Journal of Artificial Intelligence and Metaheuristic

♦ ISSN: 2833-5597 Vol: 11, No. 01, 2026 ♦ PP. 16–31

Article DOI: https://doi.org/10.54216/JAIM.110102

FULL LENGTH ARTICLE

An Intelligent Metaheuristic-Optimized Deep Learning Approach for Heart Disease Diagnosis and Patient Stratification

Khaled Sh. Gaber[1,*] **Amal H. Alharbi**[2]

[1]Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA

[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Emails: khsherif@jcsis.org · ahalharbi@pnu.edu.sa

Received: August 04, 2025 Revised: October 11, 2025 Accepted: December 07, 2025 ⋆ Corresponding author

ABSTRACT

The growing heterogeneity of cardiovascular disease presentations poses significant challenges for clinical decision support systems, particularly in identifying patient similarities and developing robust predictive models capable of supporting personalized treatment strategies, which motivates the need for advanced data-driven frameworks that can jointly exploit unsupervised learning, deep learning, and intelligent optimization. In this study, we propose a comprehensive hybrid framework that integrates unsupervised patient clustering with deep learning classification, enhanced through Fitness Greylag Goose Optimization (FGGO), where clustering is first employed to uncover latent patient subgroups and inform downstream learning, followed by the use of a Deep Learning Framework Distilled by Gradient Boosting Decision Trees (DeepGBM) as the core predictive model, and finally optimized via FGGO for automated hyperparameter tuning. The primary contribution of this work lies in the design of an FGGO-optimized DeepGBM framework that systematically improves learning stability, feature interaction modeling, and predictive robustness, while also providing a rigorous comparative evaluation against other state-of-the-art metaheuristic optimizers, including Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Dipper Throated Optimization (DDTO), and Multiverse Optimization (MVO). Experimental results demonstrate that, at the baseline stage without optimization, DeepGBM achieves an accuracy of 0.9032, sensitivity of 0.8824, specificity of 0.9195, and F-score of 0.8889, indicating strong but improvable performance on heart disease patient data. After metaheuristic optimization, the proposed FGGO + DeepGBM model exhibits a substantial performance enhancement, reaching an accuracy of 0.9795, sensitivity of 0.9747, specificity of 0.9831, positive predictive value of 0.9776, negative predictive value of 0.9809, and an F-score of 0.9761, consistently outperforming PSO + DeepGBM, GWO + DeepGBM, DDTO + DeepGBM, and MVO + DeepGBM across all evaluation metrics. These results highlight the robustness and convergence consistency of FGGO-based optimization and confirm its effectiveness in navigating complex hyperparameter search spaces. The implications of this work extend to clinical practice and intelligent healthcare systems, as the proposed framework offers a reliable and scalable solution for patient stratification and heart disease prediction, supporting more accurate, interpretable, and data-driven clinical decision-making while paving the way for future integration into personalized and precision medicine applications.

Keywords: Heart disease prediction ▪ Patient clustering ▪ Deep learning optimization ▪ Metaheuristic algorithms ▪ Clinical decision support systems

1. INTRODUCTION

Clinical decision support systems have become an integral component of contemporary healthcare infrastructures, particularly in complex and data-intensive domains such as cardiovascular medicine.

Heart disease continues to represent a major global health burden, characterized by high prevalence, diverse clinical manifestations, and substantial variability in patient outcomes [1, 2, 3]. This variability is driven by a combination of demographic factors, physiological measurements, lifestyle influences, and diagnostic indicators, which collectively create a highly heterogeneous patient population.

In such settings, traditional rule-based or population-average approaches are often insufficient to capture the nuanced differences among patients. Consequently, data-driven methodologies that can systematically analyze large volumes of clinical data and uncover meaningful patterns have gained increasing importance [4, 5].

A central objective of data-driven healthcare analytics is patient stratification, which refers to the process of grouping patients into subpopulations that share similar clinical characteristics [6, 7].

Effective stratification enables clinicians to better understand disease heterogeneity and supports more informed clinical reasoning. By identifying hidden patient subgroups, clinicians can move beyond one-size-fits-all treatment paradigms and adopt strategies that are more closely aligned with individual patient profiles. This is particularly relevant in heart disease management, where patients diagnosed under the same clinical category may exhibit markedly different risk trajectories and responses to treatment [8, 9, 10].

Patient similarity analysis constitutes a foundational concept within personalized medicine. Rather than evaluating patients in isolation, similarity-based approaches allow clinicians to contextualize an individual case by examining outcomes and clinical trajectories of patients with comparable characteristics. Such approaches facilitate evidence-informed decision-making, as treatment strategies can be guided by the observed responses of similar patient cohorts. In cardiovascular care, where treatment efficacy may vary substantially across patients, similarity-driven analysis offers a pathway toward more personalized and adaptive clinical interventions.

Within this broader analytical landscape, unsupervised learning has emerged as a powerful paradigm for medical data analysis. Unlike supervised learning approaches, which depend on labeled outcomes that may be scarce, noisy, or subjective, unsupervised methods are designed to discover latent structures directly from the data. Clustering algorithms, in particular, group patients based on intrinsic similarities among their clinical features, enabling the identification of underlying patterns without prior assumptions about class labels. In heart disease datasets, clustering can reveal clinically meaningful subgroups defined by combinations of age, blood pressure, cholesterol levels, electrocardiographic findings, and exercise-related indicators. These subgroupings can assist physicians in recognizing distinct patient profiles that may warrant different diagnostic or therapeutic considerations.

Parallel to the advancement of unsupervised techniques, deep learning has profoundly influenced the field of medical data analytics. Deep learning frameworks are capable of modeling highly nonlinear relationships and complex feature interactions, which are common in clinical data but difficult to capture using traditional statistical models [11, 12]. Their ability to learn hierarchical feature representations directly from data has led to widespread adoption in diagnostic and prognostic tasks.

Nevertheless, despite their expressive power, deep learning models do not inherently address the problem of patient heterogeneity or subgroup discovery. This limitation has motivated the development of hybrid learning strategies that integrate unsupervised clustering with supervised deep learning classifiers. By incorporating clustering-derived insights into supervised models, such hybrid frameworks aim to enhance predictive performance while maintaining clinical relevance and interpretability [13, 14].

The application of clustering and deep learning techniques to heart disease data is accompanied by several methodological and practical challenges. One of the most prominent challenges arises from the high dimensionality and heterogeneity of clinical datasets. Heart disease records typically encompass a wide range of variables, including demographic attributes, physiological measurements, laboratory results, and diagnostic indicators. Each category of features captures a distinct aspect of patient health, and their combined analysis increases both computational complexity and the risk of introducing noise into the modeling process.

Feature redundancy and correlation further complicate the analytical process. Many clinical variables are inherently correlated due to underlying physiological relationships. While such correlations may carry important medical meaning, they can adversely affect data-driven models if not properly managed.

In clustering algorithms, correlated features may distort distance measures and lead to unstable or less meaningful patient groupings. In supervised deep learning models, redundant inputs can increase model complexity without contributing additional information, potentially degrading generalization performance.

Another critical challenge is the sensitivity of deep learning models to hyperparameter configurations.

Model depth, learning rates, regularization terms, and optimization strategies all play a decisive role in determining training stability and predictive performance. Suboptimal hyperparameter choices can result in slow convergence, poor local minima, or excessive overfitting, particularly in medical datasets where sample sizes are often limited. Manual tuning of these parameters is time-consuming and prone to bias, underscoring the need for systematic and automated optimization mechanisms.

Generalization and robustness constitute additional concerns in clinical prediction tasks. Medical data are frequently affected by measurement noise, missing values, and variability in data acquisition protocols. Furthermore, privacy and ethical considerations often restrict dataset size, increasing the risk that models may overfit to idiosyncrasies of the training data. Ensuring that predictive models maintain stable performance on unseen patient records is therefore a fundamental requirement for their potential clinical adoption.

In response to these challenges, the primary objective of this study is to develop a comprehensive analytical framework for heart disease patient analysis that integrates unsupervised clustering, deep learning, and metaheuristic optimization. The first objective is to employ unsupervised clustering al-gorithms to group patients according to clinical similarity, thereby uncovering latent structures within the data. This patient stratification step aims to provide a structured repre-sentation of the population that reflects intrinsic similarities among individuals.

Building upon the clustering stage, the study seeks to construct a hybrid learning framework in which clustering-derived insights are combined with supervised deep learning classifiers. This integration is intended to leverage the com-plementary strengths of unsupervised and supervised learning, enabling more effective modeling of complex clinical rela-tionships. In parallel, advanced metaheuristic optimization techniques are incorporated to address feature selection and hyperparameter tuning in an automated and principled man-ner. By reducing feature redundancy and optimizing model configurations, the framework aims to enhance learning effi-ciency and stability.

The overarching objective is to develop a robust analytical pipeline that improves predictive capability while maintain-ing strong generalization performance. Emphasis is placed on constructing models that are not only accurate but also resilient to noise and variability inherent in clinical data.

1.1 Research Contributions

This study contributes to the field of healthcare analytics in several important ways. First, it proposes a clustering-driven analytical framework for heart disease patient stratification, demonstrating how unsupervised learning can be systematically leveraged to uncover meaningful patient subgroups. This approach provides a foundation for more personalized analysis and supports clinically informed data exploration.

Second, the study introduces optimized deep learning models that integrate metaheuristic optimization strategies with advanced learning architectures. The use of Fitness Greylag Goose Optimization (FGGO) and other optimization techniques in conjunction with DeepGBM is designed to enhance both feature relevance and model configuration, ad-dressing key limitations associated with manual tuning and high-dimensional clinical data.

Third, a rigorous comparative evaluation framework is estab-lished to assess baseline and optimized learning strategies under consistent experimental conditions. Although the de-tailed empirical findings are presented in subsequent sections, the methodological design emphasizes reproducibility, trans-parency, and fairness in comparison.

