Comparative Analysis of XGBoost, KNN, and SVM Algorithms for Heart Disease Prediction Using SMOTE-Tomek Balancing

Yuliana1)*, **Robet**2), **Leony Hoki**3)
1,2,3)Department of Informatics, STMIK Time, Medan, Indonesia
1)*yullianaa466@gmail.com, 2)robertdetime@gmail.com, 3)leony.hoki@gmail.com,

Submitted :Oct 12, 2025 | **Accepted** : Nov 10, 2025 | **Published** : Jan 02, 2026

Abstract: Heart disease remains one of the leading causes of death worldwide, making early detection crucial for improving patient outcomes. This study aims to evaluate and compare the performance of several machine learning algorithms in detecting heart disease using the 2015 BRFSS dataset, which includes responses from 253,680 individuals. The three algorithms examined are Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The data preprocessing steps involved feature encoding, class imbalance handling using the Synthetic Minority Over-sampling Technique combined with Tomek Links (SMOTE-Tomek), and hyperparameter tuning through RandomizedSearchCV. The models were assessed on a hold-out validation set using several metrics, including accuracy, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), F1-score, precision, and recall. The results demonstrated that XGBoost achieved the highest performance, with an accuracy of 94%, a ROC-AUC score of 0.98, and an F1-score of 0.94. In comparison, KNN achieved an accuracy of 87% (ROC-AUC 0.95), while SVM attained an accuracy of 79% (ROC-AUC 0.86). These findings suggest that XGBoost is a robust model for large-scale heart disease classification and holds potential for implementation in clinical decision support systems.

Keywords: Heart Disease; XGBoost; K-Nearest Neighbors; Support Vector Machine; SMOTE-Tomek

INTRODUCTION

Heart disease is a general term used to describe disorders that affect the heart's function. There are various classifications of heart diseases, including cardiovascular disease, coronary heart disease, and heart attacks (Natsir, Bakti, & Wahyuni, 2024). According to data from the World Health Organization (WHO), cardiovascular diseases, including heart diseases, accounted for approximately 32% of global deaths in 2019 (World Health Organization, 2023). The high global incidence rate, coupled with the difficulty of early prediction, highlights the urgent need for technological approaches to assist medical practitioners. Rapid and accurate diagnostic systems are crucial for preventing serious complications and improving treatment outcomes for high-risk patients (Vahlevy, Zendrato, Fadillah, & Sidiq, 2023).

The advancement of information technology has opened new opportunities in the healthcare sector, particularly in the process of digital disease diagnosis. One increasingly popular approach is the use of machine learning algorithms to develop data-driven classification systems. Machine learning is capable of processing historical data into models that can identify pattern or make predictions about patient conditions. Various well-known algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) have proven effective in classification tasks, including heart disease classification. Supervised learning serves as the primary approach in this classification process, where algorithms are trained using labeled data to recognize relationship between input features and diagnostic outcomes (Hidayat et al., 2024).

Previous studies have shown that XGBoost outperforms other boosting algorithms such as Adaptive Boosting, achieving higher accuracy and ROC-AUC values. This advantage is attributed to its ability to handle overfitting and its computational efficiency, making it effective in heart disease classification tasks (Sah, Niesa, Jafar, & Muharrom, 2025). K-Nearest Neighbors (KNN) is considered effective for classifying data based on proximity, especially when dealing with nonlinear data patterns. Meanwhile, Support Vector Machine (SVM) performs well

*name of corresponding author

in clearly separating classes with an optimal margin, even for complex data (Arif, Siregar, Faisal, & Juwita, 2024). However, a significant gap remains in the existing literature: prior work has often lacked consistent imbalance handling (such as SMOTE-Tomek) and failed to report a full suite of evaluation metrics, particularly both ROC-AUC and Precision-Recall AUC, which is crucial for imbalanced data.

To address this gap, this study provides a fair comparison of three leading models (XGBoost, KNN, and SVM) using a reproducible pipeline that includes consistent data balancing and hyperparameter tuning. The main contributions of this study are: (1) a comparative performance analysis on the large-scale BRFSS 2015 dataset; (2) the application of a complete and reproducible pipeline, including SMOTE-Tomek for balancing and RandomizedSearchCV for tuning; (3) a fair evaluation using comprehensive metrics suitable for imbalanced data, including both ROC-AUC and PR-AUC; and (4) an analysis of the feature importance to identify key risk factors.

LITERATURE REVIEW

Heart disease consists of a series of disorders that affect the heart. This includes vascular problems such as irregular heartbeat, weakened heart muscles, congenital heart defects, cardiovascular diseases, and coronary artery disease (Derisma, 2020). Evaluation based on accurate classification of heart failure risk can greatly assist patients in preventing severe heart attacks and increasing survival rates. One effective way to identify and classify heart disease is by utilizing machine learning algorithms (Nugraha, 2021). The K-Nearest Neighbors (KNN) algorithm is generally used to classify objects based on training data with the smallest difference values and the closest distance between objects (Maskuri, Sukerti, & Herdian Bhakti, 2022). This algorithm is considered both simple and effective for large datasets. The method belongs to the supervised learning category, which means the dataset requires target labels (Yogianto, Homaidi, & Fatah, 2024).

