Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 10, No. 1, March 2024, pp. 123-137
ISSN: 2338-3070, DOI: 10.26555/jiteki.v10i1.28107

Improving Performance for Diabetic Nephropathy Detection Using Adaptive Synthetic Sampling Data in Ensemble Method of Machine Learning Algorithms

Lailil Muflikhah[1], Fitra A. Bachtiar[1], Dian Eka Ratnawati[2], Riski Darmawan[1]
[1]Department of Informatics Engineering, Brawijaya University, Malang 65145, Indonesia
[2]Department of Information System, Brawijaya University, Malang 65145, Indonesia

ARTICLE INFO

Article history:

Received December 30, 2023
Revised February 18, 2024
Published March 09, 2024

Keywords:

Nephropathy;
Oversampling;
Adasyn;
Bagging;
Boosting;
Machine learning

ABSTRACT

Nephropathy is a severe diabetic complication affecting the kidneys that presents a substantial risk to patients. It often progresses to renal failure and other critical health issues. Early and accurate prediction of nephropathy is paramount for effective intervention, patient well-being, and healthcare resource optimization. This research used medical records from 500 datasets of diabetic patients with imbalanced classes. The main goal of this study is to get high-performance predictive models for nephropathy. So, this study suggests a new way to deal with the common problem of having too little or too much data when trying to predict nephropathy: adding more data through adaptive synthetic sampling (ADASYN). This technique is particularly pertinent in ensemble machine-learning methods like Random Forest, AdaBoost, and bagging (Adabag). By increasing the number of instances of minority classes, it tries to reduce the bias that comes with imbalanced datasets, which should lead to more accurate and strong predictive models in the long run. The experimental results show an improving 4% rise in performance evaluation such as precision, recall, accuracy, and f1-score, especially for the ensemble methods. Two contributions of this research are highlighted here: first, the utilization of adaptive synthetic sampling data to improve the balance and diversity of the training dataset. The second contribution is incorporating ensemble methods within machine learning algorithms to enhance the accuracy and robustness of diabetic nephropathy detection.

Corresponding Author:

Lailil Muflikhah, Department of Informatics Engineering, Brawijaya University, Malang, 65145, Indonesia
Email: lailil@ub.ac.id

1. INTRODUCTION

Diabetic nephropathy (DN) is a severe complication of diabetes mellitus and a leading cause of end-stage renal disease. Although albuminuria is a marker of DN, there is a subset of DN patients who are not characterized by high albuminuria levels. This poses a significant challenge in the early detection of the disease. Furthermore, the disease is characterized by different pathophysiological mechanisms and clinical outcomes. Hence the need for case-specific biomarkers and diagnostic criteria for personalized treatment strategies. In addition, current treatment approaches mainly target albuminuric DN, may adjust in the choice of personalized therapy to DN without albuminuria. There is thus a need to identify novel biomarkers, and develop targeted interventions to improve the clinical outcomes of this often overlooked subset of DN patients. In this study, we aim to explore the challenges and opportunities in early identification and prediction of albumin level in diabetic patients as a predictor of diabetic nephropathy. Furthermore, potential research directions may address this unmet medical need [1]. On the other hand, diabetic nephropathy is a kidney condition that results from diabetes. Europe and the United States have the highest rates of kidney failure. Diabetic nephropathy

progresses through five distinct stages. In the early stage (Phase I), hyperfiltration occurs, leading to increased GFR (glomerular filtration rate), AER (albumin excretion rate), and kidney enlargement. In Phase II, albumin excretion remains relatively normal (<30mg/24 hours), although some patients may experience hyperfiltration, which increases the risk of developing diabetic nephropathy. Phase III is characterized by microalbuminuria (30-300mg/24 hours). Phase IV is identified by a positive Difstick proteinuria test, with albumin excretion exceeding 300mg/24 hours. At this stage, there is a decrease in glomerular filtration rate (GFR), and hypertension is often detected. Stage V is the final stage of kidney disease, and End Stage Renal Disease (ESRD). At this stage dialysis is usually started when the glomerular filtration rate (GFR) decreases to 15ml/min. Nephropathy is characterized by increased urinary albumin excretion, decreased glomerular filtration rate (GFR), or both, as observed in clinical medicine and epidemiological investigations [2].

In recent years, medical diagnosis with machine learning has been widespread. Machine learning has proven useful in several medical fields, including mental health diagnosis [3], oncology [4], cancer prediction and prognosis [5], and breast cancer diagnosis [6]. Medical imaging for neurosurgery, radiation, and cancer diagnostics has used this method well too [7]. Also, medical professionals have used machine learning algorithms extensively to assist in the diagnostic process. Medical diagnostics are now more accurate and efficient thanks to evidence-based decision-making made possible by data-intensive machine learning approaches [8]. Integrating machine learning models into medical diagnostics has greatly improved the accuracy and predictive ability of diagnostic procedures, although large medical datasets and learning algorithms have been available for decades [9]. In general, machine-learning approaches can substantially improve medical diagnosis for some medical disorders and fields of study [10]. The probability of diabetic nephropathy can be determined by classifying individuals into different risk categories using machine learning algorithms. This risk category can be used by healthcare practitioners to personalize treatment plans [11]. In addition, several studies have explored the use of machine learning to identify individuals on in risk stratification. The researchers have investigated the use of image-based techniques to detect and track diabetic nephropathy with clinical information. This technology uses medical imaging techniques to detect changes associated with nephropathy. Machine learning has been used to classify diabetic nephropathy patients into different subgroups based on their disease progression. It enables the implementation of personalized medical plans and customized therapies. Many studies integrate various data sources, including microarray gene expression data sets, electronic health records, genome, and proteomic data, to build models to predict diabetic nephropathy [12]. However, one of the problems in applying machine learning for medical diagnosis is imbalanced data in constructing the classifier model.

The unequal distribution of data samples across the classroom affects learning. This makes it difficult for classifiers to learn from minority class instances efficiently [13]. To balance the distribution between the samples of the majority and minority classes, the oversampling approach creates a synthetic sample of the less represented classes to address this problem [14]. The re-sampling strategy has been used in several studies to solve the problem of class distribution that is not equal in the data. Three of these methods are up-down, over-, and under-sampling. Furthermore, research in this field deals with challenges such as understanding and accuracy of machine learning models. Machine learning algorithms are challenged by these differences, especially when it comes to categorizing minority classes correctly [15]. Oversampling methods such as Adaptive Synthetic Sampling (ADASYN) are one way to address this problem. It has been shown that unbalanced data sets can be re-balanced efficiently by using this technique, which facilitates the learning of classifiers [16]. In addition, oversampling methods such as ADASYN have proven important in improving the performance of machine learning models in the field of fraud detection, where unbalanced data sets are typical. The effectiveness of ADASYN has been proven in several applications, such as plant identification and activity identification, where it has helped produce highly accurate results on unbalanced datasets [17]. In addition, ADASYN performance evaluation against other machine learning algorithms, such as AdaBoost and XGBoost, has been carried out [18]. This assessment shows the effectiveness of ADASYN in addressing the difficulties shown by highly unbalanced data sets. This method modifies the dataset by either increasing or decreasing the number of samples to achieve a balanced distribution of classes. The objective is to mitigate bias or influence on the analysis or modeling, hence enhancing its reliability and applicability to all categories. It is crucial to acknowledge that the domain of diabetic nephropathy and machine learning is constantly progressing, with continuing investigations focused on enhancing the early identification, accuracy of prediction, and overall results for patients. Therefore, the objective of this study is to enhance the performance rate by employing the Adaptive Synthetic Sampling (ADASYN) technique to resample the data.

