ISSN: 3108-4060

ADBA

ARTIFICIAL INTELLIGENCE IN APPLIED SCIENCES

VOLUME 2, ISSUE 1, JANUARY 2026

AN INTERDISCIPLINARY JOURNAL OF COMPUTER SCIENCE

https://journals.adbascientific.com/aiapp

ISSN: 3108-4060

Artificial Intelligence in Applied Sciences

Volume: 2 – Issue No: 1 (January 2025)

EDITORIAL BOARD

ISSN: 3108-4060

Artificial Intelligence in Applied Sciences

Volume: 2 – Issue No: 1 (January 2026)

CONTENTS

https://journals.adbascientific.com/aiapp/issue/view/20

ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.1-7
https://doi.org/10.69882/adba.ai.2026011

A Genetic Algorithm-Based Traffic Light Optimization Model for Efficient Home Healthcare Service Delivery in Türkiye

Mehmet Aziz Çakmak [*,1], **Emrah Aslan** [α,2], **Yılmaz Demirhan** [β,3], **Fevzi Esen** [§,4] **and Fatma Kantaş Yılmaz** [§,5]

[*]Mardin Artuklu University, Technology Transfer Office, Mardin, Türkiye, [α]Mardin Artuklu University, Faculty of Engineering and Architecture, Mardin, Türkiye, [β]Mardin Artuklu University, Faculty of Economics and Administrative Sciences, Mardin, Türkiye, [§]University of Health Sciences, Department of Health Information Systems, Istanbul, Türkiye.

ABSTRACT Home healthcare (HHC) has become a crucial service model to address the rising needs of aging populations and patients with chronic conditions. However, efficient planning and resource allocation remain major challenges, especially in geographically dispersed regions. This study proposes a novel optimization-based operational model incorporating a traffic light algorithm to prioritize patient visits based on health status in Diyarbakır, Türkiye. The algorithm classifies patients into three categories (green, yellow, and red) allowing proactive and dynamic care management. A genetic algorithm is applied to solve the complex multi-objective routing and scheduling problem while considering numerous real-world constraints such as minimum team size, gender composition, and vehicle capacity. The model integrates demographic data from 2011–2023 and minimizes total visit duration while maximizing the number of patients served. Key decision variables include team size, staff gender distribution, patient condition, location, and travel time. The optimization process demonstrates significant improvements in performance metrics across generations, reducing penalty values and achieving more balanced, efficient outcomes. Results indicate that the model effectively aligns healthcare delivery with patient needs, operational limitations, and service quality goals. Unlike previous studies focusing mainly on cost or time, this model uniquely emphasizes clinical prioritization through color-coded patient conditions, integrating cultural and practical constraints. The study highlights the importance of tailored, region-specific solutions and offers a framework that can be adapted for broader applications. Future work should explore integrating machine learning for dynamic risk scoring and incorporating logistical elements such as traffic and real-time availability.

KEYWORDS
Traffic light algorithm
Health informatics
Optimization
Home healthcare services

INTRODUCTION

Home health care (HHC) has expanded rapidly in recent years as an alternative to hospital care in many countries due to aging populations and limited healthcare resources (TÜİK 2024). The global population aging is causing a rise in chronic diseases, ill health, and dependence, particularly among the elderly. This is a significant challenge for healthcare delivery systems due to rising healthcare and long-term care expenditures, necessitating alternative care options to address the unique needs of the elderly and their families (World Health Organization 2015; Çınar *et al.* 2025).

Home care refers to professional care delivered to individuals in their residences, aiming to enhance their quality of life and functional health status while substituting hospital care for societal considerations. It encompasses a broad spectrum of activities, ranging from preventive visits to end-of-life care (Genet *et al.* 2013; Pacal and Cakmak 2025). It involves providing medical supplies and services directly to patients inside the community, targeting many illnesses and therapy areas. Services may encompass medical, psychological, or social evaluations, wound care, medication education, pain management, illness information, physical therapy, speech therapy, medication reminders, and health promotion and prevention empowerment. Home health care is frequently more cost-efficient, convenient, and equally effective as care provided in a healthcare facility. It alleviates the burden on family members serving as caregivers and represents the most economical method to enhance access to primary healthcare services (Özüpak 2025; Cakmak *et al.* 2026). In 2020, 3 million patients received home healthcare services in the U.S.; of the 11.400 home health agencies registered, about 83.5% were classified for profit (U.S. Centers for Disease Control and Prevention 2024).

Home healthcare services require careful planning and organization to allocate nurses, schedule working hours, and manage travel routes. However, manual routing and scheduling often lead to suboptimal outcomes (Fikar and Hirsch 2017). A survey in two

Manuscript received: 16 November 2025,
Revised: 8 December 2025,
Accepted: 5 January 2026.

[1]mehmetazizcakmak21@gmail.com (**Corresponding author**).
[2]emrahaslan@artuklu.edu.tr
[3]yilmazdemirhan@artuklu.edu.tr
[4]fevzi.esen@sbu.edu.tr
[5]fatmakantas.yilmaz@sbu.edu.tr

Norwegian municipalities found that driving time accounts for 18-26% of total working hours, with an overestimation of routes (Holm and Angelsen 2014). Numerous municipalities seek cost-effective solutions to ensure home care maintains adequate quality while remaining affordable (Holm and Angelsen 2014).

In recent years, HHC routing and scheduling challenges have gained significant attention (Fikar and Hirsch 2017). Therefore, designing effective home health care routing and scheduling management is essential to alleviate the conflict between high-quality home health care and limited resources. Planning home health services involves multiple variables, including travel time to clients (parking, entering and exiting the home, supplies), visit duration (including documentation), alignment of nurses' skills with patients' expectations (medical, language, and social skills), continuity of care, staff workload balance, time sensitivity (e.g., timely insulin injections), visit sequencing, planning timelines (ranging from one day to several months), and cost assessments (Fikar and Hirsch 2017; Holm and Angelsen 2014; Yalçındağ *et al.* 2016). Consequently, numerous researchers focus on a multi-objective home healthcare routing and scheduling problem defined by conflicting objectives: reducing routing costs while improving service consistency and balancing workloads. Home healthcare managers must develop effective route plans for caregivers to provide in-person care to clients. Geographically dispersed customers must be considered, daily routes for caregivers must be determined, and the planned routes must deliver services. Home healthcare administrators prioritize reducing operational costs in formulating route plans (Trautsamwieser and Hirsch 2011).

Numerous studies have concentrated on the modeling and optimization of distributed flexible job shop scheduling problems within different systems (Luo *et al.* 2022; Du *et al.* 2022). However, studies on these problems assume that machines are always available (Xie *et al.* 2023; Zhang *et al.* 2024) and neglect the transportation activities of jobs. The methods used are diverse, paralleling the studied problem settings, and encompassing various population-based algorithms alongside local search-based procedures. Most researchers address meta-heuristic solution procedures for single-period home health care problems. The home healthcare worker scheduling challenge is complicated since it includes both the hard vehicle routing and personnel assignment issues (Mutingi and Mbohwa 2014). Koeleman *et al.* (2012) used the Markov decision process, which leads to a high-dimensional control problem. Castillo *et al.* (2024) utilized an agile algorithm to optimize route planning for providing home healthcare in Spanish rural areas. Belhor *et al.* (2023) utilizes a hybrid algorithm to enhance the routing of in-home healthcare services.

Traffic light visualizations might enhance clinical decision-making by leveraging the proven correlation between colors and corresponding therapeutic signals, which have been utilized across various therapeutic domains (Saposnik 2020). The implementation of traffic light coding for patient management has been effective in an emergency room context, where a three-tier urgency code facilitated the prioritization of patients for care (Leppäniemi and Jousela 2014; Araújo *et al.* 2021). Despite the significant importance of patient and caregiver's satisfaction, limited literature on this topic underscores the need to create a multi-objective routing and scheduling model that incorporates the interests of various stakeholders in home healthcare services (Wirnitzer *et al.* 2016). This study seeks to formulate optimization-based techniques for visit planning and vehicle routing in home healthcare services, specifically in Diyarbakır province, southeastern Türkiye.

MATERIALS AND METHODS

The section outlines the research framework employed to develop an operational plan for home healthcare services in Diyarbakır, Turkey, from 2011 to 2023. This section integrates demographic data, system variables, constraints, and a mathematical framework to optimize patient visits while minimizing visit duration. The methodology is structured into three key components: the dataset, which provides demographic and operational context; the mathematical model, which formalizes the optimization problem; and the proposed model, which details the operational strategy and constraints for service delivery. This structured approach ensures a comprehensive analysis of the home healthcare system, prioritizing patient satisfaction and operational efficiency.

Dataset

The research leverages demographic data from the Turkish Statistical Institute (TUIK) Report of 2024, which indicates that Diyarbakır province had a population of 1,818,133 in 2024, with a growth rate of 0.73% and a population density of 119 individuals per square kilometer. The demographic composition reveals that individuals aged 5-9 constitute 11.15% of the population, while older age groups show a significant decline: those aged 60-64 represent 2.60%, 65-69 account for 1.86%, 70-74 comprise 1.44%, 75-79 make up 0.94%, 80-84 constitute 0.56%, and those aged 90 and above represent 0.18%. The research population encompasses individuals receiving home healthcare services in Diyarbakır from 2011 to 2023. This retrospective study utilizes operational data from this period, focusing on patient visits and service delivery protocols.

This paper presents an operational plan for home healthcare services, a distributed healthcare system. Figure 1 illustrates home healthcare services registered in Turkey. It shows a visit protocol for patients who have received and/or are currently receiving services from 2011 to 2023. The research utilized a retrospective design.

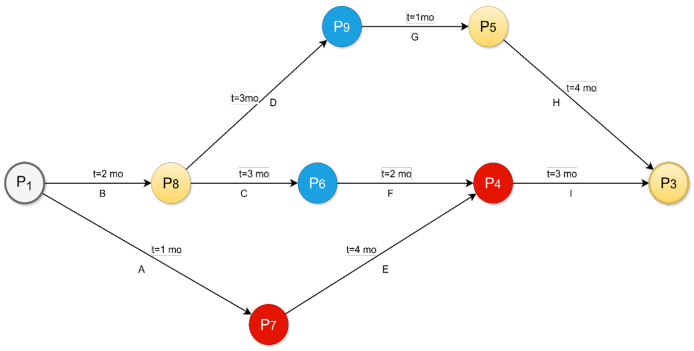


Figure 1 Home Healthcare Services Diagram

Fig 1 shows the location of the patients (P), the time between locations (t) and the possible paths (A-I) to be used during the visit. Fig 1 is a small virtual demonstration of the considered optimization problem. Teams or groups located within a healthcare facility deliver home healthcare services to individuals residing at designated addresses, whose overall health conditions differ. The variables of the system depicted in Figure 1 are as follows:

- Number of teams
- Patient's gender, Patient's overall condition (Green/Yellow/Red)
- Quantity of vehicles
- Number of male employees

- Number of female employees
- Locations of patients

The specified variables need the consideration of the following constraints in the delivery of home healthcare services:

- Teams departing from the health center must return upon the conclusion of their visits.
- Each team must comprise a minimum of four individuals.
- Each team is required to include a minimum of one female and one male staff member.
- Teams' ought to evaluate the patient's comprehensive condition during the appointment.
- Patients with like conditions ought to be attended to base on their proximity.

The traffic light-based patient classification system, adapted from established emergency triage protocols, categorizes patients according to the urgency of their health condition:

- Red: Patients with life-threatening conditions requiring immediate intervention.
- Yellow: Patients with serious but non-life-threatening conditions that can tolerate a short delay.
- Green: Patients with stable, minor conditions that can wait longer without risk of deterioration.

This classification enables proactive prioritization, ensuring critical patients are visited first while optimizing resource allocation.

Considering the specified variables and limits, the minimum visit duration must be achieved. In this setting, patient visits should occur at the most appropriate moment. The primary objective is to enhance patient satisfaction with the service. Table 1 presents the variables associated with the optimization problem modeled in the context of Home Health Services.

Mathematical framework

The mathematical framework aims to maximize patient visits while minimizing visit duration, contingent upon variables such as the patient's overall condition, location, number of teams, vehicles, and personnel. The objective function is designed to identify the optimal route and minimal visit duration for each team. Key variables are defined in Table 2, including the number of teams (E), number of patients (H), number of vehicles (C), number of male (Pm) and female (Pf) staff, patient location (Li), patient condition (Si), travel time (Tij), visit status (xek), and presence of female (yef) and male (yem) staff. The cost function incorporates these decision variables to optimize service delivery.

The objective function seeks to maximize patient visits while minimizing visit duration. The duration of the visit is contingent upon the patient's overall condition, the patients' locations, the number of teams, the number of cars, the personnel count, and the teams' locations. A cost function can be formulated in which specific variables are contingent upon decision variables. Our objective is to identify the optimal route and the minimal visit length for each team. Representation of the objective function can be formulated as follows:

$$\text{Min } Z = \sum_{e=1}^{E} \sum_{i=1}^{H} \sum_{j=1}^{H} T_{ij} \cdot x_{ei} \cdot x_{ej} \tag{1}$$

Constraints

- Each team must comprise a minimum of 4 individuals.

$$\sum_{f=1}^{P_f} y_{ef} + \sum_{m=1}^{P_m} y_{em} \geq 4, \quad \forall e \in \{1,2,3,\ldots,E\} \tag{2}$$

- Each team is required to include a minimum of one female and one male staff,

$$\sum_{f=1}^{P_f} y_{ef} \geq 1, \quad \forall e \in \{1,2,3,\ldots,E\} \tag{3}$$

$$\sum_{m=1}^{P_m} y_{em} \geq 1, \quad \forall e \in \{1,2,3,\ldots,E\} \tag{4}$$

- Each patient shall be attended by no more than one team;

$$\sum_{e=1}^{E} x_{ek} \leq 1, \quad \forall k \in \{1,2,3,\ldots,H\} \tag{5}$$

- Teams should prioritize based on the overall condition of the patients,

$$x_{ek} \geq x_{el} \quad \text{if } S_k > S_l \text{ and } L_k \text{ near } L_l \tag{6}$$

- Teams must reconvene in the center upon the conclusion of their visits,

$$\sum_{i=1}^{H} x_{ei} = \sum_{j=1}^{H} x_{ej}, \quad \forall e \in \{1,2,3,\ldots,E\} \tag{7}$$

- Vehicle capacity constraint,

$$\sum_{k=1}^{H} x_{ek} \leq C, \quad \forall e \in \{1,2,3,\ldots,E\} \tag{8}$$

EXPERIMENTAL RESULTS FINDINGS

The study's findings indicate that the penalty levels in the first population are high and show considerable variation. The optimal penalty value is 510.1, whereas the mean penalty value is 735.8. This scenario suggests that the solution is anomalous and that the system has not attained an optimal resolution. Figure 2 indicates that the initial hints of progress are evident in generation 10. The optimal penalty value decreases to 440.0, whilst the mean penalty diminishes to 648.0. This reduction signifies that the model is generating superior answers and that the optimization process has improved in efficiency compared to the the initial stage.

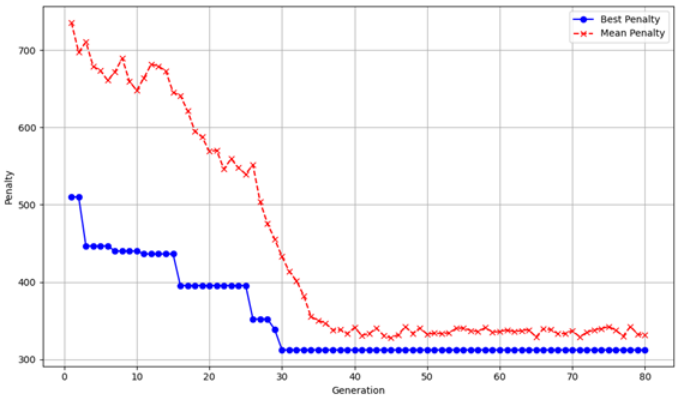


Figure 2 Optimization curve

Table 1 Key Variables Used in Home Health Services Optimization Model

Variable Name	Symbol	Definition
Number of teams	E	The total number of teams in a healthcare facility.
Number of patients	H	The total number of patients in the dataset.
Number of vehicles	C	The total number of vehicles owned by the teams.
Number of male staff	P_m	The total number of male staff in the team.
Number of female staff	P_f	The total number of female staff in the team.
Patient's location	L_i	The location of patient i.
The patient's general condition	S_i	The general condition of patient i.
Travel time	T_{ij}	Travel time from location i to location j.
Visit status	x_{ek}	The e-th team's visit status for patient k; a binary variable indicating whether team e visited patient k (1: visit, 0: no visit).
Presence of female staff	y_{ef}	Variable indicating whether female staff f is present in team e (1: present, 0: not present).
Presence of male staff	y_{em}	Variable indicating whether male staff m is present in team e (1: present, 0: not present).

In Generation 20, the penalty value decreased to 395.6, although the mean penalty was recorded at 569.1. This scenario suggests that an increase in iterations yields better solutions. Notably, from the 20th generation onward, there has been a considerable reduction in both the optimal and mean penalty values, suggesting that the model is more adeptly managing the constraints and yielding more appropriate solutions within the search space. In generation 29, the optimal penalty value declined to 338.6, while the mean penalty value fell to 455.5, signifying that the optimization process is progressing well and the model is yielding more precise solutions. The substantial reduction in the 29th generation signifies that the algorithm has transitioned into the exploitation phase and is generating solutions nearer to the optimal. In Generation 30, the penalty value was 312.2, however the mean penalty value was 433.0. This setting signifies that the model has achieved equilibrium during the optimization process, resulting in a stable optimal solution.

In the final stage, the optimal penalty value is maintained at 312.2, but the mean penalty value drops to 331.5. This condition signifies that the optimization procedure was executed with exceptional performance and that superior efficiency was attained in the model's final stage. The persistent decline in the average penalty value signifies that the solution quality has reached a stable high across the population. The data validate that the algorithm has enhanced both individual and collective solution efficacy, yielding significant efficiency. Table 2 illustrates that the quality of the solution in the optimization process enhances with advancing generations, ultimately stabilizing from the originally elevated penalty values.

To evaluate the performance of the proposed model, the random initial solution (Generation 1), where no optimization was applied, was taken as the baseline. The reduction of the best penalty value from an initial 510.1 to 312.2 in the 30th generation demonstrates that the model produces approximately 38.7% more efficient results compared to the baseline scenario.

Table 3 presents the optimized home healthcare service model's results, showing patient-to-team assignments and total visit duration for Diyarbakır (2011–2023). The sequence 1-2-3-3-1-1-1-2-3-3-1-2-3-1-1 assigns 15 patients to three teams: Team 1 (6 patients), Team 2 (3 patients), Team 3 (6 patients). This distribution optimizes patient condition prioritization, proximity, and constraints like minimum team size (four members, including one male and one female) and single-team visits. The total visit duration of 312.2146 (likely minutes) reflects minimized travel and visit times across 15 patients, averaging 20.81 minutes per visit. This efficiency, achieved via MATLAB-based optimization, aligns with the goal of enhancing patient satisfaction. The model effectively balances resource use and patient needs, but lacks baseline comparison, patient condition details, and time unit clarity. Future work should include these for robust validation.

Table 3 Solution of the new model

Definition	Assignment
Best individual (Assignment of patients to teams)	1-2-3-3-1-1-1-2-3-3-1-2-3-1-1
Best fit value (Total visit duration)	312.2146

The "Best fit value" (312.2146) presented in Table 3 represents the total visit and travel duration in minutes calculated by the model. Penalty values incorporate both the total operational time and theoretical cost points arising from violations of constraints, such as staff gender balance or vehicle capacity.

DISCUSSION

The planning and management of Home Health Services (HHS) is a multi-dimensional optimization problem that focuses on the effort to provide high quality care with limited resources. This study aims to maximize patient visits and minimize visit duration using an optimization model developed specifically for Diyarbakır province, with the aim of increasing patient satisfaction and optimizing operational efficiency in the process. The model priori-

Table 2 Optimization Output

Generation	Best Penalty	Mean Penalty	Note
1	510.1	735.8	Initial population; high penalty values, significant variance.
10	440.0	648.0	The initial indicators of improvement; a notable reduction in penalty values.
20	395.6	569.1	Substantial enhancement in solution quality; the team and restrictions are efficient.
29	338.6	455.5	A crucial phase of iterative improvement; a notable enhancement was noted.
30	312.2	433.0	The penalty value has stabilized; equilibrium has been attained in the optimization process.
80	312.2	331.5	Consistent reduction in average penalty value; exceptional performance in the model's final stage.

tized the health status of patients by classifying them into 'green', 'yellow' and 'red' categories, and planned team routes and staff assignments in an integrated manner using a genetic algorithm-based approach. This section compares the results with similar studies in the literature, discusses the innovative aspects and limitations of the model, and provides recommendations for future work. Table 4 provides detailed information on further studies and summarizes important details from the literature.

The proposed model significantly outperforms previous approaches by reducing travel distance, time, cost, and waiting time while improving fairness and health condition prioritization. Its hybrid structure, integrating a genetic algorithm with a traffic light–based prioritization system, ensures both operational efficiency and patient-centered optimization.

TD: Travel Distances, TT: Travel Time, TC: Travel Cost, WA: Waiting Time, PN: Personal Number, HC: Health Condition, FA: Fairness, NoS: Number of Nurses

The main difference of our study is that it places patients' health states at the center of the optimization process. While many studies on HHS in the literature often focus on parameters such as distance, travel time or cost, (Fikar and Hirsch 2015), the prioritization of the dynamic health status of patients is often neglected. For example, Akjiratikarl *et al.* (2007) only considered travel time and staff preferences when optimizing staff scheduling with a particle swarm optimization (PSO) algorithm; no parameters for patient health status were included in the model. Similarly, Allaoua *et al.* (2013) optimized HHS by combining staff assignment and vehicle routing problems, but did not consider patient status as a variable. In contrast, our model provides a proactive approach to care by ensuring that patients in the 'yellow' category are visited more frequently before their condition becomes 'red'. While this shows a similar flexibility to the work of (Lanzarone and Matta 2014), which considers variability in patient demands, our approach offers a more intuitive and feasible prioritization through traffic light coding.

Another important contribution of the model is the consideration of realistic constraints on the composition of teams. The requirement that each team consists of at least four people and includes at least one female and one male staff member is an approach that reflects gender balance and team dynamics. This is one of the few studies to consider the social and cultural dimensions of HHS. In the literature, studies such as Allaoua *et al.* (2013) and Braekers *et al.* (2016) have integrated staff assignment and routing problems, but such specific constraints on team composition are often ignored. In this respect, our model has a framework that is more appropriate for real-world applications.

The use of the genetic algorithm in the optimization process has shown effective results in complex and multi-objective problems. The results show that the algorithm significantly reduces the penalty values as the generations progress: the best penalty value decreased from 510.1 in the first generation to 312.2 in the 30th generation and the average penalty value decreased to 331.5 in the final stage. This confirms that the algorithm is finding better solutions in the search space and managing the constraints effectively. A similar improvement was reported by (Akjiratikarl *et al.* 2007) using PSO; however, our model provided a more comprehensive optimization by taking into account the patient's health status. Furthermore, while our initial solution showed high and variable penalty values (average 735.8), the system reached a balanced and optimal solution as iterations progressed. This proves that the model improves both individual and collective solution quality.

The model developed in this study brings together several innovative elements of HHS optimization that have not been adequately addressed in the literature. First, traffic light coding (green, yellow, red) has facilitated patient prioritization by using the association of colors with therapeutic signals in clinical decision-making processes (Bredström and Rönnqvist 2008). This approach was inspired by triage systems in emergency departments Lanzarone and Matta (2014) and adapted to the HHS context, providing a proactive strategy in patient care. Second, the fact that the model simultaneously optimizes the objective of maximizing patient visits and minimizing time provides a multi-objective optimization framework. This represents a more balanced approach compared to single objective models in the literature (e.g. studies focusing only on cost reduction). Finally, testing with demographic data specific to a particular geographic region, such as Diyarbakır, demonstrates the adaptability of the model to local conditions.

The limitations of the model are that it may not be sufficient for real-world applications as it only classifies the health status of patients into three categories of "green", "yellow" and "red"; this simple classification, which does not take into account factors such as age, chronic diseases, etc., can be improved to more precise risk scores using machine learning; furthermore, the lack of practical factors such as vehicle capacity, traffic conditions and parking

Table 4 HHS optimization studies in the literature

Author(s)	Solution	TD	NoS	TT	TC	WA	PN	FA	HC
Akjiratikarl *et al.* (2007)	Min.	↓							
Allaoua *et al.* (2013)	Min.			↓				↓	
Braekers *et al.* (2016)	Max.								↑
Fikar and Hirsch (2015)	Min.				↓				
Bredström and Rönnqvist (2008)	Min.			↓		↓			
Lanzarone and Matta (2014)	Max.								↑
Proposed Model	Max/Min	↓	↓	↓	↓	↓		↑	

space in the model limits its applicability and the integration of these dynamics; finally, the general model can be improved to more precise risk scores using machine learning, This simple classification, which does not take into account factors such as age, chronic diseases, etc., can be improved to more accurate risk scores using machine learning; furthermore, the lack of practical factors such as vehicle capacity, traffic conditions, and parking space in the model limits its applicability and the integration of these dynamics is recommended; finally, the generalizability of the model developed with data specific to Diyarbakır is limited and needs to be tested in different regions.

This study presents an innovative approach to the optimization of HHS that focuses on the health status of patients. The model, supported by a genetic algorithm, aims to increase patient and caregiver satisfaction while considering operational efficiency and integrates patient prioritization, team composition and route planning. The results show that the model overcomes the initial suboptimal solutions (high penalty values) and achieves a stable and efficient optimization process. However, to address the limitations, future work could focus on more detailed patient classifications, modelling logistical factors, and testing the model in different regions. A comprehensive and flexible optimization framework developed in this direction could contribute to a more effective implementation of HHS on a global scale.

Although the proposed optimization framework was tested specifically for Diyarbakır, its modular structure allows it to be easily adapted to other provinces with different demographic data and geographic characteristics. By updating variables such as staff count and patient density, the parametric nature of the model proves its usability as a general tool for planning home healthcare services on both a national and international scale.

CONCLUSION

This study proposed an innovative optimization-based operational model for improving the efficiency and equity of home healthcare services in Diyarbakır, Türkiye. By integrating a genetic algorithm with a traffic light prioritization mechanism, the model effectively addressed the multi-objective challenge of minimizing visit durations while maximizing the number of patients served. The algorithm classified patients into green, yellow, and red categories, ensuring that care delivery dynamically aligned with patients' clinical urgency and operational constraints. The optimization process achieved significant performance improvements across generations, reducing penalty values from 735.8 to 331.5, which indicates enhanced model stability and convergence. The findings highlight that integrating demographic data, staff composition constraints, and patient condition prioritization into a unified optimization framework can substantially enhance service quality and resource utilization in home healthcare systems. Moreover, incorporating gender-balanced team structures add a realistic and culturally sensitive dimension to service delivery. While the model demonstrated robust optimization outcomes, future research should incorporate additional real-world parameters such as traffic dynamics, vehicle capacity, and real-time patient data alongside machine learning-based dynamic risk scoring to enhance adaptability. Ultimately, the proposed framework presents a promising foundation for the development of intelligent, region-specific, and patient-centered home healthcare planning systems applicable to diverse geographic and demographic contexts.

Acknowledgments

This work was supported by the Scientific Research Projects Coordination Unit of Mardin Artuklu University. The project number is MAU.BAP.25.MMF.038.

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

Akjiratikarl, C., P. Yenradee, and P. R. Drake, 2007 Pso-based algorithm for home care worker scheduling. Computers & Industrial Engineering **53**: 559–583.

Allaoua, H., S. Borne, L. Létocart, and R. W. Calvo, 2013 A matheuristic approach for solving a home health care problem. Electronic Notes in Discrete Mathematics **41**: 471–478.

Araújo, M., P. Van Dommelen, J. Srivastava, and E. Koledova, 2021 A data-driven intervention framework for improving adherence. In *Applying the FAIR Principles to Accelerate Health Research in Europe*, pp. 23–27, IOS Press.

Belhor, M., A. El-Amraoui, A. Jemai, and F. Delmotte, 2023 Multi-objective evolutionary approach based on k-means clustering for home health care. Expert Systems with Applications **213**: 119035.

Braekers, K., R. F. Hartl, S. N. Parragh, and F. Tricoire, 2016 A bi-objective home care scheduling problem. European Journal of Operational Research **248**: 428–443.

Bredström, D. and M. Rönnqvist, 2008 Combined vehicle routing and scheduling with synchronization constraints. European Journal of Operational Research **191**: 19–31.

Cakmak, Y., I. Pacal, *et al.*, 2026 A comparative analysis of transformer architectures for automated lung cancer detection in ct images. Journal of Intelligent Decision Making and Information Science **3**: 528–539.

