Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

Effective Coronary Artery Disease Prediction Using Bayesian Optimization Algorithm and Random Forest

Muhammad Syiarul Amrullah[1*], Anny Yuniarti[2]

[1.2]Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS, Sukolilo, Surabaya 60111, Indonesia
Email: [1,*]6025232012@student.its.ac.id, [2]anny@if.its.ac.id
Correspondence Author Email: 6025232012@student.its.ac.id
Submitted: **13/07/2024**; Accepted: **08/09/2024**; Published**: 09/09/2024**

Abstract−Coronary artery disease (CAD) continues to be a major global health issue, demanding more effective diagnostic techniques. This study introduces a detailed framework for CAD detection that integrates data preprocessing, feature engineering, and model optimization to enhance diagnostic accuracy. Our methodology encompasses comprehensive data cleansing to eliminate inconsistencies, transformations for better feature representation, feature reduction to highlight relevant variables, data augmentation for balanced class distribution, and optimization strategies to boost model performance. We employed a random forest classifier, trained via 5-fold cross-validation, to develop a robust model. The efficacy of this model was tested through two key experiments: firstly, by comparing its performance on preprocessed versus raw data, and secondly, against previous studies. Results demonstrate that our model significantly surpasses the one trained on raw data, achieving an accuracy of 93.00% compared to 86.16%. Moreover, when compared with existing research, our random forest model excels with an accuracy of 93.00%, a F1 Score of 93.00%, and a recall of 94.00%. Despite the superior precision of the Hybrid PSO-EmNN model found in other research, our results are promising. They underscore the potential of advanced feature engineering to further refine the effectiveness of CAD detection models. The study concludes that meticulous data preprocessing and model optimization are crucial for enhancing CAD diagnostics. Future research should focus on incorporating more sophisticated feature engineering techniques and expanding the dataset to improve the model's precision and overall diagnostic capabilities.

Keywords: Random Forest; Bayesian Optimization; Coronary Artery Disease; Machine Learning; Feature Selection

1. INTRODUCTION

Coronary artery disease (CAD) is a serious health condition marked by narrowed or blocked coronary arteries, often leading to heart failure due to plaque buildup and inflammation [1]. CAD is a leading cause of heart failure, as it impedes the adequate flow of oxygen-rich blood to the heart muscle, potentially resulting in severe complications such as myocardial infarction or sudden cardiac death. It significantly contributes to global mortality, with early detection hampered by limitations in current diagnostic methods like the costly angiography, often inaccessible to low-income populations [2]. Conventional diagnostic approaches for coronary artery disease (CAD), such as angiography, utilize imaging modalities that offer a comprehensive visualization of the coronary arteries. Although angiography is widely regarded as the most reliable method for diagnosing coronary artery disease (CAD), it does have several drawbacks. The operation is expensive and intrusive, typically necessitating specialized facilities and expertise, which makes it less available, especially in low-income or resource-constrained environments. Moreover, the potential dangers linked to invasive treatments, such as hemorrhaging, infection, and exposure to radiation, render them less preferable for regular screening or early detection in those without symptoms. The presence of these constraints has sparked a quest for alternative diagnostic methods that are non-invasive, cost-effective, and can be easily adopted on a large scale.

Machine learning algorithms have proven effective in diagnosing CAD, utilizing data mining to extract vital information from large datasets and uncover hidden relationships, thereby improving disease detection and reducing mortality rates [3]. Through the utilization of machine learning, researchers have the ability to construct prediction models that enhance the precision of CAD diagnosis, ultimately assisting in the reduction of mortality rates linked to the condition. By incorporating machine learning (ML) into computer-aided diagnostics (CAD), a more individualized approach can be achieved, taking into account unique patient parameters including age, gender, lifestyle, and medical history. These factors have a substantial impact on the development and outcomes of diseases. However, challenges remain due to the use of outdated datasets like Z-Alizadeh, which lack critical features and adequate sample sizes. This constraint can result in partial models that do not effectively apply to various patient populations. Furthermore, the process of data preparation frequently lacks sufficient attention. Inadequate data preparation, which includes insufficient data cleaning, transformation, and augmentation, can lead to models that are unable to generate accurate and relevant predictions. This can result in incorrect outcomes and potentially detrimental therapeutic judgments[4].

Recent studies have advanced the diagnosis of coronary artery disease (CAD) using diverse machine learning models and datasets. A 2020 study [5] employing the Naive Bayes algorithm on a 303-patient dataset achieved a diagnostic accuracy of 84%, while another introduced a Hybrid Particle Swarm Optimization with Emotional Neural Network [6], achieving an accuracy of 88.34% and an F1 score of 92.12%. Additionally, an IBM SPSS Modeler-based system using random trees on the Z-Alizadeh Sani dataset outperformed other models with a 91.47% accuracy [7]. Another research utilized Random Forest and XGBoost, optimized by various hyperparameter techniques, achieving the highest accuracy of 80.20% on the Z-alisadeh dataset [8]. Meanwhile, a study employing exhaustive ensemble feature selection methods and a multi-layer perceptron classifier on multiple datasets demonstrated

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

accuracies up to 91.78% [9]. Despite these advancements, recent studies indicate that there is still a need for future research to concentrate on incorporating a wider range of datasets and advanced models to enhance the accuracy and applicability of CAD diagnostics. One of the prevailing weaknesses in prior research lies in the lack of integration between many models and various evolutionary feature selection algorithms, while certain datasets remain underutilized.

Table 1 highlights the crucial significance of data preparation in the diagnosis of CAD. The CAD dataset, which has not been widely utilised in previous studies, is subjected to standard data preparation processes including cleaning, transformation, reduction, and augmentation. This study incorporates a feature importance analysis utilising random forest approaches and assesses different machine learning models on both processed and unprocessed data, while comparing their accuracy, precision, recall, and F1 scores. The findings indicate that the stacking ensemble model, augmented with SMOTE and assessed using 5-fold cross-validation, surpasses earlier investigations, representing a noteworthy addition of this research.