Finally, the proposed framework is explicitly motivated by clinical applicability. By focusing on patient similarity, optimization-driven learning, and robust modeling, the study aims to support treatment-related decision-making and con-tribute to the broader objectives of personalized and precision medicine.

1.2 Structure of the Paper

The remainder of this paper is organized as follows. The subsequent section presents the dataset, data preprocessing procedures, and the methodological foundations of the proposed framework, including clustering techniques, deep learning models, and metaheuristic optimization strategies. This is followed by a detailed description of the experimental setup and evaluation methodology. The empirical analysis section examines the outcomes of the proposed approach. Finally, the paper concludes by summarizing the main insights and outlining directions for future research.

2. LITERATURE REVIEW

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, motivating extensive research into automated, accurate, and early diagnostic systems. Recent advances in machine learning (ML) and deep learning (DL) have significantly influenced cardiovascular disease detection by leveraging diverse biomedical signals, clinical records, and medical imaging modalities.

Electrocardiogram (ECG)-based diagnosis has been one of the most widely explored directions due to its noninvasive and cost-effective nature. A deep learning–driven framework utilizing transfer learning with lightweight pretrained models such as SqueezeNet and AlexNet, alongside a newly designed convolutional neural network (CNN), demonstrated the effectiveness of deep feature learning for ECG image classification [15]. By integrating deep features with traditional classifiers including support vector machines and Naïve Bayes, the study highlighted the complementary role of hybrid learning strategies in improving classification robustness and accuracy.

The integration of Internet of Things (IoT) technologies with machine learning has further expanded the scope of CVD prediction by enabling continuous data acquisition from wearable devices. A comparative study emphasized the limitations of conventional ML models in handling heterogeneous IoT-generated data and proposed enhanced learning strategies that improved predictive accuracy to nearly 96% [16]. This work underscored the importance of adaptive learning mechanisms for real-time cardiovascular monitoring systems.

Hybrid learning architectures combining machine learning and deep learning have also gained significant attention. An ensemble-based framework integrating CNN, LSTM, KNN, and XGBoost through majority voting achieved consistently high performance across multiple large-scale and local datasets [17]. This approach demonstrated that combining temporal modeling capabilities of LSTM with spatial feature extraction of CNNs enhances predictive reliability for cardiovascular disease risk assessment.

Beyond prediction models, comprehensive reviews have analyzed the broader application of artificial intelligence in coronary atherosclerotic heart disease diagnosis. Such reviews highlighted advancements across coronary angiography, CT angiography, intravascular imaging, cardiac MRI, and functional parameter analysis [18]. These studies emphasized both the transformative potential of AI-driven diagnostics and the ongoing challenges related to interpretability, data quality, and clinical integration.

Several studies have focused on benchmarking a wide range

of ML and DL algorithms using structured clinical datasets. Using a Kaggle heart disease dataset, one investigation systematically evaluated logistic regression, Naïve Bayes, KNN, SVM, multilayer perceptrons, and ensemble methods such as Random Forest and XGBoost, demonstrating the effectiveness of ensemble learners in improving diagnostic accuracy [19]. Similar comparative analyses on ECG data employed Gaussian Naïve Bayes, Random Forest, Logistic Regression, and Linear Discriminant Analysis to automate ECG classification, reinforcing the viability of ML-based ECG interpretation [20].

Unsupervised and semi-supervised learning strategies have also been explored to enhance classification outcomes. A study incorporating k-modes clustering with Huang initialization prior to supervised learning showed improved classification accuracy and AUC values, particularly when combined with multilayer perceptrons and cross-validation strategies [21]. This highlighted the value of clustering-based preprocessing in handling categorical clinical data.

Feature selection and dimensionality reduction have been recognized as critical steps in building efficient diagnostic systems. A deep neural network framework employing correlation-based feature selection and a Region-CNN architecture demonstrated superior performance compared to traditional ML models, particularly for coronary artery disease diagnosis [22]. These findings emphasized that optimal feature subsets can significantly reduce model complexity while improving predictive power.

Extensive comparative studies involving hyperparameter tuning and cross-validation further demonstrated that preprocessing, normalization, and optimization play pivotal roles in improving model performance. Evaluations across multiple classifiers, including AdaBoost, Extra Trees, LDA, and XGBoost, confirmed that tuned models consistently outperform their baseline counterparts [23].

Similarly, CatBoost-based approaches achieved high accuracy and F1-scores by focusing on automatic feature selection and early-stage disease detection [24].

Addressing data imbalance remains a major challenge in cardiovascular disease prediction. A comprehensive investigation deploying multiple ML and DL classifiers, including CNN and XGBoost, demonstrated that carefully optimized XGBoost models can achieve near-perfect diagnostic metrics for myocardial infarction detection [25]. This work emphasized the necessity of imbalance-aware learning strategies.

Lightweight deep learning models combined with ensemble classifiers have also been proposed for multiclass ECG classification. A tailored CNN for feature extraction coupled with an optimized weighted ensemble classifier achieved superior accuracy compared to traditional transfer learning approaches, demonstrating the effectiveness of domain-specific lightweight architectures [26].

Interpretability has emerged as a crucial concern in clinical AI systems. An interpretable wavelet convolution transformer (WCFormer) integrated wavelet theory with transformer architectures to provide physically meaningful feature extraction for heart sound analysis while maintaining high diagnostic accuracy [27]. This approach addressed clinician trust and transparency in AI-driven decision support systems.

The role of data preprocessing and optimized training strategies was further emphasized in a hybrid XGBoost framework that incorporated outlier removal and automated hyperparameter tuning using Optuna. The proposed model demonstrated high accuracy and efficiency across different train-test splits, highlighting the importance of data quality in predictive modeling [28].

Clinical decision support systems integrating machine learning with fuzzy clustering have also been explored to enhance interpretability and physician acceptance. A decision tree–based model validated by medical experts identified key predictors such as chest pain, anxiety, age, and smoking, while fuzzy clustering confirmed the relevance of these factors [29]. This study illustrated how hybrid analytical approaches can reduce diagnostic errors in real-world clinical settings.

Finally, advanced image-based diagnostic frameworks have leveraged data-driven decomposition techniques to augment deep learning models. By integrating higher order dynamic mode decomposition with CNNs for echocardiography image classification, one study achieved substantial accuracy improvements through effective data augmentation, demonstrating the potential of physics-informed feature extraction in medical imaging analysis [30].

Overall, the reviewed literature demonstrates a clear trend toward hybrid, ensemble, interpretable, and data-efficient machine learning frameworks for cardiovascular disease diagnosis, highlighting both the progress achieved and the challenges that remain in deploying reliable AI-driven clinical decision support systems.

Comparative Summary of Existing Studies To provide a structured and comparative overview of the existing research on cardiovascular disease prediction and diagnosis, Table 1 synthesizes the key characteristics of the reviewed studies. The table highlights the primary focus areas, adopted methodologies, and major contributions of each work, enabling a concise comparison across different data modalities, learning paradigms, and clinical objectives. This comparative analysis facilitates the identification of prevailing trends, methodological strengths, and research gaps in machine learning- and deep learning-based cardiovascular disease studies.

3. MATERIALS AND METHODS

3.1 Dataset Description

The dataset employed in this study consists of anonymized clinical records of patients diagnosed with heart disease and was obtained from a medical center. The data were collected as part of routine clinical assessments and subsequently anonymized to ensure patient privacy and compliance with ethical standards for medical data analysis. The primary clinical motivation for utilizing this dataset lies in its comprehensive representation of cardiovascular risk factors and diagnostic indicators that are routinely used by clinicians to assess heart disease severity and progression. As such, the dataset provides a realistic and clinically meaningful basis for exploring patient similarity, stratification, and data-driven modeling approaches in cardiovascular healthcare.

The clinical context of the dataset reflects a heterogeneous patient population, encompassing a wide range of ages, physiological conditions, and diagnostic outcomes. This hetero-

Table 1. Summary of related work on cardiovascular disease prediction and diagnosis

Ref	Focus Area	Methodology	Key Findings and Contributions
[15]	ECG-based cardiac abnormality classification	Transfer learning, custom CNN, and ML classifiers	Proposed CNN and hybrid deep feature extraction achieved near-perfect classification performance.
[16]	IoT-enabled CVD prediction	Comparative machine learning models on IoT health data	Improved prediction accuracy by addressing data heterogeneity.
[17]	General CVD risk prediction	CNN, LSTM, KNN, and XGBoost ensemble	Superior performance across multiple datasets.
[21]	CVD classification with clustering enhancement	k-modes clustering with DT, RF, MLP, and XGBoost	Improved accuracy and AUC by integrating clustering before classification.
[24]	Early-stage CVD prediction	CatBoost with automated feature selection	High accuracy and F1-score with efficient early detection.
[28]	Heart disease classification	XGBoost with Optuna hyperparameter tuning	Improved classification under outliers and heterogeneous training datasets.

geneity makes the dataset particularly suitable for unsupervised learning tasks, such as clustering, where the goal is to uncover latent patient subgroups without relying on predefined labels. Moreover, the real-world nature of the data introduces practical challenges commonly encountered in clinical analytics, including variability in measurements and overlapping clinical profiles, thereby enhancing the relevance of the proposed analytical framework.

The dataset comprises a set of numerical features that capture demographic, physiological, and diagnostic aspects of heart disease patients. Demographic attributes include patient age and sex, which are fundamental variables known to influence cardiovascular risk and disease manifestation.

These features provide essential context for understanding population-level differences and individual susceptibility to heart disease.

Clinical measurements constitute a substantial portion of the dataset and include resting blood pressure, serum cholesterol levels, and maximum heart rate achieved during exercise. These variables are routinely measured in clinical practice and serve as key indicators of cardiovascular health and functional capacity. Variations in these measurements often reflect differences in disease severity, lifestyle factors, and underlying physiological conditions, making them highly informative for patient similarity analysis.