Castillo, C., E. J. Alvarez-Palau, L. Calvet, J. Panadero, M. Viu-Roig, *et al.*, 2024 Home healthcare in spanish rural areas. Socio-Economic Planning Sciences **92**: 101828.

Du, Y., J. Li, C. Li, and P. Duan, 2022 A reinforcement learning approach for flexible job shop scheduling. IEEE Transactions on Neural Networks and Learning Systems **35**: 5695–5709.

Fikar, C. and P. Hirsch, 2015 A matheuristic for routing real-world home service transport systems. Journal of Cleaner Production **105**: 300–310.

Fikar, C. and P. Hirsch, 2017 Home health care routing and scheduling: A review. Computers & Operations Research **77**: 86–95.

Genet, N., W. Boerma, M. Kroneman, A. Hutchinson, and R. Saltman, 2013 *Home Care Across Europe: Case Studies*. World Health Organization.

Holm, S. G. and R. O. Angelsen, 2014 A descriptive retrospective study of time consumption in home care services. BMC Health Services Research **14**: 439.

Koeleman, P. M., S. Bhulai, and M. van Meersbergen, 2012 Optimal patient and personnel scheduling policies for care-at-home. European Journal of Operational Research **219**: 557–563.

Lanzarone, E. and A. Matta, 2014 Robust nurse-to-patient assignment in home care services. Operations Research for Health Care **3**: 48–58.

Leppäniemi, A. and I. Jousela, 2014 A traffic-light coding system to organize emergency surgery. British Journal of Surgery **101**: e134–e140.

Luo, Q., Q. Deng, G. Gong, X. Guo, and X. Liu, 2022 A distributed flexible job shop scheduling problem considering worker arrangement. Expert Systems with Applications **207**: 117984.

Mutingi, M. and C. Mbohwa, 2014 Multi-objective homecare worker scheduling. IIE Transactions on Healthcare Systems Engineering **4**: 209–216.

Pacal, I. and Y. Cakmak, 2025 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.

Saposnik, G. e. a., 2020 Effect of an educational intervention on therapeutic inertia. JAMA Network Open **3**: e2022227.

Trautsamwieser, A. and P. Hirsch, 2011 Optimization of daily scheduling for home health care services. Journal of Applied Operational Research **3**: 124–136.

TÜİK, 2024 Nüfus ve demografi. Online, Accessed: 16 Oct 2024.

U.S. Centers for Disease Control and Prevention, 2024 Home health care. Online.

Wirnitzer, J., I. Heckmann, A. Meyer, and S. Nickel, 2016 Patient-based nurse rostering in home care. Operations Research for Health Care **8**: 91–102.

World Health Organization, 2015 *The Growing Need for Home Health Care for the Elderly: Home Health Care for the Elderly as an Integral Part of Primary Health Care Services*. WHO Regional Office for the Eastern Mediterranean.

Xie, J., X. Li, L. Gao, and L. Gui, 2023 A hybrid genetic tabu search algorithm for distributed flexible job shop scheduling problems. Journal of Manufacturing Systems **71**: 82–94.

Yalçındağ, S., A. Matta, E. Şahin, and J. G. Shanthikumar, 2016 The patient assignment problem in home health care. Flexible Services and Manufacturing Journal **28**: 304–335.

Zhang, R., H. Yu, K. Gao, Y. Fu, and J. H. Kim, 2024 A q-learning based artificial bee colony algorithm for surgery scheduling. Swarm and Evolutionary Computation **90**: 101686.

Çınar, M., E. Aslan, and Y. Özüpak, 2025 Comparison and optimization of machine learning methods for fault detection in district heating and cooling systems. Bulletin of the Polish Academy of Sciences: Technical Sciences **73**: 1–9.

Özüpak, Y., 2025 Machine learning-based fault detection in transmission lines: A comparative study with random search optimization. Bulletin of the Polish Academy of Sciences: Technical Sciences **73**.

How to cite this article: Çakmak, M. A., Aslan, E., Demirhan, Y., Esen, F., and Yılmaz, F. K. A Genetic Algorithm-Based Traffic Light Optimization Model for Efficient Home Healthcare Service Delivery in Türkiye. *Artificial Intelligence in Applied Sciences*, 2(1), 1-7, 2026.

ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.8-14
https://doi.org/10.69882/adba.ai.2026012

Comparative Evaluation of Convolutional Neural Network Architectures for Automated Skin Cancer Classification: A Study on the ISIC 2018 Dataset

Sifeu Takougang Kingni [*,1] **and Said Baawain** [α,2]

*Department of Mechanical, Petroleum and Gas Engineering, National Advanced School of Mines and Petroleum Industries, University of Maroua, P.O. Box 46, Maroua, Cameroon, αDepartment of Mechanical Engineering, Military Technological College, Muscat, P.O. Box 262, PC 111, Sultanate of Oman.

ABSTRACT The increasing global incidence of skin cancer, particularly lethal malignant melanoma, necessitates the development of robust, automated diagnostic tools to assist dermatologists in identifying subtle pathological markers. In this study, we provide a rigorous comparative evaluation of four state-of-the-art convolutional neural network (CNN) architectures, ResNet-50, DenseNet-169, Inception-v3, and EfficientNet-B0, using the ISIC 2018 (HAM10000) dataset. Our standardized experimental pipeline utilized stratified sampling to address class imbalance, alongside a meticulous preprocessing strategy and data augmentation to ensure model generalization. Quantitative results demonstrate that EfficientNet-B0 outperformed other models, achieving a peak accuracy of 91.84% and a superior F1-Score of 0.8429, despite possessing the most compact parameter footprint of 4.02M. While ResNet-50 exhibited lower diagnostic precision, it offered the fastest inference speed (0.359 ms), highlighting a critical trade-off between accuracy and real-time operational latency. Furthermore, visual validation through Grad-CAM++ confirmed that successful predictions were driven by relevant morphological hallmarks rather than dataset artifacts. Our findings suggest that architectural optimization through compound scaling is more effective than raw model depth for dermatological tasks. Collectively, this work provides a comprehensive framework for selecting deep learning backbones for clinical triage, balancing high-precision diagnostic support with the computational constraints of real-world medical deployment.

KEYWORDS
Convolutional neural networks (CNNs)
Skin cancer classification
EfficientNet-B0
Grad-CAM++
HAM10000 dataset

INTRODUCTION

Skin cancer remains one of the most significant public health challenges globally, with its incidence rates climbing steadily over the past few decades. Among various types, malignant melanoma stands out as the most lethal form, yet it is highly curable if identified in its nascent stages (Leiter *et al.* 2020). However, the diagnostic process is inherently complex; dermatologists must distinguish between a wide array of look-alike lesions, often relying on dermoscopy to visualize sub-surface structures (Siegel *et al.* 2024). Despite the expertise of clinicians, the visual ambiguity of skin lesions, characterized by varying colors, textures, and irregular borders, introduces a level of subjectivity that can lead to diagnostic inconsistency. Consequently, there is an urgent clinical demand for objective, automated screening tools that can support early intervention and improve patient outcomes (Gloster and Neal 2006; Armstrong and Kricker 1995).

The advent of Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), has fundamentally transformed the landscape of medical image analysis (Pacal *et al.* 2024; Cakmak and Pacal 2025b; Cakmak and Maman 2025; Pacal and Cakmak 2025a). These architectures possess the unique ability to automatically learn hierarchical feature representations directly from raw pixel data, capturing intricate patterns that may be imperceptible to the human eye (Cakmak and Zeynalov 2025; Zeynalov *et al.* 2025; Pacal and Cakmak 2025b; Cakmak and Pacal 2025a). In the realm of dermatology, CNNs have shown remarkable potential in automating the classification of skin lesions across diverse diagnostic categories. By leveraging vast repositories of dermatoscopic images, such as the HAM10000 dataset, these models can be trained to recognize the subtle morphological hallmarks of malignancy with a level of precision that occasionally rivals or exceeds that of board-certified specialists (Chaurasia *et al.* 2025; Manju *et al.* 2025).

While literature is replete with various DL approaches, the selection of an optimal architecture remains a non-trivial task. Modern clinical environments require a delicate balance between high diagnostic accuracy and computational efficiency, especially when considering deployment on resource-constrained hardware or real-time diagnostic platforms. Architectures like ResNet have introduced residual learning to overcome the vanishing gradient problem, while DenseNet emphasizes feature reuse through dense connections (Karthik *et al.* 2024; Pacal *et al.* 2025). More recent

Manuscript received: 3 November 2025,
Revised: 28 December 2025,
Accepted: 20 January 2026.

[1]stkingni@gmail.com
[2]saidbaawain14@gmail.com(**Corresponding author**).

innovations, such as EfficientNet, have pushed the boundaries further by scaling depth, width, and resolution simultaneously to maximize performance while minimizing the parameter footprint. However, a comprehensive comparative analysis is necessary to determine how these varying design philosophies perform specifically on multi-class skin lesion datasets (Ozdemir and Pacal 2025).

In this study, we present a rigorous comparative evaluation of four state-of-the-art CNN architectures, ResNet-50, DenseNet-169, Inception-v3, and EfficientNet-B0, to identify the most robust backbone for skin cancer classification. Utilizing the ISIC 2018 (HAM10000) dataset, we implemented a standardized pipeline involving stratified sampling and consistent preprocessing to ensure a fair performance assessment. Beyond traditional accuracy metrics, our analysis delves into the trade-offs between model complexity (Params), computational load (Gflops), and real-world inference speed. By integrating these quantitative results with visual validation through Grad-CAM++, we aim to provide a holistic framework that not only identifies the most accurate model but also offers insights into its clinical reliability and interpretability.

RELATED WORKS

DL frameworks relying on CNNs continue to be refined through advanced optimization and ensemble strategies to enhance diagnostic precision. Farea *et al.* (2024) proposed a hybrid framework that addresses data scarcity by curating a generalized dataset from multiple sources and employing the Artificial Bee Colony (ABC) algorithm to optimize the initial weights of an Xception model. This approach aimed to mitigate the risk of local minima during training, ultimately achieving a high accuracy of 93.04% by effectively fine-tuning the learnable parameters on segmented lesion regions.

Similarly, efforts to maximize feature representation have led to the development of complex ensemble architectures. Akter et al. (2024) (r19 2024) introduced an integrated DL model that fuses the feature outputs of InceptionV3 and DenseNet121 using a weighted sum rule at the score level. Their methodology incorporated extensive data augmentation to resolve class imbalance, resulting in a robust system that achieved a detection accuracy of 92.27% on the ISIC dataset.

Addressing the "black-box" nature of deep learning, explainability has become a central focus alongside classification performance. Attallah (2024) developed "Skin-CAD," an explainable CAD system that aggregates features from multiple CNN layers and employs Principal Component Analysis (PCA) to reduce dimensionality before classification. This system not only classified lesions into seven subtypes with 97.2% accuracy but also integrated LIME (Local Interpretable Model-agnostic Explanations) to provide visual justifications for the model's predictions, thereby enhancing clinical trust.

Furthermore, lightweight CNN architecture remains vital for deployment in resource-constrained environments. Owida *et al.* (2024) designed a custom CNN architecture trained on the HAM10000 dataset, emphasizing the importance of preprocessing techniques such as morphological filtration for hair removal. Their streamlined model achieved a high efficiency of 95.23%, demonstrating that custom-built CNNs can still compete with heavier pre-trained models when data quality is rigorously managed.

The integration of meta-heuristic algorithms for hyperparameter optimization within complex neural architectures has become a prominent research direction to enhance segmentation accuracy. Ali *et al.* (2024) proposed a hybrid framework for dermoscopic image segmentation based on a fully convolutional encoder-decoder network (FCEDN) optimized via the Sparrow Search Algorithm (SpaSA). In this study, the individual wolf method and ensemble ghosting techniques were integrated into the SpaSA to maintain an effective balance between navigation and exploitation during the search process. Their proposed FCEDN-SpaSA architecture achieved high segmentation performance on datasets such as ISBI 2017 and PH2, while the adaptive CNN classification module reached a 91.67% accuracy rate with significantly lower energy, storage space, and memory access compared to conventional incremental learning techniques.

In parallel, researchers have focused on leveraging signal processing techniques to reinforce feature representation in the frequency domain. Claret *et al.* (2024) introduced an innovative approach combining Discrete Wavelet Transformation (DWT) with CNN models for enhanced skin cancer diagnosis. In this methodology, dermoscopic images are decomposed into multiple sub-images characterized by different spatial domains and independent frequencies (LL, LH, HH, HL). By utilizing the Low-Low (LL) features, which retain 50% of the relevant pixels from the original image, the model achieves effective dimensionality reduction and improved computational efficiency. Supported by a softmax activation function, this model achieved a sensitivity of 94% and a specificity of 91% on the HAM10000 dataset, significantly outperforming traditional artificial neural networks (ANN) and multilayer perceptron methods.

To address the challenges of high annotation costs and significant class imbalance in medical datasets, the synergy between active learning (AL) and optimization algorithms has emerged as a critical strategy. Mandal *et al.* (2024) proposed an efficient framework that integrates AL with Particle Swarm Optimization (PSO) to selectively identify the most informative unlabeled instances for expert annotation. This method utilizes PSO to enhance the selection process within the AL framework, ensuring that the model prioritizes training on the most uncertain and challenging samples. Experimental results using the EfficientNetV2M architecture demonstrated that the AL-PSO approach, specifically through the "Least Confidence" strategy, achieved a classification accuracy of 89.44% while requiring only 40% of the labeled training data. This approach offers a robust, cost-effective solution for clinical settings where labeled data is scarce.

MATERIALS AND METHODS

Dataset and Data Preprocessing

The foundation of this study is the HAM10000 ("Human Against Machine with 10,000 training images") dataset, which was prominently featured in the ISIC 2018 challenge (r25 2025). This benchmark repository is comprised of 10,011 high-quality dermatoscopic images collected from multiple clinical sites, representing a broad demographic and a wide variety of acquisition conditions. As showcased in Figure 1, the dataset captures the complex visual morphology of seven distinct diagnostic categories: actinic keratoses and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). Each category presents unique challenges, such as varying pigment patterns and irregular borders, which the models must learn to navigate for accurate classification.

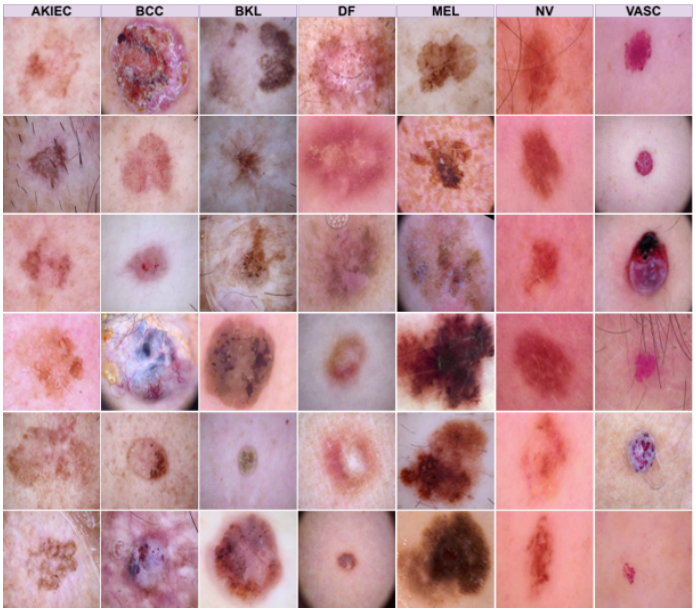


Figure 1 Representative dermoscopic image samples illustrating the morphological diversity of the seven skin lesion classes in the ISIC 2018 dataset

A significant aspect of this dataset is its inherent class imbalance, a characteristic reflective of real-world clinical distributions where benign cases often outnumber malignant ones. The precise numerical distribution of these classes across our experimental subsets is detailed in Table 1. The dataset is heavily skewed toward Melanocytic nevi (NV), which accounts for 6,705 images, whereas classes like Dermatofibroma (DF) are represented by only 115 samples. To address this and ensure the statistical validity of our performance metrics, we utilized a stratified sampling strategy to partition the data. This method meticulously preserved the original class ratios across the training (70%), validation (15%), and test (15%) sets, resulting in 7,005, 1,498, and 1,508 images respectively.

To ensure that the DL architectures could effectively extract meaningful features, we implemented a rigorous preprocessing and standardization pipeline. Raw images were first resized to a uniform resolution of 224×224 pixels to maintain consistency with the input requirements of the pre-trained backbones. Subsequently, pixel values were normalized using the global mean and standard deviation of the ImageNet dataset. This normalization is crucial for medical image analysis as it helps stabilize training dynamics and ensures that the model's attention mechanisms remain focused on lesion-specific pathological markers rather than being distracted by variations in lighting or resolution scale. By standardizing the input data in this manner, we created a level playing field for the comparative evaluation of the different CNN architectures.

Foundational Principles of Convolutional Neural Networks (CNNs)

At the heart of the recent revolution in medical image analysis lies the shift from manual, heuristic-based feature engineering to the automated, data-driven paradigm of CNNs. Unlike traditional computer vision techniques, CNNs are designed to mimic the human visual system by automatically learning hierarchical feature representations directly from raw pixel data. Through a series of specialized layers, primarily convolutional, pooling, and non-linear activation layers, these networks decompose complex dermatological structures into a multi-level abstraction of spatial patterns. Early layers typically capture low-level morphological hallmarks such as edges and color gradients, while deeper layers integrate these into high-level semantic descriptors capable of identifying the subtle diagnostic markers of malignancy. This inherent ability to preserve spatial localities while reducing dimensionality makes CNNs exceptionally robust for classifying diverse skin lesion categories (O'Shea and Nash 2015).

The mathematical rigor of these architectures is further enhanced by advanced structural innovations designed to optimize the learning process. To address the challenges of training deep networks, such as the vanishing gradient problem, modern backbones incorporate specialized design philosophies: ResNet (He *et al.* 2015) utilizes residual learning through skip connections to facilitate the flow of gradients, while DenseNet (Huang *et al.* 2017) promotes feature reuse by connecting every layer to every subsequent layer. More sophisticated frameworks, such as EfficientNet (Tan and Le 2019), employ compound scaling to balance network depth, width, and resolution, thereby maximizing diagnostic precision while maintaining a compact parameter footprint. During the training phase, these models are optimized through the minimization of a Cross-Entropy Loss function, which penalizes discrepancies between predicted and actual diagnostic labels. By employing robust optimizers like AdamW and dynamic learning rate schedulers, the network weights are iteratively refined to settle into an optimal configuration that ensures both high accuracy and clinical reliability.

Data Augmentation Strategy

We opted for a data augmentation strategy based on the default settings of the timm (PyTorch Image Models) library to improve the model's ability to generalize across different clinical scenarios. By leveraging these standard configurations, we introduced a variety of transformations such as random rotations, horizontal flips, and resized cropping, which effectively mimic the natural variability in how medical images are captured and oriented. These techniques ensure that the network does not simply memorize the training data but instead learns to recognize key pathological features regardless of their scale or position within the frame. Relying on the proven defaults of the timm framework allowed us to maintain a rigorous and reproducible training pipeline, providing strong regularization that balances complexity with the need for robust performance on unseen medical datasets (Wang *et al.* 2024; Mumuni *et al.* 2024).

Experimental Design and Training Protocol

To ensure the technical rigor and reproducibility of our experimental framework, we implemented all CNN architectures using the PyTorch library on a high-end workstation equipped with an NVIDIA RTX 5090 GPU (32GB VRAM), which provided the necessary computational power for efficient model convergence. The dataset was partitioned into training (70%, n=7,005), validation (15%, n=1,498), and testing (15%, n=1,508) subsets using a stratified sampling strategy to maintain consistent class proportions across all phases. Prior to training, each image was resized to a uniform 224×224 resolution and normalized according to ImageNet standards, a step crucial for stabilizing the learning process and ensuring the CNN backbones could focus on subtle, lesion-specific morphological features. We optimized the network parameters using the AdamW algorithm, carefully tuning the learning rate and weight decay to maintain a balance between convergence speed and generalization. The final model selection was determined by the best performance on the validation set, and the diagnostic efficacy was rigorously measured on the independent test set using

Table 1 Distribution of the HAM10000 dataset across 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

a comprehensive suite of metrics, including Accuracy, Precision, Recall, and F1-Score, across the seven diagnostic categories.

Performance Evaluation Metrics

To rigorously benchmark the diagnostic efficacy and clinical reliability of the evaluated CNN architectures, we utilized a comprehensive suite of statistical metrics derived from the multi-class confusion matrix. While overall Accuracy (1) provides a global measure of the model's classification success, the inherent class imbalance within the HAM10000 dataset, where certain benign cases significantly outnumber malignant ones, necessitates a more nuanced evaluation. To this end, we employed Precision (2) to quantify the models' predictive exactness and Recall (3), or sensitivity, to ensure the critical detection of malignant lesions that require early intervention. Given the natural trade-off between these two dimensions in dermatological screening, we prioritized the F1-Score (4) as a robust harmonic mean that balances precision and recall. Collectively, these metrics provide a holistic framework for assessing each architecture's ability to generalize across diverse pathological markers while maintaining the high level of precision required for real-world clinical decision support.

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

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

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

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

RESULTS

Quantitative Performance and Comparative Analysis

The objective evaluation of the selected CNN architectures was conducted through a multi-dimensional analysis that integrates both statistical classification metrics and computational efficiency parameters. This comprehensive assessment allows for a nuanced understanding of how different design philosophies, ranging from residual learning to compound scaling, respond to the complex morphological variations inherent in dermatoscopic imagery. The following sections detail the empirical findings derived from the independent test set, with a specific focus on the trade-offs between diagnostic precision and the technical constraints of real-time clinical deployment.

As summarized in Table 2, the experimental results reveal a distinct performance hierarchy among the evaluated models. EfficientNet-B0 emerged as the most robust architecture, achieving the highest overall Accuracy of 91.84% and a superior F1-Score of 0.8429. It is particularly noteworthy that EfficientNet-B0 achieved these results with the most compact parameter footprint in the group, utilizing only 4.02M parameters and requiring just 0.734 Gflops. This suggests that its compound scaling method, which simultaneously optimizes network depth, width, and resolution, is highly effective for capturing the intricate textural markers required for skin lesion classification.

In contrast, ResNet-50 demonstrated the lowest classification performance, with an Accuracy of 88.06% and an F1-Score of 0.7752, despite having a significantly larger parameter count of 23.52M. However, the data highlights a critical trade-off regarding operational latency: ResNet-50 recorded the fastest Inference Time (0.359 ms), whereas EfficientNet-B0 exhibited the longest latency at 5.9026 ms. This discrepancy indicates that while EfficientNet-B0 provides the most accurate diagnostic support, ResNet-50 or the mid-performing Inception-v3 (89.52% accuracy) and DenseNet-169 (90.32% accuracy) might be more suitable for high-throughput screening environments or deployment on edge-computing hardware with strict real-time requirements.

A granular look at the classification behavior of the top-performing model is provided by the confusion matrix in Figure 2. The model shows exceptional sensitivity toward NV, correctly classifying 980 instances. However, the matrix also reveals persistent diagnostic challenges; for example, 26 cases of MEL were misidentified as NV, and 13 cases of AKIEC were confused with BKL. These specific error patterns underscore the visual ambiguity between certain malignant conditions and their benign mimics, which remains a primary hurdle in automated dermatological assessment.

Table 2 Quantitative performance comparison of CNN architectures based on accuracy, precision, recall, F1-score, and computational efficiency metrics

Models	Accuracy	Precision	Recall	F1 Score	Params (M)	Gflops	Inference Time (Ms)
ResNet-50 (He *et al.* 2015)	0.8806	0.7972	0.7576	0.7752	23.52	8.2634	0.359
DenseNet-169 (Huang *et al.* 2017)	0.9032	0.8483	0.8245	0.8314	12.5	6.7169	0.7204
EfficientNet-B0 (Tan and Le 2019)	0.9184	0.8905	0.8068	0.8429	4.02	0.734	5.9026
Inception-v3 (Szegedy *et al.* 2016)	0.8952	0.8474	0.8193	0.8302	21.8	5.6719	0.4328

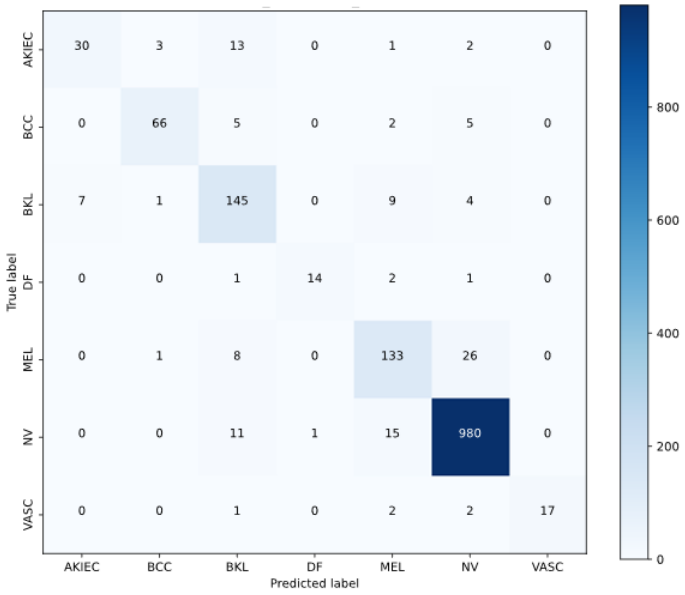


Figure 2 Confusion matrix of the EfficientNet-B0 model illustrating raw classification counts and specific inter-class error patterns across the independent test set

Visual Validation of Model Focus with Grad-CAM++

To ensure that our models were making decisions based on relevant pathological features rather than dataset artifacts, we utilized Grad-CAM++ for visual explanation. As shown in Figure 3, the "Correct Grad-CAM Predictions" row demonstrates that for successfully classified lesions across all seven categories, the model's attention was tightly localized on the lesion's core and its irregular borders. In contrast, the "Misclassified Grad-CAM Predictions" reveal that errors often stemmed from the model focusing on peripheral skin regions or being distracted by clinical artifacts like hair or skin folds, rather than the primary lesion. This visual evidence underscores the necessity of robust preprocessing and the potential of explainability tools to build clinician trust in "black-box" models.

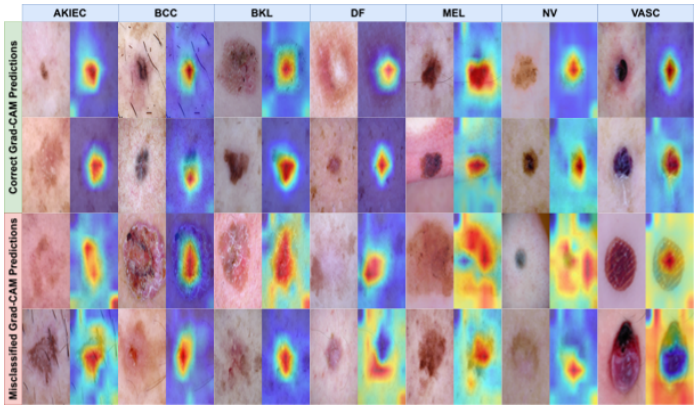


Figure 3 Visual explanation of model interpretability using Grad-CAM++ heatmaps for correct and misclassified skin lesion predictions

DISCUSSION

Interpretation of Key Findings

The results of this study reaffirm that architectural design often outweighs raw model size in medical imaging tasks. The success of EfficientNet-B0 can be attributed to its optimized balancing of network depth, width, and resolution, which proved more effective for the high-frequency textural details of dermatoscopic images than the traditional residual blocks of ResNet-50. Interestingly, the dense connectivity of DenseNet-169 provided a significant boost in recall (0.8245) compared to ResNet-50 (0.7576), suggesting that feature reuse is particularly beneficial for identifying minority classes like DF and VASC. These findings indicate that while accuracy is a valuable metric, the choice of a backbone must be dictated by the specific clinical priority, whether it be the highest possible precision (EfficientNet) or real-time processing speed (ResNet).

Clinical Implications, Limitations, and Future Directions

From a clinical standpoint, the high F1-Scores achieved by our top-performing models suggest they could serve as robust triage tools in primary care settings, potentially reducing the diagnostic

burden on specialists. However, the misclassification of melanoma as benign nevi, as seen in Figure 2, remains a primary limitation. This error is likely exacerbated by the dataset's heavy skew toward the NV class. Future research should focus on integrating cost-sensitive learning or synthetic oversampling (e.g., SMOTE) to specifically penalize "false benign" predictions for malignant cases. Additionally, while our use of Grad-CAM++ provided essential interpretability, moving toward "interpretable-by-design" architectures like Vision Transformers (ViTs) could further enhance the transparency and reliability of automated skin cancer screening in real-world practice.