To summarize, although machine learning has demonstrated significant potential in enhancing the diagnosis of coronary artery disease, there remains a considerable amount of work that has to be accomplished. Future research should give higher priority on utilizing broad and extensive datasets, including advanced models, and implementing robust data preparation approaches. Through this approach, scientists can create diagnostic tools that are more precise, dependable, and easily available. These technologies have the potential to greatly decrease the worldwide impact of CAD.

2. RESEARCH METHODOLOGY

Figure 1 illustrates the proposed preprocessing steps implemented prior to inputting the data into the Random Forest (RF) model. Data preprocessing refers to the procedure of converting unprocessed data into a structured format that may be readily utilised across different domains [10]. Data preprocessing was employed to optimize the model as part of the solution process to derive results from the hypothesis. Data preprocessing includes data cleaning, transformation, reduction, augmentation, modeling, hyperparameter tuning, and evaluation. Below is a detailed description of the stages involved in this research:

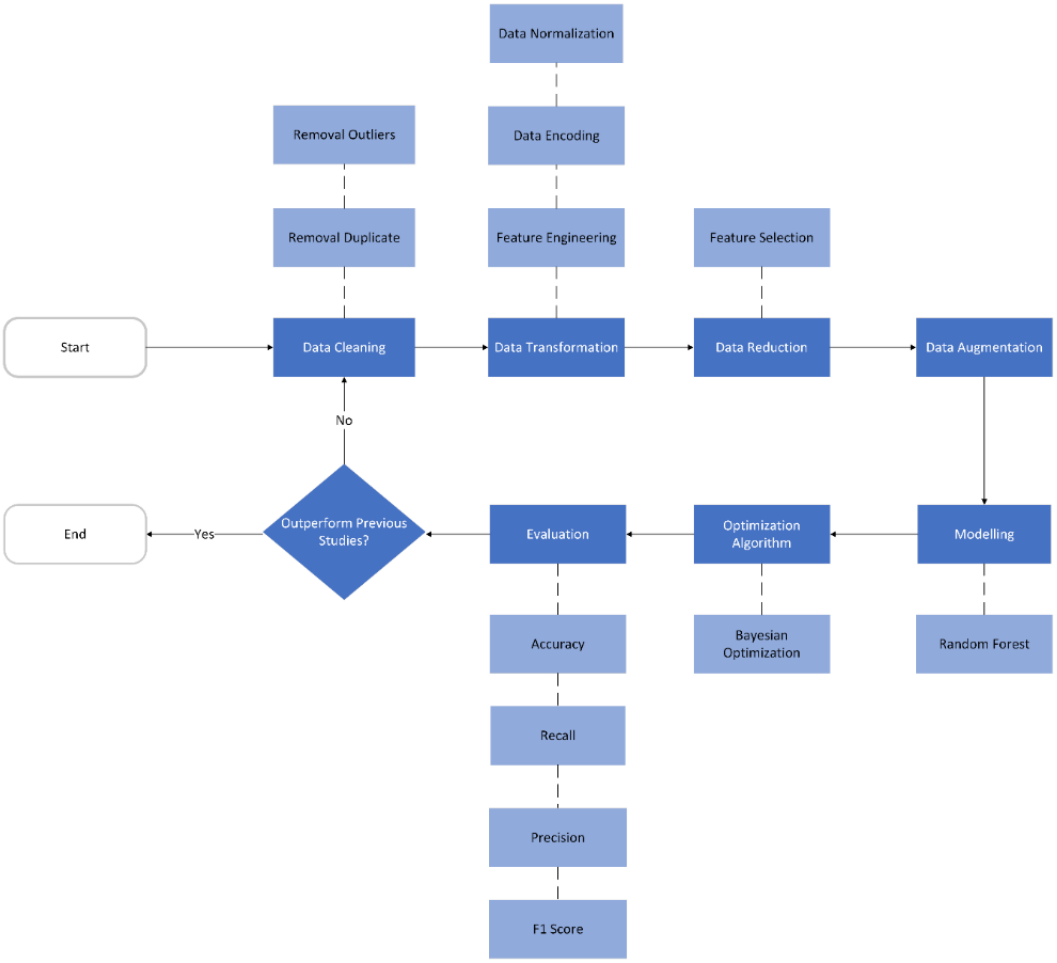

Figure 1. Flowchart of Proposed Method

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

2.1 Data Cleaning

Data cleaning is the process of identifying and repairing invalid, incorrect, or irrelevant records from a dataset [20]. This procedure ensures that the data is correct, consistent, and suitable for analysis[11]. There are several steps in data cleaning: removing duplicates and removing outliers. Duplicate removal is a regular activity for ensuring data integrity and correctness, particularly when dealing with large datasets, by removing them from the dataset [12].

Table 1. Related Work

Research	Data Cleaning	Data Transformation	Data Reduction	Data Augmentation	Optimization Algorithm
[5]			✓		✓
[6]			✓		✓
[7]		✓			
[8]			✓	✓	✓
[9]					
Proposed Method	✓	✓	✓	✓	✓

After eliminating duplicates, the subsequent task involves addressing outliers. Removing outliers is a crucial step in data preparation to ensure the accuracy and reliability of subsequent research [13]. Outliers, which are data points that significantly deviate from other observations, can distort statistical analyses and negatively impact model performance. This study employed the Interquartile Range (IQR) method.

$$K = J - L \tag{1}$$

The interquartile range (IQR) is a statistical metric that precisely measures the dispersion of data [14]. The calculation involves subtracting the value of the first quartile (Q1) from the value of the third quartile (Q3) in a dataset . The 25th percentile, sometimes referred to as Q1, represents the value that divides the lowest 50% of the dataset in half. The 75th percentile, also known as Q3, represents the value that is greater than or equal to 75% of the data points in the collection. Eq. **Kesalahan! Sumber referensi tidak ditemukan.**. has a mathematical formula to calculate the IQR(K): The interquartile range (IQR) is calculated by subtracting the first quartile (Q1/L), which represents the 25th percentile, from the third quartile (Q3/J), which represents the 75th percentile.

