated diagnosis of ear diseases using machine learning has rapidly expanded. Deep learning approaches, especially CNN-based models, have demonstrated high performance for multi-class otoscopic and tympanic membrane image classification, often surpassing non-expert human performance (Tatlı 2025; Song *et al.* 2022; Bingol 2022). Advanced CNN pipelines with hyperparameter optimization, data augmentation, and ensemble methods have achieved accuracies above 98% across multiple ear disease categories (Bingol 2022; Mihigo *et al.* 2022). TinyML applications have extended this capability to embedded edge devices. Visual TinyML applications using Edge Impulse have successfully employed compact cameras, such as otoscope modules, for real-time multi-class classification tasks. For instance, hand posture recognition, finger number detection, and tomato leaf disease identification illustrate that highly constrained devices can perform accurate classification (Heydari and Mahmoud 2025; Kwon 2023). These studies demonstrate the feasibility of deploying TinyML models for real-time, edge-based decision-making.

In healthcare, TinyML has been applied to both audio and visual modalities. Cough detection systems and on-device speech recognition employ lightweight feature extraction and optimized neural networks trained on Edge Impulse, achieving real-time, low-power inference without cloud dependency (Rana *et al.* 2022; Kwon 2023). Specifically in otolaryngology, studies on otoscopic and tympanic membrane images have explored anomaly detection and disease classification using both deep learning and TinyML approaches on otoscopic and endoscopic images (Tatlı 2025; Song *et al.* 2022; Bingol 2022). To address the limitations of static models, adaptive quantization strategies are now being developed specifically for medical wearables to maintain accuracy under extreme energy constraints (Xie and Fang 2025). Beyond medical applications, TinyML has been used for IoT and time series tasks, including predictive maintenance and environmental monitoring. Lightweight network architectures such as TinyLSTM and TinyModel have been deployed on embedded devices via Edge Impulse, demonstrating real-time inference and energy efficiency for sequential data (Mihigo *et al.* 2022; Cioflan *et al.* 2025). These studies highlight TinyML's domain-agnostic capabilities, supporting visual, auditory, and time series data on resource-constrained devices.

Despite these advances, several challenges remain in the deployment of TinyML systems. Existing literature identifies model optimization, memory management, hardware compatibility, benchmarking, and the lack of standardization as key areas requiring further improvement (Heydari and Mahmoud 2025; Schizas *et al.* 2022). In real-time medical applications, maintaining high accuracy, reliability, and low inference latency under strict energy constraints is particularly critical (Bingol 2022). Furthermore, the integration of Explainable Artificial Intelligence (XAI), including both visual explanation techniques and medical-oriented interpretability frameworks, is increasingly recognized as essential for enhancing clinical interpretability and trust in automated otoscopic diagnostic systems (Rehman *et al.* 2025; Özdilli *et al.* 2025; Tjoa and Guan 2020).

In summary, existing literature confirms that TinyML is effective for edge-based multi-class classification and anomaly detection, particularly in healthcare. However, challenges related to data diversity and model optimization remain in medical visual classification. Motivated by these gaps, the present study focuses on multi-class classification of ear conditions using TinyML, leveraging Edge Impulse optimized models deployed on microcontroller-based edge devices, extending existing visual TinyML applications to low-power medical classification on edge devices.

MATERIAL AND METHODS

The methodology of this study focuses on the design, training, and deployment of a TinyML-based machine learning model for multi-class classification of ear conditions, leveraging the Edge Impulse platform. This approach integrates modern machine learning techniques with embedded systems, enabling real-time, on-device inference on an edge device.

Dataset

In this study, a publicly available otoscopic image dataset obtained from the Kaggle (Uci Machine Learning Repository 2025) platform was used for training and evaluating the proposed TinyML-based inner ear condition classification system. The dataset consists of labeled otoscopic images representing five clinically relevant classes: Normal, AOM, CI, COM, and MYS. These categories were selected to cover a spectrum of otoscopic findings commonly encountered in clinical practice. The dataset includes color images acquired under varying illumination conditions and viewpoints, reflecting real-world variability in otoscopic examinations. As the dataset is publicly accessible and anonymized, no ethical approval or patient consent was required. Prior to model training, the dataset was uploaded to the Edge Impulse platform, where it was automatically partitioned into training and test subsets using an 80/20 split. This partitioning strategy was specifically chosen because the dataset maintains a balanced distribution across all five categories, ensuring that the test results provide a statistically significant representation of the model's generalizability. This randomized selection process ensures an unbiased evaluation of model performance and minimizes the risk of overfitting by validating the system on previously unseen data that reflects the overall composition of the dataset.

The dataset is balanced across the five categories, with approximately 600 images per class. In total, the dataset consists of 2978 images, of which 2391 were used for training and 597 for testing. The balanced distribution reduces the risk of biased learning and enhances classification performance. The detailed distribution of images across classes and data splits is summarized in Table 1. Representative sample images from each ear condition class are illustrated in Figure 1, demonstrating the visual characteristics and intra-class variability present in the dataset.

Table 1 Distribution of otoscopic images in the Kaggle dataset for inner ear condition classification (Uci Machine Learning Repository 2025)

Class Label	Inner Ear Condition	Training	Test	Total
C1 / Normal	Normal	480	120	600
C2 / AOM	Acute Otitis Media (AOM)	463	115	578
C3 / CI	Cerumen Impaction (CI)	480	120	600
C4 / COM	Chronic Otitis Media (COM)	479	121	600
C5 / MYS	Myringosclerosis (MYS)	479	121	600
	Grand Total	2391	597	2978

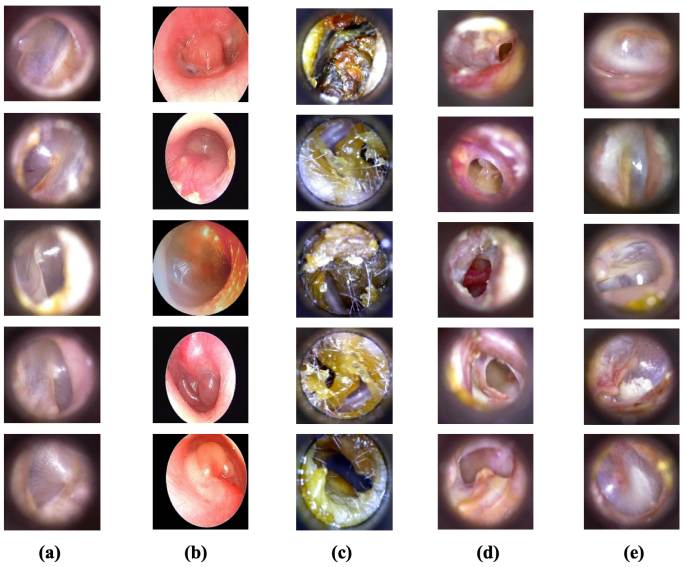


Figure 1 Representative otoscopic images from the dataset illustrating the five ear condition classes: (a) Normal, (b) AOM, (c) CI, (d) COM, and (e) MYS (Uci Machine Learning Repository 2025).

TinyML model design using Edge Impulse Platform

Edge Impulse is a comprehensive development platform specifically engineered for TinyML, providing a unified end-to-end workflow to design, optimize, and deploy machine learning models on resource-constrained hardware (Edge Impulse 2025; Warden and Situnayake 2019), with the complete cycle of model training and edge device deployment depicted in Figure 2 (Rust 2020). This systematic approach allows developers to move seamlessly from raw data acquisition to real-time inference on embedded devices. The platform integrates advanced digital signal processing (DSP) for automated feature extraction with support for optimized neural network architectures, such as Convolutional Neural Networks (CNNs) (Han *et al.* 2015). By utilizing sophisticated optimization tools like quantization and the EON Compiler, Edge Impulse minimizes the memory footprint and latency of models, enabling efficient on-device inference (Moreau 2024). This framework ensures high performance and data privacy, making it ideal for real-time medical diagnostic systems on edge devices (Hizem *et al.* 2025).

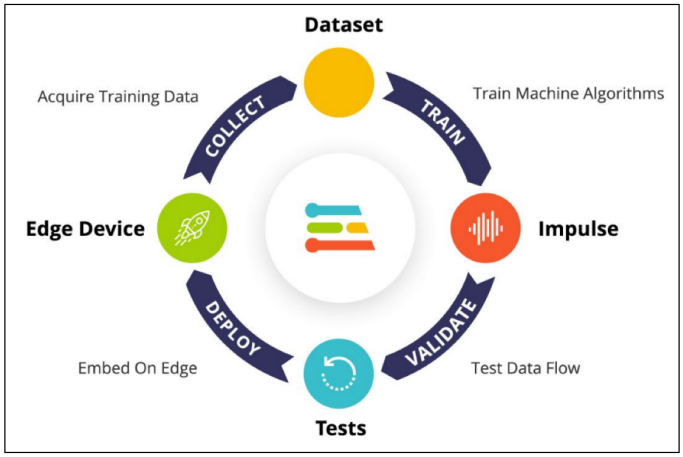


Figure 2 The end-to-end TinyML development workflow in Edge Impulse.

DESIGNED SYSTEM

Proposed Model

The systematic development of the classification system, as illustrated in the flowchart in Figure 3, follows a structured pipeline within the platform. The process begins with the data acquisition stage, which utilizes a publicly available Kaggle dataset comprising 2,978 otoscopic images categorized into five clinically relevant classes: Normal, AOM, CI, COM, and MYS. In the data preprocessing stage, all images are resized to a standardized resolution of 96×96 pixels and normalized using Edge Impulse's built-in DSP blocks. This step ensures computational efficiency while preserving discriminative visual patterns. Additionally, to improve model robustness and generalization, data augmentation techniques such as rotation, flipping, and brightness adjustment are applied during this phase.

Following the dataset upload and train-test splitting, the feature extraction and neural network design stage employs a CNN-based architecture optimized for microcontroller constraints. During model training and optimization, hyperparameters are tuned on Edge Impulse servers, and critical quantization techniques are applied to minimize the memory footprint. After rigorous validation and testing via confusion matrices, precision, and recall, the optimized model is exported as a standalone C++ library compatible with the 32-bit microcontroller. The final stages involve the development of edge device firmware and its deployment, enabling the system to execute real-time inference and classification directly on

the microcontroller.

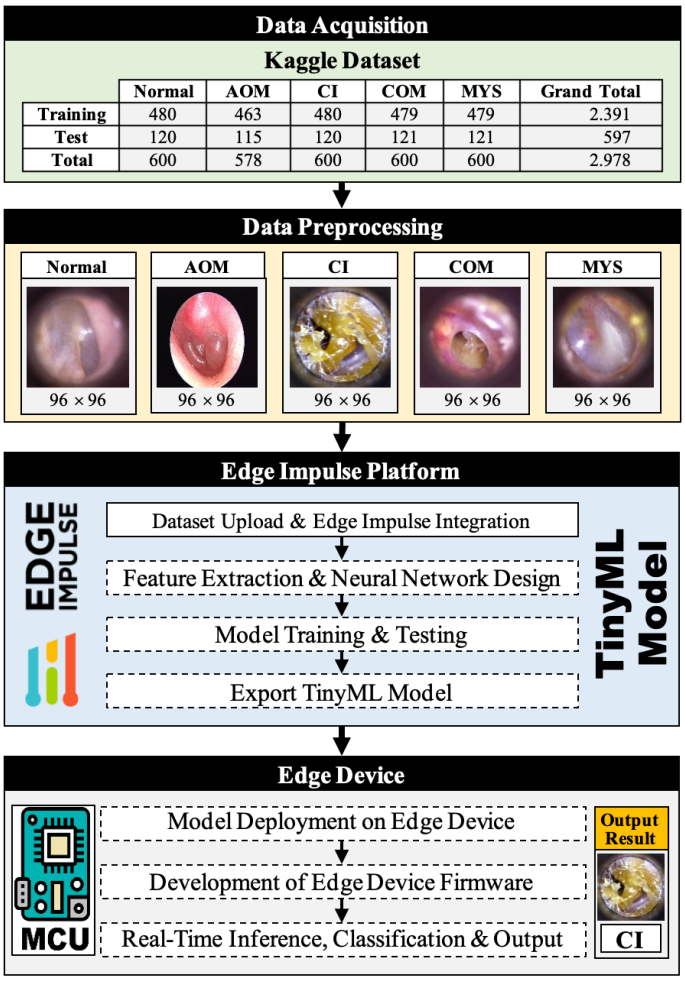


Figure 3 Flowchart of the proposed TinyML model development for ear condition classification on the Edge Impulse platform.

The structured pipeline for the diagnostic system is depicted in Figure 4, showing the transition from raw otoscopic image input to the final classification of five distinct ear conditions. The architecture begins with an image data block where raw otoscopic images are processed with a predefined 96x96 pixel resolution, a resizing step crucial for balancing the extraction of clinical features with the memory constraints of the 32-bit microcontroller hardware. This is followed by an image processing block that transforms the data into a feature set optimized for the subsequent learning phase. Finally, a neural network classifier is utilized to categorize these processed features into five diagnostic classes: AOM, CI, COM, MYS, and Normal. By utilizing this modular Impulse Design, the system ensures a streamlined data flow specifically optimized for accurate, real-time diagnostic performance on edge devices.

