Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 855-870
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

Gastroesophageal Reflux Disease Early Detection using XGBoost Method Classifier

Haura Adzkia Delfina[1], Untari Novia Wisesty *[2], and Isman Kurniawan [3]

[1]Data Science, School of Computing, Telkom University, Indonesia
[2,3]Informatics, School of Computing, Telkom University, Indonesia

Email: [2]untarinw@telkomuniversity.ac.id

Received : Nov 22, 2024; Revised : Dec 22, 2024; Accepted : Dec 26, 2024; Published : Apr 26, 2025

Abstract

Gastroesophageal reflux disease (GERD) is a clinical condition that occurs when the gastric content within the stomach rises into the esophagus. If left untreated, GERD can result in complications such as esophageal inflammation, ulcers, and even cancer. In this study, the early detection of GERD is performed using the GERD dataset obtained from the Harvard Dataverse online repository and processed with the XGBoost machine learning model. The SMOTE technique was implemented as a solution to address the data imbalance present in the dataset. In addition, this study applied Principal Component Analysis (PCA) and Pearson Correlation to select the most relevant attributes, with the aim of improving computational efficiency. The results demonstrated that feature selection through Pearson correlation and feature extraction using principal component analysis (PCA) yielded the optimal model performance when utilizing 16 attributes and 16 principal components, respectively. The XGBoost model with PCA achieves a macro average F1-score of 0.9615, while the XGBoost model with Pearson Correlation attains a value of 0.9809. Subsequently, the XGBoost model based on the original dataset yielded a macro F1-score value of 0.9568. The findings of this research indicate that the XGBoost model with the Pearson Correlation-based feature selection method has a better f1-score value than the feature extraction method with PCA or based on the original dataset with a difference in value of 0.0194 and 0.0241 respectively in enhancing the performance of the XGBoost model for early detection of GERD in this study.

Keywords : *GERD Detection, Machine Learning, PCA, Pearson Correlation, SMOTE, XGBoost.*

1. INTRODUCTION

Gastroesophageal reflux disease, or GERD, refers to a medical condition that occurs when gastrointestinal contents of the stomach ascend into the esophagus [1]-[3]. This can cause a range of symptoms including heartburn and regurgitation [4]-[6]. This condition is highly prevalent worldwide [1], [4], [7], with a prevalence of 13.98% [8]. Given this high prevalence, it is imperative to prioritize early detection to minimize the number of undetected cases and prevent more serious complications, such as esophagitis, esophageal strictures, barrett's esophagus, and esophageal adenocarcinoma (EAC) [1]-[4], [7].

The current conventional or medical approach to disease detection has inherent limitations. These limitations include an insufficient number of health personnel and diagnostic facilities in developing countries [9]-[10]. In addition, there are difficulties in detecting certain diseases, such as Fatty Liver Disease (FLD), due to costly process and highly invasive diagnosis method [11]. A machine learning approach enables disease detection through the analysis of medical information, offering a cost-effective and time-saving alternative [9], [12]-[13]. Furthermore, machine learning has proven effective in analyzing complex medical data and identify patterns or characteristics that are challenging to discern [12], [14]-[15]. Numerous studies have explored the application of machine learning for early disease

detection, including research conducted by Mazhar et al. [16] on diabetes and Kabiraj et al. [17] on breast cancer. The success of these studies highlights the potential for developing an early detection method for GERD using machine learning.

A review of the literature reveals a lack of discussion regarding the early detection of GERD using data that includes information about the symptoms experienced by patients. Some studies employed the results of supplementary examinations, such as magnetic resonance imaging (MRI) and endoscopy, for diseases with similar characteristics or utilized data from patient medical records for other diseases. Srividya and Sasi [18] demonstrated the efficacy of machine learning algorithms in GERD early detection using MRI image data. Additionally, research has been conducted on esophageal adenocarcinoma or gastric cardia [19] using electronic medical record data. However, no literature has been identified that employs a similar approach with a specific focus on GERD.

The objective of this study is to develop an early detection method for GERD using data containing information about patient symptoms sourced from research conducted by Wickramasinghe et al. [20]. This approach presents a novel opportunity for early detection of GERD. By employing machine learning on patient symptom data, it is possible to identify individuals at high risk of GERD, thereby facilitating early prevention and treatment [12]. Furthermore, the data can be utilized to examine the relationship between other factors, such as stress, alcohol consumption, sleep disorders, and their correlation with GERD.

