Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

RESEARCH PAPER OPEN ACCES Check for updates

The Impactness of SMOTE as Imbalance Class Handling for Myocardial Infarction Complication Classification using Machine Learning Approach with Data Imputation and Hyperparameter

Ahmad Tajali, Triando Hamonangan Saragih, Muhammad Itqan Mazdadi, Irwan Budiman, and Andi Farmadi

Department of Computer Science, Lambung Mangkurat University, Kalimantan Selatan, Indonesia

ABSTRACT

Myocardial Infarction (MI) is a critical medical emergency characterized by the sudden blockage of blood flow to the heart muscle, often resulting from a blood clot in a coronary artery that has been narrowed by atherosclerotic plaque buildup. This condition demands immediate attention, as prolonged disruption of blood supply can cause irreversible damage to the heart muscle. Diagnosing MI typically involves a combination of methods, including a physical examination, electrocardiogram (ECG) analysis, blood tests to measure heart-specific enzymes, and imaging techniques such as coronary angiography. Early prediction of potential MI complications is crucial to prevent severe outcomes and improve patient prognosis. This study focuses on the early prediction of MI complications through the application of machine learning classification methods. We employed algorithms such as Support Vector Machine (SVM), Random Forest, and XGBoost to analyze patient medical records and accurately predict these complications. The selection of Support Vector Machine (SVM), Random Forest, and XGBoost in this study is driven by their proven effectiveness in handling complex classification problems. To manage incomplete datasets and preserve valuable information, data imputation techniques like K-Nearest Neighbors (KNN) Imputation, Iterative Imputation, and MissForest were applied. KNN, Iterative, and MissForest imputations were chosen to handle missing data due to their effectiveness in preserving data integrity, which is crucial for accurate predictions in myocardial infarction complication studies. Additionally, Bayesian Optimization was utilized to fine-tune the hyperparameters of the models, thereby enhancing their predictive accuracy. The Iterative Imputation method yielded the best performance, particularly in SVM and XGBoost algorithms. SVM achieved 100% accuracy, precision, sensitivity, F1 score, and Area Under the Curve (AUC), while XGBoost attained 99.4% accuracy, 100% precision, 79.6% sensitivity, an F1 score of 88.7%, and an AUC of 0.898. While XGBoost and MissForest proved to be the most successful pairing, the overall effectiveness of the models suggests that Iterative Imputation and Random Forest also have potential under certain conditions.

PAPER HISTORY

Received August 02, 2024
Revised Sept. 20, 2024
Accepted Nov. 10, 2024

KEYWORDS

Myocardial Infarction;
Random Forest;
Support Vector Machine;
Extreme Gradient Boosting;
Data Imputation

CONTACT:

Triando.saragih@ulm.ac.id

1. INTRODUCTION

One of the leading causes of global mortality and hospitalization is the life-threatening heart disease known as myocardial infarction (MI) [1][2]. The diagnosis of MI is commonly performed using electrocardiography (ECG), which reflects the heart's electrophysiological activity [3]. This dangerous state stems from myocardial dysfunction, precipitating harm to the cardiac muscle tissue. Insufficient blood circulation often obstructs one or multiple arteries, affecting the cardiac musculature. As a result, timely intervention becomes imperative in the management of myocardial infarction [4]. According to data released by the World Health Organization (WHO), cardiovascular diseases claimed approximately 17.9 million lives in 2019, representing 32% of all global fatalities [5]. Given the severity of this condition, analyzing data on myocardial infarction complications plays a critical role in predicting disease progression, allowing for earlier interventions, better management strategies, and improved outcomes in patient care.

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Accurately predicting or detecting cardiovascular diseases is the primary challenge in implementing effective prevention and early intervention strategies. In the digital era, the rise of electronic health records (EHR) has made the need for efficient data mining tools to uncover valuable insights increasingly apparent [4]. The rise of digital health advancements and the accessibility of extensive data have given machine learning (ML) and data mining algorithms the potential to significantly enhance clinical decision support [6]. ML can be utilized for early screening and diagnosis, identifying risk factors for disease prevention [7], managing and monitoring treatments with improved pharmacovigilance and patient safety, and ultimately improving care and outcomes [8], particularly in cardiovascular diseases [9][10].

Machine learning has emerged as a modern predictive technique. Unlike traditional statistical tools, ML methods identify relationships between variables in a training dataset to forecast various outcomes, including mortality [11]. Random Forest, XGBoost, and Support Vector Machine (SVM) are examples of machine learning algorithms, are commonly employed to achieve accurate prediction outcomes in classification tasks that analyze patient medical record data.

Recognized for its effectiveness in high-dimensional spaces and strong performance with clear margin separation, the Support Vector Machine (SVM) is a supervised learning model utilized for data analysis in both classification and regression tasks [12]. The way SVM works is by finding the best hyperplane in a high-dimensional space to divide data points from various classes. This approach makes sure that the model effectively generalizes to new, unknown data by optimizing the margin between the closest data points of different classes, often known as support vectors.

Random Forest, an ensemble learning method, builds multiple decision trees during the training phase and outputs the mode of the classes for classification tasks. This process results in high accuracy and a strong resistance to overfitting [13].Combining the predictions of multiple decision trees, Random Forest mitigates the risk of overfitting that individual trees might exhibit. It achieves this by training each tree on a random subset of features and data points, ensuring tree diversity.

Designed for efficiency, flexibility, and portability, XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting library recognized for its high speed and robust performance in classification and regression tasks. Based on the Gradient Boosted Decision Tree (GBDT) framework, XGBoost improves the management of missing values and incorporates advanced regularization techniques to reduce overfitting, which makes it especially effective for analyzing large-scale datasets.

Addressing missing data and class imbalance represents critical challenges in creating robust predictive models, in addition to selecting appropriate machine learning algorithms. To manage incomplete datasets and prevent the loss of valuable information, data imputation techniques such as K-Nearest Neighbors (KNN) impute, Iterative impute, and MissForest are essential [15]. KNN imputation functions by locating the k-nearest neighbors to a missing value and replacing it with the mean or mode derived from these neighboring data points. For example, a missing blood pressure value might be imputed by averaging the values of patients with similar demographics [16]. Iterative imputation, commonly known as Multiple Imputation by Chained Equations (MICE), addresses missing data by performing multiple rounds of imputations. This method takes uncertainty into consideration by producing several imputed datasets. This method could iteratively fill in missing cholesterol levels while accounting for the potential correlation with other patient health metrics [17]. MissForest, which is a non-parametric imputation method, utilizes random forest algorithms to predict and substitute missing values by relying on the observed data. it can handle complex cases like imputing missing ECG readings by utilizing patterns in other variables such as age, gender, and prior health history [18]. Class imbalance, a common issue in medical datasets, occurs when there are significantly fewer instances of a minority class, such as complications, compared to a majority class, like non-complications. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is often employed. SMOTE rebalances the dataset by generating synthetic instances for the minority class (over-sampling) or by reducing the instances of the majority class (under-sampling), helping to mitigate the effects of imbalance and improve model performance.