2. METHODS

In this study, we proposed a method to get results with better performance rates. The research begins with collecting data, which means obtaining a set of important data related to the problem of diabetic nephropathy. Preprocessing is needed to improve the data so that it can be analyzed after it is collected. In this step, we do data cleaning, fill the voids with information, and make sure that all forms are the same. The Adaptive Synthetic Sampling Approach for Unbalanced Learning (ADASYN) aims to help correct class imbalances and make data more representative. Later, the researchers used advanced machine learning algorithms and statistical methods to build a classification model. This model aims to find trends, make predictions, and learn more from the data. Finally, the review step looks at the accuracy, precision, memory, and other factors of the classifier model critically. This complete review confirms the validity and reliability of the results of the study, proving the authenticity and significance of the research. Fig. 1 shows the general steps the researchers took to try to predict diabetic nephropathy. First, information is collected from the medical records of people with diabetes. Then, ADASYN is used for data pre-processing, which includes enhanced data. After that, the data added was used to create predictor models that could predict diabetic nephropathy. Finally, we see how well the proposed method works.

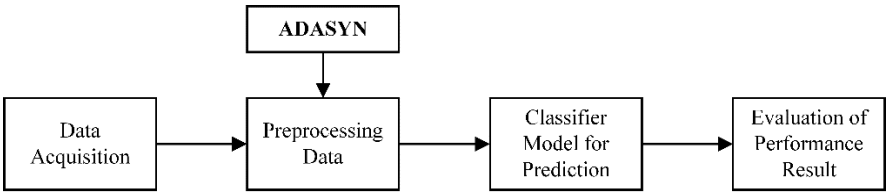


Fig. 1. General Steps in this research

The proposed method of this study is the utilization of a technique known as "data augmentation," specifically through oversampling using the Adaptive Synthetic Sampling Method for Imbalanced Data (ADASYN). The suggested method involves initiating data gathering from medical records, followed by preprocessing the data to enhance data quality. The final stage involves the implementation of a machine learning or ensemble algorithm for detection. Initially, data was gathered from diabetic individuals. The presence of missing values in the data and the significant disparity in attribute value range contribute to the noise observed. Consequently, we address missing values by utilizing KNN Imputation for numerical data and modus for categorical data values. Subsequently, we used normalization to ensure that the attribute value falls inside a predetermined range. Subsequently, to address the uneven distribution of data across each class, we employ resampling through the utilization of the ADASYN approach. Ultimately, we utilized machine learning and an ensemble technique. Fig. 2 illustrates the overall research procedure.

Then, the steps of resampling using ADASYN are designed in Fig. 3. Initially, the dataset was identified as the minority class. Then, we calculate the imbalance ratio and the amount of data for each class. For each minority class data, we find its 'k' nearest neighbors in feature space using a nearest neighbors' algorithm. After that, Synthetic samples are generated by randomly selecting a neighbor of a minority class sample and generating a new sample along the line, determined by a random value between 0 and 1. Finally, the generated synthetic data is combined with the original data.

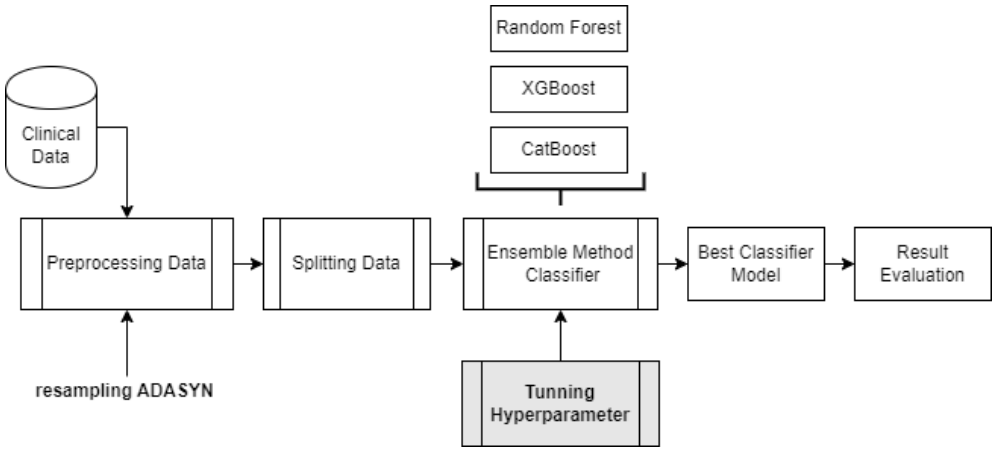


Fig. 2. The proposed method of Diabetic Nephropathy Detection

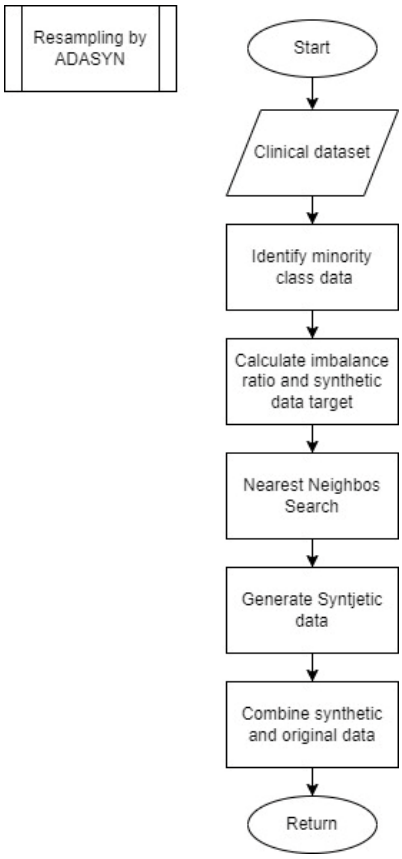
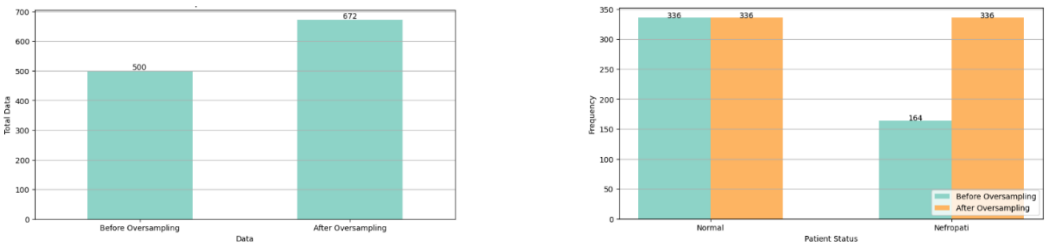


Fig. 3. Flowchart of Adaptive Synthetic Sampling Data (ADASYN)