In addition to demographic and physiological variables, the dataset includes diagnostic indicators that describe specific clinical findings. These indicators encompass chest pain type, resting electrocardiographic results, fasting blood sugar status, and the presence or absence of exercise-induced angina. Collectively, these diagnostic features capture qualitative and quantitative aspects of cardiac function and symptom presentation, providing a richer representation of patient health status. The inclusion of both continuous and discretized numerical variables ensures compatibility with clustering algorithms while preserving clinically relevant information. Figures 1 and 2 present the exploratory visual summaries used to examine feature distributions and age-related patterns before preprocessing.

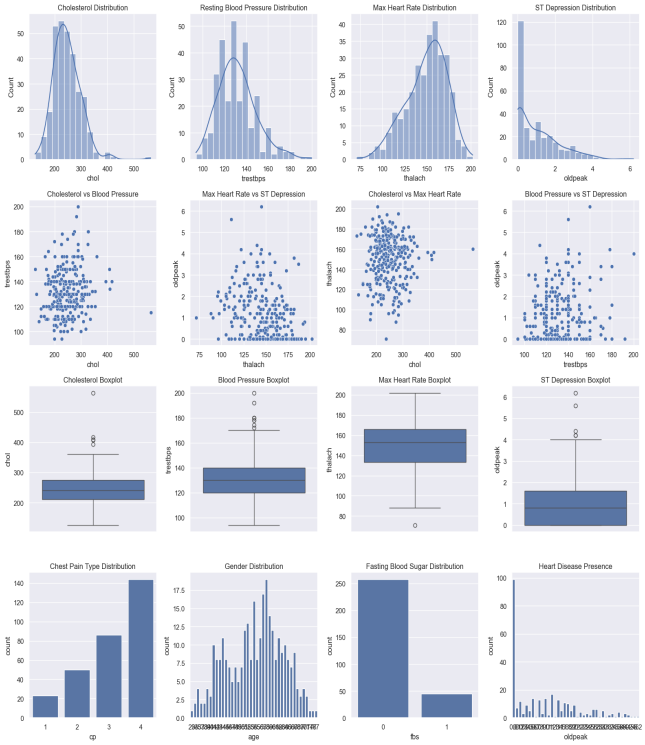

Figure 1. Exploratory data analysis visualizations of clinical and demographic features, including univariate distributions, bivariate relationships, and categorical variable frequencies related to heart disease.

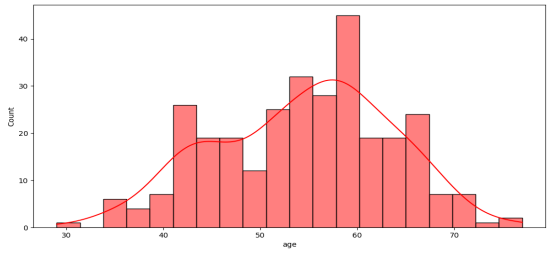

Figure 2. Distribution of patient ages in the dataset represented using a histogram with an overlaid kernel density estimate.

3.2 Data Preprocessing

Effective data preprocessing is a critical prerequisite for reliable clustering and deep learning–based analysis, particularly in the context of clinical datasets that are inherently heterogeneous and sensitive to noise. In this study, a structured preprocessing pipeline was designed to ensure data integrity, numerical compatibility with unsupervised algorithms, and balanced feature influence prior to optimization and model development.

The initial preprocessing step focused on data cleaning and validation to ensure the suitability of the dataset for clustering analysis. Since the clustering algorithms employed in this study require numerical input, all features were systematically examined to verify that they were represented in numeric form. Categorical clinical variables, such as chest pain type and resting electrocardiographic results, were already encoded numerically in the dataset, thereby ensuring compatibility with distance-based clustering methods without the need for additional transformation.

Data validation procedures were applied to identify inconsistencies, out-of-range values, or invalid entries that could adversely affect downstream analysis. Clinical attributes were checked against medically plausible ranges to confirm that recorded values aligned with expected physiological limits.

This validation step is particularly important in healthcare datasets, where data entry errors or measurement artifacts may introduce spurious patterns that distort similarity calculations. By ensuring that all variables were numerically consistent and clinically reasonable, the preprocessing stage established a reliable foundation for subsequent unsupervised and supervised learning tasks.

Following data validation, feature scaling was applied to address differences in measurement units and value ranges across clinical variables. Heart disease datasets typically include features measured on heterogeneous scales, such as age in years, blood pressure in millimeters of mercury, cholesterol in milligrams per deciliter, and heart rate in beats per minute. Without appropriate scaling, variables with larger numeric ranges can dominate distance computations in clustering algorithms and disproportionately influence model training.

To mitigate this issue, normalization or standardization techniques were employed to rescale the features to a common scale. Feature scaling ensures that each clinical attribute contributes more equally to similarity assessments and learning processes, thereby preventing bias toward high-magnitude variables. This step is essential for both unsupervised clustering, where distance metrics play a central role, and deep learning models, where balanced input distributions facilitate stable training dynamics and convergence behavior.

In addition to scaling, correlation and redundancy analysis was conducted to examine interdependencies among clinical features. Many physiological and diagnostic variables in cardiovascular datasets are inherently correlated due to underlying biological relationships. While such correlations may carry clinical significance, excessive redundancy can negatively impact clustering quality and model efficiency by inflating dimensionality without adding new information.

Pairwise correlation analysis was used to identify strongly correlated feature pairs, providing insight into potential redundancy within the dataset. Features exhibiting high correlation were carefully examined to assess their relative clinical importance and informational contribution. This analysis informed subsequent optimization stages by highlighting candidate features for reduction or prioritization, thereby supporting more compact and informative representations of patient data.

By addressing redundancy prior to optimization, the preprocessing pipeline aimed to reduce unnecessary complexity, improve computational efficiency, and enhance the interpretability of clustering outcomes.

Collectively, the data cleaning, scaling, and redundancy analysis steps ensured that the dataset was well-conditioned for the application of unsupervised clustering, deep learning models, and metaheuristic optimization techniques in the later stages of the proposed framework.

3.3 Deep Learning Models

The complex and multifactorial nature of heart disease necessitates the use of advanced modeling techniques capable of capturing nonlinear relationships, higher-order feature interactions, and latent structures within clinical data. Traditional machine learning approaches often struggle to adequately represent such complexity, particularly when dealing with heterogeneous variables that span demographic information, physiological measurements, and diagnostic indicators. For this reason, deep learning models were selected as the core predictive tools in this study, owing to their proven ability to learn rich representations directly from data and to model intricate dependencies among input features.

A total of five deep learning–based machine learning models were considered in this work. The selection of these models was guided by their methodological diversity, relevance to structured and quasi–time-series data, and their established use in healthcare and biomedical analytics. Collectively, these models provide a comprehensive comparative framework for assessing different architectural paradigms and learning mechanisms in the context of heart disease patient analysis.

boosting decision trees with neural network–based representation learning. This model is specifically designed to address the limitations of conventional deep neural networks when applied to structured tabular data, which is the predominant format of clinical datasets. DeepGBM leverages tree-based feature transformations to capture complex feature interactions and nonlinearities, while a deep neural component learns higher-level representations from these transformed features. This dual structure allows DeepGBM to effectively handle heterogeneous clinical variables, reduce sensitivity to feature scaling issues, and maintain a balance between predictive power and interpretability. These characteristics make DeepGBM particularly well suited for heart disease data, where clinical variables often exhibit nonlinear relationships and conditional dependencies.

The Dynamic Temporal Convolutional Network is included as a model capable of capturing ordered dependencies and local-to-global patterns within data. Although originally proposed for temporal and sequential modeling, DTCN architectures are also applicable to structured clinical data where implicit ordering or progression patterns may exist among features. By employing causal and dilated convolutions, DTCN models can learn both short-range and long-range dependencies without relying on recurrent connections. This property enables efficient training and stable gradient propagation, making DTCNs attractive for modeling complex dependencies among physiological measurements that may reflect disease progression or functional relationships.

The Deep Belief Network is incorporated as a representative generative–discriminative deep learning model that employs a layered probabilistic structure. DBNs are constructed by stacking restricted Boltzmann machines, enabling unsupervised pretraining of each layer before supervised fine-tuning.

This hierarchical learning process allows DBNs to extract increasingly abstract feature representations, which can be advantageous in clinical contexts where meaningful patterns may not be immediately apparent in the raw data. The inclusion of DBNs in this study provides insight into the effectiveness of probabilistic and generative feature learning approaches relative to more modern deterministic architectures.

The Convolutional Neural Network is examined due to its established success in learning localized feature interactions and hierarchical representations. While CNNs are most commonly associated with image and signal processing tasks, their convolutional structure has been successfully adapted to structured and transformed tabular data. In the context of heart disease analysis, CNNs can capture local dependencies among subsets of clinical features and learn compositional patterns that may correspond to underlying physiological relationships. As one of the most widely used deep learning architectures, CNNs serve as an important benchmark for evaluating the relative benefits of more specialized hybrid models.

The Generative Adversarial Network is included as an advanced deep learning paradigm that introduces an adversarial learning mechanism between a generator and a discriminator network. Although GANs are primarily known for their data generation capabilities, their discriminative components and learned representations have been explored in various medical data analysis tasks. The adversarial training process encourages the model to learn robust and informative feature representations, which can be beneficial in settings characterized by complex data distributions. Including GANs in the comparative analysis broadens the methodological scope of the study and allows for the examination of adversarial learning concepts in clinical prediction scenarios.

Following a comprehensive baseline evaluation of all five deep learning models, DeepGBM was selected as the primary model for subsequent optimization. This selection was based on its architectural suitability for structured clinical data, its ability to effectively capture nonlinear feature interactions, and its compatibility with feature selection and hyperparameter optimization strategies. DeepGBM's hybrid design makes it particularly amenable to enhancement through metaheuristic optimization, as both its feature transformation components and learning parameters can be systematically tuned.

Consequently, DeepGBM serves as the core predictive model within the proposed framework, while the remaining deep learning models provide valuable reference points for comparative and methodological analysis.