The Support Vector Machine (SVM) is a machine learning algorithm that identifies the optimal separating function to distinguish between classes, and it is also a component of the supervised learning approach (Adi & Wintarti, 2022). The process is carried out by selecting the optimal hyperplane and maximizing the margin between data points from different classes. In addition to classification tasks, SVM can also be applied to regression to predict continuous values (Maskuri et al., 2022). Extreme Gradient Boosting (XGBoost) can effectively handle complex and diverse medical data, meeting the requirements for timeliness and higher diagnostic accuracy (Surono, Fadli, Purwamti, & Susanto, 2025). This algorithm is a machine learning system based on tree boosting that can be optimized to build larger-scale decision trees. As an extension of gradient boosting, XGBoost is an ensemble-based decision tree method designed to accelerate processing time, even for large dataset (Sausan, Pratiwi, & Mufidah, 2024).

Previous studies on sentiment analysis of customer reviews using SVM combined with the SMOTE-Tomek Links resampling technique have shown that this approach effectively addresses the problem of data imbalance (Sumantiawan, Suseno, & Syafei, 2023). SMOTE-Tomek functions to improve the representation of minority classes while reducing noise caused by overlapping data among classes (Shabrina Assyifa & Luthfiarta, 2024). This technique contributes to creating a cleaner and more balanced dataset, while XGBoost is known for its robustness in handling complex and large-scale data (Ratantja Kusumajati, Rahmat, & Junaidi, 2024). Model evaluation on an imbalanced dataset relies not only on accuracy but also on metrics such as precision, recall, F1-score, and ROC-AUC, which better capture performance on minority classes. Other research indicates that the application of SMOTE-Tomek combined with Random Forest in diabetes classification significantly increases sensitivity and F1-score compared to models without balancing (Hairani, Anggrawan, & Priyanto, 2023).

The implementation of similar algorithms has resulted in higher AUC values in multi-class imbalance data (Sukamto, Prameswary, Royadi, & Sofia, 2025). RandomizedSearchCV in the SVM algorithm significantly improves model accuracy compared to models without tuning (Yennimar, Rasid, & Kenedy, 2023). RandomizedSearchCV can achieve nearly the same performance as GridSearchCV but with shorter training time, as demonstrated in stunting classification cases using XGBoost (Pramudhyta & Rohman, 2024).

Table 1. Comparison of Related Work on Heart Disease Classification

Research (Year)	Algorithm(s)	Dataset Size	Imbalance Handling	Focus
(Sah et al., 2025)	Boosting	270	None	Comparison of different boosting methods.
(Yogianto et al., 2024)	KNN	303	SMOTE	Optimized KNN for smaller datasets.

*name of corresponding author

This Study (2025)	**XGBoost, KNN, SVM**	**229,781**	**SMOTE-Tomek**	**Large-scale evaluation, combined SMOTE-Tomek, full stability metrics.**

METHOD

This research was conducted through a series of systematic stages designed to develop an accurate model for classifying heart disease. The research process was divided into six main phases, starting from problem identification to model performance evaluation and conclusion formulation.

Research Flowchart

The stages in this study were arranged based on a logical sequence illustrated through the flowchart in fig 1. The diagram provides an overview of the order and relationships among the main components in the research process, starting from data collection to the experiment and ensures that the analysis process is caried out systematically and consistently.

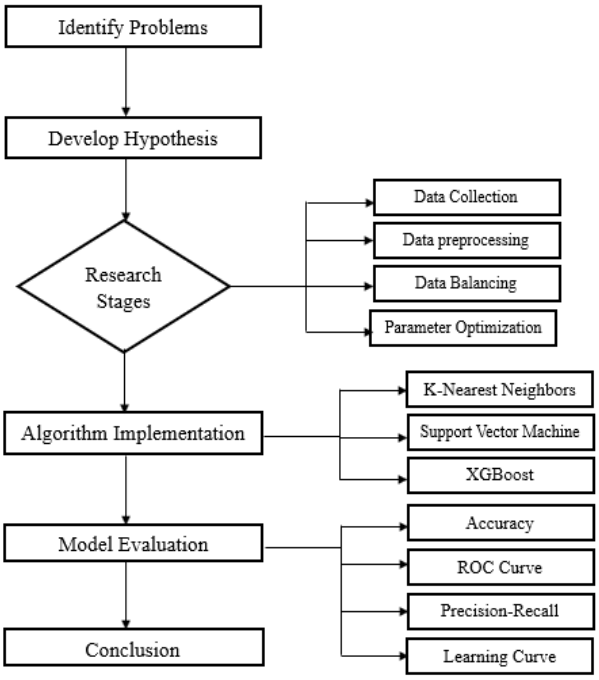

Fig 1. Research Flowchart

Data Collection

The dataset used in this study was obtained from the Kaggle platform (Centers for Disease Control and Prevention, 2017), originally sourced from the CDC BRFSS 2015 survey. The dataset is publicly available under the CC0: Public Domain license. This dataset consists of 253,680 data entries with 22 attributes, encompassing demographic information, lifestyle factors, and health conditions. The target attribute is HeartDiseaseAttack, which contains two classes:

1 = having a history of heart disease
0 = not having a history of heart disease
The data distribution based on the class dataset partition is presented in Table 2

Table 2. Distribution of BRFSS 2015 Dataset by Data Type and Class

No.	Data Type	Negative Class (0)	Positive Class (1)	Total Data
1	Training Data	181,046	21,898	202,944
2	Testing Data	22,631	2,737	25,368
3	Validation Data	22,630	2,738	25,368
	Total	**226,307**	**27,373**	**253,680**

Data Preprocessing

The data preprocessing stage was carried out to prepare the dataset according to the requirements of machine learning algorithms. The steps performed include:

a. Duplicate Removal: Eliminating identical data entries.
b. Label Encoding: Converting categorical variables into numerical form using the label encoding technique.
c. Normalization: Not performed, as all features were already on a homogeneous numerical scale.

Feature Description and Standardization

The dataset contains 22 attributes. We selected 15 features for modeling based on their relevance and the results of our feature importance analysis (see Fig. 10). The target variable was HeartDiseaseorAttack. A complete description of the attributes used is provided in Table 3.

Table 3. Description of Attributes Used in the Model

Attribute	Description	Scale / Values
HeartDiseaseorAttack	(Target) Had heart disease or attack?	0 = No, 1 = Yes
Age	Age category	1 (18-24) to 13 (80+)
GenHlth	General health status	1 = Excellent to 5 = Poor
HighBP	High blood pressure	0 = No, 1 = Yes
HighChol	High cholesterol	0 = No, 1 = Yes
Sex	Gender	0 = Female, 1 = Male
Income	Income scale	1 = <$10k to 8 = >$75k
PhysHlth	Days physical health bad	1-30 days
BMI	Body Mass Index	12-98
Education	Education level	1 = None to 6 = College grad
Smoker	Smoked 100+ cigarettes?	0 = No, 1 = Yes
Stroke	(Ever) had a stroke?	0 = No, 1 = Yes
CholCheck	Cholesterol check in 5 yrs?	0 = No, 1 = Yes
MentHlth	Days mental health bad	1-30 days
DiffWalk	Difficulty walking?	0 = No, 1 = Yes
HvyAlcoholConsump	Heavy alcohol consumption?	0 = No, 1 = Yes

A key preprocessing step that was not performed was feature scaling (e.g., normalization or standardization). This decision was based on the fact that all 22 attributes in the original BRFSS 2015 dataset are categorical or ordinal (e.g., 'Age' is in 13 categories, 'Income' is in 8 categories, 'HighBP' is 0 or 1). Because all features already exist on a similar, discrete numerical scale, standardization (which is critical for distance-based algorithms like KNN and SVM) was not required and would not impact the model outcomes.

Data Balancing

The initial data distribution indicated an imbalance between positive and negative classes. Therefore, data balancing was performed using the SMOTE-Tomek technique, which combines two methods:

a. SMOTE (Synthetic Minority Over-sampling Technique): Generates synthetic samples for the minority class.
b. Tomek Links: Removes pairs of data from different classes that are too close to each other.

This combination aims to produce a balanced and noise-free dataset, allowing the model to perform classification more effectively.

Parameter Optimization

To improve model performance, hyperparameter tuning was conducted using the RandomizedSearchCV method. This exploratory approach randomly combines parameters within a defined search space and evaluates their performance using cross-validation. This approach was chosen because it is more efficient than exhaustive search methods (grid search), especially when dealing with large-scale dataset. Given the imbalanced nature of the dataset, F1-score was used as the primary scoring metric within RandomizedSearchCV to select the best-performing hyperparameters.

Model Selection and Training

Three machine learning algorithms were used in this study:

a. Extreme Gradient Boosting (XGBoost): A decision tree-based boosting algorithm that is both efficient and accurate.
b. Support Vector Machine (SVM): Seeks the optimal hyperplane to separate classes in a high-dimensional space.
c. K-Nearest Neighbors (KNN): Classifies data based on the proximity to several nearest neighbors.

Each model was trained using preprocessed and balanced data.

Model Evaluation

After the training process was completed, the models were evaluated using validation data to measure their ability to classify unseen data. The evaluation was conducted using classification metrics, namely accuracy, precision, recall, and F1-score, as well as the ROC-AUC score. These metrics were used to assess how effectively each model generated accurate predictions, particularly in the context of a dataset with imbalanced class distributions. Each metric was calculated based on the number of correct and incorrect predictions generated by the classification results, as shown in Equations (1) to (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (4)$$

Experimental Setup

The entire processing and modeling pipeline was developed using Python 3.9 within the Jupyter Notebook environment. Core tasks, including data handling, KNN, and SVM modeling, utilized primary libraries such as scikit-learn (for distance-based and kernel methods), XGBoost (for gradient boosting), and NumPy/Pandas (for data handling). Training was conducted on a high-performance system equipped with an Intel Core i5-1135G7 processor and 16 GB RAM to ensure computational efficiency and reproducibility.