2.1. Data Acquisition

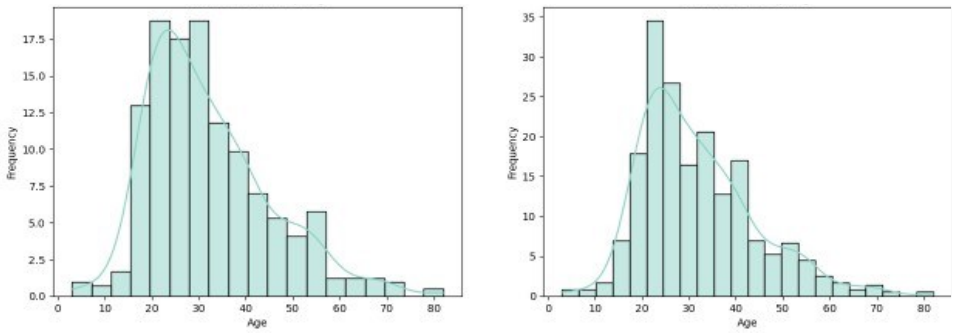
This data was collected from patients with diabetes at Saiful Anwar Hospital. It refers to a collection of information collected by individuals who have been diagnosed with diabetes nephropathy with related variables. The dataset consists of 500 lines of patient data with details of 336 normal patients and 164 patients suffering from nephropathy. The attributes used include patient information and laboratory results such as age, creatine levels, quantitative UACR, urine creatine, and urine albumin. The data ratio is unbalanced, so this study applies data oversampling. As an illustration, the number of datasets is updated to reach the highest number of data sets in both classes, as shown in Fig. 4. The comparison of all data before and after sampling is in Fig. 4 (a), and the comparisons of each class before and following samples are in Fig. 4 (b).



(a) (b)

Fig 4. Comparison of data before and after resampling (a) total dataset (b) total dataset in each class

A histogram is a visual tool commonly employed in statistics to illustrate the frequency distribution of data. A histogram is a graphical representation that illustrates the distribution of observations in various intervals or bins of a dataset. The horizontal axis of a histogram depicts the continuum of values or intervals into which the data is partitioned, sometimes referred to as bins or buckets. The vertical axis reflects the frequency or count of occurrences within each bin. Fig. 5 displays the age distribution comparison. After it applied resampling, the age values distribution is more uniform.



`**Fig. 5.** Comparison of Age distribution before and after resampling

Other visualizations of other attributes use plot boxes. The purpose of this visualization is to show the distribution of characteristics such as creatine, urine creatinine, and UACR (Urinary album concentration ratio), as shown in Fig. 6, Fig. 7, and Fig. 8. The image shows that after oversampling is applied to data sets, the distribution of data becomes more uniform in each class. By generating synthetic data points or replicating existing data from a minority class, oversampling ensures that each class is well represented in the dataset. As a result, the resulting distribution becomes more balanced, with the same number of instances in each class. This better balance reduces problems in the training of biased models and poor performance in the minority class. Therefore, with oversampling, data distribution tends to become more uniform, facilitating the training and evaluation of more powerful and reliable machine learning models in all classes.

The next step of this research is Exploration Data Analysis (EDA) using the pair plot as a graphical representation. It shows the relationship between two variables in a data set. Correlation between all pairs of numerical variables that can be indicated by scatter plot series and histograms. In diabetic nephropathy, it can reveal a correlation between the Urinary Albumin-to-Creatinine Ratio (UACR), urine creatinine, serum creatinine, and age. As an illustration, the pair plot for variables in nephropathy detection is shown in Fig. 9. Data points are scattered horizontally around a line (usually a horizontal axis), indicating that there is no visible trend between the two variables. Any value of one variable does not predict or correspond to a specific value of another variable.

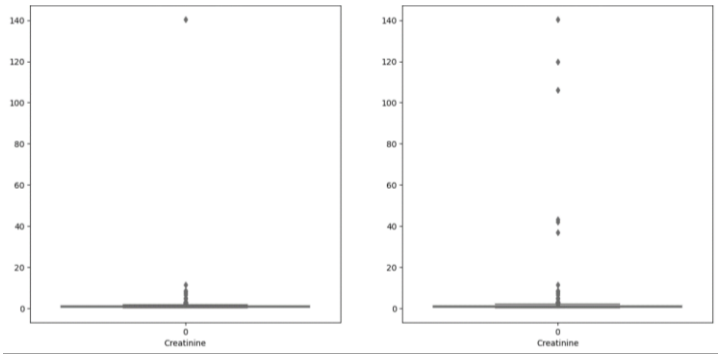


Fig. 6. Comparison of Creatinine Distribution (Boxplot) before and after resampling

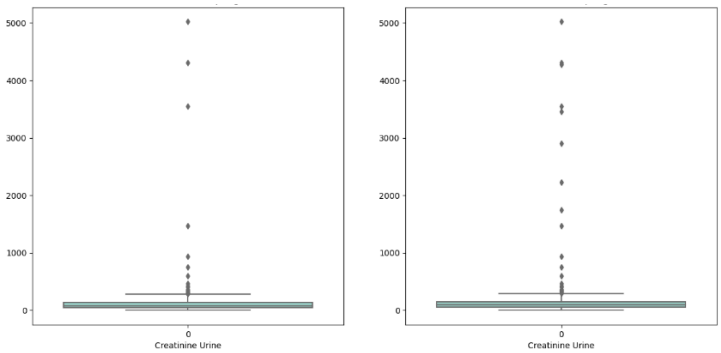


Fig. 7. Comparison of Creatine Urine Distribution (Boxplot) before and after oversampling

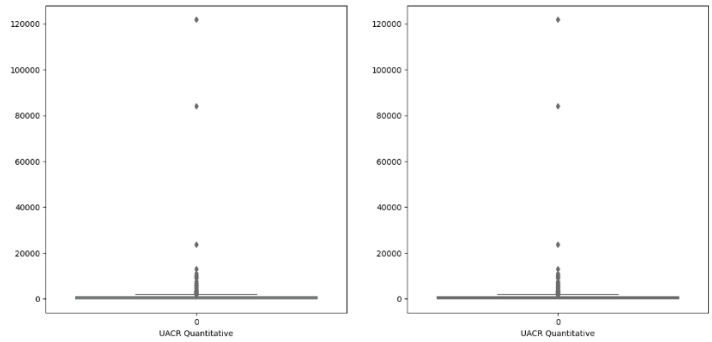


Fig. 8. Comparison of Quantitative UACR Distribution (Boxplot) Before and after oversampling

