Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal Vol. 6, No. 3, July 2024, pp: 312-321; eISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Manuscript received May 2, 2024; revised May 23, 2024; accepted May 27, 2024; date of publication July 8, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i3.452

How to cite: Edi Ismanto, Abdul Fadlil, Anton Yudhana and Kodai Kitagawa, "A Comparative Study of Improved Ensemble Learning Algorithms for Patient Severity Condition Classification", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 3, pp. 312-321, July 2024.

A Comparative Study of Improved Ensemble Learning Algorithms for Patient Severity Condition Classification

Edi Ismanto[1,2], Abdul Fadlil[2], Anton Yudhana[2], and Kodai Kitagawa[3]
[1] Department of Informatics, Universitas Muhammadiyah Riau, Pekanbaru, Indonesia
[2] Departement of Master of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3] Mechanical and Medical Engineering Course, National Institute of Technology, Hachinohe College, Hachinohe, Japan

Corresponding author: Edi Ismanto (edi.ismanto@umri.ac.id).

ABSTRACT The evolution of Electronic Health Records (EHR) has facilitated comprehensive patient record-keeping, enhancing healthcare delivery and decision-making processes. Even with these developments, there are still certain difficulties when employing ensemble machine-learning techniques to analyze EHR data. This study aims to model the classification of patient severity using EHR data. Addressing issues with dimensionality and imbalance in EHR data. and to avoid overfitting by optimizing the ensemble model. The principal component analysis (PCA) method is used to address data dimensionality issues, and the synthetic minority oversampling technique (Smote) method is used to address data imbalance issues. After that, the ensemble model's hyperparameters are optimized using the Grid Search and Random Search approaches to prevent overfitting. In light of the study findings, the ensemble model's accuracy significantly improves after correcting data imbalance and dimensionality reduction. Notably, the Gradient Boosting Machine (GBM) and CatBoost models exhibited superior performance with an accuracy of 73%, achieved through experiments involving dimensionality reduction and handling of imbalanced data. Furthermore, optimization techniques such as Grid Search and Random Search were employed to enhance the EML models. The results of model optimization revealed that the GBM + Random Search model performed the best, achieving an accuracy of 74%, followed by the XGBoost + Grid Search model with an accuracy of 73%. The GBM model also excelled in distinguishing between positive and negative classes, boasting the highest Area under Curve (AUC) value of 0.78, indicative of its superior classification capabilities compared to other models. The study's findings offer a precise severity classification that medical professionals and teams can use to make quicker and more informed clinical decisions based on a patient's condition.

INDEX TERMS Classification model, Dimensionality reduction, Electronic health record (EHR), Ensemble learning, Hyperparameter optimization (HPO), Unbalanced data.

I. INTRODUCTION

Together with improvements in information technology in the medical field, the implementation of Electronic Health Records (EHR) in hospitals has become the main focus of efforts to improve patient care standards and the efficacy of the healthcare system as a whole [1]. An electronic health record, or EHR, is a contemporary type of medical record that records a patient's entire medical history, lab results, prescriptions, and other health-related information digitally and eliminates the need for paper records [2]. A wealth of information can be gleaned from EHR data, including information on medication use, laboratory test results, chronic disease management, patient disease history, and prognostic and predictive analysis regarding the patient's health development [3]. By carefully analyzing EHR data and making use of technology, health professionals can improve patient care overall, make better decisions, and manage the disease more successfully [4]. Numerous intricate issues frequently arise when analyzing data from electronic health records (EHRs). One of the primary problems is data quality since EHR data frequently contains mistakes, missing values, or format inconsistencies [5]. The second issue is interoperability, which hinders data integration and information sharing between various EHR systems [6]. Thus, in the context of health care, choosing the appropriate Machine Learning (ML) model for analyzing EHR data is crucial. The accuracy of the predictions made by models directly influences the clinical decisions made by medical professionals. Proper models can yield more precise forecasts concerning the likelihood of an illness, the way treatment will work, and the prognosis of an individual [7]. Better interpretation of the analysis's findings is also crucial, and clear-cut models can offer a deeper understanding of the variables affecting the patient's health. Scalability and efficiency considerations are also necessary since EHR data frequently has a high volume and degree of complexity [8]. Effective data processing models can guarantee timely and accurate analysis. Other crucial factors to take into account when choosing the right model for EHR data are generalizability, suitability for the clinical setting, sustainability of the model to data variability, and analysis goals. Previous research has not thoroughly investigated the relationship between data quality, interoperability, and the selection of appropriate ML models to address prediction accuracy issues in EHRs [9, 10, 11, 12]. There remains a gap in understanding how data quality and interoperability can affect ML model performance in the context of EHRs. Additionally, the issues of data imbalance and overfitting in ML models have not been sufficiently addressed [13, 14, 17]. This study aims to tackle these problems by exploring and developing more effective and efficient ML models for analyzing EHR data, with a focus on overcoming data imbalance and mitigating overfitting. By doing so, the study seeks to provide more accurate and reliable results for medical professionals in clinical decision-making.

Research question: How can Ensemble machine learning (EML) models improve the classification of patient condition status and disease severity using EHR data? How can imbalanced and highly dimensional EHR data be resolved? How can Ensemble machine learning (EML) models be optimized for hyperparameters to solve EHR data overfitting problems? This study makes a significant contribution by initially tackling crucial concerns associated with the dimensionality of data and the imbalance within EHR datasets. Additionally, it enhances the performance accuracy of ensemble machine-learning algorithms for addressing classification tasks. Lastly, it facilitates expedited and more accurate clinical decision-making based on patient health status.

A. RELATED WORKS

This study examines several earlier investigations that address Ensemble machine learning (EML) techniques used in EHR data analysis. The purpose of the review was to obtain a deeper comprehension of the methodologies employed in health data analysis. In addition, weighing the benefits and drawbacks of every Ensemble machine learning (EML) technique. Based on trends and abnormalities in EHR data, research [13] employs ensemble machine learning models to identify and categorize different types of strokes. Concerning accuracy, the best algorithms for classification are Random Forest, Extremely Randomized Trees, and Histogram-Based Gradient Boosting. The study [14] uses big data fusion and ensemble learning algorithms to predict and classify breast cancer risk; the algorithm with the best accuracy performance is the XGBoost algorithm. To enhance the quality of care, research [15] attempts to predict the risk of falls early. The XGBoost algorithm can identify falls that occur between 54.93% and 58.01% of the time. The application of the widely used Society of Thoracic Surgeons risk score to evaluate the risk of morbidity and mortality in cardiac surgery was covered in the study [16]. The XGBoost Algorithm generated the best predictor, according to the results. In the test cohort, eXtreme Gradient Boosting consistently performed better than the Society of Thoracic Surgeons model when assessed on the index procedure.