RESULT

Classification Evaluation Technique

Model evaluation was conducted to determine the accuracy of each algorithm in classifying heart disease. In this study, the assessment was performed using stratified k-fold cross-validation, combined with SMOTE-Tomek, which was applied to the training data in each fold to address the imbalance between minority and majority classes. The evaluation did not rely solely on accuracy but also included other metrics such as precision, recall, F1-score, and ROC-AUC, since the original dataset exhibited class imbalance. In addition, a confusion matrix was used to analyze the distribution of prediction errors. At the same time, a learning curve was employed to evaluate the stability of the model during the training process.

Model Evaluation Results

Based on the training and evaluation process, a summary of the performance results of the three models is presented in Table 4.

Table 4. Model Evaluation Results with Stability Metrics (5-Fold-Cross-Validation)

Model	**Accuracy (Mean ± SD)**	**ROC-AUC (Mean ± SD)**	**F1-Score (Mean ± SD)**	**Precision (Mean ± SD)**	**Recall (Mean ± SD)**
XGBoost	**0.94**±0.0003	**0.98**±0.0001	**0.94**±0.0003	**0.99**±0.0001	0.90±0.0006
KNN	0.87±0.0007	0.95±0.0003	0.88±0.0005	0.80±0.0008	**0.98**±0.0004
SVM	0.79±0.0005	0.86±0.0010	0.80±0.0007	0.76±0.0016	0.84±0.0017

The stability of the models was explicitly evaluated through 5-fold cross-validation, with results presented in Table 3 as the Mean ± Standard Deviation (SD). All three algorithms exhibited high stability, indicated by the minimal SD across all metrics. XGBoost, in particular, showed the lowest variability (e.g., Accuracy: 0.94 ± 0.0003), confirming its robustness and reliability across different data partitions. This minimal variance validates that the model's high performance is not an artifact of data splitting.

Furthermore, due to the inherent class imbalance in the BRFSS dataset, per-class performance metrics are essential to determine the models' ability to correctly classify the minority class (heart disease patients). Table 5 presents the detailed Precision and Recall scores for both the Positive Class (1) and Negative Class (0) for all three algorithms.

Table 5. Per-Class Precision and Recall

Model	Metric	Negative Class (0)
XGBoost	Precision	0.91
	Recall	0.99
KNN	Precision	0.86
	Recall	0.76
SVM	Precision	0.79
	Recall	0.74

Evaluation Model Visualization

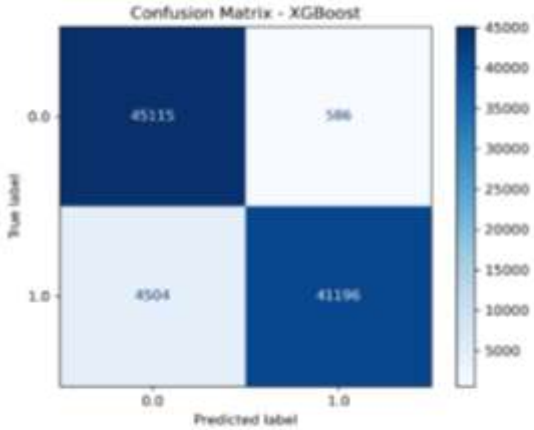

Fig 2. XGBoost Confusion Matrix, showing the distribution of correct and incorrect classification

The XGBoost model demonstrated excellent classification performance. The number of correct predictions for both the positive class (True Positive) and the negative class (True Negative) was relatively high, indicating its strong ability to identify patients with and without heart disease. Nevertheless, a number of False Negative cases still occurred, which could pose a serious risk if applied to an early detection system in real-world scenarios.

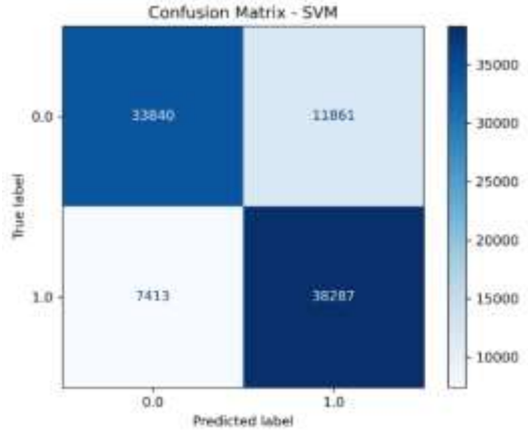

Fig 3. SVM Confusion Matrix, Showing the Distribution of Correct and Incorrect Classifications

In the SVM model, the distribution of predictions appeared relatively balanced however, the False Positive rate was slightly higher compared to XGBoost. This indicates that SVM tends to produce positive predictions for healthy individuals, leading to potential overdiagnosis.