Despite the advancements in ML and data imputation techniques, there remain challenges in effectively handling missing data and class imbalance. Prior research has frequently concentrated on discrete imputation approaches or machine learning models without incorporating sophisticated methods for hyperparameter optimization. Furthermore, there hasn't been a detailed investigation of how well these integrated techniques predict the consequences of myocardial infarction.

Improving performance requires optimizing the hyperparameters of machine learning models. Bayesian Optimization, which has become widely used for hyperparameter tuning, builds a surrogate probability model based on previous evaluation results to identify the value that minimizes the objective function [21]. Particularly beneficial for problems involving costly, non-differentiable, or complex function evaluations, Bayesian Optimization is highly effective [22].

Recent studies have demonstrated the efficacy of these methods. For instance, Ishaq et al. (2021) utilized SMOTE and effective data mining techniques to improve the prediction of heart failure patients' survival, achieving significant improvements in accuracy with Random Forest and SVM models, with Extra Tree Classifier (ETC) and

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

SMOTE achieving the highest accuracy of 92.62% [23]. Similarly, Alaa et al. (2019) implemented XGBoost with Miss Forest and Iterative imputation, demonstrating a notable accuracy in cardiovascular disease risk prediction using automated machine learning, with an AUC-ROC of 0.713 in diabetic populations [24]. Additionally, Ogunpola et al. (2024) employed SVM, RF, and XGBoost for cardiovascular disease detection, achieving a high accuracy of 98.50% [25]. Qin et al. (2020) confirmed the effectiveness of KNN impute combined with SMOTE in enhancing the prediction of chronic kidney disease, showing a significant improvement in diagnostic accuracy to 99.75% with Random Forest [26]. Furthermore, Kadhim and Radhi (2023) utilized optimized ML algorithms, including SVM and RF, to classify heart diseases, demonstrating substantial accuracy improvements through hyperparameter optimization, achieving up to 95.4% accuracy with RF [27]. These findings underscore the potential of combining advanced imputation techniques, oversampling methods, and optimization techniques with powerful ML algorithms to improve predictive performance in clinical settings.

Using class imbalance handling techniques (SMOTE) and data imputation methods (KNN, Iterative, and Miss Forest), this study attempts to assess how well a number of machine learning algorithms, including SVM, RF, and XGBoost, predict myocardial infarction complications. This study also aims to evaluate how Bayesian optimization's hyperparameter adjustment affects prediction accuracy. This study intends to improve patient outcomes and close gaps in the existing literature on prediction strategies by integrating these cutting-edge approaches to improve early detection and management of myocardial infarction complications. The research's contributions are as follows: 1. This study provides a comprehensive evaluation of combined data imputation, oversampling methods, and ML models for predicting myocardial infarction complications; 2. It highlights the effectiveness of Bayesian Optimization in enhancing the predictive performance of SVM, RF, and XGBoost models, 3. The findings offer valuable insights for improving clinical decision support systems in cardiovascular disease management, 4. By addressing challenges related to missing data and class imbalance, this research presents a robust framework applicable to various medical predictive tasks.

2. MATERIALS AND METHOD

In this study, three machine learning models—Support Vector Machine (SVM), Random Forest (RF), and XGBoost—are used to assess the performance of three data imputation techniques: K-Nearest Neighbors (KNN), Iterative imputation (MICE), and Miss Forest imputation. Each model undergoes hyperparameter tuning through Bayesian Optimization. Data preprocessing (including SMOTE for oversampling), data splitting into training and testing sets, model training, data evaluation, and data gathering from a myocardial infarction (MI) complications dataset comprise the study's five procedures. Figure 1 shows the workflow for the research.

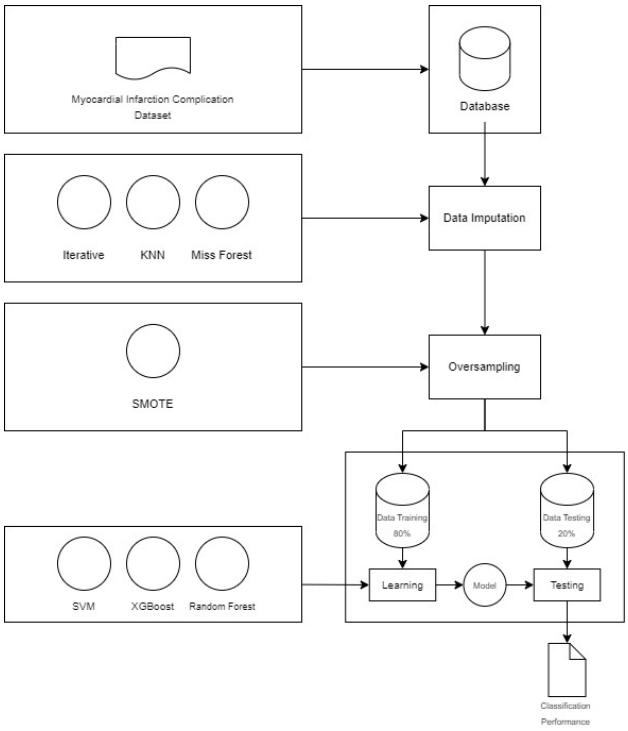


Fig. 1. Research Flowchart

A. Data Collection

The data for this study were collected from publicly available patient records, ensuring adherence to ethical standards, including anonymization to protect patient privacy and confidentiality. The dataset utilized in this study consists of 1700 instances with 111 features of patient's medical records. During the time of admission, the first day, the second day, and the third day of hospitalization, these aspects include demographic data, medical history, diagnostic test results, and clinical findings. Labels for many possible myocardial infarction (MI) consequences, including atrial fibrillation, supraventricular tachycardia, ventricular tachycardia, and pulmonary edema, are included in the dataset, which was obtained from this link. https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications is the source of the information [28].

B. Data Imputation

Data imputation addresses the problem of missing values, especially in health data, by estimating and filling gaps to maintain data integrity for accurate analysis and model training. It is crucial for preserving the quality of statistical analyses and machine learning models, preventing bias

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

and performance issues. The effectiveness of different imputation methods can be observed in Figure 2.

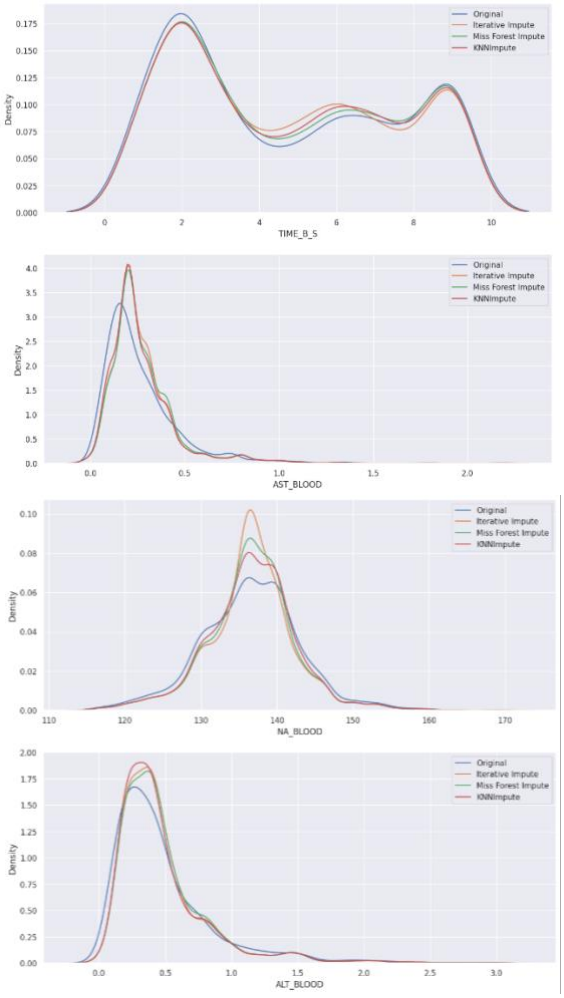


Fig. 2. Density Plots of Original and Imputed Data Across Various Medical Features

