SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

MLP Model Optimization for Heart Attack Risk Prediction: A Systematic Literature Review

Heru Supriyanto[1)*], Taqwa Hariguna[2)], Azhari Shouni Barkah[3)]
[1)2)]Master of Computer Science, Universitas Amikom Purwokerto, Indonesia, [3)]Information Technology, Universitas Amikom Purwokerto, Indonesia
[1)]herusupriyanto04@amikompurwokerto.ac.id, [2)]taqwa@amikompurwokerto.ac.id, [3)]azhari@amikompurwokerto.ac.id

Submitted :Jul 12, 2025 | **Accepted** : Aug 8, 2025 | **Published** : Aug 11, 2025

Abstract: Heart disease is a leading cause of global mortality, making accurate predictive models a clinical priority. While Multilayer Perceptron (MLP) models offer significant potential, their application is hindered by challenges in optimization, data imbalance, and interpretability. This systematic literature review aims to synthesize current research on MLP model optimization for heart disease prediction, focusing on strategies for handling class imbalance and achieving model transparency with SHapley Additive exPlanations (SHAP). Following PRISMA guidelines, a structured search was conducted across major scientific databases, including Google Scholar and IEEE Xplore. Using PICOS criteria for study selection, 30 peer-reviewed articles were chosen for in-depth analysis. The synthesis reveals two key findings: first, a clear trend towards sophisticated optimization using complex ensemble models and automated hyperparameter tuning; and second, the establishment of the Synthetic Minority Over-sampling Technique (SMOTE) as the dominant data-level solution for class imbalance. Furthermore, the application of SHAP successfully validated models by consistently identifying known clinical risk factors like age and chest pain type. This review concludes that while progress is significant, critical gaps remain. Practically, these findings can guide practitioners in developing more robust models. We recommend that future research prioritize direct comparative studies of imbalance techniques and focus on validating models with large-scale, real-world clinical data to ensure the development of trustworthy and generalizable predictive systems.

Keywords: Class Imbalance; Heart Disease Prediction; Multilayer Perceptron; SHAP; Systematic Literature Review

INTRODUCTION

Heart disease remains the foremost global health challenge, consistently ranking as the leading cause of death worldwide, which underscores the urgent need for effective predictive strategies to reduce mortality and improve patient outcomes (Ahmed, 2024; Ahsan, 2022; Katarya, 2020). In response, machine learning (ML), particularly the Multilayer Perceptron (MLP), has emerged as a crucial technology. By processing complex medical datasets, MLP models can identify hidden patterns, enabling more personalized risk assessments and serving as powerful clinical decision support tools that help transform healthcare from a reactive to a proactive system (Bhowmik, 2024; Rojek, 2024; Saboor, 2022; Wang, 2021). The development of these models is critical, as traditional diagnostic methods are often costly and may fail to capture the subtle, non-linear interactions between clinical and lifestyle factors, leading to delayed or inaccurate predictions with severe health consequences (Mehmood, 2021; Subramani, 2023).

Despite the promise of ML, two persistent challenges hinder the development of reliable predictive models. First is the critical need for high accuracy, as even minor errors can have significant consequences. Second is the technical hurdle of class imbalance, a common characteristic of medical datasets where healthy individuals (the majority class) vastly outnumber those with heart disease (the minority class). This skew causes models to develop a bias towards the majority class, undermining their primary purpose of identifying high-risk patients (Ahsan, 2022; Li, 2020). Furthermore, the "black-box" nature of many complex models raises concerns about

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

interpretability, which is essential for gaining clinical trust and ensuring fairness (Wahid et al., 2025). Therefore, building effective models requires a dual focus on optimizing predictive performance while addressing these fundamental data and transparency issues.

The objective of this systematic literature review (SLR) is to systematically identify, synthesize, and evaluate the current literature on the use of MLP models for heart disease prediction. While numerous studies have applied these models, they frequently address critical challenges such as class imbalance or model interpretability in isolation. This fragmented approach has created a research gap: the knowledge landscape lacks a cohesive understanding of the combined effects and trade-offs of different optimization, imbalance-handling, and interpretation strategies when used together.

Therefore, the primary contribution (novelty) of this work is to provide the first comprehensive, integrated analysis of three critical pillars: MLP optimization, class imbalance techniques, and SHAP-based interpretability. This leads us to re-emphasize the significant and clearly defined research gap that this paper aims to address: a systematic synthesis that cohesively analyzes the interplay between these three elements is notably absent from the literature. A holistic understanding of how optimization strategies affect models trained on imbalanced data, and how those models can be reliably interpreted, is crucial for building clinically viable systems. By systematically mapping the existing research, this paper offers a clear roadmap that identifies best practices and defines a concrete agenda for future research. To guide this investigation, this review seeks to answer the following research questions:

RQ1: What optimization techniques for MLP models have been reported in the literature for predicting heart disease risk?

RQ2: What methods for addressing class imbalance in heart disease datasets have been evaluated in studies using MLP models, and what effects have been reported on performance metrics?