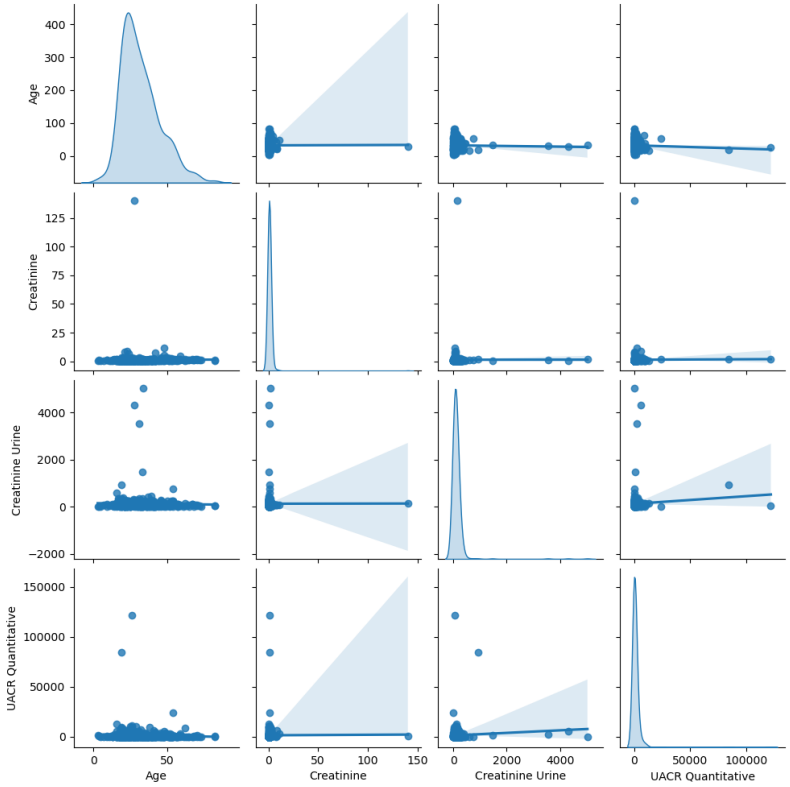


Fig. 9. Pair plot for Nephropathy Diabetic Dataset

2.2. Preprocessing Data

Noise is a disturbance, inconsistency, or error that often affects information or existing data. Data noise can occur due to several factors, such as inaccurate measurement, environmental factors, human error in data collection, or technical limitations in the data recording process. This presence of noise may interfere with accurate data analysis and interpretation. Techniques such as filtering or statistical procedures are used to reduce impact and obtain relevant information. Therefore, data cleaning is an important early stage in the research process, including filling out missing values, normalizing, and resampling data.

2.2.1. Handling Missing Values

Handling missing values is an important step in data pre-processing, especially in the context of statistical analysis and machine learning. Lost data can significantly affect the accuracy and reliability of results. One popular method for dealing with missing values is K-Nearest Neighbor Imputation (KNNI). KNN imputations are non-parametric techniques that estimate the missing value based on the closest neighbor's values in the feature space. It has been widely used in various fields such as metabolomics, epidemiology, and data-quality software because of its simplicity and effectiveness [19]. The KNNI method has shown promising results in improving the performance of classification algorithms and has been recommended as a suitable

imputation method for dealing with missing values in statistical analysis. In addition, research has shown that KNNI can be very effective in scenarios where the data mechanisms that are missing are not completely random (MCAR) or missing random (MAR) [20]. This method used statistical techniques to estimate the missing value. Efficient processing of lost values during the data preparation phase is critical to providing accurate analysis.

Related to the handling of missing values depends on the type of data used. K-Nearest Neighbors Imputation (KNNI) is a powerful technique used to deal with lost values in numerical characteristics. KNNI is a data imputation approach that uses the nearest neighbor value [13] to estimate the missing value. The steps taken in the KNN imputation are as follows:

1. To determine which numerical attributes do not have values in the collection.
2. Identify the k nearest neighbor of each observation with the missing data using the given numerical attribute.
3. Determine the distance between neighboring observations and observations with missing data. Distances can be calculated using Euclidean distances.
4. Allocate the burden to neighboring entities based on their relative distance from each other. During the imputation phase, usually, more emphasis is placed on adjacent values.
5. Calculate the unknown value by calculating the weight averages of the corresponding attribute value of the nearest neighbor k. The calculation of weight averages involves the use of coefficients that reflect the data values.
6. Iterate through each observation in the dataset that contains of missing value, through steps 2 to 5.

The problem-solving methodology employed in this study is contingent upon the data category. K-Nearest Neighbors Imputation (KNNI) is a powerful technique used for handling missing values in numerical characteristics. The KNNI is a data imputation approach that uses the values of its nearest neighbors [13] to estimate missing values. The steps involved in KNN imputation are as follows:

1. To determine which numerical attribute lacks values in the collection.
2. Identify the k nearest neighbors of each observation with missing data using the given numerical attributes.
3. Determine the distance between the neighbor observations and the observation with missing data. The distance can be calculated using the Euclidean distance.
4. Allocate weights to the neighbor entities based on their relative distances from each other. During the imputation phase, it is usual to attach more importance to adjacent values.
5. Estimate the unknown value by calculating the weighted average of the corresponding attribute values from the k nearest neighbors. The computation of the weighted average involves the utilization of coefficients that reflect the weights.
6. Iterate through each observation in the dataset that contains a missing value, as in steps 2 to 5.

Meanwhile, when working with category data, fill in the missing value using mode. This method refers to the value the category most often appears to. The imputation mode is suitable for categorical data because it is the original distribution and avoids bias. The steps for dealing with missing values in category data are as follows:

1. Identify the category attribute that has a missing value in the dataset.
2. Specify the mode, or category with the highest frequency, for each category.
3. Replace the missed value for each property with the mode value for that attribute.
4. Iterate steps 2-3 for each category attribute in the data set that does not have a value.

2.2.2. Normalization

The Min-Max scale normalization approach is used to prepare data. Normalization is a term used to describe the action of adjusting or changing data values to a consistent range. It is important to follow this procedure, especially when working with data measured using different units or scales. This work uses the Min-Max scale normalization technique for data preparation, as mentioned in the Equation (1).

$$normalized\ value = \frac{(original\ value - minimum\ value)}{(maximum - minimum)} \quad (1)$$

2.2.3. Resampling Data

Resampling techniques are widely used in medical diagnosis to address the issue of imbalanced data. Imbalanced data is a common challenge in medical datasets, and resampling methods have been successfully applied in various domains to improve classification performance [21], emphasizing the importance of resampling methods in improving the predictive power of modeling in class-imbalanced datasets, highlighting their potential application in medical literature [21]. Furthermore, Alahmari (2020) specifically investigated

the problem of class imbalance in a medical application related to autism spectrum disorder (ASD) screening, aiming to identify the ideal data resampling method that can stabilize classification performance [22].

The resampling technique aims to create a balance in the distribution of classes in the datasets. There are two main techniques for re-sampling:

a. Over-sampling refers to the technique of generating additional data points or duplicating existing ones to increase the number of instances in a minority class. By applying this approach, this model avoids bias towards the dominant class.
b. Under-sampling: This technique involves reducing the number of instances in the majority class, which is the class with the highest frequency, by selecting the data point randomly. By applying this approach, data can be reduced to the dominant class on the model and improved the ability to detect minority classes accurately.

If data cleaning and re-sampling are done at the start of the project, the data will be in satisfactory and fair condition for further analysis or modeling tasks. ADASYN is a data sampling technique widely used in machine learning that is very effective in addressing class imbalances. When there is a significant difference in the number of instances between two classes, it is mostly used for classification purposes. This could lead to the creation of models that show bias and poor performance in the minority class. ADASYN aims to balance class distribution by producing artificial samples for minority classes and adapting them to their underlying data distribution [12].

2.3. Machine Learning

Machine learning is a subdivision of artificial intelligence (AI) that allows computer systems to acquire knowledge from data so that they can make predictions without explicit programming. This method is efficient in solving complex problems and extracting valuable information from large datasets [14]. Machine learning has been used as an accurate tool in medical diagnosis, offering potential for clinical treatment. With the increasing availability of large medical datasets and advances in learning algorithms, machine-learning techniques have been applied to a wide range of medical fields, including the diagnosis of mental disorders, cancer, cardiovascular disease, and musculoskeletal conditions. These techniques have been promising in helping doctors by providing accurate prognostic predictions, disease detection, and image-based diagnosis. In addition, machine learning algorithms have become instruments in addressing challenges such as class imbalances in medical datasets, thereby improving the accuracy and reliability of predictive models [3],[23].