In this research, the challenge of missing data is particularly significant given the complexity and sensitivity of the medical records involved, which include crucial diagnostic test results and clinical findings. The missing values could lead to incomplete or biased analyses, potentially compromising the reliability of the study's conclusions. The study applied three imputation methods to address this issue: Iterative Impute, Miss Forest Impute, and KNN Impute. As illustrated in the density plots, the results indicate that subtle differences exist while each method effectively approximates the original data distribution. These differences suggest that the choice of imputation method can introduce minor biases, especially in variables with more complex distributions, such as AST_BLOOD and ALT_BLOOD. Therefore, careful consideration must be given to selecting an appropriate imputation technique to ensure the validity and robustness of the following statistical models and analyses.

1) Iterative Imputation

In statistical literature, iterative imputation, also known as multivariate imputation by chained equations (MICE), has gained recognition as a principled approach for handling missing data. Often referred to as "fully conditional specification" or "sequential regression multiple imputation," this method effectively addresses the challenge of incomplete datasets [17], [29].

This imputation technique involves running multiple regression models where the variable with missing data is predicted based on other variables in the dataset. The modeling approach depends on the variable type: for instance, logistic regression is used for binary variables, while predictive mean matching is applied to continuous variables. The process consists of four steps repeated until the best results are achieved. Initially, missing data is temporarily replaced with the mean of the observed values as a placeholder. Next, these placeholder values are reset to missing. Then, a regression model is run, where the observed values of the target variable are predicted using the other variables as predictors [33].

2) KNN Imputation

K-Nearest Neighbors (KNN) imputation uses the KNN principles to fill in missing data. This eliminates gaps by substituting the average (or the mode for categorical variables) derived from the feature space's closest neighbours for missing values. KNN finds a set of neighbouring data points with missing values in order to estimate and fill in the missing data in the context of imputation. This non-parametric approach, frequently applied to classification and regression applications, uses distance metrics such as Manhattan, Minkowski, or Euclidean to determine how close missing data items are to their closest neighbours. Following the identification of the closest neighbours, the mode for categorical variables or the average for numerical variables is used to replace the missing values.
The KNN imputation equation is shown in Eq. (1).

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k} \quad (1)$$

The KNN imputation equation is presented in (1), where $d_{i,j}$ represents the distance between the i-th and j-th data points. The variable p denotes the total number of features being compared, while Wk indicates the weight assigned to the k-th feature. Lastly, $\delta_{i,j,k}$ signifies the difference between the i-th and j-th data points for the k-th feature.

The KNN imputation approach, which can handle a variety of variables, including binary, categorical, ordered,

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

continuous, and semi-continuous distances, is used in this study with distance weighting factors. The distance between two values is calculated based on a weighted average, where each variable's contribution is considered. These weights are designed to represent the importance of each variable in the overall distance calculation.

3) MissForest Imputation

MissForest imputation estimates missing data in a dataset by utilizing the Random Forest algorithm, a non-parametric technique. This approach takes advantage of Random Forest's strength in handling complex, interrelated data, resulting in more accurate estimates for missing values. As an ensemble learning algorithm, Random Forest enhances accuracy and minimizes the risk of overfitting by combining predictions from multiple decision trees. In the MissForest method, the power of Random Forest is employed to predict and fill in missing values by analyzing the existing data in the dataset [31].

The imputation process begins by replacing missing data with the mean (for continuous variables) or the most frequent category (for categorical variables). The variable in question is then divided into two parts: the observed data, which is complete, and the missing data, which requires prediction. A Random Forest model is trained using the observed data as the response variable, with other variables serving as predictors. The missing data is then imputed with the predictions from the Random Forest. This can be mathematically represented as Eq. (2):

$$D_{\substack{imputed \\ missing}} = f(D_{observed}, D_{predictors}) \quad (2)$$

where f represents the function of the Random Forest model that predicts missing values based on the observed data and other predictor variables. This procedure is repeated for each variable until the differences between iterations are minimal [32].

C. Oversampling

When the distribution of classes is disproportionate, the data is considered imbalanced. This issue is common in many real-world datasets, where standard instances significantly outnumber abnormal ones. Classifiers trained on such datasets often become either overfitted or underfitted. To address this problem, several resampling techniques have been introduced [34].

1) Synthetic Minority Over-sampling Technique (SMOTE)

The SMOTE (Synthetic Minority Over-sampling Technique) algorithm, introduced by Chawla, addresses imbalanced data. SMOTE creates new synthetic samples through random linear interpolation between existing minority samples and their nearest neighbors. By generating a specific number of artificial minority samples, the imbalance ratio is reduced, leading to improved classification performance on the imbalanced dataset [35].

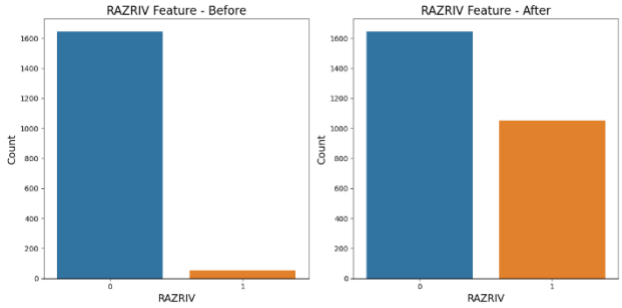


Fig. 3. Distribution of *RAZRIV* Feature Before and After SMOTE Oversampling

In this research, the "RAZRIV" feature, which is the target variable (or "y"), was identified as highly imbalanced. As shown in Fig. 3, before applying the SMOTE algorithm, the distribution of the "RAZRIV" feature was significantly skewed, with a large majority of the instances belonging to class 0 (indicating no complication) and a very small number to class 1 (indicating a complication).

To address this imbalance, SMOTE was applied to the dataset, specifically to the "RAZRIV" feature, to generate synthetic samples of the minority class (class 1). The result of this oversampling process is evident on the right side of the figure, where the distribution of the "RAZRIV" feature has become more balanced. This balance between the two classes is crucial for improving the performance of classifiers, as it allows the model to learn equally from both classes, reducing the risk of overfitting to the majority class or underfitting to the minority class.

