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

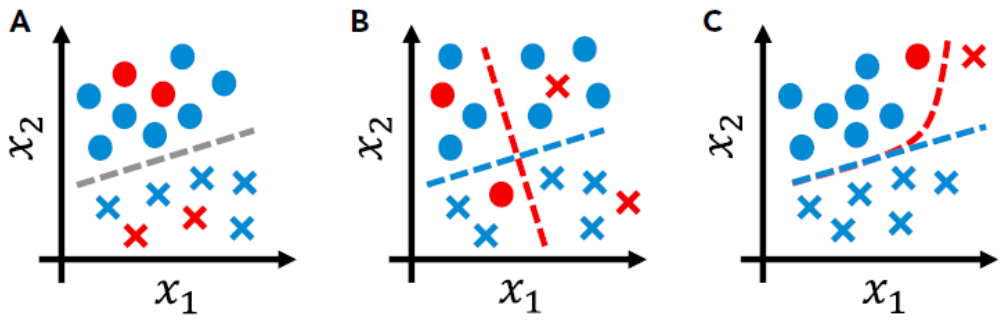


Figure 1 Cases of binary classification for creation of different decision boundaries. Blue circles (majority) and red crosses (minority) represent two patient groups of interest (Petersen *et al.* 2023; Raza *et al.* 2024).

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

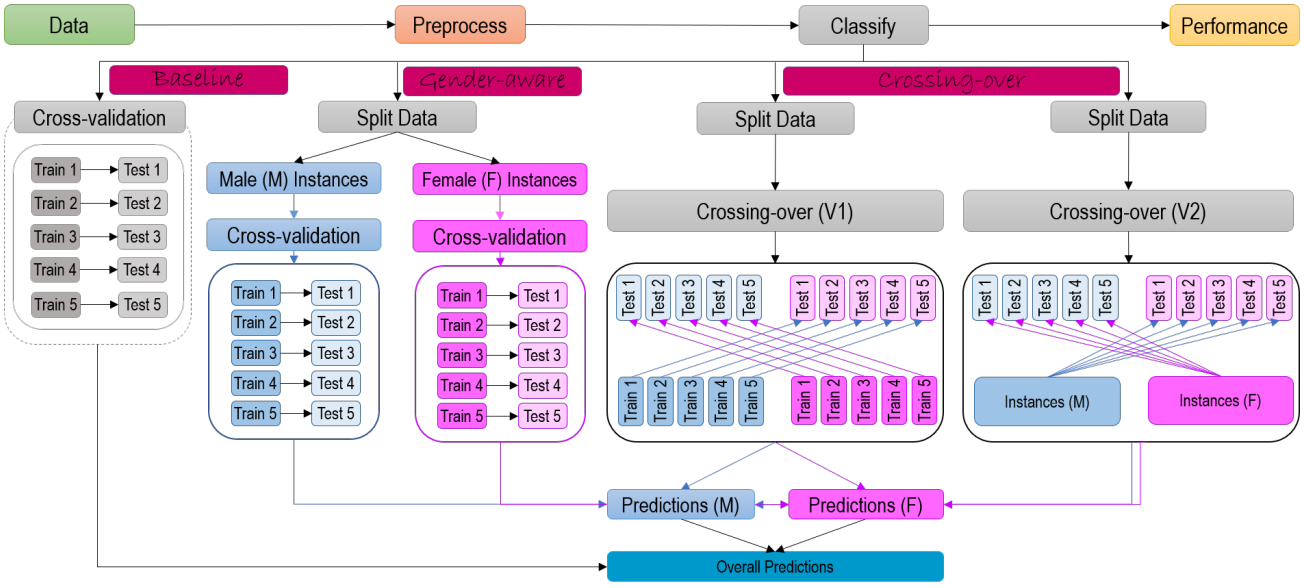


Figure 2 Flowchart of our evaluation that relies on four different pipelines.

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

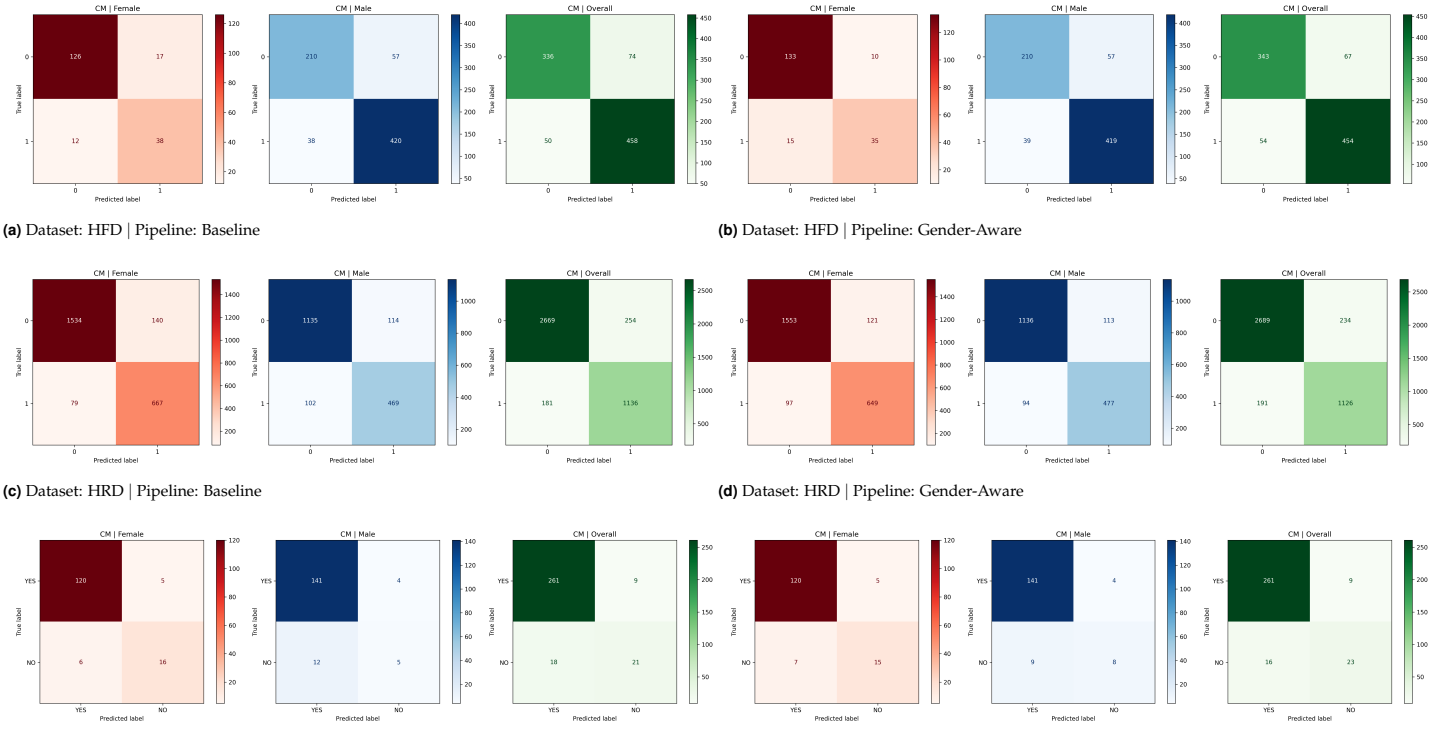
Table 2 Weighted average F1, TPR, and TNR values considering four pipelines on nine datasets across different instance groups

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.