The core of this project is the deployment of a robust diagnostic system on a resource-constrained microcontroller. As shown in Table 2, the training settings were specifically tuned for the 32-bit microcontroller. By employing a 0.0005 learning rate and INT8 quantization, the system provides real-time, on-device classification of ear conditions such as AOM, COM, CI, and MYS. This optimization ensures that the final model maintains a minimal memory footprint suitable for edge deployment while delivering high diagnostic accuracy for clinical decision support.

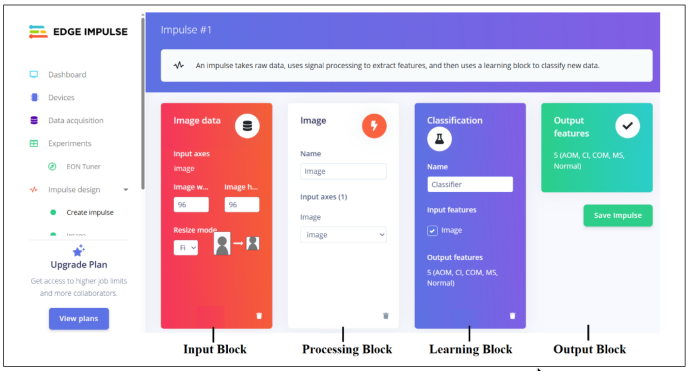


Figure 4 The impulse design and processing pipeline for otoscopic ear disease classification.

Table 2 Neural network training configuration and optimization settings for the TinyML-based ear disease classification project

Training Setting	Value/Status
Number of epochs	50
Learning Rate	0.0005
Training Processor	CPU
Validation Set Size	20%
Batch Size	32
Profile int8 Model	Active

The neural network architecture designed for the classification task is illustrated in Figure 5, featuring a Convolutional Neural Network (CNN) structure optimized for deployment on the 32-bit microcontroller. The model begins with an input layer processing 27648 features, derived from the 96 x 96 pixel input, which is then organized by a reshape layer. Feature extraction is performed through two successive stages of 2D convolutional and pooling layers, utilizing 32 filters in the first stage and 64 filters in the second, both with a 3 x 3 kernel size. To improve model generalization and prevent overfitting, two dropout layers with a rate of 0.25 are strategically integrated after the convolutional blocks. Finally, a flatten layer converts the multidimensional feature maps into a 1D vector to enable the final classification of ear conditions.

Designed Embedded System

The block diagram of the proposed TinyML-based ear condition classification system is illustrated in Figure 6. The architecture follows a sequential and modular pipeline, ensuring efficient real-time diagnostics on resource-constrained hardware.

The process begins with the Input (Camera Module) block, where an otoscope camera, supported by integrated LED illumination, captures images of the ear canal and tympanic membrane. These captured otoscopic images are transmitted directly to the Embedded Processing Unit (Edge Device / microcontroller), which serves as the central hub for local data processing. Within this unit, the Image Preprocessing stage executes resizing, noise reduction, and normalization to optimize the visual data for neural network analysis. Subsequently, the processed data is passed to the TinyML Inference Engine, utilizing an optimized model trained via Edge

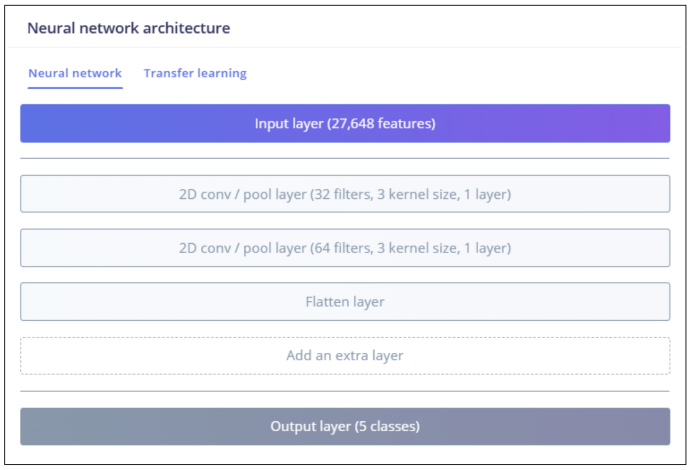


Figure 5 The neural network architecture of the TinyML model for otoscopic image classification.

Impulse.

The engine performs on-device classification, and the resulting Classified Output is forwarded to the final Output (Interface) block. This interface displays the identified class such as Normal, AOM, CI, COM, or MYS and provides diagnostic decision support along with confidence scores. By executing the entire pipeline locally, the system eliminates the need for cloud-based computation, thereby ensuring low latency and maintaining strict data privacy for point-of-care medical applications.

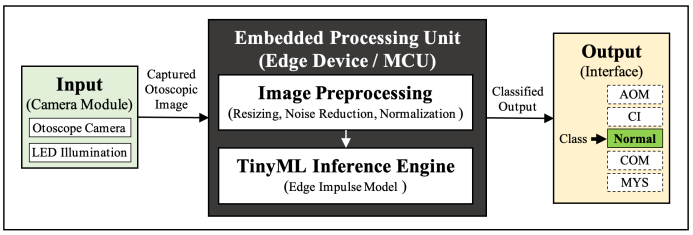


Figure 6 The block diagram of the proposed TinyML-based ear condition classification system.

As illustrated in Figure 7, the proposed TinyML-based inner ear condition classification system operates using a sequential and loop-based software flow. The process begins with system initialization, during which the embedded controller, camera module, and the pre-trained TinyML model generated using the Edge Impulse platform are loaded and configured. At this stage, camera illumination parameters are also initialized to ensure consistent lighting conditions during image acquisition.

Following initialization, the system captures an otoscopic image using the camera module under controlled illumination. The acquired image is then forwarded to the Edge Impulse image processing pipeline, where essential pre-processing operations such as resizing and normalization are applied. These steps prepare the image for efficient inference on resource-constrained embedded hardware.

Subsequently, the pre-processed image is passed to the TinyML inference engine, which classifies the image into one of the predefined ear condition classes: Normal, AOM, CI, COM, and MYS. The classification result is then presented to the user via an output interface such as LEDs, a display module, or serial communication. After a predefined waiting period, the system either captures the next image or returns to an idle state, enabling continuous or on-demand operation of the diagnostic system.

Overall, the proposed software workflow enables a low-power, real-time, and robust TinyML-based implementation on a microcontroller, supporting portable and efficient classification of ear conditions.

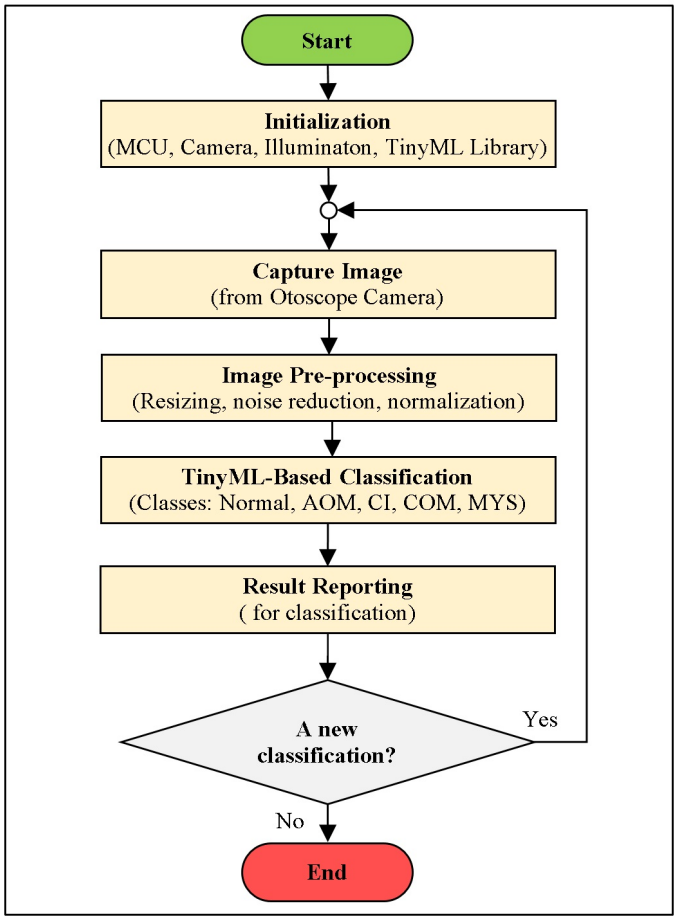


Figure 7 Software flowchart of the proposed TinyML-based ear condition classification system incorporating controlled camera illumination and Edge Impulse-based image processing.

RESULTS AND DISCUSSION

The performance of the proposed TinyML system for ear condition classification was evaluated using a comprehensive set of metrics, including accuracy, sensitivity, precision, and F1-score. These metrics, calculated from the validation set's confusion matrix, serve as a vital framework for assessing the model's diagnostic reliability in medical imaging. Accuracy measures the overall proportion of correct classifications, while sensitivity determines the model's ability to correctly identify true positive cases. Precision assesses the reliability of positive predictions, and the F1-score provides a balanced evaluation by calculating the harmonic mean of precision and sensitivity. The mathematical definitions for these metrics are presented in Equations (1–4), forming the basis for the 97.5% accuracy achieved in this study. In these equations, TP (True Positives) and TN (True Negatives) represent correctly classified instances, while FP (False Positives) and FN (False Negatives) denote the

Table 3 Performance metrics of the TinyML-based ear disease classification system

Classes	Accuracy (%)	Precision	Sensitivity	F1-Score
AOM	97.5	1.00	1.00	1.00
CI	97.5	1.00	1.00	1.00
COM	97.5	0.97	0.98	0.98
MYS	97.5	0.96	0.93	0.94
Normal	97.5	0.95	0.97	0.96
Average	97.5	0.98	0.98	0.98

misclassified instances (Alaca and Akmeşe 2025; Fawcett 2006).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{F-score} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

The diagnostic performance of the proposed TinyML system is quantitatively summarized in Table 3, which highlights the model's high classification efficacy across five distinct ear conditions. According to the data, the system achieved a robust overall accuracy of 97.5% with a consistent total precision, sensitivity, and F1-score of 0.98. Specifically, the AOM and CI categories demonstrated perfect classification performance with F1-scores of 1.00. While the MYS class exhibited the lowest relative sensitivity at 0.93 due to minor inter-class confusion with the Normal category, the overall high metrics across all labels validate the reliability of the optimized neural network for near real-time, on-device medical diagnostics as defined by Equations (1–4).

As illustrated in Figure 8, high values along the diagonal of the confusion matrix indicate strong classification accuracy. The on-device performance metrics, shown in Figure 9, demonstrate the model's efficiency on 32-bit microcontrollers. Utilizing the EON™ Compiler, the system achieves an inference time of 1482 ms, with a peak RAM usage of 240.3K and a flash usage of 243.1K.

Figure 8 Confusion matrix calculated by Edge Impulse for inner ear disease classification.

The spatial distribution is visualized in the Feature Explorer (Figure 10). The clear grouping of data points indicates that the feature extraction layers effectively identified unique patterns. Furthermore, the model achieved a perfect AUC of 1.00 as shown in Table 4, demonstrating high consistency across all categories. Finally, the model's testing phase on a separate dataset yielded an accuracy of 96.31% (Figure 11), confirming a robust and reliable classification system for otoscopic ear disease detection.

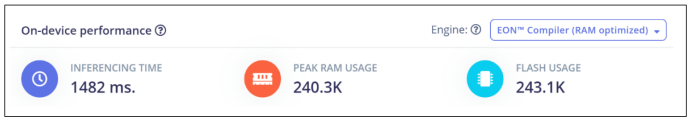


Figure 9 On-device performance metrics for inner ear disease classification on 32-bit mcu, including inference time, peak RAM usage, and flash memory usage.

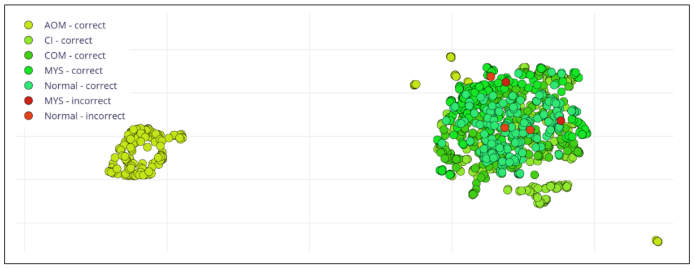


Figure 10 Feature explorer visualization of the otoscopic image dataset, illustrating the spatial separation of disease classes (AOM, CI, COM, MYS, and Normal).

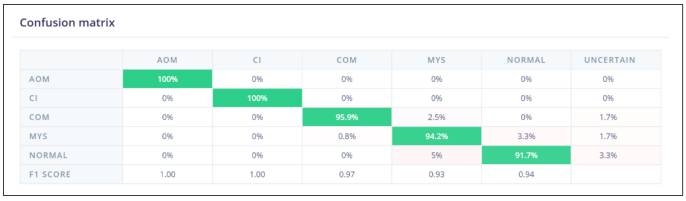


Figure 11 Model testing results showing the confusion matrix and performance metrics for five-class otoscopic disease classification.