The best model is shown by random forest algorithms in research [17], which explores the enormous potential that electronic medical records (EHR) have in creating vast and intricate medical databases that could be an effective tool for clinical research. Predicting a woman's risk of pre-eclampsia (PE) is the goal of research [18]. The model predictions were constructed using logistic regression (LR), random forest (RF), support vector machines (SVM), and extreme gradient boosting (XGBoost). The study's findings indicate that the XGboost model performs the best in terms of predictions. A study [19] uses electronic medical records to guide screening decisions for esophageal adenocarcinoma (EAC) and cardia adenocarcinoma (GCA) cancers. The study's findings indicate that the extreme gradient boosting algorithm performs discrimination the best and most accurately. With data from Electronic Medical Records (EHRs) up to two years of age, the study (Xueqin Pang) attempts to predict obesity in children aged > two to ≤ seven years. The outcome is that XGBoost outperforms all other models, yielding an AUC of 0.81 (0.001). The study [20] employed the best Random Forest (RF) Model based on the Area under the Receiver Operating Characteristic Curve results to offer fresh insights into the course of inpatients. The study (Mahesh T R) assessed the methods for diagnosing and predicting breast cancer. A variety of machine learning (ML) algorithms were employed, such as Majority-voting, eXtreme Gradient Boosting algorithm (XGBoost), Random Forest (RF), K-Nearest Neighbours (KNN), Classification and Regression Tree (CART), Logistic Regression (LR), and Naive Bayes (NB) for breast cancer classification. Based on the top three classifiers (LR, SVM, and CART), the Majority-Voting ensemble method performs better than all other individual classifiers, according to the result. The results of earlier studies indicate that the Ensemble machine learning (EML) approach performs better when analyzing EHR data. Nevertheless, overfitting to unbalanced data is a common issue with ensemble models. The issues of noise, high dimensions, and unbalanced data must be resolved to optimize the model and prevent overfitting. This is what sets

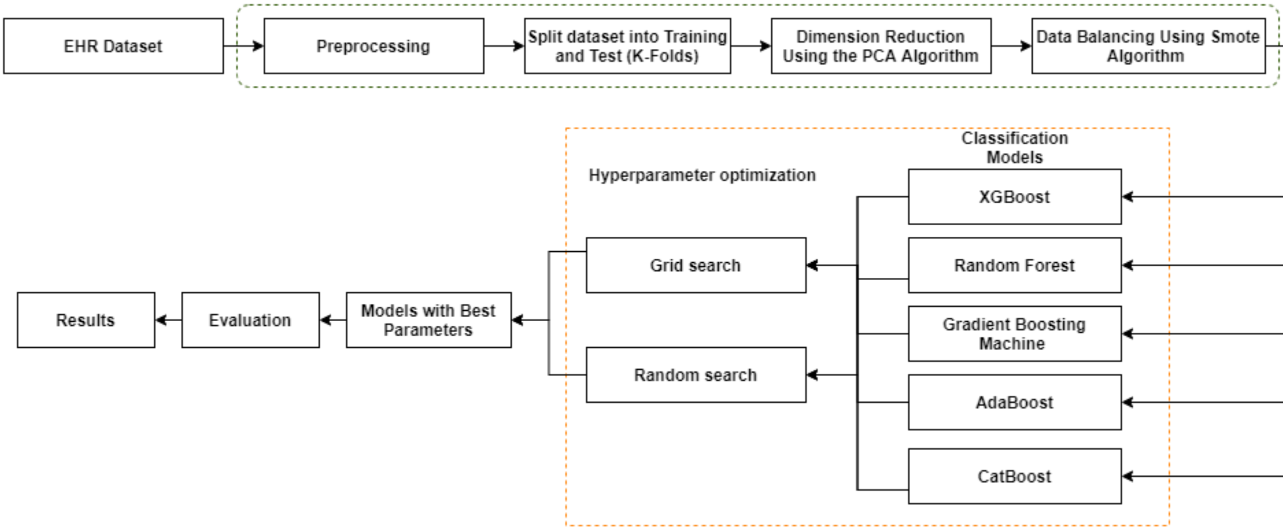


FIGURE 1. Phases of the research approach used.

TABLE 1
An explanation of the features' contents in the dataset [21]

Feature Name	Data Type	Description
Hematocrit	Float	The laboratory test results for the patient's hematocrit
Hemoglobin	Float	The laboratory test results for the patient's hemoglobin
Erythrocyte	Float	The laboratory test results for the patient's erythrocyte
Leucocyte	Float	The laboratory test results for the patient's leucocyte
Thrombocyte	Numeric	The laboratory test results for the patient's thrombocyte
Mean Corpuscular Hemoglobin (MCH)	Float	The laboratory test results for the patient's MCH
Mean Corpuscular Hemoglobin Concentration (MCHC)	Float	The laboratory test results for the patient's MCHC
Mean Corpuscular Volume (MCV)	Float	The laboratory test results for the patient's MCV
Age	Numeric	Patient's age
Sex	Nominal	Patient's gender
Severity Level	Nominal	Target label, Severe = Inpatient, Mild = Outpatient

this research apart from earlier studies. This research makes a significant contribution to the development of Ensemble machine learning (EML) approaches for Electronic health records (EHR) analysis. Focusing on handling noise, high dimensionality, and data imbalance in EHR, the study selects appropriate base models, applies complex parameter tuning, and utilizes various model optimization methods. By employing techniques like synthetic minority over-sampling, Principal component analysis (PCA), and tuning parameters, the research compares the performance of machine learning ensembles in EHR data analysis. By accurately classifying patients' severity levels, medical professionals and teams can expeditiously make more informed clinical decisions.

II. METHODOLOGY

This study specifically selected several ensemble learning models, the effectiveness of which has been assessed in previous studies. Since this aligns with our research context and objectives, we concentrate on models that demonstrate potential in managing large and unbalanced data sets. We selected ensemble learning models that offer a more accurate way to classify patient disease outcomes. This selection aims to give a thorough overview of different approaches and how well they work in comparable situations. A schematic of the research methodology is presented in FIGURE 1, which includes the various stages involved in model comparison. The objective of this study is to enhance classification performance metrics related to data imbalance issues and parameter optimization in ensemble learning models.