CONCLUSION

This research underscores that architectural design, specifically the compound scaling of depth, width, and resolution found in EfficientNet-B0, consistently outweighs raw parameter count in complex medical imaging tasks such as skin cancer classification. While EfficientNet-B0 established a new benchmark for accuracy on the HAM10000 dataset, our analysis also revealed an essential operational trade-off: the faster inference of ResNet-50 makes it highly viable for low-latency edge-computing, whereas EfficientNet's superior precision is better suited for high-stakes diagnostic support. The integration of Grad-CAM++ provided a vital layer of interpretability, confirming that our models localized on legitimate pathological markers. Despite these successes, the persistent challenge of distinguishing melanoma from common nevi due to class imbalance remains a hurdle for widespread clinical adoption. Future research should prioritize cost-sensitive learning to penalize "false benign" errors and explore "interpretable-by-design" architectures like Vision Transformers to further enhance transparency and clinician trust in automated screening systems.

Ethical standard

Not applicable.

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 they have no conflicts of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used artificial intelligence tools in order to improve the readability and language quality of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

LITERATURE CITED

2024 An integrated deep learning model for skin cancer detection using hybrid feature fusion technique.

2025 Skin cancer: Ham10000. Kaggle Dataset.

Ali, R., A. Manikandan, R. Lei, and J. Xu, 2024 A novel spasa based hyper-parameter optimized fcedn with adaptive cnn classification for skin cancer detection. Scientific Reports **14**: 9336.

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

Attallah, O., 2024 Skin-cad: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level cnns features and transfer learning. Computers in Biology and Medicine **178**.

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, 2025a Comparative analysis of transformer architectures for brain tumor classification. Exploration of Medicine **6**.

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

Cakmak, Y. and J. Zeynalov, 2025 A comparative analysis of convolutional neural network architectures for breast cancer classification from mammograms. Artificial Intelligence in Applied Sciences **1**: 28–34.

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**.

Claret, S. P. A., J. P. Dharmian, and A. M. Manokar, 2024 Artificial intelligence-driven enhanced skin cancer diagnosis: leveraging convolutional neural networks with discrete wavelet transformation. Egyptian Journal of Medical Human Genetics **25**: 50.

Farea, E., R. A. A. Saleh, H. AbuAlkebash, A. A. R. Farea, and M. A. Al-antari, 2024 A hybrid deep learning skin cancer prediction framework. Engineering Science and Technology, an International Journal **57**.

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

He, K., X. Zhang, S. Ren, and J. Sun, 2015 Deep residual learning for image recognition.

Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 2017 Densely connected convolutional networks.

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.

Leiter, U., U. Keim, and C. Garbe, 2020 Epidemiology of skin cancer: Update 2019. Advances in Experimental Medicine and Biology **1268**: 123–139.

Mandal, S., S. Ghosh, N. D. Jana, S. Chakraborty, and S. Mallik, 2024 Active learning with particle swarm optimization for enhanced skin cancer classification utilizing deep cnn models. Journal of Imaging Informatics in Medicine **38**: 2472–2489.

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)*.

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

O'Shea, K. and R. Nash, 2015 An introduction to convolutional neural networks. International Journal of Research in Applied Science and Engineering Technology **10**: 943–947.

Owida, H. A., N. Alshdaifat, A. Almaghthawi, S. Abuowaida, A. Aburomman, *et al.*, 2024 Improved deep learning architecture for skin cancer classification. Indonesian Journal of Electrical Engineering and Computer Science **36**: 501–508.

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**.

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.

Pacal, I., B. Ozdemir, J. Zeynalov, H. Gasimov, and N. Pacal, 2025 A novel cnn-vit-based deep learning model for early skin cancer diagnosis. Biomedical Signal Processing and Control **104**.

Siegel, R. L., A. N. Giaquinto, and A. Jemal, 2024 Cancer statistics, 2024. CA: A Cancer Journal for Clinicians **74**: 12–49.

Szegedy, C., V. Vanhoucke, S. Ioffe, and J. Shlens, 2016 Rethinking the inception architecture for computer vision.

Tan, M. and Q. V. Le, 2019 Efficientnet: Rethinking model scaling for convolutional neural networks.

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

Zeynalov, J., Y. Cakmak, and I. Pacal, 2025 Automated apple leaf disease classification using deep convolutional neural networks: A comparative study on the plant village dataset. Journal of Computer Science and Digital Technologies **1**: 5–17.

How to cite this article: Kingni, S. T. and Baawain, S. Comparative Evaluation of Convolutional Neural Network Architectures for Automated Skin Cancer Classification: A Study on the ISIC 2018 Dataset. *Artificial Intelligence in Applied Sciences*, 2(2), 8-14, 2026.

ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.15-26
https://doi.org/10.69882/adba.ai.2026013

Hawk-Swarmception Segmentation Network (HSCS-Net): Enhanced Liver Tumor Segmentation with Receptive Field Optimization and Clinical Data-Guided Feature Selection via PSO and GWO

Prachi Chhabra [*,1], **Sukhendra Singh** [α,2] **and Tathagat Banerjee** [β,3]

[*,α]Department of Information Technology, JSS Academy of Technical Education, Noida, [β]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India.

ABSTRACT The detection and treatment of cancerous tumors in the liver are considered one of the most challenging health issues for patients globally. There is a need for fully automated systems that can precisely detect and segment the tumorous regions in medical images. This work aims to develop an automated system which utilizes a hybrid- deep- learning framework with fusion multi-scale inception feature extraction called HSCS-NET (Hawk-Swarm Ception Segmentation Network). In the proposed model, the encoder comprises of Hawk Gating with SE (squeeze-and-excitation) attention while decoders consist of adaptive attention skip fusion which comprises of Swarm ception Residual ASPP bridges. These components allow the model to recover important details of the boundaries of the tumors irrespective to their shape, size, and the tissue contrast in CT scans.In order to improve segmentation, the HSCS-NET framework is equipped with a hybrid optimization module based on PSO (Particle Swarm Optimization) and GWO (Grey Wolf Optimization) for dynamic feature selection and reliable convergence. The model was evaluated against the 3DIRCADb1 Liver Tumor Segmentation Challenge (LiTS) dataset and significantly outperformed all other models achieving a Dice coefficient value of 0.98, Accuracy of 0.9891, and Precision of 0.9901.This marks a substantial improvement over prior models such as Christ et al.'s CNN (Dice: 0.823), Wu et al.'s Fuzzy C-means + GC (Dice: 0.83), Muhammad et al.'s ResNet (Dice: 0.87, Accuracy: 0.945, Precision: 0.93), and Kaur et al.'s PSO-PSP-Net (Accuracy: 0.9754, Precision: 0.9632).The HSCS-NET architecture is mathematically grounded, modularly extensible, and validated through rigorous cross-validation and confidence refinement. With its high segmentation performance, clinical reliability, and computational efficiency, HSCS-NET stands as a superior advancement in automated liver cancer diagnostics, reducing clinician workload and improving patient prognoses through precision imaging.

KEYWORDS
Liver tumor
Segmentation
PSO
GWO
Deep learning

INTRODUCTION

Image segmentation stands as a critical computational strategy in computer vision, offering transformative potential in medical imaging and diagnostics. Among its many clinical applications, segmentation plays an instrumental role in the accurate identification and delineation of anatomical structures, which is particularly vital in oncological assessments. Liver cancer, known for its high fatality rate and diagnostic complexity, represents one of the foremost challenges in global health. It is recognized as the second most common cause of cancer-related deaths worldwide. To evaluate liver conditions and detect malignancies, radiologists primarily rely on imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI). These modalities generate volumetric 3D scans; however, for practical implementation and computational efficiency, slice-wise 2D segmentation is often employed in both manual annotation and automated model training. Therefore, despite being

derived from 3D volumes, the use of "image segmentation" terminology is appropriate in this context. The primary methods that radiologists and oncologists traditionally use to examine liver structure and texture consist of computed tomography (CT) and magnetic resonance imaging (MRI). Multiple imaging techniques enable medical professionals to discover anomalies that function as biomarkers for liver malignancy diagnosis and ongoing observation. The combination of manual and semi-manual approaches when evaluating liver CT volume scans produces increased chances for procedural difficulties to arise. As reported by the World Health Organization, liver cancer is one of the top three causes of cancer deaths globally. Timely and precise detection of liver tumors is critical for indeed treatment and favorable prognostic outcomes. Computed Tomography (CT) imaging is

Manuscript received: 15 November 2025,
Revised: 10 December 2025,
Accepted: 12 January 2026.

[1]prachi.chhabra@gmail.com
[2]sukhendrasingh@jssaten.ac.in
[3]banerjeetathagat@gmail.com (**Corresponding author**).

instrumental in the diagnosis and staging of liver cancer. Nonetheless, the manual segmentation of the liver and the tumors associated with it from the CT images is an arduous and tedious task that is fraught with the potential for numerous errors (Cakmak *et al.* 2026). To overcome these difficulties, different automated approaches for liver and tumor segmentation have been developed, most of which emphasize the use of modern machine learning techniques in conjunction with image processing. Among these approaches, slice-wise image segmentation has been identified as an effective approach to some of the problems associated with 3D medical image segmentation. Slice-wise segmentation is the process by which a 3D volume is separated into 2D slices, each of which is then processed individually for segmentation. By converting a 3D problem into a sequence of 2D problems, this approach greatly decreases the computational burden (Pacal and Cakmak 2025).

Moreover, it permits models to consider local contextual details pertaining to each slice while also leveraging the spatial context of the full 3D volume. This has been used extensively in liver tumor segmentation since it achieves a good trade-off between computational cost and segmentation precision. The idea of slice-wise segmentation has been the subject of extensive refinement over the years, with numerous researchers contributing to this work. Christ *et al.* (2016) pioneered the use of cascaded fully convolutional neural networks (FCNs) and 3D conditional random fields (CRFs) for the automated segmentation of liver and liver lesions in CT images. Their method incorporated a cascade of FCNs to perform liver and lesion segmentation on a per-slice basis, applying CRFs to merge spatial information from neighboring slices for boundary refinement post-segmentation. This was one of the earliest works in the automation of deep learning-based liver tumor detection, demonstrating the efficacy of slice-wise segmented approaches. Many other studies have since built upon his work, demonstrating the relevance of slice-wise segmentation in clinical settings. Jin *et al.* (2020) proposed RA-UNet, a novel hybrid attention deep network aimed at extracting liver and tumor regions from CT scans.

Their approach uses an attention mechanism that helps direct the network's focus to pertinent areas in each slice, thereby increasing the accuracy of segmentation. This attention mechanism works well in CT images since the features of liver tumors are of differing intensities and textures. Slice-wise segmentation in RA-UNet helps the model to better identify the liver and tumors within each slice through parallel processing regardless of how challenging each slice may be. Jiang *et al.* (2019) also suggested AH-CNet, a hybrid deep learning model with attention and hybrid connections focused on liver tumor segmentation in CT volumes. AH-CNet processes CT images slice by slice, capturing both global and local contexts to perform the segmentation of the liver and tumors with precision. This approach allows the model to focus on specific features, such as tumor boundaries, whilst keeping in mind the inter-slice context. With the use of attention mechanisms, AH-CNet adapts to liver tumors' varying characteristics in different CT slices, improving segmentation accuracy. The past several years have witnessed great strides in the area of medical image segmentation, particularly through the use of deep learning and convolutional neural networks; CNNs, as noted by Krizhevsky *et al.* (2017) demonstrated the effectiveness of hierarchical feature learning through the use of deep CNNs for image classification. This innovation essentially caters to the progression of deep learning techniques in the domain of medical image segmentation, such as the segmentation of CT images slice-wise. Li *et al.* (2014) advanced this concept by integrating texture analysis with level-set methods for liver segmentation in CT images. Their approach utilized a multi-step method where slice-wise analysis was applied to address the challenging aspects of the liver's anatomical diversity, where texture features refined the segmentation process tailored to each slice. Li *et al.* (2018) proposed H-DenseUNet, a hybrid model of densely connected UNet tailored for liver and tumor segmentation from CT volumes. Incorporating dense connections between layers to improve feature reuse enables the model to learn complex representations, enhancing its performance. H-DenseUNet processes each CT slice independently, applying strong dense connectivity improves slice-wise segmentation accuracy and 3D CT volumetric liver tumor segmentation performance. Besides using FCNs and CNN-based approaches, the adaptive fast marching method is another approach to liver segmentation. Song *et al.* (2013) Developed a framework with an adaptive fast marching technique tailored for automatic liver segmentation in CT images, processing them slice by slice. This method is highly effective for automated liver segmentation in clinical practice because of its ability to manage a variety of complex liver shapes and heterogeneous tissue properties. The efficiency achieved through a slice-wise approach shows that each slice is processed independently, preserving accuracy while ensuring speed. Some researchers explored the use of deep learning methodologies for liver mass differentiation. Through the use of CNNs, Yasaka *et al.* (2018) classified liver masses in dynamic contrast-enhanced CT scans. Their work proved that slice-wise segmentation, where each slice was scrutinized for tumor presence, is useful for liver mass differentiation. The application of slice-wise segmentation enables the CNN to concentrate on specific features associated with each mass which enhances the model's differentiation between benign and malignant liver lesions. Liver cancer is one of the most diagnosed and deadly cancers globally; the world health organization emphasizes on this fact. In order to improve patient outcomes and lower mortality rates, early detection and accurate segmentation of liver tumors is critical (World Health Organization 2021) . Li *et al.* (2022) investigated the global burden of liver cancer and highlighted the necessity of advancements in diagnostic technologies, especially automated liver tumor segmentation systems, in order to expedite the diagnosis and treatment of the ailment. In areas that lack proficient radiologist, the application of automated segmentation techniques that use slice-wise segmentation would greatly improve the speed and precision of diagnosing liver cancer. The use of FCNs for automatic segmentation of liver tumors in multi phased contrast enhanced CT images was done by Sun *et al.* (2017). They demonstrated the successful application of slice wise segmentation in overcoming the challenges presented by multi-phase CT scans, where tumors may exhibit different appearances in each phase. Their model achieved accurate segmentation of both the liver and tumors throughout the various phases of the contrast-enhanced CT scan by processing each slice individually. In the study by Christ (2017) enhanced the understanding of convolutional neural networks in the context of medical image classification and segmentation. His approach on slice-wise segmentation, in particular, has greatly advanced the field of automated segmentation of liver and tumor tissues from CT scans. Combination of fuzzy C-means and graph cuts has proven effective for the segmentation of liver tumors in 3D CT images. Wu *et al.* (2017) applied these techniques in a slice-wise manner and achieved notable success in high accuracy segmentation. Their method is effective for complex liver tumor structures because it permits slice-by-slice tumor segmentation while maintaining spatial coherence between the slices. The work of Lu *et al.* (2020) illustrates the use of sophisticated algorithms like VGG and

extreme learning machines in the analysis of medical images. Their diagnosis of cerebral microbleeds using slice-wise segmentation displays the ability of the methodology to enhance the diagnosis and treatment strategies for many medical conditions. As has been established, slice-wise image segmentation is one of the most important techniques for liver and tumor segmentation in CT scans. Slice-wise segmentation translates 3D CT volume processing into 2D slices to lower computational workload while fully automating liver tumor segmentation. The contributions from authors such as Christ *et al.* (2016), Jin *et al.* (2020), Jiang *et al.* (2019), among others, have clearly proven the importance of deep learning and slice-wise techniques in fully automating liver tumor detection and advanced research in the domain. Given the rapid proliferation of liver cancer cases worldwide, there is an ever-growing demand for reliable and fully automated diagnostic systems, making further research into slice-wise segmentation techniques highly beneficial to liver cancer patients in regard to their treatment and prognosis.

LITERATURE REVIEW

There has been remarkable advancement in utilizing sophisticated deep learning methods for the analysis of medical images. In this review, we discuss the application of deep learning algorithms and optimization techniques in the segmentation and diagnosis of important medical disorders such as liver tumors, cerebral microbleeds, and cancers. An enhanced fuzzy C-means algorithm with graph cuts for the 3D segmentation of liver tumors in CT images marked an advancement in accurately and efficiently tumor segmenting within liver CT scans, pivotal for precision in diagnosis and treatment planning (Wu *et al.* 2017). Further study into the diagnosis of cerebral microbleeds using a VGG network with an extreme learning machine (ELM) optimized through a Gaussian map bat algorithm demonstrated the efficacy of hybrid models in enhancing diagnostic accuracy through the added complexity of microbleeds in neuroimaging (Lu *et al.* 2020). More recently, a study using MONAI and Pytorch for liver tumor segmentation in CT images highlighted the remarkable ability of transfer learning and domain-specific models to accurately identify tumors in radiological images. An optimized computer-aided diagnostic model for liver tumor detection based on InceptionV3 highlighted the need for hyperparameter optimization and model refinement in medical image analysis, especially with CT scan slices (Kaur and Kaur 2024). The ACE-SeizNet model boosts automated seizure detection through the fusion of multi-domain deep features. The integration of attention mechanisms in EEG signal processing prioritizes critical features for real-time seizure detection, demonstrating a novel application in clinical environments (Banerjee 2025a). To enhance transparency in AI models, which is pivotal for clinical validation and real-time use in cancer diagnostics, a pyramidal explainable AI framework for cervical cancer detection was proposed (Banerjee 2025b). An effective focus on diagnostic accuracy and reliability for cancer detection systems using histopathological images of lung cancer lesions was applied through a pyramidal attention network, which highlights critical tissue sample features (Banerjee 2025c). HHO-UNet-IAA, an architecture for glaucoma segmentation, employs a novel optimization-based technique which integrates attention with UNet-Inception and Harris Hawks Optimization to bolster the segmentation of glaucoma-related features in medical imaging (Banerjee *et al.* 2025). Utilizing a deep convolutional neural network, Falcon, along with transfer learning, proved effective for malaria parasite detection, showcasing the role of deep learning in disease detection with limited labeled data and pretrained models (Banerjee *et al.* 2022a). Attention mechanisms significantly improved the discrimination for detection of subtle pneumonia symptom variations, as demonstrated by the high diagnostic accuracy achieved using the attention-based discrimination model focused on mycoplasma pneumonia (Banerjee *et al.* 2022b). Further research emphasized neural network-based strategies for handling textual information highlight the significance of converting text features into vectors for machine learning in healthcare (Banerjee *et al.* 2022c). Advanced image creation strategies such as GANs are increasingly utilized within healthcare, enhancing diagnostic performance by enabling the application of deep belief convolutional networks toward pneumonia diagnoses through the augmentation of synthetic image generation techniques (Banerjee *et al.* 2021b). Integrating GANs with convolutional neural networks for improved classification of pneumonia in radiological samples demonstrates the capability of generative models to address challenges posed by limited datasets in medical imaging (Banerjee *et al.* 2021a). The study on hand sign recognition using infrared images from Leap Motion sensors contributed to novel, non-intrusive technologies in healthcare for sign language recognition aimed at aiding the hearing impaired population (Banerjee *et al.* 2021c). The application of a multi-dimensional structured neural network to analyze driving behaviors and compute a driver score advanced deep learning from the realm of healthcare to include transportation safety (Karthikeyan *et al.* 2021). The use of a single-node Hadoop cluster for small scale automation in an industrial setting illustrated the application of computational models and cluster systems in automating industrial processes, revealing the fusion of machine learning and industrial automation (Peesa *et al.* 2020). Using Resio-Inception U-Net as a thoracic organ segmentation mask enables improved segmentation of thoracic organs which aids in more precise diagnosis and treatment planning (Saminathan *et al.* 2024). An integrated method for breast cancer classification incorporating aggressiveness delineation techniques was developed which comprehensively addressed the problem of breast cancer categorization by focusing on tumor aggressiveness, a critical determinant for treatment selection (Singh *et al.* 2025a). A review post analyzing different machine learning and deep learning approaches to predicting the responses to anti-cancer drugs offered a review of approaches, which highlighted the intersection of cancer care and artificial intelligence, advancing the field of personalized medicine (Singh *et al.* 2025b). A strong prognosis model of kidney carcinoma using Swin-ViT and DeepLabV3+ with multi model transfer learning for better prognosis and characterization of the kidney carcinoma demonstrated the increasing application of transformer models and transfer learning in the analysis of medical images (Rehman *et al.* 2025). These studies highlight the significant impacts that the application of techniques of optimization and deep learning will have on medical diagnostics. It illustrates the increasing capabilities of AI in developing solutions to problems in healthcare that are more precise, more efficient, and easier to understand.

Contribution of the Study

- Study work to create a deep learning-based framework that uses varied liver tumor CT images for better region of interest identification accuracy and efficiency.
- In contrast to previous studies, which mainly depended on enhanced MRI and CT scans for narrow liver tumor type examination, this research improves feature extraction and organizational structures by employing Convolutional Neural Networks (CNNs).
- The automated system improves both diagnostic precision

and healthcare professional workflow by finding and categorizing liver tumors; therefore, it shortens the amount of time needed and work required.
- The adaptable model demonstrates the potential to develop across different inputs and circumstances, thus offering value to extended medical applications.

MATERIALS AND METHODS

Algorithm 1 Hawk-SwarmCeption Segmentation Network(HSCS-Net) presents a comprehensive and modular pipeline (showcased in Algorithm 1) designed for liver and liver tumor segmentation from CT images. The framework integrates classical image preprocessing techniques, advanced multi-scale feature extraction, attention-driven encoder-decoder design, and hybrid feature selection strategies to achieve high segmentation performance in medical imaging tasks. The pipeline starts with image preprocessing, in which each CT image is normalized by its mean and standard deviation to ensure consistency in intensity scaling across the entire dataset. After normalization, the image's contrast is enhanced using CLAHE (Contrast Limited Adaptive Histogram Equalization), which is useful in accentuating low organ contours in liver CT scans. The images are resized to a given resolution and undergo data augmentation to improve model generalization and reduce overfitting. In the multi-scale feature extraction stage, the network uses a modified version of the Inception module consisting of parallel 1×1, 3×3, and 5×5 depthwise convolutions. This model is able to capture features at different receptive fields and thus detect both fine and coarse anatomical structures. As discussed, batch normalization and ReLU activation are added after each convolutional operation to stabilize and activate the feature maps. The generated multi-scale feature maps are concatenated. They are then processed by channel attention, which adaptively adjusts the importance of each channel based on its contribution to the segmentation. Following this, the HSCS-Net is equipped with Hawk-Swarm attention-driven encoder-decoder which serves as the core structural advancement of the HSCS-Net. This module is described in Algorithm 4, which details the incorporation of Hawk gating frameworks alongside squeeze-and-excitation (SE) blocks into the encoder to improve spatial and channel feature selectivity. An Atrous Spatial Pyramid Pooling (ASPP) module is placed at the bridge between the encoder and decoder to obtain multi-scale contextual information at several dilation levels, after which Swarmception residual fusion further enhances feature representation via residual learning. The decoder employs adaptive skip fusion in which the decoder outputs are modulated with spatial attention maps and infused with the corresponding encoder features, ensuring that critical spatial details and context required for accurate segmentation are preserved. To remove limitations of the former approach, the model is augmented with a hybrid feature selection scheme which uses PSO and GWO. A binary population of feature masks is initialized representing a subset of the extracted features. In optimization, each candidate is assessed with a fitness function that combines Dice loss and Binary Cross-Entropy loss computed from the segmentation output of HSCS-Net. While PSO motivates the exploration of the feature space by updating particles with regard to personal and global best positions, GWO utilizes the social hierarchy of alpha, beta, and delta wolves to exploit the search space. The integration of both approaches facilitates and strengthens feature subset selection by minimizing redundancy and maximizing accuracy in segmentation.

The employed loss function is a weighted sum of Dice loss, Binary Cross-Entropy loss, and a structural similarity term (SSIM) which fosters overlap accuracy, pixel-wise separation fidelity, and anatomical structure preservation, respectively. Training the model with Adam optimizer applying step-wise learning rate reduction along with gradient clipping achieves stable convergence preventing explosion of the gradients during training.

Finally, the model is assessed using a comprehensive set of performance indicators including Dice coefficient, Jaccard index (IoU), Precision, Recall, and SSIM. To ensure statistical rigor, the framework implements k-fold cross validation and reports aggregated metrics such as mean, variance, and kurtosis of the performance scores across the folds. Overall, HSCS-Net framework epitomizes the integration of attention mechanisms and multi-scale feature modeling with optimization and biologically inspired reasoning, guided stratified loss functions to provide a robust, versatile, and clinically interpretable solution for liver and tumor segmentation in imaging.

Algorithm 1 Hawk-SwarmCeption Segmentation Network (HSCS-Net)

Input: Medical image dataset $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_M\}$ with ground truth masks $\mathcal{M} = \{\mathbf{M}_1, \ldots, \mathbf{M}_M\}$

Output: Segmented masks $\mathcal{M}' = \{\mathbf{M}'_1, \ldots, \mathbf{M}'_M\}$

Step 1: Image Preprocessing
for each image $\mathbf{I}_i \in \mathcal{I}$ **do**
  Normalize: $\mathbf{I}_i \leftarrow (\mathbf{I}_i - \mu)/\sigma$
  Enhance contrast: CLAHE $\leftarrow$ CLAHE($\mathbf{I}_i$)
  Resize: $\mathbf{I}_i \leftarrow$ Resize($\mathbf{I}_i, H, W$)
  Augment: $\mathbf{I}_i \leftarrow$ Augment($\mathbf{I}_i$)
end for

Step 2: Inception-based Multi-scale Feature Extraction
for each $\mathbf{I}_i \in \mathcal{I}$ **do**
  Apply 1x1, 3x3, 5x5 depthwise convolutions
  Apply batch normalization + ReLU after each
  Concatenate features: $\mathbf{F}_i = \text{Concat}(\mathbf{F}_{1\times1}, \mathbf{F}_{3\times3}, \mathbf{F}_{5\times5})$
  Channel attention: $\mathbf{F}'_i = \text{ChannelAttention}(\mathbf{F}_i)$
end for

Step 3: Hawk-Swarm Attention-Driven Encoder-Decoder (Refer: Algorithm 4)
for each $\mathbf{F}'_i$ **do**
  Encode with SE + Hawk gating: $\mathbf{E}^i = \text{HawkGateSE}(\mathbf{F}'_i)$
  Apply ASPP: $\mathbf{B}_{ASPP} \leftarrow \text{ASPP}(\mathbf{E}^L)$
  Swarmception Residual: $\mathbf{B}_{Swarm} \leftarrow \text{SwarmRes}(\mathbf{B}_{ASPP})$
  Decode via adaptive attention skip fusion (See Algorithm 3)
  Output: $\mathbf{U}_i \leftarrow \text{Decoder}(\mathbf{B}_{Swarm}, \mathbf{E}^1, \ldots, \mathbf{E}^L)$
end for

Step 4: Hybrid Feature Selection using PSO + GWO (Refer: Algorithm 3)
Initialize binary particle population $\mathcal{X}_i^{(0)} \in \{0,1\}^n$
Fitness: Dice + BCE using $HSCS(\mathbf{I}_i, \mathcal{F}_{\mathcal{X}_i})$
for each iteration $t = 1$ to T **do**
  PSO velocity and binary update
  GWO: Compute $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ from alpha/beta/delta
  Combine: $\mathcal{X}_i^{(t+1)} = \frac{\mathcal{X}_1 + \mathcal{X}_2 + \mathcal{X}_3}{3}$
  Threshold to binary $\mathcal{X}_i^{(t+1)} \in \{0,1\}^n$
  Update fitness and bests
end for
Select optimal subset $\mathcal{F}^*$ for final segmentation

Step 5: Loss Function and Optimization
Total Loss:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Dice}} + \lambda_2 \mathcal{L}_{\text{BCE}} + \lambda_3 (1 - SSIM(\mathbf{S}, \mathbf{M}))$$

Optimizer: AdamW with learning rate decay
Apply gradient clipping for training stability

Step 7: Performance Metrics and Validation
Metrics: Dice, Jaccard, Precision, Recall, SSIM via k-fold cross-validation
return Segmentation map set $\mathcal{M}'$

The Figure 1 Proposed Model framework presents a comprehensive liver segmentation pipeline utilizing deep learning and hybrid optimization techniques. It begins with image acquisition, where CT scan images are collected and split into training (70%), validation (20%), and testing (10%) subsets to ensure proper model generalization. The preprocessing stage involves image resizing, cropping, and data augmentation techniques such as rotation, flipping, and intensity variations to improve robustness. The core segmentation process is enhanced using Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO), which optimize feature selection and hyperparameters to refine segmentation accuracy.