Table 4 Validation metrics of the TinyML-based ear disease classification system

Metrics	Value
Area under ROC Curve	1.00
Weighted average Precision	0.97
Weighted average Recall	0.97
Weighted average F1 score	0.97

The proposed system enhances patient privacy by performing all inferences locally on the microcontroller, ensuring compliance with data protection regulations (e.g., GDPR) as no raw data is transmitted to the cloud. Designed as a rapid point-of-care screening tool, the device provides immediate feedback in resource-limited settings. While this proof-of-concept is promising, future transition to clinical use will require formal certifications

(e.g., FDA or CE MDR) and real-world clinical trials to validate its diagnostic efficacy and workflow integration.

CONCLUSION

This study successfully demonstrated the implementation and deployment of a TinyML-based diagnostic system for the multi-class classification of ear diseases on resource-constrained 32-bit microcontrollers. By leveraging the Edge Impulse platform and lightweight convolutional neural networks, the proposed system achieved a robust validation accuracy of 97.5% with a minimal loss of 0.15. Final testing on a separate dataset further confirmed the model's reliability, yielding a performance accuracy of 96.31%. The model exhibited exceptional precision in identifying AOM and CI categories with perfect F1-scores of 1.00, proving its effectiveness in distinguishing high-priority pathological conditions from normal otoscopic findings.

The technical evaluation highlights the system's high efficiency for edge computing. Utilizing the EON™ Compiler, the architecture was optimized to operate within the physical memory limits of embedded hardware, maintaining a peak RAM usage of 240.3K and a flash footprint of 243.1K. With an inference time of 1482 ms, the system enables near-instantaneous, on-device diagnostics without the latency, network dependency, or privacy risks associated with cloud-based AI solutions. While minor misclassification patterns were noted specifically where 5.1% of MYS cases were identified as Normal the overall aggregate F1-score of 0.98 validates the system as a robust tool for clinical support in real-world scenarios.

In conclusion, this research establishes a scalable and low-power framework for auditory healthcare diagnostics. By enabling automated and objective ear examinations directly on portable devices, this TinyML approach offers a viable solution for primary care settings and regions with limited access to otolaryngology specialists. Future work will focus on expanding the dataset to include a wider variety of pathological stages and further optimizing the model to reduce inference time on even lower-power 32-bit hardware. This study bridges the gap between complex deep learning models and practical, on-device intelligence, marking a significant step forward in the democratization of advanced medical diagnostic tools.

Ethical standard

The author has no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The author declares that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Al-Rahim Habib, A., M. Faruque, and A. M. Islam, 2022 Artificial intelligence to classify ear disease from otoscopy: A systematic review and meta-analysis. Clinical Otolaryngology **47**: 1354–1365.

Alaca, Y. and Ö. F. Akmeşe, 2025 Pancreatic tumor detection from ct images converted to graphs using whale optimization and classification algorithms with transfer learning. International Journal of Imaging Systems and Technology **35**: e70040.

Bingol, H., 2022 Classification of ome with eardrum otoendoscopic images using hybrid-based deep models, nca, and gaussian method. Traitement du Signal **39**.

Cioflan, C., J. Fonseca, X. Wang, and L. Benini, 2025 Nanohydra: Energy-efficient time-series classification at the edge. arXiv preprint arXiv:2510.20038 .

Demircan, F., M. Ekinci, Z. Cömert, and E. Gedikli, 2025 Enhanced classification of ear disease images using metaheuristic feature selection. Sakarya University Journal of Computer and Information Sciences **8**: 58–75.

Diez, P. L., J. V. Sundgaard, J. Margeta, K. Diab, F. Patou, *et al.*, 2024 Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical ct images. Computerized Medical Imaging and Graphics **113**: 102343.

Edge Impulse, 2025 Edge impulse documentation. Accessed: Sep. 01, 2025.

Fawcett, T., 2006 An introduction to roc analysis. Pattern Recognition Letters **27**: 861–874.

Han, S., H. Mao, and W. J. Dally, 2015 Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 .

Heydari, S. and Q. H. Mahmoud, 2025 Tiny machine learning and on-device inference: A survey of applications, challenges, and future directions. Sensors **25**: 3191.

Hizem, M., L. Bousbia, Y. Ben Dhiab, M. O. E. Aoueileyine, and R. Bouallegue, 2025 Reliable ecg anomaly detection on edge devices for internet of medical things applications. Sensors **25**: 2496.

Hymel, S., C. Banbury, D. Situnayake, A. Elium, C. Ward, *et al.*, 2022 Edge impulse: An mlops platform for tiny machine learning. arXiv preprint arXiv:2212.03332 .

Kwon, C. K., 2023 Development of embedded machine learning finger number recognition application using edge impulse platform. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, pp. 2697–2699, IEEE.

Livingstone, D. and J. Chau, 2020 Otoscopic diagnosis using computer vision: An automated machine learning approach. The Laryngoscope **130**: 1408–1413.

Mihigo, I. N., M. Zennaro, A. Uwitonze, J. Rwigema, and M. Rovai, 2022 On-device iot-based predictive maintenance analytics model: Comparing tinylstm and tinymodel from edge impulse. Sensors **22**: 5174.

Moreau, L., 2024 Introducing: Eon compiler ram-optimized. Accessed: Sep. 01, 2025.

Özdilli, Ö., S. Şevik, Y. Alaca, and Y. Uzunoğlu, 2025 Optimal design of flat plate fin heat sinks using a computational fluid dynamics (cfd) and deep learning (dl)-based ensemble approach with explainable artificial intelligence (xai) integration. Applied Thermal Engineering p. 127547.

Rana, A., Y. Dhiman, and R. Anand, 2022 Cough detection system using tinyml. In *2022 International Conference on Computing, Communication and Power Technology (IC3P)*, pp. 119–122, IEEE.

Rehman, Z. U., M. F. A. Fauzi, F. N. I. Lokman, M. Touhami, and L. Saim, 2025 Efficient and interpretable otoscopic image classification via distilled cnn with adaptive channel attention. IEEE Access .

Rust, E., 2020 Getting started with edge impulse. Accessed: Dec. 03, 2025.

Schizas, N., A. Karras, C. Karras, and S. Sioutas, 2022 Tinyml for

ultra-low power ai and large scale iot deployments: A systematic review. Future Internet **14**: 363.
Song, D., I. S. Song, J. Kim, J. Choi, and Y. Lee, 2022 Semantic decomposition and anomaly detection of tympanic membrane endoscopic images. Applied Sciences **12**: 11677.
Tatlı, Y., 2025 Ear pathologies using deep learning on otoscopic images. Uluslararası Sürdürülebilir Mühendislik ve Teknoloji Dergisi **9**: 51–57.
Tjoa, E. and C. Guan, 2020 A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems **32**: 4793–4813.
Uci Machine Learning Repository, 2025 Otoscopic image dataset. Kaggle, Accessed: Dec. 03, 2025.
Wang, A., H. Chen, L. Liu, K. Chen, Z. Lin, *et al.*, 2024 Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems **37**: 107984–108011.
Warden, P. and D. Situnayake, 2019 *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. O'Reilly Media.
Xie, Y. and Q. Fang, 2025 An energy-aware generative ai edge inference framework for low-power iot devices. Electronics **14**: 4086.

How to cite this article: Dişlitaş, S. TinyML-Based Machine Learning System for Multi-Class Ear Condition Classification. *Computers and Electronics in Medicine*, 3(1), 86-93, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.94-98
https://doi.org/10.69882/adba.cem.20260111

Classification of Breast Cancer with Breast X-Ray Images via Convolution Neural Networks, Vision Transformers and AlexNet

Bora Basaran [*,1] **and Ali Burak Oncul** [α,2]

*Applied Data Science, TED University, 06420, Ankara, Türkiye, αDepartment of Computer Engineering, Faculty of Engineering and Architecture, Kastamonu University, 37150, Kastamonu, Türkiye.

ABSTRACT Breast cancer is one of the most dangerous types of cancer, affecting many people long-term and leading to death. For this reason, it has become a frequently studied and emphasized topic in the medical field. Furthermore, significant advances in computer science have attracted the attention of the medical world, and computer science has also been incorporated into this challenging disease process to address it. The rapid development of artificial intelligence in recent years has led to a rapid increase in research on breast cancer. Numerous AI-based models have been developed to prevent human errors and to assist and support researchers in decision-making. In this study, one of these models was developed, and three different deep learning (DL) models were proposed to classify breast cancer as breast cancer-negative and breast cancer-positive. The study was adapted for computer vision (CV) using a Kaggle dataset called Breast Cancer, consisting of 3,383 breast tumor mammography images; the labels are 0 and 1, respectively, and the image dimensions are 640 x 640 pixels. In this study, three models were trained to classify breast cancer images: a Convolutional Neural Network (CNN), VisionTransformers (ViT), and AlexNet, trained with 45, 75, and 50 epochs, respectively. HuggingFace Space was used with these three models to classify breast cancer. The HuggingFace web application provided breast cancer classification based on the three models. Performance metrics, accuracy, loss, and execution time outperformed the CNN model, achieving a more optimal execution time (807.82 seconds), accuracy (0.9544), and loss (0.1078). The model has achieved significant success in breast cancer, and with further refinements, it is anticipated that the model will be suitable for use as a decision support system.

KEYWORDS
Cancer
Computer vision processing
Convolutional neural network
Deep learning
Vision transformers

INTRODUCTION

Detection and Classification of breast cancer plays a vital role in monitoring the health-condition of cancer-patients in early stages of cancer. To monitor disease and flow, health-professionals keep track of the breast cancer patients' mammography x-ray images. The study has been tailored to lead health-professionals with a state-of-art solution to abstain from the wrong treatments and classifying breast cancer from breast images in early stage via classification of breast x-ray images via DL and CV. Their well-known models' and their implementations that are CNN, ViT and AlexNet have been investigated in this study. The breast X-ray image dataset have been utilized. Dataset comprises 3,383 mammogram images specifically highlighting breast tumors, organized within a structured folder format. Originally exported from Roboflow, a platform dedicated to computer vision projects it serves as a valuable resource for developing and evaluating deep learning models for breast tumor detection in mammographic images.

Manuscript received: 7 October 2025,
Revised: 15 December 2025,
Accepted: 17 December 2025.

[1]bora.basaran@tedu.edu.tr (**Corresponding author**)
[2]boncul@kastamonu.edu.tr.

Recent studies have demonstrated the effectiveness of deep learning and machine learning techniques in breast cancer classification, particularly in analyzing histopathological and radiological images. Convolutional neural networks (CNNs) have been widely applied to pathology and histology datasets, achieving high accuracy in both binary and multi-class classification tasks (Liu *et al.* 2022; Golatkar *et al.* 2018; Abunasser *et al.* 2023; Nguyen *et al.* 2019). These models are capable of automatically learning hierarchical features critical for distinguishing between malignant and benign cases. Hybrid approaches that integrate handcrafted features with dense neural layers have also shown promising results, improving model robustness and interpretability (Joseph *et al.* 2022). In addition, several studies have utilized patch-based strategies to capture localized tissue structures, enhancing classification accuracy through multi-size and discriminative input representations (Li *et al.* 2019). Transfer learning techniques have been employed to leverage pre-trained models and improve generalization in settings with limited labeled data (Saber *et al.* 2021).

Classical machine learning algorithms, such as Bayes classifiers and support vector machines, continue to be explored for their simplicity and effectiveness in structured data contexts (Bazila and Thirumalaikolundusubramanian 2018; Chen *et al.* 2023; Wu and Hicks 2021; Ara *et al.* 2021). Furthermore, recent work has pro-

posed fully automated deep learning pipelines for breast cancer detection (Ghrabat *et al.* 2022), and efforts to detect metastatic cancer using large-scale deep models have demonstrated the clinical relevance of AI-based solutions (Wang *et al.* 2016). These collective advancements provide a solid foundation for developing and evaluating diverse architectures such as CNNs, Vision Transformers (ViTs), and adapted AlexNet models within the scope of breast cancer classification, as investigated in this study.

The study is divided into four parts. In Section II, the proposed method and three deep learning (DL) models used to effectively address breast cancer classification through three different model architectures are discussed. In Section III, the results and discussion are presented based on the performance of the three models, including outcomes obtained via Hugging Face Spaces. Finally, Section IV concludes the study, demonstrating the effectiveness of the proposed approach and providing suggestions for future work.

PROPOSED METHOD

It is known that medical professionals have investigate breast x-ray images to effectively keep track of the health-monitoring of patients that have signs of breast cancers. The study has been tailored via provide comparative analysis to the medical professionals in breast cancer treatment domain with the classification models that they have been carried out. The comparative study has been divided into three models that are CNN, ViT and AlexNet to provide the most successful model to abstain from the faulty treatments to the health-professionals. In this study three DL models have been trained to obtain the most successful model on classification of breast cancer-domain.