A. DATASETS

The study utilized publicly accessible patient electronic health record data from private hospitals in Indonesia [21]. Contains information from a laboratory report about the patient's health, including check results for erythrocytes, leucocytes, thrombocytes, hemoglobin, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH) and mean corpuscular hemoglobin concentration (MCHC). The dataset contains records of 4412 patient examination outcomes. The total amount of data utilized in this study was 4412. Two data classes are available: Severe and Mild. The numbers for the mild class were 2628 and the severe class 1784. A total of 3970.8, or 90% of the data, are used for training, and 441.2, or 10%, are used for testing. There are 11 features in the dataset. TABLE 1 provides comprehensive details about the features of the data. In this study, unbalanced data was handled using the Synthetic Minority Over-sampling Technique (SMOTE) method, while dimension reduction was accomplished using the Principal Component Analysis (PCA) method.

DATA PREPROCESSING

Data preprocessing involves cleaning, transforming, and organizing data to make it easier for the model to understand [22]. EHR data is processed using scaling, encoding, and dimension reduction techniques. Through the use of the k-fold cross-validation technique, the data is split into training and testing subsets: 90% for training and 10% for testing. The dimension reduction technique is implemented through the Principal component analysis (PCA) method to reduce data complexity and extract significant information from high-dimensional data sets. Oversampling techniques, such as the Synthetic minority over-sampling technique (SMOTE) method, are also employed to address data imbalances. There are two label classes: mild conditions fall under class 0, and severe conditions fall under class 1. The label data for classes 0 and 1 is shown in FIGURE 2. The ensemble learning model's hyperparameters were optimized using both the Grid search and Random search techniques.

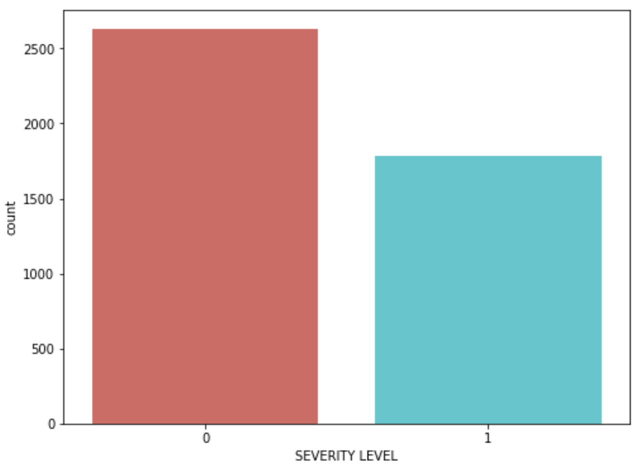


FIGURE 2. Labels for data classes 0 and 1

. K-FOLD CROSS VALIDATION

A dataset can be divided into multiple parts (folds), or K-folds, using K-fold cross-validation, where K is the desired number of parts [23]. There were K iterations in the training and evaluation process. The validation set is the set from the first fold, and the training set is the other fold in each iteration. The average of each iteration's model performance is used to determine the model's performance. Because the model is evaluated using multiple combinations of training and validation data, K-fold helps to reduce the variability of evaluation results [24, 25]. Several-fold evaluations provide a more objective view of the model's performance, particularly about its ability to generalize to new data.

D. THE PRINCIPAL COMPONENT ANALYSIS (PCA) METHOD

PCA is used to transform a dataset with many dimensions (features) into a dataset with fewer dimensions while preserving key information from the original dataset [26]. PCA lowers the dimensionality of the data by eliminating features that are not as important or contribute little to the variability of the data. PCA can be used to resolve multicollinearity in datasets where features exhibit strong correlations with one another. The formula for the PCA method is as follows [27]:

$$x_{std} = \frac{x-\mu}{\sigma} \tag{1}$$

$$\sum = \frac{1}{m}\, x_{std}^{T}\, x_{std} \tag{2}$$

$$\sum v = \lambda v \tag{3}$$

$$x_{reduced} = x_{std}\, v_k \tag{4}$$

x is an arbitrary variable for which the standard deviation is computed. μ the mean (average) of the stochastic variable x. σ the symbol for the x variable's standard deviation. Data standardization is accomplished by taking the average of each feature and dividing the result by the standard deviation, as shown in Eq. (1). m this variable represents the amount of data or the number of samples used in the computation. x_{std} is the outcome of the standard deviation computation. Here,x_{std}^{T} a row vector is transformed into a column vector through the transposition operation, represented by the T symbol. $x_{std}^{T}\, x_{std}$ it displays the outcome of multiplying the matrix x_{std}^{T} by x_{std}, which, depending on the matrix's dimensions, can result in either a matrix or a scalar. Covariance matrix Σ is computed from standardized data using Eq. (2). Using Eq. (3), determine the eigenvalues λ and eigenvectors v_k from the covariance matrix. The principal components of the data are the eigenvectors that match the largest eigenvalues. Eq. (4), shows how the chosen principal components were used to convert the original data x_{std} into a lower dimensional space.

C. K-FOLD CROSS VALIDATION

A dataset can be divided into multiple parts (folds), or K-folds, using K-fold cross-validation, where K is the desired number of parts [23]. There were K iterations in the training and evaluation process. The validation set is the set from the first fold, and the training set is the other fold in each iteration. The average of each iteration's model performance is used to determine the model's performance. Because the model is evaluated using multiple combinations of training

and validation data, K-fold helps to reduce the variability of evaluation results [24, 25]. Several-fold evaluations provide a more objective view of the model's performance, particularly about its ability to generalize to new data.

D. THE PRINCIPAL COMPONENT ANALYSIS (PCA) METHOD

PCA is used to transform a dataset with many dimensions (features) into a dataset with fewer dimensions while preserving key information from the original dataset [26]. PCA lowers the dimensionality of the data by eliminating features that are not as important or contribute little to the variability of the data. PCA can be used to resolve multicollinearity in datasets where features exhibit strong correlations with one another. The formula for the PCA method is as follows [27]:

$$x_{std} = \frac{x - \mu}{\sigma} \quad (1)$$

$$\sum = \frac{1}{m}\, x_{std}^{T}\, x_{std} \quad (2)$$

$$\sum v = \lambda v \quad (3)$$

$$x_{reduced} = x_{std}\, v_k \quad (4)$$

x is an arbitrary variable for which the standard deviation is computed. μ the mean (average) of the stochastic variable x. σ the symbol for the x variable's standard deviation. Data standardization is accomplished by taking the average of each feature and dividing the result by the standard deviation, as shown in Eq. (1). m this variable represents the amount of data or the number of samples used in the computation. x_{std} is the outcome of the standard deviation computation. Here,x_{std}^{T} a row vector is transformed into a column vector through the transposition operation, represented by the T symbol. $x_{std}^{T}\, x_{std}$ it displays the outcome of multiplying the matrix x_{std}^{T} by x_{std}, which, depending on the matrix's dimensions, can result in either a matrix or a scalar. Covariance matrix Σ is computed from standardized data using Eq. (2). Using Eq. (3), determine the eigenvalues λ and eigenvectors v_k from the covariance matrix. The principal components of the data are the eigenvectors that match the largest eigenvalues. Eq. (4), shows how the chosen principal components were used to convert the original data x_{std} into a lower dimensional space.