RQ3: How has SHAP (SHapley Additive exPlanations) been applied in the literature to interpret MLP-based heart disease models, and what key risk factors have been identified through its use?

METHOD

This study uses the Systematic Literature Review (SLR) methodology to comprehensively identify, evaluate, and synthesize existing research on MLP model optimization for heart disease prediction, with a specific focus on addressing class imbalance and model interpretability with SHAP. This review process is designed to be transparent, thorough, and replicable, adhering to established guidelines for conducting SLRs.

Search Strategy

The primary search was conducted across five leading scientific databases to ensure broad coverage of relevant literature: Google Scholar, IEEE Xplore, PubMed, ScienceDirect, and the ACM Digital Library. To maintain the relevance and currency of the findings, the search was limited to peer-reviewed articles published between January 2015 and June 2025. The search strategy was built based on a formal search string constructed from four key conceptual blocks of keywords: a) Model ("Multilayer Perceptron" OR "MLP" OR "Neural Network"); b) Problem Domain ("heart disease" OR "cardiac risk" OR "myocardial infarction"); c) Technical Challenge ("class imbalance" OR "unbalanced data" OR "SMOTE"); and d) Interpretability: ("SHAP" OR "explainable AI" OR "interpretable ML"). These blocks were combined using the "AND" operator, while terms within each block were combined with "OR" to maximize recall. This structured query was adapted as necessary to meet the specific syntax requirements of each database.

Inclusion and Exclusion Criteria

To ensure the relevance and quality of the selected literature, a set of strict inclusion and exclusion criteria was established using the PICOS framework (Population, Intervention, Comparison, Outcomes, Study Design). This framework provides a structured approach to defining the scope of the review. Detailed criteria are outlined in Table 1.

Table 1. Inclusion and Exclusion Criteria

PICOS Component	Inclusion Criteria	Exclusion Criteria
Population	Studies that use datasets to predict heart disease risk in human patients.	Studies focusing on non-human subjects or diseases other than heart disease.
Intervention	Studies that must apply a Multilayer Perceptron (MLP) model as a primary predictive tool.	Studies using other machine learning models without MLP, or theoretical papers without implementation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

PICOS Component	Inclusion Criteria	Exclusion Criteria
Comparison	Studies that ideally compare model performance before and after applying class imbalance techniques, or provide a SHAP analysis.	Studies that do not provide a basis for comparison or evaluation of model performance.
Outcomes	Studies must report model performance metrics (e.g., Accuracy, Precision, Recall, F1-Score, AUC) and/or feature importance from SHAP.	Studies that do not report quantifiable performance outcomes.
Study Design	Peer-reviewed journal articles, full conference papers, and workshop proceedings.	Grey literature (e.g., theses, patents), book chapters, abstracts, editorials, and non-English articles.

Study Selection Process

The study selection followed a multi-stage screening process guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method. The entire process is visually summarized in the PRISMA flow diagram presented in Figure 1. To ensure objectivity and minimize potential bias, the selection was conducted by two independent reviewers who were not part of the author team. The process began with the Identification phase, where an initial search in the specified databases yielded a total of 405 records (279 from Google Scholar and 126 from Scopus). Before screening, 55 duplicate records were identified and removed, leaving 350 unique records.

In the Screening phase, these 350 records were first assessed by title and abstract. During this step, 233 records were excluded as they were clearly not relevant to the research questions. This left 117 reports that were sought for retrieval for a full-text review. Of these, 34 reports could not be retrieved (e.g., no accessible full-text PDF), resulting in 83 reports that were assessed for eligibility. During the full-text eligibility assessment, a further 53 reports were excluded for specific reasons: 25 studies did not use an MLP model as their primary predictive tool, and 28 studies did not address the core concepts of class imbalance or interpretability as required by the PICOS criteria. Finally, in the Included phase, the remaining 30 studies that met all inclusion and exclusion criteria were included in the final qualitative synthesis. Any disagreements between the two reviewers during the screening stages were resolved through a consensus discussion. If a consensus could not be reached, a third independent reviewer was consulted to arbitrate and make the final decision. This rigorous, multi-stage process ensures the final selection is both comprehensive and highly relevant to the research objectives.

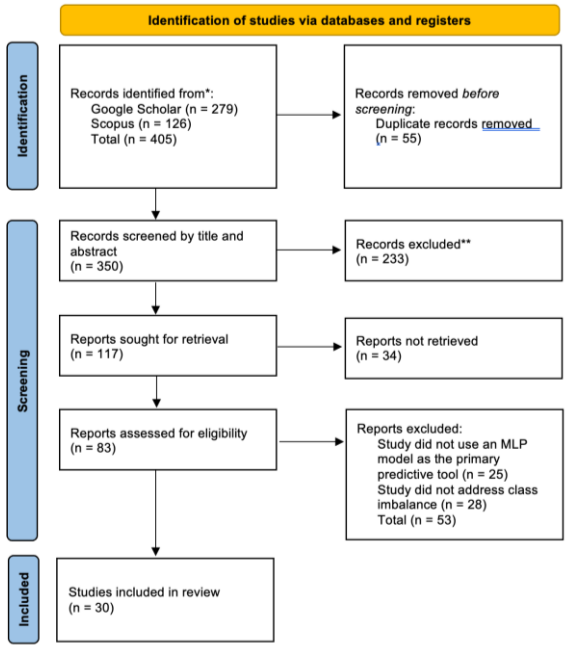

Figure 1. PRISMA Flow Diagram

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Quality Assessment

Following the final selection, a quality assessment was performed on all 30 included studies to evaluate their methodological rigor and the reliability of their findings. Each paper was assessed against a predefined set of quality criteria adapted from established SLR guidelines. The criteria included:

a. Clarity of Objectives: Does the study have clear and well-defined research questions or objectives?
b. Appropriateness of Methodology: Is the methodology (including data source, preprocessing, and model selection) appropriate and clearly described?
c. Validity of Results: Are the results presented clearly, and are the evaluation metrics appropriate for the study's objective?
d. Contribution and Conclusions: Are the conclusions supported by the results, and does the study make a clear contribution to the field?

Each study was rated on these criteria to ensure that the findings synthesized in this review are based on methodologically sound research.

Data Extraction and Synthesis

Following final selection and quality assessment, a structured data extraction form was used to systematically collect relevant information from each included study. This standardized approach ensured consistency and enabled effective comparison and synthesis of findings. The final step of this methodology is the synthesis of the extracted data. A narrative synthesis approach was used to integrate and summarize the findings, which are organized thematically based on the three research questions guiding this SLR.

RESULT

This section presents the findings from the systematic literature review. It begins with an overview of the study selection process, followed by a detailed, thematic synthesis of the findings aligned with the research questions, incorporating focused data tables for clarity.

Overview of Selected Scientific Articles

Initial literature searches confirmed that both heart attack prediction and model interpretability are highly active areas of research. A broad search for “Heart Attack Risk Prediction” on Google Scholar from 2016 to 2025 yielded many works, with an analysis of the top 200 papers showing over 119,500 citations, a high average citation rate per year (13,278). This indicates a mature and consistently impactful field of study over the past decade. In parallel, the query for “SHAP Interpretability in Machine Learning” shows a more recent but explosive growth trend. Although the first relevant papers appeared around 2019, the 200 papers analyzed in this domain have accumulated over 25,000 citations, indicating significant research interest and a rapidly forming core body of literature.

Following this initial scope determination, a formal study selection process was conducted as described in the methodology. The search across all initially defined databases yielded 350 records. After removing duplicates, the titles and abstracts of these articles were screened for relevance. During this initial screening, articles not focused on machine learning or heart disease were excluded, reducing the number of potentially relevant articles to 83. The full texts of these 83 articles were then retrieved and comprehensively assessed for eligibility based on the PICOS criteria. From this set, 53 studies were excluded for reasons such as using non-MLP models, not addressing class imbalance or interpretability, or failing to report performance results. Finally, a final set of 30 studies met all inclusion criteria and were included in the qualitative synthesis.

Synthesis of Findings for RQ1: MLP Model Optimization

Before delving into the specific optimization techniques for MLP models, it is useful to understand the broader landscape of methods employed across the 30 reviewed studies. Figure 2 provides a visual summary of the frequency of different machine learning models and optimization strategies, highlighting the prevalence of tree-based models and standard optimization practices.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