D. Data Splitting

A widely adopted method for validating models is data splitting, which involves dividing a dataset into two subsets: training and testing. Models are trained on the training data and validated using the test data. This separation ensures that model evaluation is unbiased, allowing for an accurate assessment of predictive performance while minimizing concerns about overfitting the training data [36].

Once a splitting ratio is determined, the previously mentioned data-splitting methods can be applied. A typical ratio is 80:20, where 80% of the data is allocated for training and 20% for testing. However, other ratios like 70:30, 60:40, and even 50:50 are also commonly used in practice. No definitive rule exists regarding which ratio is optimal for a particular dataset. In this research, we adopt the 80:20 data splitting ratio, where 80% of the dataset is utilized for training the model, and the remaining 20% is reserved for testing.

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

E. Hyperparameter Tuning

Solving optimization problems is a key component of machine learning. An optimization method is used to initialize and optimize the weight parameters of an ML model until the accuracy or the objective function approaches a maximum value [37]. This study employed a systematic approach for hyperparameter tuning, starting with creating a simple baseline model. This baseline model was then used for initial hyperparameter tuning, allowing the identification of the best parameters for each model. By storing these optimal parameters, the tuning process became more efficient.

After identifying the best hyperparameters, these values were applied to more advanced models. This two-step approach—first tuning the hyperparameters on simpler models and then applying them to more complex models—was crucial in minimizing computational time and resource consumption. Running advanced models for both hyperparameter search and classification simultaneously would have been highly time-intensive and impractical, so this method ensured a balance between computational efficiency and model performance.

1) Bayesian Optimization

Bayesian optimization (BO) is an iterative algorithm that uses a model to efficiently optimize objectives that are both noisy and costly to evaluate. This approach is especially useful in situations where each evaluation of the objective function is expensive and time-consuming. For instance, BO effectively fine-tun accelerator parameters, such as adjusting magnet power supplies, which can take several seconds. It also optimizes objectives that require considerable measuring time, like transverse beam emittance. Additionally, BO proves valuable in simulated environments where physics simulations demand substantial computational resources for making predictions [38].

F. Machine Learning Model

A branch of artificial intelligence (AI) called machine learning enables computer systems to learn from data and generate predictions without explicit programming. This method is especially useful for addressing complex problems and deriving valuable insights from large datasets [39]. Machine learning has become a useful tool in medical diagnostics with great promise for therapeutic treatment as learning algorithms continue to improve and massive medical datasets become more widely available. These methods, which are being used more and more in a variety of medical specialties, such as the diagnosis of cancer, heart disease, musculoskeletal disorders, and mental disorders, have proven to be useful in supporting doctors by providing precise forecasts, disease detection, and image-based diagnosis. Moreover, machine learning algorithms play a crucial role in addressing challenges such as class imbalances in medical datasets, which enhances the accuracy and reliability of predictive models [40], [41].

1) Random Forest

The Random Forest algorithm makes decisions through a series of steps organized as decision trees. Within the Random Forest, multiple decision trees are generated, each providing its prediction. The final prediction is determined by selecting the class with the majority of votes from these trees [42]. A tree-based model divides the dataset into two groups recursively based on a specific criterion, continuing this process until a predetermined stopping point is reached. The endpoints of these decision trees are referred to as leaf nodes or leaves.

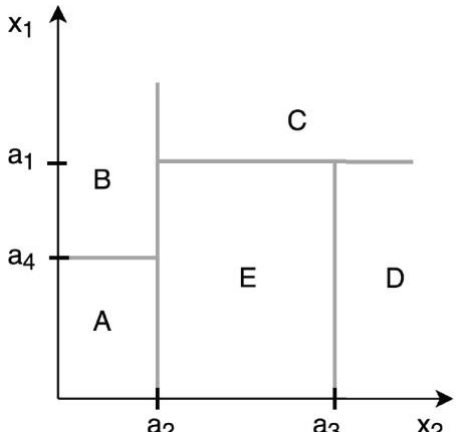


Fig. 4. Recursive binary partition of a two-dimensional subspace

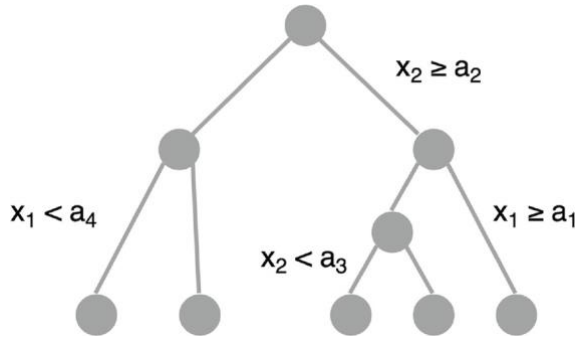


Fig. 5. A graphical representation of the decision tree in Figure 4

Figure 3 depicts a recursive partitioning of a two-dimensional input space using axis-aligned boundaries, meaning the input space is split along lines parallel to one of the axes. In this example, the first division occurs at $x_2 \geq a_2$. Subsequent splits further divide the subspaces: the left branch is split at $x_1 \geq a_4$, while the right branch is initially split at $x_1 \geq a_1$, with one of its subbranches

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

further split at $x_2 \geq a_3$. Fig. 4 visually represents the subspaces divided in Figure 3 [43].

2) Support Vector Machine (SVM)

Support Vector Machines (SVM) are used in supervised learning for both classification and regression tasks [44], [45], [46]. SVM aims to identify the hyperplane that maximizes the margin between data classes, which helps separate them linearly (Fig. 5). SVM is particularly effective for datasets with limited training samples, where traditional statistical methods may not guarantee an optimal solution [47].

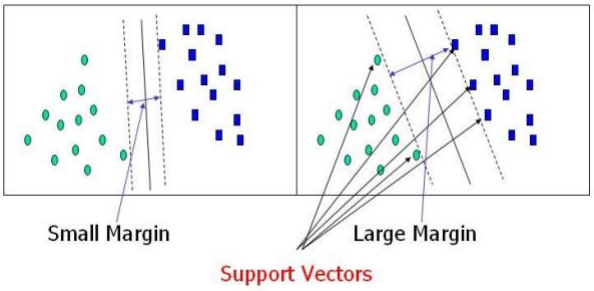


Fig. 6. Margin and Support Vectors

When linear separation is challenging, SVM uses kernel functions to project the training data into a higher-dimensional feature space, making separation more feasible. Kernel functions like the radial basis function (RBF) or polynomial kernel enhance classification accuracy, especially with small training sets [48]. Selecting an appropriate kernel function is essential for evaluating hyperplanes and minimizing classification errors. The SVM's effectiveness largely depends on the choice and size of the kernel, as well as the density of the kernel's surface, which influences smoothness [49], [50].
Unlike other machine learning algorithms, choosing the right kernel function is a complex and critical task. The strength of the SVM algorithm lies in its ability to refine itself by adjusting its kernel feature, allowing it to adapt to different dimensions of the problem. However, this adaptability can increase computational demands, particularly in comparison to other machine learning methods, depending on the problem's dimensionality [51], [52].