In this study, the XGBoost method is employed with the application of feature selection and feature extraction techniques to detect GERD using patient medical record data. This approach is selected due to the fact that it possesses several characteristics that align with the characteristics of the dataset, including symptom relatedness and demonstrated efficacy in several literature sources. Consequently, XGBoost is considered an optimal choice for predicting GERD risk. XGBoost is a machine learning algorithm widely acknowledged for its notable accuracy and ability to process complex data [21]-[22]. This aligns with the findings of Ali et al. [23], who highlighted the advantages of XGBoost in terms of speed, accuracy, interpretability, and its capacity to model intricate systems.

Furthermore, XGBoost has been demonstrated to be an effective tool for detecting other diseases, including fatty liver disease, chronic kidney disease, and coronary heart disease. In a previous study, Pei et al. [11] employed the XGBoost method to achieve an accuracy of 94.15% in the prediction of fatty liver disease. Subsequently, in other research conducted by Raihan et al. [24], a chronic kidney disease detection model was developed, achieving an accuracy of 99.16%. Additionally, a research by Zhang et al. [25] yielded an accuracy of 94.7% in predicting coronary heart disease utilizing XGBoost and feature engineering.

This research aims to make a significant contribution to the prevention and management of GERD. The implementation of an accurate early detection method is expected to reduce the number of GERD cases that are not detected in a timely manner, thereby facilitating more effective and targeted medical interventions. Furthermore, early detection facilitates the implementation of preventive measures, such as lifestyle modifications or prophylactic medication, which can mitigate the risk of severe complications in the future. It is expected that this study will facilitate a more comprehensive understanding of the potential applications of machine learning in GERD early detection based on patient symptoms, thereby establishing a foundation for the development of more effective methods in the future.

2. METHOD

This research begins with the collection of the necessary data. The data examined in this study pertains to the medical condition known as GERD, encompassing both the symptoms and the patient's overall condition. The second stage is data exploration, during which a variety of analytical techniques

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

are employed, including descriptive statistics and data visualization. The third stage is data pre-processing, which ensures the quality of the data used is optimal. The subsequent phase is data splitting, whereby the data set is partitioned into a training data set and a test data set, comprising 80% and 20% of the complete data set, respectively. In the subsequent stage, the feature selection and feature extraction techniques, utilising Pearson's correlation coefficient and Principal Component Analysis (PCA), are performed for each portion of the data. Subsequently, the training data is subjected to a SMOTE-based approach to address the issue of data imbalance. Following this, the XGBoost classification model is trained on the prepared training data. In the final stage, the XGBoost model is evaluated using the test data, with the results of previous model training and evaluation of the XGBoost model considered. The proposed methodology of this research is illustrated in Figure 1.