3.4 Metaheuristic Optimization Framework

The effectiveness of deep learning models in medical decision support systems is strongly influenced by the selection of appropriate hyperparameters. These hyperparameters govern critical aspects of the learning process, including optimization dynamics, model capacity, regularization strength, and convergence behavior. In the context of heart disease patient analysis, where datasets are often limited in size and characterized by complex, nonlinear relationships among clinical variables, improper hyperparameter configurations can significantly degrade model reliability and stability. Consequently, a systematic and intelligent optimization strategy is essential to fully exploit the representational power of deep learning architectures while mitigating risks such as overfitting and unstable training.

Hyperparameter optimization is inherently a challenging problem due to the high dimensionality, nonconvexity, and interdependence of parameters. Traditional optimization approaches, such as manual tuning or exhaustive grid search, are not only computationally expensive but also poorly suited to exploring large and continuous search spaces. These limitations are particularly pronounced in deep learning models, where even small changes in hyperparameter values can lead to substantial variations in training outcomes. To address these challenges, this study adopts a metaheuristic optimization framework that formulates hyperparameter tuning as a global optimization problem and leverages nature-inspired search strategies to identify high-quality solutions efficiently.

3.4.1 Role of Metaheuristics in Hyperparameter Optimization

Metaheuristic algorithms are designed to navigate complex optimization landscapes by balancing exploration of the global search space with exploitation of locally promising regions. This balance is especially important for deep learning hyperparameter optimization, where the loss surface is often rugged and populated with numerous local optima. By maintaining a population of candidate solutions and iteratively updating them according to adaptive rules, metaheuristics can escape local minima and explore diverse regions of the search space more effectively than deterministic methods.

In deep learning architectures applied to clinical datasets, hyperparameters such as learning rates, batch sizes, regularization coefficients, and architectural configuration parameters play a decisive role in convergence speed and predictive stability. Poorly tuned hyperparameters may lead to vanishing or exploding gradients, slow convergence, or excessive sensitivity to noise in the data. Metaheuristic optimization automates the tuning process by evaluating candidate configurations based on a predefined fitness function and progressively refining them through iterative search mechanisms. This automation not only reduces human intervention but also improves the consistency and reproducibility of model development.

Furthermore, metaheuristics are inherently flexible and model-agnostic, making them suitable for optimizing a wide range of deep learning architectures without requiring gradient information or explicit assumptions about the objective function. This property is particularly advantageous in healthcare applications, where models may involve complex training procedures and evaluation metrics.

By employing metaheuristic-based hyperparameter optimization, the proposed framework aims to enhance learning efficiency, stabilize convergence behavior, and support the development of robust predictive models capable of generalizing to unseen patient data.

3.5 Proposed FGGO-Optimized DeepGBM Framework

This study proposes an integrated optimization framework in which Fitness Greylag Goose Optimization (FGGO) is employed to enhance the DeepGBM model within a clustering-informed learning pipeline.

The proposed framework is designed to jointly address hyperparameter optimization and learning efficiency while leveraging patient stratification obtained from unsupervised clustering. By embedding search space and guides the learning process toward configurations that are more stable and suitable for complex clinical data.

3.5.1 Fitness Function Formulation

The core of the FGGO-based optimization process is the fitness function, which quantitatively evaluates the quality of each candidate solution within the population. In the context of the proposed framework, each search agent represents a candidate configuration of DeepGBM hyperparameters. The fitness function is formulated to reflect the predictive capability and stability of the DeepGBM model under a given configuration. During optimization, DeepGBM is trained using the hyperparameters encoded by each agent, and its performance on the validation subset is used as the objective value.

The optimization objective is defined in a minimization or maximization form depending on the selected evaluation criterion. By consistently applying the same fitness evaluation protocol across all agents, optimization process to explicitly favor configurations that lead to improved convergence behavior and robust learning, while discouraging unstable or poorly generalized solutions. The fitness-driven nature of FGGO enables adaptive movement of search agents toward promising regions of the hyperparameter space.

3.5.2 Optimization Workflow Integrating Clustering and Classification

The proposed FGGO-Optimized DeepGBM framework follows a structured workflow that integrates unsupervised clustering with supervised classification and metaheuristic optimization. Initially, patient data are preprocessed and subjected to clustering in order to uncover latent patient subgroups based on clinical similarity. The resulting clustered representation serves as an informative foundation for subsequent classification, ensuring that the learning process accounts for intrinsic data structure and patient heterogeneity.

Following clustering, DeepGBM is employed as the primary classification model. Rather than relying on fixed or manually selected hyperparameters, FGGO is invoked to optimize the DeepGBM configuration. Each FGGO agent encodes a candidate set of DeepGBM hyperparameters, and the population collectively explores the search space through iterative updates governed by exploration and exploitation mechanisms. During each iteration, agents are evaluated using the fitness function, and the best-performing agent guides the search process.

Algorithm 1 presents the detailed pseudocode of the proposed FGGO algorithm. The algorithm begins by initializing a population of search agents along with FGGO control parameters and the maximum number of iterations. The objective function is evaluated for all agents, and the best agent is identified as the current leader. The population is then divided into exploration and exploitation groups, enabling adaptive search behavior.

During the optimization process, agents in the exploration group update their positions using multiple stochastic and fitness-driven strategies, allowing the algorithm to investigate diverse regions of the search space. In contrast, agents in the exploitation group focus on refining solutions around the best-performing configurations. The algorithm dynamically adjusts the size of these groups based on convergence behavior, increasing exploration when stagnation is detected and reinforcing exploitation when improvement is observed. This adaptive mechanism helps balance global search and local refinement throughout the optimization process.

At each iteration, boundary constraints are enforced to ensure that candidate solutions remain within the feasible search space. The fitness values are recalculated, parameters are updated, and the iteration counter is incremented until the termination criterion is met. The final output of the algorithm is the best-performing agent, which corresponds to the optimized DeepGBM hyperparameter configuration.

This configuration is then used to train the final DeepGBM model for heart disease prediction.

Algorithm 1Proposed Fitness Greylag Goose Optimization (FGGO) Algorithm

1: Initialize FGGO population X_i (i = 1, 2, ..., n), population size n, maximum iterations t_{max}, and objective function F_n

2: Initialize FGGO parametersa, A, C, b, l, c, r 1, r 2, r 3, r 4, r 5, w, w 1, w 2, w 3, w 4, A 1, A 2, A 3, C 1, C 2, C 3 3: Calculate objective function F_n for each agent X_i 4: SetPas the best agent position 5: Divide population into exploration group(n 1)and exploitation group(n 2) 6: whilet≤t max do 7: fori= 1ton 1 do 8: iftmod 2 = 0then 9: ifr 3 <0.5then 10: if|A|<1then 11: X(t+ 1) =X *(t)-A·|C·X*(t)-X(t)| 12: else 13: Select three random agentsX Paddle1, X Paddle2, X Paddle3 14: z= 1- (t tmax)2

15: X(t+ 1) = w1 w1+w2+w3 XPaddle1 +z w2 w1+w2+w3 (XPaddle2-XPaddle3)

16: end if 17: else 18: X(t+ 1) =w 4|X*(t)-X(t)|ebl cos(2πl) + [2w1(r4 +r 5)]X *(t) 19: end if 20: else 21: X(t+ 1) =X(t) +D(1 +z)w(X-X Flock1) 22: end if 23: end for 24: fori= 1ton 2 do 25: iftmod 2 = 0then 26: ComputeX 1, X 2, X 3 using sentry agents 27: X(t+ 1) =X 3 i

28: else 29: X(t+ 1) =X(t) +D(1 +z)w(X-X Flock1) 30: end if 31: end for 32: Calculate objective function F_n for all agents 33: Update parameters and incrementt 34: ifbestF n unchanged for two iterationsthen 35: Increasen 1, decreasen 2 36: end if 37: end while 38: returnbest agentP The proposed FGGO-Optimized DeepGBM framework thus establishes a tightly coupled optimization and learning process, in which clustering-informed data representation, deep learning classification, and adaptive metaheuristic optimization operate cohesively. This integration enables the development of a robust and efficient predictive system tailored to the complexities of heart disease patient data.

3.5.3 State-of-the-Art Metaheuristic Algorithms

To comprehensively assess the impact of different optimization strategies, several state-of-the-art metaheuristic algorithms were incorporated into the proposed framework. These algorithms represent diverse optimization philosophies and search dynamics, allowing for a systematic comparison of their effectiveness in deep learning hyperparameter tuning.

Fitness Greylag Goose Optimization (FGGO) is a population-based metaheuristic inspired by the collective movement and migratory behavior of greylag geese. FGGO emphasizes fitness-driven adaptation, where candidate solutions adjust their trajectories based on both individual performance and group-level information. The algorithm dynamically balances exploration and exploitation by modulating movement strategies in response to fitness feedback, enabling it to efficiently

traverse high-dimensional and nonlinear search spaces. In this study, FGGO plays a central role due to its strong capability to guide the search toward high-quality hyperparameter configurations while maintaining sufficient population diversity to avoid premature convergence.

Particle Swarm Optimization (PSO) is a classical swarm intelligence algorithm that models the social behavior observed in flocks of birds and schools of fi sh. In PSO, each particle represents a candidate solution that updates its position in the search space based on its own historical best position and the globally best position identified by the s warm. This cooperative learning mechanism allows PSO to converge rapidly toward promising regions of the search space while retaining exploratory behavior through stochastic velocity updates. PSO is widely recognized for its simplicity, computational efficiency, and effectiveness in continuous optimization problems, making it a strong baseline for hyperparameter optimization in deep learning models.

The Grey Wolf Optimizer (GWO) is inspired by the leadership hierarchy and hunting strategy of grey wolves in nature. GWO organizes candidate solutions into a hierarchical structure, where the best-performing solutions guide the search process and influence the movement of other candidates.