3) Extreme Gradient Boosting (XGBoost)

Initially introduced by Chen et al. [53], XGBoost is a scalable tree-boosting system that has gained significant attention for its remarkable efficiency and predictive accuracy. It was prominently utilized in Kaggle's Higgs sub-signal recognition contest, and its popularity has only grown since then. XGBoost improves upon the traditional Gradient Boosting Decision Trees (GBDT) algorithm [54], leveraging a collection of decision trees to handle both classification and regression tasks effectively.

Chen et al. [53] introduced XGBoost as equation (3) described. During each step of the gradient boosting process, the residual is adjusted to improve the accuracy of the previous predictor by optimizing the specified loss function. The term $f_k(x_i)$ refers to the predicted value for the i-th sample generated by the k-th tree. The set of functions f_k is learned by minimizing the following objective function (Eq. (3)):

$$Obj = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (3)$$

Where:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \omega^2 \quad (4)$$

In Eq. (3), the first term, l, is used in the classification tree and is defined as (Eq. (5)):

$$(\hat{y}_i, y_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) ln\left(1 + e^{-\hat{y}_i}\right) \quad (5)$$

This term represents the loss function, which measures the difference between the predicted value $\hat{y}_i$ and the actual target value y_i. The second term, Ω, refers to the regularization component (Eq. (4), which evaluates the complexity of the tree f_k. In this context, γ and λ are regularization parameters, while T is the number of leaves and λ represents the vector of values assigned to each leaf [55].

G. Performance Metrics

Confusion matrices are a fundamental tool for evaluating the performance of classification models in machine learning. By presenting the actual versus predicted outcomes in a structured format, they enable a comprehensive analysis of the model's performance. This framework helps identify not only how often a model correctly or incorrectly classifies data, but also the nature of those errors.

The key metrics in confusion matrices include True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN), which are essential for understanding various aspects of classification accuracy. TPs refer to the correct predictions of the positive class, while FNs occur when the model fails to identify a positive instance. Conversely, FPs indicate instances where the model incorrectly predicts a positive outcome, and TNs capture the accurate identification of negative instances [38]. As summed up in Table 1, these keywords offer important insights into the model's prediction capabilities' advantages and disadvantages across various classes.

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

Table 1. Confusion Matrix

Actual Class	Predicted Class	
	True	False
True	True Positive (TP)	False Negative (FN)
False	False Positive (FP)	True Negative (TN)

To evaluate the performance of the model based on the confusion matrix, several key metrics are used:

Accuracy: This represents the ratio of correctly classified instances (both positive and negative) to the total instances, as expressed in Eq. (6).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \quad (6)$$

Sensitivity: Also known as recall, this metric indicates the proportion of actual positives correctly identified by the model, shown in Eq. (7).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

Precision: This metric reflects the proportion of positive identifications that are correct, as given in Eq. (8).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

F1 Score: A harmonic mean of precision and recall, providing a balance between the two, as represented in Eq. (9).

$$\text{F1} = \frac{2*\text{precision}*\text{Recall}}{\text{precision+Recal}} \quad (9)$$

Additionally, the Area Under the Curve (AUC) offers a valuable measure of a model's ability to differentiate between positive and negative instances. A higher AUC value suggests a better-performing model, with values ranging from 0 to 1, where 1 indicates perfect classification. The AUC can be mathematically modeled as shown in Eq. (10).

$$\text{AUC} = \frac{\left(\frac{\text{TP}}{\text{TP+FN}}\right)\text{x}\left(\frac{\text{TN}}{\text{TN+FP}}\right)}{2} \quad (10)$$

The interpretation of the AUC score is significant in assessing the model's discriminative ability, with higher AUC values corresponding to better classification performance. Table 2 categorizes model performance based on AUC values, helping to determine the effectiveness of the classification model [38].

Table 2. Categories of results from classification based on AUC values

AUC Values	Category
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
0.50 – 0.60	Failure

3. RESULTS

This section focuses on evaluating the classification algorithms SVM, Random Forest, and XGBoost, while also incorporating data imputation techniques like KNN Imputation, Iterative Imputation, and MissForest Imputation. Furthermore, Bayesian Optimization is employed to enhance the hyperparameters of these algorithms. The main objective is to assess the effectiveness of these classification models in predicting complications associated with myocardial infarction. Various performance metrics—including precision, sensitivity, accuracy, F1-score, and the Area Under the Curve (AUC) of the ROC curve—are utilized to achieve this.

Fig. 7. APerformance Matrics Comparison of SVM, Random Forest, and XGBoost with Different Data Imputation Method.

In this study, the dataset is split into training and testing sets with an 80:20 ratio, ensuring that the models are trained on a substantial portion of the data while leaving enough data for a robust evaluation of their predictive performance. This ratio strikes a balance between providing the model with sufficient information for learning and retaining a representative test set to evaluate its generalization capabilities. Given the issue of imbalanced classes in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data to create a more balanced distribution of classes, thereby improving the models' ability to accurately classify minority classes. This 80:20 split also helps mitigate overfitting by ensuring that the model's performance is assessed on data it has not seen during training, leading to more reliable predictions in real-world scenarios.

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 6, No. 4, pp. 227-239, November 2024; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v6i4.13

The performance results of the models with different imputation techniques are presented in Fig. 6. The evaluation results, as seen in the figure, demonstrate a detailed comparison of various metrics across different models and imputation techniques, highlighting the effectiveness of each approach in improving prediction accuracy. As shown in Figue 7, the performance metrics of all algorithms using various data imputation methods are notably strong, with results approaching near-perfect scores of 100%.

A more detailed breakdown can be seen in Table 3, where the XGBoost algorithm, when combined with MissForest data imputation, stands out as the top performer. Specifically, it achieved an accuracy of 99.6%, a sensitivity of 100%, a precision of 99%, an AUC (Area Under the Curve) score of 99.6%, and an F1 score of 99.5%.

Table 3. Performance Metrics Result

Model	Imputation	Performance Metrics				
		AUC	F1	Acc	Sensitivity	Precision
SVM	Iterative	0.995	0.995	0.996	0.99	1.0
	Miss Forest	0.989	0.99	0.987	1.0	0.968
	KNN	0.944	0.926	0.937	0.981	0.876
RF	Iterative	0.966	0.964	0.972	0.935	0.995
	Miss Forest	0.956	0.954	0.964	0.912	1.0
	KNN	0.954	0.951	0.962	0.912	0.994
Xgboost	Iterative	0.993	0.99	0.992	0.995	0.986
	Miss Forest	0.996	0.995	0.996	1.0	0.99
	KNN	0.994	0.993	0.994	0.995	0.99

4. DISCUSSION

The results of this study shed important light on how well the machine learning algorithms SVM, Random Forest, and XGBoost perform when used to the prediction of myocardial infarction-related complications. We sought to determine the best strategy for improving prediction accuracy by utilizing a variety of data imputation techniques, including KNN, Iterative, and MissForest Imputation, and optimizing model performance using Bayesian Optimization. The inclusion of SMOTE, a technique for balancing the dataset, proved essential for addressing the inherent class imbalance in the data, particularly by ensuring that minority classes were well represented during training.