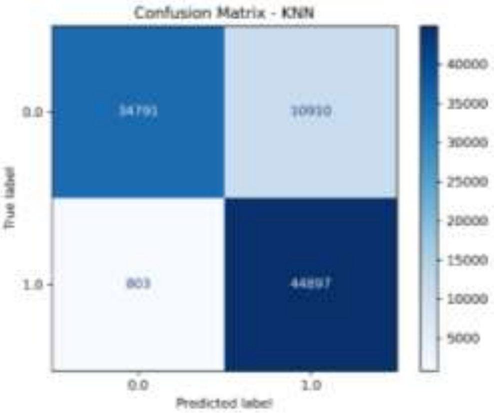

Fig 4. KNN Confusion Matrix, Showing the Distribution of Correct and Incorrect Classifications

The KNN model exhibited relatively lower performance. The number of False Positive and False Negative cases remained relatively high, indicating the limitations of this algorithm in handling complex datasets such as heart disease survey data. This could be attributed to the sensitivity of the KNN data scale and distribution.

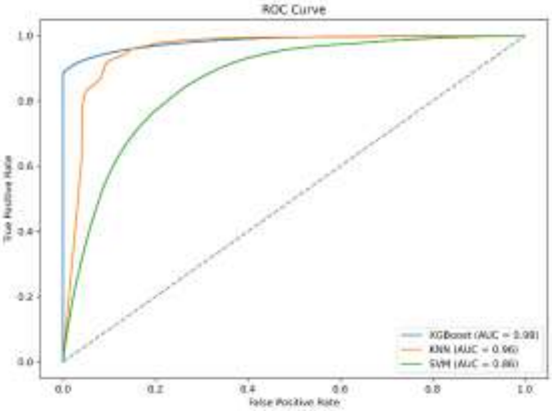

Fig 5. Comparative ROC Curves, illustrating the models' class-distinction ability (AUC)

The ROC curve illustrates the discriminative ability of each classification model on the dataset. The closer the curve approaches the upper-left corner of the graph, the better the model's performance. The XGBoost model achieved the best performance with an AUC value of 0.98, indicating an almost perfect ability to distinguish between positive and negative classes. The KNN model followed with an AUC value of 0.95, demonstrating relatively high performance. Meanwhile, the SVM model obtained an AUC value of 0.86, which although still considered good, was lower that the other two models. These differences indicate that XGBoost possesses the most balanced sensitivity and specificity in the context of heart disease classification.

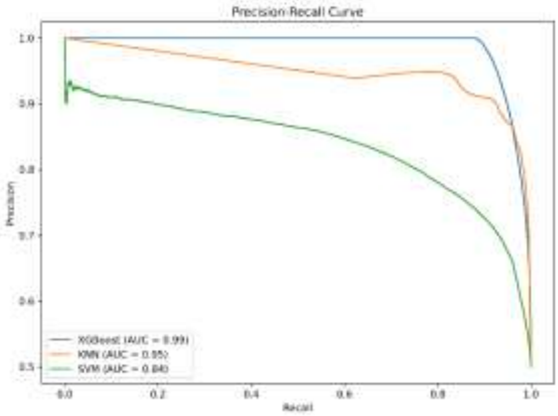

Fig 6. Comparative Precision-Recall Curves, Showing Model Performance on the Imbalanced Positive Class

The curve shown in fig 6 indicates that XGBoost achieved the best performance with an AUC value of 0.99, demonstrating a very high ability to maintain precision as recall increases. The KNN model exhibited fairly good performance with an AUC value of 0.95, while the SVM model lagged behind with an AUC of 0.84. These results show that XGBoost excels in identifying heart disease cases, complementing the findings from the ROC curve, particularly when prioritizing the detection of positive classes in an imbalanced dataset.

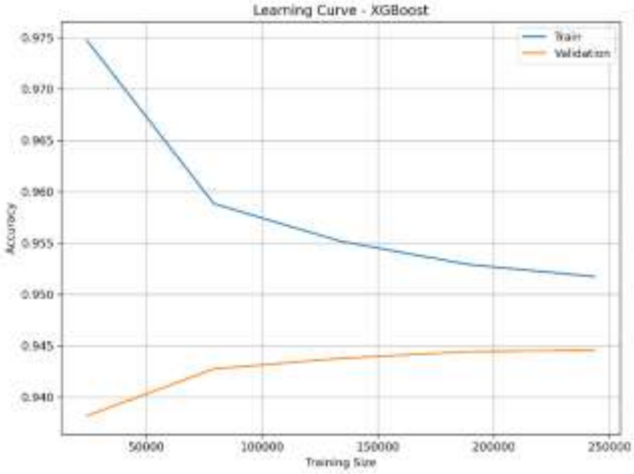

Fig 7. XGBoost Learning Curve, Indicating Model Stability (Train vs Validation)

The XGBoost learning curve shows stable performance on both training and validation data. The training accuracy slightly decreased as the amount of data increased, while the validation accuracy gradually and consistently improved. This indicates that the model possesses excellent generalization capability and does not experience overfitting, as the gap between training and validation accuracy continues to narrow.