Three DL models CNN, ViT and AlexNet have been trained in this manner with different hyperparameters. The images that exist on the dataset are images with the 640x640 pixel size that are each Breast X-ray images which those datasets are generally known as mammography data. Since 224x224 pixel size images have been used on classification in health-care domain; the images have been resized to the intended pixel size. Three models CNN, ViT and AlexNet will be investigated in a comparative manner, and they will provide the most optimal model to successfully classification of breast cancer patients via negative and positive that the classes of dataset have been spitted into two parts that are 0 and 1 respectively. To comparison of performances on each model, metrics that are accuracy, loss and time comparison has been investigated and based on those metrics; models' have been deployed on HuggingFace space via option of selection of model. Via HuggingFace space, health professionals will be uploading their breast x-ray images and they will be monitoring the health-condition of their patients effectively and the results they will obtain will be successfully classified images which will be breast cancer negative and breast cancer positive respectively.

The first architecture used in this study is a lightweight Custom Convolutional Neural Network (CNN), built from scratch for binary classification of breast X-ray images. It processes 224×224×3 input images through three convolutional blocks, each comprising a Conv2D layer with ReLU activation and a max-pooling layer. The blocks use 32, 64, and 32 filters respectively, all with 3×3 kernels. After the final block, the feature maps are flattened into an 86528-dimensional vector, followed by a dense layer with 128 ReLU units. A final sigmoid-activated neuron outputs the binary classification, cancerous or non-cancerous. This model offers an efficient and interpretable design suitable for real-time or resource-constrained clinical applications.

The second model employs a Vision Transformer (ViT), introducing attention-based mechanisms for classifying 224×224×3 breast X-ray images. The image is divided into 196 non-overlapping 16×16 patches, each flattened and embedded into a 64-dimensional space with positional encoding. These embeddings pass through two Transformer Encoder blocks containing LayerNorm, Multi-Head Self-Attention, and Feed Forward Networks with residual connections. A global average pooling layer aggregates the information, followed by a dense layer (128 units, ReLU) and a final sigmoid-activated neuron for binary cancer prediction. ViT enables the model to capture global context and subtle spatial patterns beyond the reach of traditional CNNs.

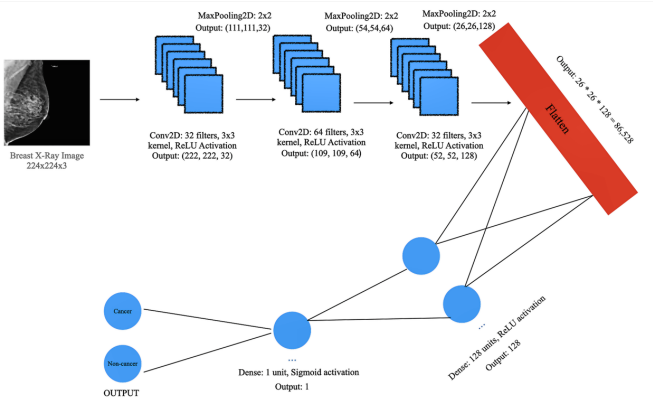
The third model adapts AlexNet, a classical deep CNN, for breast cancer classification. It processes 224×224×3 images through five convolutional layers with filters ranging from 96 to 384, using 11×11, 5×5, and 3×3 kernels, followed by max-pooling. The extracted features are flattened into a 9216-dimensional vector and passed through two fully connected layers with 4096 ReLU units and dropout (0.5). A final sigmoid-activated layer provides binary output. This adaptation highlights the effectiveness of traditional deep networks in medical imaging tasks when domain-optimized (Figure 1).

RESULTS AND DISCUSSION

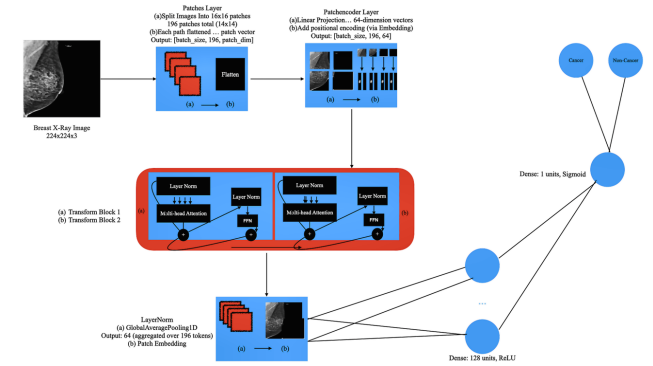
Carrying out three (DL) models provide health-professionals to choose to compare the results with three models' outcomes and their sight. Furtherly study has been conducted with the results of breast cancer negative and positive respectively on HuggingFace space on a use of medical professionals. Both breast cancer negative and positive images have been investigated, and their potential results' have been analyzed in a comparative manner to provide valuable insight into medical and computer science integrated domain which solely focusses on cancer detection. In Study results' that have been observed that some results' have been classified as faulty even though high accuracy metrics' that have been employed. To abstain from the wrong classification, the corresponding output images as breast cancer negative and positive will be carried out with even the sight of the medical professionals for the most correct outcome that will be derived from the study which will be in use of medical professionals that the breast x-ray images that they have in their hand and patients'.

In this study for provide an example two breast x-ray images from the dataset have been selected to compare the performances of classification that one of the images are selected from breast cancer negative and that the other image has been selected from breast cancer positive and their classification performance on HuggingFace space have been investigated. According to the HuggingFace results that they have been obtained in positive breast-cancer results based on breast x-ray images there are no misclassification on models CNN, ViT and AlexNet respectively. However, there is a minor fault classification in negative breast-cancer result in the AlexNet model; even though high classification accuracy that has been obtained in Training set. In that circumstance the sight of medical professional in chest major is a must that they are provided in (Figure 2).

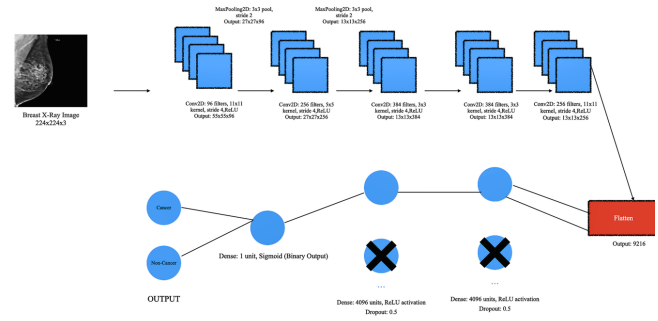
Due to the misclassification on breast-cancer-negative image on the AlexNet model must keep trackt of the models' performance with several metrics. Several metrics in this domain are based on classification metrics due to the classification study has been tailored that the metrics that are in use will be accuracy and loss respectively. Accuracy and losses of the system will keep track of during each of epochs during execution of training models re-



(a) CNN Architecture

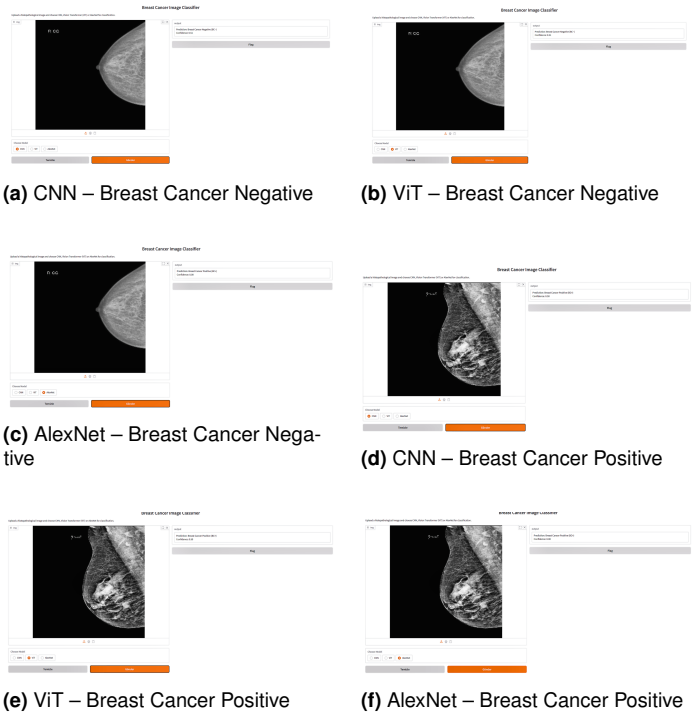


(b) ViT Architecture



(c) AlexNet Architecture

Figure 1 Model Architectures of Models that they have trained during the Breast Cancer Classification study.



(a) CNN – Breast Cancer Negative

(b) ViT – Breast Cancer Negative

(c) AlexNet – Breast Cancer Negative

(d) CNN – Breast Cancer Positive

(e) ViT – Breast Cancer Positive

(f) AlexNet – Breast Cancer Positive

Figure 2 HuggingFace Breast Cancer Classification Space which one Breast Cancer Negative and one Breast Cancer Positive Images have been investigated.

spectively on training and validation sets to observe how model behaves with the inputs that the inputs are breast x-ray images. Loss function of binary cross-entropy will be observed due to it is a binary classification study that the sigmoid activation function activates the neurons at the last layer of each model. Each models' accuracy and loss metrics have been played a vital role in breast-cancer classification study with the accuracies that are over 0,9 in each of the models' separately. Loss functions have been diminished based on epochs in each model which is a shown that the models behave well under several conditions in both validation and training splits. Additionally, validation performances of models have been trained; strong even though plot of them seems not that promising. Even though breast-cancer classification has been trained and the outcomes have been provided correct classification which will also lead health-professionals on the domain of cancer classification which will save many human-beings' life in early diagnosis of breast cancer. Accuracy and Loss plots of each model on training and validation sets have been provided in (Figure 3).

Table 1 presents a comprehensive comparison of the three deep learning models CNN, Vision Transformer (ViT), and AlexNet, in terms of classification accuracy, loss, training epochs, parameter counts, and execution time during breast X-ray image classification. Among the three models, the Custom CNN achieved the highest accuracy of 95.44% with a low loss of 0.1078 in just 45 epochs. This performance demonstrates its strong ability to extract relevant features from medical images using a lightweight architecture. Additionally, the CNN model required approximately 807.82 seconds of execution time and had around 11.1 million trainable parameters, striking a good balance between performance and computational efficiency.

The ViT model, on the other hand, while introducing advanced attention mechanisms, achieved a lower accuracy of 91.56% with a higher loss of 0.1987 despite training for 75 epochs. Notably, the ViT had the fewest trainable parameters (236,737) and the lowest model complexity, yet it required 1365.79 seconds to execute nearly double the time of the CNN. This indicates that although ViT excels at capturing long-range dependencies, it may not be as well-suited for small-scale medical datasets or classification tasks where local features dominate. AlexNet, the most complex model in terms of architecture and parameters (46.75 million trainable parameters), achieved an accuracy of 95.20%, which is comparable to CNN, with a slightly better loss of 0.1072.

However, this performance came at a significant computational cost, it recorded the highest execution time of 2597.42 seconds,

Table 1 Models and Their Performances have been analyzed in a comparative manner in Breast Cancer Classification study

Metrics / Model	CNN	ViT	AlexNet
Accuracy	0.9544	0.9156	0.9520
Epochs	45	75	50
Loss	0.1078	0.1987	0.1072
Trainable Parameters	11.169.089	236.737	46.751.105
Total Parameters	11.169.089	236.737	46.751.105
Execution Time	807.82	1365.79	2597.42

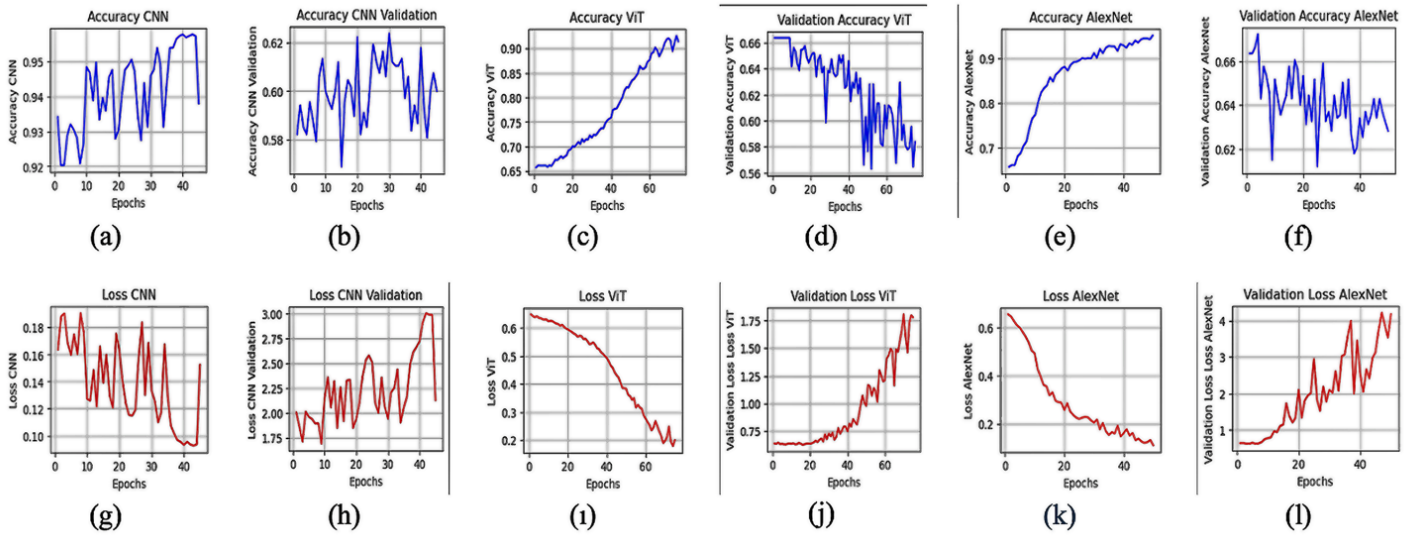


Figure 3 Accuracy and Loss Visualizations of models that they have been trained on Breast Cancer Classification study. (a) CNN – Accuracy (Train), (b) CNN – Accuracy (Validation), (c) ViT – Accuracy (Train), (d) ViT – Accuracy (Validation), (e) AlexNet – Accuracy (Train), (f) AlexNet – Accuracy (Validation), (g) CNN – Loss (Train), (h) CNN – Loss (Validation), (ı) ViT – Loss (Train), (j) ViT – Loss (Validation), (k) AlexNet – Loss (Train), (l) AlexNet – Loss (Validation).

more than three times that of CNN. This highlights the trade-off between deep architectural complexity and practical applicability in real-time or resource-limited clinical environments. In summary, the CNN model consistently outperformed both the Vision Transformer (ViT) and AlexNet architectures across multiple evaluation metrics, including accuracy, computational efficiency, and execution time, establishing it as the most optimal choice for the breast cancer classification task in this study.