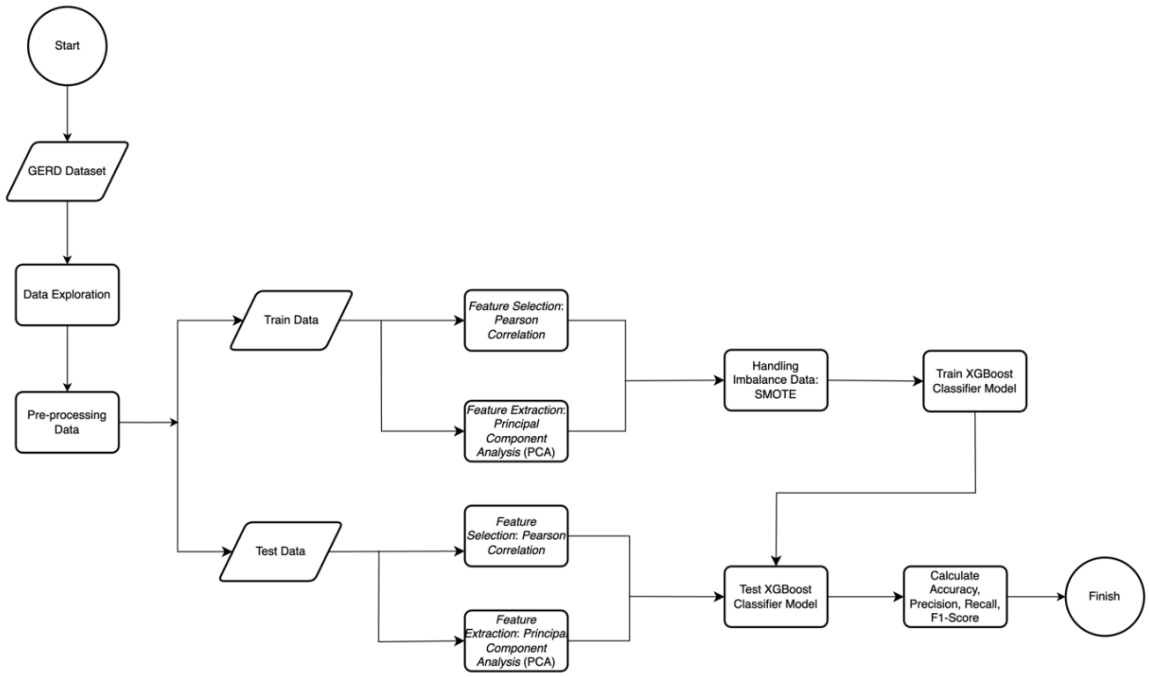

Figure 1. Proposed Methodology.

2.1. Data Collection

The data utilized in the study were obtained from the online data repository service site, Harvard Dataverse, which was released in the article, “The association between symptoms of gastroesophageal reflux disease and perceived stress: A countrywide study of Sri Lanka” was conducted by Wickramasinghe et al. [20]. The dataset consists of 69 columns and 1,200 rows of data. The objective column or feature of this study is “GORDgrade”, which has classes 1 (non-probable GERD) and 2 (probable GERD).

2.2. Data Exploration

In research on disease detection, an exploratory approach is necessary to analyze and implement appropriate models based on the findings in the data [26]. Data exploration provides an overview of the data to facilitate the understanding of trends and patterns [27]. At this stage, several analytical techniques, such as descriptive statistics and data visualization, are performed.

2.3. Data Pre-processing

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

Data pre-processing is the initial stage of the data analysis process, during which raw data is prepared for subsequent analysis or modelling. The pre-processing stage is beneficial for ensuring the quality of the data is optimal [28] and suitable for modelling purposes, as well as facilitating greater comprehension of the data. The process of data cleaning involves the removal of redundant feature columns, thereby reducing the computational load [29]. Furthermore, during the data cleaning stage, imputation is conducted on columns that produce NaN values following data type conversion. Subsequently, data scaling is conducted to ensure that the range of values within each feature is consistent [30], thereby facilitating optimal utilization in the feature extraction stage through Principal Component Analysis (PCA).

2.4. Data Splitting

Data splitting is a technique used for machine learning model validation [31]. At this stage, training data is utilized to train and build machine learning models so that the models can learn and understand patterns in the data. Then, test data is used to test the trained model and evaluate the performance of the model that has been implemented.

2.5. Feature Selection and Feature Extraction

Feature extraction and feature selection are techniques used to reduce data dimensions in research. This stage is necessary because the dataset utilized in the study comprises a large number of columns or attributes. The objective of this data dimension reduction technique is to reduce the computational burden while maintaining information on the initial attributes [32]. The process of feature selection involves the identification of relevant or highly correlated attributes [33], without altering the values of those selected. Feature extraction, on the other hand, entails reducing the number of dimensions in a dataset by transforming the data and extracting crucial features [34]. At this stage, feature selection techniques will be employed using the Pearson Correlation coefficient and the Principal Component Analysis (PCA), with the aim of extracting relevant features from the given dataset.

2.5.1. Pearson Correlation

One of the techniques that can be used for feature selection is Pearson Correlation [35]. Pearson Correlation Coefficient is a technique for calculating coefficients used to explain bivariate relationships [36]. Pearson Correlation Coefficient (r) is employed in the feature selection phase to identify those features that are highly correlated with the target column [37]. Pearson coefficient values range from -1 to 1. If the r value is greater than zero, it indicates a positive correlation between the two variables. Conversely, if r value less than zero, it signifies a negative correlation. An r value of zero indicates that the two variables are not correlated [37]. The formula to calculate the r value is presented in Equation 1, where x and y represent the first and second variables, respectively, and N is the number of variable pairs for which the correlation is calculated [37].

