BUKTI KORESPONDENSI
ARTIKEL JURNAL NASIONAL TERAKREDITASI

Judul Artikel : The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning

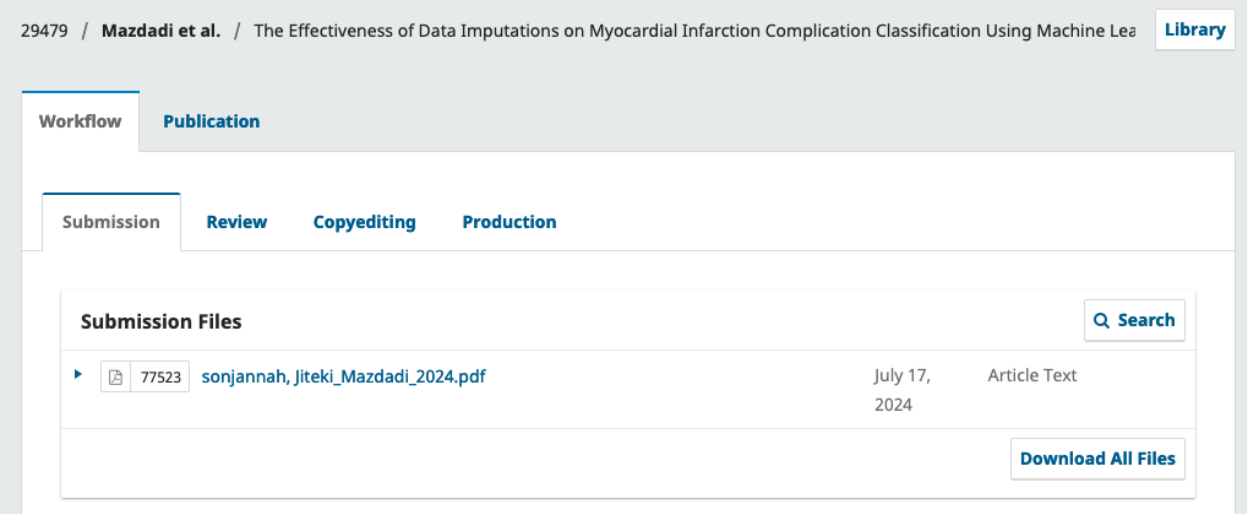
Jurnal : Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, Vol. 10 No. 3 (2024): September

Penulis : Muhammad Itqan Mazdadi, Triando H Saragih, Irwan Budiman, Andi Farmadi, Ahmad Tajali

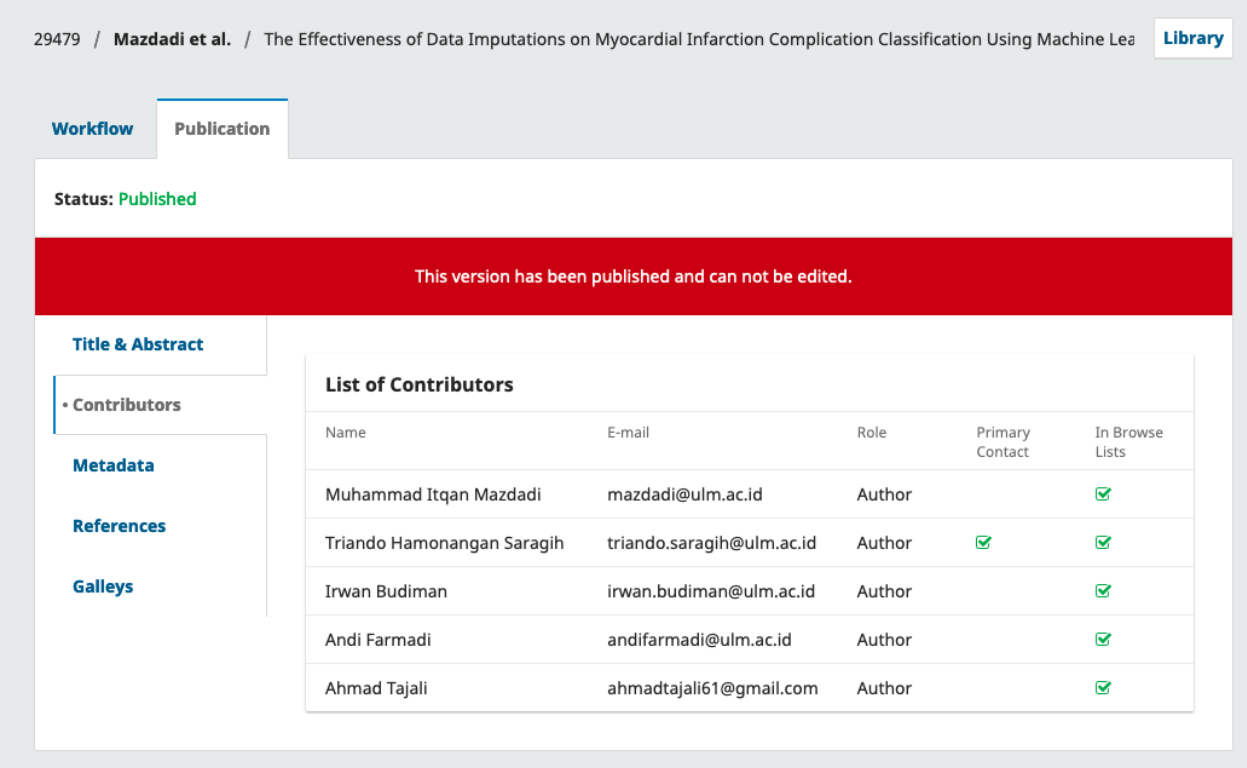
Link : https://journal.uad.ac.id/index.php/JITEKI/article/view/29479

No.	Perihal	Tanggal
1	Submit	17 Juli 2024
2	Notifikasi dan Review untuk Revisi	11 Agustus 2024
3	Submit Revisi	14 Agustus 2024
4	Notifikasi Diterima	2 September 2024
5	Publikasi Terbit di Jurnal	4 September 2024

Bukti Submit



Kontributor



Jurnal yang disubmit

Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 8, No. 2, June 2022, pp. xx-xx
ISSN: 2338-3070, DOI: 10.26555/jiteki.v8i2.xxxx

The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning

Muhammad Itqan Mazdadi[1], Triando Hamonangan Saragih[1], Irwan Budiman[1], Andi Farmadi[1], Ahmad Tajali[1]
[1] Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Indonesia

ARTICLE INFO

Article history:

Received July 28, 2021
Revised August 28, 2021
Published September 28, 2021

Keywords:

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

ABSTRACT

Complications arising from Myocardial Infarction represented a critical medical emergency triggered by the blockage of blood flow to the heart muscle. This scenario arised when the coronary arteries, responsible for delivering blood to the heart, are suddenly blocked, mainly due to a blood clot within a narrowed artery caused by the accumulation of atherosclerotic plaque. Diagnosis involved a thorough assessment including a physical examination, electrocardiogram (ECG) evaluation, blood sample analysis to determine specific heart enzyme levels, and often the use of imaging techniques such as coronary angiography. Proactive anticipation of acute myocardial health complications can help mitigate adverse outcomes. This proactive approach involved early prediction through the application of classification methods. Classification, a key process, entailed organizing objects or data into distinct classes based on identifiable characteristics. Machine learning algorithms like Support Vector Machine (SVM), Random Forest, and XGBoost were employed to ensure accurate predictive outcomes from patient medical records for classification purposes. Techniques such as K-Nearest Neighbors (KNN) imputation, Iterative imputation, and Miss Forest played a crucial role in managing incomplete datasets to avoid losing important information. Optimizing machine learning models' hyperparameters was vital for improving their performance. Bayesian Optimization has become a widely used method for tuning these hyperparameters by creating a model based on past evaluation results to minimize the objective function. Iterative Imputation method showed excellent performance in SVM and XGBoost algorithms for classification tasks. SVM achieved 100% accuracy, precision, sensitivity, F1 test score, and AUC. XGBoost achieved 99.4% accuracy, 100% precision, 79.6% sensitivity, F1 score of 88.7%, and AUC of 0.898. KNN Imputation in SVM yields similar resulted to Iterative Imputation in SVM, while Random Forest showed poor classification outcomes due to data imbalance causing overfitting.

Corresponding Author:

Triando Hamonangan Saragih, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Indonesia

Email: triando.saragih@ulm.ac.id

1. INTRODUCTION

The complications that arise from Myocardial Infarction constitute a critical medical emergency condition precipitated by the obstruction of blood flow to the heart muscle[1]. This occurrence manifests when the coronary arteries responsible for supplying blood to the heart become abruptly blocked, primarily due to the formation of a blood clot within an artery that is already constricted as a result of the accumulation of

atherosclerotic plaque[2]. Consequently, the segment of the heart muscle deprived of an adequate blood supply commences a process of cellular demise owing to the insufficiency of oxygen and essential nutrients. The classic manifestations associated with complications of myocardial infarction encompass intense chest discomfort, difficulty in breathing, as well as feelings of nausea and episodes of vomiting[3]. The diagnostic approach involves a comprehensive evaluation encompassing a physical examination, electrocardiogram (ECG) assessment, analysis of blood samples to ascertain the levels of specific heart enzymes, and frequently, the utilization of imaging modalities such as coronary angiography[4]. Prompt initiation of therapeutic interventions holds paramount importance in mitigating the extent of irreversible cardiac injury and enhancing the overall prognosis of the affected individual. These interventions encompass the prompt administration of thrombolytic agents to dissolve the clot, subsequent implementation of coronary procedures like angioplasty or coronary artery bypass graft surgery, in addition to the implementation of post-treatment care strategies and the institution of long-term therapeutic regimens aimed at averting recurrent incidences of myocardial infarction[5].

One method of averting acute complications related to myocardial health is through the proactive anticipation of such occurrences. Anticipating these complications can be achieved by engaging in early predictive measures. Early prediction, as a fundamental approach, entails the application of classification techniques[6], [7]. Classification, a pivotal procedure, involves the systematic categorization of objects or data into distinct classes or groups based on identifiable characteristics or specific attributes. Across diverse domains, the practice of classification serves as a vital tool for the purpose of structuring information in a coherent manner, facilitating comprehension and enabling thorough analysis[8]. The process of classification can be executed through manual intervention by human operators or through automated means utilizing computational algorithms, particularly in scenarios involving voluminous datasets characterized by intricate complexities. Algorithms that are frequently utilized in the process of making predictions fall under the realm of Machine Learning, a subset of artificial intelligence that focuses on developing systems and algorithms that can learn and improve from experience without being explicitly programmed[9]. These algorithms are designed to analyze data, recognize patterns, and make intelligent decisions or predictions based on the information provided, thus enabling machines to perform tasks or make decisions that would typically require human intervention or intelligence. Machine Learning is a common set of algorithms frequently employed in the process of classification[10].

Machine learning, as a subset of artificial intelligence (AI), is dedicated to the advancement of algorithms and methodologies that empower computers to acquire knowledge and formulate predictions or conclusions by analyzing data. Unlike traditional programming methods, machine learning involves the training of computers through datasets, enabling them to identify patterns and autonomously reach decisions. The field of machine learning is in a state of continuous development, playing a vital role in addressing intricate challenges and enhancing productivity across different industries. Its significance is steadily increasing as it proves to be instrumental in tackling complex problems and streamlining operations in various sectors of the economy. Support Vector Machine[11], [12], Random Forest[6], [13], and Extreme Gradient Boosting (Xgboost)[14], [15] are frequently employed Machine Learning techniques for making classifications.

Support Vector Machine (SVM), Random Forest, and XGBoost are machine learning algorithms utilized to ensure the accuracy of predictive outcomes derived from patient medical records for classification purposes. SVM functions as a supervised learning model that scrutinizes data for classification and regression analysis, recognized for its efficacy in high-dimensional spaces and robust performance in achieving clear margin separation[16], [17]. Random Forest, on the other hand, operates as an ensemble learning technique that generates multiple decision trees during training and outputs the mode of the classes for classification tasks, providing high accuracy and resistance against overfitting[13], [18]. XGBoost, also known as Extreme Gradient Boosting, stands as an optimized distributed gradient boosting library formulated for efficiency, flexibility, and portability, renowned for its superior speed and effectiveness in classification and regression tasks[19]. By employing the Gradient Boosted Decision Tree (GBDT) algorithm framework, XGBoost improves the handling of missing values and provides enhanced regularization techniques to avoid overfitting, making it very effective for extensive data analysis[20], [21].

In conjunction with selecting appropriate machine learning models, managing missing data and class imbalances pose significant challenges in building robust predictive models. Methods for data imputation like K-Nearest Neighbors (KNN) imputation[22], Iterative imputation[23], [24], and Miss Forest[25] are crucial in managing incomplete datasets to prevent the loss of important information from missing values. KNN impute operates by identifying the k-nearest neighbors to a missing value and replacing it based on the mean or mode of these neighbors[26]. Iterative impute, also referred to as Multiple Imputation by Chained Equations (MICE), conducts multiple rounds of imputations, taking into account the uncertainty of missing data by generating

various imputed datasets[27]. Miss Forest, a non-parametric imputation approach, employs random forest algorithms to predict and substitute missing values based on observed data[28].

Despite the progress made in ML and data imputation methodologies, effectively managing missing data continues to pose challenges. Prior research has often concentrated on individual imputation techniques or machine learning models without incorporating advanced strategies for hyperparameter optimization[17]. Additionally, the comparative efficacy of these integrated approaches in forecasting complications related to myocardial infarction has not been thoroughly explored. Optimizing the hyperparameters of machine learning models is essential for enhancing their performance. Bayesian Optimization has emerged as a widely utilized method for tuning machine learning hyperparameters, constructing a surrogate model based on past evaluation outcomes of the target to determine the value that minimizes the objective function. Particularly advantageous for problems with costly (high duration), non-differentiable, or complex function evaluations, Bayesian optimization proves to be highly effective[29].

This study seeks to compare the effectiveness of various machine learning algorithms, including SVM, Random Forest, and XGBoost, in predicting myocardial infarction complications while incorporating data imputation techniques (KNN impute, Iterative impute, and Miss Forest). Additionally, it aims to evaluate the impact of hyperparameter optimization using Bayesian Optimization on predictive accuracy. The integration of these advanced techniques is expected to enhance early detection and management of myocardial infarction complications, thus improving patient outcomes and addressing the gaps in existing literature on predictive approaches.

2. METHODS

This research process requires evaluating the efficacy of three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and XGBoost, each utilizing three distinct data imputation methods, specifically K-Nearest Neighbors (KNN) imputation, Iterative imputation (MICE), and Miss Forest imputation. All the models undergo assessment with hyperparameter adjustment through Bayesian Optimization. This study is split into five successive stages, involving data collection using a dataset on MI complications, data partitioning using k-fold cross-validation, model training, and evaluation of assessment results. The research progression undertaken in this investigation is illustrated in Figure 1.

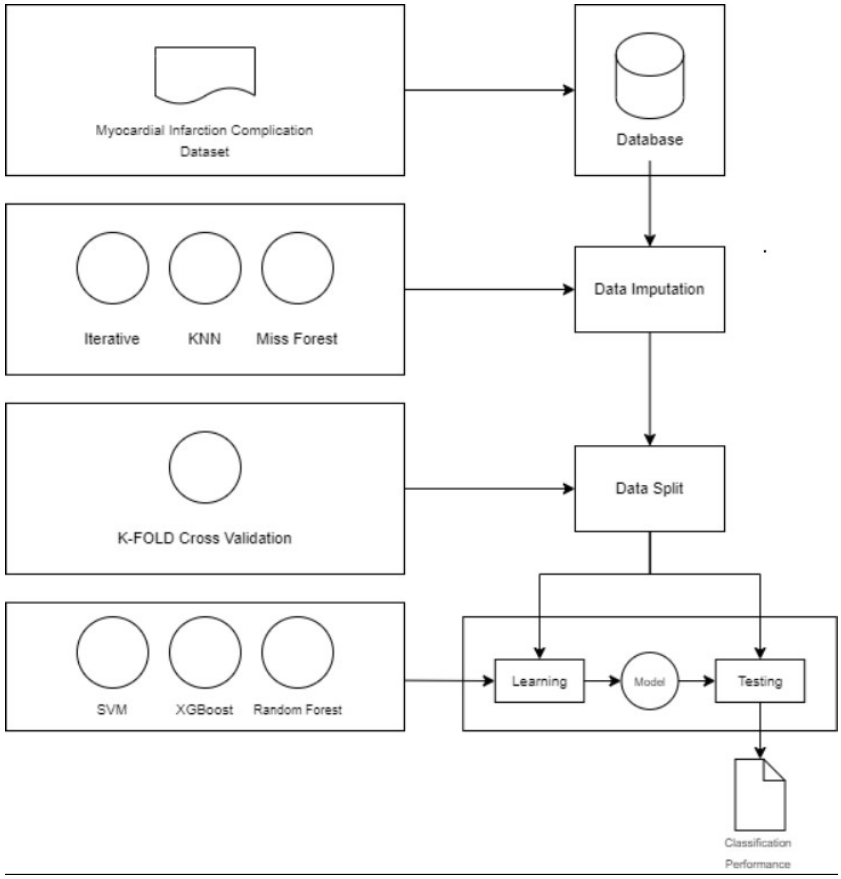


Fig. 1. Research Flowchart

2.1. Data Collection

The dataset analyzed in this research comprises 1700 instances containing 111 attributes related to the medical histories of patients. These attributes cover demographic details, medical background, results of diagnostic tests, and clinical observations during different phases of hospital stay, including admission, first day, second day, and third day. Additionally, the dataset contains annotations for various potential complications of myocardial infarction (MI) like atrial fibrillation, supraventricular tachycardia, ventricular tachycardia, pulmonary edema, and others. The information is obtained from https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications.