The CNN's ability to effectively capture spatial hierarchies and local features contributed to its superior classification performance, making it well-suited for medical image analysis where precision is critical. While the ViT model demonstrated promise with its relatively minimal number of trainable parameters, its higher execution overhead and longer processing times present practical challenges for real-time or resource-constrained diagnostic environments. On the other hand, AlexNet, despite achieving comparable accuracy to CNN, proved to be computationally demanding and slower, limiting its feasibility for rapid clinical deployment. These findings highlight the trade-offs between model complexity, accuracy, and operational efficiency, emphasizing the CNN's balance as ideal for breast cancer detection workflows that require both reliability and speed. Future work could explore hybrid models or lightweight transformer variants to potentially combine the strengths of these architectures while mitigating their respective limitations.

CONCLUSION

In this study, three deep learning models trained from scratch that are respectively CNN, ViT and AlexNet have been analyzed in a comparative manner with the Breast X-ray images to classify the input images based on breast cancer negative and positive. The study has been conducted with a dataset of breast cancer x-ray images that are in use for computer vision and deep learning studies that the dataset consists of 3,383 breast x-ray images. Study has been tailored via using three deep learning models' and their respective implementations seperately and after training of models' HuggingFace space environment has been used to deployment of the study which the study has been classified several breast x-ray images based on breast cancer classification which classify images breast cancer negative and breast cancer positive respectively. To compare the performance of three proposed models' performance metrics of accuracy, loss and execution time have been analyzed in

both training and validation sets respectively.

According to the outcomes of the results, medical professionals should keep track of the information of CNN preferable, apart from CNN; they may rely on the trust of the ViT model. However as discussed in the begining of the results section AlexNet model have proven some misclassifications even though its high accuracy and losses obtained during training. Accuracy and loss performances of the models' proven the suggestion that the medical Professionals should rely on the CNN models' trust in the classification of breast cancer x-ray images based on cancer. Outcomes have been proven that the 0,9544 accuracy and 0,1078 loss performance observed in CNN model which outperformed both ViT and AlexNet models respectively. In further studies It should be analyzed that the AlexNet model should be trained with higher number of epochs with more amount of execution time or in comparison, different models' apart from AlexNet should be investigated to performance comparison on breast cancer domain.

Acknowledgments

Preliminary findings of this study were presented as an oral abstract entitled "Evrişimsel Sinir Ağları, Görüntü Dönüştürücüler ve AlexNet ile Meme X-Ray Görüntüleriyle Meme Kanserinin Sınıflandırılması" at the 5th International Turkish World Engineering and Science Congress (Abstract No: 172092).

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Abunasser, B. S., M. R. J. Al-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, 2023 Convolution neural network for breast cancer detection and classification using deep learning. Asian Pacific journal of cancer prevention: APJCP **24**: 531.

Ara, S., A. Das, and A. Dey, 2021 Malignant and benign breast cancer classification using machine learning algorithms. In *2021 international conference on artificial intelligence (ICAI)*, volume 2021, pp. 97–101, IEEE Islamabad, Pakistan.

Bazila, B. B. A. and P. Thirumalaikolundusubramanian, 2018 Comparison of bayes classifiers for breast cancer classification. Asian Pacific Journal of Cancer Prevention **19**: 2917–2922.

Chen, H., N. Wang, X. Du, K. Mei, Y. Zhou, *et al.*, 2023 Classification prediction of breast cancer based on machine learning. Computational intelligence and neuroscience **2023**: 6530719.

Ghrabat, M. J. J., Z. A. Hussien, M. S. Khalefa, Z. A. Abduljabba, V. O. Nyangaresi, *et al.*, 2022 Fully automated model on breast cancer classification using deep learning classifiers. Indonesian Journal of Electrical Engineering and Computer Science **28**: 183–191.

Golatkar, A., D. Anand, and A. Sethi, 2018 Classification of breast cancer histology using deep learning. In *International conference image analysis and recognition*, pp. 837–844, Springer.

Joseph, A. A., M. Abdullahi, S. B. Junaidu, H. H. Ibrahim, and H. Chiroma, 2022 Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer). Intelligent Systems with Applications **14**: 200066.

Li, Y., J. Wu, and Q. Wu, 2019 Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning. IEEE Access **7**: 21400–21408.

Liu, M., L. Hu, Y. Tang, C. Wang, Y. He, *et al.*, 2022 A deep learning method for breast cancer classification in the pathology images. IEEE Journal of Biomedical and Health Informatics **26**: 5025–5036.

Nguyen, C. P., A. H. Vo, and B. T. Nguyen, 2019 Breast cancer histology image classification using deep learning. In *2019 19th international symposium on communications and information technologies (ISCIT)*, pp. 366–370, IEEE.

Saber, A., M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, 2021 A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique. IEEe Access **9**: 71194–71209.

Wang, D., A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, 2016 Deep learning for identifying metastatic breast cancer.

Wu, J. and C. Hicks, 2021 Breast cancer type classification using machine learning. Journal of personalized medicine **11**: 61.

How to cite this article: Basaran, B., and Oncul, A. B. Classification of Breast Cancer with Breast X-Ray Images via Convolution Neural Networks, Vision Transformers and AlexNet. *Computers and Electronics in Medicine*, 3(1), 94-98, 2026.

ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.99-108
https://doi.org/10.69882/adba.cem.20260112

Adaptive–Scaled Digital Watermarking in Color Medical Imaging

İrem Özmen [α,1], **Zeynep Çetinkaya** [α,2], **Fahrettin Horasan** [α,3], **Fatih Varçın** [β,4] **and Shaobo He** [*,5]

[α]Department of Computer Engineering, Kırıkkale University, Kırıkkale, Türkiye, [β]Department of Computer Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye, [*]School of Automation and Electronic Information, Xiangtan University, Xiangtan, 411105, China.

ABSTRACT Nowadays digital environment, the rapid duplication and unauthorized use of color images shared online have led to increasing concerns regarding copyright infringement and data security. Therefore, the development of effective and robust methods for protecting digital content has become critically important. In this study, an integrated digital watermarking method is proposed to ensure copyright protection and data security for both color and medical images. In the proposed approach, watermark embedding is performed by jointly applying Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD) on the blue (B) color channel of the cover image. The watermark image is preprocessed using a dimension reduction technique and optimized through the Grey Wolf Optimization (GWO) algorithm. The optimized watermark is then embedded into the cover image, and inverse transformations are applied to obtain the watermarked image. The performance of the proposed algorithm is evaluated on standard test images under various attack scenarios. Imperceptibility is assessed using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), while robustness is measured using the Normalized Correlation (NC) and Bit Error Rate (BER) metrics. Experimental results demonstrate that the proposed method achieves high imperceptibility and successfully preserves watermark information under different attacks. Comparisons with related studies in the literature indicate that the proposed approach is competitive and, in some cases, superior in terms of both imperceptibility and robustness. Consequently, the proposed method provides an effective and reliable digital watermarking solution for color images.

KEYWORDS
Data hiding
Color image
Medical image watermarking
Dimension reduction

INTRODUCTION

With the advancement of the technological era, the increasing volume of digital data has made the storage, transmission, and confidentiality of this data more critical (Gull and Parah 2024; Awasthi *et al.* 2024). This situation has led to a growing importance of security systems developed to ensure data security, which are generally classified into data encryption and data hiding techniques (Çelik and Doğan 2021). Cryptography, a subfield of data encryption, is a method that encrypts content to prevent it from being read by unauthorized parties and ensures that it is accessible only to authorized users (Çelik and Yalçın 2023; Çelik and Doğan 2021). Steganography, as a subcategory of data hiding, aims to preserve data confidentiality by concealing the very existence of the data (Awasthi *et al.* 2024; Çelik and Doğan 2021).

In digital image watermarking, copyright protection allows the verification of content ownership by embedding a hidden watermark into an image (Awasthi *et al.* 2024; Çelik and Doğan 2021). During the watermarking process applied to digital data, the modifications introduced into the data must remain imperceptible to third parties. In other words, the impact of the watermarking process on the original digital data should be minimized as much as possible (Nawaz *et al.* 2025). From this perspective, digital watermarking provides an effective solution for ensuring data security (Wang *et al.* 2023). The digital watermarking process consists of two main stages: watermark embedding and watermark extraction (Mamuti 2019). In the watermark embedding stage, the watermark information is embedded into the cover signal (Awasthi *et al.* 2024; Yağcıoğlu and Sondaş 2021), while in the extraction and verification stage, the watermark is retrieved by the receiver and the verification process is performed. The stages of the digital watermarking process are illustrated in Figure 1.

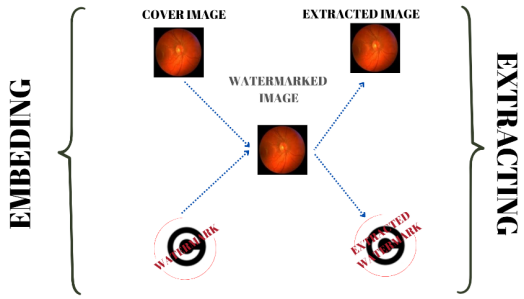


Figure 1 Watermarking scheme (embedding and extraction)

Watermarking methods are classified according to various criteria in order to meet different application requirements, and this study focuses on approaches used for color images (Wang *et al.*

Manuscript received: 14 October 2025,
Revised: 7 January 2026,
Accepted: 7 January 2026.

[1]irem_ozmen@kku.edu.tr
[2]zeynepcetinkaya@kku.edu.tr
[3]fhorasan@kku.edu.tr (**Corresponding author**)
[4]fatihvarcin@subu.edu.tr
[5]heshaobo@xtu.edu.cn

2023; Mahto and Singh 2021; Sharma *et al.* 2023). This classification is based on the type of cover signal, watermark perceptibility, embedding domain, robustness, and data extraction method (Nawaz *et al.* 2025; Mahto and Singh 2021).

The first criterion used in the classification of watermarking methods is the type of the cover signal. Different signal types such as text, image, audio, and video can be used as cover signals in the watermarking process (Awasthi *et al.* 2024; Mahto and Singh 2021). Another classification criterion is whether the watermark is perceptible to the human visual system. In this context, watermarking methods are divided into two groups: visible and invisible watermarking (Awasthi *et al.* 2024; Mahto and Singh 2021; Sharma *et al.* 2023). Invisible watermarks are generally used in applications such as copyright protection, image authentication, and privacy protection (Mahto and Singh 2021).

The domain in which the watermarking process is performed is also an important classification criterion (Wang *et al.* 2023; Mahto and Singh 2021). A watermark can be embedded either in the spatial domain or in the transform domain of the cover signal (Hu 2024; Sharma *et al.* 2023). In the Least Significant Bit (LSB) technique, which is one of the spatial domain methods, watermark bits are stored in the least significant bits of the pixels of the carrier image (Gull and Parah 2024; Wang *et al.* 2016). In transform-domain methods, the image is represented using specific mathematical transformations; examples include techniques based on DWT, DCT, and SVD (Hu 2024; Sharma *et al.* 2023; Wang *et al.* 2016). The ability of transform-domain methods to effectively exploit the principles of the human visual system and visual perception characteristics has made them highly popular (Hu 2024). In line with these advantages, this study proposes a watermarking method based on transform-domain techniques.

Watermarking techniques are also classified according to their robustness against external attacks. In this context, they are generally categorized into three main groups: fragile, semi-fragile, and robust watermarking methods (Mahto and Singh 2021; Anand and Singh 2021; Singh *et al.* 2023). Robust watermarks exhibit resistance to both intentional and unintentional modifications applied to the image (Awasthi *et al.* 2024). In contrast, fragile watermarks are easily degraded even by minor alterations. Semi-fragile watermarking methods provide resistance to certain types of attacks while enabling the detection of content manipulations (Zainol *et al.* 2021; Sharma *et al.* 2024). Finally, watermarking systems are classified into three groups blind, semi-blind, and non-blind based on the information required during the watermark extraction process (Awasthi *et al.* 2024; Sharma *et al.* 2023). In blind watermarking methods, the original image is not required at the extraction stage; semi-blind watermarking methods require only the original watermark information, whereas non-blind watermarking methods require both the cover image and the watermark information (Awasthi *et al.* 2024; Sharma *et al.* 2023; Zainol *et al.* 2021).