E. THE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE) METHOD

The Synthetic Minority Over-sampling Technique (SMOTE) approach is employed in this study to address the issue of class imbalance in the dataset [28]. SMOTE helps to address the problem of class imbalance by increasing the representation of the minority class and preventing the model from favoring the majority class in its predictions [29]. SMOTE reduces the risk of overfitting in the model by producing more variable synthetic data [30]. The SMOTE method formula is as follows [31]:

$$diff_{vector} = x_{iNN} - x_i \quad (5)$$

$$x_{new} = x_i + r * diff_{vector} \quad (6)$$

In Eq. (5), x_{iNN} is the closest neighbor, and x_i is a minority sample. Eq. (6), represents the freshly generated synthetic data x_{new} which was appended to the original dataset to enhance the representation of the minority class.

F. THE XGBOOST ALGORITHM

An ensemble boosting algorithm called extreme gradient boosting (XGBoost) builds a stronger model by combining multiple weak models, or weak learners. To enhance the model, XGBoost makes use of the loss function's gradient as a reference. By highlighting incorrect samples in subsequent iterations, this approach focuses on fixing prediction errors from the previous model [32].

G. THE RANDOM FOREST ALGORITHM

The Random Forest (RF) ensemble learning technique merges several decision tree models into a single, more powerful model. Because Random Forests employ randomization during their creation, they are generally less prone to overfitting than single-decision trees. The Random Forest algorithm uses two levels of randomization [20].

H. THE GRADIENT BOOSTING MACHINE (GBM) ALGORITHM

The Gradient Boosting technique is implemented by the Gradient Boosting Machine (GBM), which uses the gradient of the loss function to continuously correct previous model prediction errors. A new model is added to the ensemble with each GBM iteration, and this new model is adjusted based on the residual (residual error) of the preceding model. The GBM regularization features that help prevent overfitting include using a lower learning rate and adding regularization to the loss function. Large and complex data sets are well handled by GBM. Because the GBM learning process is iterative, it tends to be slower than some other ensemble methods [33].

I. THE ADABOOST ALGORITHM

Adaptive Boosting (AdaBoost) uses an ensemble in which a weak model is the core component. The primary step in AdaBoost is giving every data sample a unique weight. The sample weights are rearranged in such a way that samples with incorrect classifications are assigned a higher weight and samples with correct classifications are assigned a lower weight. The final result is determined by taking the weighted average of the prediction results from each weak model and calculating the majority of votes (classification) or the average prediction (regression) of all the models. AdaBoost's sensitivity to noise or anomalies in the data may cause learning to be disrupted and unstable models to be produced [33].

J. THE CATBOOST ALGORITHM

CatBoost, or categorical boosting, focuses mostly on data that are classified or have categorical features. This algorithm is made to work well with data that has categorical features, avoiding the need for extra preprocessing like one-hot encoding. Additionally, this algorithm automatically accounts for the interactions between features, producing a more accurate model [33]. The Gradient Boosting approach, on which CatBoost is based, creates a more powerful model by aggregating a large number of weak models, or weak learners.

K. THE HYPERPARAMETER OPTIMIZATION (HPO)

Hyperparameter optimization (HPO) is the process of identifying the ideal values for a model's hyperparameters. Optimal hyperparameter values have a substantial impact on the model's accuracy and performance, which makes hyperparameter optimization crucial. To optimize the model, different combinations of hyperparameter values are tried, and cross-validation techniques are used to assess the model's performance [24]. The two hyperparameter optimization methods used in this study are the Grid Search and Random Search methods.

The Grid Search method's objective is to test every possible combination of hyperparameter values from a predetermined parameter space. The model will be trained using the training data for each combination of hyperparameters that are tested, and cross-validation techniques will be used to assess the model's performance [34]. After evaluating every possible combination of hyperparameters, Grid Search will choose the combination that yields the best results based on the evaluation metrics that have been chosen [35].

A Random Search starts with a set of hyperparameter values that are randomly selected from a predefined parameter space. Random Search samples at random without adhering to a specific pattern, in contrast to Grid Search, which methodically tests each combination of hyperparameter values in the grid [36]. The flexibility of Random Search is greater than that of Grid Search because it is not sensitive to the number of values that have been determined for each hyperparameter.

L. EVALUATION OF MODEL PERFORMANCE

A model performance evaluation was carried out to gain a better understanding of the model's ability to handle classifications related to the status of patients' health conditions [37]. This is important to ensure that the built model can generalize well on test data. Several metrics are used in the model's evaluation [38]. Equations (7), (8), (9), and (10) [39].

$$Accuracy = \frac{Number\ of\ Correct\ Classifications}{Total\ Number\ of\ Classifications} \tag{7}$$

$$Recall = \frac{\sum True\ positives\ (TP)}{\sum True\ positives(TP) + \sum False\ negatives(FN)} \tag{8}$$

$$Precission = \frac{\sum True\ positives(TP)}{\sum False\ positives(FP) + \sum True\ positives(TP)} \tag{9}$$

$$F1\ Score = 2\ x\ \frac{\sum Recall\ x\ \sum Precission}{\sum Recall + \sum Precission} \tag{10}$$

The model's ability to consistently produce accurate classifications is measured by equation accuracy (7). Recall equation (8) quantifies the number of positive classes that the model was able to correctly identify. The precision of equation (9) is the number of accurate positive predictions. The harmonic average of recall and precision is represented by the F1 Score equation (10).