$$JB = L - 1.5 * K \tag{2}$$

$$KB = J + 1.5 * K \tag{3}$$

To identify outliers, the IQR and Outlier Boundaries (Upper Bound and Lower Bound) are needed. As shown in Eq. **Kesalahan! Sumber referensi tidak ditemukan.**, the lower bound (LB) is determined by subtracting 1.5 times the interquartile range (IQR) from the first quartile (Q1). Eq. **Kesalahan! Sumber referensi tidak ditemukan.**. presents the formula for calculating the upper bound (UB), which is obtained by adding 1.5 times the IQR to the third quartile (Q3). Any data point below the lower bound or above the upper bound is considered an outlier. In the end, any data point that exists beyond the limits of the outliers is eliminated.

2.2 Data Transformation

Data transformation refers to the procedure of changing data from one format, structure, or standard to another render it appropriate for analysis, storage, or interaction with other systems [15]. There are several steps in data transformation: Feature Engineering, Data Encoding and Data Normalization. Feature engineering is a method of utilizing domain expertise to generate new input features from existing data, with the aim of enhancing the performance of machine learning models [16]. The groupings are categorized into two to three levels: low, normal, and high. The discretization ranges specified in the Braunwald heart book [17]. Data encoding refers to the procedure of transforming data from one format or representation to another in order to guarantee its effective and efficient processing, transmission, or storage across various systems and applications [18].

After the data encoding phase, the next crucial step was data normalization, which is a key technique used in data preprocessing. Normalization is a process that seeks to standardize the values of numerical characteristics in a dataset [19]. This helps to reduce the danger of any individual feature having a disproportionate influence on the model because of its magnitude. The study utilized min-max normalization, a commonly used method, to linearly convert each characteristic to a predetermined range, often ranging from 0 to 1 [20]. This transformation maintains

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

the relative relationships between data points, guaranteeing that all attributes have an equal impact on subsequent analysis or modeling procedures.

$$V = (V - \min(v))/(\max(v) - \min(v)) \quad (4)$$

Equation **Kesalahan! Sumber referensi tidak ditemukan.** represents the mathematical expression for min-max normalization, a method that rescales the original data (v) to a normalized value (V) within a predetermined range. The formula entails the subtraction of the minimum value of the feature (min(v)) from the original data point, followed by the division of the resulting value by the range of the feature (max(v) - min(v)). This method efficiently rescales the data to a predefined range, typically from 0 to 1, guaranteeing that all features have an equal contribution in subsequent analysis.

2.3 Data Reduction

Data reduction refers to the procedure of converting numerical or alphabetical digital data into a rectified, organized, and streamlined state. The objective of this method is to decrease the amount of data while maintaining its crucial information, hence enhancing its manageability and efficiency for storage, processing, and analysis. In this step, only feature selection is performed. Feature selection involves choosing a subset of specific characteristics from an immense dataset to construct machine learning models [21]. This research uses a Random Forest classifier to determine feature importance. Although Random Forest is not explicitly designed for feature selection, it is commonly employed for this task since it can effectively assess the value of each variable in predicting an outcome. Features with an importance score exceeding 2% were selected for further analysis [22].

2.4 Data Augmentation

Data augmentation is a method used to artificially expand the quantity and variety of a training dataset by generating enhanced reproductions of existing data [23]. In this step used SMOTE algorithm. SMOTE (Synthetic Minority Over-sampling Technique) is a statistical method used to address class imbalance in datasets by creating artificial samples for the minority class, a process known as oversampling [24]. This approach is especially valuable in situations when the dataset exhibits a substantial disparity between the classes, which can have a detrimental effect on the efficacy of machine learning models. There are several steps to perform oversampling using SMOTE. First, identify the minority class. Then, determine the number of synthetic samples and find the k-nearest neighbors. Lastly, generate synthetic samples with Eq. **Kesalahan! Sumber referensi tidak ditemukan.**.

$$S = n + R.(N - n) \quad (5)$$

Eq. **Kesalahan! Sumber referensi tidak ditemukan.** shows formula to generate syntetic sample. Generate syntetic sample involves a process where a new data point (S) is created by combining a randomly selected sample from the minority class (n) with a weighted contribution from its nearest neighbors (N) within the same minority class. The weight assigned to each neighbor is determined by a random number (R) between 0 and 1, ensuring variability in the synthesized data.

2.5 Modelling

The modeling phase in machine learning entails the training, and validation of a machine learning model to acquire knowledge from the data and provide precise predictions [25]. This stage is vital since it has a direct influence on the performance and efficacy of the model. The algorithm model used in this research is the Random Forest (RF). Random Forest (RF) is a machine learning algorithm that is commonly employed for both classification and regression tasks [26].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

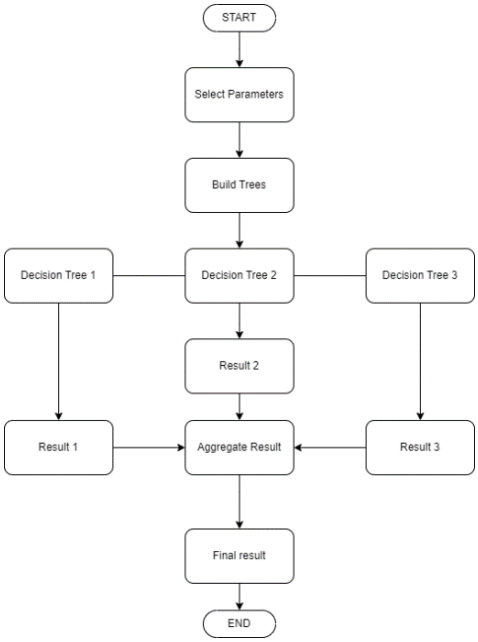

Figure 2. Flowchart: Random Forest Process

Figure 2. illustrates the Random Forest algorithm's workflow, starting with the initialization phase where the number of trees and features are selected. Each tree is built through bootstrap sampling of the dataset, followed by recursive node splitting based on randomly selected features until a stopping criterion is met. The individual trees are finalized, and their predictions are aggregated using majority voting for classification tasks or averaging for regression tasks, culminating in the final output. This structured process ensures robust and accurate predictions by leveraging the ensemble learning approach of Random Forests.