2.3.1. Naïve Bayes

Naive Bayes is a classification algorithm that is widely used in medical research for its simplicity, efficiency, and effectiveness [24]. This method is good for data related to statistical diagnosis, thus making it very valuable in medical science [25]. The naive Bayes method is a machine learning classification probabilistic technique. It is based on the Bayes theorem. This approach assesses the probability of a hypothesis by examining the observed data, such as the quality of the data. Naive Bayes uses probability for categorization, prediction, and training. It can be applied to a multinomial distribution, Gaussian, or Bernoulli. Some of its advantages are simplicity, user-friendliness, high dimensions, and powerful performance, even when assuming independence. This approach is also used for the diagnosis and treatment of medical conditions [26].

2.3.2. K-Nearest Neighbour (KNN)

The K-nearest neighbor algorithm (KNN) has been widely used in medical diagnosis due to its effectiveness as a non-parametric classifier. The KNN algorithm works by identifying the nearest K neighbor to a particular data point and then classifying the data according to the majority class of the neighbor. In the context of medical datasets, the selection of distance functions significantly affects the accuracy of the classification algorithm of KNN [27]. Furthermore, the KNN has been applied in the diagnosis of heart disease, where it demonstrates its ability to accurately classify patients based on relevant characteristics [28]. The K-Nearest Neighbors method is a remote-based machine learning technique that is supervised for classification and regression, effective in low-dimensional spaces but computing-intensive for large data sets [29].

2.3.3. Decision Tree (DT)

Machine learning techniques known as decision trees are used to solve classification and regression issues. Decision trees are a popular and widely used classification method in the field of machine learning and artificial intelligence (Begenova & Avdeenko, 2018). They are known for their effectiveness in classification, high speed, and ease of interpretation [30]. Decision trees are commonly used for data classification and are particularly well-suited for environments exposed to threats, making them valuable for risk management and

decision-making processes (Ziegel, 2003). They are also recognized for their ability to provide explainable machine learning, explicitly showing how different features contribute to predictions [31]. Decision trees have been applied in various fields, including healthcare, where they offer new strategies to improve the quality of care [32]. Additionally, decision trees are used for regression tasks, estimating an objective variable from an explanatory variable [33]. Overall, decision trees play a crucial role in data mining, offering a visually intuitive way to represent data partitioning and classification [34]. The tree-growing algorithm divides data into subsets based on input characteristics, with each leaf representing a prediction or class label. The algorithm selects the optimal feature during training using criteria like Gini impurity, entropy, or mean squared error. The tree-growing process ends at the end of nodes [35].

2.3.4. Linier Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a powerful statistical method widely used in medical diagnosis. LDA has been applied in various medical fields, such as spinal lesions diagnosis [36], coronary artery disease prognosis [37], and bearing fault diagnosis [38]. LDA is a machine learning technique used to detect diabetic nephropathy in this study. It requires a dataset with class labels and predictor variables. LDA reduces dimensionality by finding the best linear combination of predictors, maximizes class separability by maximizing between-class variance, and then uses it for classification based on learned linear discriminant functions [39].

2.3.5. Support Vector Machine (SVM)

Support Vector Machines (SVMs) have gained significant attention in medical diagnostics due to their ability to handle complex and high-dimensional data. This method is suitable for disease diagnosis and prognosis. SVM has been widely used in the medical field, offering efficient and accurate diagnosis of a variety of diseases, including liver disease and heart disease [40]. In addition, SVM has been applied in disease infection detection in health care [41]. The use of SVMs in medical diagnostics is more supported by their ability to handle unbalanced data so that it is suitable for classification tasks with unequal class distribution [42]. The accuracy and efficiency of the SVM in medical diagnosis have been demonstrated in various studies, with a precision of 98.96% in classifying anxiety, depression, and stress in social media users [43]. Overall, SVMs offer a strong and effective approach to medical diagnosis, leveraging their ability to handle complex data and deliver accurate classification results. Support Vector Machine (SVM) is a supervised machine learning technique used for classification and regression problems. This method maximizes the margin between classes and identifies the optimal hyperplane for the data point. The kernel and regularization parameters are the primary hyperparameter. SVMs are efficient, adjustable, and resistant to over-fitting, but require high computational costs and are sensitive to the choice of kernels and hyperparameters [44].

2.4. Ensemble Machine Learning

An ensemble classifier is more reliable than a single classifier, thus achieving a classification accuracy of 99.5% on test data [45]. In some previous studies, the Ensemble method by combining several models could improve prediction accuracy and overall performance [46]. Ensemble methods in machine learning are widely used involving the merger of many different models to build more robust predictive models. The Ensemble approach aims to exploit the collective intelligence of several models by combining predictions to produce more accurate predictions than a single model. The Ensemble approach usually begins with collecting a series of basic models, which are also called weak classifications. Basic models cover a variety of machine-learning approaches, including decision trees, support vector machines, logistical regression, and neural networks [20]. Each fundamental model has its advantages and disadvantages. This method is to combine several models or algorithms to improve prediction performance, such as ensemble techniques.

- Voting: In this methodology, many models provide predictions, and the final prediction is chosen by a majority vote (for classification) or an average (for regression).
- Bagging, also known as Bootstrap Aggregating, is a technique that involves training many iterations of the basic model using different parts of the training data. It reduces variability and reduces the risk of over-fit.
- Random Forest: A particular form of bagging that uses decision trees as its underlying model. It combines random by selecting a random subset of features during each split.
- Boosting is a technique that aims to improve the performance of low classification by giving greater attention to samples that were previously misclassified. This approach involves the aggregation of predictions produced by several models, in which each model is specifically trained to correct the inaccuracies made by the previous model. AdaBoost, also known as Adaptive Boosting, is a method widely

used for boosting. It sets a varying burden on training instances, focusing on emphasizing the importance of misclassified data.

2.5. Performance Evaluation

Accuracy, along with precision, recall, and F1 score, is a widely used metric for assessing the effectiveness of a classification model. Below are the steps to compute accuracy, precision, recall, and F1 score for the identification of nephropathy. The accuracy in Equation (2) refers to the percentage of individuals correctly classified in the dataset, which includes both true positives and true negatives. The accuracy of a classification model can be calculated using the formula:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

where TP, TN, FP, and FN indicate the number of true positives, true negatives, false positives, and false negatives, respectively. False positives are instances where nephropathy is incorrectly detected, leading to the prediction of normal or healthy persons. Precision refers to the proportion of individuals accurately recognized as having nephropathy among all those predicted to have nephropathy. The formula for precision is TP divided by the sum of TP and FP, as represented by the equation (3).

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

Then, the proportion of true positives (those who are predicted to have nephropathy and do so) among all people who have indicated nephropathy is known as recall.