III. RESULTS

The model in this study is trained and tested using the Python programming language. To determine the degree of dependence or relationship between two or more features in a dataset, feature correlation is used. The correlation between data features is displayed in FIGURE 3. Three phases comprised the research experiments: the first involved training and testing the model without the use of PCA, or Smote; the second stage involved using PCA, and SMOTE techniques. In phase three, Grid Search and Random Search methods are used to optimize the model. The following section explains the study findings

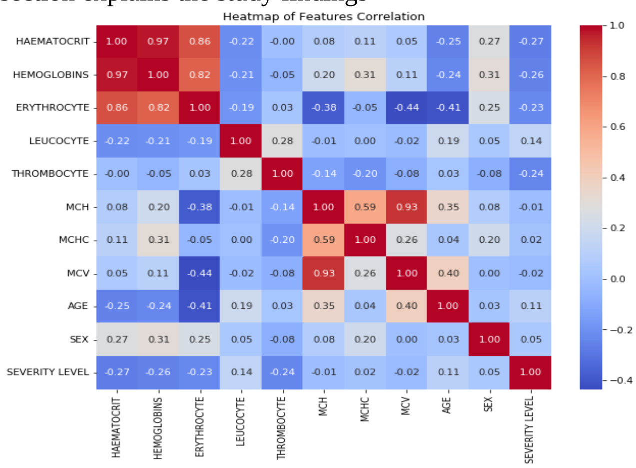


FIGURE 3. The correlation between data feature

A. BASELINE MODEL EXPERIMENTS FOR ENSEMBLE MACHINE LEARNING (EML)

The five Ensemble machine learning (EML) models used are XGBoost, Random Forest (RF), Gradient Boosting Machine (GBM), AdaBoost, and CatBoost. The model was trained and tested using a cross-validation technique with a k-fold of 10. TABLE 2 displays the accuracy, recall, precision, and F1-Score values for the model.

TABLE 2
Performance evaluation results of the EML model without optimization

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.69	0.69	0.69	0.68
RF	0.70	0.69	0.70	0.69
GBM	0.72	0.71	0.72	0.71

AdaBoost	0.71	0.70	0.71	0.70
CatBoost	0.71	0.71	0.71	0.71

When measured against other models, the accuracy of the Gradient Boosting Machine (GBM) and CatBoost models is superior. The accuracy obtained by the CatBoost and Gradient Boosting Machine (GBM) models was 71% and 72%. The Gradient Boosting Machine (GBM) outperforms CatBoost when evaluated in terms of precision, recall, and F1-Score.

B. ENSEMBLE MACHINE LEARNING (EML) MODEL EXPERIMENTS WITH PCA ANDA SMOTE

The objective of this experiment's second phase was to address data imbalances and reduce the dimensions of the data. Data imbalances are handled with the SMOTE method, while data dimensions are reduced using the PCA method. Next, the model is trained and tested. TABLE 3 displays the obtained accuracy, precision, recall, and F1-Score results.

TABLE 3
Performance evaluation results of the EML model following the application of SMOTE and PCA.

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.71	0.70	0.71	0.70
RF	0.72	0.71	0.72	0.71
GBM	0.73	0.73	0.73	0.72
AdaBoost	0.72	0.71	0.72	0.71
CatBoost	0.73	0.72	0.73	0.72

The Ensemble machine learning (EML) model's accuracy increased as a result of using the PCA and SMOTE methods. The accuracy of the CatBoost and Gradient Boosting Machine (GBM) models was 73%. The Gradient Boosting Machine (GBM) model outperforms the CatBoost model when compared based on precision, recall, and F1-Score results. The classification outcomes of the Gradient Boosting Machine (GBM) model are shown in FIGURE 4. Results of the CatBoost model's classification are shown in FIGURE 5.

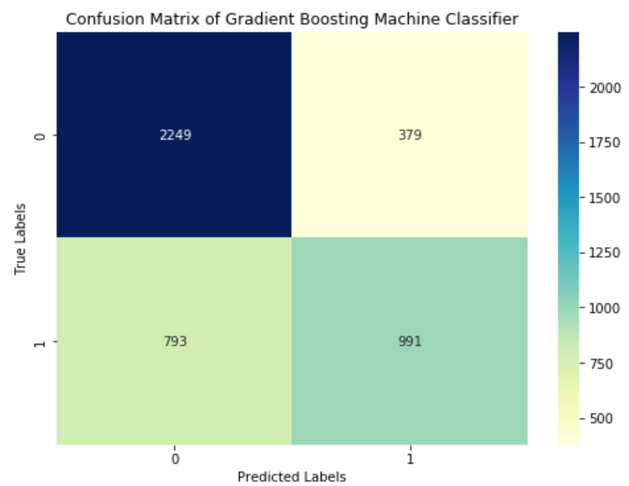


FIGURE 4. Results of classification using the Gradient Boosting Machine (GBM) model.

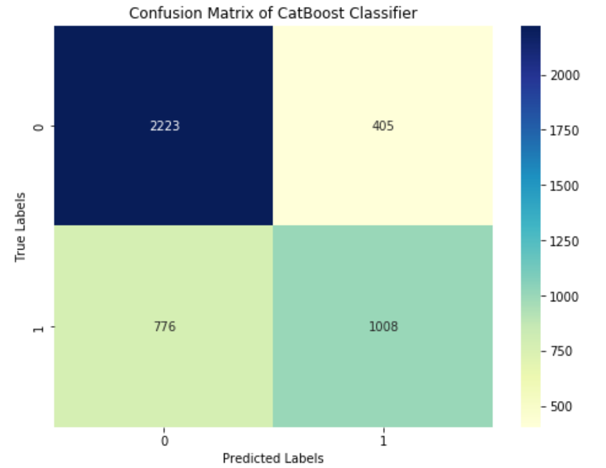


FIGURE 5. Classification outcomes of the CatBoost model

C. ENSEMBLE MACHINE LEARNING (EML) MODEL EXPERIMENTS WITH OPTIMIZATION

At this point, the experiment proceeded by employing Grid Search and Random Search methods to optimize the model. The model employs 0.1, 0.01, and 0.001 as the values for the learning rate parameter. The performance outcomes of the EML model, which was optimized through the application of Grid Search and Random Search techniques, are shown in TABLE 4.