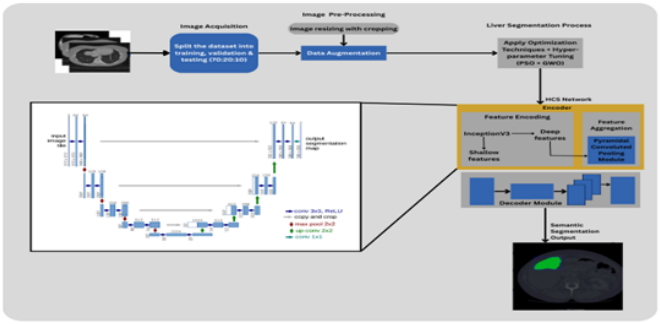


Figure 1 Proposed Model framework presents a comprehensive liver segmentation pipeline utilizing deep learning and hybrid optimization techniques

The segmentation model, referred to as HCS Network (Hybrid Convolutional Segmentation Network), utilizes an encoder-decoder structure in which the encoder performs hierarchical feature extraction and combines shallow and deep feature learning through a pyramid pooling module which enables extraction of multi- scale features. After upscaling the feature maps, the decoder incorporates skip connections to facilitate accurate boundary detection. The final output after semantic segmentation highlights the liver in green, clearly delineating it from the surrounding tissues. This framework is optimized using hybrid strategies along with multi-scale feature extraction and attention-based feature refinement to increase accuracy of segmentation, thus making it viable for medical imaging in the diagnosis of liver diseases and detection of tumors.

Data Description

The proposed approach is assessed using two datasets: 3DIRCADb1 (Bilic *et al.* 2019) and the Liver Tumor Segmentation Challenge (LiTS) (Soler *et al.* 2010). The LiTS dataset consists of abdominal CT and MRI images, with 20 sets of CT scans utilized for both training and testing purposes. Pixel-wise segmentation is conducted, with 70% of the data allocated for training, 20% for validation, and 10% for testing.

Optimization

Altering parameters for Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO) improves the accuracy for liver tumor segmentation. PSO enhances segmentation by mimicking avian social behavior, searching for better strategies to optimize fitness loss. GWO achieves a manual balance between exploration and exploitation in pursuit of optimal segmentation results by emulating wolf hunting strategies. The novelty of utilizing these techniques lies in their ability to significantly improve the accuracy of segmentation and their flexibility in adapting to various imaging scenarios, successfully addressing challenges such as irregular tumor shapes and varying contrasts.

Hawk-Swarmception Segmentation Network

The Hawk-SwarmCeption architecture is an advanced model that merges the Encoder Decoder Module and Inception network designs by integrating cross-feature connections to enhance the efficiency of segmentation. Our combined architecture uses Encoder Decoder Module and Inception network elements to fix common problems with basic image segmentation methods (illustrated in Algorithm 2). The Encoder Decoder Module encoder-decoder design helps medical image segmentation tasks because it preserves spatial data and locates structures accurately. The network connects the encoder directly to the decoder so the feature information is protected and stays intact during the process. The Encoder Decoder Module model keeps precise anatomical information through every stage to make accurate segmentations within liver tumor areas.

Besides Encoder Decoder Module functionality, the inception network brings multi-scale feature extraction power by combining different size convolution filters into single network. The model can detect both small and large tumor details better by processing images at various levels. Parallel convolutions help the network learn context from both small and large areas which makes it better at processing different liver tumor types. The hybrid model uses combined network designs that better detects tumors while making more precise feature extractions. This design allows the network to accurately capture features of different sizes and levels of complexity, hence improving its capacity to understand a wide range of patterns and fine details inside the image. Inception modules, as observed in InceptionV3, use filters of different sizes (1x1, 3x3, 5x5) and pooling operations to achieve a comprehensive analysis of the visual features.

Algorithm 2 HSCS-NET: Liver and Liver Tumor Segmentation Algorithm

Input: CT image $\mathbf{I} \in \mathbb{R}^{H \times W}$, Ground truth label $\mathbf{L} \in \mathbb{R}^{H \times W}$
Output: Segmentation map $\mathbf{S} \in \mathbb{R}^{H \times W}$

Step 1: Input Preparation
  Load CT scan image $\mathbf{I}$
  Load corresponding ground truth mask $\mathbf{L}$

Step 2: Multi-Scale Feature Extraction via Inception Modules
for each level $i \in \{1, 2, \ldots, N\}$ **do**
  Apply Inception block: $\mathbf{F}_{\text{inc}}^{i} \leftarrow \mathcal{M}_{\text{inc}}^{i}(\mathbf{I})$
end for
Concatenate multi-scale features: $\mathbf{F}_{\text{inc}} \leftarrow \text{Concat}(\mathbf{F}_{\text{inc}}^{1}, \ldots, \mathbf{F}_{\text{inc}}^{N})$

Step 3: Cross-Feature Integration with Encoder
for each encoder layer $i \in \{1, 2, \ldots, N\}$ **do**
  $\mathbf{F}_{\text{enc}}^{i} \leftarrow \mathcal{E}^{i}(\mathbf{I})$ ▷ Extract features from encoder
  $\mathbf{F}_{\text{cross}}^{i} \leftarrow \text{Fuse}(\mathbf{F}_{\text{inc}}^{i}, \mathbf{F}_{\text{enc}}^{i})$ ▷ Cross-feature concatenation or summation
end for

Step 4: Decoding and Segmentation Reconstruction
Decode fused features using decoder: $\mathbf{S} \leftarrow \mathcal{D}(\mathbf{F}_{\text{cross}}^{1}, \ldots, \mathbf{F}_{\text{cross}}^{N})$
return S ▷ Final liver and tumor segmentation mask

Algorithm 2 HSCS-NET Network Algorithm for Liver and Liver Tumor Segmentation, showcases that the Inception modules are embedded into the Encoder Decoder Module framework, particularly within the encoder section. This integration enhances the model's feature extraction capabilities by combining detailed multi-scale features from the Inception modules with the spatial context preserved by the Encoder Decoder Module's skip connections.

Hawk-Swarm Optimization framework

The Haris-Swarm Optimization Framework presents a novel hybrid approach for feature selection by integrating two powerful swarm intelligence algorithms, Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO) (defined in Algorithm 3). This dual-strategy mechanism is designed to identify the most relevant and informative features from a high-dimensional feature space extracted from CT scans, prior to liver and tumor segmentation using the Hawk-SwarmCeption Segmentation Network (HSCS-Net). Feature selection plays a critical role in eliminating redundancy,

reducing overfitting, and improving computational efficiency, especially in medical imaging tasks where large and complex datasets are prevalent.

As a binary vector $X_i \in \{0,1\}^n$, where each bit indicates the inclusion or exclusion of a corresponding feature. A population of such vectors is initialized randomly, forming the candidate solutions. Each solution is evaluated using a fitness function that measures segmentation performance, typically the Dice coefficient or Intersection over Union (IoU), after feeding the selected subset of features into HSCS-Net. The PSO component drives global exploration by updating each particle's velocity and position based on its personal best and the global best solution found so far. The velocity update equation incorporates inertia as well as cognitive and social terms, steering particles towards more promising areas within the search space. Following this, each particle's position is transformed into a binary vector using a sigmoid-based thresholding method.

At the same time, the GWO component captures local exploitation by simulating the social structure of grey wolves. The top three solutions which are referred to as alpha, beta, and delta are used to steer the remaining population with adaptive coefficient-based distance updates. A candidate solution is drawn towards the leading wolves, and the average of the modified positions determines where the search agent will be placed. This GWO mechanism enhances convergence by reinforcing top consensus divergence while maintaining diversity. After every iteration, both global best, which is PSO, and top wolves, GWO, are refreshed with the latest fitness score. By integrating PSO and GWO within a single framework, both the exploration and exploitation phases are optimally conducted in selecting the feature subset F^* .After selection, the subset is used as input for HSCS-Net which applies multi-scale Inception modules and cross-feature skip connections in an encoder-decoder framework to accurately segment the liver and tumors. The network's ability to disregard clinically irrelevant features enhances the accuracy and efficiency of the segmentation process and reduces the complexity of the training phase. Apart from these advantages, the Haris-Swarm architecture greatly improves the generalization capability of HSCS-Net, and simultaneously provides a robust, flexible, and scalable approach to biomedical image analysis.

Algorithm 3 introduces a biologically inspired optimization approach that combines Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) to perform intelligent feature selection before final segmentation is carried out by the HSCS-Net. This hybridization ensures the selection of the most informative and discriminative features, which ultimately improves segmentation performance while reducing redundant or irrelevant data.

The pipeline begins with the initialization phase, where a population of binary solution vectors $X_i^{(0)}) \in (0,1)^n$ is generated. Each vector represents a subset of the entire feature matrix $F = f_1, f_2, \ldots, f_n$, with a '1' indicating selection and a '0' indicating exclusion of the corresponding feature. For each solution vector, the subset of features it represents is used to perform segmentation via the HSCS-Net, and the segmentation output is evaluated using performance metrics such as Dice Score or Intersection over Union (IoU). This evaluation serves as the fitness function J determining how suitable each solution is. The best-performing individual in the population is selected as the global best (GB) in the context of PSO, while the top three solutions are labelled as B_1,B_2,B_3 and are used for GWO updates, reflecting the alpha, beta, and delta wolves in the optimization hierarchy.The core of the algorithm lies in Step 1: Hybrid Optimization, where both PSO and GWO strategies are applied iteratively over a defined number of iterations T. In the PSO update, each solution adjusts its position based on its own previous best solution and the global best, incorporating random perturbations controlled by coefficients c_1 and c_2. The velocity vector is updated, and the position is transformed using a sigmoid activation function to produce a binary decision through thresholding. This ensures that the updates remain within a binary search space suitable for feature selection tasks.Concurrently, the GWO update computes new candidate solutions based on the simulated hunting behaviour of grey wolves. Distances from each of the top three wolves (B_1, B_2, B_3) to the current solution are calculated, and new positions X_1, X_2, X_3 are computed by modulating these distances with coefficients A and C, which are themselves randomly generated to simulate exploration and exploitation. The final updated position for the candidate solution is derived by averaging the three positions, and this average is again thresholder to produce a binary feature mask. The fitness of the updated solution is then reevaluated using the HSCS-Net segmentation output.After completing all iterations, the algorithm proceeds to Step 2: Feature Selection and Segmentation. Here, the best solution vector Xbest from the final iteration defines the optimal subset of features F^*. These selected features are then passed to the HSCS-Net model to generate the final segmentation output S. This step ensures that only the most relevant features, as determined by the hybrid optimization process, contribute to the segmentation, resulting in better performance and reduced computational overhead. In summary, this hybrid PSO-GWO algorithm serves as a powerful metaheuristic wrapper around the HSCS-Net segmentation model. It systematically explores the feature space to eliminate redundancy, improve segmentation accuracy, and enhance model generalization. The biologically inspired design effectively balances exploration and exploitation, yielding robust feature selection tailored to the underlying segmentation objective.

Hawk-Swarm Attention-Driven Encoder-Decoder Architecture

The Hawk-Swarm Attention-Driven Encoder-Decoder is the main framework of HSCSNet. It is aimed at improving the segmentation task by incorporating hierarchical attentionand context at multiple scales. This architecture includes three major parts: (1) Encoder with Hawk gating and SE attention paced squeeze-and-excitation (SE) attention, (2) Bridge block with Swarmception residual fusion and ASPP, and (3) Decoder with adaptive skip fusion which is highlighted in these illustrations. During the encoder phase, the input feature tensor F in $RHWC$, obtained from previous Inception modules, goes through hierarchical down-sampling through a series of strided convolutional layers. For each level I, the encoder applies a standard convolution with stride s=2 to reduce spatial dimensions and extract coarse features. These downsampled features, denoted as E^l are then refined using squeeze-and-excitation (SE) attention. The SE module computes a global descriptor z^l via global average pooling (GAP), which is subsequently transformed using a two-layer fully connected network with non-linearities (ReLU and sigmoid) to produce channel attention weights s^l. These weights recalibrate the feature maps via channel-wise scaling, allowing the network to emphasize more informative channels and suppress less relevant ones. Building upon this, a novel Hawk attention gate is introduced to perform spatial modulation. The recalibrated feature maps E^l are passed through a convolutional gating unit, where a batch normalization layer and a sigmoid activation produce a spatial gate G^l. This gate performs element-wise multiplication with the recalibrated features, resulting in spatially attentive representations $E^{(}l_{)}$. These modulated features are stored

Algorithm 3 Hybrid PSO-GWO Feature Selection Integrated with HSCS-Net for Liver and Tumor Segmentation

Input: Extracted feature matrix $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, population size P, number of iterations T, CT image $\mathbf{I}$

Output: Final segmentation mask $\mathbf{S} \in \mathbb{R}^{H \times W}$

Initialization:

Initialize a population of binary solution vectors $\mathcal{X}_i^{(0)} \in \{0,1\}^n$ for $i = 1, 2, \ldots, P$

Each vector $\mathcal{X}_i$ encodes feature selection: $\mathcal{F}_i = \{f_j \in \mathcal{F} \mid \mathcal{X}_{i,j} = 1\}$

Evaluate the initial fitness of each particle/wolf:

$$\text{Fitness}(\mathcal{X}_i) = \mathcal{J}(HSCS(\mathbf{I}, \mathcal{F}_i)) = \text{Dice Score, IoU, etc.}$$

Identify initial global best (PSO): $\text{GB} \leftarrow \arg\max \text{Fitness}(\mathcal{X}_i)$

Identify top three wolves (GWO): $\text{B}_1, \text{B}_2, \text{B}_3 \leftarrow$ Top-3 sorted $\mathcal{X}_i$

Step 1: Hybrid Optimization via PSO and GWO

for each iteration $t = 1$ to T **do**

  for each particle/wolf $i = 1$ to P **do**

    PSO Update: Update velocity and position

$$v_i^{(t+1)} = w \cdot v_i^{(t)} + c_1 r_1 \cdot (p_i - \mathcal{X}_i^{(t)}) + c_2 r_2 \cdot (\text{GB} - \mathcal{X}_i^{(t)})$$

$$\mathcal{X}_i^{(t+1)} = \text{sigmoid}(v_i^{(t+1)}) > \theta \quad \text{(binary update)}$$

    GWO Update: Compute new position from top wolves

$$D_1 = |C_1 \cdot \text{B}_1 - \mathcal{X}_i^{(t)}|, \quad X_1 = \text{B}_1 - A_1 \cdot D_1$$

$$D_2 = |C_2 \cdot \text{B}_2 - \mathcal{X}_i^{(t)}|, \quad X_2 = \text{B}_2 - A_2 \cdot D_2$$

$$D_3 = |C_3 \cdot \text{B}_3 - \mathcal{X}_i^{(t)}|, \quad X_3 = \text{B}_3 - A_3 \cdot D_3$$

$$\mathcal{X}_i^{(t+1)} = \frac{X_1 + X_2 + X_3}{3}$$

    Threshold $\mathcal{X}_i^{(t+1)}$ to binary: $\mathcal{X}_i^{(t+1)} \in \{0,1\}^n$

    Re-evaluate fitness:

$$\text{Fitness}(\mathcal{X}_i^{(t+1)}) = \mathcal{J}(HSCS(\mathbf{I}, \mathcal{F}_i))$$

  end for

  Update personal bests and global best $\text{GB} \leftarrow \arg\max \text{Fitness}(\mathcal{X}_i)$

  Update top 3 wolves $\text{B}_1, \text{B}_2, \text{B}_3 \leftarrow \text{Sorted}(\mathcal{X}_i)$

end for

Step 2: Feature Selection and Segmentation

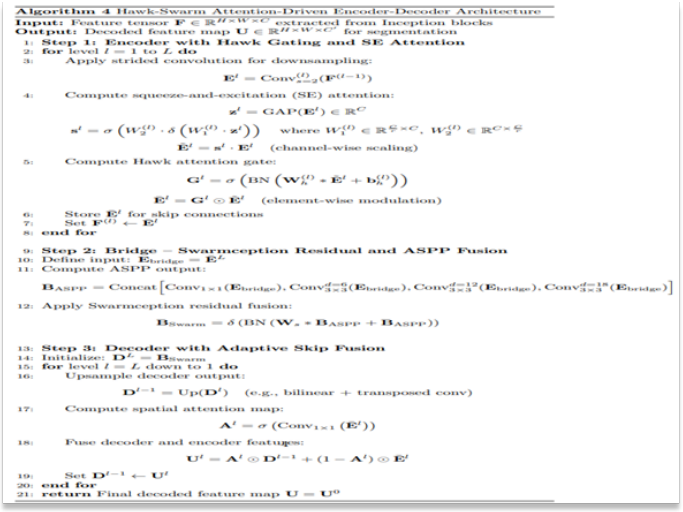
Extract optimal subset: $\mathcal{F}^* = \mathcal{F}_{\mathcal{X}_{\text{best}}}$

Run HSCS-Net on selected features:

$$\mathbf{S} = HSCS^{3}(\mathbf{I}, \mathcal{F}^*)$$

return S as the final segmented mask

for skip connections and forwarded through the encoder hierarchy, creating a multiscale and attention-enriched encoding.



Algorithm 4 Hawk-Swarm Attention-Driven Encoder-Decoder Architecture

Input: Feature tensor $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ extracted from Inception blocks
Output: Decoded feature map $\mathbf{U} \in \mathbb{R}^{H \times W \times C'}$ for segmentation

Step 1: Encoder with Hawk Gating and SE Attention
for level $l = 1$ to L **do**
  Apply strided convolution for downsampling:

$$\mathbf{E}^l = \text{Conv}_{s=2}^{(l)}(\mathbf{F}^{(l-1)})$$

  Compute squeeze-and-excitation (SE) attention:

$$\mathbf{z}^l = \text{GAP}(\mathbf{E}^l) \in \mathbb{R}^C$$

$$\mathbf{s}^l = \sigma\left(W_2^{(l)} \cdot \delta\left(W_1^{(l)} \cdot \mathbf{z}^l\right)\right) \quad \text{where } W_1^{(l)} \in \mathbb{R}^{\frac{C}{r} \times C},\ W_2^{(l)} \in \mathbb{R}^{C \times \frac{C}{r}}$$

$$\mathbf{E}^l = \mathbf{s}^l \cdot \mathbf{E}^l \quad \text{(channel-wise scaling)}$$

  Compute Hawk attention gate:

$$\mathbf{G}^l = \sigma\left(\text{BN}\left(\mathbf{W}_h^{(l)} * \mathbf{E}^l + \mathbf{b}_h^{(l)}\right)\right)$$

$$\mathbf{E}^l = \mathbf{G}^l \odot \mathbf{E}^l \quad \text{(element-wise modulation)}$$

  Store $\mathbf{E}^l$ for skip connections
  Set $\mathbf{F}^{(l)} \leftarrow \mathbf{E}^l$
end for

Step 2: Bridge – Swarmception Residual and ASPP Fusion
Define input: $\mathbf{E}_{\text{bridge}} = \mathbf{E}^L$
Compute ASPP output:

$$\mathbf{B}_{\text{ASPP}} = \text{Concat}\left[\text{Conv}_{1\times1}(\mathbf{E}_{\text{bridge}}), \text{Conv}_{3\times3}^{d=6}(\mathbf{E}_{\text{bridge}}), \text{Conv}_{3\times3}^{d=12}(\mathbf{E}_{\text{bridge}}), \text{Conv}_{3\times3}^{d=18}(\mathbf{E}_{\text{bridge}})\right]$$

Apply Swarmception residual fusion:

$$\mathbf{B}_{\text{Swarm}} = \delta\left(\text{BN}\left(\mathbf{W}_s * \mathbf{B}_{\text{ASPP}} + \mathbf{B}_{\text{ASPP}}\right)\right)$$

Step 3: Decoder with Adaptive Skip Fusion
Initialize: $\mathbf{D}^L = \mathbf{B}_{\text{Swarm}}$
for level $l = L$ down to 1 **do**
  Upsample decoder output:

$$\mathbf{D}^{l-1} = \text{Up}(\mathbf{D}^l) \quad \text{(e.g., bilinear + transposed conv)}$$

  Compute spatial attention map:

$$\mathbf{A}^l = \sigma\left(\text{Conv}_{1\times1}\left(\mathbf{E}^l\right)\right)$$

  Fuse decoder and encoder features:

$$\mathbf{U}^l = \mathbf{A}^l \odot \mathbf{D}^{l-1} + (1 - \mathbf{A}^l) \odot \mathbf{E}^l$$

  Set $\mathbf{D}^{l-1} \leftarrow \mathbf{U}^l$
end for
return Final decoded feature map $\mathbf{U} = \mathbf{U}^0$

The bridge stage operates at the bottleneck of the encoder-decoder architecture. It begins with defining the final encoder output $E^(l_)$ as the bridge input E_bridge. The architecture then applies Atrous Spatial Pyramid Pooling (ASPP) with multiple dilation rates (6, 12, 18) to extract features at varying receptive fields, enhancing the context sensitivity of the model without increasing computation. The output from ASPP is passed through a custom Swarmception residual fusion module. This fusion block performs a residual aggregation of the ASPP features, followed by batch normalization and a ReLU activation, effectively learning deeper interactions among multiscale context representations.

In the decoder stage, the architecture initiates the reconstruction of segmentation masks from the bottleneck representation. Starting from the deepest level, the decoder performs upsampling through a combination of bilinear interpolation and transposed convolutions to increase spatial resolution. At each decoding level, the model retrieves the corresponding encoder feature $E^(l_)$ stored from the encoding phase. It computes a spatial attention map A^l using a 1×11 11×1 convolution followed by a sigmoid activation applied to the encoder feature. This attention map guides the adaptive fusion of encoder and decoder features. The final decoded feature map U^l at each level is generated as a weighted combination of decoder output and encoder features, where A^l controls the balance between new predictions and spatially rich skip features. This ensures that fine details suppressed during down-sampling are adaptively reintroduced during reconstruction.The decoder concludes by outputting the final feature map $U = U^0$, which contains the segmentation-relevant spatial and contextual information necessary to produce the final prediction. This architecture, through its combined use of SE and spatial attention (Hawk gating), context-enhancing ASPP, and adaptive skip connections, is designed to preserve structural integrity and improve boundary delineation in medical image segmentation.

The HSCS-NET network as shown by Algorithm 2 works as follows:

The Inception modules, with their diverse convolutional filters, extract rich and varied features from the input CT images. These features capture different spatial resolutions and semantic details.

Cross-Feature Connection, Features from the Inception modules are passed through cross-feature connections to the corresponding layers in Encoder Decoder Module 's decoder. This cross-connection ensures that the detailed multi-scale features are integrated with the spatial information from the encoder.

As the information from the Inception modules merges with the Encoder Decoder Module decoder, the network can utilize both high-level contextual information and fine-grained details for segmentation tasks. This hybrid approach results in more accurate and precise segmentation of liver and liver tumors.

RESULTS AND DISCUSSION

This section evaluates the quantitative and qualitative assessment of our liver segmentation method, focusing on evaluation of the measurements of Dice Coefficient, Accuracy, and Precision. The assessment of efficacy of our technique takes into account the hepatic region's division with these factors. The Dice Coefficient quantifies the overlap between prediction and actual liver region segmentation while Accuracy evaluates the over-all correctly segmented areas over the total area. Precision evaluates the segmented area counted as detected where stream of liver recognition prediction occurs while truly positive cases only comprise of a subset. All these metrics mentioned above form a composite description covering the strengths and weaknesses of our segmentation method. The training process of HSCS-Net uses systematic approaches aimed at maximum accuracy in the segmentation tasks of medical images. Image enhancement involves a series of processes starting from normalization, then resizing followed by data augmentation to increase distribution using methods such as elastic deformation, contrast adjustment, and rotation and flipping. To ensure unbiased evaluation, the dataset is divided into three equal parts comprising of training, validation, and testing in the ratio of 80:10:10. An initialization step forms a network having four encoder blocks and five decoding blocks constructed upon the Inception backbone network. A feature selection optimization strategy uses particle swarm optimization and grey wolf optimization together for maintaining discriminate spatial features during selection. Feature representation quality increases through the combination of skip connections and attention mechanisms which maintain vital information. The algorithm 4 uses dice loss together with cross-entropy loss as the training mechanism to achieve accurate segmentation and discrimination between different classes. The training utilizes Adam optimizer with 1e-4 as the initial learning rate but incorporates cosine annealing scheduling to control the learning rate dynamics across its sessions. The network performs forward propagation to generate an output that backpropagation modifies according to its weights. The algorithm stops training when the dice coefficient on validation data fails to enhance during ten successive epochs in order to avoid overfitting. The top model selection happens through validation metric assessments after which the test set receives evaluation through dice coefficient measures alongside accuracy and precision and recall and jaccard similarity metrics. HSCS-Net to effectively obtain spatial features while optimizing features and achieving the highest possible accuracy in medical image segmentation tasks.

Ablation Study

The ablation study table analyzes different model configurations by evaluating their performance through Dice Coefficient, Accuracy and Precision statistics. The different model versions use varying numbers of encoder and decoder layers as well as backbone architectures. The study investigates the impact of model depth and feature selection strategies, specifically the integration

of Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO), on segmentation performance.

The HSCS-Net model which uses four encoder layers followed by five decoder layers obtains maximum performance results on all three metrics when coupled with the Inception backbone and PSO/GWO feature selection. The model achieves highest performance values of Dice Coefficient 0.98 and Accuracy 0.9891 combined with Precision 0.9901.

The segmentation performance decreases as the model architecture becomes shallower when using 3-encoder and 3-decoder configurations because deeper networks enable more effective spatial feature extraction and representation. A 3-encoder, 3-decoder model built with ConvexNet obtains a Dice Coefficient measurement of 0.94 however this value remains significantly lower than Dice Coefficient outcomes achieved with either 4-encoder or 5-encoder model architecture designs. The segmentation performance improves when encoder and decoder layers become deeper because deeper networks enable better identification of complex spatial patterns in the input data.

Among available backbone architectures Inception-based models show higher performance than both ResNet and ConvexNet models. The top non-Inception model combines ResNet architecture with a 4-encoder along with 5-decoder structure to reach Dice Coefficient performance at 0.96 andPrecision at 0.9793 with Accuracy at 0.9782. The performance achieved by Inception-based models surpasses other networks despite exhibiting overall measured results that are slightly weaker than the Inception pipelines. The incorporation of PSO and GWO significantly enhances performance by optimizing feature selection. This is particularly evident in models that use both optimization techniques, as they consistently outperform configurations without them. The results affirm that hybrid feature selection methods play a crucial role in refining feature extraction and improving classification robustness.

The Figure 2 visualizes the performance comparison of different model configurations based on three key metrics: Dice Coefficient, Accuracy, and Precision. The proposed HSCS-Net (4Enc-5Dec) achieves the highest scores across all three metrics, demonstrating superior segmentation performance. Inception-based architectures with Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO) show strong results, outperforming ResNet and ConvNet variants. Notably, configurations with additional encoder and decoder layers (4Enc-4Dec and 4Enc-5Dec) tend to enhance performance, highlighting the importance of deeper architectures in improving segmentation accuracy. The plot illustrates a clear trend where hybrid optimization techniques (PSO + GWO) contribute to higher precision and robustness in segmentation.

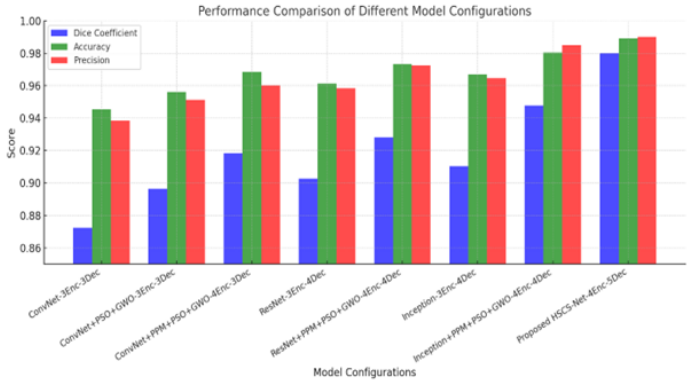


Figure 2 Ablation Study Performance

Quantitative Analysis

The performance of the proposed HSCS-NET Network in detecting liver and liver tumors from CT scan slices was assessed using confusion matrices. Figure 3 Confusion Matrix displays the confusion matrix, which highlights the detection accuracy across various liver and liver tumor scenarios within the dataset.

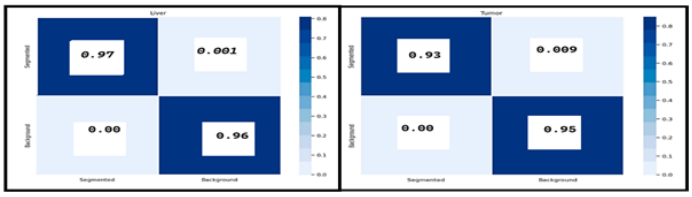


Figure 3 Confusion Matrix

In Table 2 State of the art analysis, showcases the comparison between proposed HSCS-NET Network and other state of the art techniques. The combination of high Dice Coefficient, accuracy, and precision underscores the effectiveness of the HSCS-NET Network model in liver tumor segmentation, demonstrating its ability to provide accurate and reliable diagnostic support.