Among the algorithms evaluated, XGBoost paired with MissForest Imputation emerged as the superior model, achieving near-perfect performance across key metrics such as accuracy (99.6%), sensitivity (100%), and precision (99%). This suggests that the combination of XGBoost's advanced gradient-boosting methodology and MissForest's ability to handle missing data effectively contributed significantly to the model's predictive strength. Additionally, the model's high AUC score of 99.6% highlights how strong it is at differentiating across classes, which makes it a good fit for challenging classification tasks like forecasting the consequences of myocardial infarction.

In comparison to other studies that can be seen in Table 4, our results exhibit superior performance across key metrics, particularly with the integration of XGBoost and MissForest Imputation, yielding 99.6% accuracy, 100% sensitivity, and an AUC of 0.996. For example, Ishaq et al. (2021) reported an accuracy of 92.62% using the Extra Tree Classifier (ETC) with SMOTE, while Ogunpola et al. (2024) achieved an accuracy of 98.50% with SVM, Random Forest (RF), and XGBoost for cardiovascular disease detection. Although these studies demonstrate strong results, our application of Bayesian Optimization, SMOTE, and advanced imputation methods consistently led to higher accuracy and improved predictive performance. This comparison highlights the effectiveness of combining XGBoost with MissForest Imputation for tackling complex healthcare prediction challenges, such as myocardial infarction complications.

Despite the promising outcomes, several limitations need to be acknowledged. First, the computational cost associated with MissForest Imputation and Bayesian Optimization is relatively high, especially when applied to large datasets. In real-time clinical settings, where rapid decisions are crucial, this could pose a challenge. Another limitation is the use of KNN Imputation, which consistently underperformed across all models. This highlights that simpler imputation methods may not be suitable for handling datasets with complex variable interactions, such as those seen in healthcare applications.

Table 4. Comparison of Performance Metrics Result

Study	Model & Method	Accuracy (%)
This Study (2024)	XGBoost + MissForest, SMOTE, Bayesian Opt	99.6
Ishaq et al. [23]	ETC + SMOTE	92.62
Ogunpula et al. [25]	SVM, RF, XGBoost	98.50
Qin et al. [26]	Random Forest + KNN Impute, SMOTE	99.75
Kadhim and Radhi [27]	Random Forest, hyperparameter tuning	95.4

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Moreover, the study is limited to three machine learning models—SVM, Random Forest, and XGBoost—and three imputation techniques. While these are widely used, exploring a broader range of models, including deep learning algorithms or ensemble models, might yield additional insights. Finally, the dataset used in this study was relatively small and sourced from a single database. Larger, more diverse datasets from multiple sources would improve the generalizability of the results.

The findings of this study have several important implications for healthcare and machine learning. First, they confirm that advanced algorithms like XGBoost, when paired with robust imputation techniques such as MissForest, can significantly enhance the predictive accuracy of models in clinical settings. This improvement in prediction accuracy is critical for timely intervention in patients at risk of complications from myocardial infarction, potentially reducing mortality and improving patient outcomes.

Moreover, the study highlights the importance of addressing data imbalance and missing data, both of which are common in medical datasets. The success of SMOTE in balancing the dataset and MissForest in handling missing data underscores the need for thoughtful data preprocessing techniques to ensure reliable model performance. Additionally, the high accuracy achieved through Bayesian Optimization suggests that careful hyperparameter tuning can further improve the performance of machine learning models in healthcare.

Finally, the results suggest that future research should focus on scalability and computational efficiency to make these methods more applicable to real-time clinical decision-making. Researchers should also investigate the benefits of combining feature selection techniques with existing models to reduce complexity and improve interpretability. By continuing to explore new imputation methods, data balancing techniques, and more powerful machine learning models, the predictive capabilities of healthcare algorithms can be further enhanced.

Future research should explore additional methods that could further enhance the predictive performance of machine learning models in healthcare-related tasks, particularly in the context of myocardial infarction complications. One promising area is incorporating feature selection techniques, which can help identify the most relevant variables for prediction, potentially improving model efficiency and reducing computational complexity. Additionally, experimenting with ensemble methods or hybrid models that combine the strengths of multiple algorithms may yield even better results. Further investigation into more advanced imputation techniques and exploring alternative data balancing methods beyond SMOTE could also enhance the handling of imbalanced datasets.

5. CONCLUSION

Three popular machine learning algorithms—SVM, Random Forest, and XGBoost—were evaluated in this study for their ability to predict myocardial infarction-related complications. The models were assessed using various data imputation methods, including KNN, Iterative, and MissForest Imputation, with their hyperparameters optimized through Bayesian Optimization. The results demonstrated that all three algorithms achieved strong performance across multiple metrics, such as accuracy, sensitivity, precision, F1 score, and AUC. However, XGBoost, in combination with MissForest Imputation, emerged as the most effective approach, achieving near-perfect performance across key metrics such as AUC score of 99.6%, F1 (99.5%), accuracy (99.6%), sensitivity (100%), and precision (99%), and showcasing its robustness in handling missing data and imbalanced classes.

These results reflect the importance of oversampling imbalanced datasets. This research, therefore, offers a better understanding of the effectiveness of various data imputation methods and machine learning algorithms. Future development prospects could involve feature selection methods and combining several classification models.

REFERENCES

[1] A. Chakraborty, S. Chatterjee, K. Majumder, R. N. Shaw, and A. Ghosh, “A Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning,” Advanced Computing and Intelligent Technologies, pp. 257–267, Jul. 2021, doi: https://doi.org/10.1007/978-981-16-2164-2_21.

[2] U. B. Baloglu, M. Talo, O. Yildirim, R. S. Tan, and U. R. Acharya, “Classification of myocardial infarction with multi-lead ECG signals and deep CNN,” Pattern Recognition Letters, vol. 122, pp. 23–30, May 2019, doi: https://doi.org/10.1016/j.patrec.2019.02.016.

[3] Z. Wang, L. Qian, C. Han, and L. Shi, “Application of multi-feature fusion and random forests to the automated detection of myocardial infarction,” Cognitive Systems Research, vol. 59, pp. 15–26, Jan. 2020, doi: https://doi.org/10.1016/j.cogsys.2019.09.001.

[4] Imen Boudali, Sarra Chebaane, and Yassine Zitouni, “A predictive approach for myocardial infarction risk assessment using machine learning and big clinical data,” Healthcare analytics, vol. 5, pp. 100319–100319, Jun. 2024, doi: https://doi.org/10.1016/j.health.2024.100319.

[5] S. Kaptoge et al., “World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions,” The Lancet Global Health, vol. 7, no. 10, pp. e1332–e1345, Oct. 2019, doi: https://doi.org/10.1016/s2214-109x(19)30318-3.

[6] A. Pina, M. P. Macedo, and R. Henriques, “Clustering Clinical Data in R,” Methods in Molecular Biology (Clifton, N.J.), vol. 2051, pp. 309–343, 2020, doi: https://doi.org/10.1007/978-1-4939-9744-2_14.

[7] A. Pina, Maria João Meneses, Inês Sousa-Lima, R. Henriques, João Filipe Raposo, and M. Paula Macedo, “Big data and machine learning to tackle diabetes management,” European Journal of Clinical Investigation, vol. 53, no. 1, Nov. 2022, doi: https://doi.org/10.1111/eci.13890.