This hierarchy-driven mechanism promotes focused exploitation of high-quality regions while still enabling exploration through subordinate agents. GWO's ability to maintain a balance between convergence speed and search diversity makes it suitable for optimizing deep learning hyperparameters characterized by interdependent and nonlinear effects.

Dipper Throated Optimization (DDTO) is a nature-inspired algorithm that simulates the diving and foraging behavior of dipper-throated birds. The algorithm alternates between exploration phases, where candidate solutions search broadly across the solution space, and exploitation phases, where they concentrate on refining promising s olutions. This adaptive switching mechanism enables DDTO to respond dynamically to the fitness landscape and avoid stagnation. In the context of hyperparameter optimization, such adaptive behavior is particularly valuable for identifying narrow optimal regions that may otherwise be overlooked by more rigid search strategies.

Multiverse Optimization (MVO) draws inspiration from cosmological theories involving multiple interacting universes. In MVO, candidate solutions are conceptualized as universes that exchange information through probabilistic mechanisms associated with white holes, black holes, and wormholes.

These interactions allow high-quality solutions to influence others while preserving diversity across the population. MVO's unique information-sharing strategy facilitates both global exploration and local exploitation, making it well suited for high-dimensional optimization problems such as deep learning hyperparameter tuning. this study provides into a unified optimization.

this study provides a comprehensive evaluation of nature-inspired optimization strategies for enhancing deep learning architectures in heart disease patient analysis. The diversity of search mechanisms and adaptive behaviors represented by FGGO, PSO, GWO, DDTO, and MVO enables a thorough investigation of how different optimization philosophies influence learning stability, convergence behavior, and overall model robustness in a clinically relevant setting.

3.6 Evaluation Metrics

The evaluation of predictive models in healthcare applications requires a comprehensive and carefully designed set of performance metrics that reflect both statistical correctness and clinical relevance.

In heart disease analysis, the consequences of misclassification can be significant, as false negatives may delay critical treatment while false positives may lead to unnecessary medical procedures and patient anxiety. For this reason, model assessment in this study is not limited to a single metric but instead relies on a collection of complementary classification measures that together provide a multidimensional view of model behavior.

The selected evaluation metrics are widely adopted in medical machine learning studies and are particularly suitable for binary classification problems in clinical decision support systems. These metrics quantify different aspects of predictive performance, including overall correctness, sensitivity to diseased cases, specificity toward healthy cases, and the reliability of positive and negative predictions.

By jointly analyzing these measures, the evaluation framework ensures that the developed models are assessed not only in terms of predictive accuracy but also in terms of their potential clinical utility and safety.

Let TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. A true positive corresponds to a patient with heart disease who is correctly identified by the model, whereas a true negative represents a healthy individual correctly classified as non-diseased. False positives and false negatives denote incorrect predictions that may carry different clinical risks. Based on these quantities, the evaluation metrics used in this study are formally defined in Table 2.

Table 2. Classification performance metrics and mathematical formulations

Metric	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity (TPR)	$\frac{TP}{TP+FN}$
Specificity (TNR)	$\frac{TN}{TN+FP}$
PPV	$\frac{TP}{TP+FP}$
NPV	$\frac{TN}{TN+FN}$
F-Score	$\frac{2\times PPV\times Sensitivity}{PPV+Sensitivity}$

Accuracy Accuracy= TP+TN TP+TN+FP+FN Sensitivity (TPR) Sensitivity= TP TP+FN Specificity (TNR) Specificity= TN TN+FP Positive Predictive Value (PPV) PPV= TP TP+FP Negative Predictive Value (NPV) NPV= TN TN+FN F-Score FrameScore=work,2× PPV×Sensitivity PPV+Sensitivity Accuracy provides a global measure of classification performance by quantifying the proportion of correctly classified instances among all predictions. While accuracy is intuitive and easy to interpret, it can be misleading in medical datasets where class imbalance is common. In such cases, a model may achieve high accuracy by favoring the majority class while performing poorly on clinically critical minority cases. Therefore, accuracy is interpreted in conjunction with other, more

discriminative metrics.

Sensitivity, also known as the true positive rate, measures the model's ability to correctly identify patients who actually have heart disease. High sensitivity is essential in clinical screening and diagnostic applications, as it reduces the likelihood of missed diagnoses. From a medical standpoint, sensitivity directly relates to patient safety, since false negatives may result in delayed intervention or disease progression without appropriate treatment.

Specificity, or true negative rate, quantifies the model's ability to correctly classify non-diseased individuals. This metric is crucial for avoiding unnecessary diagnostic procedures and treatments in healthy patients. In clinical practice, high specificity contributes to efficient resource utilization and reduces the psychological and economic burden associated with false alarms.

Positive Predictive Value evaluates the reliability of positive predictions by measuring the proportion of predicted positive cases that are truly diseased. PPV is particularly relevant from a physician's perspective, as it reflects the confidence that can be placed in a positive model output. Similarly, Negative Predictive Value measures the reliability of negative predictions, indicating the likelihood that a patient predicted as healthy is indeed free from heart disease. Both PPV and NPV are highly informative in real-world clinical settings, where predictive confidence directly influences decision-making.

The F-score provides a balanced assessment by combining sensitivity and precision into a single metric.

By harmonizing the trade-off between false positives and false negatives, the F-score offers a concise summary of classification performance, particularly in scenarios where class imbalance exists. This metric is valuable for comparing models under different optimization strategies, as it captures both detection capability and prediction reliability.

Collectively, the evaluation metrics summarized in Table 2 establish a robust and clinically meaningful assessment framework. Their combined use ensures that the proposed models are evaluated from multiple perspectives, aligning quantitative performance analysis with the practical requirements and ethical considerations of healthcare decision support systems.

4. EXPERIMENTAL RESULTS

4.1 Baseline Model Performance

This subsection provides an in-depth analysis of the baseline predictive performance of the deep learning models prior to the application of any optimization or feature selection mechanisms. Establishing a baseline is a critical step in experimental evaluation, as it offers a clear reference against which the effectiveness of metaheuristic optimization and hybrid modeling strategies can later be quantified.

All models were trained and evaluated under identical conditions, using the same dataset partitions, preprocessing pipeline, and evaluation metrics, ensuring a fair and unbiased comparison.

Table 3 presents the quantitative performance of the five deep learning models in terms of accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score.

Table 3. Baseline performance comparison of deep learning models without optimization

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F-Score
DeepGBM	0.9032	0.8824	0.9195	0.8955	0.9091	0.8889
DTCN	0.8901	0.8657	0.9091	0.8815	0.8966	0.8735
DBN	0.8758	0.8507	0.8953	0.8636	0.8851	0.8571
CNN	0.8627	0.8358	0.8837	0.8485	0.8736	0.8421
GAN	0.8455	0.8151	0.8690	0.8279	0.8588	0.8214

The results clearly indicate that DeepGBM achieves the strongest baseline performance among all evaluated models. Specifically, DeepGBM attains an accuracy of 0.9032, demonstrating that more than 90% of patient instances are correctly classified without any optimization. In terms of clinical sensitivity, DeepGBM achieves a sensitivity of 0.8824, indicating a strong ability to correctly identify patients with heart disease. This is complemented by a high specificity of 0.9195, reflecting reliable discrimination of non-diseased individuals. Furthermore, DeepGBM records a positive predictive value of 0.8955 and a negative predictive value of 0.9091, highlighting the robustness and reliability of both positive and negative predictions. The resulting F-score of 0.8889 confirms a well-balanced trade-off between sensitivity and precision, reinforcing DeepGBM's suitability for structured clinical data.

The Dynamic Temporal Convolutional Network demonstrates competitive but slightly inferior performance compared to DeepGBM. DTCN achieves an accuracy of 0.8901, which is approximately 1.3% lower than DeepGBM. Its sensitivity of 0.8657 indicates a modest reduction in detecting true heart disease cases, while its specificity of 0.9091 remains relatively high. The PPV and NPV values of 0.8815 and 0.8966, respectively, suggest stable predictive confidence; however, the F-score of 0.8735 reflects a noticeable decline in balanced performance relative to DeepGBM. These results suggest that while DTCN is effective in modeling structured dependencies, it may require further tuning to fully capture complex clinical feature interactions.

The Deep Belief Network exhibits moderate baseline performance, achieving an accuracy of 0.8758.

Its sensitivity of 0.8507 and specificity of 0.8953 indicate reasonable discriminative capability, though both metrics are consistently lower than those observed for DeepGBM and DTCN. The PPV and NPV values of 0.8636 and 0.8851 further illustrate this performance gap. The F-score of 0.8571 suggests that while DBN benefits from hierarchical feature learning, its probabilistic structure may limit its ability to fully exploit the nonlinear relationships present in the heart disease dataset without optimization.

The Convolutional Neural Network records an accuracy of 0.8627, reflecting a further decline in baseline performance. Its sensitivity of 0.8358 indicates reduced effectiveness in identifying diseased patients, while its specificity of 0.8837 remains acceptable but lower than that of the higher-performing models. The PPV and NPV values of 0.8485 and 0.8736 demonstrate moderate predictive reliability.

The F-score of 0.8421 highlights the limitations of conventional CNN architectures when applied directly to tabular clinical data, where spatial locality assumptions may not align well with feature relationships.

The Generative Adversarial Network exhibits the weakest baseline performance among the evaluated models. GAN achieves an accuracy of 0.8455, with sensitivity and specificity values of 0.8151 and 0.8690, respectively. These results indicate a comparatively higher rate of both false negatives and false positives. The PPV of 0.8279 and NPV of 0.8588 further suggest reduced confidence in prediction outcomes. The lowest F-score of 0.8214 underscores the difficulty of leveraging adversarial learning frameworks for direct classification tasks in structured clinical datasets without specialized adaptation.

Overall, the baseline results reveal clear performance stratification among the deep learning models.