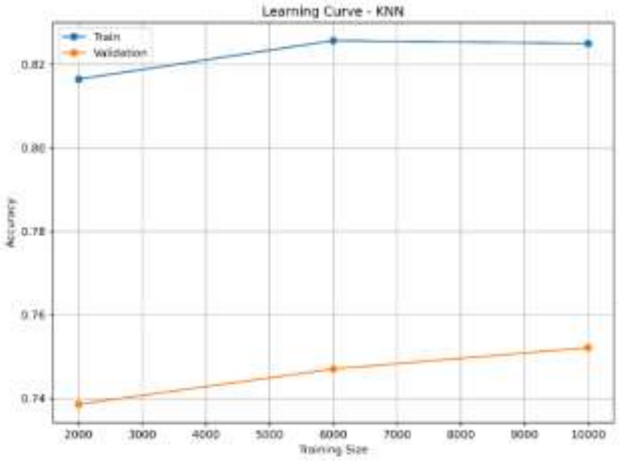

Fig 8. KNN Learning Curve, Indicating Model Stability (Train vs Validation)

The KNN model exhibited a stable pattern of increasing validation accuracy but tended to experience overfitting (High Variance), as indicated by a significant gap between the training and validation curves. This suggests that the model fits too closely to the training data but struggles to generalize on unseen data, which is typical for distance-based models on large datasets.

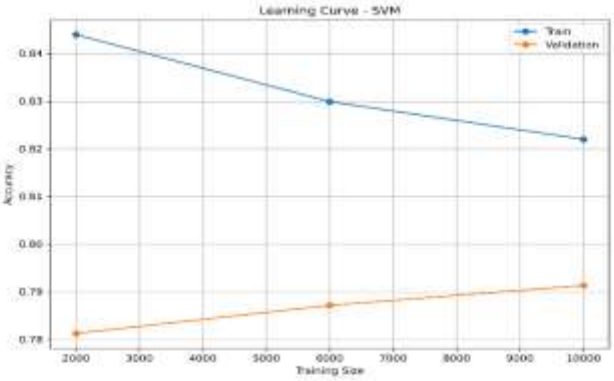

Fig 9. SVM Learning Curve, Indicating Model Stability (Train vs Validation)

In the SVM model, the learning curve indicates that as the amount of training data increases, the validation accuracy gradually improves. However, the large, persistent gap between the two curves suggests apparent symptoms of overfitting (High Variance). The model fits too closely to the complex training data but performs less optimally on new data, indicating poor generalization.

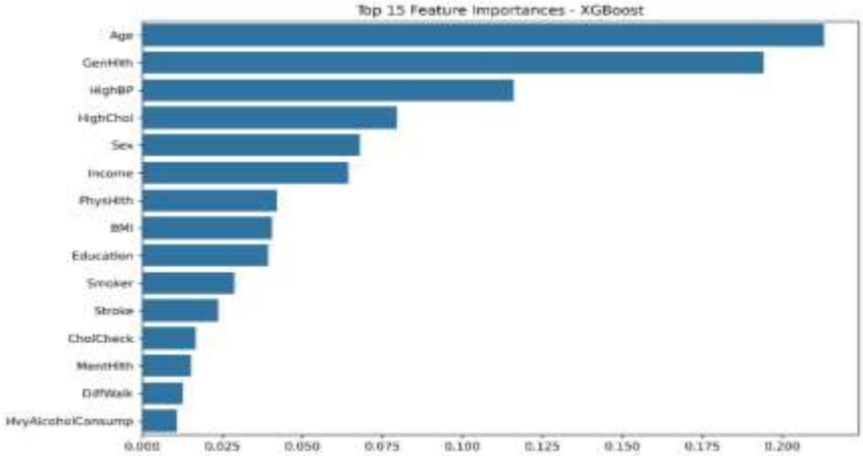

Fig 10. XGBoost Feature Importance, Showing the Top 15 Most Influential Features for Classification

The analysis results show that age (Age) is the most dominant factor in determining predictions, followed by general health condition (GenHlth) and high blood pressure (HighBP). These factors are widely recognized as key indicators of heart disease risk, and thus, the model's validation reinforces their medical relevance. In addition, attributes such as high cholesterol level (HighChol), gender (Sex), and income (Income), also contribute significantly. Interestingly, behavioral factors such as smoking (Smoker) and heavy alcohol consumption (HvyAlcoholConsum) were also recorded, although with lower influence, indicating that the model considers not only clinical factors but also individual lifestyle aspects.

DISCUSSIONS

This study evaluated three machine learning algorithms, Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), to classify heart disease using the BRFSS 2015 dataset. The results showed that XGBoost achieved the highest performance, with an accuracy of 94%, a ROC-AUC of 0.98, and an F1-score of 0.94. These outcomes indicate that XGBoost has excellent capability in distinguishing between individuals with and without heart disease, supported by its robustness in handling class imbalance and preventing overfitting. The learning curve analysis also confirmed the model's stability, with a smooth convergence between training and validation accuracy. This result is particularly significant when compared to the baseline model. For comparison, the XGBoost model trained on the original, imbalanced data without hyperparameter tuning only achieved an accuracy of 90.51%. This demonstrates that the application of SMOTE-Tomek and RandomizedSearchCV was efficient in improving model performance.