TABLE 4
Results of the EML model's performance evaluation following optimization

Model	Accuracy	Precision	Recall	F1-Score
XGBoost + Grid Search	0.73	0.73	0.73	0.73
XGBoost + Random Search	0.72	0.72	0.72	0.72
RF + Grid Search	0.70	0.70	0.70	0.70
RF + Random Search	0.72	0.72	0.72	0.72
GBM + Grid Search	0.73	0.74	0.73	0.73
GBM + Random Search	0.74	0.74	0.74	0.74
AdaBoost + Grid Search	0.69	0.69	0.69	0.69
AdaBoost + Random Search	0.69	0.69	0.69	0.69
CatBoost + Grid Search	0.72	0.72	0.72	0.72
CatBoost + Random Search	0.72	0.72	0.72	0.72

The EML model optimization experiment that employed Grid Search and Random Search yielded two optimal models: Gradient Boosting Machine (GBM) and XGBoost. Overall evaluation results between the Gradient Boosting Machine (GBM) and XGBoost indicate that the GBM model is the better model. The Gradient Boosting Machine (GBM) model experienced a 74% increase in accuracy following Random Search technique optimization. The classification outcomes of the Gradient Boosting Machine (GBM) model following Random Search optimization are shown in FIGURE 6.

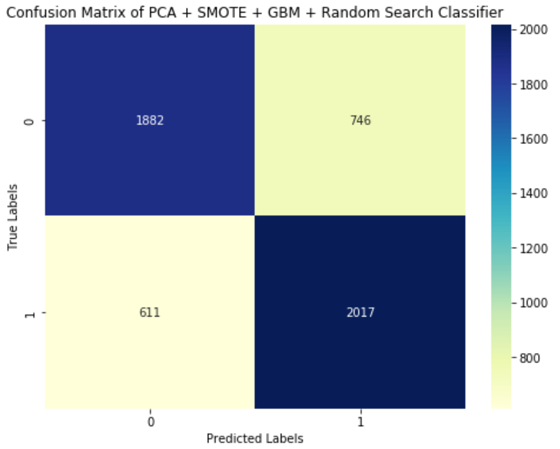


FIGURE 6. Results of classifying the Gradient Boosting Machine (GBM) model using optimization for Random Search.

IV. DISCUSSION

Through a comprehensive comparative analysis of multiple ensemble machine learning models on EHR datasets, this study focuses on minority class performance, which is important in this field, as well as model hyperparameter tuning to classify patient severity condition status. Our analysis highlights the impact of dimensionality, data imbalance, and hyperparameter tuning in classification modeling and demonstrates the varying degrees of efficacy among these models in the classification of outcomes for minority classes.

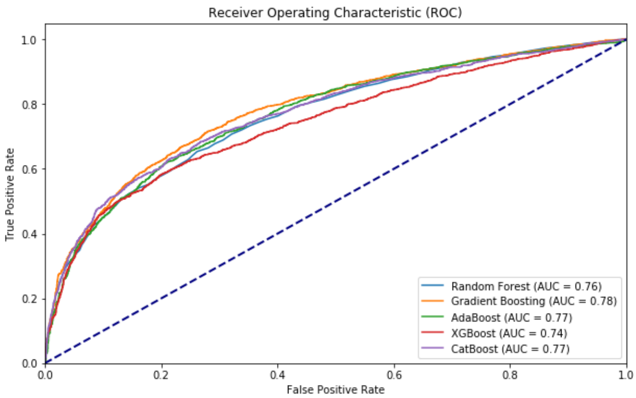


FIGURE 7. The ROC-AUC value for the EML model

After handling unbalanced data and dimension reduction, the accuracy performance results of the EML model show a significant change. We highlight the results of GMB and CatBoost models that have superior performance with an obtained accuracy of 73%, based on experiments with dimensionality reduction and handling of imbalanced data. Additionally, we experimented by improving the EML model. We use optimization techniques such as Grid Search and Random Search to enhance the EML model. According to the results of the EML model optimization, the GBM + Random Search model performed the best, achieving an accuracy of 74%; XGBoost + Grid Search came in second place, achieving an accuracy of 73%. FIGURE 7 shows the results of the Receiver Operating Characteristic - Area under the Curve (ROC-AUC) of the used EML model. The GBM model performs the best in distinguishing between positive and negative classes with the highest AUC value of 0.78. This indicates that the GBM model has better classification capabilities compared to other models. When juxtaposed with the study conducted by [30, 33], the GBM model in this research exhibited superior enhancements in performance. On the other hand, both AdaBoost and CatBoost models have the same AUC value of 0.77, showing comparable performance in class discrimination. Meanwhile, the RF and XGBoost models, although still acceptable in performance, have lower AUC values of 0.76 for the RF model and 0.74 for the XGBoost model. Due in part to the large sample size in this study, the machine learning ensemble model continues to perform slowly. A deeper investigation will be conducted using the deep learning (DL) approach to conduct additional research. However, the study's findings offer a precise severity classification that medical professionals and teams can use to make quicker and more informed clinical decisions based on a patient's condition.

V. CONCLUSION

The study conducted a thorough comparative analysis of various ensemble machine learning (EML) models using Electronic Health Record (EHR) datasets. The focus was on evaluating the performance of minority classes, crucial in this domain, and optimizing model hyperparameters for classifying patient severity conditions. The analysis emphasized the impact of dimensionality, data imbalance, and hyperparameter tuning on classification modeling, showcasing differing levels of efficacy among these models in predicting outcomes for minority classes. After addressing data imbalance and reducing dimensionality, the accuracy of the EML models showed significant improvement. Notably, the GBM and CatBoost models exhibited superior performance with an accuracy of 73%, achieved through experiments involving dimensionality reduction and handling of imbalanced data. Furthermore, optimization techniques such as Grid Search and Random Search were employed to enhance the EML models. The results of model optimization revealed that the GBM + Random Search model performed the best, achieving an accuracy of 74%, followed by the XGBoost + Grid Search model with an accuracy of 73%. The GBM model also excelled in distinguishing between positive and negative classes, boasting the highest AUC value of 0.78, indicative of its superior classification capabilities compared to other models. Conversely, both AdaBoost and CatBoost models showed comparable performance with an AUC value of 0.77, while the RF and XGBoost models exhibited slightly lower AUC values of 0.76 and 0.74, respectively, although still maintaining acceptable performance levels.

Disclosure of interest:

The authors state that there are no conflicting interests to reveal.

Consent to Publish:

All participants whose identifying information appears in this article gave their consent for it to be published.

Data Availability:

Datasets are accessible to all. Accessible via the following link: https://data.mendeley.com/datasets/7kv3rctx7m/1

REFERENCES

[1] G. Lombardo, C. Couvert, M. Kose, A. Begum and C. Spiertz, "Electronic health records (EHRs) in clinical research and platform trials: Application of the innovative EHR-based methods developed by EU-PEARL," *Journal of Biomedical Informatics,* vol. 148, no. 104553, pp. 1-8, 2023.