While all models demonstrate a reasonable capacity to learn from heart disease data, none achieves uniformly optimal performance across all metrics. The consistent superiority of DeepGBM across accuracy, specificity, NPV, and F-score highlights its robustness and suitability as a core predictive model. At the same time, the observed performance gaps and metric trade-offs across all models emphasize the limitations of baseline configurations, particularly with respect to hyperparameter sensitivity and feature redundancy. These findings strongly motivate the adoption of metaheuristic optimization strategies in subsequent sections to further enhance predictive accuracy, balance clinical trade-offs, and improve model generalization.

A comprehensive evaluation of classification models requires the simultaneous analysis of multiple performance metrics in order to capture complementary aspects of predictive behavior. In this study, model performance is assessed using accuracy, sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), negative predictive value (NPV), and F-score, which together provide a balanced view of discriminative capability and clinical reliability. Figure 3 presents swarm plots for these evaluation metrics across the investigated models, enabling a clear comparison of performance variability and central tendencies.

The visualization incorporates mean values and standard deviation bounds for each metric, thereby facilitating the identification of models that consistently perform above or below the average range.

Such graphical analysis is particularly valuable for highlighting robustness, stability, and trade-offs among models, which are critical considerations in selecting appropriate machine learning approaches for medical decision-support systems.

Comparative performance analysis across machine learning models is essential for understanding trade-offs between different evaluation metrics, particularly in medical classification tasks where both overall correctness and class-specific reliability are critical. Accuracy provides a global measure of correct predictions, while specificity (true negative rate) reflects a model's ability to correctly identify non-diseased cases, which is especially important for minimizing false positives in clinical decision-making. Figure 4 illustrates a dual-axis visualization that simultaneously presents accuracy values as scatter points and specificity values as a line plot across the evaluated models.

By combining these two metrics within a single visualization, the figure enables direct comparison of model behavior

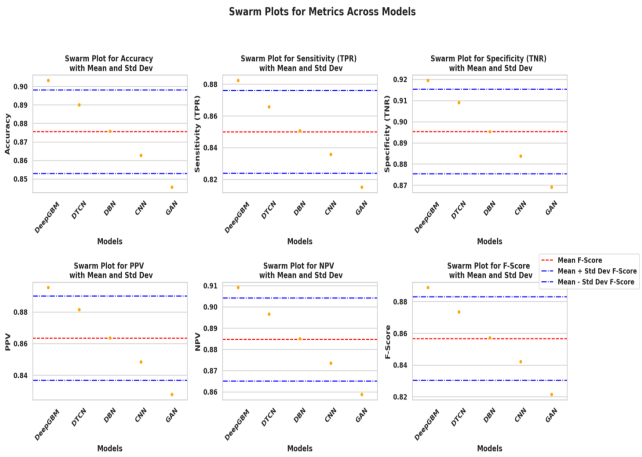

Figure 3. Swarm plots illustrating the distribution of performance metrics across different classification models, including accuracy, sensitivity (TPR), specificity (TNR), positive predictive value (PPV), negative predictive value (NPV), and F-score. Dashed and dash-dotted lines represent the mean and mean±standard deviation, respectively.

with respect to general performance and negative-class discrimination. This integrated representation facilitates the identification of models that achieve a favorable balance between accuracy and specificity, thereby supporting informed model selection for reliable and clinically meaningful heart disease prediction systems.

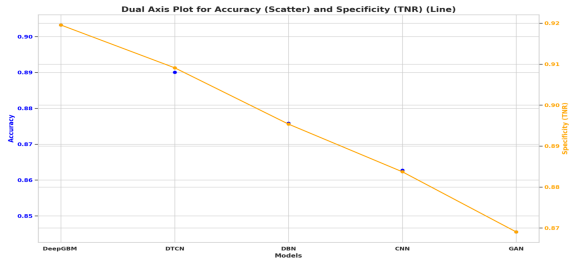

Figure 4. Dual-axis visualization of model performance, where accuracy is represented using scatter points on the left y-axis and specificity (true negative rate) is shown as a line plot on the right y-axis across different classification models.

Assessing the statistical distribution of evaluation metrics is an important step in validating the reliability and robustness of comparative model analysis. Many parametric statistical techniques assume that performance metrics follow an approximately normal distribution; therefore, verifying this assumption is essential before conducting further inferential analysis. Figure 5 presents quantile–quantile (Q–Q) plots for key performance metrics, including accuracy, sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), negative predictive value (NPV), and F-score.

The Q–Q plots compare the empirical quantiles of each metric against the corresponding theoretical quantiles of a normal distribution, allowing for visual assessment of normality, symmetry, and potential deviations such as skewness or outliers. A close alignment of data points with the reference line indicates approximate normality, thereby supporting the use of mean-based comparisons and standard deviation measures in the subsequent performance evaluation of the classification models.

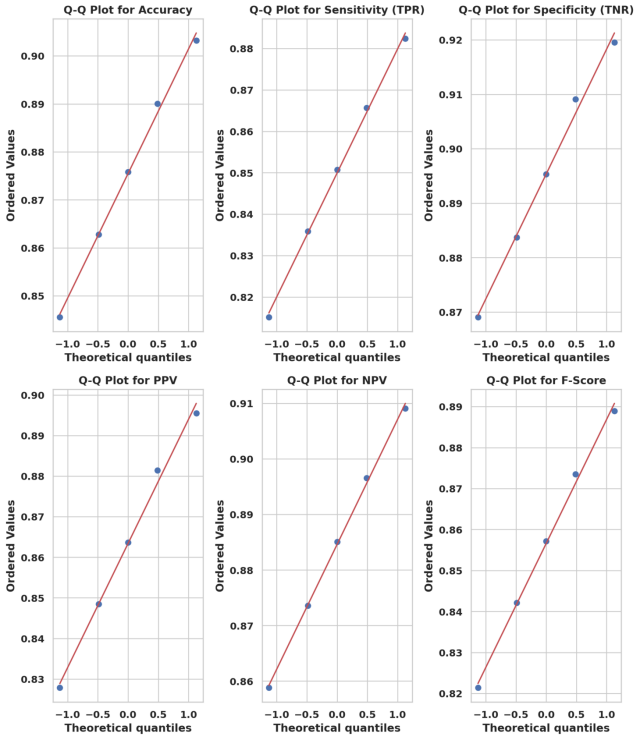

Figure 5. Quantile–quantile (Q–Q) plots for performance evaluation metrics, including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score.

4.2 Optimized Model Performance

This subsection presents a detailed analysis of the predictive performance achieved after integrating metaheuristic optimization strategies with the DeepGBM model. The primary objective of this analysis is to quantify the impact of metaheuristic-driven hyperparameter optimization on classification effectiveness and to examine how different optimization algorithms influence model behavior across all evaluation metrics. By comparing optimized models against one another under identical experimental conditions, this study provides a rigorous assessment of optimization efficacy and stability.

Table 4 summarizes the performance of DeepGBM optimized using Fitness Greylag Goose Optimization (FGGO), Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Dipper Throated Optimization (DDTO), and Multiverse Optimization (MVO).

Table 4. Performance comparison of optimized DeepGBM models using different metaheuristic algorithms

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F-Score
FGGO + DeepGBM	0.9795	0.9747	0.9831	0.9776	0.9809	0.9761
PSO + DeepGBM	0.9700	0.9633	0.9749	0.9663	0.9727	0.9648
GWO + DeepGBM	0.9656	0.9574	0.9716	0.9612	0.9687	0.9593
DDTO + DeepGBM	0.9621	0.9535	0.9683	0.9567	0.9660	0.9551
MVO + DeepGBM	0.9578	0.9478	0.9651	0.9518	0.9621	0.9498

The results in Table 4 clearly demonstrate that all metaheuristic optimization techniques lead to substantial performance improvements over the baseline DeepGBM configuration across all evaluation metrics. Among the optimized models, FGGO + DeepGBM achieves the highest performance consistently. Specifically, it records an accuracy of 0.9795, indicating that nearly 98% of patient instances are correctly classified. This represents a marked improvement in overall predictive capability and highlights the effectiveness of FGGO in navigating the hyperparameter search space.

From a clinical perspective, FGGO + DeepGBM attains a sensitivity of 0.9747, reflecting an exceptional ability to correctly identify patients with heart disease. This high sensitivity is complemented by a specificity of 0.9831, demonstrating strong discrimination of non-diseased individuals. The positive predictive value of 0.9776 and negative predictive value of 0.9809 further confirm the reliability of the model's predictions. The resulting F-score of 0.9761 indicates a highly balanced trade-off between precision and recall, underscoring the robustness of the FGGO-optimized model.

The PSO + DeepGBM configuration also exhibits strong performance, achieving an accuracy of 0.9700. Its sensitivity and specificity values of 0.9633 and 0.9749, respectively, indicate reliable detection of both diseased and non-diseased cases. With an F-score of 0.9648, PSO-based optimization demonstrates effective convergence behavior, though it remains consistently below FGGO across all metrics.

Similarly, GWO + DeepGBM achieves an accuracy of 0.9656, with sensitivity and specificity values of 0.9574 and 0.9716. While these results confirm the effectiveness of leadership-based optimization mechanisms, the slight reductions in PPV (0.9612), NPV (0.9687), and F-score (0.9593) suggest comparatively reduced consistency relative to FGGO and PSO.

The DDTO + DeepGBM model records an accuracy of 0.9621, with sensitivity and specificity values of 0.9535 and 0.9683. Although the adaptive exploration–exploitation behavior of DDTO yields notable performance gains, its F-score of 0.9551 indicates a marginally lower balance between false positives and false negatives compared to the higher-ranked optimizers.

The MVO + DeepGBM configuration achieves an accuracy of 0.9578, with sensitivity of 0.9478 and specificity of 0.9651. While MVO-based optimization still substantially improves DeepGBM performance relative to the baseline, its F-score of 0.9498 is the lowest among the optimized models, reflecting comparatively reduced predictive balance.