Table 2 State of the art analysis

References Existing and Proposed	Dice Coefficient	Accuracy	Precision
Christ (2017) – CNN-based	0.823	–	–
Sun *et al.* (2017) – FCNs-based	–	–	–
Wu *et al.* (2017) – Fuzzy C-means & GC	0.830	–	–
Lu *et al.* (2020) – VGG & ELM	0.670	–	–
Muhammad and Zhang (2024) – ResNet	0.870	0.945	0.930
Kaur and Kaur (2024) – PSO-PSP-Net + InceptionV3	–	0.9754	0.9632
Proposed – HSCS-NET Network	0.980	0.9891	0.9901

Table 2 showcases the results of existing and proposed methods using Dice Coefficient, Accuracy and Precision. Previous CNN and/or clustering-based methods Christ (2017); Wu *et al.* (2017) achieved Dice scores around 0.82–0.83 whereas Lu *et al.* (2020) reported a Dice Coefficient of 0.67. New deep learning models perform better, including Muhammad and Zhang (2024) with a Dice score of 0.87, an accuracy of 0.945, and Kaur and Kaur (2024) achieving high accuracy (0.9754) and precision (0.9632). The developed HSCS-NET yields better results compared to all the methods and achieves 0.98 in Dice Coefficient, accuracy of 0.9891 and a specificity of 0.9772.

Table 1 HSCS-NET Ablation Study

Model Configuration	Encoder Depth	Decoder Depth	Backbone	PSO + GWO Optimization	Dice Coefficient ↑	Accuracy ↑	Precision ↑
Baseline ConvNet Segmentation	3	3	ConvexNet		0.8723	0.9452	0.9384
ConvNet + Hybrid Optimization	3	3	ConvexNet		0.8964	0.9561	0.9513
ConvNet + PPM + Hybrid Optimization	4	3	ConvexNet		0.9182	0.9685	0.9602
ResNet-based Segmentation	3	4	ResNet		0.9027	0.9613	0.9584
ResNet + PPM + Hybrid Optimization	4	4	ResNet		0.9281	0.9734	0.9725
Inception-based Segmentation	3	4	Inception		0.9104	0.9668	0.9647
Inception + PPM + Hybrid Optimization	4	4	Inception		0.9476	0.9803	0.9851
Proposed – HSCS-Net	4	5	Inception		0.9800	0.9891	0.9901

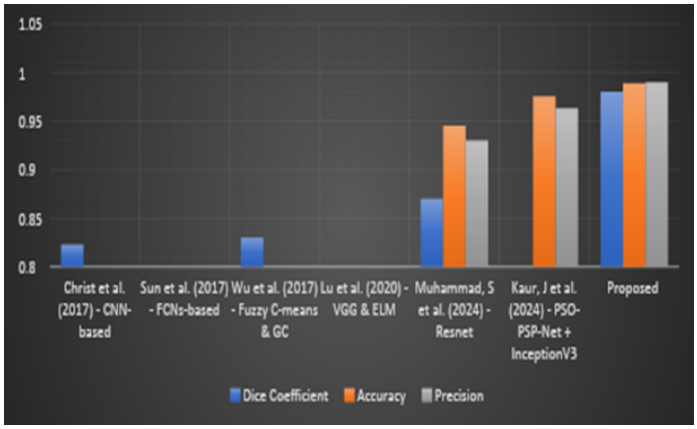


Figure 4 Comparison of our proposed model with existing state of the art methods

The graph illustrates a comparative performance analysis of existing state-of-the-art methods and the proposed HSCS-NET model based on Dice Coefficient, Accuracy, and Precision. The horizontal axis represents different methods, while the vertical axis indicates the corresponding performance scores. Earlier approaches show comparatively lower metric values, reflecting limited segmentation accuracy. Recent deep learning–based methods demonstrate noticeable improvement; however, variations in reported metrics indicate inconsistent performance across models.

The proposed HSCS-NET consistently achieves the highest values across all metrics, forming the peak of the graph. This clear performance gap highlights the effectiveness and robustness of the proposed approach over existing methods. Overall, the graphical trend confirms that HSCS-NET provides superior segmentation accuracy, precision, and reliability compared to both traditional and recent deep learning techniques.

Qualitative analysis

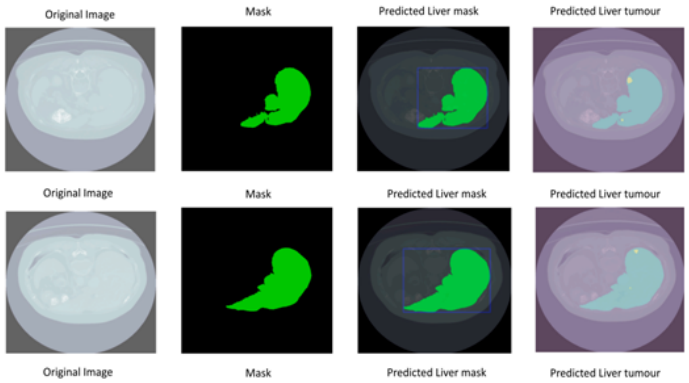


Figure 5 Liver Prediction

Figure 5 Liver Prediction demonstrates the deployment of semantic segmentation for liver and tumor lesion detection on the LiTS dataset. This figure highlights the system implementation together with its output. Semantic segmentation represents an essential medical imaging technique which divides every image pixel so healthcare specialists can distinguish healthy liver tissue from malignant lesions. The illustration shows the initial photographs alongside the generated masks used for segmenting key areas alongside model projection outputs. The masks efficiently display

the boundaries of liver tissue together with tumor locations which provides a clear visual assessment ability for both original images and segmented outputs. The illustration demonstrates how the model effectively separates distinct tissues which remains essential for doctors to make exact diagnosis and plan treatment strategies when treating liver cancer. Results from model forecasts help determine its effectiveness in operational healthcare establishments.

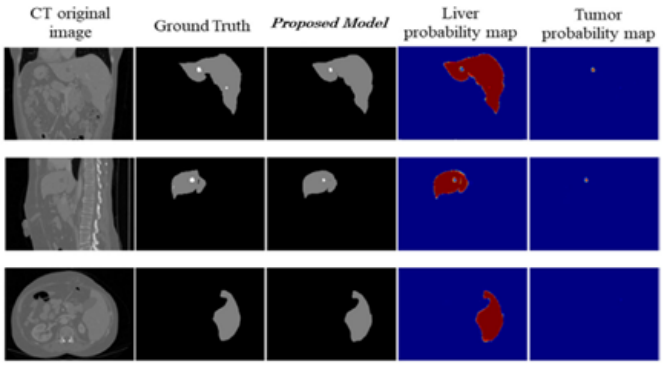


Figure 6 Liver Tumor Prediction

Dimensional Dual Path-Convolutional neural networks (TDP-CNN) together with Fully connected conditional Random fields segment both liver and tumors as shown in Figure 6 Liver Tumor Prediction. The presented segmentation outcomes display several performance outcomes. A small segment of tumor tissue shows how the model performs well for detecting restricted tumor areas and keeping them distinct. The model demonstrates its accuracy through the identification of major tumors by showing its ability to precisely retrieve substantial tumor tissues while preserving unharmed liver areas. The model demonstrates its effectiveness in intricate lesion cases through the inclusion of a sample which contains multiple tumor locations. The TDP-CNN model demonstrates reliable capabilities for detecting liver cancers accurately because of its adjustable precision in medical diagnosis and therapeutic strategy assessment.

CONCLUSION

Applying the HSCS-NET Network model achieved impressive results in terms of liver tumor detection throughput the segmentation of liver tumors achieving a Dice Coefficient of 0.98 and an accuracy of 0.9891. The model demonstrated better pixel classification than the unnamed benchmarks from other studies. The model achieved high precision performance of 0.9901 which demonstrated the model?s ability to identify tumor regions with very few misclassifications of non-tumor regions. The HSCS-NET Network demonstrated solid overall performance in both quantitative and qualitative assessments benchmarked against current state-of-the-art models. While the results with the HSCS-NET Network are encouraging, the domain adaptation problem, dataset variability, and the integration of clinical settings into the model still need additional effort. Future research in these areas could focus on creating hybrid models with domain adaptation and transfer learning techniques to improve generalization across diverse medical imaging databases. The incorporation of multiple image data types into a single system requires more efficient computing systems in order to seamlessly integrate multiple imaging data types; with this, the clinical application of CNN-based models will become more reliable. The clinical value of deep learning models as diagnostic tools for liver cancer will increase as these initiatives are undertaken.

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.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used artificial intelligence tools in order to improve the readability and language quality of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

LITERATURE CITED

Banerjee, T., 2025a Attentive CNN EEG or ACE-SeizNet: An Attention-Enhanced CNN Model for Automated EEG-Based Seizure Detection .

Banerjee, T., 2025b IMATX: An Integrated Multi-Context Pyramidal Framework for Explainable AI Predictions .

Banerjee, T., 2025c Towards Automated and Reliable Lung Cancer Detection Using DY-FSPAN. Computational Biology and Chemistry p. 108500.

Banerjee, T., D. Batta, A. Jain, S. Karthikeyan, H. Mehndiratta, *et al.*, 2021a Deep Belief CNN with GAN-Based Diagnosis of Pneumonia. In *ICEEE 2021*, Springer.

Banerjee, T., D. Butta, A. Jain, K. S. Biradar, R. R. Koripally, *et al.*, 2021b Deep Belief CNN for Diagnosis of Pneumonia. In *ICAECT 2021*, IEEE.

Banerjee, T., A. Jain, S. C. Sethuraman, S. C. Satapathy, S. Karthikeyan, *et al.*, 2022a Deep CNN (Falcon) and Transfer Learning-Based Approach to Detect Malarial Parasite. Multimedia Tools and Applications **81**: 13237–13251.

Banerjee, T., Y. F. Khan, T. Rafiq, S. Singh, R. Wason, *et al.*, 2025 HHO-UNet-IAA: Harris Hawks Optimization Based UNet Architecture for Glaucoma Segmentation. International Journal of Information Technology .

Banerjee, T., A. Sharma, K. Charvi, S. Raman, and S. Karthikeyan, 2022b Attention-Based Discrimination of Mycoplasma Pneumonia. In *ICCIDE 2021*, Springer.

Banerjee, T., A. Sharma, K. Charvi, S. Raman, R. G. Regalla, *et al.*, 2022c Journey of Letters to Vectors Through Neural Networks. In *ICDAM 2021*, Springer.

Banerjee, T., K. V. P. Srikar, S. A. Reddy, K. S. Biradar, R. R. Koripally, *et al.*, 2021c Hand Sign Recognition Using Infrared Imagery. In *ICIPTM 2021*, IEEE.

Bilic, P., P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, *et al.*, 2019 The Liver Tumor Segmentation Benchmark (LiTS). arXiv preprint arXiv:1901.04056 .

Cakmak, Y., I. Pacal, *et al.*, 2026 A comparative analysis of transformer architectures for automated lung cancer detection in ct images. Journal of Intelligent Decision Making and Information Science **3**: 528–539.

Christ, P. F., 2017 *Convolutional Neural Networks for Classification and Segmentation of Medical Images*. Ph.D. thesis, Technische Universität München, Munich, Germany.

Christ, P. F., M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, *et al.*, 2016 Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 415–423, Athens, Greece, Springer.

Jiang, H., T. Shi, Z. Bai, and L. Huang, 2019 AHCNet: An Application of Attention Mechanism and Hybrid Connection for Liver Tumor Segmentation in CT Volumes. IEEE Access **7**: 24898–24909.

Jin, Q., Z. Meng, C. Sun, H. Cui, and R. Su, 2020 RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans. Frontiers in Bioengineering and Biotechnology **8**: 1471.

Karthikeyan, S., S. Gopikrishnan, D. Batta, and T. Banerjee, 2021 Double Helical Ensemble Neural Network to Analyze Driving Pattern. Turkish Journal of Computer and Mathematics Education **12**: 6447–6458.

Kaur, J. and P. Kaur, 2024 PSO-PSP-Net + InceptionV3: An Optimized Hyper-Parameter Tuned CAD Model for Liver Tumor Detection. Biomedical Signal Processing and Control **95**: 106442.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2017 ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM **60**: 84–90.

Li, D., L. Liu, J. Chen, H. Li, and Y. Yin, 2014 A Multistep Liver Segmentation Strategy by Combining Level Set Based Method with Texture Analysis for CT Images. In *International Conference on Orange Technologies*, pp. 109–112, Xi'an, China, IEEE.

Li, Q., M. Cao, L. Lei, F. Yang, H. Li, *et al.*, 2022 Burden of Liver Cancer: From Epidemiology to Prevention. Chinese Journal of Cancer Research **34**: 554.

Li, X., H. Chen, X. Qi, Q. Dou, C.-W. Fu, *et al.*, 2018 H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. IEEE Transactions on Medical Imaging **37**: 2663–2674.

Lu, S., K. Xia, and S.-H. Wang, 2020 Diagnosis of Cerebral Microbleed via VGG and Extreme Learning Machine Trained by Gaussian Map Bat Algorithm. Journal of Ambient Intelligence and Humanized Computing **14**: 5395–5406.

Muhammad, S. and J. Zhang, 2024 Segmentation of Liver Tumors by MONAI and PyTorch in CT Images with Deep Learning Techniques. Applied Sciences **14**: 5144.

Pacal, I. and Y. Cakmak, 2025 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.

Peesa, R. B., A. Satpathy, S. Karthikeyan, M. Bisht, T. Banerjee, *et al.*, 2020 Single Node Hadoop Cluster for Small Scale Industrial Automation. In *ICCCA 2020*, IEEE.

Rehman, A., T. Mahmood, and T. Saba, 2025 Robust Kidney Carcinoma Prognosis Using Swin-ViT and DeepLabV3+. Applied Soft Computing **170**: 112518.

Saminathan, K., T. Banerjee, D. P. Rangasamy, and M. Vimal Cruz, 2024 Segmentation of Thoracic Organs Using Resio-Inception U-Net. Current Gene Therapy **24**: 217–238.

Singh, D. P., T. Banerjee, P. Kour, D. Swain, and Y. Narayan, 2025a CICADA (UCX): Automated Breast Cancer Classification. Computational Biology and Chemistry p. 108368.

Singh, D. P., P. Kour, T. Banerjee, and D. Swain, 2025b Review of Machine Learning Models for Anti-Cancer Drug Response Prediction. Archives of Computational Methods in Engineering .

Soler, L. *et al.*, 2010 3D-IRCADb-01: A 3D Imaging Dataset of Liver Tumors. IRCAD, Strasbourg, France.

Song, X., M. Cheng, B. Wang, S. Huang, X. Huang, *et al.*, 2013 Adaptive Fast Marching Method for Automatic Liver Segmentation from CT Images. Medical Physics **40**: 091917.

Sun, C., S. Guo, H. Zhang, J. Li, M. Chen, *et al.*, 2017 Automatic Segmentation of Liver Tumors from Multiphase Contrast-Enhanced CT Images Based on FCNs. Artificial Intelligence in Medicine **83**: 58–66.

World Health Organization, 2021 Cancer. Available online.

Wu, W., S. Wu, Z. Zhou, R. Zhang, and Y. Zhang, 2017 3D Liver Tumor Segmentation in CT Images Using Improved Fuzzy C-Means and Graph Cuts. BioMed Research International p. 5207685.

Yasaka, K., H. Akai, O. Abe, and S. Kiryu, 2018 Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-Enhanced CT. Radiology **286**: 887–896.

How to cite this article: Chhabra, P., Singh, .S and Banerjee, T. Hawk-Swarmception Segmentation Network (HSCS-Net): Enhanced Liver Tumor Segmentation with Receptive Field Optimization and Clinical Data-Guided Feature Selection via PSO and GWO. *Artificial Intelligence in Applied Sciences*, 2(3), 15-26, 2026.

ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.27-31
https://doi.org/10.69882/adba.ai.2026014

A Comparative Evaluation of QLoRA and AdaLoRA for Parameter-Efficient Fine-Tuning of Large Language Models on Medical Textbook Question Answering

Seda Bayat Toksoz *,1 **and Gultekin Isik** α,2
*,α Department of Computer Engineering, Igdir University, Igdir, Turkiye.

ABSTRACT

Parameter-efficient fine-tuning methods have emerged as practical solutions for adapting large language models to specialized domains while minimizing computational overhead. This study presents a systematic comparison of two prominent approaches, QLoRA and AdaLoRA, for fine-tuning instruction-tuned language models on medical textbook question answering. We evaluated both methods using two backbone architectures, Llama-3-8B-Instruct and Qwen2-7B-Instruct, on a dataset comprising 6,500 question-answer pairs derived from 13 authoritative medical textbooks spanning diverse clinical and biomedical disciplines. Our experiments demonstrate that QLoRA consistently outperforms AdaLoRA under single-epoch training conditions, achieving validation perplexity values of 1.085 and 1.086 for Llama-3 and Qwen2, respectively, compared to AdaLoRA's 1.125 and 1.169. These results correspond to relative validation loss reductions of 30.8% for Llama-3 and 47.5% for Qwen2 when using QLoRA over AdaLoRA. Both methods maintained comparable trainable parameter counts, approximately 167 million for Llama-3 and 161 million for Qwen2, representing roughly 3.5% of total model parameters. Our findings indicate that QLoRA provides more stable convergence behavior within limited training budgets, while AdaLoRA's adaptive rank allocation mechanism may require extended training schedules to realize its theoretical advantages. These results offer practical guidance for deploying parameter-efficient fine-tuning in medical natural language processing applications where computational resources are constrained.

KEYWORDS

Parameter-efficient fine-tuning
QLoRA
AdaLoRA
Large language models
Medical question answering
Low-rank adaptation

INTRODUCTION

Large language models have demonstrated remarkable capabilities across diverse natural language processing tasks, establishing new performance benchmarks in text generation, comprehension, and reasoning (Brown *et al.* 2020; Chowdhery *et al.* 2023). The medical domain presents a particularly compelling application area, where accurate information retrieval and question answering can support clinical decision-making, medical education, and patient care (Thirunavukarasu *et al.* 2023; Singhal *et al.* 2023). However, the computational demands of training and deploying these models at scale present significant barriers, especially for healthcare institutions operating under resource constraints (Karabacak and Margetis 2023).

The standard approach of full fine-tuning, which updates all model parameters during training, becomes prohibitively expensive as model sizes grow into the billions of parameters. A model containing 7 to 8 billion parameters requires substantial GPU memory merely for inference, and fine-tuning such models with full precision necessitates specialized hardware configurations that remain inaccessible to many research groups and clinical settings. This computational barrier has motivated the development of parameter-efficient fine-tuning techniques, which seek to adapt pretrained models to downstream tasks while modifying only a

Manuscript received: 24 October 2025,
Revised: 28 December 2025,
Accepted: 20 January 2026.

[1] seda.bayat@igdir.edu.tr (**Corresponding author**).
[2] gultekin.isik@igdir.edu.tr

small fraction of the total parameters (Lialin *et al.* 2023; Han *et al.* 2024). Similar approaches have been successfully applied in other domains, including financial sentiment analysis using parameter-efficient methods (Bayat Toksoz and Isik 2025).

Low-Rank Adaptation, or LoRA, introduced by Hu *et al.* (2022), represents a seminal contribution to this field. The core insight underlying LoRA is that the weight updates during fine-tuning can be effectively approximated by low-rank matrices, thereby dramatically reducing the number of trainable parameters. Rather than updating the full weight matrix W, LoRA decomposes the update into two smaller matrices: DeltaW = BA, where B and A are low-rank matrices with the rank r chosen to be much smaller than either dimension. This formulation maintains the expressiveness needed for task adaptation while reducing memory requirements by orders of magnitude.

Building upon LoRA, Dettmers *et al.* (2024) proposed QLoRA, which combines low-rank adaptation with 4-bit quantization of the base model weights. QLoRA introduces several technical innovations, including a novel 4-bit NormalFloat data type optimized for normally distributed weights, double quantization to reduce memory overhead from quantization constants, and paged optimizers to handle memory spikes during training. These contributions enable fine-tuning of models with tens of billions of parameters on single consumer-grade GPUs, democratizing access to large-scale language model adaptation.

An alternative approach, AdaLoRA, proposed by Zhang *et al.* (2023), extends the LoRA framework by introducing adaptive budget allocation. Rather than assigning fixed ranks to all weight matrices, AdaLoRA parameterizes the weight updates using singular value decomposition and dynamically adjusts the rank of each layer based on learned importance scores. The method prunes less important singular values during training, theoretically allocating more capacity to layers that contribute most significantly to task performance. This adaptive mechanism promises improved parameter efficiency by concentrating trainable parameters where they matter most.

This study addresses three primary research questions. First, we investigate how QLoRA and AdaLoRA compare in terms of validation loss and perplexity when fine-tuning instruction-tuned language models on medical textbook question answering. Second, we examine whether AdaLoRA's dynamic rank allocation mechanism provides measurable advantages over QLoRA's fixed-rank approach under practical training constraints. Third, we assess the consistency of these findings across different backbone model architectures to determine whether our conclusions generalize beyond specific model families.

Our experimental contributions are threefold. We present the first systematic comparison of QLoRA and AdaLoRA specifically targeting medical textbook question answering, using a curated dataset spanning 13 medical textbooks and covering major clinical and biomedical disciplines. We provide detailed analyses of training dynamics, convergence behavior, and parameter efficiency for both methods across two distinct backbone architectures. Finally, we offer practical recommendations for practitioners seeking to deploy parameter-efficient fine-tuning in medical natural language processing applications.

RELATED WORKS

Large Language Models in Healthcare

The application of transformer-based language models to healthcare tasks has progressed through several developmental phases. Early work focused on domain-specific pretraining, with BioBERT (Lee *et al.* 2020) demonstrating that continued pretraining on biomedical literature improves performance on named entity recognition, relation extraction, and question answering tasks within the biomedical domain. ClinicalBERT (Huang *et al.* 2019) extended this approach to clinical notes, showing that exposure to electronic health record data during pretraining enhances model understanding of clinical language patterns and medical terminology.

The emergence of instruction-tuned large language models opened new possibilities for medical applications. Singhal *et al.* (2023) introduced Med-PaLM, demonstrating that appropriately prompted large language models can approach physician-level performance on medical licensing examination questions. Subsequent work explored fine-tuning strategies specifically tailored to medical dialogue and consultation scenarios. ChatDoctor (Li *et al.* 2023) applied supervised fine-tuning to create a medical conversational agent, while MedAlpaca (Han *et al.* 2023) demonstrated effective medical adaptation using instruction-following datasets derived from clinical guidelines and medical literature.

Parameter-Efficient Fine-Tuning Methods

Parameter-efficient fine-tuning encompasses a diverse family of techniques designed to adapt pretrained models while updating only a small subset of parameters. Comprehensive surveys by Lialin *et al.* (2023) and Han *et al.* (2024) provide taxonomies of these approaches, broadly categorizing them into adapter-based methods, prompt-based methods, and reparameterization-based methods.

Adapter methods insert small trainable modules between frozen pretrained layers. The original adapter formulation by Houlsby *et al.* (2019) demonstrated that adding bottleneck layers with as few as 3% additional parameters could achieve competitive performance on diverse natural language understanding benchmarks. Prompt-based methods modify the input representation rather than the model architecture. Prefix tuning (Li and Liang 2021) prepends trainable continuous vectors to the input sequence, while prompt tuning (Lester *et al.* 2021) optimizes task-specific soft prompts.

QLoRA (Dettmers *et al.* 2024) combines low-rank adaptation with aggressive quantization, introducing the 4-bit NormalFloat format that preserves information content while reducing memory footprint. AdaLoRA (Zhang *et al.* 2023) addresses a potential limitation of standard LoRA, namely the assumption that all weight matrices benefit equally from a given rank allocation. The method parameterizes weight updates as DeltaW = PAQ, where A is a diagonal matrix of singular values that can be pruned based on importance scores computed during training.

METHODS

Dataset Description

Our experiments utilize the MedicalTextbook_QA dataset, which comprises question-answer pairs extracted from 13 medical textbooks representing core disciplines in medical education and clinical practice. Table 1 presents the complete list of source textbooks along with their respective subject domains. Each textbook contributes 500 question-answer pairs to the dataset, yielding a total of 6,500 instances that cover anatomy, biochemistry, cell biology, gynecology, histology, immunology, neurology, obstetrics, pathology, pediatrics, pharmacology, physiology, and psychiatry.

Table 1 Medical textbooks comprising the MedicalTextbook_QA dataset, organized by clinical and basic science domains

Textbook	Domain	Samples
Gray's Anatomy	Anatomy	500
Lippincott Biochemistry	Biochemistry	500
Alberts Cell Biology	Cell Biology	500
Novak's Gynecology	Gynecology	500
Ross Histology	Histology	500
Janeway's Immunology	Immunology	500
Adams Neurology	Neurology	500
Williams Obstetrics	Obstetrics	500
Robbins Pathology	Pathology	500
Nelson Pediatrics	Pediatrics	500
Katzung Pharmacology	Pharmacology	500
Levy Physiology	Physiology	500
DSM-5	Psychiatry	500
Total		6,500

The dataset was partitioned into training and validation subsets, allocating 5,000 instances for training and 500 instances for validation. This split was performed after shuffling with a fixed random seed to ensure reproducibility across experimental conditions.

Backbone Models

We selected two instruction-tuned language models as backbone architectures for our comparative evaluation: Meta-Llama-3-8B-Instruct and Qwen2-7B-Instruct. Both models represent current state-of-the-art open-weight language models optimized for instruction following and dialogue applications. Table 2 summarizes the key specifications of both models.

Table 2 Specifications of backbone models used in experimental evaluation

Specification	Llama-3-8B	Qwen2-7B
Total Parameters	8.03B	7.62B
Quantized Params	4.71B	4.51B
Pretraining Tokens	15T	7T
Attention Type	Grouped-Query	Grouped-Query
Position Encoding	RoPE	RoPE

Parameter-Efficient Fine-Tuning Configurations

QLoRA Configuration QLoRA was implemented using the standard formulation where low-rank adapters are applied to frozen quantized base model weights. The base model weights were quantized to 4-bit precision using the NormalFloat4 data type. The low-rank adaptation matrices were configured with rank r = 64 and scaling factor alpha = 16, yielding an effective learning rate scaling of alpha/r = 0.25 applied to the adapter outputs. Dropout regularization was applied to the adapter layers with probability 0.05 to prevent overfitting.

AdaLoRA Configuration AdaLoRA was configured to enable dynamic rank allocation through importance-based pruning of singular values. The initial rank was set to r_init = 64, matching the QLoRA configuration, with a target rank of r_target = 8 representing an 87.5% reduction in rank through the pruning process. Orthogonal regularization with weight 0.5 was applied to encourage orthogonality between the left and right singular vector matrices.

Training Procedure

All experiments were conducted on a single NVIDIA A100-SXM4-80GB GPU. The training procedure employed the AdamW optimizer with a learning rate of 2e-4 and weight decay of 0.01. A cosine learning rate schedule was applied with a warmup period comprising 3% of total training steps. Each model was trained for a single epoch with a per-device batch size of 4 and gradient accumulation over 4 steps, yielding an effective batch size of 16.

Evaluation Metrics

Model performance was assessed using validation loss and perplexity as primary evaluation metrics. Perplexity, defined as the exponential of the validation loss, offers an interpretable measure of model uncertainty, with lower values indicating better predictive performance. The relationship between these metrics is expressed as: Perplexity = exp(L_val), where L_val denotes the average validation loss.

RESULTS

Overall Performance Comparison

Table 3 presents the complete experimental results across all model and method combinations. QLoRA consistently achieved lower validation loss and perplexity values compared to AdaLoRA for both backbone models. On the Llama-3-8B-Instruct backbone, QLoRA attained a validation loss of 0.0814 and perplexity of 1.085, compared to AdaLoRA's validation loss of 0.1177 and perplexity of 1.125. This represents a relative reduction of 30.8% in validation loss when using QLoRA over AdaLoRA.

The performance gap was more pronounced on the Qwen2-7B-Instruct backbone. QLoRA achieved a validation loss of 0.0821 and perplexity of 1.086, while AdaLoRA produced a validation loss of 0.1563 and perplexity of 1.169. The relative improvement in validation loss for QLoRA over AdaLoRA reached 47.5% on this backbone.

Parameter Efficiency Analysis

Both methods achieved comparable trainable parameter counts. For Llama-3-8B, QLoRA utilized 167.77 million trainable parameters while AdaLoRA used 167.79 million parameters, representing approximately 3.56% of the total 4.71 billion parameters in the quantized model. For Qwen2-7B, the trainable parameter counts were 161.48 million for QLoRA and 161.49 million for AdaLoRA, corresponding to approximately 3.58% of the 4.51 billion total parameters. The minimal difference in trainable parameters between methods indicates that the performance disparities observed cannot be attributed to differences in model capacity.