There are three fundamental properties that determine the performance of digital watermarking systems: imperceptibility, robustness, and capacity. Imperceptibility refers to the similarity between the cover image and the watermarked image obtained after the watermarking process. In other words, it indicates the level of distortion introduced into the cover image by the watermark embedding operation . Robustness is defined as the similarity between the original watermark image and the watermark extracted from the watermarked image, and it represents the degree of degradation affecting the watermark during the watermarking process (Awasthi *et al.* 2024; Hu 2024; Zainol *et al.* 2021). Capacity denotes the amount of watermark data that can be embedded into the cover image. Establishing an appropriate balance among these parameters is of great importance for the overall performance of watermarking systems (Hu 2024).

In watermarking schemes, the relationship among robustness, imperceptibility, and capacity follows a trade-off structure (Begum and Uddin 2020). For instance, as the data payload increases, i.e., as capacity is enhanced, the robustness of the watermark decreases while its perceptibility increases. Similarly, improving imperceptibility may lead to a reduction in both robustness and capacity. Maintaining all parameters at optimal levels within this trade-off triangle is one of the primary objectives of digital watermarking systems (Begum and Uddin 2020; Qasim *et al.* 2018). These trade-off components are illustrated in Figure 2.

Figure 2 Trade-off components (imperceptibility, robustness, and capacity)

In watermarking methods, not only the structure of the watermark image but also the type of the cover (host) image is an important classification criterion that affects the design and performance of the algorithm. While cover images may be grayscale or color, watermark images can be used in binary, grayscale, or color formats depending on the application (Wang *et al.* 2023; Sharma *et al.* 2023). These different data structures of the cover and watermark images directly influence performance metrics such as capacity, imperceptibility, and robustness of watermarking methods.

Color cover images are represented in the RGB color space and have a three-dimensional matrix structure consisting of rows, columns, and color channels. Within this structure, each layer corresponds to one of the red (R), green (G), and blue (B) color components. Each color channel can be treated as a grayscale image, and the watermarking process can be applied separately to these channels (Hu 2024; Mamuti 2019). This multi-channel structure enables higher data embedding capacity during the watermarking process (Wang *et al.* 2023; Hu 2024).

The watermark image, on the other hand, can be selected as binary, grayscale, or color depending on the application. Embedding strategies and extraction methods vary according to the data structure of the watermark (Wang *et al.* 2023; Sharma *et al.* 2023). In this study, color images are preferred as cover images, while a grayscale image is used as the watermark. This choice aims to exploit the capacity advantage provided by color cover images while establishing a balanced structure between imperceptibility and robustness through the use of a grayscale watermark.

Performance evaluation in watermarking methods is generally carried out based on imperceptibility and robustness criteria (Awasthi *et al.* 2024). To assess the imperceptibility performance of watermarking schemes, Peak Signal-to-Noise Ratio (PSNR) and

Structural Similarity Index (SSIM) metrics are widely used for image quality and similarity analysis. For robustness evaluation, the Normalized Correlation (NC) metric is employed to measure the similarity between the extracted watermark and the original watermark, while the Bit Error Rate (BER) metric is used to assess the degree of robustness (Karmouni *et al.* 2024; Çelik and Yalçın 2023; Melman and Evsutin 2023). The PSNR metric is calculated using the following Equation (1).

$$\mathrm{PSNR}(H, H_w) = 10 \log_{10} \left(\frac{\mathrm{Max}_H^2}{\mathrm{MSE}} \right) \tag{1}$$

Here, H denotes the cover image, while Hw represents the watermarked image. The Mean Squared Error (MSE) represents the average pixel-wise error between the cover and watermarked images (Eltoukhy *et al.* 2023). MSE is defined by Equation (2):

$$\mathrm{MSE} = \frac{1}{N^2} \sum_{x=1}^{n} \sum_{y=1}^{n} [H(x,y) - H_w(x,y)]^2 \tag{2}$$

The SSIM, which is a structural similarity measure, is defined by Equation (3):

$$\mathrm{SSIM}(H, H_w) = \frac{(2\mu_H \mu_{H_w} + C_1)(2\sigma_{HH_w} + C_2)}{(\mu_H^2 + \mu_{H_w}^2 + C_1)(\sigma_H^2 + \sigma_{H_w}^2 + C_2)} \tag{3}$$

Here, μ_H and μ_{H_w} denote the local mean values of the cover image H and the watermarked image H_w, respectively, while σ_H and σ_{H_w} represent their corresponding standard deviations. The constants C_1 and C_2 are introduced to ensure numerical stability during computation. For an ideally imperceptible watermarking algorithm, the SSIM value is expected to be close to 1 (Horasan *et al.* 2019; Mousavi *et al.* 2014).

To evaluate robustness in watermarking schemes, the watermarked image is subjected to various potential attacks during testing (Mousavi *et al.* 2014). These attacks are generally classified as geometric and conventional attacks. Commonly applied attack types include rotation, cropping, copy–paste, deletion, median filtering, JPEG compression, salt-and-pepper noise, and Gaussian filtering operations (Li *et al.* 2023; Khare and Srivastava 2021; Horasan *et al.* 2019).

The NC metric, which is used for robustness evaluation, is defined by Equation (4):

$$\mathrm{NC} = \frac{\sum_{i=1}^{P} \sum_{j=1}^{Q} w(i,j)\, w'(i,j)}{\sum_{i=1}^{P} \sum_{j=1}^{Q} [w(i,j)]^2} \tag{4}$$

In this equation, $w(i,j)$ represents the original watermark image, while $w'(i,j)$ denotes the extracted watermark image (Mohanarathinam *et al.* 2020; Anand and Singh 2021). An NC value greater than 0.75 under attack conditions and close to 1 under ideal conditions is generally expected. As the NC value approaches unity, the watermarking method is regarded as robust and resilient (Yurttakal and Horasan 2022; Priyadarshini and Naik 2024).

The BER metric, which is used to evaluate robustness, represents the ratio of incorrectly extracted bits to the total number of embedded bits. BER values range between 0 and 1, where a value of 0 indicates error-free extraction [38]. BER is defined by Equation (5):

$$\mathrm{BER} = \frac{\text{Number of Errors}}{\text{Total Number of Transmitted Bits}} \tag{5}$$

LITERATURE REVIEW

The rapid production and sharing of visual data in digital environments have led to increasing issues related to copyright infringement and unauthorized use. In this context, digital image watermarking has emerged as an effective solution for ownership verification, preservation of content integrity, and prevention of unauthorized copying. Although early studies primarily focused on grayscale images, the widespread use of color images in real-world applications has necessitated the adaptation of watermarking methods to multi-channel and more complex structures. The presence of multiple channels in color images offers the potential to increase watermark embedding capacity, while also introducing new challenges due to inter-channel correlations and compression processes.

Studies in the field of color image watermarking have predominantly focused on classical transform-domain-based approaches. In this regard, Wang *et al.* (2023) proposed a blind color image watermarking method using mid-frequency coefficients in the two-dimensional Discrete Cosine Transform (2D-DCT) domain. To enhance watermark security, a two-stage encryption process based on affine transformation and Arnold transform was applied. In addition, imperceptibility and robustness were improved by employing variable embedding coefficients across different blocks and layers. Although such DCT-based methods provide high visual quality, a significant portion of approaches operating in the RGB color space do not explicitly address the effects of color space transformation and quantization applied during the JPEG compression process. This limitation may lead to notable performance degradation of RGB-based watermarks, particularly under JPEG attacks.

The neglect of spectral relationships among RGB channels has emerged as one of the fundamental weaknesses of color image watermarking methods in the literature. Addressing this issue, Hu (2024) demonstrated that the conversion to the YCbCr color space and the quantization of chroma components during the JPEG compression process have destructive effects on watermark information embedded in RGB channels. To mitigate this problem, the proposed synergistic compensation (SC) approach aims to enhance JPEG robustness by suppressing quantization errors in channels where the watermark is not embedded.

Other studies aiming to model inter-channel relationships in a more holistic manner have focused on quaternion-based approaches. Wang *et al.* (2025) proposed a watermarking method that preserves linear correlations among RGB channels by employing a split quaternion matrix model and double-layer singular value decomposition. Similarly, Wang *et al.* (2013) introduced a blind watermarking method based on the quaternion Fourier transform, targeting high robustness against both color-based attacks and geometric distortions. Although such methods provide strong robustness, they involve practical limitations such as increased computational cost and additional information requirements.

More recent studies in the literature have emphasized optimization-based and hybrid-domain approaches to improve the imperceptibility–robustness trade-off. Sharma *et al.* (2023) and Sahir *et al.* (2025) optimized scaling factors in DWT–SVD and DWT–HD–SVD frameworks, respectively, using the Artificial Bee Colony (ABC) algorithm, thereby enhancing both visual quality and robustness against attacks. Agarwal and Singh (2022) employed a genetic algorithm in the DCT domain to select optimal pixel groups, achieving notable improvements in PSNR and NCC values, particularly across RGB channels. Ahmadi *et al.* (2021) proposed a blind dual watermarking approach based on

DWT–SVD and Particle Swarm Optimization (PSO), addressing both copyright protection and integrity authentication within a unified framework; robust watermarking performed specifically on the blue channel resulted in high capacity and robustness.

Among approaches focusing on lower computational cost, Su and Chen (2018) proposed a blind color image watermarking method that directly utilizes DC coefficients in the spatial domain instead of the transform domain, achieving satisfactory robustness against JPEG compression and noise attacks. This method stands out in terms of processing speed and implementation simplicity. Roy and Pal (2019) introduced a DWT–SVD-based approach relying solely on the luminance component in the YCbCr color space; although the method demonstrates robustness under various attacks, its non-blind structure and single-channel utilization limit its capacity. In the context of medical image security, Eltoukhy *et al.* (2023) developed a watermarking method based on Slant–SVD–QFT transforms combined with one-time pad (OTP) encryption, providing very high visual quality and security. Finally, Su *et al.* (2024) presented a low-latency, high-security, and fully blind fusion-domain watermarking scheme using graph-based transformation and PSO.mOverall, a review of the literature reveals that the field of color image watermarking has witnessed a wide diversity of methods and significant advancements, ranging from classical transform-domain approaches to quaternion-based models, and from heuristic optimization algorithms to graph-based and learning-assisted techniques. These studies offer various advantages in terms of imperceptibility, robustness, security, and computational efficiency.

However, it is noteworthy that the effects of attacks originating from color space transformations and quantization processes such as those introduced by JPEG compression are often addressed indirectly or treated as secondary issues in many studies, particularly in RGB-based methods. Therefore, approaches that explicitly consider the interaction between color space conversion and compression processes provide a complementary and strengthening perspective to existing methods.

THEORICAL BACKGROUND

Discrete Wavelet Transform (DWT)

The DWT is an effective signal processing technique that enables multilevel analysis by decomposing images into their frequency components. Through the LL, LH, HL, and HH sub-bands obtained after applying DWT, the fundamental structural information and detailed components of an image can be separated (Wang *et al.* 2016; Othman and Zeebaree 2020). The high time–frequency localization capability provided by DWT allows the watermark to be embedded efficiently without significantly degrading image quality. Moreover, the ability to extract the watermark without requiring the original image makes DWT a suitable method for applications such as compression and noise reduction (Othman and Zeebaree 2020).

In DWT-based watermarking approaches for color images, the image is first decomposed into RGB or YCbCr color spaces, and DWT is applied separately to each color channel to obtain the corresponding sub-bands. By selecting appropriate sub-bands for watermark embedding, visual quality can be preserved while robustness against signal-processing-based attacks is enhanced (Karmouni *et al.* 2024). The DWT-based approach for color images is schematically illustrated in Figure 3.

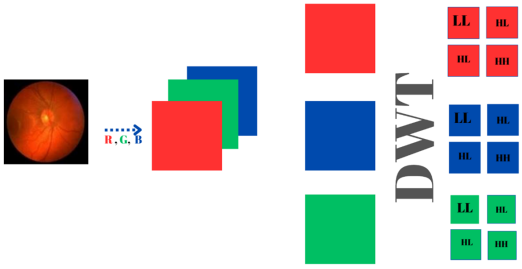


Figure 3 DWT-based approach for color images

Singular Value Decomposition (SVD)

SVD is a powerful linear algebra technique that is widely used in image processing and, in particular, in digital watermarking due to its ability to preserve visual quality and provide robustness against various distortions [19]. SVD decomposes an image matrix into three matrices (U, S, V^T) where the singular values in S capture the intrinsic properties of the image and exhibit high stability under common image processing operations (Mohanarathinam *et al.* 2020; Zainol *et al.* 2021).

In the literature, SVD is commonly combined with the DWT in adaptive watermarking algorithms that allow watermark embedding in both low- and high-frequency components. Such approaches contribute to the development of robust watermarking systems capable of withstanding attacks such as contrast and brightness adjustments, cropping, and noise addition (Mohanarathinam *et al.* 2020). The application of SVD on images is illustrated in Figure 4.