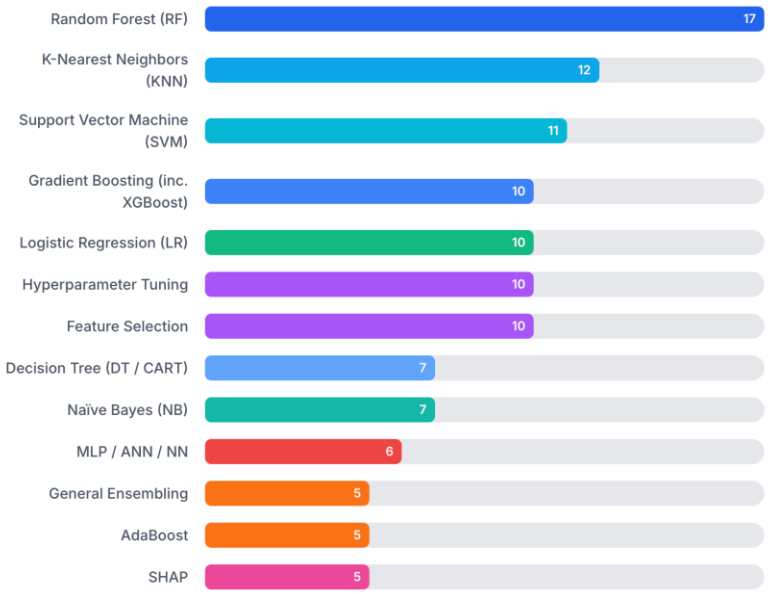

Figure 2. Frequency of Different Models and Optimizations

An analysis of the methods used across the 30 selected studies reveals a landscape dominated by well-established machine learning algorithms, with tree-based models being particularly prevalent. As shown in the graph of method distribution, Random Forest (RF) is the most frequently employed model, appearing in 17 studies. Its popularity likely stems from its inherent robustness, its strong performance on tabular data without extensive tuning, and its ability to handle non-linear relationships effectively. This is followed by other traditional models like K-Nearest Neighbors (KNN) (12 studies) and Support Vector Machines (SVM) (11 studies). Interestingly, despite being a primary focus of this review, MLP and its variants (ANN/NN) appear less frequently (6 studies). This suggests that while neural networks are a key area of research, their higher complexity, greater need for hyperparameter tuning, and "black-box" nature lead many researchers to still rely on the proven performance and relative simplicity of traditional ensemble and instance-based learners for this prediction task.

A critical finding is that common optimization techniques are applied broadly across these popular models to elevate their performance beyond baseline levels. Hyperparameter Tuning and Feature Selection are the most common optimization strategies, each appearing in 10 studies. Their prevalence indicates a widespread understanding that off-the-shelf models are insufficient for clinical-grade accuracy. The influence of these methods on the results is direct and significant: systematic tuning, for example through GridSearchCV as used by Bhatt et al (2023) and Saboor et al (2022), directly impacts performance by identifying the optimal model configuration, leading to higher accuracy. Similarly, feature selection improves results by reducing noise and model complexity, which can increase accuracy and decrease computational cost, as demonstrated by Ahmad et al. (2022). The near-universal use of k-fold cross-validation (Ali et al., 2021; Rani et al., 2021) further ensures that these optimized results are generalizable and not simply an artifact of overfitting.

While less frequent, the studies that do focus on MLP and other neural networks often explore more specialized and novel optimization strategies. This is likely because standard gradient-based optimization can be more challenging to configure for NNs and may be prone to local minima. Researchers therefore turn to advanced methods to unlock the full potential of these models. For instance, Nandy et al. (2023) and Patro et al. (2021) introduced bio-inspired approaches like Swarm Optimization and the Salp Swarm Algorithm to more effectively explore the vast solution space of network weights and biases. Furthermore, Y. Li et al. (2020) designed a custom "P-S loss" function, which directly influences the results by forcing the model to learn more effectively from the specific challenges of imbalanced ECG data. This suggests that while the broader field relies on tuning established models, a dedicated subset of research is focused on creating highly tailored and architecturally innovative neural network solutions to achieve state-of-the-art performance.

Table 2 summarizes the findings related to RQ1: What optimization techniques for Multilayer Perceptron (MLP) models have been reported in the literature for predicting heart disease risk? It includes studies that applied specific optimization methods such as hyperparameter tuning, feature selection, advanced model architectures like ensembles, and robust validation protocols.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Summary of Model Optimization and Validation Techniques

Author, Year	Model(s)	Key Optimization Technique(s)	Validation Method
Ahmad et al., 2022	RF, DT, GBC, SVM, etc.	Sequential Feature Selection (SFS)	5-fold CV
Ahmed & Husien, 2024	Ensemble/Hybrid Models	Literature Review on Ensembling (Voting, AdaBoost)	N/A (Review)
Ali et al., 2021	MLP, KNN, RF, DT, etc.	Standard Training	10-fold CV
Bhatt et al., 2023	MLP, RF, XGBoost, DT	GridSearchCV	k-fold CV
Chandrasekhar & Peddakrishna, 2023	RF, KNN, LR, etc.	GridSearchCV, Soft Voting Ensemble (SVE)	5-fold CV
El-Hasnony et al., 2022	Label Ranking Classifier	Grid Search	10-fold CV
Katarya & Meena, 2021	MLP, ANN, DNN, RF, etc.	Comparative Analysis	N/A (Comparative)
Y. Li et al., 2020	Custom Ensemble NN	Custom Loss Function (P-S Loss), DDAG Ensemble	5-fold CV
J. Li et al., 2022	XGBoost, LR, RF, SVM	LASSO Feature Selection	70/30 Split
X. Li et al., 2023	RF, AdaBoost, SVM, etc.	Select K Best Feature Selection	80/20 Split
Lu et al., 2022	XGBoost	GridSearchCV	70/20/10 Split
Mehmood et al., 2021	Custom CNN	LASSO Feature Selection	30/70 Split
Nandy et al., 2023	Custom 3-Layer ANN	Swarm Optimization	200/103 Split
Nowfal et al., 2025	Ensemble (XGB Meta)	Stacking Ensemble	N/A
Patro et al., 2021	NN, SVM, NB, KNN	Bayesian Optimization, Salp Swarm Algorithm	80/20 Split
Rajendran & Karthi, 2022	Ensemble (NB+LR)	Entropy-based Feature Engineering (EFE)	10-fold CV
Rani et al., 2021	RF, SVM, Adaboost, etc.	Hybrid Feature Selection (GA+RFE)	10-fold CV
Rath et al., 2022	Ensemble (SOM-AE)	Deep Learning Ensemble	70/30 Split
Rojek et al., 2024	LR, RF, KNN, etc.	RandomizedSearchCV, GridSearchCV	80/20 Split
Saboor et al., 2022	SVM, AdaBoost, RF, etc.	GridSearchCV, Data Standardization	10-fold CV
Somantri & Wanti, 2024	NB, GLM	Modified Genetic Algorithm (M-GA)	10-fold CV
Subramani et al., 2023	Stacking Ensemble (inc. MLP)	Optuna Framework, GBDT+SHAP Selection	80/20 Split
Taher & Abdulazeez, 2023	Literature Review	Review of Feature Selection (Lasso, Relief)	N/A (Review)
Wang et al., 2021	MLP, XGBoost, SVM, etc.	RFE Feature Selection, Grid Search	5-fold CV