Table 3 Comprehensive experimental results comparing QLoRA and AdaLoRA across backbone models. Lower values indicate better performance for all metrics except trainable parameters

Backbone	Method	Train Loss	Val. Loss	Perp.	Param (M)
Llama-3-8B	QLoRA	0.0817	0.0814	1.085	167.77
Llama-3-8B	AdaLoRA	0.1234	0.1177	1.125	167.79
Qwen2-7B	QLoRA	0.0822	0.0821	1.086	161.48
Qwen2-7B	AdaLoRA	0.1986	0.1563	1.169	161.49

DISCUSSION

Interpretation of Results

Our experimental findings reveal a consistent performance advantage for QLoRA over AdaLoRA under single-epoch training conditions on medical textbook question answering. Several factors may explain this pattern.

First, the single-epoch training budget may be insufficient for AdaLoRA's adaptive rank allocation to reach its optimal configuration. AdaLoRA's pruning schedule progressively reduces ranks from the initial value of 64 to the target value of 8 between the warmup period at 10% of training and the finalization point at 70% of training. Extended training over multiple epochs would allow the model more time to stabilize after each rank reduction, potentially enabling AdaLoRA to realize its theoretical advantages.

Second, the medical question answering task may not exhibit the layer-wise importance heterogeneity that AdaLoRA is designed to exploit. If the importance of different weight matrices is relatively uniform across the model for this particular task, adaptive rank allocation provides no advantage over fixed-rank approaches.

Implications for Medical NLP

Our results have practical implications for deploying parameter-efficient fine-tuning in medical natural language processing applications. The consistent performance advantage of QLoRA suggests that this method represents a reliable choice for healthcare institutions seeking to adapt large language models to medical domains under computational constraints. The 4-bit quantization employed by QLoRA substantially reduces memory requirements, enabling fine-tuning on consumer-grade hardware without sacrificing performance on medical question answering tasks.

Limitations

This study has several limitations. First, our evaluation relied exclusively on perplexity and loss metrics. Second, the single-epoch training schedule may not represent optimal training conditions for either method. Third, we employed default hyperparameters without extensive optimization. Fourth, our experiments were limited to two backbone models. Fifth, the MedicalTextbook_QA dataset represents a specific question answering format that may not transfer directly to other medical NLP tasks.

CONCLUSION

This study presented a systematic comparison of QLoRA and AdaLoRA for parameter-efficient fine-tuning of large language models on medical textbook question answering. Our experiments across two backbone architectures, Llama-3-8B-Instruct and Qwen2-7B-Instruct, demonstrate that QLoRA consistently outperforms AdaLoRA under single-epoch training conditions. QLoRA achieved validation perplexity values of 1.085 and 1.086 for the two backbones respectively, compared to AdaLoRA's 1.125 and 1.169, representing relative validation loss reductions of 30.8% and 47.5%.

Both methods maintained comparable trainable parameter counts at approximately 3.5% of total model parameters, indicating that the performance differences stem from how each method utilizes its parameter budget rather than from differences in model capacity. Our analysis suggests that QLoRA's fixed-rank approach provides more stable convergence behavior within limited training budgets, while AdaLoRA's adaptive rank allocation mechanism may require extended training schedules to realize its theoretical advantages. These findings offer practical guidance for medical NLP practitioners: QLoRA represents a reliable and effective choice for adapting large language models to medical domains when computational resources are constrained.

Acknowledgments

This work was supported by Igdir University.

Ethical standard

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

Availability of data and material

The MedicalTextbook_QA dataset is available through the Hugging Face Hub at https://huggingface.co/datasets/winder-hybrids/MedicalTextbook_QA.

Conflicts of interest

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

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that generative artificial intelligence (AI) tools were used during the preparation of this manuscript. Specifically, AI assistance was utilized for language editing, text refinement, and formatting purposes. The authors take full responsibility for the content and have carefully reviewed and verified all AI-assisted outputs.

LITERATURE CITED

Bayat Toksoz, S. and G. Isik, 2025 Parameter-efficient fine-tuning of llama models for financial sentiment classification. Cluster Computing **29**: 41.

Brown, T. B., B. Mann, N. Ryder, *et al.*, 2020 Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901.

Chowdhery, A., S. Narang, J. Devlin, *et al.*, 2023 Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**: 1–113.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer, 2024 Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115.

Han, T., L. C. Adams, J.-M. Papaioannou, *et al.*, 2023 Medalpaca: An open-source collection of medical conversational ai models and training data. arXiv preprint .

Han, Z., C. Gao, J. Liu, *et al.*, 2024 Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint .

Houlsby, N., A. Giurgiu, S. Jastrzebski, *et al.*, 2019 Parameter-efficient transfer learning for nlp. In *Proceedings of ICML*, volume 97, pp. 2790–2799.

Hu, E. J., Y. Shen, P. Wallis, *et al.*, 2022 Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Huang, K., J. Altosaar, and R. Ranganath, 2019 Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint .

Karabacak, M. and K. Margetis, 2023 Embracing large language models for medical applications. Cureus **15**: e39305.

Lee, J., W. Yoon, S. Kim, *et al.*, 2020 Biobert: A pre-trained biomedical language representation model. Bioinformatics **36**: 1234–1240.

Lester, B., R. Al-Rfou, and N. Constant, 2021 The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pp. 3045–3059.

Li, X. L. and P. Liang, 2021 Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the ACL*, pp. 4582–4597.

Li, Y., Z. Li, K. Zhang, *et al.*, 2023 Chatdoctor: A medical chat model fine-tuned on llama. Cureus **15**: e40895.

Lialin, V., V. Deshpande, and A. Rumshisky, 2023 Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint .

Singhal, K., S. Azizi, T. Tu, *et al.*, 2023 Large language models encode clinical knowledge. Nature **620**: 172–180.

Thirunavukarasu, A. J., D. S. J. Ting, *et al.*, 2023 Large language models in medicine. Nature Medicine **29**: 1930–1940.

Zhang, Q., M. Chen, A. Bukharin, *et al.*, 2023 Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*.

How to cite this article: Toksoz, S. B. and Isik, G. A Comparative Evaluation of QLoRA and AdaLoRA for Parameter-Efficient Fine-Tuning of Large Language Models on Medical Textbook Question Answering. *Artificial Intelligence in Applied Sciences*, 2(4),27-31, 2026.

ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.32-36
https://doi.org/10.69882/adba.ai.2026015

Bridging the Gap Between Theoretical Performance and Clinical Utility in Multi-Class Skin Lesion Diagnosis

Furkan Sönmez *,1 **and Fevzi Das** α,2

*Department of Computer Engineering, Igdir University, Igdir, Turkiye, αDepartment of Architecture and Town Planning, Igdir University, 76000 Igdir, Turkiye.

ABSTRACT The escalating global incidence of skin cancer necessitates the development of robust, objective, and automated diagnostic systems capable of augmenting clinical decision-making. This study presents a rigorous comparative analysis of four landmark Convolutional Neural Network (CNN) architectures, ResNet-101, MobileNet-v3-Large, EfficientNet-B5, and Inception-v4, evaluated against the expansive and heterogeneous ISIC 2019 dataset. Comprising 25,331 high-resolution images across eight diagnostic categories, the dataset presents significant morphological challenges due to inherent visual ambiguity and class imbalance. Our findings reveal that EfficientNet-B5 achieves the highest predictive robustness with a peak accuracy of 0.8968 and an F1-score of 0.8458, leveraging its sophisticated compound scaling approach to capture subtle malignant markers. Concurrently, MobileNet-v3-Large demonstrated exceptional efficiency, yielding a nearly identical accuracy of 0.8965 with a minimal computational load of 0.4307 GFLOPs, making it a prime candidate for edge-computing applications. Despite its higher theoretical complexity, ResNet-101 provided the fastest real-world inference latency at 0.5032 ms, indicating superior hardware optimization. While these results underscore the transformative potential of deep learning in dermatology, misclassification patterns between melanoma and melanocytic nevi highlight persistent challenges in navigating fine-grained morphological boundaries. Ultimately, this research provides a holistic framework for selecting optimal architectural backbones based on specific clinical deployment constraints, bridging the gap between theoretical model performance and practical utility.

KEYWORDS
Skin lesion classification
Deep learning in dermatology
ISIC 2019 dataset
Computational efficiency
Model benchmarking

INTRODUCTION

Skin cancer represents a formidable global health crisis, with its prevalence reaching unprecedented levels over the last few decades (Gloster and Neal 2006; Armstrong and Kricker 1995; Leiter *et al.* 2020). Among the diverse spectrum of cutaneous malignancies, malignant melanoma is particularly notorious for its aggressive metastatic potential; however, it remains highly treatable when intercepted in its incipient stages. Despite the availability of advanced dermatoscopic techniques, the clinical diagnosis of skin lesions is fraught with challenges. The morphological overlap between benign and malignant pathologies, coupled with the subtle variations in pigment patterns and border irregularities, introduces a significant degree of intra-observer subjectivity. Consequently, there is an imperative clinical need for robust, objective, and automated diagnostic systems capable of augmenting a clinician's decision-making process (Madan *et al.* 2010).

Manuscript received: 3 November 2023,
Revised: 28 December 2023,
Accepted: 20 January 2024.

[1]furkansonmez2024@gmail.com
[2] fevzi.das@igdir.edu.tr (**Corresponding author**).

The paradigm shift in medical image analysis has been primarily driven by the maturation of Deep Learning (DL), specifically through the evolution of Convolutional Neural Networks (CNNs) (Aruk *et al.* 2026; Cakmak and Pacal 2025a; Pacal and Cakmak 2025). These computational frameworks have demonstrated an extraordinary capacity to autonomously distill high-level hierarchical features from complex dermatological datasets, often identifying diagnostic biomarkers that elude manual visual inspection (Attallah 2024; Cakmak and Pacal 2025b; Cakmak 2025; Cakmak and Maman 2025). The transition from traditional hand-crafted feature engineering to end-to-end deep learning architecture has allowed for more nuanced classification across multi-class pathologies. Central to this progress is the availability of large-scale, annotated repositories such as the ISIC 2019 dataset, which provides a more diverse and challenging benchmark than its predecessors by incorporating a broader range of lesion categories and imaging conditions.

However, selecting an optimal architectural backbone for clinical deployment is not merely a question of peak accuracy; it requires a multifaceted evaluation of the trade-offs between predictive power and computational overhead. In modern medical environments, where real-time inference and integration into mobile or edge-computing platforms are increasingly vital, architec-

tural efficiency is as quintessential as diagnostic sensitivity. While heavyweight models offer high-capacity feature extraction, lighter architecture provides the agility required for point-of-care applications. Systematic benchmarking across varying architectural paradigms is therefore essential to bridge the gap between theoretical model performance and practical clinical utility.

In this research, we conduct a rigorous comparative analysis of four landmark CNN architectures: ResNet-101, MobileNet-v3-Large, EfficientNet-B5, and Inception-v4. By leveraging the ISIC 2019 dataset, our study investigates how distinct design philosophies, ranging from the residual learning of ResNet to the sophisticated compound scaling of EfficientNet, address the inherent complexities of multi-class skin lesion classification. Beyond traditional accuracy metrics, we scrutinize these models through the lens of computational complexity (GFLOPs) and parameter efficiency. This multifaceted evaluation ensures that the selected models are not only statistically robust but also optimized for the practical constraints of clinical environments, offering a balanced framework for high-stakes medical decision-making.

RELATED WORKS

The pursuit of high-precision diagnostic tools has led researchers to explore diverse architectural optimizations and learning strategies. In this context, Musthafa *et al.* (2024) demonstrated the efficacy of ensemble learning and model checkpoints, utilizing architectures like InceptionV3 and ResNet50 to achieve more stable and robust classification across various lesion types. Their work highlights the importance of strategic model saving and optimization to capture the most representative features during the training process. Building on the theme of architectural exploration, Rafeeque and Abini (2024) conducted a performance comparison between landmark models such as VGG16 and DenseNet, emphasizing how the depth and connectivity of these networks directly influence their ability to discern subtle malignant patterns in multi-class environments.

As models become more complex, the need for transparency in clinical decision-making has become paramount. Addressing this, Shah *et al.* (2024) introduced an explainable AI (XAI) framework that integrates Convolutional Neural Networks (CNNs) with Particle Swarm Optimization (PSO). By leveraging XAI, they provided a means to visualize the diagnostic markers driving the model's predictions, thereby bridging the gap between "black-box" algorithms and clinical interpretability. Similarly, Kumar *et al.* (2024) proposed SCCNet, a dedicated architecture for multi-class classification that provides estimated disease probabilities. Their approach moves beyond simple labels, offering a probabilistic output that aligns more closely with the nuanced nature of dermatological assessment.

The challenge of data variability and feature refinement has also been a focal point of recent research. . Surya *et al.* (2025) investigated the impact of rigorous image preprocessing and data augmentation on models like AlexNet and VGG, demonstrating that enhancing the quality of the input data is as critical as the choice of the network itself for melanoma detection. In a more specialized approach, Ali *et al.* (2024) presented a novel Fully Convolutional Encoder-Decoder Network (FCEDN) combined with SpaSA-based hyper-parameter optimization. Their work illustrates how adaptive CNN classification can be refined through automated optimization to handle the morphological diversity inherent in skin lesions. Finally, Sabir and Mehmood (2024) explored the boundaries of transfer learning, investigating how pre-trained networks can be effectively fine-tuned for dermatological tasks. Their findings underscore that while deeper architectures offer significant representational power, the strategic application of knowledge from broader domains is essential for achieving high generalization in specialized medical datasets. Collectively, these studies establish a foundation for balancing predictive accuracy, computational feasibility, and clinical transparency.

MATERIALS AND METHODS

Dataset and Data Preprocessing

The foundational pillar of this investigation is the ISIC 2019 dataset (r19 2019), which serves as an expansive and heterogeneous repository of 25,331 high-resolution dermatoscopic images. This corpus is uniquely characterized by its profound morphological diversity, spanning eight distinct diagnostic categories: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevus (NV), Squamous Cell Carcinoma (SCC), and Vascular lesions (VASC). As visualized in Figure 1, these pathologies exhibit significant visual ambiguity and intra-class variations, presenting a formidable challenge for automated diagnostic systems to distinguish between benign and malignant lesions with high precision.

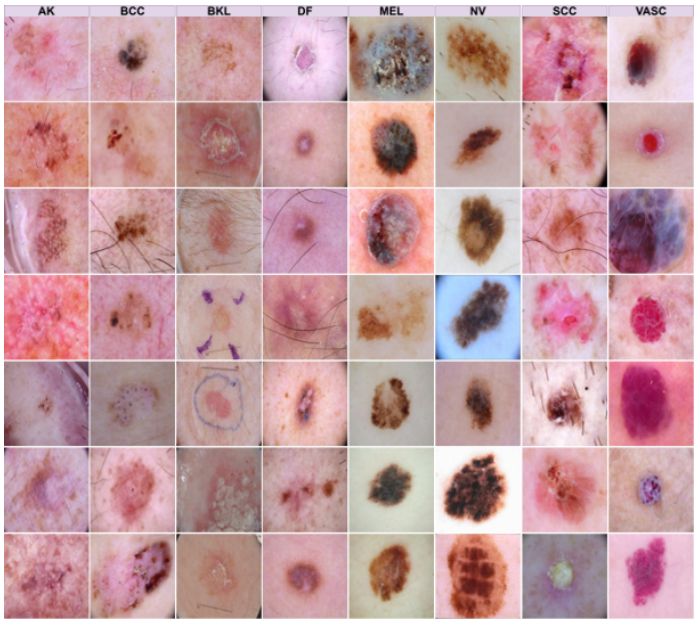


Figure 1 Representative dermatoscopic images from the ISIC 2019 dataset illustrating the morphological diversity and visual ambiguity across eight diagnostic categories

To ensure the statistical integrity of our comparative analysis and to promote robust model generalization, we implemented a rigorous stratified partitioning of the available data. The dataset was systematically divided into training (70%), validation (15%), and testing (15%) subsets, ensuring that the underlying distribution of lesion types remained consistent across all experimental phases. The precise numerical distribution of these samples across the eight diagnostic classes, culminating in a grand total of 25,331 images, is comprehensively documented in Table 1.

Table 1 Statistical distribution of the ISIC 2019 dataset across training, validation, and testing subsets for each of the eight skin lesion classes

Class Name	Total	Train	Val	Test
BKL	2624	1836	393	395
DF	239	167	35	37
VASC	253	177	37	39
AK	867	606	130	131
MEL	4522	3165	678	679
BCC	3323	2326	498	499
NV	12875	9012	1931	1932
SCC	628	439	94	95
GRAND TOTAL	25331	17728	3796	3807

Architectural Design Philosophies

To evaluate the trade-offs between predictive capacity and computational demand, our methodology benchmarks four landmark Convolutional Neural Network (CNN) architectures, each representing a distinct evolution in deep learning research: ResNet-101 (He *et al.* 2015), MobileNet-v3-Large (Howard *et al.* 2019), EfficientNet-B5 (Tan and Le 2019), and Inception-v4 (Szegedy *et al.* 2017). ResNet-101 leverages deep residual learning to facilitate the optimization of high-capacity networks by addressing the vanishing gradient problem. In contrast, MobileNet-v3-Large is specifically engineered for resource-constrained environments, utilizing depthwise separable convolutions to achieve high-speed inference with minimal parameter overhead. EfficientNet-B5 introduces a sophisticated compound scaling approach that simultaneously optimizes network depth, width, and resolution to maximize diagnostic sensitivity. Finally, Inception-v4 utilizes multi-scale convolutional modules to capture complex spatial hierarchies. Together, these models provide a comprehensive spectrum of design philosophies, ranging from the parameter-heavy residual blocks of ResNet to the agile, mobile-centric architecture of MobileNet.

Data Augmentation Strategy

A critical challenge inherent in dermatological imaging is the pronounced class imbalance, particularly the dominance of Melanocytic Nevus (NV) samples. To mitigate this and enhance the generalization capabilities of the models, we implemented an extensive data augmentation strategy. This pipeline included geometric transformations, such as random rotations, spatial scaling, and horizontal/vertical flipping, designed to simulate the varied orientations and perspectives encountered in clinical dermoscopy. These techniques ensure that the networks learn to distill invariant diagnostic features rather than memorizing dataset-specific noise, thereby bridging the gap between theoretical performance and practical clinical utility (Wang *et al.* 2024).

Performance Evaluation Metrics

To rigorously assess the diagnostic efficacy of the benchmarked architectures, we utilize a multifaceted suite of performance metrics that account for both statistical robustness and clinical applicability. The primary metric, Accuracy, provides an overall measure of the model's ability to correctly classify lesions across all eight diagnostic categories as defined in (1). However, given the high-stakes nature of dermatological screening, we extend our evaluation to include Precision and Recall to specifically measure the fidelity of malignancy detection and the model's sensitivity to true positive cases, as formulated in (2) and (3) respectively. To reconcile the potential trade-offs between these two metrics, especially in the context of the inherent class imbalance found within the ISIC 2019 dataset, the F1-Score is employed as a harmonic mean (4), ensuring a balanced representation of diagnostic performance across minority and majority classes. Beyond these conventional accuracy-based metrics, our analysis incorporates critical computational parameters such as model complexity (Params), floating-point operations (GFLOPs), and real-world inference latency (ms) to determine the practical feasibility of deploying these models within resource-constrained or real-time medical environments.

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

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

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

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

RESULTS

Quantitative Performance and Comparative Analysis

The empirical evaluation of the four benchmarked CNN architectures reveals a compelling spectrum of performance, highlighting the trade-offs between depth-wise feature extraction and computational agility. As detailed in Table 2.

EfficientNet-B5 emerged as the superior model in terms of predictive robustness, achieving a peak Accuracy of 0.8968 and a leading F1-Score of 0.8458. These results underscore the efficacy of the compound scaling method in distilling high-level diagnostic markers from the heterogeneous ISIC 2019 dataset. Interestingly, MobileNet-v3-Large demonstrated nearly identical accuracy (0.8965) and a highly competitive F1-Score (0.8383), despite maintaining a significantly smaller parameter footprint of only 4.21M compared to the 28.36M of EfficientNet-B5.

The comparative analysis further illustrates the impact of architectural design on computational overhead. While ResNet-101 and Inception-v4 provided respectable performance with accuracies of 0.8521 and 0.8697 respectively, they exhibited considerably higher computational costs, with ResNet-101 reaching 15.7288 GFLOPs. From a clinical deployment perspective, MobileNet-v3-Large presented a highly favorable profile, achieving the lowest GFLOP count (0.4307) while maintaining high diagnostic sensitivity. However, it is noteworthy that ResNet-101 offered the fastest real-world inference time at 0.5032 ms, suggesting that its residual structure remains highly optimized for parallel processing on modern GPU hardware.

Analysis of Model Predictions and Error Patterns

To gain deeper insight into the classification behavior of our most accurate model, we scrutinized the EfficientNet-B5 predictions through the confusion matrix presented in Figure 2. The model exhibited exceptional proficiency in identifying NV, correctly classifying 1,866 out of 1,932 test samples, which is vital given the

Table 2 Comparative performance analysis of ResNet-101, MobileNet-v3-Large, EfficientNet-B5, and Inception-v4 based on diagnostic metrics and computational complexity

Models	Accuracy	Precision	Recall	F1 Score	Params (M)	GFLOPs	Inference Time (ms)
ResNet-101	0.8521	0.8042	0.7490	0.7729	42.52	15.7288	0.5032
MobileNet-v3-Large	0.8965	0.8774	0.8093	0.8383	4.21	0.4307	1.9880
EfficientNet-B5	0.8968	0.8805	0.8166	0.8458	28.36	4.6551	3.4748
Inception-v4	0.8697	0.8333	0.7799	0.8046	41.16	12.2450	0.5516

prevalence of this category in clinical practice. Similarly, BCC and MEL showed high true positive rates, with 465 and 545 correct identifications respectively.

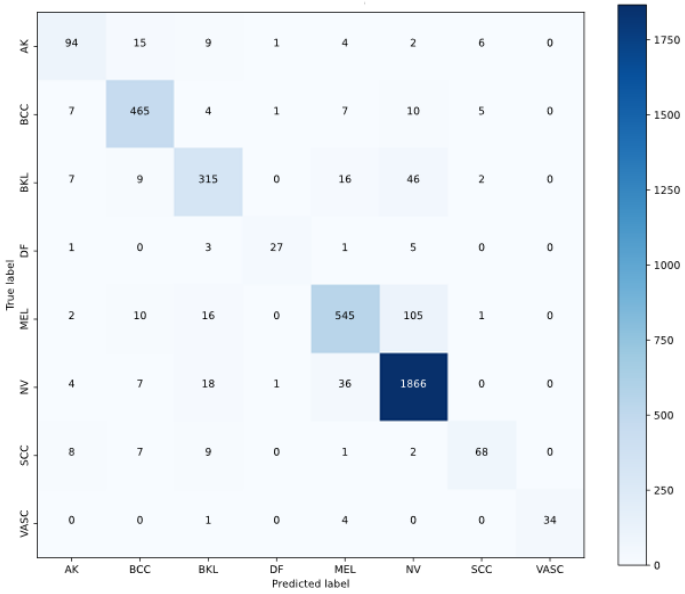


Figure 2 Confusion matrix of the EfficientNet-B5 model on the ISIC 2019 test set, highlighting classification performance and common morphological overlaps

Despite these strengths, the matrix reveals critical areas of morphological overlap that contribute to misclassification. A notable percentage of MEL cases were incorrectly labeled as NV, a common diagnostic pitfall in dermatology due to the subtle pigmentary transitions between benign and malignant melanocytic lesions. Furthermore, AK and SCC exhibited mutual confusion, likely stemming from their shared epithelial origin and similar scaling textures observed in dermatoscopic imagery.

A qualitative assessment of the classification outcomes, as illustrated in Figure 3, provides a visual context for both the successes and the limitations of the automated pipeline. The "True Predictions" panels demonstrate the model's ability to successfully capture the hallmark features of various pathologies, such as the distinct vascular structures in VASC and the characteristic pigment networks in MEL. Conversely, the "Misclassified Predictions" highlight the profound visual ambiguity inherent in skin lesion analysis. In several instances, lesions with atypical presentations, such as Melanoma mimicking the symmetry of a Benign Keratosis, led to false negative results. These visual findings emphasize that while CNNs possess remarkable feature-extraction capabilities, the clinical reliability of such systems is highly dependent on the model's ability to navigate the fine-grained morphological boundaries that often blur the distinction between benignity and malignancy.

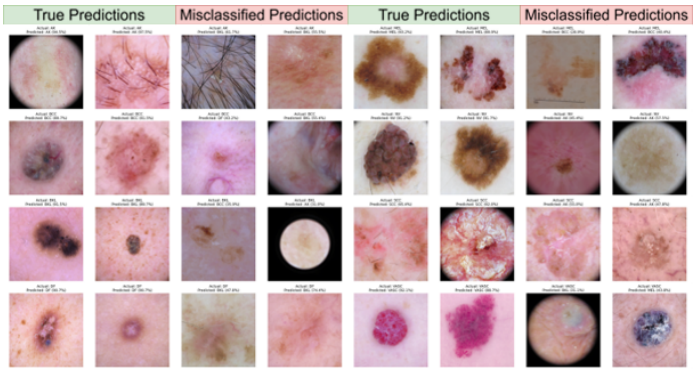


Figure 3 Qualitative assessment of classification results showcasing successful true predictions and representative misclassified samples with associated confidence levels

DISCUSSION

Interpretation of Key Findings

Our comparative analysis underscores a pivotal trade-off between predictive robustness and computational efficiency. EfficientNet-B5 emerged as the most accurate architecture, achieving a peak Accuracy of 0.8968 and a leading F1-Score of 0.8458. This superiority is largely attributed to its compound scaling philosophy, which systematically optimizes network depth, width, and resolution to capture the subtle morphological hallmarks of malignancy. However, the performance of MobileNet-v3-Large is arguably more noteworthy from a practical standpoint; it achieved a nearly identical Accuracy of 0.8965 while utilizing only 4.21M parameters and the lowest recorded computational load of 0.4307 GFLOPs.

An intriguing observation from our results is the discrepancy between theoretical complexity (GFLOPs) and real-world Inference Time. While ResNet-101 exhibited significantly higher GFLOPs than the mobile-centric models, it yielded the fastest inference latency at 0.5032 ms. This suggests that the classic residual structure remains exceptionally well-optimized for the parallel processing capabilities of modern GPU hardware. These findings imply that the "best" model is highly dependent on the target deployment environment, whether it be a high-throughput clinical workstation or a resource-constrained edge device.

Clinical Implications and Diagnostic Challenges

The error patterns identified through our confusion matrix analysis in Figure 2 reflect the same diagnostic pitfalls encountered by board-certified dermatologists. The frequent misclassification of MEL as NV highlights the critical challenge posed by subtle pigmentary transitions and "look-alike" lesions. Similarly, the mutual confusion between AK and SCC suggests that the shared epithelial origins of these pathologies lead to overlapping visual textures that remain difficult for even sophisticated CNNs to resolve.

Qualitative assessments further reveal that while models successfully identify hallmark features like vascular structures, they remain susceptible to atypical presentations. As seen in Figure 3, instances of malignancy mimicking benign symmetry can lead to dangerous false negatives. These results emphasize that while automated tools provide powerful decision support, their clinical reliability is fundamentally limited by the fine-grained morphological boundaries of skin cancer.

Limitations and Future Directions

Despite the high performance achieved, this study faces limitations inherent in the ISIC 2019 dataset, primarily the significant class imbalance where NV accounts for over half of the samples. While our data augmentation strategies mitigated this effect, future work should explore more advanced synthetic data generation or cost-sensitive learning to further improve minority class detection. Additionally, moving beyond purely visual data to incorporate patient metadata, such as age, anatomical site, and clinical history, could provide the contextual cues necessary to resolve the visual ambiguities identified in this research.

CONCLUSION

This research has established a comprehensive benchmarking framework for multi-class skin lesion classification by evaluating the intricate trade-offs between architectural design philosophies and diagnostic efficacy. Our investigation underscores that while EfficientNet-B5 provides the most robust predictive power, the agile architecture of MobileNet-v3-Large offers a remarkably viable alternative for resource-constrained point-of-care applications, maintaining high sensitivity with a significantly reduced parameter footprint. Furthermore, the observation that ResNet-101 delivers the fastest inference latency despite higher GFLOPs highlights that real-world clinical utility is as much a product of hardware-specific optimization as it is of algorithmic complexity. Although the deep learning models evaluated here effectively navigate profound visual ambiguities, the persistent confusion between malignant melanoma and benign nevi remains a significant diagnostic hurdle that mirrors clinician-level difficulties. To transcend these limitations, future research must move beyond purely visual data to integrate patient metadata and explore advanced learning strategies to address the inherent class imbalances found in large-scale dermatological repositories. In summary, this study provides the quantitative evidence and qualitative insights necessary to guide the selection of deep learning backbones for the next generation of automated dermatological screening tools.

Acknowledgments

This study was supported by the Turkish Health Institutes Presidency (TÜSEB) within the scope of the 2025-A1-01 Call, Project No. 44675. The authors gratefully acknowledge TÜSEB for its financial support and scientific contribution to this research. The experimental computations were carried out using the computing infrastructure of the Artificial Intelligence and Big Data Application and Research Center, Igdir University, which provided essential resources for model training and evaluation.