[2] Y. Ramakrishnaiah, N. Macesic, G. I. Webb, A. Y. Peleg and S. Tyagi, "EHR-QC: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes," *Journal of Biomedical Informatics,* vol. 147, no. 104509, pp. 1-10, 2023.

[3] H. Niu, O. A. Omitaomu, M. A. Langston, M. Olama, O. Ozmen, H. B. Klasky, A. Laurio, M. Ward and J. Nebeker, "EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records," *Journal of Biomedical Informatics,* vol. 150, no. 104605, pp. 1-11, 2024.

[4] D. Kotecha, F. W. Asselbergs, S. Achenbach, S. D. Anker, D. Atar, C. Baigent, A. Banerjee, B. Beger, G. Brobert, B. Casadei, C. Ceccarelli, M. R. Cowie, F. Crea, M. Cronin, S. Denaxas and A. Derix, "CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research," *The Lancet Digital Health*, vol. 4, no. 10, p. e698, 2022.

[5] D. A. Martinez, S. R. Levin, E. Y. Klein, C. R. Parikh, S. Menez, R. A. Taylor and J. S. Hinson, "Early Prediction of Acute Kidney Injury in the Emergency Department With Machine-Learning Methods Applied to Electronic Health Record Data," *Annals of Emergency Medicine,* vol. 76, no. 4, pp. 501-514, 2020.

[6] A. M. Cogan, T. M. Haltom, S. L. Shimada, B. P. McGinn and G. M. Fix, "Understanding patients' experiences during transitions from one electronic health record to another: A scoping review," *PEC Innovation,* vol. 4, no. 100258, pp. 1-5, 2024.

[7] B. J. H. Yarborough and S. P. Stumbo, "Patient perspectives on acceptability of, and implementation preferences for, use of electronic health records and machine learning to identify suicide risk," *General Hospital Psychiatry,* vol. 70, pp. 31-37, 2021.

[8] E. L. Eisenstein, M. N. Zozus, M. Y. Garza, H. J. Lanham, B. Adagarla, A. Walden, D. K. Benjamin, K. O. Zimmerman and K. R. Kumar, "Assessing clinical site readiness for electronic health record (EHR)-to-electronic data capture (EDC) automated data collection," *Contemporary Clinical Trials,* vol. 128, p. 107144, 2023.

[9] W. Hurst, B. Tekinerdogan, T. Alskaif, A. Boddy and N. Shone, "Securing electronic health records against insider-threats: A supervised machine learning approach," *Smart Health,* vol. 26, no. 100354, pp. 1-14, 2022.

[10] K. Shoenbill, Y. Song, M. Craven, H. Johnson, M. Smith and E. A. Mendonca, "Identifying patterns and predictors of lifestyle modification in electronic health record documentation using statistical and machine learning methods," *Preventive Medicine,* vol. 136, no. 106061, pp. 1-11, 2020.

[11] A. Banerjee, A. Dashtban, S. Chen, L. Pasea, J. H. Thygesen, G. Fatemifar, B. Tyl, T. Dyszynski, F. W. Asselbergs, L. H. Lund, T. Lumbers, S. Denaxas and H. Hemingway, "Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study," *European Union Innovative Medicines,* vol. 5, no. 378, pp. 1-10, 2023.

[12] C. Ye, J. Li, S. Hao, M. Liu, H. Jin, L. Zheng, M. Xia, B. Jin, C. Zhu, S. T. Alfreds and X. B. Ling, "Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm," *International Journal of Medical Informatics,* vol. 137, no. 104105, pp. 1-7, 2020.

[13] A. Srinivas and J. P. Mosiganti, "A brain stroke detection model using soft voting based ensemble machine learning classifier," *Measurement: Sensors,* vol. 29, no. 100871, pp. 1-7, 2023.

[14] V. Jaiswal, P. Saurabh, U. K. Lilhore, M. Pathak, S. Simaiya and S. Dalal, "A breast cancer risk predication and classification model with ensemble learning and big data fusion," *Decision Analytics Journal,* vol. 8, no. 100298, pp. 1-13, 2023.

[15] C. Ye, J. Li, S. Hao, M. Liu, H. Jin, L. Zheng, M. Xia, B. Jin and X. B. Ling, "Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm," *International Journal of Medical Informatics,* vol. 137, no. 104105, pp. 1-7, 2020.

[16] A. J. Weiss, A. S. Yadaw, D. L. Meretzky, M. A. Levin, D. H. Adams, K. McCardle, G. Pandey and R. Iyengar, "Machine learning using institution-specific multi-modal electronic health records improves mortality risk prediction for cardiac surgery patients," *Risk Scores: Evolving Technology,* vol. 14, pp. 214-251, 2023.

[17] A. M. McKnite, K. M. Job, R. Nelson, C. M. Sherwin, K. M. Watt and S. C. Brewer, "Medication based machine learning to identify subpopulations of pediatric hemodialysis patients in an electronic health record database," *Informatics in Medicine Unlocked,* vol. 34, no. 101104, pp. 1-8, 2022.

[18] Y.-x. Li, X.-p. Shen, C. Yang, Z.-z. Cao, R. Du, M.-d. Yu, J.-p. Wang and M. Wang, "Novel electronic health records applied for prediction of pre-eclampsia: Machine-learning algorithms," *Pregnancy Hypertension: An International Journal of Women's Cardiovascular Health,* vol. 26, no. 102-109, pp. 102-109, 2021.

[19] J. H. Rubenstein, S. Fontaine, P. W. MacDonald, J. A. Burns, R. R. Evans, M. E. Arasim, J. W. Chang, E. M. Firsht and A. K. Waljee, "Predicting Incident Adenocarcinoma of the Esophagus or Gastric Cardia Using Machine Learning of Electronic Health Records," *Gastroenterology,* vol. 165, p. 1420–1429, 2023.

[20] M. Herrero-Zazo, T. Fitzgerald, V. Taylor, H. Street, A. N. Chaudhry, J. R. Bradley, E. Birney and V. L. Keevil, "Using machine learning to model older adult inpatient trajectories from electronic health records data," *iScience,* vol. 26, no. 105876, pp. 1-15, 022.

[21] M. Sadikin, "Mendeley Data," 05 05 2020. [Online]. Available: https://data.mendeley.com/datasets/7kv3rctx7m/1. [Accessed Monday 05 2024].

[22] A. M. Cogan, T. M. Haltom, S. L. Shimada, J. A. Davila, B. P. McGinn and G. M. Fix, "Understanding patients' experiences during transitions from one electronic health record to another: A scoping review," *PEC Innovation,* vol. 4, no. 100258, pp. 1-5, 2024.