2.6 Hyperparameter Tuning

Hyperparameter tuning involves the selection of the most suitable settings for a machine learning model's hyperparameters in order to maximize its performance on a certain task . Hyperparameters are adjustable parameters that determine the behavior of the learning algorithm and are established prior to the commencement of the training process . Model parameters, in contrast to data-learned parameters, exhibit differences. Bayesian optimization is used to perform hyperparameter tuning in this step. It is a technique used to optimize objective functions that are costly to assess.

Bayesian optimization is commonly employed in machine learning to fine-tune hyperparameters and is especially efficient for functions that are intricate, noisy, non-convex, and require significant computational resources to analyze. Table 2 shows the boundaries of parameter values used for hyperparameter tuning.

Table 2. Parameter Boundaries

Parameter	Bound
n_estimators	(10,100)
max_depth	(5,50)
min_samples_split	(2,11)
min_samples_leaf	(1,11)
max_features	(1,64)
criterion	(gini, entropy)

2.7 Evaluation

After hyperparameter tuning, the next step is evaluation. Evaluation involves assessing the performance of the model using various metrics such as accuracy, precision, recall, and F1 score. Accuracy is a metric that measures the proportion of correct predictions among the total number of predictions . Eq. **Kesalahan! Sumber referensi tidak ditemukan.** shows the formula to measure accuracy, where K represents accuracy, N is the number of correct predictions, and S is the total number of predictions.

$$K = \frac{N}{S} \tag{6}$$

Secondly, another evaluation metric besides accuracy is precision. Precision is a metric that measures how many predicted positives are actually positive . Eq. **Kesalahan! Sumber referensi tidak ditemukan.** shows the formula for precision, where A represents precision, LP is true positive, and JP is false positive.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

$$A = \frac{LP}{LP+JP} \tag{7}$$

Thirdly, another evaluation metric besides precision is recall. Recall is a metric that measures how many actual positives were correctly predicted . Eq. **Kesalahan! Sumber referensi tidak ditemukan.** shows the formula for recall, where B represents recall, LP is true positive, and JL is false negative.

$$B = \frac{LP}{LP+JL} \tag{8}$$

Lastly, another evaluation metric besides recall is the F1 score. The F1 score is a harmonic mean of precision and recall, providing a single metric that balances both . Eq. **Kesalahan! Sumber referensi tidak ditemukan.** shows the formula for the F1 score, where C represents the F1 score, A is precision, and B is recall.

$$C = 2\ X\ \frac{A\ X\ B}{A+B} \tag{9}$$

3. RESULT AND DISCUSSION

This section presents the findings derived from the conducted experiments. The research involved experiments at every step of the preprocessing phase. The detailed results of these experiments are as follows:

3.1 Data Cleaning

Firstly, this study performed a check for duplicate entries and removed any that were found. Figure 3 clearly illustrates the distribution of duplicate and unique rows within the dataset. A thorough analysis revealed no duplicate entries, underscoring the unique nature of each row. In total, 303 unique rows were identified, indicating a diverse and rich dataset that did not require the removal of duplicate data.

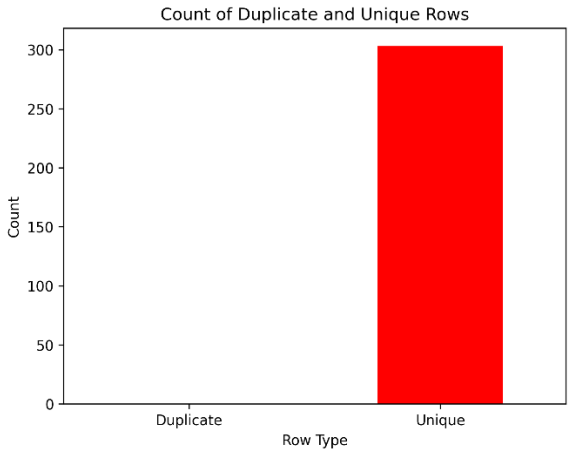

Figure 3. Count of Duplicate Data

After eliminating duplicates, the subsequent task involves addressing outliers. Figure 4 illustrates the percentage of outliers per feature, highlighting those with an outlier prevalence exceeding 5% of the data. While some outliers were removed, both training and evaluation performance declined, suggesting that not all outliers were successfully eliminated. To address this, feature engineering was applied to features with outlier rates exceeding 2%, specifically FBS, BUN, TG, Na, EF-TTE, ESR, PR, PLT, WBC, BP, HB, and HDL. These features were categorized into distinct groups based on their respective value ranges. Other features, such as BMI, LDL, Lymph, Weight, K, and Neut, had their outliers removed. Figure 5 shows the results after removing outliers.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