[8] A. Pina et al., “Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning,” European Journal of Preventive Cardiology, vol. 27, no. 15, pp. 1639–1646, Feb. 2020, doi: https://doi.org/10.1177/2047487319898951.

[9] M. Oliveira, J. Seringa, Fausto José Pinto, R. Henriques, and T. Magalhães, “Machine learning prediction of mortality in Acute

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Myocardial Infarction," BMC Medical Informatics and Decision Making, vol. 23, no. 1, Apr. 2023, doi: https://doi.org/10.1186/s12911-023-02168-6.

[10] L. A. Barrett, S. N. Payrovnaziri, J. Bian, and Z. He, 'Building Computational Models to Predict One-Year Mortality in ICU Patients with Acute Myocardial Infarction and Post Myocardial Infarction Syndrome', AMIA Jt Summits Transl Sci Proc, vol. 2019, pp. 407–416, May 2019.

[11] U. B. Baloglu, M. Talo, O. Yildirim, R. S. Tan, and U. R. Acharya, "Classification of myocardial infarction with multi-lead ECG signals and deep CNN," Pattern Recognition Letters, vol. 122, pp. 23–30, May 2019, doi: https://doi.org/10.1016/j.patrec.2019.02.016.

[12] None Shahmirul Hafizullah Imanuddin, None Kusworo Adi, and None Rahmat Gernowo, "Sentiment Analysis on Satusehat Application Using Support Vector Machine Method," Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 5, no. 3, pp. 143–149, Jul. 2023, doi: https://doi.org/10.35882/jeemi.v5i3.304.

[13] D. Dhiyaussalam, A. Wibowo, F. A. Nugroho, E. A. Sarwoko, and I. M. A. Setiawan, "Classification of Headache Disorder Using Random Forest Algorithm," in ICICoS 2020 - Proceeding: 4th International Conference on Informatics and Computational Sciences, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ICICoS51170.2020.9299105.

[14] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, Aug. 2020, doi: https://doi.org/10.1007/s10462-020-09896-5.

[15] C. Guo, C. Liu, and W. Yang, "Iterative missing value imputation based on feature importance," arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.08005.

[16] K. M. Fouad, M. M. Ismail, A. T. Azar, and M. M. Arafa, "Advanced methods for missing values imputation based on similarity learning," PeerJ Computer Science, vol. 7, p. e619, Jul. 2021, doi: https://doi.org/10.7717/peerj-cs.619.

[17] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," BMC Medical Research Methodology, vol. 20, no. 1, Jul. 2020, doi: https://doi.org/10.1186/s12874-020-01080-1.

[18] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," BMC Medical Research Methodology, vol. 20, no. 1, Jul. 2020, doi: https://doi.org/10.1186/s12874-020-01080-1.

[19] Tanapol Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," vol. 16, no. 1, Apr. 2023, doi: https://doi.org/10.1186/s13040-023-00330-4.

[20] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," Information Sciences, vol. 505, pp. 32–64, Dec. 2019, doi: https://doi.org/10.1016/j.ins.2019.07.070.

[21] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," Geoscience Frontiers, vol. 12, no. 1, pp. 469–477, Jan. 2021, doi: https://doi.org/10.1016/j.gsf.2020.03.007.

[22] L. Hansen, Mikkel Stokholm-Bjerregaard, and Petar Durdevic, "Modeling phosphorous dynamics in a wastewater treatment process using Bayesian optimized LSTM," Computers & Chemical Engineering, vol. 160, pp. 107738–107738, Apr. 2022, doi: https://doi.org/10.1016/j.compchemeng.2022.107738.

[23] A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," IEEE Access, vol. 9, pp. 39707–39716, 2021, doi: https://doi.org/10.1109/access.2021.3064084.

[24] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," PLOS ONE, vol. 14, no. 5, p. e0213653, May 2019, doi: https://doi.org/10.1371/journal.pone.0213653.

[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, p. 144, Jan. 2024, doi: https://doi.org/10.3390/diagnostics14020144.

[26] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," IEEE Access, vol. 8, pp. 20991–21002, 2020, doi: https://doi.org/10.1109/ACCESS.2019.2963053.

[27] M. Abood Kadhim and A. M. Radhi, "Heart disease classification using optimized Machine learning algorithms," Iraqi Journal for Computer Science and Mathematics, pp. 31–42, Feb. 2023, doi: https://doi.org/10.52866/ijcsm.2023.02.02.004.

[28] Golovenkin,S.E., Shulman,V.A., Rossiev,D.A., Shesternya,P.A., Nikulina,S.Yu., Orlova,Yu.V., and Voino-Yasenetsky,V.F.. (2020). Myocardial infarction complications. UCI Machine Learning Repository. https://doi.org/10.24432/C53P5M.

[29] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," International Journal of Methods in Psychiatric Research, vol. 20, no. 1, pp. 40–49, Feb. 2011, doi: https://doi.org/10.1002/mpr.329.

[30] A. Fadlil, Herman, and D. Praseptian M, "K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 6, no. 4, pp. 570–576, Aug. 2022, doi: 10.29207/resti.v6i4.4173.

[31] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," BMC Med. Res. Methodol., vol. 20, no. 1, p. 199, 2020, doi: 10.1186/s12874-020-01080-1.

[32] S. Hong, Y. Sun, H. Li, and H. S. Lynn, "Influence of parallel computing strategies of iterative imputation of missing data: a case study on missForest," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.48550/arxiv.2004.11195.

[33] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya, "MICE vs PPCA: Missing data imputation in healthcare," Informatics in Medicine Unlocked, vol. 17, p. 100275, 2019, doi: https://doi.org/10.1016/j.imu.2019.100275.

[34] G. S. Thejas, Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets," Machine Learning with Applications, p. 100267, Jan. 2022, doi: https://doi.org/10.1016/j.mlwa.2022.100267.

[35] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," Scientific Reports, vol. 11, no. 1, p. 24039, Dec. 2021, doi: https://doi.org/10.1038/s41598-021-03430-5.

[36] V. R. Joseph, "Optimal ratio for data splitting," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 15, no. 4, pp. 531–538, Apr. 2022, doi: https://doi.org/10.1002/sam.11583.

[37] R. Roussel et al., "Bayesian optimization algorithms for accelerator physics," Physical Review Accelerators and Beams, vol. 27, no. 8, Aug. 2024, doi: https://doi.org/10.1103/physrevaccelbeams.27.084801.

[38] D. Fitria, T. H. Saragih, D. Kartini, and F. Indriani, "Classification of Appendicitis in Children Using SVM with KNN Imputation and SMOTE Approach to Improve Prediction Quality," J. Electron. Electromed. Eng. Med. Informatics, vol. 6, no. 3, pp. 302–311, 2024, doi: https://doi.org/10.35882/jeeemi.v6i3.470.

[39] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," Data Mining and Knowledge Discovery Handbook, pp. 853–867, 2019, doi: https://doi.org/10.1007/0-387-25465-x_40.