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

The F1 score is the harmonic mean of precision and recall and provides a balanced measure of the two metrics.

$$F1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \tag{5}$$

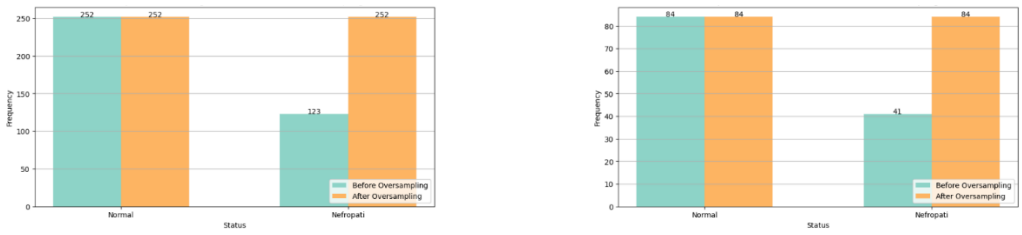
Note that these metrics should be evaluated together to get a comprehensive understanding of how well the nephropathy detection model performs. Different thresholds for detection may lead to different values of these metrics, and the choice of threshold should be based on the specific application and goals of the model.

3. RESULTS AND DISCUSSION

3.1. Experimental Result

This research is applied using 500 data sets of diabetic patients from medical records in Saiful Anwar Hospital. By splitting the dataset, we allocate 75% (375 samples) of the dataset for training and 25% (125 samples) for testing. The detailed dataset before and after oversampling is shown in Fig. 10. The data in the minority class is increased to the same number of data as in the majority class.

The performance measures (recall, precision, and f1 score) are evaluated to represent machine learning, including K-Nearest Neighbor, Naïve Bayes, SVM, Decision Tree, LDA, Random Forest, Bagging, and AdaBoost. The performance results for all algorithms are shown in Fig. 11. The ensemble methods, including Random Forest, Bagging, and AdaBoost, are stable and obtain a high performance of more than 90%. By resampling, the proposed method has obtained an increase of 4% for all the performance rates, as shown in Fig. 9. On the other hand, LDA has the lowest performance evaluation, either with or without oversampling implementation. This method is not suitable for implementation in this dataset.



(a) (b)

Fig. 10. Comparison of Before and After Oversampling (a) Training data (b) Testing dat

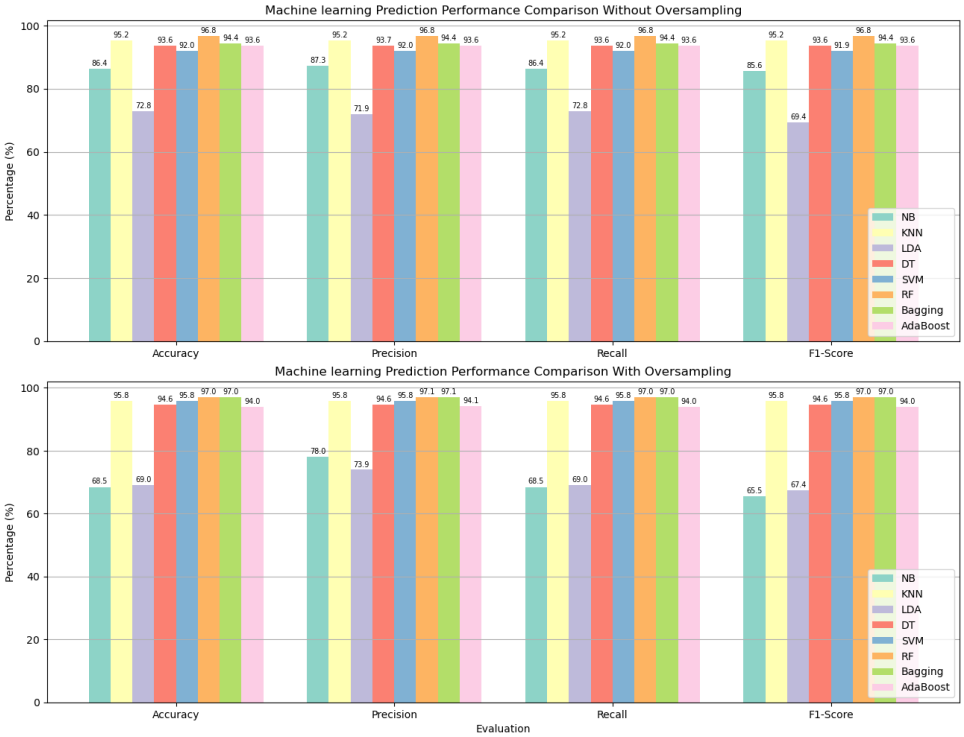


Fig. 11. Performance Comparison of Naïve Bayes, KNN, LDA, DT, SVM, RF, Bagging, and AdaBoost

3.2. Discussion

The purpose of this work was to identify neuropathy by employing an ensemble learning technique in conjunction with the Adaptive Synthetic Sampling (ADASYN) resampling method. This methodology was used to tackle the issue of class imbalance and finally achieve reliable and effective classification with optimal performance. The experimental results confirm the efficacy of the proposed methodology. The ensemble technique, which merged many base classifiers, demonstrated resilience in dealing with the complicated and complex characteristics of neuropathy identification.

In performance evaluation of medical diagnostics, the concepts of true positive, true negative, false positive, and false negative are fundamental in assessing the accuracy and reliability of diagnostic tests. True positive refers to cases where the test correctly identifies the presence of a condition, such as detecting Diabetic Nephropathy disease in patients. Conversely, true negatives denote cases where the absence of a condition is correctly identified, as seen in the analysis of true negatives. On the other hand, a false positive occurs when the test incorrectly indicates the presence of a condition, as observed in the evaluation of the application. Finally, a false negative occurs when the absence of a condition is incorrectly identified, as evidenced in the study on application. These concepts are crucial in understanding the diagnostic accuracy and reliability of medical tests, thereby influencing clinical decision-making and patient care [47],[48]. The incorporation of various base classifiers into the ensemble facilitated improved generalization and acquisition of knowledge from multiple parts of the input data, hence boosting the overall performance metrics, as shown in Table 1. The experimental result showed that there was a significant increment in the total number of TN after the data was applied over-sampling. It affects to performance evaluation measurements such as precision, recall, and f1-score.

Table 1. Confusion Matrix of Ensemble Method: Random Forest, Bagging, and AdaBoost

Resampling	Random Forest				Bagging				AdaBoost			
	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN
Without Oversampling	82	2	2	39	80	3	4	38	80	4	4	37
With Oversampling	80	1	4	83	80	1	4	83	80	6	4	78

Furthermore, the implementation of ADASYN resampling had a crucial role in reducing the problem of class imbalance commonly found in neuropathy datasets. The capability of ADASYN to produce synthetic samples for under-represented classes resulted in a more equitable distribution within the dataset. The

technique of rebalancing was crucial in improving the learning process by eliminating bias towards the majority class in classifiers. This resulted in more precise and representative training of the models. An impressive result of our work was the attainment of stability in high-performance categorization. The combination of the ensemble technique with ADASYN consistently exhibited improved performance across multiple evaluation measures, including accuracy, precision, recall, and F1-score.

Stability is essential in practical clinical applications, where the detection of neuropathy needs to be consistent and dependable for precise diagnosis and timely action. Overall, the combination of the ensemble method and ADASYN resampling demonstrated a highly effective and reliable methodology for detecting neuropathy, resulting in consistent and superior classification performance. The results of this work provide a basis for future research, which may involve investigating more complex ensemble strategies and incorporating advanced resampling approaches. Ultimately, this will contribute to improving the accuracy of diagnosing neuropathy.