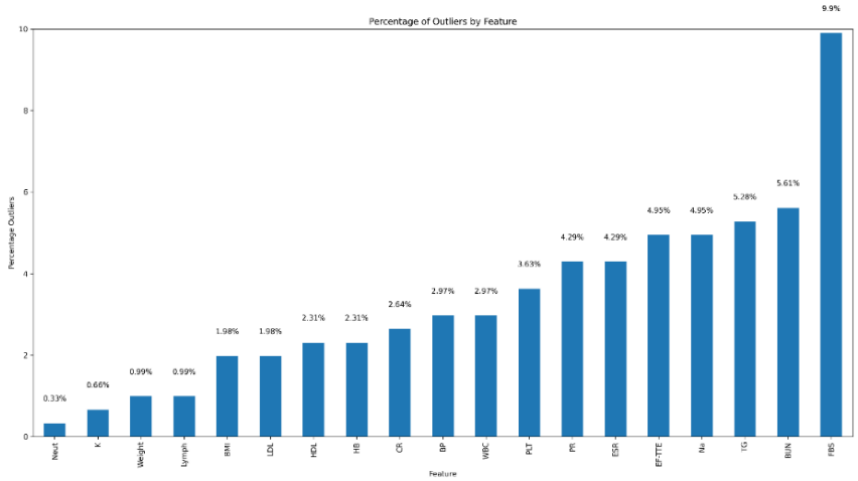

Figure 4. Percentage of Outliers

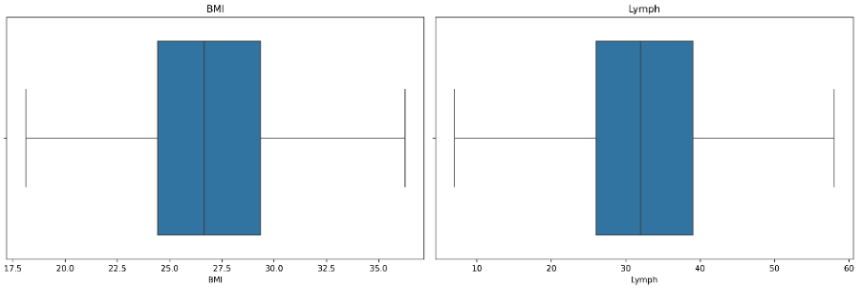

Figure 5. Outlier Removal Results

3.2 Data Transformation

After data cleaning, the subsequent step involves data transformation, beginning with feature creation. **Figure 6** shows an example of feature creation, where data is categorized into two or three groups: Low, Mid, and High. This categorization enables the differentiation of data points based on their values, thereby facilitating more effective analysis. By creating these categorical features, we can better capture the underlying patterns and relationships within the data, ultimately improving the performance and interpretability of subsequent models.

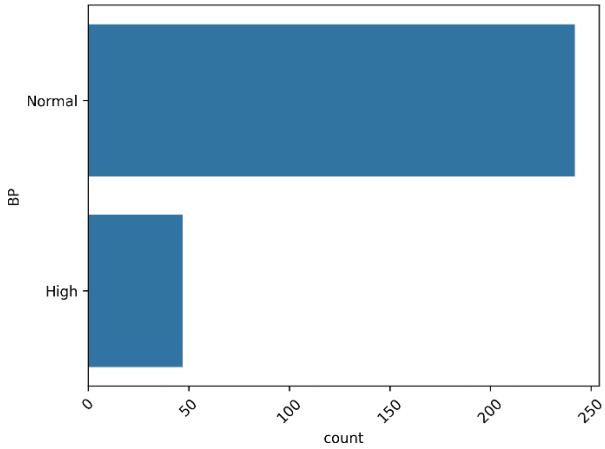

Figure 6. feature Creation Results

After the creation of new features, the subsequent step involves the crucial process of data encoding. Figure 7 displays the distribution of blood pressure categories before and after the application of label encoding, a technique used to convert categorical data into numerical values that can be processed by machine learning algorithms. In this context, the blood pressure categories were transformed into numerical values to facilitate model training and evaluation. Specifically, the category labeled 'Normal' was assigned the numerical value of 1, whereas the category labeled 'High' was assigned the numerical value of 0. The encoded data is depicted in Figure 7, which clearly illustrates the distribution changes pre- and post-encoding. This step is pivotal in preparing the dataset for subsequent analysis and model development, ensuring that all features are in a suitable format for computational processing.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

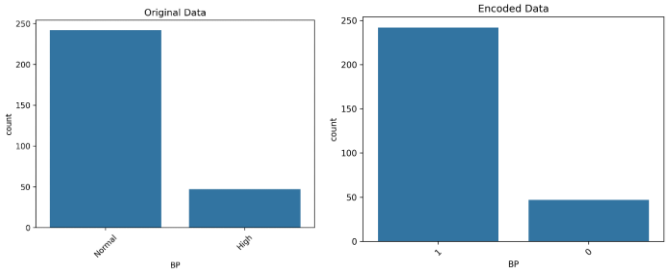

Figure 7. Encoding Data Results

After the data encoding phase, the next crucial step was data normalization. Figure 8 clearly depicts the data distribution both before and after the normalization process, providing a visual comparison that highlights the effectiveness of this technique. The pre-normalization data exhibited significant variability in scale, which could potentially lead to skewed model performance. Post-normalization, the data values were rescaled to fall within a standard range, thereby mitigating the risk of bias and enhancing the model's learning capability.