2.2. Iterative Imputation

Iterative Imputation is a methodology utilized for managing missing data within datasets. It involves the gradual replacement of missing values through the application of predictive models. This process comprises multiple stages that are executed iteratively until the missing values are resolved and stabilized.

The progression of the iterative imputation approach is structured as follows[30]:

1. Start: Load the dataset containing missing values.
2. Initialization: Substitute the missing values with preliminary estimations such as mean, median, or mode.
3. Iterations:
   - For each specific feature i with missing values:
     - Segment the data into target features (feature i) and predictor features (remaining features).
     - Develop a predictive model (e.g., regression, decision tree, etc.) to forecast the value of feature i.
     - Utilize the model for predicting and replacing the missing values in the features i.
   - Assess convergence:
     - Cease if the imputed values exhibit minimal alteration (convergence).
     - Otherwise, repeat this stage.
4. End: Generate the dataset with the replaced missing values.

2.3. MissForest Imputation

MissForest imputation is an approach that employs the Random Forest algorithm as a non-parametric technique for addressing missing values within a dataset. This method leverages the capabilities of Random Forest in managing intricate and interconnected data to offer precise estimations for the missing values. Random Forest, functioning as an ensemble learning algorithm, merges forecasts from numerous decision trees to enhance precision and mitigate overfitting. MissForest harnesses the potential of Random Forest to anticipate missing values by considering the available dataset values[28].

Similar to iterative imputation, MissForest operates in an iterative manner. During each cycle, the Random Forest model is trained using other features in the dataset to predict the missing values. Through the utilization of MissForest imputation, more accurate estimations can be used to replace missing values in the dataset, thereby enabling a more dependable subsequent analysis and modeling process.

The progression of the MissForest imputation approach is structured as follows[25]:

1. Firstly, the identification of missing values involves determining their location and quantity within the dataset.
2. Subsequently, missing values are filled with initial estimates (such as mean, median, or mode) to initiate the iterative process.
3. The iterative process entails the segmentation of data into target features (specific feature with missing values) and predictor features (other features), followed by training a Random Forest model to forecast the value of the target feature based on the other features. The model is then employed to predict and fill the missing values in the target feature. The convergence is evaluated by assessing the magnitude of change in the imputed values; if minimal (indicating convergence), the process is halted, otherwise, it is repeated.
4. Ultimately, the final imputed outcomes from the iteration are utilized to substitute the missing values in the dataset.

2.4. KNN Imputation

K-Nearest Neighbors (KNN) imputation is a technique employed to address the absence of data values within a dataset by leveraging the principles of KNN. This approach involves replacing the missing values with the average (or mode for categorical variables) of the closest neighbors in the feature space. KNN, a non-parametric algorithm commonly utilized for both classification and regression tasks, is utilized in imputation

to identify a set of k neighboring data points that lack missing values in order to infer and substitute the missing values. In this process, KNN employs a distance metric (such as Euclidean, Manhattan, or Minkowski) to locate the nearest neighbors of data points with missing values. Subsequently, the missing values are imputed with the average (for numerical data) or mode (for categorical data) of these identified nearest neighbors.

Through the application of KNN imputation, the gaps in the dataset can be filled in a manner that leverages the localized similarities between data points, leading to more dependable estimates that align with the prevailing data patterns[22].

The equation of KNN imputation can be seen in (1).

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

This research employs KNN imputation utilizing distance weighting parameters, which have the capability to manage binary, categorical, ordered, continuous, and semi-continuous distance variables. The calculation of the distance between two values involves a weighted mean of the contribution of each variable, with the weights intended to reflect the significance of the respective variable.

2.5. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a machine learning technique employed to categorize a given set of training data along with associated labels. The optimal decision boundary is characterized by having the greatest distance and margin between the two data classes. SVM identifies the most suitable hyperplane for data segregation[17], [31].

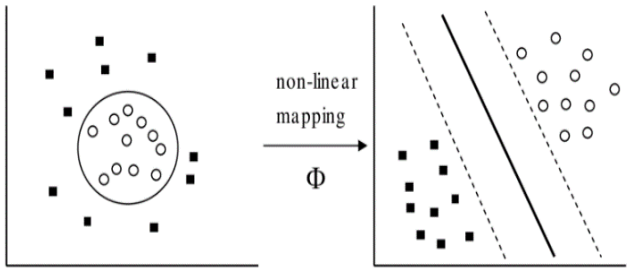


Fig. 2. SVM Model Generation

To effectively divide the data into two distinct linear classes, SVM seeks out the ideal hyperplane by enhancing the separation or margin between the hyperplane and the nearest data points from each class[12], [32].

In this particular investigation, SVM utilized a "kernel" setting of Polynomial with a regulatory parameter denoted as "C" set to 1. The subsequent equation outlines the SVM classification as well as the parameters relevant to the polynomial function.

$$K(x_i, x_j) = (x_i, x_j + c)^d \quad (2)$$

Here, the regulatory parameter is designated as c, while d signifies the polynomial degree, and K(xi, xj) corresponds to the kernel function.

2.6. Random Forest

The Random Forest algorithm is based on the concept of decision-making driven by a sequence of decisions structured in a decision tree format. Several decision trees are developed within the Random Forest structure, with each tree producing its own predictive outcomes. Eventually, the predictive class that receives the highest number of votes is selected as the ultimate prediction. A deeper comprehension of the Random Forest's framework can be attained by analyzing its structure. The architecture of Random Forest can be seen in Figure 3[10], [33].

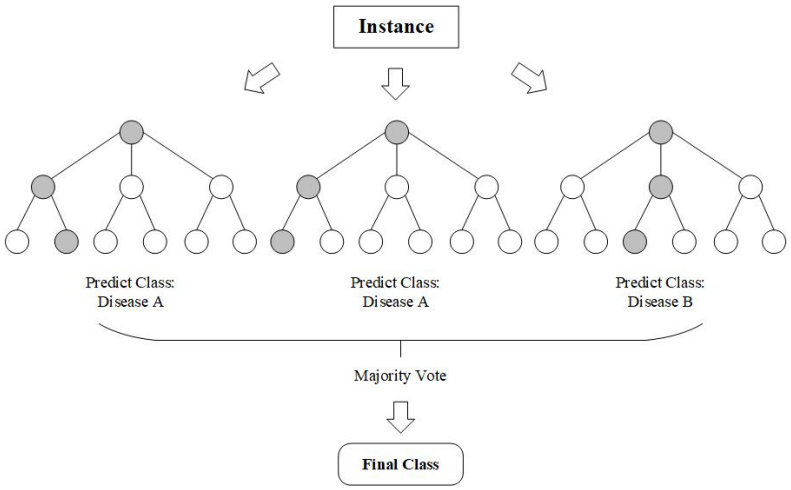


Fig. 3. Random Forest Model Architecture

Two techniques, namely bagging and random subspace, can be utilized for the construction of a Random Forest model. The subsequent section will elaborate on the steps involved in developing a Random Forest model in the field[34], [35]:

1. Utilizing the bootstrapping method to perform random resampling is a strategy that involves employing a sample size identical to that of the training data.
2. The random subspace technique entails selecting K attributes from a set of M attributes, where K is a value less than M, typically corresponding to the square root of M.
3. The development of a decision tree involves using bootstrap samples and previously selected attributes.
4. To attain the desired outcome, it is essential to repeat steps 1 to 3 multiple times in order to shape the tree accordingly. The quantity of trees within the Random Forest model is determined by assessing the out-of-bag error rate (OOB).

2.7. Extreme Gradient Boosting (Xgboost)

The XGBoost principle entails the development of an ensemble-based algorithm that amalgamates ensemble learning and decision trees[36]. When employing the XGBoost method, the concept of ensemble learning plays a crucial role in influencing the training process for the subsequent generation of trees. This influence is manifested in the addition of the residual outcome from the previous training process as a new threshold for the creation of a new tree. Such a process serves to diminish the likelihood of overfitting that may arise from the generation of new trees. Upon reaching the maximum number of iterations, the final output value is designated as the ultimate result. The architecture of XGBoost is visually depicted in Figure 4, showcasing its underlying structure and components[37], [38].

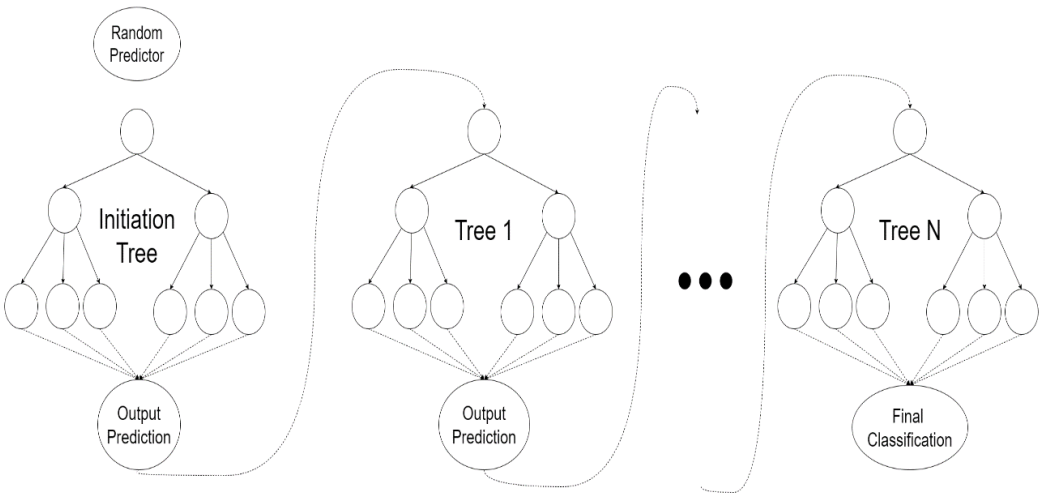


Fig. 4. Extreme Gradient Boosting Model Architecture

A XGBoost model can be created through the process of forming trees and executing an ensemble learning method. The steps involved in developing a XGBoost model include the following[39], [40]:

1. The initialization phase begins by making a prediction for the 0-th tree, which is set to be equal to 0. This initial prediction sets the foundation for the subsequent steps in the model development process.
2. Next, the Splitting Mode is determined by the algorithm, which involves the calculation and traversal of all leaf node gain values until the maximum gain score relative to the root node is obtained. This step is crucial for identifying the optimal splitting points within the tree structure.
3. Following the determination of the Splitting Mode, the current binary leaf node set is established by continuing the calculation process until the gain score becomes negative or another stopping condition is met. This iterative process helps in refining the structure of the tree for better predictive accuracy.
4. Subsequently, the predicted value of the entire leaf node is calculated based on the information gathered from the previous steps. This predicted value serves as the basis for making decisions on how to further optimize the model for better performance.
5. A new tree is then established using the latest prediction result as the threshold, with the condition that the value is greater than the threshold. This process is repeated iteratively until the maximum number of trees specified for the model is reached, ensuring a comprehensive ensemble of trees is created.
6. Finally, the ultimate result of the XGBoost model is determined by calculating the output values of the latest node in the ensemble. This final step brings together the individual predictions of each tree to generate a collective output that represents the overall predictive power of the model.

2.8. Bayesian Optimization

Bayesian Optimization is a method for optimizing objective functions that are unknown and costly to evaluate, based on a probabilistic model. This technique is particularly valuable for tackling optimization challenges where direct assessment of the objective function is time-consuming or expensive, such as hyperparameter tuning in machine learning.

Bayesian Optimization involves several key steps[41]:

1. Prior Model: A probabilistic prior model, typically a Gaussian Process (GP), is established to represent the objective function. The GP is favored for its adaptability in capturing intricate functions and its ability to offer predictive uncertainty.
2. Observation Data: Begin with a small set of initial observation data, including appropriate inputs and outputs. The objective function is assessed at randomly chosen starting points or based on prior knowledge.
3. Construct Surrogate Model: Develop a surrogate model using the available observational data. This model aims to mimic the true objective function and provides a probabilistic approximation of the output.
4. Acquisition Function: Define an acquisition function that utilizes the surrogate model to identify the next point for evaluation. The acquisition function is crafted to balance exploration (exploring less-known regions) and exploitation (exploring areas expected to yield optimal outcomes).
5. Acquisition Function Optimisation: optimizing the acquisition function to determine the next input point for evaluation.
6. Evaluation and Update: Assessing the objective function at the new input point, updating the observation dataset with the new data.
7. Iteration: Iterating through the process from model construction to evaluation and update until a predefined stopping criterion is met, such as a maximum number of iterations or convergence.

2.9. Performance Metrics

In machine learning, the assessment of the combined model's classification performance is typically achieved by employing confusion matrices. These matrices offer a more effective means of displaying outcomes in classification problems, offering insights into both actual and predicted classification results.

Terms such as False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP) are commonly utilized within the context of confusion matrices. The terms are defined in Table 1[12].

Table 1. Confusion Matrix

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

The evaluation matrix under consideration incorporates these confusion matrix parameters to assess each parameter's performance.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F1 = \frac{2 * precision * Recall}{precision + Recal} \tag{6}$$

Utilizing a mathematical formula that combines the curves, the Area Under the Curve (AUC) may be construed as the likelihood that the classification model will accurately differentiate between positive and negative instances. The method of categorization suggests that if chosen randomly, positive instances will yield higher rankings than negative ones. Consequently, an increased AUC signifies an enhanced capability of the classification model in effectively distinguishing between positive and negative categories. The primary objective in crafting an efficient classification model is to maximize the AUC value.

The AUC metric spans from 0 to 1, where a higher AUC denotes superior model performance. AUC can be modeled mathematically in (7).

$$AUC = \frac{\left(\frac{TP}{TP + FN}\right) x \left(\frac{TN}{TN + FP}\right)}{2} \tag{7}$$

Moreover, the AUC value's interpretation reflects the model's competence in distinguishing between positive and negative categories. Furthermore, AUC serves as a valuable instrument for model selection and comparison, enabling practitioners to assess the relative efficacy of different classifiers. The classification quality assessment based on the AUC value is illustrated in Table 2[12].

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 AND DISCUSSION

This section displays the assessment of the SVM, Random Forest, and Xgboost classification algorithms, integrating data imputation methods like KNN Imputation, Iterative Imputation, and MissForest Imputation, as well as Bayesian Optimization for hyperparameter optimization. The objective is to assess the efficacy of classification algorithms in identifying myocardial infarction complications. Different assessment measurements like precision, sensitivity, accuracy, F1-score, and AUC of the ROC curve are applied in this examination. The evaluation aims to compare the performance of machine learning algorithms and gauge the impact of data imputation. In this study, k-fold cross validation is employed for splitting the data due to imbalanced data classes[42].

3.1. Testing Results with K-Fold value 2

This part presents the empirical results derived from the machine learning classification model utilizing a k-fold value of 2.

Table 3. Classification Result using K-Fold value 2

Model	Imputation Method	Peformance Metrics				
		AUC	F1	Accuracy	Sensitivity	Precision
SVM	Iterative	0.589	0.25	0.965	0.188	0.375
	MissForest	0.500	0.00	0.969	0.000	nan
	KNN	0.530	0.111	0.969	0.062	0.500
Random Forest	Iterative	0.500	0.00	0.969	0.000	nan
	MissForest	0.500	0.00	0.969	0.000	nan
	KNN	0.500	0.00	0.969	0.000	nan
Xgboost	Iterative	0.500	0.00	0.969	0.000	nan
	MissForest	0.531	0.118	0.971	0.062	1,000
	KNN	0.531	0.118	0.971	0.062	1,000

Evaluation of the machine learning classification model using a k-fold value of 2 indicates a high level of accuracy. The model's accuracy rate of 97.1% demonstrates its capability in effectively categorizing the data. Nonetheless, the outcomes of additional performance metrics reveal a subpar level of performance. Within the SVM method utilizing iterative imputation, the AUC result reached its peak at 0.589. Subsequently, a further test will be carried out employing a k-fold value of 3. A comparison of performance metrics for all strategies utilized is presented in Figure 5.