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Synthesis of Findings for RQ2: Handling Class Imbalance

The challenge of class imbalance was identified as a central theme and significant obstacle in most of the reviewed papers. The most prevalently used strategy to mitigate this issue is the Synthetic Minority Over-sampling Technique (SMOTE), a data-level solution. Its popularity likely stems from its straightforward implementation and its direct, intuitive approach to solving data scarcity. The primary influence of SMOTE on the results is a significant improvement in the model's sensitivity and recall for the high-risk patient class. By generating synthetic instances, it provides the model with more minority class examples to learn from, directly addressing the bias caused by the skewed data distribution. This effect was consistently demonstrated in studies like Ali et al. (2021), where its application was a critical step in achieving high performance, and Rani et al. (2021), who noted improvements across multiple performance metrics.

Furthering this data-level approach, some studies explored more sophisticated over-sampling techniques. For example, Wang et al. (2021) combined SMOTE with Edited Nearest Neighbors (SMOTE+ENN). This hybrid technique improves upon standard SMOTE by not only creating synthetic samples but also "cleaning" the data space by removing majority class samples that are close to the decision boundary, which can lead to a more robust and well-defined training set. The widespread adoption of SMOTE and its variants, as also confirmed in the systematic review by Ahsan & Siddique (2022), establishes it as the de facto standard data-level solution in the literature, valued for its accessibility and immediate impact on minority class detection.

Although SMOTE is the most common method, the reviewed literature indicates that a one-size-fits-all solution is insufficient, prompting researchers to explore innovative algorithm-level techniques. These methods are less common, likely due to their higher implementation complexity and the need for a deeper understanding of the underlying learning algorithm. Unlike SMOTE, which alters the data, these techniques influence the results by modifying the model's learning process itself. For instance, Mehmood et al. (2021) adjusted the class weight ratio during training, which forces the model to pay a higher penalty for misclassifying the minority class. Similarly, Y. Li et al. (2020) designed a custom loss function to compel their model to focus on harder-to-classify minority examples. This approach can be highly effective and avoids the potential risk of introducing noise or artificial patterns that can sometimes occur with synthetic data generation.

Table 3 summarizes the findings related to RQ2: What methods for handling class imbalance in heart disease datasets have been evaluated in studies using MLP models, and what are their reported effects on performance metrics? It includes studies that explicitly applied or analyzed techniques to manage imbalanced data.

Table 3. Summary of Class Imbalance Techniques and Their Reported Effects

Author, Year	Imbalance Technique Used	Reported Effect on Performance
Ahsan & Siddique, 2022	SLR Finding: SMOTE is most used; algorithm-level methods emerging.	N/A (Review paper summarizing trends).
Ali et al., 2021	SMOTE (Data-Level)	Enabled models (KNN, DT, RF) to achieve 100% accuracy, sensitivity, and specificity.
Y. Li et al., 2020	Algorithm-Level (Custom Loss, DDAG structure)	Achieved high average sensitivity (89.25%) on a severely imbalanced dataset.
Mehmood et al., 2021	Algorithm-Level (Class Weighting)	Contributed to the CNN model achieving 97% accuracy for binary classification.
Rani et al., 2021	SMOTE (Data-Level)	Improved performance across all classifiers on metrics like accuracy, sensitivity, and F-Measure.
Wang et al., 2021	SMOTE+ENN (Data-Level Hybrid)	Used to create a robustly balanced training set for the MLP and XGBoost models.