4. CONCLUSION

Detecting nephropathy based on clinical data was applied using an ensemble method of machine learning. In this research, we applied ensemble methods, including Random Forest, AdaBoost, and Bagging algorithm, by aggregating the classifier as leaner. Due to imbalanced data distribution in each class, we proposed resampling using Adaptive Synthetic Sampling Data (ADASYN). By over-sampling data using ADASYN, the proposed method achieved a high performance of 90% and above. It means that the performance evaluation increased to 4% for precision, recall, and f1-score.

Acknowledgments

This research is supported financially by the Faculty of Computer Science, Brawijaya University, through the Research Grant Program "Hibah Doktor Lektor Kepala 2023" under contract number: 2120/UN10.F15/PN/2023 was dated June 6, 2023.

REFERENCES

[1] N. Samsu, "Diabetic Nephropathy: Challenges in Pathogenesis, Diagnosis, and Treatment," *BioMed Research International*, vol. 2021, p. e1497449, 2021, https://doi.org/10.1155/2021/1497449.

[2] I. H. de Boer, T. C. Rue, Y. N. Hall, P. J. Heagerty, N. S. Weiss, and J. Himmelfarb, "Temporal Trends in the Prevalence of Diabetic Kidney Disease in the United States," *JAMA*, vol. 305, no. 24, pp. 2532–2539, 2011, https://doi.org/10.1001/jama.2011.861.

[3] R. de Filippis *et al.*, "Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review," *NDT*, vol. 15, pp. 1605–1627, 2019, https://doi.org/10.2147/NDT.S202418.

[4] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001, https://doi.org/10.1016/S0933-3657(01)00077-X.

[5] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform*, vol. 2, pp. 59–77, 2007, https://doi.org/10.1177/117693510600200030.

[6] Z. Zhou, "Breast Cancer Diagnosis with Machine Learning," *Highlights in Science, Engineering and Technology*, vol. 9, pp. 73–75, 2022, https://doi.org/10.54097/hset.v9i.1718.

[7] S. Wang, J. Chen, and S. Gong, "Extracting critical data from medical images based on machine learning," in *Second International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2022)*, pp. 516–520, 2022, https://doi.org/10.1117/12.2659592.

[8] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015, https://doi.org/10.1126/science.aaa8415.

[9] V. Lai, S. Carton, and C. Tan, "Harnessing explanations to bridge ai and humans," *arXiv preprint arXiv:2003.07370*, 2020, https://doi.org/10.48550/arXiv.2003.07370.

[10] R. Prabha, G. A. Senthil, D. A. Lazha, D. VijendraBabu, and M. D. Roopa, "A Novel Computational Rough Set Based Feature Extraction For Heart Disease Analysis," In *I3CAC 2021: Proceedings of the First International Conference on Computing, Communication and Control System, I3CAC 2021*, p. 371, 2021, https://doi.org/10.4108/eai.7-6-2021.2308575.

[11] A. Gholaminejad, M. Fathalipour, and A. Roointan, "Comprehensive analysis of diabetic nephropathy expression profile based on weighted gene co-expression network analysis algorithm," *BMC Nephrology*, vol. 22, no. 1, p. 245, 2021, https://doi.org/10.1186/s12882-021-02447-2.

[12] N. Kaur, S. Bhattacharya, and A. J. Butte, "Big Data in Nephrology," *Nat Rev Nephrol*, vol. 17, no. 10, 2021, https://doi.org/10.1038/s41581-021-00439-x.

[13] H. He and Y. Ma, Eds. *Imbalanced Learning: Foundations, Algorithms, and Applications*. 1st edition. Hoboken. New Jersey: Wiley-IEEE Press, 2013, https://doi.org/10.1002/9781118646106.

[14] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, pp. 853–867, 2005, https://doi.org/10.1007/0-387-25465-X_40.
[15] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions," *ACM Comput. Surv.*, vol. 52, no. 4, p. 79:1-79:36, 2019, https://doi.org/10.1145/3343440.
[16] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, 2008, https://doi.org/10.1109/IJCNN.2008.4633969.
[17] V. K. Gajjar, A. K. Nambisan, and K. L. Kosbar, "Plant Identification in a Combined-Imbalanced Leaf Dataset," *IEEE Access*, vol. 10, pp. 37882–37891, 2022, https://doi.org/10.1109/ACCESS.2022.3165583.
[18] W. Chu, C. S. Ho, and P. H. Liao, "Comparison of different predicting models to assist the diagnosis of spinal lesions. *Informatics for Health and Social Care*, vol. 47, no. 1, pp. 92-102, 2022, https://doi.org/10.1080/17538157.2021.1939355.
[19] J. P. Dekermanjian, E. Shaddox, D. Nandy, D. Ghosh, and K. Kechris, "Mechanism-aware imputation: a two-step approach in handling missing values in metabolomics," *BMC Bioinformatics*, vol. 23, no. 1, p. 179, 2022, https://doi.org/10.1186/s12859-022-04659-1.
[20] J. Huang *et al.*, "Cross-validation based *K* nearest neighbor imputation for software quality datasets: An empirical study," *Journal of Systems and Software*, vol. 132, pp. 226–252, 2017, https://doi.org/10.1016/j.jss.2017.07.012.
[21] P. H. Lee, "Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets," *International Journal of Environmental Research and Public Health*, vol. 11, no. 9, 2014, https://doi.org/10.3390/ijerph110909776.
[22] F. Alahmari, "A Comparison of Resampling Techniques for Medical Data Using Machine Learning," *J. Info. Know. Mgmt.*, vol. 19, no. 01, p. 2040016, 2020, https://doi.org/10.1142/S021964922040016X.
[23] O. Kudina and B. de Boer, "Co-designing diagnosis: Towards a responsible integration of Machine Learning decision-support systems in medical diagnostics," *Journal of Evaluation in Clinical Practice*, vol. 27, no. 3, pp. 529–536, 2021, https://doi.org/10.1111/jep.13535.
[24] J. C. Tesoro, "A Semantic Approach of the Naïve Bayes Classification Algorithm," *IJATCSE*, vol. 9, no. 3, pp. 3287–3294, 2020, https://doi.org/10.30534/ijatcse/2020/125932020.
[25] V. Melinda, R. Primartha, A. Wijaya, and M. I. Jambak, "Optimization Naive Bayes Algorithm Using Particle Swarm Optimization in the Classification of Breast Cancer," In *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, pp. 362-369, 2020, https://doi.org/10.2991/aisr.k.200424.055.
[26] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J Supercomput*, vol. 77, no. 5, pp. 5198–5219, 2021 https://doi.org/10.1007/s11227-020-03481-x.
[27] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, p. 1304, 2016, https://doi.org/10.1186/s40064-016-2941-7.
[28] J. Zeniarja, A. Ukhifahdhina, and A. Salam, "Diagnosis Of Heart Disease Using K-Nearest Neighbor Method Based On Forward Selection," *Journal of Applied Intelligent System*, vol. 4, no. 2, 2019, https://doi.org/10.33633/jais.v4i2.2749.
[29] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets," *SN Appl. Sci.*, vol. 1, no. 12, p. 1559, 2019, https://doi.org/10.1007/s42452-019-1356-9.
[30] N. Rochmawati *et al.*, "Covid Symptom Severity Using Decision Tree," in *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)*, pp. 1–5, 2020, https://doi.org/10.1109/ICVEE50212.2020.9243246.
[31] H. E. C. Cao, R. Sarlin, and A. Jung, "Learning Explainable Decision Rules via Maximum Satisfiability," *IEEE Access*, vol. 8, pp. 218180–218185, 2020, https://doi.org/10.1109/ACCESS.2020.3041040.
[32] L. Keikes *et al.*, "Conversion of a colorectal cancer guideline into clinical decision trees with an assessment of validity," *International Journal for Quality in Health Care*, vol. 33, no. 2, p. mzab051, 2021, https://doi.org/10.1093/intqhc/mzab051.
[33] K. Yawata, Y. Osakabe, T. Okuyama, and A. Asahara, "QUBO Decision Tree: Annealing Machine Extends Decision Tree Splitting," in *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 355–364, 2022, https://doi.org/10.1109/ICKG55886.2022.00052.
[34] S. Tsang, B. Kao, K. Y. Yip, W.-S. Ho, and S. D. Lee, "Decision Trees for Uncertain Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 64–78, 2011, https://doi.org/10.1109/TKDE.2009.175.
[35] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: a review," *Procedia Manufacturing*, vol. 35, pp. 698–703, 2019, https://doi.org/10.1016/j.promfg.2019.06.011.
[36] S. M. S. Dashti and S. F. Dashti, "An expert system to diagnose spinal disorders," *arXiv preprint arXiv:2302.03625*, 2023, https://doi.org/10.48550/arXiv.2302.03625.
[37] S. Shariatnia, M. Ziaratban, A. Rajabi, A. Salehi, K. Abdi Zarrini, and M. Vakili, "Modeling the diagnosis of coronary artery disease by discriminant analysis and logistic regression: a cross-sectional study," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 85, 2022, https://doi.org/10.1186/s12911-022-01823-8.
[38] H. Ye, P. Wu, Y. Huo, X. Wang, Y. He, X. Zhang, and J. Gao, "Bearing Fault Diagnosis Based on Randomized Fisher Discriminant Analysis," *Sensors*, vol. 22, no. 21, p. 8093, 2022, https://doi.org/10.3390/s22218093.