The superior performance of XGBoost can be attributed to its Gradient boosting framework, which optimizes multiple weak learners into a strong predictive model. In contrast, SVM achieved a moderate performance with an accuracy of 78.91% and an ROC-AUC of 0.86, indicating that it was less effective in capturing nonlinear relationships within the large-scale BRFSS dataset. The KNN algorithm achieved an accuracy of 87.19% and ROC-AUC of 0.95, but was more sensitive to data noise and variations in data distribution. These findings highlight that XGBoost offers more consistent and accurate classification results compared to SVM and KNN when applied to imbalanced health data.

In a clinical context, the False Negative (FN) rate carries the highest risk, as it represents patients with heart disease who are incorrectly predicted as healthy (no heart disease). This type of error is highly undesirable as it can lead to delayed or missed vital treatment. The model's recall on the Positive Class (Class 1), which measures the rate of FN, is critically important. While XGBoost achieved a high recall of 0.90, it still resulted in approximately 10% of actual heart disease cases being misclassified as healthy. This rate, although lower than other models, must be addressed in a real-world implementation. Conversely, a False Positive (FP) error, where a healthy patient is incorrectly classified as sick, is less critical but leads to unnecessary follow-up procedures. The high precision score of XGBoost on the positive class (0.99, see Table 4) indicates that the model excels at minimizing False Positives, demonstrating strong diagnostic reliability.

The results of this study align with the findings of (Sah et al., 2025), who reported that XGBoost outperformed other boosting algorithms such as Gradient Boosting and Adaptive Boosting in heart disease classification. Similarly, (Rahman, Agusman, & Sutabri, 2024) and (Hidayat et al., 2024) demonstrated that machine learning algorithms are effective in identifying cardiovascular diseases, with XGBoost achieving higher accuracy and generalization capability. Furthermore, (Andani, Triloka, Irianto, & Nugroho, 2025) also emphasized that ensemble and distance-based algorithms such as KNN and Random Forest can perform well for medical classification tasks, supporting the relevance of comparing multiple models as conducted in this study. These comparisons confirm that integrating XGBoost with data balancing and parameter optimization yields a more reliable classification model.

Nevertheless, this study has several limitations. The BRFSS dataset used in this research is based on self-reported survey data, which may include bias or inaccuracies in health-related responses. Furthermore, the evaluation was conducted only on one dataset without external validation using real clinical data, which may affect generalizability. Future research could enhance model reliability by incorporating a multi-year dataset, clinical health records, and real-time medical data to improve early detection and decision-support systems.

CONCLUSION

This study successfully implemented the Synthetic Minority Over-sampling Technique combined with Tomek Links (SMOTE-Tomek) to address class imbalance, followed by an evaluation of three machine learning algorithms: Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The goal was to develop a robust model for predicting heart disease risk using the large BRFSS 2015 dataset.

The results unequivocally demonstrated that XGBoost achieved the highest performance with an Accuracy of 94% and a ROC-AUC score of 0.98. A subsequent 5-fold cross-validation confirmed the model's exceptional stability, indicated by the minimal standard deviation across all metrics (e.g., Accuracy: 0.94 ± 0.0003). This low variability validates the model's robustness and strong potential for generalization. Furthermore, the model demonstrated a high recall for the positive class (0.90), which is crucial for minimizing the critical risk of False Negatives (FNs) in a clinical setting. This finding suggests that XGBoost is the most suitable model among the three evaluated for developing a reliable initial screening tool for heart disease.

The primary limitation of this study lies in the self-reported nature of the BRFSS dataset, which may introduce bias or inaccuracies in the input variables. Therefore, the model's high performance cannot be directly generalized as 'ready for clinical use' without further validation. Future work should focus on: (1) Validating the XGBoost model using clinical datasets obtained from prospective patient studies; and (2) Investigating the integration of this model into Electronic Medical Record (EMR) systems to provide real-time risk assessment for healthcare providers.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to STMIK TIME Medan for the facilities and support provided during the completion of this research. Appreciation is also extended to the Department of Informatics Engineering for their guidance and contribution throughout the study. The authors would also like to thank all individuals who have provided valuable insights and assistance in data analysis and manuscript preparation.

REFERENCES

Adi, S., & Wintarti, A. (2022). Komparasi Metode Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Dan random forest (RF) Untuk Prediksi Penyakit Gagal Jantung. *MATHunesa: Jurnal Ilmiah Matematika*, *10*(2), 258–268. https://doi.org/10.26740/mathunesa.v10n2.p258-268

Andani, M., Triloka, J., Irianto, S. Y., & Nugroho, H. W. (2025). Comparison of K-Nearest Neighbor, Naive Bayes, Random Forest Algorithms for Obesity Prediction. *Sinkron*, *9*(1), 502–510. https://doi.org/10.33395/sinkron.v9i1.14478

Arif, S. N. N., Siregar, A. M., Faisal, S., & Juwita, A. R. (2024). Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(3), 1617–1626. https://doi.org/10.30865/mib.v8i3.7844