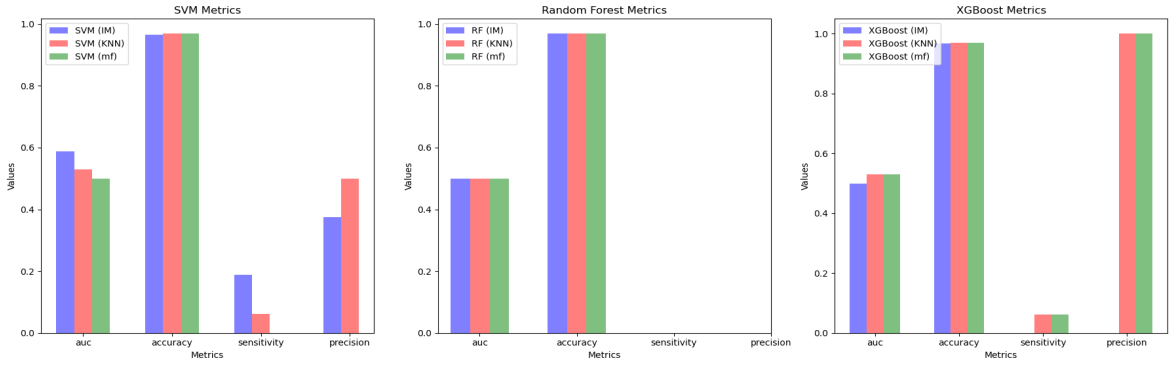


Fig. 5. Comparision of Machine Learning Methods using K-Fold Value 2

3.2. Testing Results with K-Fold value 3

This part presents the empirical results derived from the machine learning classification model utilizing a k-fold value of 3.

Table 4. Classification Result using K-Fold value 3

Model	Imputation Method	Peformance Metrics				
		AUC	F1	Accuracy	Sensitivity	Precision
SVM	Iterative	1.000	1.000	1.000	1.000	1.000
	MissForest	0.712	0.568	0.979	0.426	0.852
	KNN	1.000	1.000	1.000	1.000	1.000
Random Forest	Iterative	0.500	0.000	0.968	0.000	nan
	MissForest	0.500	0.000	0.968	0.000	nan
	KNN	0.500	0.000	0.968	0.000	nan
Xgboost	Iterative	0.898	0.887	0.994	0.796	1.000
	MissForest	0.750	0.667	0.984	0.500	1.000
	KNN	0.722	0.615	0.982	0.444	1.000

Evaluation of machine learning classification models using a k-fold value of 3 yielded positive outcomes across various methodologies. Within the SVM approach, all performance metrics demonstrated optimal outcomes when employing Iterative and KNN imputation techniques. The Xgboost method also exhibited favorable results, achieving a maximum accuracy of 99.4% and an AUC of 0.898, placing it within the good

range. Conversely, the random forest method displayed notably inadequate results despite utilizing a k-fold value of 3. A comparison of performance metrics for all strategies utilized is presented in Figure 6.

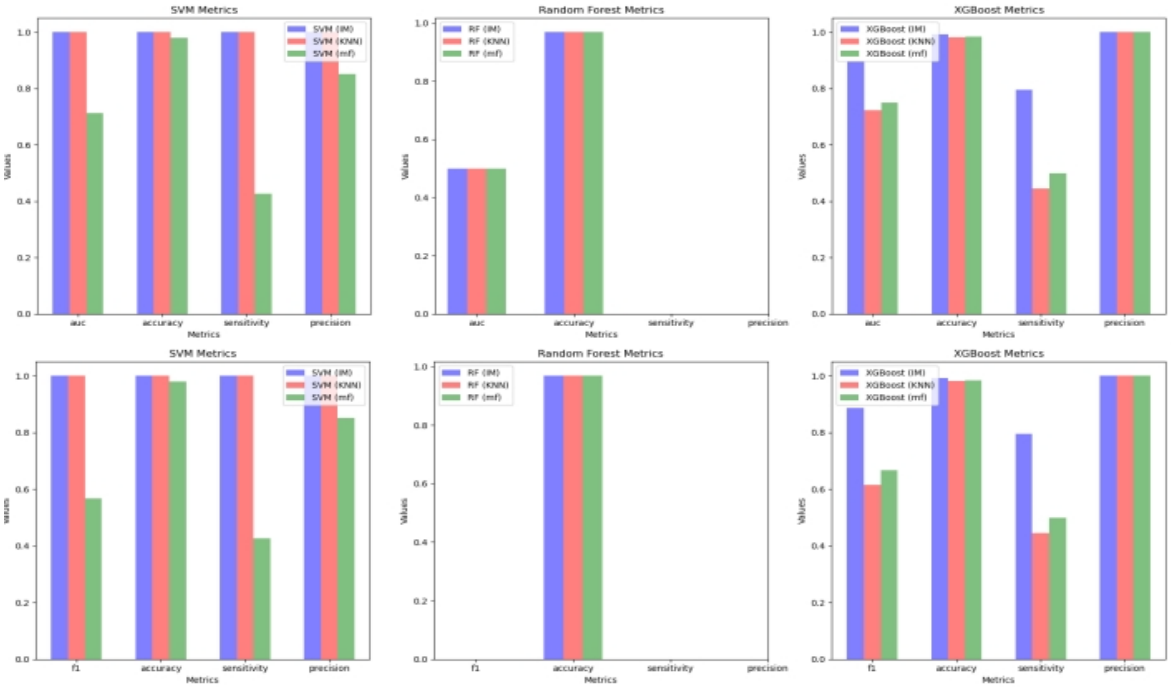


Fig. 6. Comparision of Machine Learning Methods using K-Fold Value 3

3.3. Discussion

The assessment findings indicate that the utilization of imputation techniques proved to be effective in yielding satisfactory outcomes for the SVM and Xgboost algorithms. In the SVM algorithm, exemplary results were achieved in terms of accuracy, precision, sensitivity, and F1 scores of 100%, along with an AUC of 1.00, when employing the iterative and knn imputation techniques with a k-fold of 3. These outcomes demonstrated an enhancement compared to a k-fold of 2. Conversely, in the Xgboost algorithm, optimal outcomes were observed with the iterative imputation technique, showcasing an accuracy of 99.4%, precision of 100%, sensitivity of 79.6%, F1 score of 88.7%, and an AUC of 0.898. The results suggest that Xgboost can yield favorable outcomes when utilizing a k-fold value of 3, overcoming overfitting issues associated with imbalanced data. Nevertheless, the outcomes for Random Forest were found to be unsatisfactory, as indicated by an AUC value of 0.5, signifying its failure in addressing overfitting concerns within the dataset.

Upon comparing the various methodologies applied, it is evident that the Iterative Imputation Method stands out as the most effective approach for handling missing data concerns. Conversely, Random Forest exhibited subpar results due to its AUC value of 0.5, despite achieving high accuracy levels. These results imply that the prevalence of the majority class significantly influences the high accuracy rates through correct classification. Moving forward, additional research is warranted to explore the implementation of data balancing techniques, intended to equalize the representation of minority class data with that of the majority class. By addressing data imbalances, it is possible to mitigate overfitting issues, consequently leading to improved classification outcomes[37], [42].

4. CONCLUSION

According to the findings presented earlier, the Iterative Imputation technique demonstrates superior performance in SVM and Xgboost algorithms for classification tasks. SVM achieves perfect accuracy, precision, sensitivity, F1 test score of 100%, and AUC of 1.00. XGBoost accomplishes 99.4% accuracy, 100% precision, 79.6% sensitivity, F1 score of 88.7%, and AUC of 0.898. Similarly, KNN Imputation in SVM yields identical outcomes to Iterative Imputation in SVM. However, poor classification results are observed with Random Forest due to data class imbalance leading to overfitting.

In forthcoming studies, it is imperative to incorporate class balancing techniques like SMOTE in order to enhance the efficacy of the Random Forest algorithm and to support imputation approaches such as MissForest and KNN Imputation. The utilization of class balancing methods is anticipated to address the issue of overfitting during the classification process.

Acknowledgments

This work was supported in part by LPPM of Lambung Mangkurat University under Contract Number 1090.56/UN8.2/PG/2024. We want to say thank you to Rector of Lambung Mangkurat University and Head of LPPM of Lambung Mangkurat University.

5. REFERENCES

[1] A. A. Damluji *et al.*, "Mechanical Complications of Acute Myocardial Infarction: A Scientific Statement From the American Heart Association," *Circulation*, vol. 144, no. 2, Jul. 2021, doi: 10.1161/CIR.0000000000000985.

[2] N. R. Stephens, C. S. Restrepo, S. S. Saboo, and A. J. Baxi, "Overview of complications of acute and chronic myocardial infarctions: revisiting pathogenesis and cross-sectional imaging," *Postgrad. Med. J.*, vol. 95, no. 1126, pp. 439–450, Aug. 2019, doi: 10.1136/postgradmedj-2018-136279.

[3] N. Dewaswala and R. D. Chait, "A Complication of Acute Myocardial Infarction," *HCA Healthc. J. Med.*, vol. 1, no. 3, Jul. 2020, doi: 10.36518/2689-0216.1054.

[4] J. Zalewski, K. Nowak, P. Furczynska, and M. Zalewska, "Complicating Acute Myocardial Infarction. Current Status and Unresolved Targets for Subsequent Research," *J. Clin. Med.*, vol. 10, no. 24, p. 5904, Dec. 2021, doi: 10.3390/jcm10245904.

[5] E. Moras *et al.*, "Complications in Acute Myocardial Infarction: Navigating Challenges in Diagnosis and Management," *Hearts*, vol. 5, no. 1, pp. 122–141, Mar. 2024, doi: 10.3390/hearts5010009.

[6] C. Browne *et al.*, "Multivariate random forest prediction of poverty and malnutrition prevalence," *PLoS One*, vol. 16, no. 9 September, pp. 1–23, 2021, doi: 10.1371/journal.pone.0255519.

[7] M. Rostami, K. Berahmand, and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00398-3.

[8] Y. Suzuki, A. Suzuki, S. Nakamura, T. Ishikawa, and A. Kinoshita, "Machine learning model estimating number of COVID-19 infection cases over coming 24 days in every province of South Korea (XGBoost and MultiOutputRegressor)," *medRxiv*, p. 2020.05.10.20097527, Jan. 2020, doi: 10.1101/2020.05.10.20097527.

[9] J. R. Saura, B. R. Herraez, and A. Reyes-Menendez, "Comparing a traditional approach for financial brand communication analysis with a big data analytics technique," *IEEE Access*, vol. 7, pp. 37100–37108, 2019, doi: 10.1109/ACCESS.2019.2905301.

[10] 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.

[11] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[12] 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.

[13] T. H. Saragih, V. N. Wijayaningrum, and M. Haekal, "Jatropha Curcas Disease Identification using Random Forest," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 1, p. 9, Apr. 2021, doi: 10.26555/jiteki.v7i1.20141.

[14] N. Zhai, P. Yao, and X. Zhou, "Multivariate time series forecast in industrial process based on XGBoost and GRU," vol. 2020, no. X, pp. 1397–1400, 2020, doi: 10.1109/ITAIC49862.2020.9338878.

[15] C. Chen *et al.*, "Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier," *Comput. Biol. Med.*, vol. 123, p. 103899, Aug. 2020, doi: 10.1016/j.compbiomed.2020.103899.

[16] J. Zhou, P. Yang, P. Peng, M. Khandelwal, and Y. Qiu, "Performance Evaluation of Rockburst Prediction Based on PSO-SVM, HHO-SVM, and MFO-SVM Hybrid Models," *Mining, Metall. Explor.*, vol. 40, no. 2, pp. 617–635, 2023, doi: 10.1007/s42461-022-00713-x.

[17] H. U, K. V, and A. K, "Breast cancer segmentation using hybrid HHO-CS SVM optimization techniques," *Multimed. Tools Appl.*, 2024, doi: 10.1007/s11042-023-18025-7.

[18] B. Thomas and C. J., "Random forest application on cognitive level classification of E-learning content," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 4, p. 4372, Aug. 2020, doi:

10.11591/ijece.v10i4.pp4372-4380.
[19] J. Wang, Q. Cheng, and Y. Dong, "An XGBoost-based multivariate deep learning framework for stock index futures price forecasting," *Kybernetes*, May 2022, doi: 10.1108/K-12-2021-1289.
[20] M. D. Guillen, J. Aparicio, and M. Esteve, "Gradient tree boosting and the estimation of production frontiers," *Expert Syst. Appl.*, vol. 214, no. April 2022, p. 119134, 2023, doi: 10.1016/j.eswa.2022.119134.
[21] L. Nespoli and V. Medici, "Multivariate Boosted Trees and Applications to Forecasting and Control," no. 2017, 2020, [Online]. Available: http://arxiv.org/abs/2003.03835.
[22] 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.
[23] D. Jarrett, B. Cebere, T. Liu, A. Curth, and M. van der Schaar, "HyperImpute: Generalized Iterative Imputation with Automatic Model Selection," *Proc. Mach. Learn. Res.*, vol. 162, pp. 9916–9937, 2022.
[24] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative Robust Semi-Supervised Missing Data Imputation," *IEEE Access*, vol. 8, pp. 90555–90569, 2020, doi: 10.1109/ACCESS.2020.2994033.
[25] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 53–62, Feb. 2022, doi: 10.21512/comtech.v13i1.7388.
[26] A. R. Ismail, N. Z. Abidin, and M. K. Maen, "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare," *J. Robot. Control*, vol. 3, no. 2, pp. 143–152, Feb. 2022, doi: 10.18196/jrc.v3i2.13133.
[27] V. Nassiri, G. Molenberghs, G. Verbeke, and J. Barbosa-Breda, "Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets," *Am. Stat.*, vol. 74, no. 2, pp. 125–136, Apr. 2020, doi: 10.1080/00031305.2018.1543615.
[28] 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.
[29] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019, doi: 10.11989/JEST.1674-862X.80904120.
[30] H. I. Oberman, S. van Buuren, and G. Vink, "Missing the Point: Non-Convergence in Iterative Imputation Algorithms," no. 1996, pp. 3–5, 2021, doi: 10.5334/dsj-2017-037.Van.
[31] H. Saputra, D. Stiawan, and H. Satria, "Malware Detection in Portable Document Format (PDF) Files with Byte Frequency Distribution (BFD) and Support Vector Machine (SVM)," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 4, pp. 1144–1153, 2023, doi: 10.26555/jiteki.v9i4.27527.
[32] W. J. Bidul, S. Surono, and T. B. Kurniawan, "Comparative Evaluation of Feature Selection Methods for Heart Disease Classification with Support Vector Machine," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 2, pp. 265–278, 2024, doi: 10.26555/jiteki.v10i2.28647.
[33] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," *Reliab. Eng. Syst. Saf.*, vol. 200, p. 106931, 2020, doi: 10.1016/j.ress.2020.106931.
[34] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J. Promot. Commun. Stat. Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.
[35] A. Khairunnisa, K. A. Notodiputro, and B. Sartono, "A Comparative Study of Random Forest and Double Random Forest Models from View Points of Their Interpretability," *Sci. J. Informatics*, vol. 11, no. 1, pp. 207–218, Feb. 2024, doi: 10.15294/sji.v11i1.48721.
[36] M. Dwinanda, N. Satyahadewi, and W. Andani, "CLASSIFICATION OF STUDENT GRADUATION STATUS USING XGBOOST ALGORITHM," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 3 SE-Articles, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1785-1794.
[37] D. Arifah, T. H. Saragih, D. Kartini, M. Muliadi, and M. I. Mazdadi, "Application of SMOTE to Handle Imbalance Class in Deposit Classification Using the Extreme Gradient Boosting Algorithm," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 2, pp. 396–410, 2023, doi: 10.26555/jiteki.v9i2.26155.
[38] D. Pebrianti, H. Kurniawan, L. Bayuaji, and R. Rusdah, "XgBoost Hyper-Parameter Tuning Using Particle Swarm Optimization for Stock Price Forecasting," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 4, pp. 1179–1195, 2023, doi: 10.26555/jiteki.v9i4.27712.
[39] T. H. Saragih, R. Ramadhani, M. I. Mazdadi, and M. Haekal, "Energy Efficiency in Buildings Using Multivariate Extreme Gradient Boosting," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Dec. 2022, pp. 1–5, doi: 10.1109/ICIC56845.2022.10006902.

[40] T. H. Saragih and M. I. Mazdadi, "Comparison of Air Quality Prediction using Random Forest and Gradient Boosting Tree," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, Dec. 2023, pp. 1–5, doi: 10.1109/ICIC60109.2023.10382104.
[41] V. Nguyen, "Bayesian Optimization for Accelerating Hyper-Parameter Tuning," in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Jun. 2019, pp. 302–305, doi: 10.1109/AIKE.2019.00060.
[42] T. R. Mahesh *et al.*, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9005278.