Figure 4 Application of Singular Value Decomposition

Dimension Reduction

Dimension reduction is an approach that aims to obtain a more compact representation that preserves the essential structure of the data by eliminating components that are considered noise or do not carry meaningful information in large datasets. This process not only reduces computational cost but also increases data processing speed and enhances system robustness by suppressing noise present in the signal. As a result, more meaningful relationships can be revealed, and in applications such as watermarking, the imperceptibility performance can be improved (Yurttakal and Horasan 2022; Horasan *et al.* 2019; Çetinkaya and Horasan 2025). The general structure of the dimension reduction approach is illustrated in Figure 5.

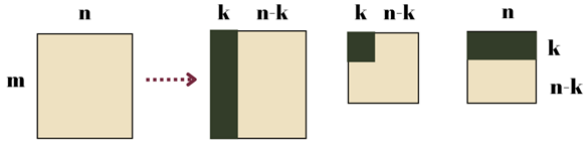


Figure 5 Dimension Reduction

Optimization Algorithm

Optimization is a process that aims to obtain the best (optimal) value of an objective function defined under certain constraints (Melman and Evsutin 2023). In this study, Grey Wolf Optimization (GWO) is examined and applied. GWO is an optimization algorithm inspired by the social hierarchy and hunting behavior of grey wolf packs, proposed by Mirjalili et all. (Seyedali and Andrew 2014). In this method, candidate solutions are represented by alpha, beta, delta, and omega wolves, and the search process is guided according to this hierarchical structure (Mirjalili 2015; Seyedali and Andrew 2014). The position update mechanism, which models the encircling and attacking phases of prey, together with a linearly decreasing control parameter, enables the algorithm to gradually transition from the exploration phase to the exploitation phase (Rodríguez *et al.* 2017; Seyedali and Andrew 2014).

MATERIAL AND METHOD

Embedding Process

In this study, color images in RGB format were selected as cover images, and these images were decomposed into three color channels: red (R), green (G), and blue (B). The digital watermarking process was performed solely on the B channel, and the matrix corresponding to this channel was obtained. The DWT was applied to the obtained B matrix, and among the subbands obtained from the transformation, the low-frequency LL band was selected for processing. The SVD algorithm was applied to the LL band to obtain a singular value matrix from this band. Similarly, the rank-k SVD algorithm was applied to the data to be used as the watermark for dimension reduction, and the singular value matrix of the watermark was computed. These two singular value matrices obtained from the cover image and the watermark data were summed prior to the merging operation to form a new singular value matrix. The resulting new singular values were reconstructed using the inverse SVD process.

The Inverse DWT was applied to the LL band obtained from this process to yield the B channel matrix containing the digital watermark. In the final stage, the original R and G channels were combined with the watermarked B channel to successfully generate the watermarked color image. The stages of the DWT–SVD based watermark embedding process performed on the B channel of the cover image are schematically illustrated in Figure 6.

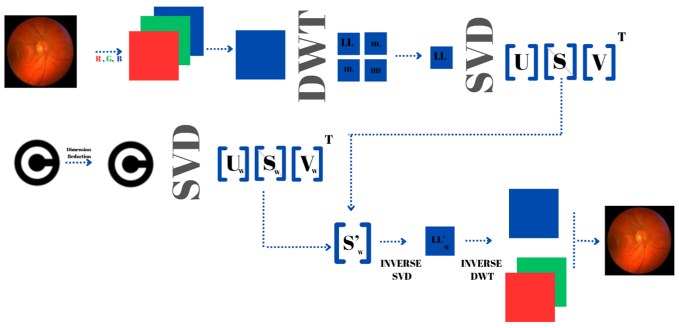


Figure 6 Embedding Process

Extraction Process

After obtaining the watermarked image, the image was decomposed into RGB color channels, and all operations were performed solely on the B channel. The DWT was applied to this channel, decomposing the image into four subbands: LL, LH, HL, and HH.

SVD was applied to the low-frequency LL band of the watermarked image, and the corresponding singular value matrix was obtained as a result of this operation. For watermark extraction, the singular value matrix of the original cover image was subtracted from the singular value matrix obtained from the watermarked image, thereby isolating the singular values corresponding exclusively to the embedded watermark.

These isolated singular values were combined with the other two SVD matrices of the watermark obtained during the embedding stage, and the watermark image was reconstructed using the Inverse SVD method. As a result of this process, the watermark image was successfully retrieved. The extracted watermark image has the same dimensions as the dimension reduced watermark used in the embedding stage. The overall workflow of the proposed watermark extraction process is illustrated in Figure 7.

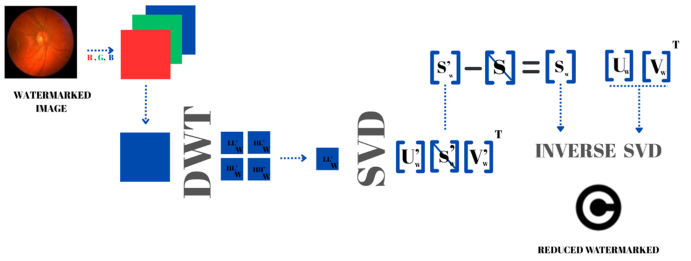


Figure 7 Extraction Process

Optimization Process

In this study, an optimization problem was defined to determine the optimal values of the parameters used in the RGB multi-level DWT–SVD based watermarking method. The optimization process aims to achieve a balance between the robustness and imperceptibility metrics obtained under different attack scenarios.

The optimization problem involves two fundamental decision variables: k, representing the rank value employed in the SVD of the watermark, and α, denoting the scaling coefficient used to embed the watermark into the singular values of the cover image.

When the constraints are examined in detail:

- **Rank (k) Constraint:** Given that the watermark image has dimensions $h_w \times w_w$, the maximum meaningful rank value is defined in Equation (6).

$$k_{\max} = \min(h_w, w_w) \tag{6}$$

To prevent the adverse effects of excessively small or excessively large rank values, it has been constrained within the range specified in Equation 7.

$$0.1\,k_{\max} \leq k \leq 0.75\,k_{\max} \tag{7}$$

This constraint prevents both excessive reduction of the watermark and overly aggressive embedding.

- **Scaling Coefficient (α) Constraint**: This controls the effect of the watermark on the cover image. This parameter is defined in the continuous interval in Equation 8.

$$\alpha_{\min} \leq \alpha \leq \alpha_{\max} \tag{8}$$

This range has been determined experimentally to provide a reasonable balance between imperceptibility and robustness.

Each candidate solution (k, α) was evaluated using NC, BER, SSIM and PSNR metrics under multiple attack scenarios. The results of these metrics reflect both the extractability of the watermark and the perceptual quality of the cover image. When these metrics are examined:

- PSNR Normalization: Since the PSNR value is measured in dB, it has been normalized to be evaluated on the same scale as the other metrics, as given in Equation 9.

$$\text{normalizedPSNR} = \begin{cases} 1, & \text{PSNR} \geq 37, \\ \dfrac{\text{PSNR}}{37}, & \text{PSNR} < 37. \end{cases} \tag{9}$$

  This threshold was determined based on the 37 dB reference, which is widely accepted in the literature for high perceptual quality.
- Attack-Based Average Performance: For a candidate solution (k, α), the averages of the metrics obtained under all attack scenarios are defined as follows Equations 10-13, N denotes the total number of attacks employed.

$$\overline{\text{NC}} = \frac{1}{N}\sum_{i=1}^{N} \text{NC}_i \tag{10}$$

$$\overline{\text{invBER}} = \frac{1}{N}\sum_{i=1}^{N} (1 - \text{BER}_i) \tag{11}$$

$$\overline{\text{SSIM}} = \frac{1}{N}\sum_{i=1}^{N} \text{SSIM}_i \tag{12}$$

$$\overline{\text{PSNR}}_{\text{norm}} = \frac{1}{N}\sum_{i=1}^{N} \text{PSNR}_{\text{norm},i} \tag{13}$$

In this study, the fitness function is defined in Equation 14 to minimize the difference between the average robustness and imperceptibility metrics across all attack scenarios

$$\text{fit}(k, \alpha) = \frac{1}{N}\sum_{i=1}^{N} |\text{NC}_i + \text{invBER}_i - \text{SSIM}_i - \text{normalizedPSNR}_i| \tag{14}$$

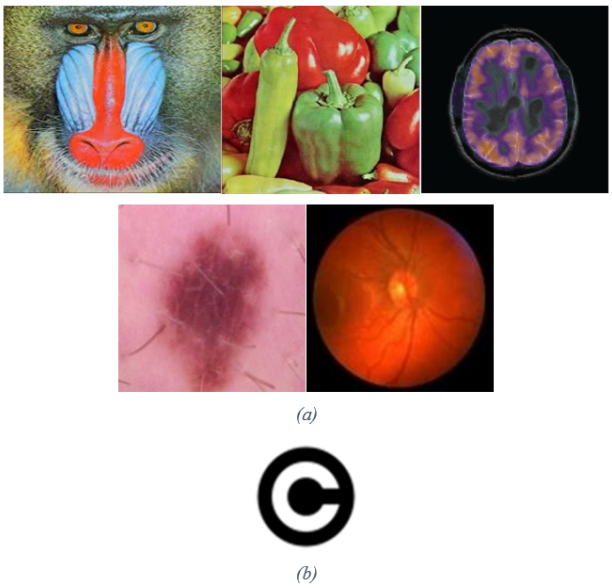
The primary objective of this formulation is to ensure that the terms representing robustness, $NC_i + invBER_i$ and those representing imperceptibility, $SSIM_i + normalizedPSNR_i$ are as close to each other as possible. The optimization problem under the defined fitness function and constraints is expressed in Equation 15 as follows.

$$(k^*, \alpha^*) = \arg\min_{k,\alpha} \text{fit}(k, \alpha) \tag{15}$$

Through this structure, the optimization process is directed toward producing a balanced solution between robustness and imperceptibility, rather than excessively optimizing a single metric.

EXPERIMENTAL RESULTS

To evaluate the performance of the proposed watermarking algorithm, both medical and standard benchmark color images were employed as cover images. As illustrated in Figure 1(a), the selected cover images include Melanoma, Cerebral, Retinal Fundus, Baboon, and Peppers, which are widely used benchmark images in digital watermarking studies. All cover images have a spatial resolution of 512 × 512 pixels.



(a)

(b)

Figure 8 (a) Color Cover Images [Baboon, Peppers, Cerebral, Melanoma, Retinal Fundus] (b) Watermark [copyright]

Table 1 presents the optimal k parameter and the optimal α values obtained for different cover images and watermark sizes. When all cover images are considered, it is observed that the optimal k value decreases as the watermark size becomes smaller. This indicates that the rank parameter is directly dependent on the watermark dimensions. The optimal α parameter, namely the watermark embedding coefficient, shows an increasing trend as the watermark size decreases. This result indicates that a stronger embedding operation is required to maintain robustness when smaller watermarks are used. The obtained findings clearly demonstrate the drawbacks of using fixed parameter values and highlight the necessity of optimization-based parameter selection.

To evaluate the performance of the proposed DWT–SVD-based watermarking method using the optimized parameters, NC, BER, SSIM, and PSNR metrics were employed. The performance evaluation results obtained for different watermark sizes are presented in Table 1.

When the values presented in Table 2 are examined, it is observed that NC $\geq$ 0.97 is achieved for different cover images and watermark sizes. The BER values are very close to zero (with a maximum BER of 0.0070), indicating that the watermark can be extracted with a negligible level of error. These results demonstrate that the proposed method provides high watermark extractability across different cover images and watermark sizes. Moreover, the high NC and low BER values obtained particularly for medical images further support the suitability of the proposed method for applications involving sensitive data.

Secondly, when evaluated in terms of watermark size, a clear increase in PSNR values is observed as the watermark size decreases. The fact that all obtained PSNR values are above 37 dB, which is widely accepted in the literature as the threshold for high perceptual quality and imperceptibility, indicates that the proposed method achieves satisfactory imperceptibility performance. This result shows that smaller watermarks introduce less distortion to the cover image and confirms the effectiveness of the proposed approach in terms of imperceptibility. An examination of the SSIM values reveals that they remain at approximately 0.999 throughout all experiments. Variations in either the cover image or the

Table 1 Optimized parameter values (k and α) for different watermark sizes and cover images

Cover Image	Watermark	Opt. k	Opt. α
Melanoma	256×256	182	0.0346
	128×128	80	0.0305
	64×64	42	0.0202
Cerebral	256×256	154	0.0173
	128×128	72	0.0183
	64×64	30	0.0284
Retinal Fundus	256×256	122	0.0077
	128×128	56	0.0137
	64×64	46	0.0229
Baboon	256×256	126	0.0090
	128×128	76	0.0120
	64×64	38	0.0115
Peppers	256×256	142	0.0206
	128×128	72	0.0181
	64×64	46	0.0219

Table 2 Performans evaluation for different watermark sizes

Cover Image	Watermark	NC	BER	PSNR	SSIM
Melanoma	256×256	0.9962	0.0006	37.042	0.9997
	128×128	0.9991	0.0022	44.064	0.9999
	64×64	0.9820	0.0034	51.440	0.9999
Cerebral	256×256	0.9994	0.0002	43.043	0.9994
	128×128	0.9990	0.0016	48.507	0.9998
	64×64	0.9962	0.0022	50.517	0.9999
Retinal Fundus	256×256	0.9893	0.0017	49.621	0.9996
	128×128	0.9964	0.0025	50.251	0.9997
	64×64	0.9963	0.0009	51.323	0.9998
Baboon	256×256	0.9983	0.0003	48.255	0.9999
	128×128	0.9923	0.0029	51.175	0.9999
	64×64	0.9700	0.0070	59.587	0.9999
Peppers	256×256	0.9995	0.0002	41.493	0.9996
	128×128	0.9979	0.0017	48.767	0.9999
	64×64	0.9964	0.0034	51.275	0.9999

watermark size do not lead to a significant decrease in SSIM values.