[23] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express,* vol. 8, p. 183–188, 2022.

[24] S. Kucheryavskiy, O. Rodionova and A. Pomerantsev, "Procrustes cross-validation of multivariate regression models," *Analytica Chimica Acta,* vol. 1255, no. 341096, pp. 1-10, 2023.

[25] M. T. R, V. K. V, D. K. V, O. Geman, M. Margala and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics,* vol. 4, no. 100247, pp. 1-10, 2023.

[26] T. Wang, Y. Xie, Y.-S. Jeong and M. K. Jeong, "Dynamic sparse PCA: a dimensional reduction method for sensor data in virtual metrology," *Expert Systems With Applications,* vol. 251, no. 123995, pp. 1-12, 2024.

[27] L. C. D. Nkengfack, D. Tchiotsop, R. Atangana, B. S. Tchinda, V. Louis-Door and D. Wolf, "A comparison study of polynomial-based PCA, KPCA, LDA and GDA feature extraction methods for epileptic and eye states EEG signals detection using kernel

machines," *Informatics in Medicine Unlocked,* vol. 26, no. 100721, pp. 1-16, 2021.

[28] A. Imakura, M. Kihira, Y. Okada and T. Sakurai, "Another use of SMOTE for interpretable data collaboration analysis," *Expert Systems With Applications,* vol. 228, no. 120385, pp. 1-10, 2023.

[29] A. Arafa, N. El-Fishawy, M. Badawy and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences,* vol. 34, pp. 5059-5074, 2022.

[30] I. Czarnowski, "Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams," *Journal of Computational Science,* vol. 61, no. 101614, pp. 1-10, 2022.

[31] M. Umer, S. Sadiq, M. M. S. Missen, Z. Hameed, Z. Aslam, M. A. Siddique and M. NAPPI, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognition Letters,* vol. 150, p. 250–257, 2021.

[32] A. J. Weiss, A. S. Yadaw, D. L. Meretzky, M. A. Levin, D. H. Adams, K. McCardle, G. Pandey and R. Iyengar, "Machine learning using institution-specific multi-modal electronic health records improves mortality risk prediction for cardiac surgery patients," *Evolving Technology,* vol. 14, pp. 214-251, 2023.

[33] A. Khoder and F. Dornaika, "Ensemble learning via feature selection and multiple transformed subsets: Application to image classification," *Applied Soft Computing,* vol. 113, no. 108006, pp. 1-14, 2021.

[34] M. Ogunsanya, J. Isichei and S. Desaia, "Grid Search Hyperparameter Tuning in Additive Manufacturing Processes," *Manufacturing Letters,* vol. 35, pp. 031-1042, 2023.

[35] Y. Zhao, W. Zhang and X. Liu, "Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting," *Applied Soft Computing,* vol. 154, no. 111362, pp. 1-20, 2024.

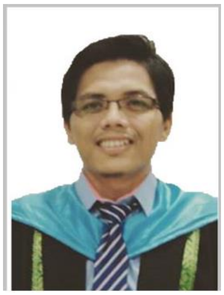
[36] B. Xi, G. Xiong, K. A. Kozin, C. He, T. S. Tamir, Y. Song, X. Liu and Z. Shen, "Modeling and random search optimization for the polysilicon CVD reactor," *Results in Control and Optimization,* vol. 13, no. 100320, pp. 1-10, 2023.

[37] J. M. Beinecke, a. Anders, T. Schurrat and D. Heider, "Evaluation of machine learning strategies for imaging confirmed prostate cancer recurrence prediction on electronic health records," *Computers in Biology and Medicine,* vol. 143, no. 105263, pp. 1-9, 2022.

[38] A. Theissler, M. Thomas, M. Burch and F. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowledge-Based Systems,* vol. 247, no. 108651, pp. 1-16, 2022.

[39] D. Valero-Carreras, J. Alcaraz and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Computers and Operations Research,* vol. 152, no. 106131, pp. 1-12, 2023.

AUTHOR BIOGRAPHY



Edi Ismanto is currently pursuing a Master's Degree in Electrical Engineering at Ahmad Dahlan University. He is also employed by the Informatics Study Programme at the Muhammadiyah University of Riau. Edi has a profound interest in research involving Deep Learning (DL), Machine Learning (ML), and Artificial Intelligence (AI). With a strong background in electrical engineering and informatics, he focuses on developing and applying advanced technologies to solve various real-world challenges. With a high dedication to advancing knowledge and technology, Edi aims to make significant contributions to both the academic and industrial communities. He is actively involved in scientific publications and research collaborations, hoping to drive innovation and progress in the fields of AI and related technologies.

Abdul Fadlil received the B.Eng. in Physic – Electrical & Insrumentation and M.Eng. in Electrical Engineering from Universitas Gadjah Mada in 1992 and 2000, respectively. He also received his Ph.D. in Electrical Engineering from Universiti Teknologi Malaysia in 2006.

He is currently as Professor with the Electrical Engineering Department, Universitas Ahmad Dahlan (UAD), Yogyakarta, Indonesia. His current research interests include pattern recognition, image processing, and artificial intelligence. He is actively involved in scientific publications and research collaborations, hoping to drive innovation and progress in the fields of AI and related technologies. He can be contacted at email: fadlil@mti.uad.ac.id

Anton Yudhana obtained an Electrical Engineering B.Eng degree from the Surabaya Institute of Technology. Earned a Master of Engineering in Electrical Engineering from Gadjah Mada University. In Electrical Engineering from Universiti Teknologi Malaysia.

He is currently as Professor with the Electrical Engineering Department, Universitas Ahmad Dahlan (UAD), Yogyakarta, Indonesia. His current research interests include intelligent networks, intelligent control, and artificial intelligence. He is actively involved in scientific publications and research collaborations, hoping to drive innovation and progress in the fields of AI and related technologies. He can be contacted at email: eyudhana@ee.uad.ac.id

Kodai Kitagawa received a Ph.D. in Engineering from Kyushu Institute of Technology, Japan, in 2022. Since 2022, he has been an assistant professor at the Mechanical and Medical Engineering Course, Department of Industrial Systems Engineering, National Institute of Technology, Hachinohe College. His current research interests include wearable sensors, human movement, gait measurement, physical therapy, and ergonomics.

He is a member of the Institute of Electrical and Electronics Engineers (IEEE), the Society of Instrument and Control Engineers (SICE), and the Society of Physical Therapy Science (SPTS