A comparative ranking of optimization algorithms based on overall performance reveals a clear hierarchy among the evaluated metaheuristic strategies. FGGO consistently ranks first across all evaluation metrics, followed by PSO, GWO, DDTO, and MVO. This ranking is not limited to a single metric but is observed uniformly across accuracy, sensitivity, specificity, predictive values, and F-score, indicating stable and reliable optimization behavior.

The superior performance of FGGO-based optimization can be attributed to its fitness-driven adaptive search mechanism, which effectively balances global exploration and local exploitation throughout the optimization process. The consistently high values achieved across all metrics suggest that FGGO not only identifies optimal hyperparameter configurations but also promotes robust generalization and stable learning dynamics. In contrast, while other optimizers demonstrate strong performance, their results exhibit slightly greater variability across metrics, indicating differences in convergence consistency.

Overall, the comparative analysis confirms that FGGO provides the most robust and consistent optimization strategy for DeepGBM within the proposed framework. Its ability

to achieve superior performance across all evaluation metrics underscores its suitability for optimizing deep learning models in complex clinical prediction tasks, where reliability, stability, and balanced decision-making are of paramount importance.

Evaluating classification models using multiple performance metrics simultaneously is essential for obtaining a holistic understanding of their predictive capabilities, particularly in medical decision-support systems where trade-offs between metrics may exist. Radar plots offer an effective visualization framework for comparing several metrics at once, enabling intuitive assessment of overall model balance and dominance patterns. Figure 6 illustrates a radar plot summarizing the performance of different hybrid optimization–classification models across six key evaluation criteria, namely accuracy, sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), negative predictive value (NPV), and F-score.

By projecting each metric onto a common radial scale, the figure facilitates direct visual comparison of the strengths and weaknesses of each model configuration. Larger and more uniform polygonal areas indicate models with consistently strong performance across all evaluation dimensions, thereby supporting informed selection of robust and clinically reliable predictive models.

Analyzing the variation of performance metrics across different model configurations is essential for understanding relative strengths and performance trends among competing approaches. When multiple hybrid models are evaluated using the same classifier, visualizing metric progression across models helps reveal systematic improvements or degradations attributable to the underlying optimization strategies. Figure 7 presents time-series–style line plots for key evaluation metrics, including accuracy, sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), negative predictive value (NPV), and F-score, across the investigated hybrid optimization–DeepGBM models.

Although the horizontal axis represents discrete model configurations rather than temporal evolution, the line-based visualization effectively highlights comparative trends and relative ordering of model performance.

The line-based representation highlights trade-offs among models, thereby supporting informed selection of robust hybrid frameworks for heart disease prediction.

A detailed comparative analysis of model performance metrics is essential for identifying strengths, weaknesses, and overall robustness of competing hybrid learning frameworks. Heatmaps provide an effective visualization tool for simultaneously examining multiple evaluation metrics across several models, as they enable rapid identification of performance patterns through color intensity and numerical annotations. Figure 8 presents an annotated heatmap summarizing the performance of hybrid optimization–DeepGBM models with respect to accuracy, sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), negative predictive value (NPV), and F-score.

The color gradients and embedded metric values facilitate precise comparison across models, highlighting both dominant configurations and subtle performance variations. Such visual

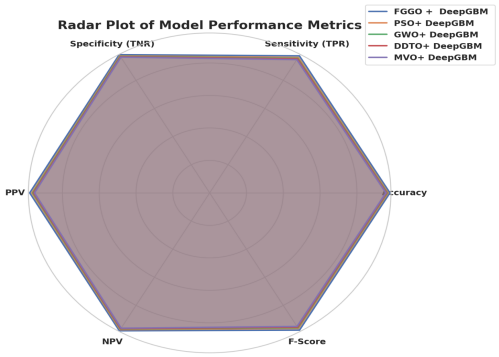

Figure 6. Radar plot comparing the performance of hybrid optimization–DeepGBM models across multiple evaluation metrics, including accuracy, sensitivity (TPR), specificity (TNR), positive predictive value (PPV), negative predictive value (NPV), and F-score.

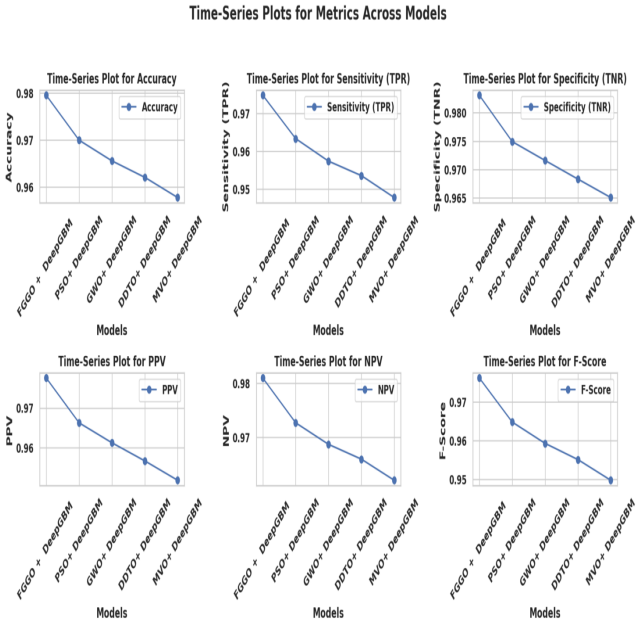

Figure 7. Time-series–style line plots illustrating the variation of performance metrics across hybrid optimization–DeepGBM models, including accuracy, sensitivity (TPR), specificity (TNR), positive predictive value (PPV), negative predictive value (NPV), and F-score.

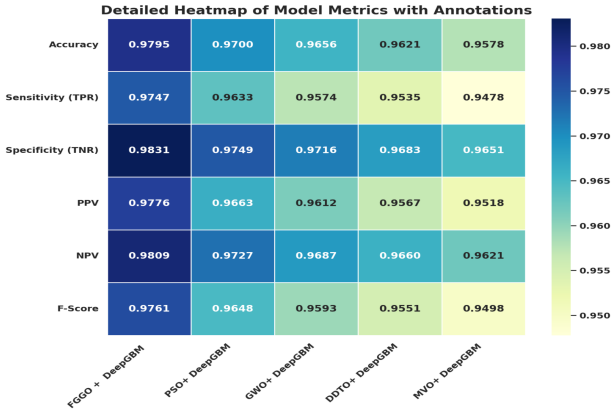

Figure 8. Annotated heatmap illustrating the performance of hybrid optimization–DeepGBM models across multiple evaluation metrics, including accuracy, sensitivity (TPR), specificity (TNR), positive predictive value (PPV), negative predictive value (NPV), and F-score. Color intensity and numerical annotations indicate the relative magnitude of each metric.

analysis supports informed interpretation of trade-offs among metrics and aids in selecting models that achieve balanced and clinically reliable performance in heart disease prediction tasks.

5. CONCLUSION AND FUTURE WORK

This study presented a comprehensive and integrated analytical framework for heart disease patient analysis that combines unsupervised clustering, deep learning–based classification, and metaheuristic optimization. The central objective was to address the challenges posed by patient heterogeneity, nonlinear clinical feature interactions, and hyperparameter sensitivity in deep learning models. Through the incorporation of unsupervised clustering, the framework enabled the discovery of latent patient subgroups based on clinical similarity, providing a structured representation of the patient population that supports more informed learning and interpretation.

A key finding of this work is the substantial improvement achieved through the application of Fitness Greylag Goose Optimization (FGGO) to the Deep Learning Framework Distilled by Gradient Boosting Decision Trees (DeepGBM). The optimization process systematically enhanced model performance across all evaluation metrics by effectively tuning hyperparameters and stabilizing the learning process.

Compared to the baseline DeepGBM configuration, the FGGO-optimized model demonstrated markedly improved predictive accuracy, sensitivity, specificity, and balanced performance as reflected by the F-score. These improvements confirm that the joint use of clustering-driven data representation and clinical prediction tasks.

The findings of this study have important implications for clinical practice and the development of intelligent healthcare systems. The use of patient clustering provides clinicians with a data-driven mechanism to identify subgroups of patients who share similar demographic, physiological, and diagnostic characteristics. Such stratification can support treatment personalization by enabling physicians to draw insights from the outcomes and responses of clinically similar patients, thereby moving beyond population-level averages toward more individualized care strategies.

Moreover, the optimized predictive framework proposed in this study is well suited for integration into clinical decision support systems. By combining robust deep learning classification with automated optimization, the framework can deliver reliable and consistent predictions that assist clinicians in risk assessment and diagnostic decision-making. The high predictive reliability achieved through FGGO optimization enhances trust in model outputs, which is a critical factor for adoption in real-world clinical environments. As a result, the proposed approach has the potential to contribute to improved diagnostic accuracy, more efficient resource allocation, and enhanced patient outcomes.

While the results of this study are promising, several avenues for future research can further extend and strengthen the proposed framework. One important direction involves the exploration of advanced clustering techniques that can capture dynamic and evolving patient profiles. Such techniques may enable longitudinal patient monitoring and more refined subgroup discovery, particularly in settings where patient characteristics change over time.

Another key area for future work is the investigation of real-time clinical deployment considerations.

This includes assessing computational efficiency, scalability, and system integration challenges associated with deploying FGGO-optimized deep learning models in hospital information systems or wearable health monitoring platforms. Addressing these practical considerations is essential for translating research findings into operational clinical tools.

Additionally, the proposed framework can be extended to other cardiovascular conditions and chronic diseases, such as diabetes, hypertension, and respiratory disorders, where patient heterogeneity and complex feature interactions similarly affect predictive modeling. Evaluating the generalizability of the framework across diverse medical domains would further validate its robustness and applicability.

Finally, future research may focus on integrating the proposed approach with smart healthcare infrastructures and AI-driven monitoring systems. By combining optimized predictive models with real-time data streams from electronic health records, wearable sensors, and Internet of Medical Things devices, the framework could support continuous risk assessment and proactive clinical intervention, thereby contributing to the advancement of intelligent and adaptive healthcare ecosystems.