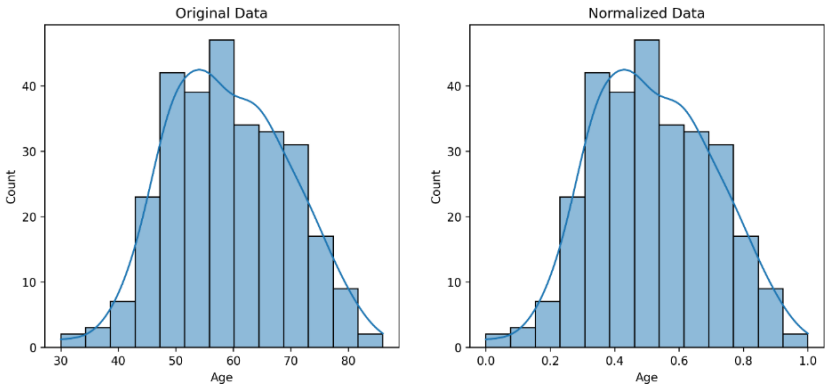

Figure 8. Results of Data Normalization

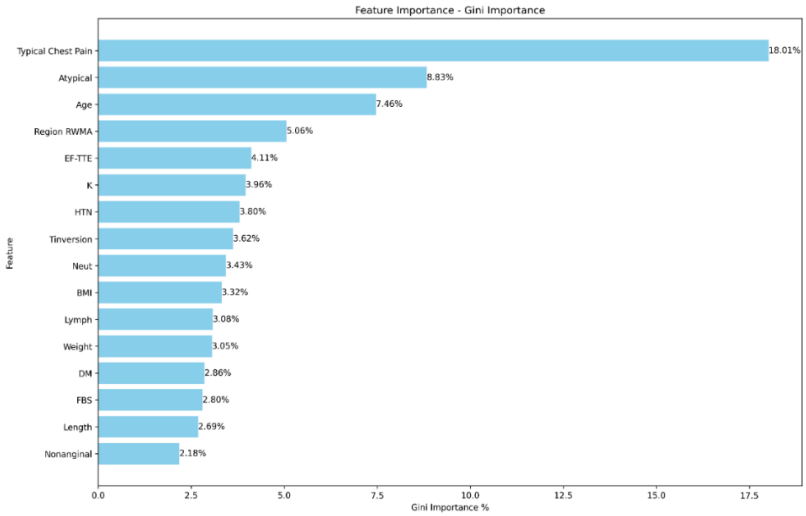

Figure 9. Feature Importance - Gini Importance

3.3 Data Reduction

Following the data transformation, the next step is data reduction. Figure 9 illustrates the features selected by the random forest algorithm, highlighting the significance of each feature within the CAD diagnostic model. This selection process leverages all the features available in the dataset to identify those most influential in predicting coronary artery disease. Among the features, the variable "Typical Chest Pain" demonstrates a particularly noteworthy contribution, accounting for 18.01% of the model's predictive power. This substantial contribution underscores the importance of "Typical Chest Pain" in the diagnostic process, indicating its potential as a critical indicator in the assessment of coronary artery disease.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

3.4 Data Augmentation

Following the data reduction, the next step involves data augmentation. Figure 10 compares the class frequencies before and after the application of oversampling techniques, highlighting the successful balancing of the dataset achieved through the addition of synthetic samples to the minority class.

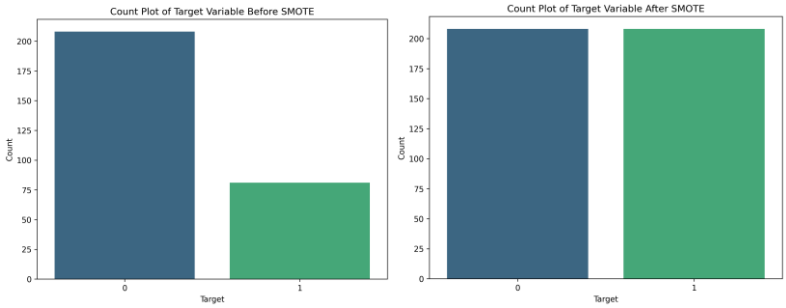

Figure 10. Comparison of Data Before and After Oversampling

3.5 Modelling

Following all preprocessing steps, the next phase is modeling. In this study, the data was shuffled and split into 5 folds to perform cross-validation, ensuring that the model's performance is robust and generalizable. The cross-validation technique partitions the dataset into five subsets, sometimes known as 'folds,' where one fold is designated as the test set and the remaining folds are utilised as the training set. The method is iterated five times, ensuring that each fold is utilised as a test set once. This approach helps mitigate overfitting and provides a comprehensive evaluation of the model's performance on unseen data. Figure 11 shows the model performance metrics: accuracy is 92%, precision is 91%, recall is 92%, and the F1 score is 92%."

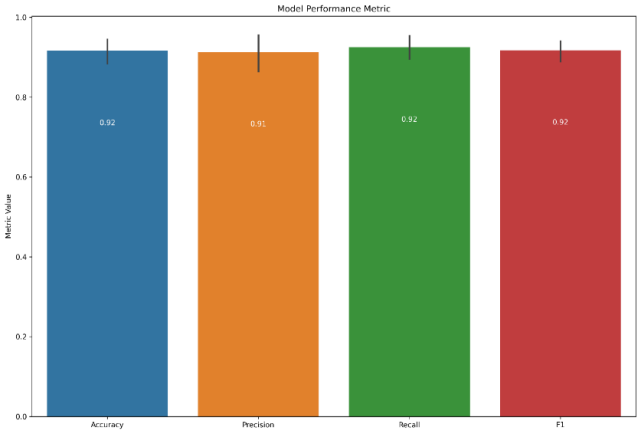

Figure 11. Performance Metrics of the Model

3.6 Optimization Algorithm

Four parameters are used in the tuning process: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features and criterion. Bayesian optimization iteratively searches within the defined bounds to find the maximum accuracy. Once the optimal accuracy is achieved, Bayesian optimization provides the model with the best parameters to use. Table 3 displays the best parameters identified through Bayesian optimization.

Table 3. Optimized Parameters

Parameter	Bound
n_estimators	53
max_depth	46
min_samples_split	2
min_samples_leaf	1
max_features	1
criterion	entropy

Figure 12 represents the comparative evaluation of performance before and after optimisation. The results unequivocally indicate that the optimisation method had a substantial and beneficial effect on the model's

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

performance. Before optimization, the model's evaluation metrics were acceptable but had potential for enhancement. Following the implementation of Bayesian optimisation, all of these indicators shown significant improvement.