Synthesis of Findings for RQ3: SHAP in Model Interpretation

A growing and impactful subset of the reviewed literature leverages SHapley Additive exPlanations (SHAP) to move beyond predictive accuracy and provide model transparency. The popularity of this approach stems from its crucial role in fostering clinical trust and utility. These studies consistently use SHAP to deconstruct "black-box" models and identify the most influential features driving their predictions. A significant finding from this synthesis is the strong convergence of key clinical risk factors identified across various models and datasets. For example, demographic factors such as age are universally recognized as highly influential, cited as primary predictors of mortality or disease presence by J. Li et al. (2022), X. Li et al. (2023), and Wang et al. (2021). The influence of this finding on the results is profound; by confirming that models are learning from well-established clinical factors, SHAP provides essential validation that is a prerequisite for any real-world clinical adoption.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Beyond demographics, SHAP analysis consistently highlights the importance of specific clinical and physiological markers. In a study on heart failure mortality, J. Li et al. (2022) found that average blood urea nitrogen (BUN) was the single most important predictor. Similarly, Wang et al. (2021) used SHAP to highlight that NT-proBNP and NYHA class are the primary predictors of 3-year mortality. Research by Subramani et al. (2023) further confirmed the consistency of SHAP, noting that its feature importance results aligned with those from a GBDT model and highlighted ST Slope as the primary predictor. This finding resonates with the study by Ozcan & Peker (2023), whose more transparent CART model also identified ST Slope as the most important feature. This cross-validation of feature importance across different model types and interpretability methods strengthens confidence in these clinical indicators and explains why SHAP is a popular choice: it provides a robust, model-agnostic way to verify that a complex model is learning patterns consistent with simpler, more trusted models.

Symptomatic and categorical features are also consistently identified as critical drivers of model prediction. Subramani et al. (2023) highlights Chest Pain Type as one of the most important features identified by SHAP. This is highly consistent with the findings of Yazdani et al. (2021), who, using a weighted associative rule mining (WARM) approach, identified asymptomatic chest pain as a key component in all high-confidence rules for predicting heart disease. The consistent identification of this feature by disparate methods like SHAP and associative rule mining underscores its fundamental importance. This influences the results by demonstrating that these advanced models are capable of quantifying the impact of subjective, patient-reported symptoms, effectively bridging the gap between qualitative clinical assessment and quantitative risk prediction.

Table 4. Key Predictive Features Identified by SHAP and Other Interpretability Methods

Author, Year	Interpretability Method	Top Identified Predictive Features
Bhowmik et al., 2024	Feature Importance	Age, cholesterol, blood pressure, gender, smoking
J. Li et al., 2022	SHAP	Blood Urea Nitrogen (BUN), age, systolic BP
X. Li et al., 2023	SHAP	Heavy Metals (Cesium, Lead), age, gender
Lu et al., 2022	SHAP	Gender, Blood Pressure (Systolic & Diastolic)
Nowfal et al., 2025	Associativity Analysis	Chest Pain, Heart Rate, Slope Rate
Ozcan & Peker, 2023	CART (Rule-Based)	ST Slope, Oldpeak, Chest Pain Type
Rojek et al., 2024	Feature Importance	Heart rate, age, BMI, cholesterol
Subramani et al., 2023	SHAP	ST Slope, Chest Pain Type
Wang et al., 2021	SHAP	Age, NT-proBNP, NYHA class
Yazdani et al., 2021	WARM (Rule-Based)	Asymptomatic chest pain, exercise-induced angina

DISCUSSIONS

This section analyzes the principal findings synthesized from the 30 reviewed studies, offers a critical evaluation of the current state of the literature, identifies key research gaps that have emerged from the analysis, and proposes concrete directions for future research.

Analysis of Principal Findings

The synthesis of the reviewed literature reveals several significant trends in the application of machine learning for heart disease prediction. Firstly, the field is clearly maturing beyond the use of simple, out-of-the-box algorithms towards more complex and optimized solutions. This is evidenced by the widespread adoption of sophisticated ensemble methods, such as the stacking model proposed by Subramani (2023) and the soft voting ensemble by Chandrasekhar & Peddakrishna (2023). This trend suggests a collective understanding that single models are often insufficient to capture the intricate, non-linear relationships present in cardiovascular data. Furthermore, the consistent use of automated hyperparameter tuning frameworks like GridSearchCV (Bhatt et al., 2023; Saboor et al., 2022) indicates a move towards more rigorous and reproducible research practices aimed at maximizing predictive power.

Secondly, there is a clear and widespread acknowledgment of class imbalance as a fundamental challenge. The overwhelming prevalence of the Synthetic Minority Over-sampling Technique (SMOTE) (Ali et al., 2021; Rani

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

et al., 2021; Wang et al., 2021) establishes it as the de facto standard data-level solution. The implication is that SMOTE is perceived as a reliable and effective method for improving model sensitivity to the under-represented high-risk class. However, the emergence of more advanced techniques, such as the algorithm-level class weighting used by Mehmood (2021), suggests a growing recognition that simply altering the data distribution may not always be the optimal approach.

Finally, the application of SHAP for model interpretability marks a pivotal shift from a singular focus on accuracy to a dual focus on performance and trustworthiness. The most powerful implication from the SHAP-based findings is the strong alignment between the models' learned feature importance and established clinical knowledge. The consistent identification of risk factors such as age (Li et al., 2022; Wang et al., 2021) and chest pain type (Subramani et al., 2023) provides crucial validation that these complex models are learning clinically relevant patterns. From a practical medical standpoint, this is a critical finding. It implies that clinicians can begin to trust these AI systems as diagnostic aids, as the models' reasoning aligns with established medical science. For example, a model that flags a patient based on factors like age and ST Slope can be more readily accepted by a doctor, potentially leading to faster and more confident decision-making in a clinical setting.