[40] R. de Filippis et al., "Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review," Neuropsychiatric Disease and Treatment, vol. Volume 15, pp. 1605–1627, Jun. 2019, doi: https://doi.org/10.2147/ndt.s202418.

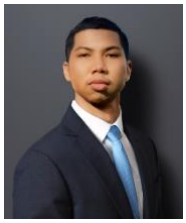
[41] Olya Kudina and Bas de Boer, "Co-designing diagnosis: Towards a responsible integration of Machine Learning decision-support systems in medical diagnostics," Journal of Evaluation in Clinical

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

Practice, vol. 27, no. 3, pp. 529–536, Jan. 2021, doi: https://doi.org/10.1111/jep.13535.
[42] D. Chaerul, E. Saputra, Y. Maulana, T. A. Win, R. Phann, and W. Caesarendra, "Implementation of Machine Learning and Deep Learning Models Based on Structural MRI for Identification of Autism Spectrum Disorder," J. Ilm. Tek. Elektro Komput. dan Inform., vol. 9, no. 2, pp. 307–318, 2023, doi: 10.26555/jiteki.v9i2.26094.
[43] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal: Promoting Communications on Statistics and Stata, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: https://doi.org/10.1177/1536867x20909688.
[44] C. Wang, Y. Zhang, J. Song, Q. Liu, and H. Dong, "A novel optimized SVM algorithm based on PSO with saturation and mixed time-delays for classification of oil pipeline leak detection," Systems Science & Control Engineering, vol. 7, no. 1, pp. 75–88, Jan. 2019, doi: https://doi.org/10.1080/21642583.2019.1573386.
[45] M. Fan, L. Wei, Z. He, W. Wei, and X. Lu, "Defect inspection of solder bumps using the scanning acoustic microscopy and fuzzy SVM algorithm," Microelectronics Reliability, vol. 65, pp. 192–197, Oct. 2016, doi: https://doi.org/10.1016/j.microrel.2016.08.010.
[46] S. Long, X. Huang, Z. Chen, S. Pardhan, and D. Zheng, "Automatic Detection of Hard Exudates in Color Retinal Images Using Dynamic Threshold and SVM Classification: Algorithm Development and Evaluation," BioMed Research International, vol. 2019, pp. 1–13, Jan. 2019, doi: https://doi.org/10.1155/2019/3926930.
[47] Abd Elkarim, I. S., & J. Agbinya,, "A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification," AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES, 2019, doi: https://doi.org/10.22587/ajbas.2019.13.12.10.
[48] Z. Sun, K. Hu, T. Hu, J. Liu, and K. Zhu, "Fast Multi-Label Low-Rank Linearized SVM Classification Algorithm Based on Approximate Extreme Points," IEEE Access, vol. 6, pp. 42319–42326, 2018, doi: https://doi.org/10.1109/access.2018.2854831.
[49] S. Talukdar et al., "Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review," Remote Sensing, vol. 12, no. 7, p. 1135, Apr. 2020, doi: https://doi.org/10.3390/rs12071135.
[50] A. Al-Zebari and A. Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection," IEEE Xplore, Nov. 01, 2019. https://ieeexplore.ieee.org/abstract/document/8965542.
[51] B. A. Khalaf, S. A. Mostafa, A. Mustapha, M. A. Mohammed, and W. M. Abduallah, "Comprehensive Review of Artificial Intelligence and Statistical Approaches in Distributed Denial of Service Attack and Defense Methods," IEEE Access, vol. 7, pp. 51691–51713, 2019, doi: https://doi.org/10.1109/ACCESS.2019.2908998.
[52] Y. Zhang et al., "Research and Application of AdaBoost Algorithm Based on SVM," 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), May 2019, doi: https://doi.org/10.1109/itaic.2019.8785556.
[53] T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785.
[54] Y. Xu, X. Zhao, Y. Chen, and Z. Yang, "Research on a Mixed Gas Classification Algorithm Based on Extreme Random Tree," Applied Sciences, vol. 9, no. 9, p. 1728, Apr. 2019, doi: https://doi.org/10.3390/app9091728.
[55] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interactive Learning Environments, pp. 1–20, May 2021, doi: https://doi.org/10.1080/10494820.2021.1928235.

.

AUTHOR BIOGRAPHY



Ahmad Tajali, a student at Lambung Mangkurat Ahmad Tajali is a dedicated student at Lambung Mangkurat University, where he has been pursuing his education in the Department of Computer Science since 2021. His primary research interest lies in the field of Data Science, encompassing areas such as data analysis, machine learning, and predictive modeling. Ahmad is deeply passionate about exploring innovative methodologies to address real-world problems through data-driven approaches. Beyond academics, he actively engages in collaborative research projects to enhance his practical skills and contribute to the growing body of knowledge in Data Science. Ahmad aims to leverage his expertise to make meaningful contributions to the technological advancements in his field. Email: ahmadtajali61@gmail.com.

Triando Hamonangan Saragih, currently holding the position of a lecturer within the Department of Computer Science at Lambung Mangkurat University, is heavily immersed in the realm of academia, with a profound focus on the multifaceted domain of Data Science. His academic pursuits commenced with the successful completion of his bachelor's degree in Informatics at the esteemed Brawijaya University, located in the vibrant city of Malang, back in the year 2016. Building upon this foundational achievement, he proceeded to further enhance his scholarly credentials by enrolling in a master's program in Computer Science at Brawijaya University, Malang, culminating in the conferral of his advanced degree in 2018. The research field he is involved in is Data Science. Email: triando.saragih@ulm.ac.id. Orcid ID: 0000-0003-4346-3323.

Muhammad Itqan Mazdadi, a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. Before becoming a lecturer, he completed his undergraduate program in the Computer Science Department at Lambung Mangkurat University In 2013. He then completed his master's degree from Department of Informatics at Islamic Indonesia University, Yogyakarta. Currently, he serves as the Secretary of the Computer Science Department at Lambung Mangkurat University. Email: mazdadi@ulm.ac.id. Orcid ID: 0000-0002-8710-4616.

Irwan Budiman, He is a lecturer at Lambung Mangkurat University and currently serves as the Coordinator in the Department of Computer Science, Faculty of Mathematics and Natural Sciences. He earned his Bachelor's Degree in Informatics Engineering from Islam Indonesia University, Yogyakarta. Subsequently, he completed his Master's studies in information systems at Diponegoro University, Semarang. His research interests include data mining, human-computer interaction, applied business intelligence, and e-government. Email: irwan.budiman@ulm.ac.id. Orcid ID: 0000-0002-0514-7429.

Andi Farmadi, a senior lecturer in the Computer Science program at Lambung MangkuratUniversity. He has been teaching since 2008 and currently serves as the Head of the Data Science Lab since 2018. He completed his undergraduate studies at Hasanuddin University and his graduate studies at Bandung Institute of Technology. His research area, up to the present, focuses on Data Science. One of his research projects, along with other researchers, published in the International Conference of Computer

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.

and Informatics Engineering (IC2IE), is titled "Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," and this research was published in 2021. Email: andifarmadi@ulm.ac.id. Orcid ID: 0009-0009-0926-8082.

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.