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/salviohexia/isic-2019-skin-lesion-images-for-classification?resource=download.

Conflicts of interest

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

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used artificial intelligence tools in order to improve the readability and language quality of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

LITERATURE CITED

2019 Isic 2019 skin lesion images for classification. Kaggle Dataset, Accessed December 23, 2025.

Ali, R., A. Manikandan, R. Lei, and J. Xu, 2024 A novel spasa based hyper-parameter optimized fcedn with adaptive cnn classification for skin cancer detection. Scientific Reports **14**.

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**.

Attallah, O., 2024 Skin-cad: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level cnns features and transfer learning. Computers in Biology and Medicine **178**.

Cakmak, Y., 2025 Machine learning approaches for enhanced diagnosis of hematological disorders. Computational Systems and Artificial Intelligence **1**: 8–14.

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, 2025a Comparative analysis of transformer architectures for brain tumor classification. Exploratory Medicine **6**.

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

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

He, K., X. Zhang, S. Ren, and J. Sun, 2015 Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Howard, A., M. Sandler, G. Chu, *et al.*, 2019 Searching for mobilenetv3. arXiv preprint .

Kumar, V. A., C. Chandana, G. Supraja, *et al.*, 2024 Sccnet: Skin cancer detection and multi-class classification using deep cnn model with estimated disease probabilities. SN Computer Science **5**.

Leiter, U., U. Keim, and C. Garbe, 2020 Epidemiology of skin cancer: Update 2019. Advances in Experimental Medicine and Biology **1268**: 123–139.

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

Musthafa, M. M., M. T R, V. K. V, and S. Guluwadi, 2024 Enhanced skin cancer diagnosis using optimized cnn architecture and checkpoints for automated dermatological lesion classification. BMC Medical Imaging **24**.

Pacal, I. and Y. Cakmak, 2025 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.

Rafeeque, S. and M. A. Abini, 2024 Performance comparison of skin cancer detection and classification using cnn. In *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*.

Sabir, R. and T. Mehmood, 2024 Classification of melanoma skin cancer based on image data set using different neural networks. Scientific Reports **14**.

Shah, S. A. H., S. T. H. Shah, R. Khaled, *et al.*, 2024 Explainable ai-based skin cancer detection using cnn, particle swarm optimization and machine learning. Journal of Imaging **10**: 332.

Surya, V., B. H. Gollavilli, U. H. Nagarajan, *et al.*, 2025 Cloud-based cnn for automated skin cancer detection and classification in healthcare. International Journal of Science and Engineering Applications **14**: 40–45.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. A. Alemi, 2017 Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence .

Tan, M. and Q. V. Le, 2019 Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint .

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

How to cite this article: Sönmez, F. B. and Das, F. Bridging the Gap Between Theoretical Performance and Clinical Utility in Multi-Class Skin Lesion Diagnosis. *Artificial Intelligence in Applied Sciences*, 2(5),32-36, 2026.

ADBA
Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.37-48
https://doi.org/10.69882/adba.ai.2026016

Predicting Upper Respiratory Tract Infections: The Role of Weather Data and Explainable AI

Türker Berk Dönmez [*,1], **Mustafa Kutlu** [α,2] **and Chris Freeman** [β,3]

*Sakarya University of Applied Sciences, Faculty of Technology, Department of Biomedical Engineering, Sakarya 54050, Turkey, [α]Department of Mechatronics Engineering, Sakarya University of Applied Sciences, Sakarya, Türkiye, [β]University of Southampton, School of Electronics and Computer Science (ECS), Southampton SO17 1BJ, United Kingdom.

ABSTRACT Upper Respiratory Tract Infections (URTIs) are a global health issue, affecting myriad individuals and encompassing infections of the nose, sinuses, pharynx, or larynx. The diverse symptoms and varying severity of URTIs, coupled with their potential to be influenced by meteorological factors, underscore the importance of understanding the interplay between weather conditions and URTI incidence. This research, conducted in the Pamukova District from December 15, 2020, to December 31, 2022, delves into this relationship by integrating weather data from Meteoblue and patient data from the Pamukova Family Medicine Center. The comprehensive data cleaning, harmonization, and preprocessing considered the 3 or 5 preceding days in alignment with the URTI incubation period. Utilizing the Catboost machine learning model on two separate datasets, the study revealed enhanced performance with a 5-day data frame. The model yielded 67 true positives, 24 true negatives, 8 false positives, and 22 false negatives, resulting in an F1-score of 0.6154, an accuracy of 75.21%, precision and recall values of 0.75 and 0.5217, respectively, and an AUC value of approximately 0.7768. These results emphasize the critical role of an extended temporal frame in understanding the connection between environmental factors and URTI incidence, offering substantial insights for the development of targeted public health interventions in the Pamukova District.

KEYWORDS
Explainable AI
Upper respiratory tract infections
Weather data

INTRODUCTION

When the upper respiratory tract is thought of, the anatomical formations located in the head and neck and above the thorax, especially the nose, paranasal sinuses, nasopharynx, oropharynx and larynx, are thought of (Tu *et al.* 2013). Viral infections of the upper respiratory tract, also known as the common cold (acute chorea, cold, common cold), are the most common respiratory diseases in humans. They are occurred on average 2-4 times a year in adults and every year in children. URTIs, which can be seen 6-10 times a year on average, are usually self-limiting and rarely lead to complications (Derbyshire and Calder 2021).

URTIs may be brought on by a variety of diseases, including bacterial, viral, and fungal infections (Kolawole and Idris 2020). The rhinovirus and coronavirus are two of the viruses that are most commonly discovered in connection with URTIs. They may be spread by direct contact with respiratory droplets from ill persons or by touching contaminated surfaces and then coming in contact with the face. The frequency of URTIs might also be impacted by the weather and other environmental variables (Stolz *et al.* 2019). URTIs are placing a heavy strain on both the healthcare system and the general public. These infections can raise the chance of acquiring later-life illnesses like wheeze, asthma, as well as recurring doctor visits, parental stress, and other health issues (Chen *et al.* 2021).

In Turkey, URTI is the most common condition among children and adults, according to the Turkey Health Survey 2022. The results of the survey show that in the previous six months, URTI accounted for 31.3% of illnesses seen in children aged 0–6 years and 27.1% of illnesses seen in children aged 7–14 years. Furthermore, with a prevalence of 9.6% in the previous 12 months, URTI ranks fourth among the illnesses experienced by people over the age of 15 years. According to this findings, URTI is a significant public health issue in Turkey, and preventive and therapeutic efforts are required to lessen the burden it places on the populace .

Many academics have been interested in the connection between weather and URTIs which are common conditions that can be brought on by a variety of bacterial or viral agents and affect the nose, sinuses, throat, larynx, or trachea (Zeru *et al.* 2020). The productivity, quality of life, and cost of healthcare of those who are affected by URTIs can all be significantly impacted. In order to create effective preventative and treatment plans, it is crucial to understand the factors that affect the frequency and severity of URTIs. As they can modify the host's immune response, the pathogen's survival and transmission, and the exposure to other respiratory irritants and allergens, weather conditions are one of the environmental factors that may alter the risk and outcome of URTIs (D'Amato *et al.* 2018). The epidemiology and pathophysiol-

Manuscript received: 5 November 2025,
Revised: 26 December 2025,
Accepted: 18 January 2026.

[1]turkerberkdonmez@yahoo.com
[2]mkutlu@subu.edu.tr
[3] cf@ecs.soton.ac.uk (**Corresponding author**).

ogy of URTIs can be affected by a variety of meteorological factors, including temperature, humidity, precipitation, wind speed, and air quality (Zou *et al.* 2021).

The Eccles and Wilkinson (2015) research found a link between exposure to cold air and a greater frequency of URTI. Cold air may influence the nasal mucosa, the body's first line of defense against viral viruses, which might explain this. Cold air may slow down and make less efficient mucociliary clearance, the process of removing mucus and trapped particles from the respiratory system. Cold air may also impair the capacity of immune cells in the nasal cavity, such as macrophages, to ingest and destroy viruses. The link between cold and URTI has also been shown to be stronger in northern than in southern areas in UK research, indicating that the influence of cold air on URTI may vary by geographic location. However, rather than happening immediately after the temperature shift, the effect of cold air on URTI occurs two to three weeks later. The cold air also has less of an impact on URTI than it does on lower respiratory infections like pneumonia or bronchitis. In conclusion, exposure to cold air may impair the mucosal immune system, raising the danger of upper respiratory tract viral infections .

In order to learn more about the connections between various meteorological factors and URTIs and lower respiratory tract infections (LRTIs), Falagas et al. retrospectively analyzed meteorological and clinical data from the Attica region of Greece. The incidence of URTIs was observed to positively correlate with cold weather conditions, peaking when the weekly average temperature fell below 10 degrees Celsius, according to the researchers. In addition, they discovered that LRTIs were more commonly associated with chilly temperatures than URTIs were. The association between cold weather and URTIs, according to researchers, is caused by a number of factors, including direct effects of cold on the viability and infectivity of the viruses that cause URTIs, indirect effects of cold on immune and respiratory system function, and direct effects of cold on people's behavior (Falagas *et al.* 2008).

A comprehensive analysis of the impact of meteorological and air pollution factors on respiratory diseases in Linyi, China, was conducted. According to the study, a 0.31 increase in the concentration of NO2 is associated with a rise in pneumonia cases. Similarly, increased levels of PM2.5 and PM_{10}—specifically, by 0.23 and 0.24, respectively—were linked to higher pneumonia incidence. Low temperature and humidity levels, particularly a decrease in daily average temperature and humidity, were associated with a reduction in chronic lower respiratory diseases and pneumonia cases. Conversely, these same factors increased the incidence of acute upper respiratory infections by 0.04 and 0.05. High wind speeds also correlated positively with respiratory diseases. The SVR model used in the study showed a significant prediction potential, with an R^2 value of 0.308 for pneumonia, highlighting the intricate relationship between environmental factors and respiratory health (Yang *et al.* 2023).

Lim *et al.* (2023) conducted a comprehensive study on forecasting URTIs using high-dimensional time series data and forecast combinations. Their research indicated that a 1-week lag in lower temperature is associated with a significant increase in URTI attendances. Similarly, past relative humidity and absolute humidity levels showed notable effects on URTI forecasts. For example, a 1% increase in relative humidity decreased URTI attendances by approximately 3-4%, while an increase in absolute humidity at longer forecast windows (4-8 weeks ahead) was associated with a decrease in URTI attendances. The study also highlighted the superior predictive performance of forecast combinations, with mean absolute percentage errors ranging from 10% to 25% across different horizons. These findings emphasize the intricate relationship between climatic factors and URTI incidence, providing valuable insights for public health resource planning and outbreak preparedness .

Jhuo *et al.* (2019) conducted a comprehensive study on predicting trends in influenza and associated pneumonia in Taiwan using machine learning. Their research utilized meteorological parameters, such as temperature and relative humidity, and air pollution parameters, including PM 2.5 and CO, alongside the number of acute upper respiratory infection (AURI) outpatients as inputs. They used data from December 2009 to December 2017 and made predictions for January 2010 to January 2018. Patients were categorized into low, moderate, and high volume levels. The multilayer perceptron (MLP) model developed in their study achieved an accuracy of 81.16% for the elderly population and 77.54% for the overall population. The study found that larger data sets from bigger areas improved accuracy, whereas lower accuracy was observed for children aged 0-4 years due to fewer samples and less exposure to environmental factors. These findings underscore the intricate relationship between environmental factors and the incidence of influenza and pneumonia .

A thorough investigation of the effects of climatic factors on URTIs was undertaken by Makinen et al. According to their research, a $1^{\circ}C$ drop in temperature is associated with a 4-5% rise in URTI incidence. Similar to this, low humidity levels—more precisely, those below 40%—were linked to an increase in instances. High wind speeds were similarly linked to an increased risk of URTIs despite having received less research. Additionally, it was proposed that a minor increase in infections may be associated with situations of declining barometric pressure. These observations highlight the complex interaction between weather and the frequency of URTIs (Mäkinen *et al.* 2009).

Kern *et al.* (2016) explored the relationship between weather data and the incidence of ophthalmological conditions using model-agnostic methods. Through Spearman's correlation analysis, they examined clinical data from the University Eye Hospital Munich from January 2014 to July 2015. They linked patient visits to weather variables like sunshine duration, temperature, and wind speed, finding a weekly increase of one sunshine hour correlated with an additional patient visit per week ($\rho = 0.44$, $P < 0.01$). Temperature increase of 1°C correlated with 2.6 more patients per week ($\rho = 0.29$, $P < 0.01$). Specifically, higher temperatures and longer durations of sunshine were positively correlated with increased visits for conditions like conjunctivitis and foreign body injuries. The model-agnostic approach allowed them to uncover significant correlations without being constrained by underlying data structure assumptions .

Santhanam *et al.* (2024) extended the application of model-agnostic methods by incorporating machine learning models to predict daily acute ischemic stroke (AIS) admissions based on weather data. Employing techniques such as Support Vector Machines (SVR), Random Forests (RF), and Extreme Gradient Boosting (XGB), they effectively managed the complex, nonlinear relationships between environmental factors and health outcomes. Their study identified maximum air pressure as a critical predictive variable, with extreme temperature conditions and stormy conditions also playing significant roles. The XGB model's robust predictive capability was evidenced by a low mean absolute error (MAE) of 1.21 cases/day on the test set, supporting better healthcare resource allocation and preparedness .

Mansour *et al.* (2023) employed the Lorenz equation and numer-

ical techniques like the Runge-Kutta method to develop a novel chaotic system for forecasting respiratory disease outbreaks, using a model-agnostic approach to integrate weather variables such as maximum temperature, air pressure, and humidity with patient data from the Pamukova Region. By utilizing a NARX network for input-output data processing, they established a high correlation coefficient of 90.16% between predicted and actual patient numbers. Their findings suggest a robust framework for employing chaotic systems in real-time health warning systems, potentially enhancing preemptive responses to environmental health risks. This model-agnostic methodology underlines the adaptability of chaotic systems in predicting complex health-related events .

In this study, the necessity is underscored by the global health impact of URTIs and the limited understanding of how weather conditions influence their incidence. By focusing on the Pamukova District, this research provides localized insights using model-agnostic SHAP (SHapley Additive exPlanations) values to reveal how specific weather conditions affect URTI patients. Integrating weather data with patient records and utilizing advanced machine learning models, the study highlights the importance of considering an extended temporal frame for accurate predictions. These findings are crucial for developing effective public health interventions and offer significant insights that can be applied to similar contexts globally, demonstrating how studies in smaller provinces can contribute to the bigger picture of improving public health outcomes.There is still a need for a more thorough understanding that takes into account regional variations, seasonal patterns, and potential interactions with other health determinants. The existing literature has given valuable insights into how specific meteorological factors affect the prevalence of URTIs (Mansour *et al.* 2023). In order to fully understand the epidemiology of URTIs, which continue to place a heavy strain on healthcare systems and communities, a comprehensive approach is required. This study contributes to the creation of more effective public health treatments and policies meant to lessen the effects of URTIs by investigating these aspects holistically.

MATERIALS AND METHODS

Data Collection and Preprocessing

Data on URTIs from the Pamukova District were combined with corresponding meteorological data. The datasets were then cleaned, harmonized, and assessed for outliers before being analyzed.

Weather Data Meteoblue was used to gather weather information, which included metrics for maximum, minimum, and mean temperatures, sunlight duration, shortwave radiation levels, precipitation, snowfall, humidity levels, cloud cover, air pressure, and wind speeds. The dataset is exclusive to the Pamukova area and runs from December 15, 2020, through December 31, 2022.

Patient Data Patients' information was gathered for this study with Sakarya University's ethical permission (E-71522473-050.01.04-15185-157). The research, which covers the period from January 1, 2021, to December 31, 2022, is concentrated on the Pamukova District in Sakarya Province. The information was given by the Pamukova Family Medicine Center, which is closed on weekends and major holidays. The clinic treated 52,792 patients in total during the course of 484 workdays, 4,454 of whom had upper respiratory tract infections (ICD codes J09–J18). As a result, during the course of the trial, URTIs accounted for around 9.2% of all patient visits.

Preprocessing Data and Model Training The weather data underwent various preprocessing. The averages, standard deviations, and value gaps of the data from the previous 3 or 5 days, including the current day, were calculated. This approach was chosen because the average incubation period for any URTI agent varies between approximately 1-5 days. The average number of patients was calculated as approximately 9.2 patients per day. The number of patients distributed over the days was classified binarily as above average and below average and was determined as the main target.

The study utilized a range of meteorological attributes to predict Upper Respiratory Tract Infections (URTIs). These attributes include the mean temperature (*meantemp*), mean humidity (*meanhumidity*), mean pressure (*meanpressure*), mean wind speed (*meanwind*), and mean sunshine duration (*sunshine*) measured over the last 3 or 5 days. Additionally, shortwave radiation (*radiation*), total precipitation (*precipitation*), and snowfall amount (*snowfall*) were considered. Cloud cover (*cloudcover*) was also included, along with calculated variables such as the standard deviation of min-max values (*minmaxSDtemp*, *minmaxSDhumidity*, *minmaxSDpressure*, *minmaxSDwind*), the standard deviation of average values (*meanSDtemp*, *meanSDhumidity*, *meanSDpressure*, *meanSDwind*), and the value range of the highest and lowest values (*VRtemp*, *VRhumidity*, *VRpressure*, *VRwind*) over the specified period. The *minmaxSD* values were calculated by measuring the standard deviation of the minimum and maximum values on a daily basis within the specified period.

Twenty-one different variables were created for these two separate datasets and are shown in detail in Table 1. These two datasets were evaluated with various machine learning models as shown in Table 2, and among them, the Catboost model showed the highest success.

75% of the data for each of the two sets is used for training and 25% is utilized for testing, resulting in a 75/25 split of the data. Additionally, the "Discussions" section includes findings that shed light on the model's stability.

CatBoost has been chosen as the main model for more investigation. CatBoost stands out for its practical efficiency and simplicity of use, qualities that are especially well-aligned with the study goals, even if other models show equivalent performance. CatBoost is thought to be the best solution for the challenges associated with illness forecasting using meteorological data because it combines great computing speed with powerful predictive skills.

Categorical Boosting (CatBoost) CatBoost, which stands for *Category Boosting*, is a gradient boosting library developed by Yandex. It has gained popularity in the machine learning community for its performance and its built-in support for categorical features, thus eliminating the need for extensive preprocessing like one-hot encoding or label encoding.

The mathematical foundation of CatBoost is rooted in the gradient boosting framework. The primary objective of gradient boosting is to optimize a cumulative objective function, which is a sum of a loss function and a regularization term (Prokhorenkova *et al.* 2018). CatBoost introduces several enhancements to the traditional gradient boosting technique:

$$\mathcal{L}(\mathbf{y}, \mathbf{F}) = \sum_{i=1}^{N} l(y_i, F(x_i)) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

Where $\mathbf{y}$ is the vector of true labels, $\mathbf{F}$ is the ensemble model, l is a differentiable convex loss function, f_k are the individual trees, and Ω is a regularization term.

Table 1 Attribution of the dataset for 3 or 5-day metrics

No.	Abbreviation	Name of the attribute	Units
1	meantemp	Last 3 or 5-day mean temperature	℃
2	meanhumidity	Last 3 or 5-day mean humidity	%
3	meanpressure	Last 3 or 5-day mean pressure	hPa
4	meanwind	Last 3 or 5-day mean wind speed	km/h
5	sunshine	Last 3 or 5-day mean sunshine duration	min
6	radiation	Last 3 or 5-day mean shortwave radiation	W/m²
7	precipitation	Last 3 or 5-day mean total precipitation	mm
8	snowfall	Last 3 or 5-day mean snowfall amount	cm
9	cloudcover	Last 3 or 5-day mean total cloud cover	%
10	minmaxSDtemp	Standard deviation of min-max temperature values in the last 3 or 5 days	Calculated
11	meanSDtemp	Standard deviation of the average temperature over the last 3 or 5 days	Calculated
12	VRtemp	Range of the highest and lowest temperature values in the last 3 or 5 days	Calculated
13	minmaxSDhumidity	Standard deviation of min-max humidity values in the last 3 or 5 days	Calculated
14	meanSDhumidity	Standard deviation of the average humidity over the last 3 or 5 days	Calculated
15	VRhumidity	Value range of the highest and lowest humidity values in the last 3 or 5 days	Calculated
16	minmaxSDpressure	Standard deviation of min-max pressure values in the last 3 or 5 days	Calculated
17	meanSDpressure	Standard deviation of the average pressure over the last 3 or 5 days	Calculated
18	VRpressure	Value range of the highest and lowest pressure values in the last 3 or 5 days	Calculated
19	minmaxSDwind	Standard deviation of min-max wind speed values in the last 3 or 5 days	Calculated
20	meanSDwind	Standard deviation of the average wind speed over the last 3 or 5 days	Calculated
21	VRwind	Value range of the highest and lowest wind speed values in the last 3 or 5 days	Calculated

A standout feature of CatBoost is its treatment of categorical features. The algorithm leverages a technique called ordered boosting, which involves random permutations to prevent overfitting. Another notable method is mean encoding, where categories are replaced with the average target value for that category, with certain regularization techniques applied to avoid overfitting.

For model interpretation, CatBoost offers built-in support for SHAP values, making it easier to explain the predictions and understand feature importances. This integration of SHAP values is particularly beneficial as it offers a consistent methodology for model interpretation without needing external tools (Chelgani *et al.* 2023).

In practice, CatBoost has proven to be competitive with other gradient boosting implementations, often outperforming them, especially when dealing with datasets with a high number of categorical features. Its efficiency, coupled with its user (Bentéjac *et al.* 2021). Table 3 in this research displays the hyperparameters for Catboost that were chosen.

Table 2 Model Metrics and Confusion Matrices

Model	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
CatBoost	0.7521	0.7500	0.5217	0.6154	67 8 22 24
XGBoost	0.7273	0.6667	0.5652	0.6118	62 13 20 26
Extra Trees	0.7273	0.7097	0.4783	0.5714	66 9 24 22
Random Forest	0.7107	0.6571	0.5000	0.5679	63 12 23 23
LightGBM	0.6942	0.6047	0.5652	0.5843	58 17 20 26
Explainable Boosting Machine	0.6860	0.6111	0.4783	0.5366	61 14 24 22
Logistic Regression	0.6777	0.6061	0.4348	0.5063	62 13 26 20
Adaboost	0.6446	0.5385	0.4565	0.4941	57 18 25 21
Decision Tree	0.6281	0.5116	0.4783	0.4944	54 21 24 22
KNN	0.6281	0.5111	0.5000	0.5055	53 22 23 23
Support Vector Machine	0.6198	0.5000	0.3478	0.4103	59 16 30 16
Naive Bayes	0.6116	0.4915	0.6304	0.5524	45 30 17 29

Table 3 Hyperparameter values for the CatBoost model

Hyperparameter	Value
iterations	1000
depth	6
l2_leaf_reg	3.0
model_size_reg	0.5
border_count	254

Model Interpretation with SHAP Shapley Additive Explanations (SHAP) provides a robust methodology for understanding and interpreting the output of any machine learning model. Drawing its foundation from cooperative game theory, SHAP was proposed in 2017 with an ambition to unify the various methods of model interpretation. By allocating an "importance value" to each feature, SHAP gives an indication of how much each feature contributes to a given prediction. This methodology serves as a consistent and locally accurate lens through which we can understand model behavior (Kavzoglu and Teke 2022).

The mathematical foundation of SHAP is rooted in the Shapley value from cooperative game theory. To calculate the SHAP value for a particular feature, denoted as f_i, we consider a set of all features, F, and all the potential feature subsets, S, that can be created after removing the i-th feature. The equation is:

$$f_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \times (|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right) \quad (2)$$

In this equation, $f_{S \cup \{i\}}$ and f_S represent the predictions of models trained with feature sets $S \cup \{i\}$ and S respectively. The terms

$x_{S\cup\{i\}}$ and x_S denote the values of the input features in the sets $S \cup \{i\}$ and S.

While SHAP provides a comprehensive methodology, the direct computation of Shapley values can be intensive, especially when dealing with a large number of features. To address this, SHAP offers approximations such as Shapley sampling and Shapley quantitative influence.

SHAP values can be interpreted from both global and local perspectives. On a global scale, features with consistently high absolute SHAP values across many samples generally have a greater influence on model predictions. Conversely, at the local level, for a given prediction, SHAP values provide information on the variables (Kannangara *et al.* 2022).

The versatility of SHAP is one of its standout attributes. It can be seamlessly applied to a multitude of models, ranging from decision trees and ensemble methods to neural networks. However, it's worth noting that the computational demand of SHAP can sometimes be a bottleneck, especially when the model has a vast number of features or when dealing with large datasets.

RESULTS

Results of Classification

In our quest to measure the potential of environmental variables in predicting the occurrence of Upper Respiratory Tract Infections (URTIs), we applied the Catboost machine learning model on two individually built datasets, incorporating both meteorological and patient data. For the dataset structured around metrics from the prior 3 days, the model defined 61 true positives and 21 true negatives, while misidentifying 14 and 25 examples as false positives and negatives, respectively. This resulted in an F1-score of around 0.519, an accuracy rate of 67.77%, and precision and recall values of 0.6 and 0.4565, respectively. Moreover, the model's AUC value, a vital statistic evaluating its discriminative power, was at around 0.6861. (Figure 1a and 1c)

Contrastingly, when the model was trained on data covering the prior 5 days, the results were considerably improved. The confusion matrix indicated 67 true positives, 24 true negatives, 8 false positives, and 22 false negatives. The measures indicated improvement across the board: an F1-score of 0.6154, an accuracy of 75.21%, and precision and recall values of 0.75 and 0.5217, respectively. The AUC value likewise experienced a boost, reaching roughly 0.7768. These findings underline the Catboost model's heightened competence with a longer 5-day data frame as compared to a 3-day one. It emphasizes the benefit of adopting an extended temporal frame while examining the link between environmental elements and URTI occurrences in the Pamukova District. Gleaning from this, specific public health interventions may be designed, aiming to limit the effect of URTIs on the population. (Figure 1b and 1d)

Explaining Model with SHAP

Understanding the choices of complicated models, such as Catboost, is vital for exposing the role of numerous environmental factors in forecasting Upper Respiratory Tract Infections. SHapley Additive exPlanations values give a complete measure of feature importance, providing a better comprehension of the model's decision-making process.

For the model developed using meteorological data from the prior three days, shortwave radiation emerged as the most significant feature with a SHAP value of 0.50, underlining its substantial influence on predicting Upper Respiratory Tract Infections. Other significant determinants included the variability in temperature

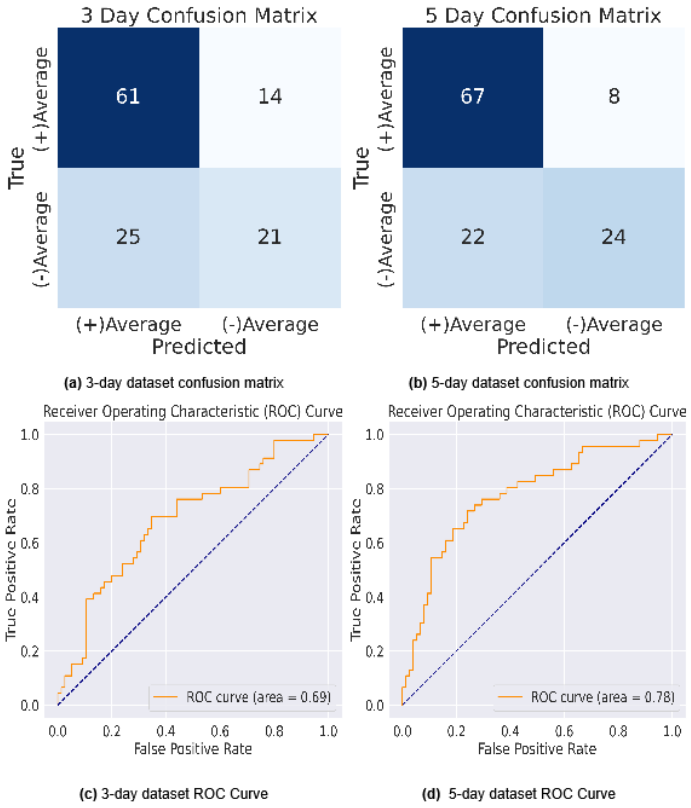
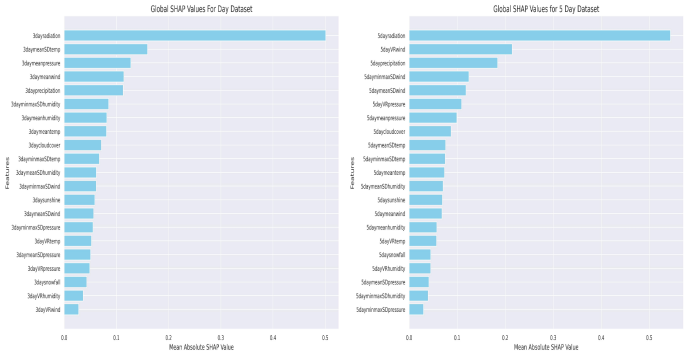


Figure 1 Confusion Matrices and ROC Curves for 2 dataset

over these three days with a SHAP value of 0.16, the average atmospheric pressure (SHAP value: 0.13), and the average wind speed (SHAP value: 0.11), further emphasizing the role of these environmental conditions in the model's predictions. Features including the length of sunlight (SHAP value: 0.06), the fluctuation in wind speed (SHAP value: 0.06), and snowfall during the same time (SHAP value: 0.04) also affected the model's outputs, albeit to a lesser amount as shown in Figure 2a.