Critical Evaluation of the Literature

Despite the advancements, a critical evaluation of the 30 reviewed studies reveals several methodological limitations. The most significant of these is an over-reliance on a small number of publicly available datasets, particularly the UCI Cleveland dataset (Ali et al., 2021; Jindal et al., 2021). While valuable for benchmarking, models trained on these datasets may lack generalizability to diverse, real-world clinical populations. Another prevalent issue is the inconsistency in reporting and the potential for inflated performance metrics, with some studies claiming near-perfect accuracy (Ali et al., 2021; Ahmad et al., 2022), which is unlikely to be reproducible in a clinical environment.

When comparing this review to other SLRs in a similar domain, our findings both align and offer unique contributions. For instance, our conclusion regarding the prevalence of SMOTE and the challenge of interpretability aligns with the broader findings of Ahsan & Siddique (2022). However, their review covered a wider range of ML models and heart conditions. Our SLR provides a more focused and deeper analysis specifically on the optimization of MLP models and the emerging, critical role of SHAP for interpretability, a combination not previously synthesized in detail. This focused scope allows for a more granular identification of research gaps specific to the application of neural networks in this domain.

Furthermore, it is important to acknowledge the potential for publication bias in the body of literature reviewed. There is a well-known tendency in scientific publishing to favor studies that report positive results or high accuracy scores. This bias may lead to an overestimation of the true effectiveness of certain models, as studies with negative or less impressive results may be less likely to be published and, therefore, were not included in our review. Additionally, our search was limited to articles published in English, which may have excluded relevant research from non-English speaking regions. These factors represent inherent limitations in the available literature that should be considered when interpreting the findings of this review.

Research Gaps Identified from the Review

Based on the explicit findings of this review, several critical research gaps have been identified, organized according to the research questions that guided this SLR.

a. Pertaining to RQ1 (MLP Model Optimization): A primary gap is the lack of standardized architectures and reporting for MLP models, making it difficult to replicate findings or establish a baseline for what constitutes an optimal architecture.
b. Pertaining to RQ2 (Handling Class Imbalance): There is a significant gap in the comparative analysis of imbalance techniques specifically for MLP models. It remains unclear whether it is more effective to modify the data (e.g., with SMOTE) or the algorithm itself.
c. Pertaining to RQ3 (SHAP Interpretability): The application of SHAP is predominantly limited to generating global feature importance rankings. There is a substantial, unexplored opportunity to use SHAP for more granular, local-level analysis, such as investigating why a model misclassifies specific patients.

Addressing the Gaps and Future Directions

To advance the field, future research should be strategically directed at addressing the identified gaps. Future work should focus on establishing standardized benchmarking for MLP models and conducting head-to-head comparative studies of imbalance techniques. The application of XAI must also mature from confirmation to exploration, leveraging SHAP for in-depth error analysis and local-level interpretation. Finally, the field must prioritize generalizability by shifting towards large-scale, real-world EHR data and prospective cohorts for both training and, critically, external validation.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

CONCLUSION

This systematic literature review synthesized research on MLP model optimization for heart disease prediction, focusing on the dual challenges of class imbalance and interpretability. Key findings reveal a field moving toward advanced optimization through ensemble methods and automated tuning, with SMOTE being the standard for class imbalance. The application of SHAP has proven effective in validating models against established clinical knowledge. The main contribution of this review is a comprehensive roadmap highlighting both progress and critical gaps, though the study is limited by potential publication bias and its focus on English-language literature. Based on these findings, we recommend that future research prioritize rigorous comparative studies of imbalance techniques, deeper local-level analysis using explainable AI, and, most importantly, robust validation on large-scale, real-world clinical data. Embracing these directions is essential for developing the next generation of predictive systems that are not only accurate but also trustworthy and clinically valuable.

REFERENCES

Ahmad, G. N., Ullah, S., Algethami, A., Fatima, H., & Akhter, S. Md. H. (2022). Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection. *IEEE Access*, *10*, 23808–23828. https://doi.org/10.1109/ACCESS.2022.3153047

Ahmed, M., & Husien, I. (2024). Heart Disease Prediction Using Hybrid Machine Learning: A Brief Review. *Journal of Robotics and Control (JRC)*, *5*(3), Article 3. https://doi.org/10.18196/jrc.v5i3.21606

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289. https://doi.org/10.1016/j.artmed.2022.102289

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2), Article 2. https://doi.org/10.3390/a16020088

Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M. R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies*, *4*(2), Article 2. https://doi.org/10.32996/bjns.2024.4.2.5

Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4), Article 4. https://doi.org/10.3390/pr11041210

El-Hasnony, I. M., Elzeki, O. M., Alshehri, A., & Salem, H. (2022). Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. *Sensors*, *22*(3), Article 3. https://doi.org/10.3390/s22031184

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012072. https://doi.org/10.1088/1757-899X/1022/1/012072