[39] H. Zhao, Z. Lai, H. Leung, and X. Zhang, "Linear Discriminant Analysis," in *Feature Learning and Understanding: Algorithms and Applications*, pp. 71–85, 2020, https://doi.org/10.1007/978-3-030-40794-0_5.
[40] D. Devikanniga, A. Ramu, and A. Haldorai, "Efficient Diagnosis of Liver Disease using Support Vector Machine Optimized with Crows Search Algorithm," *EAI Endorsed Transactions on Energy Web*, vol. 7, no. 29, 2020, https://doi.org/10.4108/eai.13-7-2018.164177.
[41] J. I. Park, D. Z. Bliss, C.-L. Chi, C. W. Delaney, and B. L. Westra, "Knowledge Discovery With Machine Learning for Hospital-Acquired Catheter-Associated Urinary Tract Infections," *CIN: Computers, Informatics, Nursing*, vol. 38, no. 1, p. 28, 2020, https://doi.org/10.1097/CIN.0000000000000562.
[42] J. Liu and E. Zio, "Integration of feature vector selection and support vector machine for classification of imbalanced data," *Applied Soft Computing*, vol. 75, pp. 702–711, 2019, https://doi.org/10.1016/j.asoc.2018.11.045.
[43] T. M. Wijiasih, R. N. S. Amriza, and D. A. Prabowo, "The Classification of Anxiety, Depression, and Stress on Facebook Users Using the Support Vector Machine," *JISA(Jurnal Informatika dan Sains)*, vol. 5, no. 1, 2022, https://doi.org/10.31326/jisa.v5i1.1273.
[44] M. Awad and R. Khanna, "Support Vector Machines for Classification," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pp. 39–66, 2015, https://doi.org/10.1007/978-1-4302-5990-9_3.
[45] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," *IEEE Access*, vol. 9, pp. 67488–67500, 2021, https://doi.org/10.1109/ACCESS.2021.3075140.
[46] E. Yaman and A. Subasi, "Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification," *BioMed Research International*, vol. 2019, pp. 1–13, 2019, https://doi.org/10.1155/2019/9152506.
[47] T. Yokoo *et al.*, "Linearity, Bias, and Precision of Hepatic Proton Density Fat Fraction Measurements by Using MR Imaging: A Meta-Analysis," *Radiology*, vol. 286, no. 2, pp. 486–498, 2018, https://doi.org/10.1148/radiol.2017170550.
[48] J. D. Pleil, M. A. G. Wallace, M. A. Stiegel, and W. E. Funk, "Human biomarker interpretation: the importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates," *Journal of Toxicology and Environmental Health, Part B*, vol. 21, no. 3, pp. 161–180, 2018, https://doi.org/10.1080/10937404.2018.1490128.

BIOGRAPHIES OF AUTHORS

Lailil Muflikhah received a B.Sc. degree in computer science from the Institute Technologi Sepuluh Nopember, an M.Sc. degree in computer science from the Universiti Technology of Petronas, Malaysia, and a Ph.D. degree in Bioinformatics from The University of Brawijaya. She is currently an Associate Professor with the Department of Informatics Engineering in the Faculty of Computer Science, University of Brawijaya. Her research interests include soft computing, machine learning, and intelligent systems. She can be contacted at email: lailil@ub.ac.id.

Fitra Abdurrachman Bachtiar is a lecturer in the Faculty of Computer Science at Brawijaya University. He graduated in Electrical Engineering from Brawijaya University in 2008 and from Ritsumeikan University Graduate School of Science and Engineering in 2011. He received the Dr. Eng from Ritsumeikan University in 2016. His research interests are affective computing, affective engineering, intelligent systems, and data mining. He can be contacted at fitra.bachtiar@ub.ac.id.

Dian Eka Ratnawati is an educator at the Faculty of Computer Science, Brawijaya University, Indonesia. She earned her bachelor's and master's degrees from the Sepuluh Nopember Institute of Technology (ITS). Her master's studies focused on Computer Science at ITS, with research centered on Hybrid Genetic Algorithms for job shop scheduling. Continuing her academic journey, Dian pursued her Ph.D. at Brawijaya University (UB), delving into the field of Bioinformatics. Her research specifically explored the utilization of Neural Networks (ELM) and PSO Optimization for drug prediction through SMILES code. Dian Eka Ratnawati made significant contributions to the field of PSO Optimization algorithms. She successfully developed two crucial parameters that introduced innovation into PSO, namely the adaptive inertia weight known as Modified SAIW and the adaptive acceleration coefficient known as Modified SBAC. These discoveries have enhanced the efficiency of the PSO algorithm, enriching the knowledge landscape in the field of computational science. She can be contacted at dian_ilkom@ub.ac.id.

Riski Darmawan is a fresh graduate from the Faculty of Computer Science, Brawijaya University, Indonesia. He earned his bachelor's degree in 2023. He is passionate in Data Science, aiming to leverage analytical skills and creativity in data manipulation and contribute to the company's success. He enjoys learning new things and always seeks to improve his skills and knowledge. He can be contacted at riskidarmawan@student.ub.ac.id.