BIOGRAPHY OF AUTHORS

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

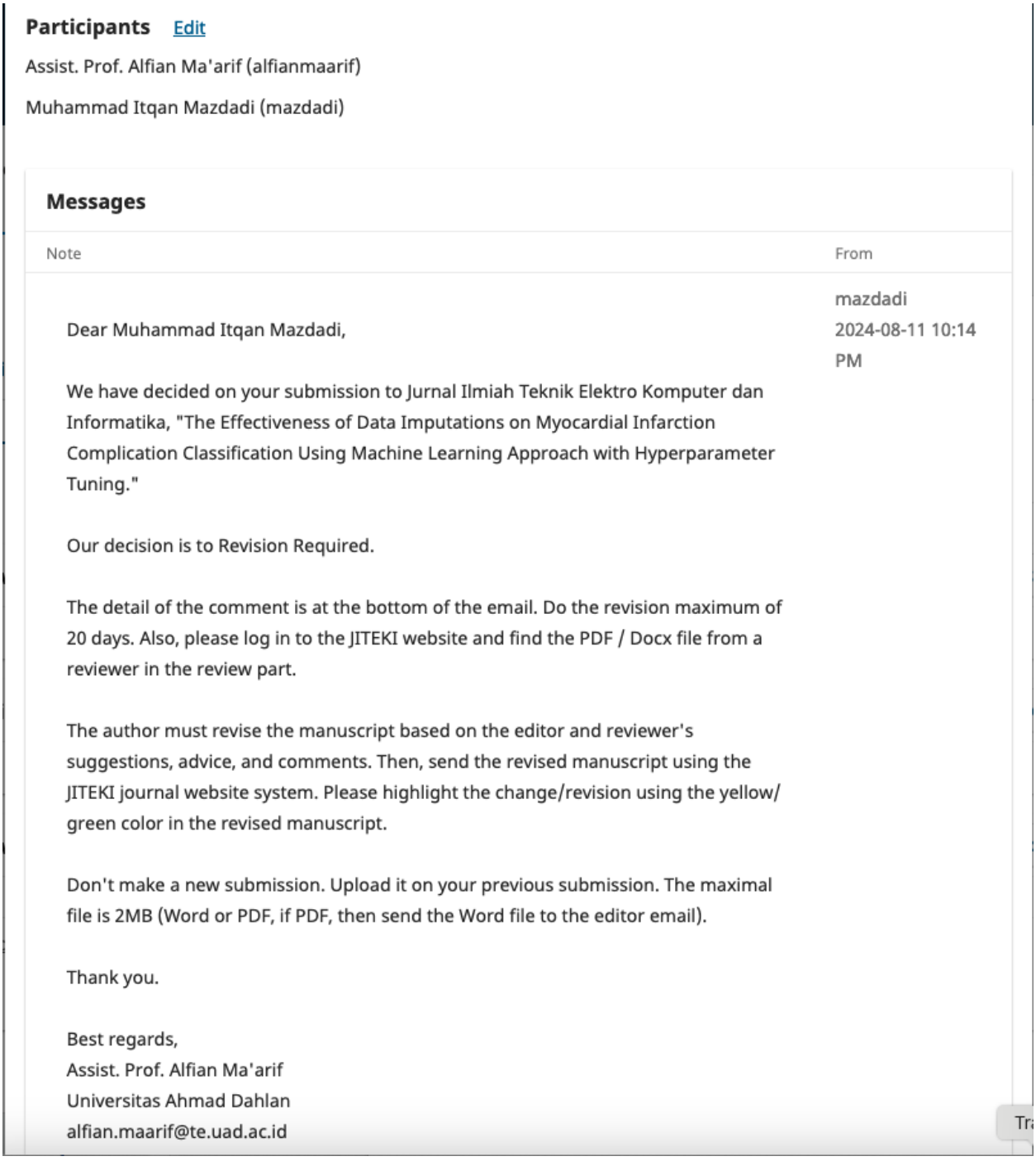
Triando Hamonangan Saragih, currently holding the position of a lecturer within the Department of Computer Science at Lambung Mangkurat University, is heavily immersed in the realm of academia, with a profound focus on the multifaceted domain of Data Science. His academic pursuits commenced with the 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

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 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

Ahmad Tajali, a student at Lambung Mangkurat University who began her education in 2021 in the Department of Computer Science. Her current research field is Data Science. Email: ahmadtajali61@gmail.com.

Notifikasi dan Hasil Review



Participants Edit

Assist. Prof. Alfian Ma'arif (alfianmaarif)

Muhammad Itqan Mazdadi (mazdadi)

Messages

Note	From
Dear Muhammad Itqan Mazdadi, We have decided on your submission to Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, "The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning." Our decision is to Revision Required. The detail of the comment is at the bottom of the email. Do the revision maximum of 20 days. Also, please log in to the JITEKI website and find the PDF / Docx file from a reviewer in the review part. The author must revise the manuscript based on the editor and reviewer's suggestions, advice, and comments. Then, send the revised manuscript using the JITEKI journal website system. Please highlight the change/revision using the yellow/ green color in the revised manuscript. Don't make a new submission. Upload it on your previous submission. The maximal file is 2MB (Word or PDF, if PDF, then send the Word file to the editor email). Thank you. Best regards, Assist. Prof. Alfian Ma'arif Universitas Ahmad Dahlan alfian.maarif@te.uad.ac.id	mazdadi 2024-08-11 10:14 PM

Dear Muhammad Itqan Mazdadi,

We have decided on your submission to Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, "The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning."

Our decision is to Revision Required.

The detail of the comment is at the bottom of the email. Do the revision maximum of 20 days. Also, please log in to the JITEKI website and find the PDF / Docx file from a reviewer in the review part.

The author must revise the manuscript based on the editor and reviewer's suggestions, advice, and comments. Then, send the revised manuscript using the JITEKI journal website system. Please highlight the change/revision using the yellow/green color in the revised manuscript.

Don't make a new submission. Upload it on your previous submission. The maximal file is 2MB (Word or PDF, if PDF, then send the Word file to the editor email).

Thank you.

Best regards,
Assist. Prof. Alfian Ma'arif
Universitas Ahmad Dahlan
alfian.maarif@te.uad.ac.id

==========

Comment of Editor:

>> Please state the research contribution at the end of the introduction and abstract, for example, "The research contribution is..." We need at least two research contributions.

The Method contains an explanation of the research method and the proposed method. Add the research flowchart in this section.

In the Results and Discussion section, Please compare the result with similar methods from previous works (including citations) to enhance the research contribution.

The minimal References are 50 from an English Journal published 5 years ago and used IEEE Style.

Use Mendeley, Zotero, or EndNote to make the references.

The Figure must be in PNG format with 600-1200 dpi. Please don't crop or use snipping and print-screen tools. Otherwise, the image resolution will be compromised. The images using vector format are better.

The Tabel must be made using the Insert Table feature, not from the cropping table as an image.

The Equation must be made using Insert Equation. We suggest using 3 columns of the table to help and then making the equation in the middle of the column and the number of the equation on the right-side column. The variable in the equation must be given information.

All figures and tables must be cited in paragraphs before and explained; please give some explanation, information, or analysis. All figures and tables must be given by some analysis in at least one paragraph.

Use Grammarly to check your manuscript. The free Grammarly is enough for fixing some typos and grammar mistakes. Proofreading is recommended to increase the quality of the English language and writing.

Use the template journal and ensure the plagiarism percentage is under 25%, or the manuscript will be rejected.

==========
Comment of Reviewers
>>

Reviewer A:

TITLE AND ABSTRACT - Provide comments and recommendations for the title and abstract.:
Complications from Myocardial Infarction (MI) represent a critical medical emergency caused by the blockage of blood flow to the heart muscle, primarily due to a blood clot in a coronary artery narrowed by atherosclerotic plaque. Diagnosing MI involves physical examination, electrocardiogram (ECG) evaluation, blood sample analysis for specific heart enzyme levels, and imaging techniques such as coronary angiography. Proactively predicting acute myocardial complications can mitigate adverse outcomes, and this study focuses on early prediction using classification methods.

Machine learning algorithms such as Support Vector Machine (SVM), Random Forest, and

XGBoost were employed to classify patient medical records accurately. Techniques like K-Nearest Neighbors (KNN) imputation, Iterative imputation, and Miss Forest were used to handle incomplete datasets, preserving vital information. Hyperparameter optimization, crucial for model performance, was performed using Bayesian Optimization, which minimizes the objective function by modeling past evaluations.

Results demonstrated that the Iterative Imputation method yielded excellent performance with SVM and XGBoost algorithms. SVM achieved 100% accuracy, precision, sensitivity, F1 score, and AUC. XGBoost reached 99.4% accuracy, 100% precision, 79.6% sensitivity, an F1 score of 88.7%, and an AUC of 0.898. KNN Imputation with SVM showed results similar to Iterative Imputation with SVM, while Random Forest exhibited poor classification outcomes due to data imbalance causing overfitting.

INTRODUCTION - Provide comments and recommendations about the state of the art, novelty, and contribution of the research.:
Complications from Myocardial Infarction (MI) constitute a critical medical emergency precipitated by the obstruction of blood flow to the heart muscle. This blockage occurs when coronary arteries, responsible for supplying blood to the heart, become suddenly blocked, primarily due to a blood clot within an artery narrowed by the accumulation of atherosclerotic plaque. Consequently, the segment of the heart muscle deprived of adequate blood supply begins to experience cellular death due to the lack of oxygen and essential nutrients.

Classic symptoms of MI complications include intense chest discomfort, difficulty in breathing, nausea, and vomiting. Diagnosis involves a comprehensive evaluation, including physical examination, electrocardiogram (ECG) assessment, blood sample analysis for specific heart enzymes, and often imaging techniques such as coronary angiography. Prompt therapeutic interventions are crucial to mitigate irreversible cardiac injury and improve patient prognosis. These interventions may include thrombolytic agents to dissolve the clot, coronary procedures like angioplasty or coronary artery bypass graft surgery, and long-term therapeutic regimens to prevent recurrent MI incidents.

Proactive anticipation of acute myocardial complications can significantly improve outcomes. Early prediction can be achieved using classification techniques, which involve organizing data into distinct classes based on identifiable characteristics. Classification is a vital tool across various domains, structuring information coherently and enabling thorough analysis. This can be executed manually or through automated means using computational algorithms, particularly for large and complex datasets.

Machine Learning (ML), a subset of artificial intelligence, focuses on developing algorithms that enable computers to learn from data, recognize patterns, and make intelligent decisions. ML algorithms such as Support Vector Machine (SVM), Random Forest, and XGBoost are commonly used for classification tasks, including the prediction of MI complications.

Support Vector Machine (SVM): A supervised learning model that excels in high-dimensional spaces and is effective for classification and regression analysis.
Random Forest: An ensemble learning technique that generates multiple decision trees during training and outputs the mode of the classes for classification tasks, providing high accuracy and resistance to overfitting.
XGBoost: An optimized distributed gradient boosting library known for its superior speed and effectiveness in classification and regression tasks, handling missing values efficiently and offering enhanced regularization techniques.
Managing missing data and class imbalances are significant challenges in building robust predictive models. Techniques such as K-Nearest Neighbors (KNN) imputation, Iterative imputation, and Miss Forest are crucial for handling incomplete datasets and preserving important information.

KNN Imputation: Replaces missing values based on the mean or mode of the k-nearest neighbors.
Iterative Imputation: Also known as Multiple Imputation by Chained Equations (MICE), generates multiple imputed datasets, considering the uncertainty of missing data.
Miss Forest: A non-parametric approach using random forest algorithms to predict and substitute missing values based on observed data.

METHOD - Provide comments and recommendations for the method.:
The methods section is comprehensive and covers the essential aspects of the research. However, it could benefit from some restructuring for clarity and conciseness. Here are some detailed comments and suggestions:

General Suggestions:
Flow and Organization:

Reorganize the sections to maintain a logical flow. For instance, start with data collection, followed by data preprocessing (imputation methods), and then proceed to the machine learning models and evaluation techniques.
Consider using subheadings consistently to improve readability.
Detail Level:

Ensure that the descriptions of methods are concise yet detailed enough for reproducibility.
Where possible, avoid overly technical language and explain terms clearly for a broader audience.
Consistency:

Use consistent terminology and formatting throughout the section. For instance, if you use “KNN” in one place, avoid using “k-nearest neighbors” elsewhere unless defining it.
Specific Section Suggestions:

Data Collection:

This section is clear and well-defined.
Consider providing a brief description of the dataset source (UCI Machine Learning Repository) to contextualize the data.
Imputation Methods:

Merge the descriptions of iterative imputation, MissForest, and KNN imputation into a single “Data Imputation Methods” section.
Simplify the iterative process description to avoid redundancy. For instance, outline the common steps and then briefly mention the unique aspects of each method.
Machine Learning Models:

Separate the descriptions of the machine learning models (SVM, Random Forest, XGBoost) into distinct subsections.
Simplify the descriptions and focus on how each model is applied in the study rather than explaining basic principles that are widely known.
Hyperparameter Optimization:

The description of Bayesian Optimization is clear but could be shortened. Focus on how it’s applied in your study rather than a detailed explanation of the method itself.
Performance Metrics:

This section is well-detailed. Ensure that each metric is clearly defined and its relevance to the study is explained.
Consider providing a brief example or explanation for terms like TP, TN, FP, and FN if the target audience might not be familiar with them.

RESULTS AND DISCUSSION - Provide comments and recommendations for the research results and discussion.:
Summary of Findings
The results section provides a detailed analysis of the performance of the SVM, Random Forest, and XGBoost classification algorithms, each coupled with different data imputation methods (KNN, Iterative, and MissForest) and hyperparameter optimization via Bayesian Optimization. The metrics used for evaluation include accuracy, precision, sensitivity, F1-score, and AUC of the ROC curve. The analysis is conducted using k-fold cross-validation with k-values of 2 and 3.

Observations and Critique
High Accuracy but Low AUC and F1-scores:

In the k-fold value of 2, the models showed high accuracy (~97%) but low AUC and F1-scores. This discrepancy suggests that the models might be overfitting or that the dataset is highly imbalanced, leading to misleading accuracy metrics. High accuracy in

imbalanced datasets often indicates that the model is correctly predicting the majority class but failing to predict the minority class.

Improved Performance with k=3:

The results significantly improved with a k-fold value of 3, particularly for the SVM and XGBoost models. This improvement highlights the importance of choosing an appropriate value for k in cross-validation to obtain a more reliable performance assessment.

Effectiveness of Imputation Techniques:

Iterative and KNN imputation methods yielded superior results compared to MissForest, especially when paired with the SVM and XGBoost algorithms. This indicates that these imputation methods may be better suited for this specific dataset.

Underperformance of Random Forest:

The Random Forest algorithm consistently performed poorly, with an AUC of 0.5 across different imputation methods and k-values. This suggests that Random Forest may not be suitable for this particular task, or it might require further tuning or preprocessing adjustments.

Potential Overfitting:

The perfect scores (100%) observed in the SVM with iterative and KNN imputation for k=3 might indicate overfitting. It would be beneficial to investigate this further by using additional evaluation metrics or validation techniques.

Suggestions for Improvement

Addressing Class Imbalance:

Implement techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling) to balance the dataset before training the models. This can help improve the model's performance on minority classes and provide a more accurate evaluation of its efficacy.

Ensemble Methods:

Consider using ensemble methods that combine predictions from multiple models to improve robustness and performance. Techniques like stacking, blending, or voting can help leverage the strengths of different models.

Hyperparameter Tuning:

While Bayesian Optimization was used for hyperparameter tuning, further exploration with other optimization techniques such as Grid Search or Random Search might uncover better hyperparameter configurations.

Model Interpretability:

Incorporate model interpretability techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to understand the

contribution of different features to the model's predictions. This can provide insights into why certain models perform better than others.
Cross-Validation Strategy:

Experiment with different values of k in k-fold cross-validation to determine the most reliable k-value for this dataset. Additionally, consider using stratified k-fold cross-validation to ensure each fold is representative of the overall class distribution.
Advanced Imputation Techniques:

Explore advanced imputation techniques such as autoencoders or GANs (Generative Adversarial Networks) for missing data imputation. These methods can capture complex patterns in the data and potentially provide better imputations.
Additional Metrics:

Include more evaluation metrics such as Matthews correlation coefficient (MCC) or balanced accuracy, which are more informative for imbalanced datasets. These metrics can provide a more nuanced understanding of model performance.
Visualization:

Enhance the visualization of results by including more detailed plots, such as precision-recall curves, to complement ROC curves. These visual aids can help in better understanding the trade-offs between different performance metrics.

CONCLUSIONS - Provide comments and recommendations for conclusions.:
Emphasis on Best Performers:

The conclusion correctly emphasizes the superior performance of the Iterative Imputation technique with SVM and XGBoost. This provides clear direction on the most effective methods identified in the study.
Acknowledgment of Random Forest’s Poor Performance:

The conclusion acknowledges the poor performance of the Random Forest algorithm due to class imbalance, which is important for transparency and directing future research efforts.
Need for Class Balancing Techniques:

The recommendation to incorporate class balancing techniques such as SMOTE is well-founded, given the observed issues with class imbalance and overfitting.
Suggestions for Improvement
Highlighting the Importance of Validation:

Mention the importance of further validation, particularly to ensure that the perfect scores observed with SVM are not due to overfitting. This could involve additional validation techniques or external validation datasets.

Future Research Directions:

While the need for class balancing techniques is addressed, it would be beneficial to outline more specific future research directions. For instance, suggesting the exploration of other advanced imputation methods, or hybrid models that combine multiple algorithms, could provide a clearer path forward.
Broader Implications:

Consider discussing the broader implications of these findings. For example, how might these results impact the practical implementation of machine learning models in medical diagnostics or other real-world applications?
Addressing Limitations:

Briefly mention any limitations of the study that should be addressed in future research. This might include the need for larger or more diverse datasets, or the exploration of other machine learning algorithms that were not considered in this study.
Detailed Recommendations:

Provide more detailed recommendations on how to implement the suggested class balancing techniques. For instance, briefly explain how SMOTE or other techniques could be integrated into the existing workflow

REFERENCES - Provide comments and recommendations for reference.:
-

DECISION RECOMMENDATION - Choose the decision recommendations.:
Minor Revision
Major Revision

GENERAL COMMENT - Provide overall comments and recommendations of the paper. Give feedback for the paper's structure, conformity to the template, clarity, grammatical construction, etc.:
- Cite all Figures and Tables in the paragraph and explain them.
- Please adjust/modify the paper following the article journal template. See the PDF published paper.