Derisma. (2020). Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 4). Retrieved from http://jurnal.polibatam.ac.id/index.php/JAIC

Hairani, H., Anggrawan, A., & Priyanto, D. (2023). Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. *International Journal on Informatics Visualization*, *7*(1), 258–264. https://doi.org/10.30630/joiv.7.1.1069

Hidayat, R., Sy, Y. S., Sujana, T., Husnah, M., Saputra, H. T., & Okmayura, F. (2024). Implementasi Machine Learning Untuk Prediksi Penyakit Jantung Menggunakan Algoritma Support Vector Machine. *BIOS : Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *5*(2), 161–168. https://doi.org/10.37148/bios.v5i2.152

Maskuri, M. N., Sukerti, K., & Herdian Bhakti, R. M. (2022). Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *4*(1).

Natsir, Fitra. M., Bakti, R. Y., & Wahyuni, T. (2024). Analisis Deteksi Dini Penyakit Jantung dengan Pendekatan Support Vector Machine pada Data Pasien. *Arus Jurnal Sains Dan Teknologi (AJST)*, *2*(2), 437–446. https://doi.org/10.57250/ajst.v2i2.669

Nugraha, W. (2021). Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi. *Jurnal SIGMATA*, *9*(2), 82–89. https://doi.org/10.31704/sigmata.v9i2.214

Pramudhyta, N. A., & Rohman, M. S. (2024). Perbandingan Optimasi Metode Grid Search dan Random Search dalam Algoritma XGBoost untuk Klasifikasi Stunting. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(1), 19–29. https://doi.org/10.30865/mib.v8i1.6965

Rahman, H., Agusman, R., & Sutabri, T. (2024). MODEL PREDIKSI PENYAKIT JANTUNG MENGGUNKAN MACHINE LEARNING. In *Tata Sutabri Jurnal Ilmiah Betrik* (Vol. 15). https://doi.org/10.36050/46ccvp37

Ratantja Kusumajati, F., Rahmat, B., & Junaidi, A. (2024). Implementation Of Balancing Data Method Using Smotetomek In Diabetes Classification Using XGBoost. *Jurnal Ilmiah Kursor*, *12*(4), 201–212. https://doi.org/10.21107/kursor.v12i4.410

Sah, A., Niesa, C., Jafar, R. R., & Muharrom, M. (2025). Analisis Model Prediksi Penyakit Jantung Menggunakan Adaptive Boosting, Gradient Boosting, dan Extreme Gradient Boosting. *Jurnal Ilmiah FIFO*, *17*(1), 46–56. https://doi.org/10.22441/fifo.2025.v17i1.006

Sausan, Pratiwi, D. M., & Mufidah, L. (2024). Perbandingan Metode Decision Tree Classifier dan XGBoost Classifier Dalam Memprediksi Penyakit Jantung. *CENTIVE: Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2024*, *4*(1), 991–1000.

Shabrina Assyifa, D., & Luthfiarta, A. (2024). SMOTE-Tomek Re-sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-class Classification. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(2), 151–160. https://doi.org/10.25139/inform.v9i2.8410

Sukamto, T. F., Prameswary, C. L., Royadi, D., & Sofia, D. (2025). Diabetes Disease Prediction on Unbalanced Data Using SMOTE-Tomek Links and Random Forest Algorithm. *G-Tech: Jurnal Teknologi Terapan*, *9*(3), 1194–1203. https://doi.org/10.70609/g-tech.v9i3.7164

Sumantiawan, D. I., Suseno, J. E., & Syafei, W. A. (2023). Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction. *J. Sistem Info. Bisnis*, *13*(1), 1–9. https://doi.org/10.21456/vol13iss1pp1-9

Surono, M., Fadli, M., Purwamti, D., & Susanto, E. R. (2025). Hybrid XGBoost-SVM Model untuk Sistem Pendukung Keputusan dalam Prediksi Penyakit Diabetes. *INSOLOGI: Jurnal Sains Dan Teknologi*, *4*(3), 443–454. https://doi.org/10.55123/insologi.v4i3.5410

Vahlevy, A. D., Zendrato, L. P. E., Fadillah, R., & Sidiq, J. R. (2023). Tinjauan Literatur Sistematik pada Sistem Pakar untuk Diagnosa Penyakit Manusia. *Jurnal Artificial Inteligent Dan Sistem Penunjang Keputusan*, *1*(1). Retrieved from https://garuda.kemdikbud.go.id/.

World Health Organization. (2023). Cardiovascular diseases (CVDs). Retrieved November 8, 2025, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yennimar, Rasid, A., & Kenedy, S. (2023). Implementation Of Support Vector Machine Algorithm With Hyper-Tuning Randomized Search In Stroke Prediction. In *Journal of Information Systems and*

Computer Science Prima) (Vol. 6). https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v6i2.3479

Yogianto, A., Homaidi, A., & Fatah, Z. (2024). Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung. *G-Tech: Jurnal Teknologi Terapan*, *8*(3), 1720–1728. https://doi.org/10.33379/gtech.v8i3.4495