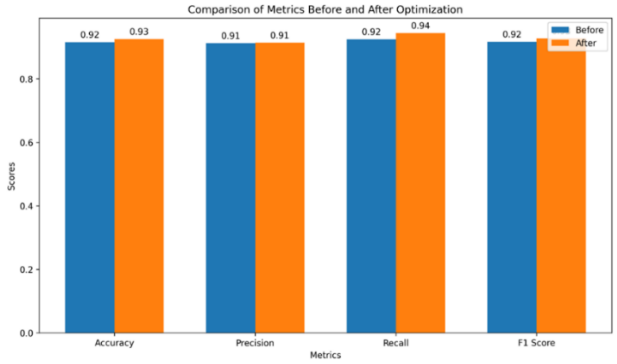

Figure 12. Metric Comparison Before and After Optimization

3.7 Evaluation

This section provides a presentation of the findings derived from the conducted experiments. This research consisted of two experiments: Comparison of the Proposed Approach with Unprocessed Data, and another comparing the proposed approach with related work.

a. Comparison Between Processed and Unprocessed Data
The suggested model, when using all selected and processed features, offers the highest attainable dependability and accuracy based on the CAD dataset. The performance of the suggested model in CAD diagnosis is evaluated in comparison to another model that utilises all characteristics and unprocessed data. According to Figure 13 shows that the model with selected features and processed data achieves better accuracy compared to the model with all features and unprocessed data. Table 4 shows accuracy model each fold between processed data and unprocessed data.

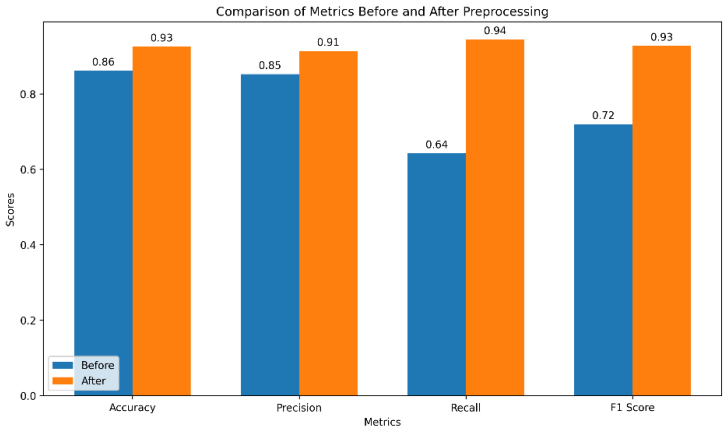

Figure 13. Comparison of Metrics Before and After Processing

According to Table 4, the treated data shows improved results in the majority of folds. This demonstrates the advantageous nature of performing data preprocessing prior to training the model.

Table 4. Performance Metrics by Fold

Processed Data				
Fold	Accuracy	Precision	Recall	F1 Score
1	90.00%	85.10%	**97.56%**	90.90%
2	91.56%	95.12%	88.63%	91.76%
3	92.77%	88.88%	97.56%	93.02%
4	92.77%	**95.34%**	91.11%	93.18%
5	**95.18%**	92.30%	97.29%	**94.73%**
Unprocessed Data				
1	85.24%	81.25%	68.42%	74.28%
2	78.68%	80%	42.10%	55.17%
3	85.24%	90%	52.94%	66.66%
4	90.00%	75%	85.71%	80.00%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

5	91.66%	100%	72.22%	83.87%

b. Comparison Between the Proposed Method and Related Work
The literature models constructed using algorithms like as Naive Bayes and Hybrid PSO-EmNN have been evaluated using the Z-alisadeh sani dataset. Table 5 demonstrates that the suggested model yields the most precise outcomes when compared to other models found in the literature. The precision scores of the MLP model are higher than those of the suggested model. Thus, it may be inferred that the proposed method is the optimal literary model. Conversely, the random forest model achieves the lowest level of accuracy.

Table 5. Comparative Evaluation of Metrics: Proposed Method vs. Related Work

Method	Accuracy	Precision	Recall	F1 Score
Naïve Bayes [5]	83.00%	64.00%		78.00%
Hybrid PSO-EmNN [6]	88.34%	92.37%		92.12%
Random Trees [7]	91.47%			
Random Forest [8]	80.20%		84.00%	
Multi Layer Perceptron [9]	91.78%	**95.14%**	93.50%	
Proposed Method	**93.00%**	91.00%	**94.00%**	**93.00%**

3.8 Discussion

In this research, the employment of the random forest algorithm along with advanced data preprocessing techniques has proven effective in the detection of coronary artery disease (CAD), as evidenced by the notable improvement in accuracy and F1 score when compared to models trained on raw data. This improvement highlights the critical role of thorough data preparation in enhancing machine learning model performance. A central aspect of our discussion involves comparing our results with those obtained from other studies, particularly those employing different methodologies such as the Hybrid PSO-EmNN model. While our model excels in overall accuracy and recall, it is noted that the Hybrid PSO-EmNN model achieves higher precision. This discrepancy opens up avenues for further research, specifically in exploring hybrid models that might combine the strengths of random forest and PSO-EmNN techniques to achieve balanced performance across all metrics.

Another significant point of discussion is the impact of data diversity on model performance. Our findings suggest that incorporating a more varied dataset could potentially improve precision, a hypothesis that aligns with current understanding in the field of machine learning where data diversity often correlates with model robustness. Furthermore, while our study provides promising results, it also acknowledges limitations, including the scope of data used. Expanding the dataset not only in size but also in variability could help in validating the model's effectiveness across different demographics and clinical settings. Additionally, continuous refinement of feature engineering techniques, especially those tailored towards specific characteristics of CAD, could further enhance the model's diagnostic capabilities.Future work will focus on these areas, aiming to refine the predictive model and explore the integration of additional machine learning techniques. This sustained effort is expected to push the boundaries of current CAD diagnostic methods and potentially establish new benchmarks in the accuracy and reliability of CAD detection using machine learning.