Reviewer B:

TITLE AND ABSTRACT - Provide comments and recommendations for the title and abstract.:
The maximum number of keywords permitted is five.

Title Clarity and Brevity: The title is verbose and somewhat cumbersome, making it difficult to quickly grasp the focus of the study. Simplifying and condensing the title while retaining the key elements would improve readability and impact. Additionally, the term "Effectiveness of Data Imputations" is vague; specifying the type of effectiveness (e.g., accuracy improvement, error reduction) would be clearer.

Abstract Structure: The abstract is dense and lacks a clear structure that separates background, methods, results, and conclusions. A more organized abstract with distinct sections would enhance readability and ensure that key information is easily accessible. Currently, the abstract reads as a continuous block of text, which can be overwhelming for the reader.

Technical Jargon and Readability: The abstract contains technical jargon that may not be easily understood by readers unfamiliar with the domain. Terms like "Bayesian Optimization," "Iterative Imputation," and "AUC" should be briefly explained or simplified to improve accessibility. Balancing technical details with readability is crucial for a broader audience.

Lack of Contextual Background: The abstract jumps directly into the complications of myocardial infarction without providing adequate background on the significance and prevalence of the issue. A brief context regarding the importance of predicting myocardial infarction complications would set the stage for the study's relevance and urgency.

Insufficient Explanation of Methods: The abstract mentions several machine learning algorithms and imputation techniques but lacks detailed explanation of their implementation and significance. Providing a concise overview of how these methods were applied and why they were chosen would clarify their role in the study.

Inconsistent Use of Terminology: The abstract inconsistently uses terms such as "data imputation" and "imputation methods," which can be confusing. Consistent terminology throughout the abstract would improve coherence and understanding.

Overemphasis on Specific Metrics: While the abstract highlights specific performance metrics for SVM and XGBoost, it does not adequately explain the implications of these results or compare them meaningfully. Additionally, the presentation of 100% accuracy and other perfect scores for SVM may raise skepticism without a discussion of potential overfitting or validation methods used.

Omission of Limitations and Challenges: The abstract does not address any limitations or challenges encountered during the study, such as issues with data imbalance or overfitting in Random Forest. Acknowledging these aspects would provide a more balanced and transparent view of the study's findings.

Keywords Selection: The keywords listed are relevant but could be expanded to include terms like "hyperparameter tuning" and "classification methods" to better capture the scope of the study. Additionally, more specific keywords related to the dataset or study context could improve searchability.

Absence of Practical Implications: The abstract fails to discuss the practical implications of the findings, such as how the results could be applied in clinical settings or inform future research. Highlighting the real-world impact of the study would enhance its significance and appeal to practitioners and policymakers.

INTRODUCTION - Provide comments and recommendations about the state of the art, novelty, and contribution of the research.:
Excessive Detail in Background Information: The introduction provides an extensive background on myocardial infarction and its complications, including detailed descriptions of diagnostic and treatment methods. While this information is relevant, it could be more concise. The overemphasis on medical details detracts from the focus on the study's objectives and the significance of machine learning techniques in predictive analytics.

Redundancy and Lack of Focus: The introduction repeatedly mentions the role of classification and machine learning in predictive tasks without sufficiently connecting these aspects to the specific research problem. This redundancy can make it challenging for readers to discern the primary focus of the study, which is the integration of machine learning and data imputation for myocardial infarction complications.

Insufficient Justification for Method Selection: Although the introduction discusses various machine learning algorithms and data imputation techniques, it lacks a clear justification for why these specific methods were chosen for this study. An explanation of why SVM, Random Forest, and XGBoost were selected over other algorithms, and why the particular imputation methods were chosen, would provide better context and strengthen the rationale.

Limited Discussion of Hyperparameter Optimization: The introduction briefly mentions hyperparameter optimization using Bayesian Optimization but does not delve into its specific relevance or how it will be integrated into the study. A more detailed explanation of how hyperparameter optimization will be applied and its expected impact on model performance would enhance the reader's understanding.

Weak Transition to Research Objectives: The transition from background information to the research objectives is abrupt. The introduction provides substantial information on the general use of machine learning and data imputation techniques but does not smoothly transition into how these techniques will specifically address the research gap related to myocardial infarction complications.

Lack of Clear Research Gap: While the introduction outlines the general challenges of data

imputation and machine learning in predictive modeling, it does not clearly articulate the specific research gap that this study aims to fill. Defining the precise gap in the literature or practical application that this research addresses would clarify the study's relevance.

Inadequate Literature Context: The introduction references prior research on imputation techniques and machine learning but does not adequately situate the study within the existing body of literature. Providing a more thorough review of recent studies, including their limitations and how this research will build upon or differ from them, would better contextualize the study.

Ambiguity in Study Contribution: The introduction mentions that the study aims to improve predictive accuracy and address gaps in the literature but lacks specificity in describing the expected contributions. Clearly stating the anticipated impact of the research on the field of myocardial infarction prediction and management would provide a stronger rationale for the study.

Insufficient Detail on Methodology Integration: The introduction describes various machine learning algorithms and imputation techniques in isolation but does not detail how these methods will be combined or evaluated together in the study. Explaining how these methodologies will be integrated and assessed collectively would enhance the understanding of the research design.

Overemphasis on Technical Terms: The introduction uses technical terms and jargon extensively, which might alienate readers who are not experts in the field. Simplifying language and ensuring that technical terms are defined or explained would improve accessibility and readability.

METHOD - Provide comments and recommendations for the method.:
Justification for Method Choices: The Methods section describes the choice of algorithms and imputation techniques but lacks a thorough justification for why these specific methods were selected over others. There is no discussion on how these choices were guided by the problem at hand or how they compare to alternative methods, which diminishes the robustness of the methodological rationale.

Details on Hyperparameter Tuning: The use of Bayesian Optimization for hyperparameter tuning is mentioned, yet the section does not detail the specific hyperparameters tuned for each model or the ranges explored. This omission hinders the reproducibility of the study and the understanding of how hyperparameter settings might influence model performance.

Dataset Preprocessing: The manuscript provides an overview of the dataset but fails to address the preprocessing steps undertaken before imputation. Details on handling outliers, categorical variables, or skewed distributions are missing, which are crucial for assessing the appropriateness and effectiveness of the imputation methods used.

Cross-Validation Procedure: Although k-fold cross-validation is used, the manuscript does not specify the value of k or provide a rationale for its choice. The impact of this choice on model evaluation and performance is not discussed, which is essential for understanding the generalizability of the results.

Suitability of Performance Metrics: The performance metrics used for model evaluation are listed, but the choice of these metrics is not well justified in the context of predicting myocardial infarction complications. Metrics such as precision-recall curves might offer more insight in this context, given potential class imbalances, but are not discussed.

Imputation Methods Limitations: The description of imputation methods is thorough, yet there is no discussion on their limitations or potential drawbacks. For instance, KNN imputation is sensitive to the choice of distance metric and k, and MissForest can be computationally intensive. Addressing these issues would provide a more comprehensive view of the methods' applicability.

Overfitting and Model Complexity: The potential for overfitting is not discussed in relation to the complexity of the models used. With complex models like Random Forests and XGBoost, strategies for managing overfitting, such as regularization techniques, are crucial but are not mentioned.

Integration of Imputation and Models: The process of integrating different imputation techniques with machine learning models is not clearly explained. The manuscript lacks details on how imputed datasets are incorporated into model training and evaluation, which affects the clarity of the methodological approach.

Performance Metrics Computation: While performance metrics are defined, the manuscript does not provide details on how these metrics are computed from confusion matrices or the thresholds used for classification. More information on these aspects would enhance the interpretation of the results.

Bayesian Optimization Details: The description of Bayesian Optimization is clear, yet it lacks specific examples of its application to hyperparameter tuning for the models used. Details on the acquisition function and its impact on optimization would offer a deeper understanding of the optimization process.

RESULTS AND DISCUSSION - Provide comments and recommendations for the research results and discussion.:
Inconsistency in Performance Metrics: The Results section presents performance metrics that vary significantly between different k-fold values and imputation methods. For instance, the reported accuracy rates are exceptionally high for k-fold 3 but appear inconsistent or overly optimistic for k-fold 2. This inconsistency raises questions about the reliability of the results, especially since the metrics do not always align with expectations

based on the imputation methods and algorithms used. A more detailed analysis of why these variations occur is necessary to ensure the results' robustness.

Lack of Explanation for High Accuracy Rates: The results indicate exceptionally high accuracy rates, particularly for SVM and Xgboost models with k-fold 3. However, these high accuracy rates are not sufficiently explained or contextualized. Without a clear explanation of why the models perform so well, there is a risk that these results could be due to overfitting or an imbalance in class representation rather than genuine predictive performance. The manuscript should discuss how high accuracy aligns with other performance metrics and address the potential for misleading results due to class imbalance.

Sparse Analysis of Random Forest Performance: The Random Forest algorithm consistently shows subpar results across different k-fold values and imputation methods, with AUC values at 0.5. The discussion lacks a thorough analysis of why Random Forest performs poorly. Possible reasons, such as model parameters, the nature of the data, or specific challenges associated with Random Forest in this context, should be explored in more detail to provide a comprehensive understanding of its underperformance.

Insufficient Exploration of Imputation Methods Impact: The discussion acknowledges the effectiveness of imputation techniques but does not delve deeply into how different imputation methods specifically impact the performance of each algorithm. While Iterative Imputation is highlighted as effective, there is limited discussion on why it outperforms others or how each method's imputation strategy interacts with the models' performance. A more detailed exploration of the imputation methods' impact would help in understanding their role in the results.

Lack of Statistical Significance Testing: The Results section does not mention any statistical significance testing of the performance metrics. Without significance testing, it is challenging to determine whether observed differences between models or k-fold values are statistically meaningful or could be due to random chance. Including statistical tests would strengthen the validity of the comparisons and conclusions drawn.

Inadequate Discussion on Overfitting: Although overfitting is briefly mentioned in the context of Random Forest, there is insufficient discussion on how overfitting was assessed or managed in the study. For instance, the manuscript does not elaborate on how the models' performance was validated or how hyperparameter tuning was optimized to avoid overfitting. More detailed insights into these aspects would enhance the credibility of the findings.

Comparison Figures Lack Detail: Figures 5 and 6, which compare the performance metrics of different methodologies, are referenced but not discussed in detail. The discussion would benefit from a deeper analysis of these figures, including how they illustrate the strengths and weaknesses of each approach and any notable trends or anomalies.

Absence of Practical Implications: The discussion focuses heavily on technical performance metrics but does not address the practical implications of the findings. For example, how do these results translate into practical benefits for identifying myocardial infarction complications? Including a discussion on the practical relevance of the results would provide a more comprehensive understanding of the study's impact.

No Mention of Limitations: The Results and Discussion sections do not address any limitations of the study. A discussion of potential limitations, such as data quality, model assumptions, or the generalizability of the findings, is crucial for providing a balanced view of the results and guiding future research directions.

CONCLUSIONS - Provide comments and recommendations for conclusions.:
Overemphasis on Perfect Metrics: The conclusion highlights that SVM achieved perfect accuracy, precision, sensitivity, F1 score, and AUC of 1.00, and Xgboost achieved high performance metrics. However, it does not address the potential issue of overfitting or the robustness of these metrics. Perfect scores are unusual in practical scenarios and may suggest overfitting or data leakage, which should be acknowledged and addressed to avoid misleading conclusions.

Insufficient Analysis of KNN Imputation Results: The conclusion notes that KNN Imputation yields identical outcomes to Iterative Imputation in SVM but does not provide sufficient analysis or context. It is essential to explain why KNN Imputation performs similarly and whether it has any advantages or disadvantages compared to Iterative Imputation. Without this analysis, the conclusion may appear incomplete and lacking in critical insight.

Generalization of Findings: The conclusion generalizes the effectiveness of Iterative Imputation and its application across different algorithms without acknowledging the limitations or potential variability in other contexts. It should address how these findings might be specific to the dataset used and the implications for generalizability to other datasets or domains.

Lack of Discussion on Random Forest's Poor Performance: The conclusion briefly mentions that Random Forest had poor classification results due to class imbalance leading to overfitting but does not delve into the underlying reasons or propose specific solutions beyond class balancing techniques. A more detailed exploration of why Random Forest underperformed and how it can be effectively improved would provide a more comprehensive conclusion.

Absence of Practical Recommendations: While the conclusion suggests incorporating class balancing techniques like SMOTE in future studies, it does not provide practical recommendations on how these techniques should be implemented or integrated with existing methods. Detailed guidance on applying these techniques would enhance the conclusion's applicability and usefulness for future research.

No Discussion of Limitations: The conclusion fails to address any limitations of the study, such as dataset constraints, methodological limitations, or assumptions made during the analysis. Including a discussion on the study's limitations would provide a more balanced view and help guide future research directions.

Future Research Directions Are Limited: Although the conclusion suggests the use of class balancing techniques, it lacks a broader perspective on other potential future research directions. For example, exploring other imputation methods, comparing different algorithms, or investigating the impact of feature selection on performance could provide a more comprehensive outlook on future work.

No Mention of Implications for Clinical Practice: The conclusion focuses on technical performance but does not address the practical implications for clinical practice or real-world applications. Discussing how the findings can impact myocardial infarction diagnostics or treatment would make the conclusions more relevant and impactful.

REFERENCES - Provide comments and recommendations for reference.:
-

DECISION RECOMMENDATION - Choose the decision recommendations.:
Major Revision

GENERAL COMMENT - Provide overall comments and recommendations of the paper. Give feedback for the paper's structure, conformity to the template, clarity, grammatical construction, etc.:
The maximum number of keywords permitted is five.

Reviewer C:

TITLE AND ABSTRACT - Provide comments and recommendations for the title and abstract.:
- The abstract is not representative of the content and contributions of the paper. The abstract does not seem to convey the rigor of research properly.
- The abstract should contain Objectives, Contributions, Methods/Analysis, Results, Findings, and Novelty /Improvement.
- Abstract must have 150-250 words that consist of 1-2 sentences about the Introduction, problem, and solution; 1-2 sentences about the research contribution (write the research contribution is...); 2-3 sentences about the method; 4-5 sentences about the result; and 1-2 sentences about conclusions.

INTRODUCTION - Provide comments and recommendations about the state of the art, novelty, and contribution of the research.:
- The research contributions of the paper should be articulated more clearly.
- Aside from the aim stated in the title, the research gap and the goals of the research are not specified, which leads to the reader missing the significance of the research.
- The introduction is poorly written and it does not properly refer to previously published studies. The authors need to carefully review the published literature, identify the gaps in the literature, and propose their approach to fill the gap.
- The introduction section must contain the research problem, solution, state of the art, novelty, literature review from previous research, and research contribution (the most important).
- Write the research contribution in the last part of the Introduction, such as “The research contribution is....” At least there are two research contributions.

METHOD - Provide comments and recommendations for the method.:
- Commonly, there are research flow, research diagrams, system block diagrams, control system block diagrams, hardware diagrams, pseudocode, or flowcharts in the method section.
- The figure must be clear, detailed, not blurry, easy to read and provide proper information.
- A flowchart should be added to the article to show the research methodology.
- Much more explanations and interpretations must be added to the method, which is not enough at all.

RESULTS AND DISCUSSION - Provide comments and recommendations for the research results and discussion.:
- In the results section, provide a comparison to a similar method from previous works (including citation) to enhance research contributions (compare the result with previous research).
- All figures and tables must be clear, detailed, not blurry, and easy to read. Each figure and table must be given a comprehensive explanation in at least one paragraph of analysis (crucial).
- The discussion section needs to be described scientifically. Kindly frame it along the following lines:
i. Main findings of the present study
ii. Comparison with other studies
iii. Implication and explanation of findings
iv. Strengths and limitations
- It is suggested to compare the results of the present study with previous studies and analyze their results completely.

CONCLUSIONS - Provide comments and recommendations for conclusions.:
- Add future work so that it can motivate other researchers to continue the research.
- Update the conclusion to include the newly formulated theoretical contributions;
- Mention the limitations of the study and prospects for future research;

- Summarize the key results in a compact form and re-emphasize their significance;
- Summarize how the article contributes to new knowledge in the domain.

REFERENCES - Provide comments and recommendations for reference.:
- The minimal number of references is 50 from Science Direct, IEEE Xplore, Springer Link, MDPI Scilit, or Scopus databases. Cited references must be taken from the journal. Each should have a Digital Object Identifier (DOI) or permanent link. The references were published in the last five years.
- The references must use the IEEE Style.