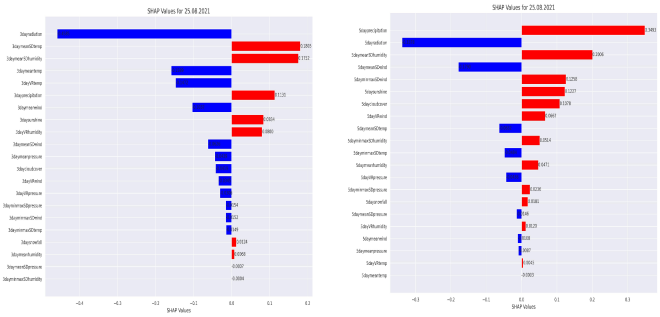
When assessing the model trained with data covering the preceding five days, shortwave radiation consistently appeared as the dominating feature with a SHAP score of 0.54. This reinforces the persistent relevance of shortwave radiation levels in predicting Upper Respiratory Tract Infection prevalence over an extended duration. Other parameters, such as the variation in wind speed (SHAP value: 0.21), precipitation levels (SHAP value: 0.18), and the difference between the lowest and highest wind speeds over these five days (SHAP value: 0.12), also played a considerable part in the model's classifications. However, features like snowfall (SHAP value: 0.05), the variance in humidity (SHAP value: 0.04), and the difference between the minimum and maximum atmospheric pressures over this period (SHAP value: 0.03) exhibited a relatively reduced influence in the five-day dataset compared to the three-day one as shown in Figure 2b.

Drawing from these findings, it's obvious that although certain climatic factors, like shortwave radiation, continuously increase the incidence of Upper Respiratory Tract Infections, others change in their relevance dependent on the observational period. This sophisticated knowledge reveals the delicate association between environmental conditions and Upper Respiratory Tract Infection incidences, underlining the necessity for region-specific, data-driven interventions, especially in locations like the Pamukova District.



(a) Global SHAP Values for 3-Day Dataset **(b)** Global SHAP Values for 5-Day Dataset

Figure 2 Global SHAP Values for 3-Day and 5-Day Datasets



(a) Local SHAP values for 25.08.2021 from 3-day dataset **(b)** Local SHAP values for 25.08.2021 from 5-day dataset

Figure 3 Local SHAP values analysis for 25.08.2021

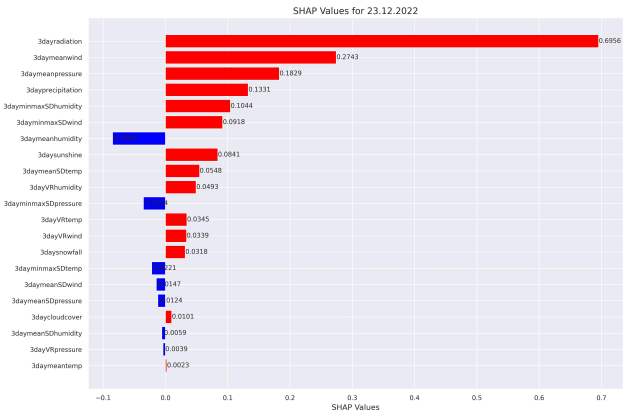
The predictions of machine learning models may be understood by SHAP values, since both global and local explanations can be offered by them. How each characteristic impacts the model's output on average is explained by global explanations, while how each feature affects the model's output for a given instance is indicated by local explanations. The 2 days on which common and accurate forecasts were provided by both models are investigated in this case: 25.08.2021 and 23.12.2022. These days are noteworthy since diverse seasons and weather conditions are represented by them, and it is desirable to understand how these differences were recorded by the models. The table provides meteorological readings and SHAP values of the data for these two days. The relative value of each attribute for the two models and the two days can be compared by evaluating the table, and which features contributed favorably or adversely to the predictions may be discovered. It can also be seen how the meteorological data impact the SHAP values, and how the nonlinear and interactive effects of the characteristics on the model's output are represented by them.

For 25.08.2021, placed in the warmer season, the model generated a True Negative forecast, suggesting that URTI diagnoses were below average—a prediction that was successfully determined by the model. Shortwave radiation on this day was significantly high, and its accompanying negative SHAP values of -0.46 (3-day model) and -0.34 (5-day model) imply that excessive radiation levels lowered the chance of URTIs. This makes shortwave radiation one of the most significant factors for this day's prognosis. Additionally, the lack of precipitation, as represented in its zero actual value, bore positive SHAP values of 0.11 (3-day model) and 0.35 (5-day model), designating it as another important contributor.

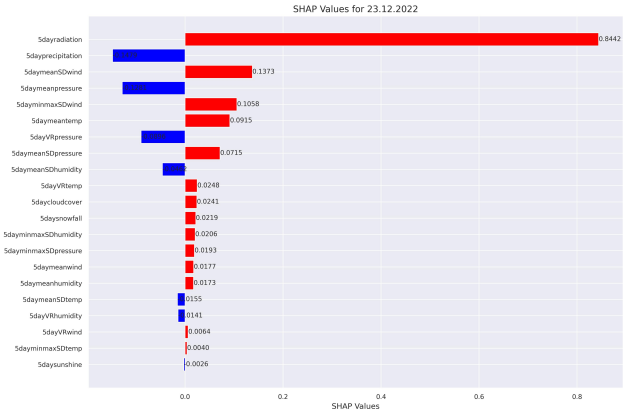
Wind speed and its variations also appear relevant. The average wind speed, with its SHAP value of 0.11 for the 3-day model, shows that wind could have had a role in raising URTI prevalence. Similarly, the variation in humidity across the three days, albeit tiny in real value, nonetheless exhibited a positive SHAP value of 0.05 (3-day model), showing its subtle effect on the model's predictions.

On 23.12.2022, a day indicative of the colder season, the model returned a True Positive prediction, accurately projecting an above-average URTI diagnosis. The reduced shortwave radiation value for this day, followed with its noticeable positive SHAP value of 0.84 (5-day model), clearly stands out as the most impactful variable, highlighting its crucial significance in the model's predictions. Atmospheric pressure, however constant across both models in actual values, displayed differing SHAP values, indicating its subtle effect.

Furthermore, the variation in wind speed for the 5-day model, with its SHAP value of 0.21, also emerges as an important driver, illustrating the model's sensitivity to varying wind conditions throughout this winter day.



a) Local SHAP values for 23.12.2022 from 3-day dataset



b) Local SHAP values for 23.12.2022 from 5-day dataset

Figure 4 Local SHAP values analysis for 23.12.2022

Piecing together these findings, it's evident that the model pre-

cisely evaluates many climatic parameters when forecasting URTIs. Shortwave radiation, precipitation, and wind-related factors constantly appear as the most relevant drivers, altering the model's predictions throughout various seasons. This research underlines the complex, multiple character of environmental determinants on health outcomes and underscores the need of examining a spectrum of environmental variables, particularly when creating tailored treatments for varied weather conditions and seasons.

SHAP Dependences for Variables

Using SHAP dependency plots, it was showed how the model output is impacted by critical meteorological factors. This variation is not formed entirely by an individual component but is also impacted by interactions with other weather-related variables. In each figure, the SHAP value and the individual variable value are depicted on the axes, and the feature with the most noticeable interaction impact is indicated by the color of the dots. Through these representations, crucial meteorological components were discovered, and their complicated interaction was recognized.

Both the 3-day and 5-day datasets were employed to produce the SHAP dependency plots, revealing insights into short-term weather patterns and their possible cumulative consequences. By merging information from both periods, a thorough knowledge of the climatic conditions' immediate and prolonged consequences was produced. However, it's crucial to remember that although an instructive summary is offered by these plots, they have not been submitted to in-depth statistical analysis. Their major objective is to illustrate the link between certain weather conditions and the predictions provided by our model.

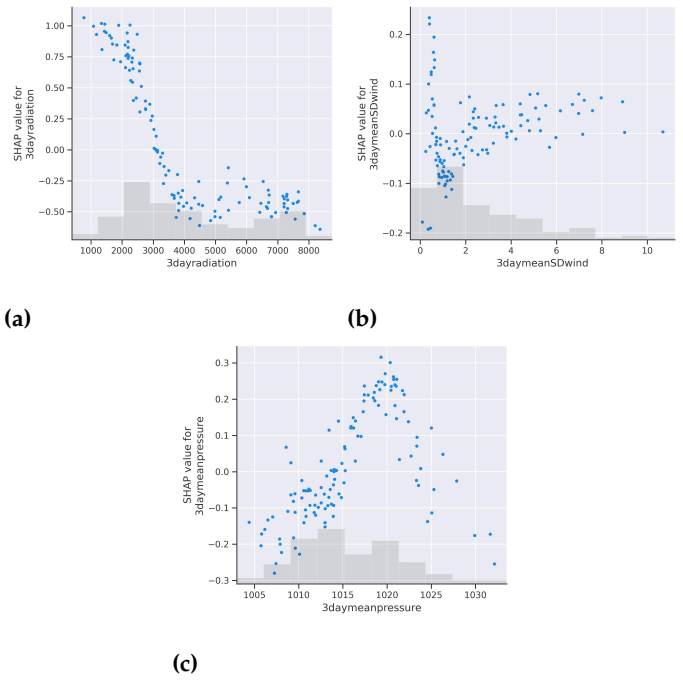
A constant trend was found across both periods in the Shortwave Radiation dependency plots shown in Figures 6a and 6e for the 3-day and 5-day datasets respectively. Looking at both graphs, it is seen that values below 3000 W/m2 trigger an increase in the number of patients, while values above trigger a decrease in the number of patients. This discovery was confirmed by the remarkable SHAP values reported for shortwave radiation in both data sets.

In Figure 6f, the impact of average wind speed fluctuations on forecasts is shown by the Standard Deviation of Average Wind Speed in the Last 3 Days. Although there is no obvious sign of an increase in the standard deviation, it seems that the low wind speed change within 3 days triggered the number of patients. Its important role is emphasized by the corresponding SHAP values.

The significant impact of wind dynamics on model predictions is highlighted by the Value Range of Highest and Lowest Wind Speed Values in the Last 5 Days, as shown in Figure 6g. It has been observed that all "wind speed range" within 5 days being below 20 km/h triggers an increase in the number of diseases, while being above 30 km/h triggers a decrease in the number of patients. This discovery was supported by the prominent SHAP values in both datasets.

The importance of atmospheric pressure in shaping model outputs is illustrated in the Last 3-Day Average Pressure dependency plot, which can be seen in Figure 6h. Looking at the graph, it . as seen that the 3-day average pressure value being below 1015 hPa pushed the number of patients to a decreasing trend, and being between 1015 and 1020 hPa triggered a relative increase in the number of patients. The relevance was further emphasized by the SHAP value in the 3-day data set.

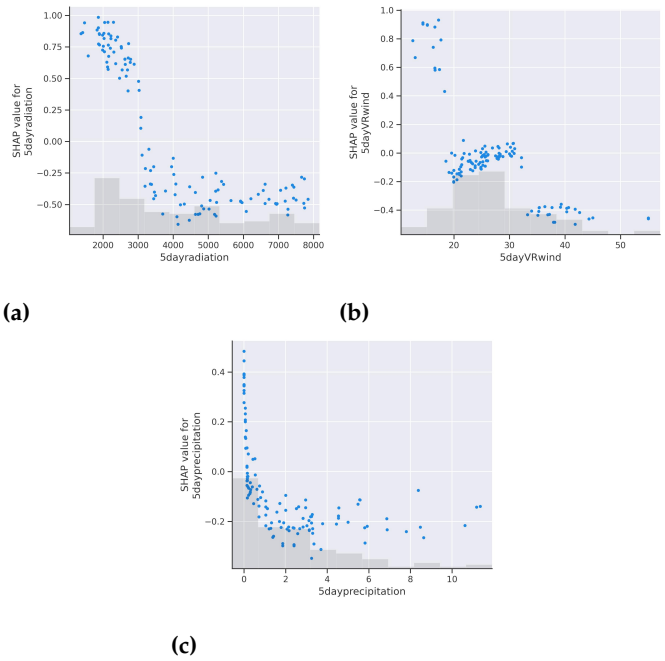
Finally, the impact of precipitation on our model's predictions over the 5-day period is clarified in the Precipitation dependency plot in Figure 6d. Looking at the graph, it is seen that the 5-day



(a) (b) (c)

Figure 5 Dependence plots for variables **a)** shortwave radiation (3-day), **b)** mean SD wind (3-day), **c)** mean pressure (3-day)

average rainfall amount being different from zero tends to reduce the number of patients relatively.



(a) (b) (c)

Figure 6 Dependence plots for variables **a)** shortwave radiation (5-day), **b)** VR wind (5-day), **c)** precipitation (5-day)

DISCUSSION

This research aims to bridge the gap in understanding the complicated association between climatic factors and the prevalence of

Table 4 Actual Values and SHAP Values table for 25.08.2021 and 23.12.2022

	25.08.2021 (True Negative)				23.12.2022 (True Positive)			
Feature	3-Day A.V.	3-Day SHAP	5-Day A.V.	5-Day SHAP	3-Day A.V.	3-Day SHAP	5-Day A.V.	5-Day SHAP
meantemp	23.22	-0.16	22.48	-0.00	6.88	0.00	6.19	0.09
meanhumidity	77.03	0.01	76.58	0.05	63.53	-0.08	73.78	0.02
meanpressure	1012.54	-0.04	1013.31	-0.01	1019.02	0.18	1024.24	-0.13
meanwind	9.70	-0.10	10.96	-0.01	6.84	0.27	11.35	0.02
sunshine	650.60	0.08	660.33	0.12	554.56	0.08	361.82	-0.00
radiation	6635.25	-0.46	6618.57	-0.34	2561.42	0.70	1987.19	0.84
precipitation	0.00	0.11	0.00	0.35	0.00	0.13	2.52	-0.15
snowfall	0.00	0.01	0.00	0.02	0.00	0.03	0.22	0.02
cloudcover	44.85	-0.04	41.04	0.11	10.56	0.01	44.31	0.02
minmaxSDtemp	5.25	-0.01	5.60	-0.05	5.99	-0.02	5.15	0.00
meanSDtemp	0.72	0.18	1.11	-0.06	1.95	0.05	1.86	-0.02
VRtemp	11.83	-0.15	14.99	0.00	14.97	0.03	14.97	0.02
minmaxSDhumidity	1.03	-0.00	27.89	0.05	7.18	0.10	22.23	0.02
meanSDhumidity	1.16	0.18	1.76	0.20	9.20	-0.01	14.82	-0.05
VRhumidity	2.61	0.08	66.00	0.01	18.39	0.05	58.00	-0.01
minmaxSDpressure	1.61	-0.02	2.21	0.02	4.32	-0.04	8.03	0.02
meanSDpressure	1.62	-0.00	1.57	-0.01	4.11	-0.01	7.16	0.07
VRpressure	3.89	-0.03	7.10	-0.04	10.20	-0.00	24.20	-0.09
minmaxSDwind	2.14	-0.02	13.49	0.13	1.11	0.09	9.02	0.11
meanSDwind	1.91	-0.06	2.14	-0.18	1.56	-0.01	5.97	0.14
VRwind	4.52	-0.03	30.65	0.07	3.02	0.03	25.77	0.01

URTIs in the Pamukova District, utilizing modern machine learning methods. Several critical discoveries arose from this analysis, having substantial implications for both the scientific community and public health measures.

The results of this study underscore the significant improvements achieved by extending the temporal frame of environmental data from 3 to 5 days when predicting URTIs using the Catboost model. The extended data frame led to a substantial increase in predictive performance, with the F1-score improving from 0.519 to 0.6154, accuracy from 67.77% to 75.21%, precision from 0.6 to 0.75, recall from 0.4565 to 0.5217, and the AUC value from 0.6861 to 0.7768. These results highlight the advantage of considering a broader temporal context, which captures more comprehensive environmental patterns that influence URTI occurrences.

Additionally, the integration of SHAP values significantly enhances the interpretability of the Catboost model, providing clear insights into feature importance. Key findings revealed that shortwave radiation was the most influential predictor, with its SHAP value increasing from 0.50 in the 3-day model to 0.54 in the 5-day model. Other important factors included variations in wind speed, precipitation levels, and atmospheric pressure. The improved interpretability and predictive accuracy underscore the potential of this method for developing effective, data-driven public health interventions. This approach not only improves the robustness of predictions but also enables targeted strategies tailored to specific environmental conditions, ultimately contributing to better health outcomes in regions like the Pamukova District.

The continuous relevance of shortwave radiation across both 3-day and 5-day datasets highlights its severe influence on URTI incidences. This coincides with some recent study, which has highlighted the possible immunomodulatory effects of sun radiation, possibly altering virus transmission and susceptibility. The negative link identified between high shortwave radiation levels and URTIs shows that greater sunshine exposure, and maybe its related vitamin D synthesis, could give some protection against URTIs. This underscores the necessity of addressing regional and seasonal changes when creating public health interventions.

The association between wind dynamics and URTIs is complicated. While wind may scatter respiratory droplets, possibly lowering transmission, it can also worsen respiratory symptoms and increase exposure to allergens. Our results highlighting the impact of wind speed changes, particularly over extended periods, may open the way for more nuanced study addressing the interactions between wind patterns, allergen dispersion, and URTI occurrences. Specifically, the 5-day model showed that variations in wind speed, with a SHAP value of 0.21, were significant predictors of URTIs. Additionally, the dependency plot analysis revealed that wind speed ranges below 20 km/h increased the number of cases, whereas ranges above 30 km/h decreased the number of cases, further emphasizing the importance of wind dynamics in predicting URTI occurrences.

Atmospheric pressure, another crucial element in our model, has been little examined in connection to respiratory diseases. Our results reveal prospective pathways for investigation into how atmospheric pressure could alter air quality, respiratory function, and therefore, susceptibility to infections. The observed association between pressure levels and URTIs could also indicate indirect consequences, such as behavioral changes in reaction to climatic circumstances. For instance, the 3-day average atmospheric pressure below 1015 hPa was associated with a decrease in URTI cases, whereas pressure between 1015 and 1020 hPa triggered a relative increase, as shown in the SHAP dependency plot. The SHAP values for mean atmospheric pressure were -0.04 for the 3-day model and 0.18 for the 5-day model, indicating its significant but complex role in influencing URTI rates.

Precipitation appeared as a key variable across the 5-day period. This is in accordance with prior research, which has generally correlated damp circumstances with higher virus survival and transmission, particularly in cold settings. Our findings indicate that the presence of precipitation in the 5-day dataset, with a SHAP value of 0.18, was a significant factor in predicting URTIs. The dependency plot demonstrated that non-zero precipitation values tended to reduce the number of URTI cases. This lays the ground for further extensive investigations that might shed light on how various precipitation types affect URTIs, emphasizing the importance of understanding the specific climatic conditions that promote or hinder the transmission of respiratory infections.

The strength of this work comes in its holistic approach, incorporating a variety of climatic factors and applying powerful machine learning methods to decode their cumulative influence on URTIs. CatBoost, with its capacity to handle categorical characteristics without preprocessing, emerged as a useful tool, offering insights with great accuracy. However, some limits must be noted. While the research caught a broad variety of environmental factors, additional possible confounders such as indoor air quality, personal habits, and vaccination rates were not evaluated. The reliance on data from a specific area further restricts the generalizability of the results. Furthermore, although SHAP values give a comprehensive comprehension of feature relevance, they do not always suggest causation.

Future research should seek to replicate and build upon these results in other geographical locations, integrating additional possible confounders and examining causal processes. Longitudinal research covering longer time periods might give insights into the long-term impact of climatic factors on URTIs. There's also a need for in-depth investigation of the molecular and physiological pathways via which these environmental influences impact respiratory health.

In conclusion, our analysis underlines the complicated interaction of climatic circumstances in producing URTI patterns. As the global community grapples with respiratory illnesses, information from such research are vital. They not only increase our knowledge but also give actionable information for public health authorities, allowing the creation of tailored treatments that account the particular environmental and climatic context of a place.

CONCLUSION AND FUTURE WORK

This research provides a complete examination of the interaction between several climatic conditions and the prevalence of URTIs in the Pamukova District of Sakarya Province, Turkey. Through the application of the CatBoost machine learning model, we discovered that certain environmental characteristics, notably shortwave radiation, precipitation, and wind-related variables, continuously emerge as key drivers in forecasting URTI occurrences.

Notably, shortwave radiation's consistent effect throughout varied seasons highlights its importance as a critical element. This underlines the delicate balance between the environment and human health, indicating that although certain elements have a consistent influence, others fluctuate dependent on the time range studied. Additionally, the model's precise detection of detailed weather patterns, particularly when employing a 5-day observing period, shows that protracted environmental circumstances could play a more essential role in determining URTIs than previously believed.

The use of SHAP values greatly improved our comprehension,

allowing for both global and local interpretations of the model's predictions. These numbers not only strengthened the findings reached from the model's raw outputs but also offered insight on the complicated connections between numerous weather factors.

The results of this study uncover various exciting paths for both additional research and practical applications. The usage of SHAP values in our study has shown to be a rewarding venture, enabling a granular comprehension of the multiple meteorological aspects that contribute to the occurrence of URTIs. While our work has found a plethora of discoveries, it's obvious that the full potential of SHAP values and other machine learning interpretability tools needs to be utilized, especially in the context of URTIs and weather conditions. Some of the planned futureworks are:

- Enhancing Patient-Centric Reporting Using SHAP Values: Presently, SHAP values enable us to discover which climatic parameters in our dataset play a crucial role in forecasting URTI occurrences. An extension of this method would be to describe the percentage contribution of each variable, supplying a more explicit and accurate grasp of the risk factors. This revised technique might provide healthcare practitioners with a personalized strategy to predict URTI spikes, basing their preventative actions around the most relevant weather circumstances.
- Integration of Additional Data Sources: Beyond meteorological considerations, incorporating statistics relating to air pollution, pollen counts, and other environmental variables might increase the forecasting powers of the model. This would offer a more thorough view of the environmental triggers of URTIs.
- Expansion of Geographical Scope: Delving into different geographical locations, both within Turkey and worldwide, could show whether the linkages detected in the Pamukova District are generally applicable or feature regional quirks.
- Temporal Analysis using SHAP: By applying SHAP values in a temporal context, it would be feasible to discover patterns linked to the seasonality of URTIs and how various meteorological conditions interact through time to impact URTI prevalence.
- Real-time URTI Predictive Systems: Capitalizing on the insights obtained, there's a chance to create real-time prediction systems that can estimate URTI prevalence based on present and impending climatic conditions, thereby helping healthcare institutions to plan appropriately.

Ethical standard

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

Availability of data and material

Participants' information was gathered from January 2020 to December 2022 at the Pamukova Family Health Centre, where T.B.D. is affiliated as MD, after receiving ethical approval from the Sakarya University of Applied Sciences Ethical Committee (E-26428519-044-77759). The Pamukova Family Health Centre provided the dataset for URTI. T.B.D., the study's corresponding author, will provide the data that back up its conclusions upon request. Please be aware, though, that the data cannot be made public because it contains details that would jeopardize the research participants' right to privacy even if it was anonymized.

Conflicts of interest

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

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used artificial intelligence tools in order to improve the readability and language quality of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

LITERATURE CITED

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz, 2021 A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review **54**: 1937–1967.

Chelgani, S. C., H. Nasiri, A. Tohry, and H. Heidari, 2023 Modeling industrial hydrocyclone operational variables by shap-catboost-a "conscious lab" approach. Powder Technology **420**: 118416.

Chen, X., L. Huang, Q. Li, M. Wu, L. Lin, *et al.*, 2021 Exposure to environmental tobacco smoke during pregnancy and infancy increased the risk of upper respiratory tract infections in infants: a birth cohort study in wuhan, china. Indoor air **31**: 673–681.

Derbyshire, E. J. and P. C. Calder, 2021 Respiratory tract infections and antibiotic resistance: a protective role for vitamin d? Frontiers in Nutrition **8**: 652469.

D'Amato, M., A. Molino, G. Calabrese, L. Cecchi, I. Annesi-Maesano, *et al.*, 2018 The impact of cold on the respiratory tract and its consequences to respiratory health. Clinical and translational allergy **8**: 1–8.

Eccles, R. and J. Wilkinson, 2015 Exposure to cold and acute upper respiratory tract infection. Rhinology **53**: 99–106.

Falagas, M. E., G. Theocharis, A. Spanos, L. A. Vlara, E. A. Issaris, *et al.*, 2008 Effect of meteorological variables on the incidence of respiratory tract infections. Respiratory medicine **102**: 733–737.

Jhuo, S.-L., M.-T. Hsieh, T.-C. Weng, M.-J. Chen, C.-M. Yang, *et al.*, 2019 Trend prediction of influenza and the associated pneumonia in taiwan using machine learning. In *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–2, IEEE.

Kannangara, K. P. M., W. Zhou, Z. Ding, and Z. Hong, 2022 Investigation of feature contribution to shield tunneling-induced settlement using shapley additive explanations method. Journal of Rock Mechanics and Geotechnical Engineering **14**: 1052–1063.

Kavzoglu, T. and A. Teke, 2022 Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (xgboost) and natural gradient boosting (ngboost). Arabian Journal for Science and Engineering **47**: 7367–7385.

Kern, C., K. Kortüm, M. Müller, F. Raabe, W. J. Mayer, *et al.*, 2016 Correlation between weather and incidence of selected ophthalmological diagnoses: a database analysis. Clinical ophthalmology pp. 1587–1592.

Kolawole, O. M. and O. O. Idris, 2020 Erythromycin resistance in bacterial isolates from patients with respiratory tract infections in ikere-ekiti, nigeria. Annals of Science and Technology **5**: 49–57.

Lim, J. T., K. B. Tan, J. Abisheganaden, and B. L. Dickens, 2023 Forecasting upper respiratory tract infection burden using high-dimensional time series data and forecast combinations. PLOS Computational Biology **19**: e1010892.

Mäkinen, T. M., R. Juvonen, J. Jokelainen, T. H. Harju, A. Peitso, *et al.*, 2009 Cold temperature and low humidity are associated with increased occurrence of respiratory tract infections. Respiratory medicine **103**: 456–462.

Mansour, M., T. B. Donmez, M. KUTLU, and C. Freeman, 2023 Respiratory diseases prediction from a novel chaotic system. Chaos Theory and Applications **5**: 20–26.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 2018 Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31**.

Santhanam, N., H. E. Kim, D. Ruegamer, A. Bender, S. Muthers, *et al.*, 2024 Machine learning-based forecasting of daily acute ischemic stroke admissions using weather data. medRxiv pp. 2024–07.

Stolz, D., E. Papakonstantinou, L. Grize, D. Schilter, W. Strobel, *et al.*, 2019 Time-course of upper respiratory tract viral infection and copd exacerbation. European Respiratory Journal **54**.

Tu, J., K. Inthavong, G. Ahmadi, J. Tu, K. Inthavong, *et al.*, 2013 The human respiratory system. Computational fluid and particle dynamics in the human respiratory system pp. 19–44.

Yang, J., X. Xu, X. Ma, Z. Wang, Q. You, *et al.*, 2023 Application of machine learning to predict hospital visits for respiratory diseases using meteorological and air pollution factors in linyi, china. Environmental Science and Pollution Research **30**: 88431–88443.

Zeru, T., H. Berihu, G. Buruh, and H. Gebrehiwot, 2020 Magnitude and factors associated with upper respiratory tract infection among under-five children in public health institutions of aksum town, tigray, northern ethiopia: an institutional based cross-sectional study. Pan African Medical Journal **36**.

Zou, Z., C. Cheng, and S. Shen, 2021 The complex nonlinear coupling causal patterns between pm2. 5 and meteorological factors in tibetan plateau: A case study in xining. IEEE Access **9**: 150373–150382.

How to cite this article: Dönmez, T. B.,Kutlu, M. and Freeman, C. Predicting Upper Respiratory Tract Infections: The Role of Weather Data and Explainable AI. *Artificial Intelligence in Applied Sciences*, 2(6),37-48, 2026.