4. CONCLUSION

This research introduces a machine learning algorithm along with a comprehensive data analysis procedure to effectively detect coronary artery disease (CAD). The study involves multiple stages of data preparation including cleansing, transformation, reduction, augmentation, and the application of optimization techniques. We employed the random forest algorithm, trained with 5-fold cross-validation, to develop a robust model for CAD diagnosis. The research was structured around two main experiments: firstly, a comparison of our methodology using preprocessed data versus raw data, and secondly, a comparison of our approach with previous studies in the field. Results indicate that our random forest model, when trained on preprocessed data, significantly outperforms the version trained on raw data, achieving an accuracy of 93.00% as opposed to 86.16%. Further comparisons with past research show that our model not only maintains a high accuracy rate of 93.00% but also achieves a F1 Score of 93.00% and a recall of 94.00%. These metrics underscore the model's effectiveness in diagnosing CAD accurately. Although a Hybrid PSO-EmNN model exhibited a higher precision in another study, our findings are still encouraging. They suggest that further enhancements could be achieved through additional feature engineering and by expanding the dataset with more diverse data examples. Future work will focus on improving the precision by incorporating a larger dataset and continuing to refine our feature engineering techniques. This approach promises to enhance the predictive capabilities of our model, potentially setting a new benchmark in CAD diagnosis using machine learning.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

ACKNOWLEDGMENT

We like to extend our heartfelt appreciation to Institut Teknologi Sepuluh Nopember for furnishing the essential resources and facilities for this research endeavor. We express our sincere gratitude to Dr. Anny Yuniarti, S.Kom, M.Comp.Sc. our valued colleague and corresponding author, for their important assistance and support throughout this project. We would like to express our gratitude to the participants who generously dedicated their time and provided valuable insights for this research.

REFERENCES

[1] R. Moretti *et al.*, "Common Shared Pathogenic Aspects of Small Vessels in Heart and Brain Disease," *Biomedicines*, vol. 10, no. 5, 2022, doi: 10.3390/biomedicines10051009.
[2] P.-S. Huang *et al.*, "An artificial intelligence-enabled ECG algorithm for the prediction and localization of angiography-proven coronary artery disease," *Biomedicines*, vol. 10, no. 2, p. 394, 2022.
[3] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Comput Methods Programs Biomed*, vol. 192, p. 105400, 2020.
[4] F. Bodendorf, M. Sauter, and J. Franke, "A mixed methods approach to analyze and predict supply disruptions by combining causal inference and deep learning," *Int J Prod Econ*, vol. 256, p. 108708, 2023.
[5] S. S. Alotaibi *et al.*, "Automated prediction of Coronary Artery Disease using Random Forest and Naïve Bayes," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2020, pp. 109–114. doi: 10.1109/ICACSIS51025.2020.9263159.
[6] A. H. Shahid and M. P. Singh, "A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network," *Biocybern Biomed Eng*, vol. 40, no. 4, pp. 1568–1585, 2020, doi: https://doi.org/10.1016/j.bbe.2020.09.005.
[7] J. H. Joloudari *et al.*, "Coronary Artery Disease Diagnosis; Ranking the Significant Features Using a Random Trees Model," *Int J Environ Res Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17030731.
[8] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomed Signal Process Control*, vol. 70, p. 103033, 2021, doi: https://doi.org/10.1016/j.bspc.2021.103033.
[9] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," *Comput Stand Interfaces*, vol. 84, p. 103706, 2023, doi: https://doi.org/10.1016/j.csi.2022.103706.
[10] S. F. Pane and M. S. Amrullah, "Systematic Literature Review: Analisa Sentimen Masyarakat terhadap Penerapan Peraturan ETLE," *Journal of Applied Computer Science and Technology*, vol. 4, no. 1, pp. 65–74, Jul. 2023, doi: 10.52158/jacost.v4i1.493.
[11] M. Hosseinzadeh *et al.*, "Data cleansing mechanisms and approaches for big data analytics: a systematic study," *J Ambient Intell Humaniz Comput*, pp. 1–13, 2023.
[12] X. Wu, W. Zheng, X. Xia, and D. Lo, "Data quality matters: A case study on data label correctness for security bug report prediction," *IEEE Transactions on Software Engineering*, vol. 48, no. 7, pp. 2541–2556, 2021.
[13] B. Dastjerdy, A. Saeidi, and S. Heidarzadeh, "Review of applicable outlier detection methods to treat geomechanical data," *Geotechnics*, vol. 3, no. 2, pp. 375–396, 2023.
[14] D. P. Misra, O. Zimba, and A. Y. Gasparyan, "Statistical data presentation: a primer for rheumatology researchers," *Rheumatol Int*, vol. 41, no. 1, pp. 43–55, 2021.
[15] B. Diène, J. J. P. C. Rodrigues, O. Diallo, E. L. H. M. Ndoye, and V. V Korotaev, "Data management techniques for Internet of Things," *Mech Syst Signal Process*, vol. 138, p. 106564, 2020.
[16] X. Zhang, Y. Han, W. Xu, and Q. Wang, "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture," *Inf Sci (N Y)*, vol. 557, pp. 302–316, 2021.
[17] D. L. Mann, D. P. Zipes, P. Libby, and R. O. Bonow, *Braunwald's Heart Disease E-Book: A Textbook of Cardiovascular Medicine*. Elsevier Health Sciences, 2014. [Online]. Available: https://books.google.co.id/books?id=1R44BAAAQBAJ
[18] V. Barannik, S. Sidchenko, N. Barannik, and V. Barannik, "Development of the method for encoding service data in crypto-compression image representation systems," *Eastern-European Journal of Enterprise Technologies*, vol. 3, no. 9, p. 111, 2021.
[19] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2020, pp. 729–735.
[20] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," *Data (Basel)*, vol. 6, no. 2, p. 11, 2021.
[21] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022.
[22] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 785−796
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5554

[23] Y. Zhou, F. Dong, Y. Liu, Z. Li, J. Du, and L. Zhang, "Forecasting emerging technologies using data augmentation and deep learning," *Scientometrics*, vol. 123, pp. 1–29, 2020.
[24] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach Learn*, pp. 1–21, 2023.
[25] S. Studer *et al.*, "Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology," *Mach Learn Knowl Extr*, vol. 3, no. 2, pp. 392–413, 2021.
[26] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, p. 102051, 2020.