DECISION RECOMMENDATION - Choose the decision recommendations.:
Major Revision

GENERAL COMMENT - Provide overall comments and recommendations of the paper. Give feedback for the paper's structure, conformity to the template, clarity, grammatical construction, etc.:
- The authors should ask for the help of a native English-speaking proof reader because there are some typos and linguistic mistakes that should be fixed.
- The topic is interesting and important. However, there are several key areas that need more work prior to publication.
- However, in its present form, the manuscript contains several weaknesses. Appropriate revisions to all of the points should be undertaken to justify recommendations for publication.

Jurnal Ilmiah Teknik Elektro Komputer dan Informatika
email: alfianmaarif@ee.uad.ac.id
website: www.journal.uad.ac.id/index.php/JITEKI

Bukti Jurnal yang Diperbaiki

Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 8, No. 2, June 2022, pp. xx-xx
ISSN: 2338-3070, DOI: 10.26555/jiteki.v8i2.xxxx

The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning

Muhammad Itqan Mazdadi[1], Triando Hamonangan Saragih[1], Irwan Budiman[1], Andi Farmadi[1], Ahmad Tajali[1]
[1] Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Indonesia

ARTICLE INFO

Article history:

Received July 28, 2021
Revised August 28, 2021
Published September 28, 2021

Keywords:

Myocardial Infarction;
Machie Learning;
Classification;
Data Imputation;
Bayesian Optimization

ABSTRACT

Complications from Myocardial Infarction (MI) represent a critical medical emergency caused by the blockage of blood flow to the heart muscle, primarily due to a blood clot in a coronary artery narrowed by atherosclerotic plaque. Diagnosing MI involves physical examination, electrocardiogram (ECG) evaluation, blood sample analysis for specific heart enzyme levels, and imaging techniques such as coronary angiography. Proactively predicting acute myocardial complications can mitigate adverse outcomes, and this study focuses on early prediction using classification methods. Machine learning algorithms such as Support Vector Machine (SVM), Random Forest, and XGBoost were employed to classify patient medical records accurately. Techniques like K-Nearest Neighbors (KNN) imputation, Iterative imputation, and Miss Forest were used to handle incomplete datasets, preserving vital information. Hyperparameter optimization, crucial for model performance, was performed using Bayesian Optimization, which minimizes the objective function by modeling past evaluations. The contribution to this study is to see how much influence data imputation has on classification using machine learning methods on missing data and to see how much influence the optimization method has when performing hyperparameter tuning. Results demonstrated that the Iterative Imputation method yielded excellent performance with SVM and XGBoost algorithms. SVM achieved 100% accuracy, precision, sensitivity, F1 score, and AUC. XGBoost reached 99.4% accuracy, 100% precision, 79.6% sensitivity, an F1 score of 88.7%, and an AUC of 0.898. KNN Imputation with SVM showed results similar to Iterative Imputation with SVM, while Random Forest exhibited poor classification outcomes due to data imbalance, causing overfitting.

Commented [TS1]: improvements to the abstract to clarify the research content

Corresponding Author:

Triando Hamonangan Saragih, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Indonesia

Email: triando.saragih@ulm.ac.id

1. INTRODUCTION

Complications from Myocardial Infarction (MI) constitute a critical medical emergency precipitated by the obstruction of blood flow to the heart muscle[1]. This blockage occurs when coronary arteries, responsible for supplying blood to the heart, become suddenly blocked, primarily due to a blood clot within an artery narrowed by the accumulation of atherosclerotic plaque[2]. Consequently, the segment of the heart muscle deprived of adequate blood supply begins to experience cellular death due to the lack of oxygen and essential nutrients[3]. Classic symptoms of MI complications include intense chest discomfort, difficulty in breathing, nausea, and vomiting[4]. Diagnosis involves a comprehensive evaluation, including physical examination, electrocardiogram (ECG) assessment, blood sample analysis for specific heart enzymes, and often imaging techniques such as coronary angiography[5]. Prompt therapeutic interventions are crucial to mitigate

irreversible cardiac injury and improve patient prognosis. These interventions may include thrombolytic agents to dissolve the clot, coronary procedures like angioplasty or coronary artery bypass graft surgery, and long-term therapeutic regimens to prevent recurrent MI incidents. Proactive anticipation of acute myocardial complications can significantly improve outcomes. Early prediction can be achieved using classification techniques, which involve organizing data into distinct classes based on identifiable characteristics. Classification is a vital tool across various domains, structuring information coherently and enabling thorough analysis. This can be executed manually or through automated means using computational algorithms, particularly for large and complex datasets.[6].

Commented [TS2]: A more complete explanation of myocardial infarction complications

One method of averting acute complications related to myocardial health is through the proactive anticipation of such occurrences. Anticipating these complications can be achieved by engaging in early predictive measures. Early prediction, as a fundamental approach, entails the application of classification techniques[7], [8]. Classification, a pivotal procedure, involves the systematic categorization of objects or data into distinct classes or groups based on identifiable characteristics or specific attributes. Across diverse domains, the practice of classification serves as a vital tool for the purpose of structuring information in a coherent manner, facilitating comprehension and enabling thorough analysis[9]. The process of classification can be executed through manual intervention by human operators or through automated means utilizing computational algorithms, particularly in scenarios involving voluminous datasets characterized by intricate complexities. Algorithms that are frequently utilized in the process of making predictions fall under the realm of Machine Learning, a subset of artificial intelligence that focuses on developing systems and algorithms that can learn and improve from experience without being explicitly programmed[10]. These algorithms are designed to analyze data, recognize patterns, and make intelligent decisions or predictions based on the information provided, thus enabling machines to perform tasks or make decisions that would typically require human intervention or intelligence. Machine Learning is a common set of algorithms frequently employed in the process of classification[11].

Machine Learning (ML), a subset of artificial intelligence, focuses on developing algorithms that enable computers to learn from data, recognize patterns, and make intelligent decisions. Unlike traditional programming methods, machine learning involves the training of computers through datasets, enabling them to identify patterns and autonomously reach decisions. The field of machine learning is in a state of continuous development, playing a vital role in addressing intricate challenges and enhancing productivity across different industries. Its significance is steadily increasing as it proves to be instrumental in tackling complex problems and streamlining operations in various sectors of the economy. Support Vector Machine[12], [13], Random Forest[7], [14], and Extreme Gradient Boosting (Xgboost)[15], [16] are frequently employed Machine Learning techniques for making classifications.

Commented [TS3]: Explanation of Machine Learning in the background

Support Vector Machine (SVM), Random Forest, and XGBoost are machine learning algorithms utilized to ensure the accuracy of predictive outcomes derived from patient medical records for classification purposes. SVM functions as a supervised learning model that scrutinizes data for classification and regression analysis, recognized for its efficacy in high-dimensional spaces and robust performance in achieving clear margin separation[17], [18]. Random Forest, on the other hand, operates as an ensemble learning technique that generates multiple decision trees during training and outputs the mode of the classes for classification tasks, providing high accuracy and resistance against overfitting[14], [19]. XGBoost, also known as Extreme Gradient Boosting, stands as an optimized distributed gradient boosting library formulated for efficiency, flexibility, and portability, renowned for its superior speed and effectiveness in classification and regression tasks[20]. By employing the Gradient Boosted Decision Tree (GBDT) algorithm framework, XGBoost improves the handling of missing values and provides enhanced regularization techniques to avoid overfitting, making it very effective for extensive data analysis[21], [22].

In conjunction with selecting appropriate machine learning models, managing missing data and class imbalances pose significant challenges in building robust predictive models. Methods for data imputation like K-Nearest Neighbors (KNN) imputation[23], Iterative imputation[24], [25], and Miss Forest[26] are crucial in managing incomplete datasets to prevent the loss of important information from missing values. KNN impute operates by identifying the k-nearest neighbors to a missing value and replacing it based on the mean or mode of these neighbors[27]. Iterative impute, also referred to as Multiple Imputation by Chained Equations (MICE), conducts multiple rounds of imputations, taking into account the uncertainty of missing data by generating various imputed datasets[28]. Miss Forest, a non-parametric imputation approach, employs random forest algorithms to predict and substitute missing values based on observed data[29].

Despite the progress made in ML and data imputation methodologies, effectively managing missing data continues to pose challenges. Prior research has often concentrated on individual imputation techniques or machine learning models without incorporating advanced strategies for hyperparameter optimization[18].

Additionally, the comparative efficacy of these integrated approaches in forecasting complications related to myocardial infarction has not been thoroughly explored. Optimizing the hyperparameters of machine learning models is essential for enhancing their performance[30]. Bayesian Optimization has emerged as a widely utilized method for tuning machine learning hyperparameters, constructing a surrogate model based on past evaluation outcomes of the target to determine the value that minimizes the objective function[31]. Particularly advantageous for problems with costly (high duration), non-differentiable, or complex function evaluations, Bayesian optimization proves to be highly effective[32], [33].

This study seeks to compare the effectiveness of various machine learning algorithms, including SVM, Random Forest, and XGBoost, in predicting myocardial infarction complications while incorporating data imputation techniques (KNN impute, Iterative impute, and Miss Forest). Additionally, it aims to evaluate the impact of hyperparameter optimization using Bayesian Optimization on predictive accuracy. The integration of these advanced techniques is expected to enhance early detection and management of myocardial infarction complications, thus improving patient outcomes and addressing the gaps in existing literature on predictive approaches. The contribution to this study is to see how much influence data imputation has on classification using machine learning methods on missing data and to see how much influence the optimization method has when performing hyperparameter tuning.

Commented [TS4]: Adding an explanation regarding the contribution to the research conducted

2. METHODS

This research process requires evaluating the efficacy of three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and XGBoost, each utilizing three distinct data imputation methods, specifically K-Nearest Neighbors (KNN) imputation, Iterative imputation (MICE), and Miss Forest imputation. All the models undergo assessment with hyperparameter adjustment through Bayesian Optimization. This study is split into five successive stages, involving data collection using a dataset on MI complications, data partitioning using k-fold cross-validation, model training, and evaluation of assessment results. The research progression undertaken in this investigation is illustrated in Fig. 1.

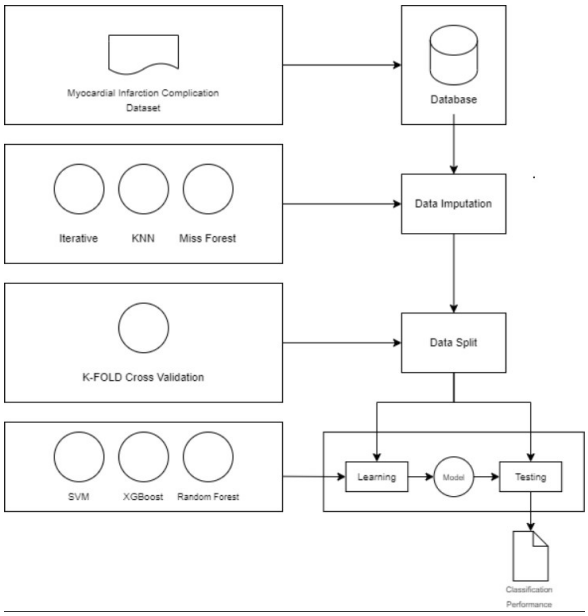


Fig. 1. Research Flowchart

2.1. Data Collection

The dataset analyzed in this research comprises 1700 instances containing 111 attributes related to the medical histories of patients. These attributes cover demographic details, medical background, results of diagnostic tests, and clinical observations during different phases of hospital stay, including admission, first day, second day, and third day. Additionally, the dataset contains annotations for various potential complications of myocardial infarction (MI) like atrial fibrillation, supraventricular tachycardia, ventricular tachycardia, pulmonary edema, and others. The information is obtained from https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications.

2.2. Iterative Imputation

Iterative Imputation is a methodology utilized for managing missing data within datasets. It involves the gradual replacement of missing values through the application of predictive models. This process comprises multiple stages that are executed iteratively until the missing values are resolved and stabilized.

The progression of the iterative imputation approach is structured as follows[34]:

1. Start: Load the dataset containing missing values.
2. Initialization: Substitute the missing values with preliminary estimations such as mean, median, or mode.
3. Iterations:
   - For each specific feature i with missing values:
     - Segment the data into target features (feature i) and predictor features (remaining features).
     - Develop a predictive model (e.g., regression, decision tree, etc.) to forecast the value of feature i.
     - Utilize the model for predicting and replacing the missing values in the features i.
   - Assess convergence:
     - Cease if the imputed values exhibit minimal alteration (convergence).
     - Otherwise, repeat this stage.
4. End: Generate the dataset with the replaced missing values.

2.3. MissForest Imputation

MissForest imputation is an approach that employs the Random Forest algorithm as a non-parametric technique for addressing missing values within a dataset. This method leverages the capabilities of Random Forest in managing intricate and interconnected data to offer precise estimations for the missing values. Random Forest, functioning as an ensemble learning algorithm, merges forecasts from numerous decision trees to enhance precision and mitigate overfitting. MissForest harnesses the potential of Random Forest to anticipate missing values by considering the available dataset values[29].

Similar to iterative imputation, MissForest operates in an iterative manner. During each cycle, the Random Forest model is trained using other features in the dataset to predict the missing values. Through the utilization of MissForest imputation, more accurate estimations can be used to replace missing values in the dataset, thereby enabling a more dependable subsequent analysis and modeling process.

The progression of the MissForest imputation approach is structured as follows[26]:

1. Firstly, the identification of missing values involves determining their location and quantity within the dataset.
2. Subsequently, missing values are filled with initial estimates (such as mean, median, or mode) to initiate the iterative process.
3. The iterative process entails the segmentation of data into target features (specific feature with missing values) and predictor features (other features), followed by training a Random Forest model to forecast the value of the target feature based on the other features. The model is then employed to predict and fill the missing values in the target feature. The convergence is evaluated by assessing the magnitude of change in the imputed values; if minimal (indicating convergence), the process is halted, otherwise, it is repeated.
4. Ultimately, the final imputed outcomes from the iteration are utilized to substitute the missing values in the dataset.

2.4. KNN Imputation

K-Nearest Neighbors (KNN) imputation is a technique employed to address the absence of data values within a dataset by leveraging the principles of KNN. This approach involves replacing the missing values with the average (or mode for categorical variables) of the closest neighbors in the feature space. KNN, a non-parametric algorithm commonly utilized for both classification and regression tasks, is utilized in imputation to identify a set of k neighboring data points that lack missing values in order to infer and substitute the missing

values. In this process, KNN employs a distance metric (such as Euclidean, Manhattan, or Minkowski) to locate the nearest neighbors of data points with missing values. Subsequently, the missing values are imputed with the average (for numerical data) or mode (for categorical data) of these identified nearest neighbors.

Through the application of KNN imputation, the gaps in the dataset can be filled in a manner that leverages the localized similarities between data points, leading to more dependable estimates that align with the prevailing data patterns[23].

The equation of KNN imputation can be seen in (1).

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

This research employs KNN imputation utilizing distance weighting parameters, which have the capability to manage binary, categorical, ordered, continuous, and semi-continuous distance variables. The calculation of the distance between two values involves a weighted mean of the contribution of each variable, with the weights intended to reflect the significance of the respective variable.

2.5. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a machine learning technique employed to categorize a given set of training data along with associated labels. The optimal decision boundary is characterized by having the greatest distance and margin between the two data classes. SVM identifies the most suitable hyperplane for data segregation[18], [35].

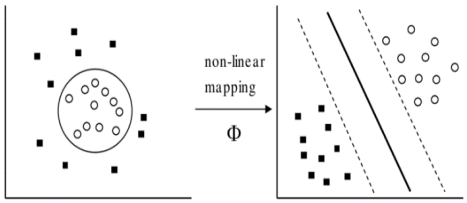


Fig. 2. SVM Model Generation[13]

Based on Fig. 2, to effectively divide the data into two distinct linear classes, SVM seeks out the ideal hyperplane by enhancing the separation or margin between the hyperplane and the nearest data points from each class[13], [36].

In this particular investigation, SVM utilized a "kernel" setting of Polynomial with a regulatory parameter denoted as "C" set to 1. The subsequent equation outlines the SVM classification as well as the parameters relevant to the polynomial function. The equation of SVM kernel can be seen in (2).

$$K(x_i, x_j) = (x_i, x_j + c)^d \quad (2)$$

Here, the regulatory parameter is designated as c, while d signifies the polynomial degree, and K(xi, xj) corresponds to the kernel function.

2.6. Random Forest

The Random Forest algorithm is based on the concept of decision-making driven by a sequence of decisions structured in a decision tree format. Several decision trees are developed within the Random Forest structure, with each tree producing its own predictive outcomes. Eventually, the predictive class that receives the highest number of votes is selected as the ultimate prediction. A deeper comprehension of the Random Forest's framework can be attained by analyzing its structure. The architecture of Random Forest can be seen in Fig. 3 [11], [37].