Data Availability The dataset used in this study is publicly available on Kaggle athttps: //www.kaggle.com/datasets/kingabzpro/heart-disease-patients.

Declarations •Acknowledgments Not applicable.

•Conflict of interest/Competing interests The authors declare that they have no conflicts of interest to report regarding the present study.

•Ethics approval and consent to participate Not applicable.

•Consent for publication Not applicable.

•Funding No Fund

REFERENCES

[1] B. Almadani, H. Kaisar, I. R. Thoker, and F. Aliyu, “A systematic survey of distributed decision support systems in healthcare”, Systems, vol. 13, no. 3, 2025, issn: 2079-8954.doi: 10.3390/ systems 13030157. [Online]. Available: https://www.mdpi.com/2079-8954/13/3/157.

[2] X. Gao, P. He, Y. Zhou, and X. Qin, “Artificial intelligence applications in smart healthcare: A survey”, Future Internet, vol. 16, no. 9, 2024, issn: 1999-5903.doi: 10.3390/fi 16090308. [Online]. Available: https://www.mdpi.com/1999-5903/16/9/308.

[3] G. Santangelo et al., “The global burden of valvular heart disease: From clinical epidemiology to management”, Journal of Clinical Medicine, vol. 12, no. 6, 2023, issn: 2077-0383.doi: 10.3390/jcm 12062178. [Online]. Available: https://www.mdpi.com/2077-0383/12/6/2178.

[4] J. M. Bastos, B. Colaço, R. Baptista, C. Gavina, and R. Vitorino, “Innovations in heart failure management: The role of cutting-edge biomarkers and multiomics integration”, Journal of Molecular and Cellu-

lar Cardiology Plus, vol. 11, p. 100290, 2025, issn: 2772-9761.doi: https:// doi.org/10.1016/j. jmccpl. 2025. 100290. [Online]. Available: https:// www. sciencedirect.com/science/article/pii/S 2772976125000091.

[5] O. T, ica and O. T, ica, "Molecular diagnostics in heart failure: From biomarkers to personalized medicine", Diagnostics, vol. 15, no. 14, 2025, issn: 2075-4418.doi: 10.3390/diagnostics 15141807. [Online]. Available: https://www.mdpi.com/2075-4418/15/14/1807.

[6] G. Lyu, "Data-driven decision making in patient management: A systematic review, "BMC Medical Informatics and Decision Making, vol. 25, no. 1, p. 239, 2025. doi: https://doi.org/ 10.1186/s 12911-025-03072-x.

[7] L. Abualigah et al., "Artificial intelligence-driven translational medicine: A machine learning framework for predicting disease outcomes and optimizing patient-centric care", Journal of Translational Medicine, vol. 23, no. 1, p. 302, 2025. doi: https://doi.org/10.1186/s 12967- 025-06308-6.

[8] F. Epelde, "Heterogeneity in heart failure with preserved ejection fraction: A systematic review of phenotypic classifications and clinical implications", Journal of Clinical Medicine, vol. 14, no. 14, 2025, issn: 2077-0383.doi: 10.3390/jcm 14144820. [Online]. Available: https://www. mdpi.com/2077-0383/14/14/4820.

[9] G. Guglielmi, S. Moscatelli, G. Rocchetti, P. Agostoni, M. Chessa, and M. Mapelli, "Cardiopulmonary exercise testing in congenital heart disease: A never-ending story from paediatrics to adult life", Children, vol. 12, no. 9, 2025, issn: 2227-9067.doi: 10.3390/children 12091175. [Online]. Available: https://www.mdpi.com/2227-9067/12/9/1175.

[10] J.-S. Hulot et al., "Heart failure improvement, remission, and recovery: A european journal of heart failure expert consensus document", European Journal of Heart Failure, vol. 27, no. 10, pp. 1807–1819, 2025. doi: https://doi.org/10.1002/ejhf.3732.

[11] A. Alsayat et al., "Enhancing cardiac diagnostics: A deep learning ensemble approach for precise ecg image classification", Journal of Big Data, vol. 12, no. 1, p. 7, 2025. doi: https://doi.org/10.1186/s 40537-025-01070-4.

[12] D.-I. Kasartzian and T. Tsiampalis, "Transforming cardiovascular risk prediction: A review of machine learning and artificial intelligence innovations", Life, vol. 15, no. 1, 2025, issn: 2075-1729.doi: 10.3390/life 15010094. [Online]. Available: https://www.mdpi.com/2075- 1729/15/1/94.

[13] C. Ding, T. Yao, C. Wu, and J. Ni, "Advances in deep learning for personalized ecg diagnostics: A systematic review addressing inter-patient variability and generalization constraints", Biosensors and Bioelectronics, vol. 271, p. 117073, 2025, issn: 0956-5663.doi: https://doi.org/10.1016/j.bios.2024.117073. [Online]. Available: https://www.sciencedirect.com/science/ article/pii/S0956566324010807.

[14] S. M. Srinivasan and V. Sharma, "Applications of ai in cardiovascular disease detection—a review of the specific ways in which ai is being used to detect and diagnose cardiovascular diseases, "AI in Disease Detection: Advancements and Applications, pp. 123–146, 2025. doi: https://doi.org/10.1002/9781394278695.

[15] M. B. Abubaker and B. Babayiğit, "Detection of cardiovascular diseases in ecg images using machine learning and deep learning methods", IEEE Transactions on Arti-ficial Intelligence, vol. 4, no. 2, pp. 373–382, 2023. doi: 10.1109/ TAI.2022.3159505.

[16] S. Subramani et al., "Cardiovascular diseases prediction by machine learning incorporation with deep learning", Frontiers in medicine, vol. 10, p. 1150933, 2023. doi: https://doi.org/10.3389/fmed.2023.1150933.

[17] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, "Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models", European Journal of Medical Research, vol. 29, no. 1, p. 455, 2024. doi: https://doi.org/10.1186/s 40001-024- 02044-7.

[18] H. Lu et al., "Research progress of machine learning and deep learning in intelligent diag-nosis of the coronary atherosclerotic heart dis-ease", Computational and Mathematical Methods in Medicine, vol. 2022, no. 1, p. 3016532, 2022. doi: https:// doi.org/10.1155/2022/3016532.

[19] K. Vayadande et al., "Heart disease prediction using ma-chine learning and deep learning algorithms, " in 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022, pp. 393–401.doi: 10.1109/CISES 54857.2022.9844406.

[20] S. Matin Malakouti, "Heart disease classification based on ecg using machine learning models, " Biomedical Signal Processing and Control, vol. 84, p. 104796, 2023, issn: 1746-8094.doi: https:// doi.org/10.1016/j.bspc.2023.104796. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S1746809423002288.

[21] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques", Algorithms, vol. 16, no. 2, 2023, issn: 1999-4893.doi: 10.3390/ a16020088. [Online]. Avail-able: https://www.mdpi.com/1999-4893/16/2/88.

[22] S. Mall and J. Singh, "Heart diagnosis using deep neural network, " in 2023 International Conference on Computational Intelligence and Knowledge Econ-omy (ICCIKE), 2023, pp. 7–12. doi: 10.1109/ICCIKE 58312.2023.10131696.

[23] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improv-ing prediction of human heart disease using ma-chine learning algorithms", Mobile Information Sys-tems, vol. 2022, no. 1, p. 1410169, 2022. doi: https:// doi.org/10.1155/2022/1410169.

[24] N. A. Baghdadi, S. M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis", Journal of Big Data, vol. 10, no. 1, p. 144, 2023. doi: https://doi.org/10.1186/s 40537-023-00817-1.

[25] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine learning-based predictive models for detection of cardiovascular diseases", Diagnostics, vol. 14, no. 2, 2024, issn: 2075-4418.doi: 10.3390/diagnostics 14020144. [Online]. Available: https://www.mdpi.com/ 2075-4418/14/2/144.

[26] M. N. Hasan, M. A. Hossain, and M. A. Rahman, "An ensemble based lightweight deep learning model for the prediction of cardiovascular diseases from electrocardiogram images, " Engineering Applications of Artificial Intelligence, vol. 141, p. 109782, 2025, issn: 0952-1976. doi: https:// doi.org/10.1016/j.engappai. 2024. 109782. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624019419.

[27] S. Wang, J. Hu, Y. Du, X. Yuan, Z. Xie, and P. Liang, "Wcformer: An interpretable deep learning framework for heart sound signal analysis and automated diagnosis of cardiovascular diseases", Expert Systems with Applications, vol. 276, p. 127238, 2025, issn: 0957-4174.doi: https:// doi.org/10.1016/j.eswa. 2025. 127238. [Online]. Available: https:// www. sciencedirect.com/science/article/pii/S0957417425008607.

[28] S. Dhanka and S. Maini, "A hybridization of xgboost machine learning model by optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets", International Journal of Cardiology, vol. 420, p. 132757, 2025, issn: 0167-5273.doi: https:// doi.org/10.1016/j.ijcard. 2024. 132757. [Online]. Available: https:// www. sciencedirect.com/science/article/pii/S0167527324013792.

[29] M. Dehghani Saryazdi and A. Mostafaeipour, "Identification and validation of key predictive factors for heart attack diagnosis using machine learning and fuzzy clustering", Engineering Applications of Artificial Intelligence, vol. 142, p. 109968, 2025, issn: 0952-1976.doi: https:// doi.org/10.1016/j.engappai. 2024. 109968. [Online]. Available: https:// www. sciencedirect.com/science/article/pii/S0952197624021274.

[30] N. Groun et al., "A novel data augmentation tool for enhancing machine learning classification: A new application of the higher order dynamic mode decomposition for improved cardiac disease identification", Results in Engineering, vol. 25, p. 104143, 2025, issn: 2590-1230.doi: https:// doi.org/10.1016/j. rineng. 2025. 104143. [Online]. Available: https:// www. sciencedirect.com/science/article/pii/S 2590123025002312.