Katarya, R., & Meena, S. K. (2021). Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. *Health and Technology*, *11*(1), 87–97. https://doi.org/10.1007/s12553-020-00505-7

Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart Disease Prediction using Hybrid machine Learning Model. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 1329–1333. https://doi.org/10.1109/ICICT50816.2021.9358597

Li, J., Liu, S., Hu, Y., Zhu, L., Mao, Y., & Liu, J. (2022). Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study. *Journal of Medical Internet Research*, *24*(8), e38082. https://doi.org/10.2196/38082

Li, X., Zhao, Y., Zhang, D., Kuang, L., Huang, H., Chen, W., Fu, X., Wu, Y., Li, T., Zhang, J., Yuan, L., Hu, H., Liu, Y., Zhang, M., Hu, F., Sun, X., & Hu, D. (2023). Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: Findings of the US NHANES from 2003 to 2018. *Chemosphere*, *311*, 137039. https://doi.org/10.1016/j.chemosphere.2022.137039

Li, Y., He, Z., Wang, H., Li, B., Li, F., Gao, Y., & Ye, X. (2020). CraftNet: A deep learning ensemble to diagnose cardiovascular diseases. *Biomedical Signal Processing and Control*, *62*, 102091. https://doi.org/10.1016/j.bspc.2020.102091

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15027
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Lu, S., Chen, R., Wei, W., Belovsky, M., & Lu, X. (2022). Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions. *AMIA Annual Symposium Proceedings*, *2021*, 813–822.

M. R., S., Devasthali, S., Mishra, S., Saha, S., Jain, N., & Manjunath, T. C. (2024, June 5). *A Machine Learning-based Predictive Analytics System for Enhancing Cardiovascular Health using Heart-Sage Concepts. | EBSCOhost*. https://openurl.ebsco.com/contentitem/gcd:181690671?sid=ebsco:plink:crawler&id=ebsco:gcd:181690671

Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T., & Masood, M. (2021). Prediction of Heart Disease Using Deep Convolutional Neural Networks. *Arabian Journal for Science and Engineering*, *46*(4), 3409–3422. https://doi.org/10.1007/s13369-020-05105-1

Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications*, *35*(20), 14723–14737. https://doi.org/10.1007/s00521-021-06124-1

Nowfal, S. H., Sengan, S., G, J. S. D., Bhatta, S., V, S., & B, V. (2025). The Diagnosis of Heart Attacks: Ensemble Models of Data and Accurate Risk Factor Analysis Based on Machine Learning. *Journal of Machine and Computing*, 589–599. https://doi.org/10.53759/7669/jmc202505046

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*, 100130. https://doi.org/10.1016/j.health.2022.100130

Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*, *26*, 100696. https://doi.org/10.1016/j.imu.2021.100696

Rajendran, R., & Karthi, A. (2022). Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers. *Expert Systems with Applications*, *207*, 117882. https://doi.org/10.1016/j.eswa.2022.117882

Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263–275. https://doi.org/10.1007/s40860-021-00133-6

Rath, A., Mishra, D., Panda, G., Satapathy, S. C., & Xia, K. (2022). Improved heart disease detection from ECG signal using deep learning based ensemble model. *Sustainable Computing: Informatics and Systems*, *35*, 100732. https://doi.org/10.1016/j.suscom.2022.100732

Rojek, I., Kotlarz, P., Kozielski, M., Jagodziński, M., & Królikowski, Z. (2024). Development of AI-Based Prediction of Heart Attack Risk as an Element of Preventive Medicine. *Electronics*, *13*(2), Article 2. https://doi.org/10.3390/electronics13020272

Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, *2022*(1), 1410169. https://doi.org/10.1155/2022/1410169

Somantri, O., & Wanti, L. P. (2024). A proposed model using Naïve Bayes and generalized linear models for early detection of heart attack risk. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(2), Article 2. https://doi.org/10.11591/ijeecs.v33.i2.pp1169-1176

Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1150933

Taher, H. A., & Abdulazeez, A. M. (2023). Machine Learning Approaches for Heart Disease Detection: A Comprehensive Review. *International Journal of Research and Applied Technology (INJURATECH)*, *3*(2), Article 2. https://doi.org/10.34010/injuratech.v3i2.12052

Wahid, A. M., Hariguna, T., & Karyono, G. (2025). Optimization of Recommender Systems for Image-Based Website Themes Using Transfer Learning. *Journal of Applied Data Sciences*, *6*(2), Article 2. https://doi.org/10.47738/jads.v6i2.671

Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Liu, Y., Han, Q., & Zhang, Y. (2021). Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Computers in Biology and Medicine*, *137*, 104813. https://doi.org/10.1016/j.compbiomed.2021.104813

*name of corresponding author

Yazdani, A., Varathan, K. D., Chiam, Y. K., Malik, A. W., & Wan Ahmad, W. A. (2021). A novel approach for heart disease prediction using strength scores with significant predictors. *BMC Medical Informatics and Decision Making*, *21*(1), 194. https://doi.org/10.1186/s12911-021-01527-5

*name of corresponding author