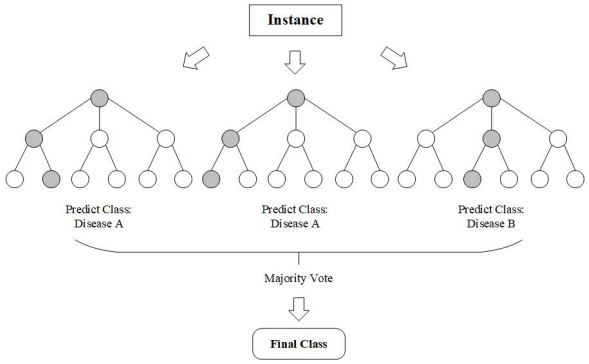


Fig. 3. Random Forest Model Architecture[38]

Two techniques, namely bagging and random subspace, can be utilized for the construction of a Random Forest model. The subsequent section will elaborate on the steps involved in developing a Random Forest model in the field[39], [40]:

1. Utilizing the bootstrapping method to perform random resampling is a strategy that involves employing a sample size identical to that of the training data.
2. The random subspace technique entails selecting K attributes from a set of M attributes, where K is a value less than M, typically corresponding to the square root of M.
3. The development of a decision tree involves using bootstrap samples and previously selected attributes.
4. To attain the desired outcome, it is essential to repeat steps 1 to 3 multiple times in order to shape the tree accordingly. The quantity of trees within the Random Forest model is determined by assessing the out-of-bag error rate (OOB).

2.7. Extreme Gradient Boosting (Xgboost)

The XGBoost principle entails the development of an ensemble-based algorithm that amalgamates ensemble learning and decision trees[41]. When employing the XGBoost method, the concept of ensemble learning plays a crucial role in influencing the training process for the subsequent generation of trees. This influence is manifested in the addition of the residual outcome from the previous training process as a new threshold for the creation of a new tree. Such a process serves to diminish the likelihood of overfitting that may arise from the generation of new trees. Upon reaching the maximum number of iterations, the final output value is designated as the ultimate result. The architecture of XGBoost is visually depicted in Fig. 4, showcasing its underlying structure and components[42], [43].

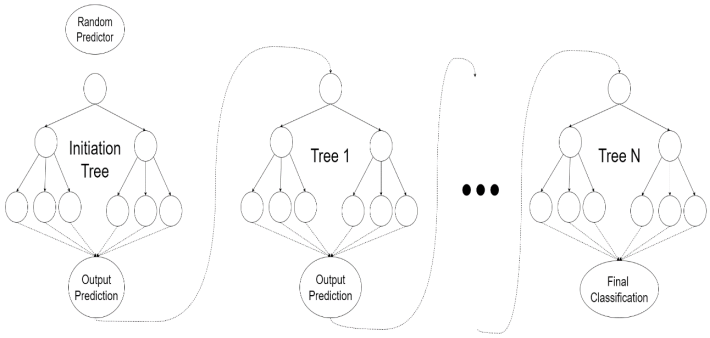


Fig. 4. Extreme Gradient Boosting Model Architecture[44]

A XGBoost model can be created through the process of forming trees and executing an ensemble learning method. The steps involved in developing a XGBoost model include the following[38], [44]:

1. The initialization phase begins by making a prediction for the 0-th tree, which is set to be equal to 0. This initial prediction sets the foundation for the subsequent steps in the model development process.
2. Next, the Splitting Mode is determined by the algorithm, which involves the calculation and traversal of all leaf node gain values until the maximum gain score relative to the root node is obtained. This step is crucial for identifying the optimal splitting points within the tree structure.
3. Following the determination of the Splitting Mode, the current binary leaf node set is established by continuing the calculation process until the gain score becomes negative or another stopping condition is met. This iterative process helps in refining the structure of the tree for better predictive accuracy.
4. Subsequently, the predicted value of the entire leaf node is calculated based on the information gathered from the previous steps. This predicted value serves as the basis for making decisions on how to further optimize the model for better performance.
5. A new tree is then established using the latest prediction result as the threshold, with the condition that the value is greater than the threshold. This process is repeated iteratively until the maximum number of trees specified for the model is reached, ensuring a comprehensive ensemble of trees is created.
6. Finally, the ultimate result of the XGBoost model is determined by calculating the output values of the latest node in the ensemble. This final step brings together the individual predictions of each tree to generate a collective output that represents the overall predictive power of the model.

2.8. Bayesian Optimization

Bayesian Optimization is a method for optimizing objective functions that are unknown and costly to evaluate, based on a probabilistic model. This technique is particularly valuable for tackling optimization challenges where direct assessment of the objective function is time-consuming or expensive, such as hyperparameter tuning in machine learning.

Bayesian Optimization involves several key steps[45]:

1. Prior Model: A probabilistic prior model, typically a Gaussian Process (GP), is established to represent the objective function. The GP is favored for its adaptability in capturing intricate functions and its ability to offer predictive uncertainty.
2. Observation Data: Begin with a small set of initial observation data, including appropriate inputs and outputs. The objective function is assessed at randomly chosen starting points or based on prior knowledge.
3. Construct Surrogate Model: Develop a surrogate model using the available observational data. This model aims to mimic the true objective function and provides a probabilistic approximation of the output.
4. Acquisition Function: Define an acquisition function that utilizes the surrogate model to identify the next point for evaluation. The acquisition function is crafted to balance exploration (exploring less-known regions) and exploitation (exploring areas expected to yield optimal outcomes).
5. Acquisition Function Optimisation: optimizing the acquisition function to determine the next input point for evaluation.
6. Evaluation and Update: Assessing the objective function at the new input point, updating the observation dataset with the new data.
7. Iteration: Iterating through the process from model construction to evaluation and update until a predefined stopping criterion is met, such as a maximum number of iterations or convergence.

2.9. Performance Metrics

In machine learning, the assessment of the combined model's classification performance is typically achieved by employing confusion matrices. These matrices offer a more effective means of displaying outcomes in classification problems, offering insights into both actual and predicted classification results.

Terms such as False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP) are commonly utilized within the context of confusion matrices. True Positive (TP)is the test predicts "positive," and the result is actually positive. True Negative (TN) is the test predicts "negative," and the result is actually negative. False Positive (FP) is the test predicts "positive," but the result is actually negative. False Negative (FN) is the test predicts "negative," but the result is actually positive[46]. The terms are defined in Table 1.

Commented [TS5]: More complete explanation regarding True Positive, True Negative, False Positive, False Negative

Table 1. Confusion Matrix[47]

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

The evaluation matrix under consideration incorporates these confusion matrix parameters to assess each parameter's performance[48].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$F1 = \frac{2 * precision * Recall}{precision + Recal} \quad (6)$$

Utilizing a mathematical formula that combines the curves, the Area Under the Curve (AUC) may be construed as the likelihood that the classification model will accurately differentiate between positive and negative instances. The method of categorization suggests that if chosen randomly, positive instances will yield higher rankings than negative ones. Consequently, an increased AUC signifies an enhanced capability of the classification model in effectively distinguishing between positive and negative categories. The primary objective in crafting an efficient classification model is to maximize the AUC value[49].

The AUC metric spans from 0 to 1, where a higher AUC denotes superior model performance. AUC can be modeled mathematically in (7).

$$AUC = \frac{\left(\frac{TP}{TP + FN}\right) x \left(\frac{TN}{TN + FP}\right)}{2} \quad (7)$$

Moreover, the AUC value's interpretation reflects the model's competence in distinguishing between positive and negative categories. Furthermore, AUC serves as a valuable instrument for model selection and comparison, enabling practitioners to assess the relative efficacy of different classifiers. The classification quality assessment based on the AUC value is illustrated in Table 2[50].

Table 2. Categories of results from classification based on AUC values[50]

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 AND DISCUSSION

The results section provides a detailed analysis of the performance of the SVM, Random Forest, and XGBoost classification algorithms, each coupled with different data imputation methods (KNN, Iterative, and MissForest) and hyperparameter optimization via Bayesian Optimization. The metrics used for evaluation include accuracy, precision, sensitivity, F1-score, and AUC of the ROC curve. The analysis is conducted using

k-fold cross-validation with k-values of 2 and 3. The evaluation aims to compare the performance of machine learning algorithms and gauge the impact of data imputation. In this study, k-fold cross validation is employed for splitting the data due to imbalanced data classes[51].

Commented [TS6]: Explanation of the tests conducted in this study

3.1. Testing Results with K-Fold value 2

This part presents the empirical results derived from the machine learning classification model utilizing a k-fold value of 2.

Table 3. Classification Result using K-Fold value 2

Model	Imputation Method	Peformance Metrics				
		AUC	F1	Accuracy	Sensitivity	Precision
SVM	Iterative	0.589	0.25	0.965	0.188	0.375
	MissForest	0.500	0.00	0.969	0.000	nan
	KNN	0.530	0.111	0.969	0.062	0.500
Random Forest	Iterative	0.500	0.00	0.969	0.000	nan
	MissForest	0.500	0.00	0.969	0.000	nan
	KNN	0.500	0.00	0.969	0.000	nan
Xgboost	Iterative	0.500	0.00	0.969	0.000	nan
	MissForest	0.531	0.118	0.971	0.062	1,000
	KNN	0.531	0.118	0.971	0.062	1,000

Based on Table 1, evaluation of the machine learning classification model using a k-fold value of 2 indicates a high level of accuracy. The model's accuracy rate of 97.1% demonstrates its capability in effectively categorizing the data. Nonetheless, the outcomes of additional performance metrics reveal a subpar level of performance. Within the SVM method utilizing iterative imputation, the AUC result reached its peak at 0.589. Subsequently, a further test will be carried out employing a k-fold value of 3. A comparison of performance metrics for all strategies utilized is presented in Fig. 5.

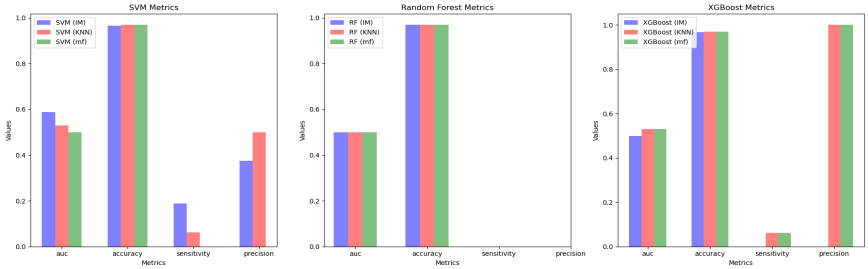


Fig. 5. Comparision of Machine Learning Methods using K-Fold Value 2

3.2. Testing Results with K-Fold value 3

This part presents the empirical results derived from the machine learning classification model utilizing a k-fold value of 3.

Table 4. Classification Result using K-Fold value 3

Model	Imputation Method	Peformance Metrics				
		AUC	F1	Accuracy	Sensitivity	Precision
SVM	Iterative	1.000	1.000	1.000	1.000	1.000
	MissForest	0.712	0.568	0.979	0.426	0.852
	KNN	1.000	1.000	1.000	1.000	1.000
Random Forest	Iterative	0.500	0.000	0.968	0.000	nan
	MissForest	0.500	0.000	0.968	0.000	nan
	KNN	0.500	0.000	0.968	0.000	nan

	Iterative	0.898	0.887	0.994	0.796	1.000
Xgboost	MissForest	0.750	0.667	0.984	0.500	1.000
	KNN	0.722	0.615	0.982	0.444	1.000

Based on Table 4, the results significantly improved with a k-fold value of 3, particularly for the SVM and XGBoost models. This improvement highlights the importance of choosing an appropriate value for k in cross-validation to obtain a more reliable performance assessment. Within the SVM approach, all performance metrics demonstrated optimal outcomes when employing Iterative and KNN imputation techniques. The Xgboost method also exhibited favorable results, achieving a maximum accuracy of 99.4% and an AUC of 0.898, placing it within the good range. The Random Forest algorithm consistently performed poorly, with an AUC of 0.5 across different imputation methods and k-values. This suggests that Random Forest may not be suitable for this particular task, or it might require further tuning or preprocessing adjustments. Iterative and KNN imputation methods yielded superior results compared to MissForest, especially when paired with the SVM and XGBoost algorithms. This indicates that these imputation methods may be better suited for this specific dataset. A comparison of performance metrics for all strategies utilized is presented in Fig. 6.

Commented [TS7]: Additional explanation regarding the results of the 3-fold k-value test

Commented [TS8]: Explanation of the analysis results obtained using the Random Forest method

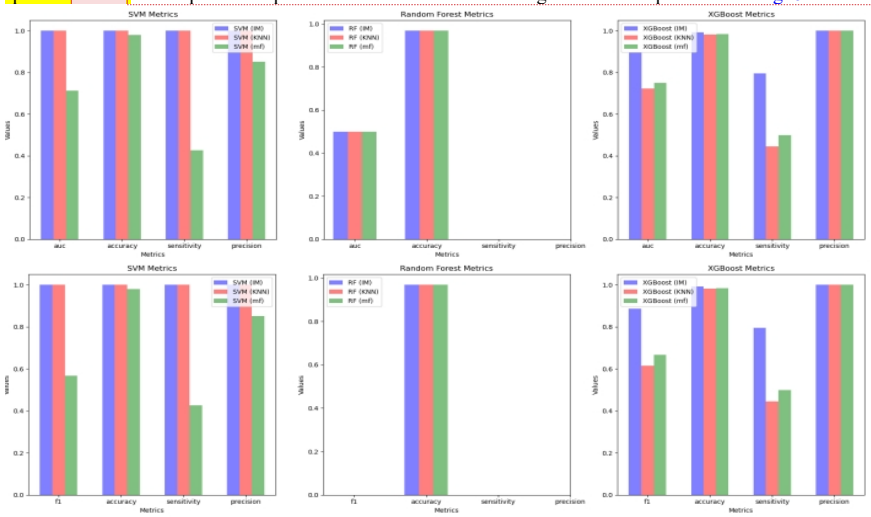


Fig. 6. Comparision of Machine Learning Methods using K-Fold Value 3

3.3. Discussion

The assessment findings indicate that the utilization of imputation techniques proved to be effective in yielding satisfactory outcomes for the SVM and Xgboost algorithms. In the SVM algorithm, exemplary results were achieved in terms of accuracy, precision, sensitivity, and F1 scores of 100%, along with an AUC of 1.00, when employing the iterative and knn imputation techniques with a k-fold of 3. These outcomes demonstrated an enhancement compared to a k-fold of 2. Conversely, in the Xgboost algorithm, optimal outcomes were observed with the iterative imputation technique, showcasing an accuracy of 99.4%, precision of 100%, sensitivity of 79.6%, F1 score of 88.7%, and an AUC of 0.898. The results suggest that Xgboost can yield favorable outcomes when utilizing a k-fold value of 3, overcoming overfitting issues associated with imbalanced data. Nevertheless, the outcomes for Random Forest were found to be unsatisfactory, as indicated by an AUC value of 0.5, signifying its failure in addressing overfitting concerns within the dataset.

Upon comparing the various methodologies applied, it is evident that the Iterative Imputation Method stands out as the most effective approach for handling missing data concerns. Conversely, Random Forest exhibited subpar results due to its AUC value of 0.5, despite achieving high accuracy levels. These results imply that the prevalence of the majority class significantly influences the high accuracy rates through correct classification. The perfect scores (100%) observed in the SVM with iterative and KNN imputation for k=3 might indicate overfitting. It would be beneficial to investigate this further by using additional evaluation

metrics or validation techniques. Moving forward, additional research is warranted to explore the implementation of data balancing techniques, intended to equalize the representation of minority class data with that of the majority class. Implement techniques such as SMOTE (Synthetic Minority Over-sampling Technique)[51] or ADASYN (Adaptive Synthetic Sampling)[52], [53] to balance the dataset before training the models. This can help improve the model's performance on minority classes and provide a more accurate evaluation of its efficacy. While Bayesian Optimization was used for hyperparameter tuning, further exploration with other optimization techniques such as Grid Search[54] or Random Search[55] might uncover better hyperparameter configurations.

Commented [TS9]: Explanation of overfitting SVM results

Commented [TS10]: Solution to handle the problem of SVM overfitting results

4. CONCLUSION

According to the findings presented earlier, the Iterative Imputation technique demonstrates superior performance in SVM and Xgboost algorithms for classification tasks. SVM achieves perfect accuracy, precision, sensitivity, F1 test score of 100%, and AUC of 1.00. XGBoost accomplishes 99.4% accuracy, 100% precision, 79.6% sensitivity, F1 score of 88.7%, and AUC of 0.898. Similarly, KNN Imputation in SVM yields identical outcomes to Iterative Imputation in SVM. However, poor classification results are observed with Random Forest due to data class imbalance leading to overfitting.

In forthcoming studies, it is imperative to incorporate class balancing techniques like SMOTE and ADASYN in order to enhance the efficacy of the Random Forest algorithm and to support imputation approaches such as MissForest and KNN Imputation. The utilization of class balancing methods is anticipated to address the issue of overfitting during the classification process. While Bayesian Optimization was used for hyperparameter tuning, further exploration with other optimization techniques such as Grid Search or Random Search might uncover better hyperparameter configurations.

Commented [TS11]: Making additional suggestions for further research using data imputation

Acknowledgments

This work was supported in part by LPPM of Lambung Mangkurat University under Contract Number 1090.56/UN8.2/PG/2024. We want to say thank you to Rector of Lambung Mangkurat University and Head of LPPM of Lambung Mangkurat University.

5. REFERENCES

[1] A. A. Damluji *et al.*, "Mechanical Complications of Acute Myocardial Infarction: A Scientific Statement From the American Heart Association," *Circulation*, vol. 144, no. 2, Jul. 2021, doi: 10.1161/CIR.0000000000000985.