This finding indicates that the proposed method preserves not only pixel-level similarity but also the structural integrity of the image. Since maintaining structural integrity is of critical impor-

Table 3 NC Values obtained against different attacks (watermark: 64x64)

Type of Attacks	Noise Density	Retinal Fundus	Baboon	Peppers
No Attack	–	0.9963	0.9700	0.9964
Gaussian low-pass filter	3×3	0.9625	0.8089	0.9313
Median	3×3	0.9912	0.8978	0.9669
Rescaling	0.25–4	0.9486	0.8083	0.9337
Gaussian noise	0.001	0.9053	0.8809	0.8306
Salt and pepper noise	0.001	0.9180	0.9448	0.9727
Speckle noise	0.001	0.9797	0.9452	0.9767
JPEG compression	50	0.9323	0.8242	0.9503
JPEG2000 compression	12	0.9951	0.9072	0.9879
Sharpening attack	0.8	0.9575	0.8370	0.9203
Average filter	3×3	0.9626	0.8087	0.9404

tance in medical imaging applications, the obtained results further support the applicability of the proposed method to such sensitive use cases.

Table 3 presents the robustness evaluation of the proposed method against different types of attacks using Retinal Fundus, Baboon, and Peppers as cover images when the watermark size is fixed at 64×64. The robustness performance is assessed in terms of the NC metric.

The conducted attack analyses demonstrate that the proposed DWT–SVD-based watermarking method exhibits a high level of robustness against various attack types. Owing to its frequency-domain characteristics, the method successfully preserves NC values, particularly under JPEG2000 compression, speckle noise, and salt-and-pepper noise attacks. On the other hand, Gaussian noise and geometric rescaling attacks are observed to have a more limiting effect on performance, especially for images with high texture complexity, such as Baboon. Nevertheless, even under these challenging attack conditions, the obtained NC values remain at acceptable levels, indicating that the embedded watermark can still be reliably extracted.

Overall, the experimental results indicate that the proposed watermarking method achieves high robustness and imperceptibility across different cover images and watermark sizes. As the watermark size decreases, an improvement in perceptual image quality is observed. The fact that all obtained values remain above the thresholds commonly accepted in the literature confirms that the proposed approach provides a stable, reliable, and generalizable performance under a wide range of attack scenarios.

CONCLUSION

In this study, a hybrid digital watermarking method based on dimensionality reduction and optimization is proposed to ensure copyright protection and data security for both color and medical images. In the proposed approach, DWT and SVD are jointly applied only to the B channel of the color host image, while the watermark image is preprocessed through dimensionality reduction. The optimal embedding parameters are determined using the GWO algorithm. In this way, a balanced trade-off between imperceptibility and robustness is achieved.

Experimental studies were conducted on both standard test images and medical images under various watermark sizes and different attack scenarios. The obtained results demonstrate that the proposed method provides high imperceptibility under all test conditions. The fact that PSNR values remain above the threshold commonly accepted in the literature across all experiments, and that SSIM values are consistently around 0.999, indicates that the watermarking process has a negligible effect on image quality and structural integrity. This is particularly important for medical imaging applications, where preserving visual fidelity and structural information is critical. Robustness analyses reveal that the proposed method exhibits strong resistance against both classical signal processing attacks and compression- and noise-based attacks. The NC values obtained under different attack types are mostly above 0.90, confirming that the embedded watermark can be reliably extracted. In addition, BER values remaining close to zero clearly demonstrate the effectiveness of the frequency-domain-based structure of the method as well as the optimization process.

Furthermore, the analysis of optimized parameters for different watermark sizes shows that the use of fixed parameters may be insufficient in terms of performance. Thanks to the GWO-based optimization approach, the optimal rank value and embedding strength are automatically determined for each host image and watermark size, significantly enhancing the generalizability and adaptability of the method. The obtained findings support the importance of optimization-based approaches in improving the performance of digital watermarking systems. In conclusion, the proposed digital watermarking method offers high imperceptibility, strong robustness, and balanced performance characteristics. It provides an effective and reliable solution for application areas such as medical image security, copyright protection, and the prevention of unauthorized use of sensitive data. In future work, the proposed method will be evaluated in different color spaces (e.g., YCbCr, HSV), integrated with deep learning-based optimization approaches, and improved to reduce computational cost for real-time systems.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Agarwal, N. and P. K. Singh, 2022 Discrete cosine transforms and genetic algorithm based watermarking method for robustness and imperceptibility of color images for intelligent multimedia applications. Multimedia tools and applications **81**: 19751–19777.

Ahmadi, S. B. B., G. Zhang, M. Rabbani, L. Boukela, and H. Jelodar, 2021 An intelligent and blind dual color image watermarking for authentication and copyright protection. Applied Intelligence **51**: 1701–1732.

Anand, A. and A. K. Singh, 2021 Watermarking techniques for medical data authentication: a survey. Multimedia Tools and Applications **80**: 30165–30197.

Awasthi, D., A. Tiwari, P. Khare, and V. K. Srivastava, 2024 A comprehensive review on optimization-based image watermarking techniques for copyright protection. Expert Systems with Applications **242**: 122830.

Begum, M. and M. S. Uddin, 2020 Digital image watermarking techniques: a review. Information **11**: 110.

Çelik, H. and N. Doğan, 2021 K-en az anlamlı bitlere dayalı kaotik bir harita kullanan renkli görüntü steganografisi. Politeknik Dergisi **26**: 679–692.

Çelik, S. and N. Yalçın, 2023 Kriptografi ve görüntü steganografi tabanlı bir veri gizleme uygulaması: Sten 0.1. Bilgisayar Bilimleri ve Teknolojileri Dergisi **4**: 56–66.

Çetinkaya, Z. and F. Horasan, 2025 An optimized watermarking technique for medical images using dimension reduction. In *2025 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, pp. 1–6, IEEE.

Eltoukhy, M. M., A. E. Khedr, M. M. Abdel-Aziz, and K. M. Hosny, 2023 Robust watermarking method for securing color medical images using slant-svd-qft transforms and otp encryption. Alexandria Engineering Journal **78**: 517–529.

Gull, S. and S. A. Parah, 2024 Advances in medical image watermarking: a state of the art review. Multimedia Tools and Applications **83**: 1407–1447.

Horasan, F., H. Erbay, F. Varçın, and E. Deniz, 2019 Alternate low-rank matrix approximation in latent semantic analysis. Scientific Programming **2019**: 1095643.

Hu, H.-T., 2024 Synergistic compensation for rgb-based blind color image watermarking to withstand jpeg compression. Journal of Information Security and Applications **80**: 103673.

Karmouni, H., M. A. Tahiri, I. Dagal, H. Amakdouf, M. O. Jamil, *et al.*, 2024 Secure and optimized satellite image sharing based on chaotic eπ map and racah moments. Expert Systems with Applications **236**: 121247.

Khare, P. and V. K. Srivastava, 2021 A secured and robust medical image watermarking approach for protecting integrity of medical images. Transactions on Emerging Telecommunications Technologies **32**: e3918.

Li, Y., J. Li, U. A. Bhatti, J. Ma, D. Li, *et al.*, 2023 Robust zero-watermarking algorithm for medical images based on orb and dct. In *2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, pp. 282–289, IEEE.

Mahto, D. K. and A. K. Singh, 2021 A survey of color image watermarking: State-of-the-art and research directions. Computers & Electrical Engineering **93**: 107255.

Mamuti, M., 2019 *Tibbi Görüntü Güvenliği İçin Yeni bir Sayisal Damgalama Yöntemi*. Master's thesis, Sakarya Universitesi (Turkey).

Melman, A. and O. Evsutin, 2023 Image data hiding schemes based on metaheuristic optimization: a review. Artificial Intelligence Review **56**: 15375–15447.

Mirjalili, S., 2015 How effective is the grey wolf optimizer in training multi-layer perceptrons. Applied intelligence **43**: 150–161.

Mohanarathinam, A., S. Kamalraj, G. Prasanna Venkatesan, R. V. Ravi, and C. Manikandababu, 2020 Retracted article: Digital watermarking techniques for image security: a review. Journal of Ambient Intelligence and Humanized Computing **11**: 3221–3229.

Mousavi, S. M., A. Naghsh, and S. Abu-Bakar, 2014 Watermarking techniques used in medical images: a survey. Journal of digital imaging **27**: 714–729.

Nawaz, S. A., J. Li, D. Li, M. U. Shoukat, U. A. Bhatti, *et al.*, 2025 Medical image zero watermarking algorithm based on dual-tree complex wavelet transform, alexnet and discrete cosine transform. Applied Soft Computing **169**: 112556.

Othman, G. and D. Q. Zeebaree, 2020 The applications of discrete wavelet transform in image processing: A review. Journal of Soft Computing and Data Mining **1**: 31–43.

Priyadarshini, P. and K. Naik, 2024 Privacy protection and authentication of electronic patient information using hashing and multi watermarking technique. Multimedia Tools and Applications **83**: 89893–89930.

Qasim, A. F., F. Meziane, and R. Aspin, 2018 Digital watermarking: Applicability for developing trust in medical imaging workflows state of the art review. Computer Science Review **27**: 45–60.

Rodríguez, L., O. Castillo, J. Soria, P. Melin, F. Valdez, *et al.*, 2017 A fuzzy hierarchical operator in the grey wolf optimizer algorithm. Applied Soft Computing **57**: 315–328.

Roy, S. and A. K. Pal, 2019 A hybrid domain color image watermarking based on dwt–svd. Iranian Journal of Science and Technology, Transactions of Electrical Engineering **43**: 201–217.

Sahir, M., T. Bekkouche, F. Belilita, and N. Amardjia, 2025 An optimized color image watermarking scheme based on hd and svd in dwt domain. Engineering, Technology & Applied Science Research **15**: 21639–21646.

Seyedali, M. and L. Andrew, 2014 Grey wolf optimizer advances engineering software. Elsevier .

Sharma, S., H. Sharma, J. B. Sharma, and R. C. Poonia, 2023 A secure and robust color image watermarking using nature-inspired intelligence. Neural Computing and Applications **35**: 4919–4937.

Sharma, S., J. J. Zou, G. Fang, P. Shukla, and W. Cai, 2024 A review of image watermarking for identity protection and verification. Multimedia Tools and Applications **83**: 31829–31891.

Singh, R., M. Saraswat, A. Ashok, H. Mittal, A. Tripathi, *et al.*, 2023 From classical to soft computing based watermarking techniques: A comprehensive review. Future Generation Computer Systems **141**: 738–754.

Su, Q. and B. Chen, 2018 Robust color image watermarking technique in the spatial domain. Soft Computing **22**: 91–106.

Su, Q., F. Hu, X. Tian, L. Su, and S. Cao, 2024 A fusion-domain intelligent blind color image watermarking scheme using graph-based transform. Optics & Laser Technology **177**: 111191.

Wang, D., F. Yang, and H. Zhang, 2016 Blind color image watermarking based on dwt and lu decomposition. Journal of Information Processing Systems **12**: 765–778.

Wang, G., T. Jiang, D. Zhang, and V. Vasil'ev, 2025 Color image watermarking scheme based on singular value decomposition of split quaternion matrices. Journal of the Franklin Institute **362**: 107508.

Wang, H., Z. Yuan, S. Chen, and Q. Su, 2023 Embedding color watermark image to color host image based on 2d-dct. Optik **274**: 170585.

Wang, X.-y., C.-p. Wang, H.-y. Yang, and P.-p. Niu, 2013 A robust blind color image watermarking in quaternion fourier transform domain. Journal of Systems and Software **86**: 255–277.

Yağcıoğlu, H. and A. Sondaş, 2021 Çoklu görsel nesnelere veri gizleme (steganografi). Kocaeli Üniversitesi Fen Bilimleri Dergisi **4**: 1–5.

Yurttakal, A. H. and F. Horasan, 2022 Kesik tekil değer ayrışımı ve ayrık dalgacık dönüşümü kullanılarak boyut indirgeme tabanlı dayanıklı dijital görüntü damgalama. Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi **22**: 761–768.

Zainol, Z., J. S. Teh, M. Alawida, A. Alabdulatif, *et al.*, 2021 Hybrid svd-based image watermarking schemes: a review. IEEE Access **9**: 32931–32968.

How to cite this article: Özmen, İ., Çetinkaya, Z., Horasan, F., Varçın, F., and He, S. Adaptive–Scaled Digital Watermarking in Color Medical Imaging. *Computers and Electronics in Medicine*, 3(1), 99-108, 2026.