[2] N. R. Stephens, C. S. Restrepo, S. S. Saboo, and A. J. Baxi, "Overview of complications of acute and chronic myocardial infarctions: revisiting pathogenesis and cross-sectional imaging," *Postgrad. Med. J.*, vol. 95, no. 1126, pp. 439–450, Aug. 2019, doi: 10.1136/postgradmedj-2018-136279.

[3] J. Yang *et al.*, "Current status of emergency medical service use in ST-segment elevation myocardial infarction in China: Findings from China Acute Myocardial Infarction (CAMI) Registry," *Int. J. Cardiol.*, vol. 406, p. 132040, Jul. 2024, doi: 10.1016/j.ijcard.2024.132040.

[4] N. Dewaswala and R. D. Chait, "A Complication of Acute Myocardial Infarction," *HCA Healthc. J. Med.*, vol. 1, no. 3, Jul. 2020, doi: 10.36518/2689-0216.1054.

[5] J. Zalewski, K. Nowak, P. Furczynska, and M. Zalewska, "Complicating Acute Myocardial Infarction. Current Status and Unresolved Targets for Subsequent Research," *J. Clin. Med.*, vol. 10, no. 24, p. 5904, Dec. 2021, doi: 10.3390/jcm10245904.

[6] E. Moras *et al.*, "Complications in Acute Myocardial Infarction: Navigating Challenges in Diagnosis and Management," *Hearts*, vol. 5, no. 1, pp. 122–141, Mar. 2024, doi: 10.3390/hearts5010009.

[7] C. Browne *et al.*, "Multivariate random forest prediction of poverty and malnutrition prevalence," *PLoS One*, vol. 16, no. 9 September, pp. 1–23, 2021, doi: 10.1371/journal.pone.0255519.

[8] M. Rostami, K. Berahmand, and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00398-3.

[9] Y. Suzuki, A. Suzuki, S. Nakamura, T. Ishikawa, and A. Kinoshita, "Machine learning model estimating number of COVID-19 infection cases over coming 24 days in every province of South Korea (XGBoost and MultiOutputRegressor)," *medRxiv*, p. 2020.05.10.20097527, Jan. 2020, doi: 10.1101/2020.05.10.20097527.

[10] J. R. Saura, B. R. Herraez, and A. Reyes-Menendez, "Comparing a traditional approach for financial brand communication analysis with a big data analytics technique," *IEEE Access*, vol. 7, pp. 37100–37108, 2019, doi: 10.1109/ACCESS.2019.2905301.

[11] 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.
[12] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
[13] 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.
[14] T. H. Saragih, V. N. Wijayaningrum, and M. Haekal, "Jatropha Curcas Disease Identification using Random Forest," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 1, p. 9, Apr. 2021, doi: 10.26555/jiteki.v7i1.20141.
[15] N. Zhai, P. Yao, and X. Zhou, "Multivariate time series forecast in industrial process based on XGBoost and GRU," vol. 2020, no. X, pp. 1397–1400, 2020, doi: 10.1109/ITAIC49862.2020.9338878.
[16] C. Chen *et al.*, "Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier," *Comput. Biol. Med.*, vol. 123, p. 103899, Aug. 2020, doi: 10.1016/j.compbiomed.2020.103899.
[17] J. Zhou, P. Yang, P. Peng, M. Khandelwal, and Y. Qiu, "Performance Evaluation of Rockburst Prediction Based on PSO-SVM, HHO-SVM, and MFO-SVM Hybrid Models," *Mining, Metall. Explor.*, vol. 40, no. 2, pp. 617–635, 2023, doi: 10.1007/s42461-022-00713-x.
[18] H. U, K. V, and A. K, "Breast cancer segmentation using hybrid HHO-CS SVM optimization techniques," *Multimed. Tools Appl.*, 2024, doi: 10.1007/s11042-023-18025-7.
[19] B. Thomas and C. J., "Random forest application on cognitive level classification of E-learning content," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 4, p. 4372, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4372-4380.
[20] J. Wang, Q. Cheng, and Y. Dong, "An XGBoost-based multivariate deep learning framework for stock index futures price forecasting," *Kybernetes*, May 2022, doi: 10.1108/K-12-2021-1289.
[21] M. D. Guillen, J. Aparicio, and M. Esteve, "Gradient tree boosting and the estimation of production frontiers," *Expert Syst. Appl.*, vol. 214, no. April 2022, p. 119134, 2023, doi: 10.1016/j.eswa.2022.119134.
[22] L. Nespoli and V. Medici, "Multivariate Boosted Trees and Applications to Forecasting and Control," no. 2017, 2020, [Online]. Available: http://arxiv.org/abs/2003.03835.
[23] 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.
[24] D. Jarrett, B. Cebere, T. Liu, A. Curth, and M. van der Schaar, "HyperImpute: Generalized Iterative Imputation with Automatic Model Selection," *Proc. Mach. Learn. Res.*, vol. 162, pp. 9916–9937, 2022.
[25] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative Robust Semi-Supervised Missing Data Imputation," *IEEE Access*, vol. 8, pp. 90555–90569, 2020, doi: 10.1109/ACCESS.2020.2994033.
[26] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 53–62, Feb. 2022, doi: 10.21512/comtech.v13i1.7388.
[27] A. R. Ismail, N. Z. Abidin, and M. K. Maen, "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare," *J. Robot. Control*, vol. 3, no. 2, pp. 143–152, Feb. 2022, doi: 10.18196/jrc.v3i2.13133.
[28] V. Nassiri, G. Molenberghs, G. Verbeke, and J. Barbosa-Breda, "Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets," *Am. Stat.*, vol. 74, no. 2, pp. 125–136, Apr. 2020, doi: 10.1080/00031305.2018.1543615.
[29] 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.
[30] P. Eleftheriadis, S. Leva, and E. Ogliari, "Bayesian Hyperparameter Optimization of stacked Bidirectional Long Short-Term Memory neural network for the State of Charge estimation," *Sustain. Energy, Grids Networks*, vol. 36, p. 101160, Dec. 2023, doi: 10.1016/j.segan.2023.101160.

[31] L. Tani and C. Veelken, "Comparison of Bayesian and particle swarm algorithms for hyperparameter optimisation in machine learning applications in high energy physics," *Comput. Phys. Commun.*, vol. 294, p. 108955, Jan. 2024, doi: 10.1016/j.cpc.2023.108955.
[32] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019, doi: 10.11989/JEST.1674-862X.80904120.
[33] R. Herrera Casanova and A. Conde, "Enhancement of LSTM models based on data pre-processing and optimization of Bayesian hyperparameters for day-ahead photovoltaic generation prediction," *Comput. Electr. Eng.*, vol. 116, p. 109162, May 2024, doi: 10.1016/j.compeleceng.2024.109162.
[34] H. I. Oberman, S. van Buuren, and G. Vink, "Missing the Point: Non-Convergence in Iterative Imputation Algorithms," no. 1996, pp. 3–5, 2021, doi: 10.5334/dsj-2017-037.Van.
[35] H. Saputra, D. Stiawan, and H. Satria, "Malware Detection in Portable Document Format (PDF) Files with Byte Frequency Distribution (BFD) and Support Vector Machine (SVM)," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 4, pp. 1144–1153, 2023, doi: 10.26555/jiteki.v9i4.27527.
[36] W. J. Bidul, S. Surono, and T. B. Kurniawan, "Comparative Evaluation of Feature Selection Methods for Heart Disease Classification with Support Vector Machine," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 2, pp. 265–278, 2024, doi: 10.26555/jiteki.v10i2.28647.
[37] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," *Reliab. Eng. Syst. Saf.*, vol. 200, p. 106931, 2020, doi: 10.1016/j.ress.2020.106931.
[38] T. H. Saragih and M. I. Mazdadi, "Comparison of Air Quality Prediction using Random Forest and Gradient Boosting Tree," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, Dec. 2023, pp. 1–5, doi: 10.1109/ICIC60109.2023.10382104.
[39] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J. Promot. Commun. Stat. Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.
[40] A. Khairunnisa, K. A. Notodiputro, and B. Sartono, "A Comparative Study of Random Forest and Double Random Forest Models from View Points of Their Interpretability," *Sci. J. Informatics*, vol. 11, no. 1, pp. 207–218, Feb. 2024, doi: 10.15294/sji.v11i1.48721.
[41] M. Dwinanda, N. Satyahadewi, and W. Andani, "CLASSIFICATION OF STUDENT GRADUATION STATUS USING XGBOOST ALGORITHM," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 3 SE-Articles, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1785-1794.
[42] D. Arifah, T. H. Saragih, D. Kartini, M. Muliadi, and M. I. Mazdadi, "Application of SMOTE to Handle Imbalance Class in Deposit Classification Using the Extreme Gradient Boosting Algorithm," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 2, pp. 396–410, 2023, doi: 10.26555/jiteki.v9i2.26155.
[43] D. Pebrianti, H. Kurniawan, L. Bayuaji, and R. Rusdah, "XgBoost Hyper-Parameter Tuning Using Particle Swarm Optimization for Stock Price Forecasting," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 4, pp. 1179–1195, 2023, doi: 10.26555/jiteki.v9i4.27712.
[44] T. H. Saragih, R. Ramadhani, M. I. Mazdadi, and M. Haekal, "Energy Efficiency in Buildings Using Multivariate Extreme Gradient Boosting," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Dec. 2022, pp. 1–5, doi: 10.1109/ICIC56845.2022.10006902.
[45] V. Nguyen, "Bayesian Optimization for Accelerating Hyper-Parameter Tuning," in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Jun. 2019, pp. 302–305, doi: 10.1109/AIKE.2019.00060.
[46] G. Phillips *et al.*, "Setting nutrient boundaries to protect aquatic communities: The importance of comparing observed and predicted classifications using measures derived from a confusion matrix," *Sci. Total Environ.*, vol. 912, p. 168872, Feb. 2024, doi: 10.1016/j.scitotenv.2023.168872.
[47] Y. Wang, Y. Jia, Y. Tian, and J. Xiao, "Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring," *Expert Syst. Appl.*, vol. 200, p. 117013, Aug. 2022, doi: 10.1016/j.eswa.2022.117013.
[48] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.
[49] J. Y. Verbakel *et al.*, "ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models," *J. Clin. Epidemiol.*, vol. 126, pp. 207–216, Oct. 2020, doi: 10.1016/j.jclinepi.2020.01.028.
[50] J. Xu, "Comparing multi-class classifier performance by multi-class ROC analysis: A nonparametric

approach," *Neurocomputing*, vol. 583, p. 127520, May 2024, doi: 10.1016/j.neucom.2024.127520.
[51] T. R. Mahesh *et al.*, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9005278.
[52] Y. He, J. Zhou, C. Cao, S. Wang, and H. Fu, "Detection of electricity theft based on Minimal Gated Memory network combined adaptive synthesis sampling and decision tree," *Sustain. Energy, Grids Networks*, vol. 39, p. 101415, Sep. 2024, doi: 10.1016/j.segan.2024.101415.
[53] T. A. Assegie, A. O. Salau, K. Sampath, R. Govindarajan, S. Murugan, and B. Lakshmi, "Evaluation of Adaptive Synthetic Resampling Technique for Imbalanced Breast Cancer Identification," *Procedia Comput. Sci.*, vol. 235, pp. 1000–1007, 2024, doi: 10.1016/j.procs.2024.04.095.
[54] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction," *Clean. Eng. Technol.*, vol. 15, p. 100664, Aug. 2023, doi: 10.1016/j.clet.2023.100664.
[55] N. A. Pérez-Padilla *et al.*, "Optimizing trigger timing in minimal ovarian stimulation for In Vitro fertilization using machine learning models with random search hyperparameter tuning," *Comput. Biol. Med.*, vol. 179, p. 108856, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108856.

BIOGRAPHY OF AUTHORS

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

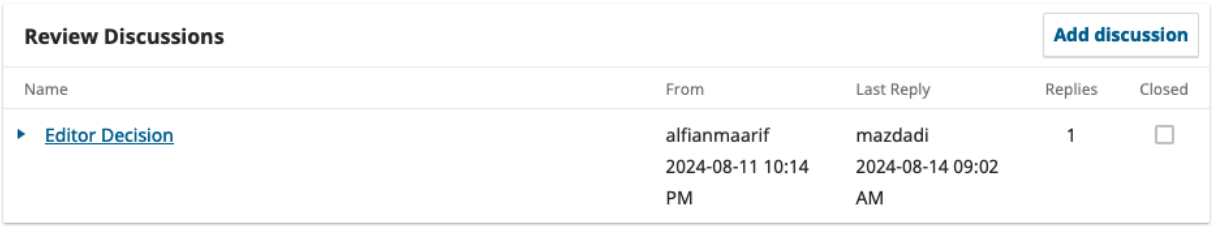
Triando Hamonangan Saragih, currently holding the position of a lecturer within the Department of Computer Science at Lambung Mangkurat University, is heavily immersed in the realm of academia, with a profound focus on the multifaceted domain of Data Science. His academic pursuits commenced with the 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

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 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

	**Ahmad Tajali,** a student at Lambung Mangkurat University who began his education in 2021 in the Department of Computer Science. His current research field is Data Science. Email: ahmadtajali61@gmail.com.

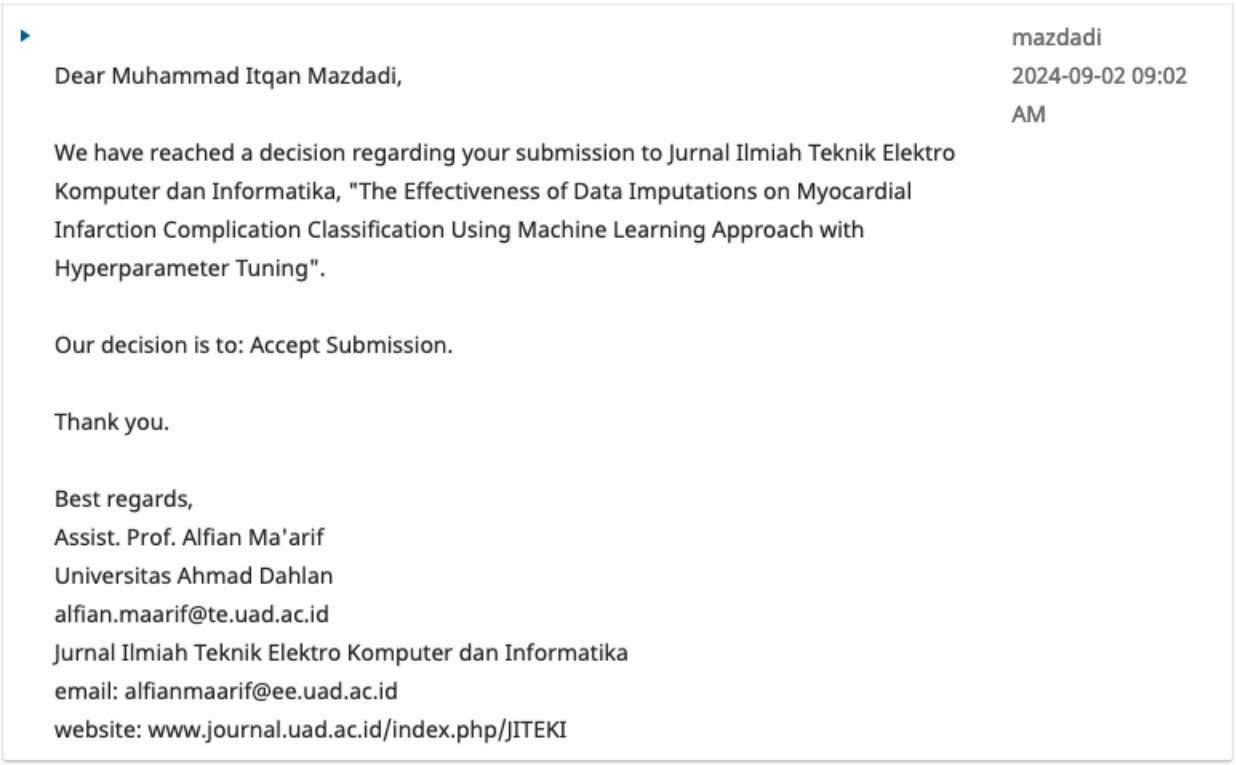
Bukti Submit Revisi



Review Discussions — Add discussion

Name	From	Last Reply	Replies	Closed
Editor Decision	alfianmaarif 2024-08-11 10:14 PM	mazdadi 2024-08-14 09:02 AM	1	☐

Bukti Notifikasi Diterima



mazdadi
2024-09-02 09:02 AM

Dear Muhammad Itqan Mazdadi,

We have reached a decision regarding your submission to Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, "The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning".

Our decision is to: Accept Submission.

Thank you.

Best regards,
Assist. Prof. Alfian Ma'arif
Universitas Ahmad Dahlan
alfian.maarif@te.uad.ac.id
Jurnal Ilmiah Teknik Elektro Komputer dan Informatika
email: alfianmaarif@ee.uad.ac.id
website: www.journal.uad.ac.id/index.php/JITEKI