$$r = \frac{N \sum xy - (\sum x)\,(\sum y)}{\sqrt{\{N \sum x^2 - (\sum x)^2\}\,\{N \sum y^2 - (\sum y)^2\}}} \qquad (1)$$

2.5.2. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a technique that can be utilized for the purpose of feature extraction. The PCA reduces the dimensionality of data and creates a new variables set [38], called principal components, which are used to represent the data in a more condensed form. The objective of PCA is to transform data into a smaller number of dimensions by extracting the most important features, thereby reducing the size of the dataset and accelerating the computational process in research. PCA standardizes the value of each attribute to ensure a uniform range of values, then calculates the

covariance matrix to determine the correlation between attributes. Subsequently, the eigenvectors are sorted based on the eigenvalues obtained from the covariance matrix to identify the principal components in the data. The resulting analysis is a set of principal components in the form of new features, which are linear combinations of the initial attributes [37].

2.6. Synthetic Minority Oversampling Technique (SMOTE)

In this research, it is necessary to address the unbalanced data issue, given that the target column in the dataset, 'GORDgrade', exhibits an unbalanced distribution of class values. To address this issue, this research employs Synthetic Minority Oversampling Technique (SMOTE) on the training data. SMOTE is an oversampling-based data handling technique that is employed to address the issue of an unbalanced data distribution between classes [39].

The main concept of SMOTE algorithm is to create new samples that are the result of random oversampling of minority classes based on their k-nearest neighbor and then added to the dataset repeatedly [25]. SMOTE technique begins with the selection of a random sample from the minority class and the determination of its k-nearest neighbors. Subsequently, the distance between the sample and each nearest neighbor is calculated and multiplied by a random number between 0 and 1, thereby generating synthetic data. This process is iterated until the number of minority class samples is equivalent to that of the majority class [37].

2.7. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) represents a machine learning algorithm that leverages ensemble learning techniques. Extreme Gradient Boosting (XGBoost) represents a further development of the gradient boosting method, which previously demonstrated good performance but was limited in its ability to train models representing complex systems within a reasonable time frame. XGBoost method employs the generation of individual trees in parallel, a strategy that reduces training time and enhances model performance [23].

One of the key advantages of XGBoost is the tree pruning feature, which enhances its efficacy relative to previous techniques [23]. Tree pruning is a method used to reduce decision tree size by removing nodes in order to avoid overfitting the training results [23]. Furthermore, there is a parallelization feature that is used to sort each block of data in parallel [23]. Additionally, XGBoost has a cache-aware feature which enables gradient statistics to be stored at each vertex, thus reducing unnecessary processing time [23]. The illustration of XGBoost method can be seen in Figure 2.

In this research, several main parameters are used for binary classification in the XGBoost model, such as objective function, evaluation metric, label encoder usage, and random state. The objective function uses binary logistic because the target column in the study has 2 classes, namely probable gerd and non-probable gerd for GERD disease. Then the evaluation metric uses logloss because it is suitable for binary classification problems. Label Encoder Usage is set to False to avoid automatic label encoding by XGBoost. In addition, Random State is set to 42 to ensure consistent and reproducible training results in each experiment.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 855-870
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

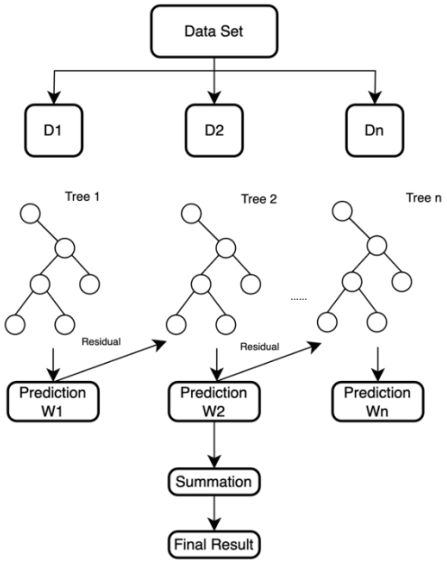

Figure 2. XGBoost Illustration.

2.8. Model Evaluation

In order to assess the classification model performance, the confusion matrix, which provides a matrix for the evaluation of classification models performance, is employed [40]. Confusion matrix shows the results of model predictions of test data in the form of four categories, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The true negative (TN) represents the number of negative samples that have been correctly identified as such by the model [40]. Conversely, the True Positive (TP) indicator denotes the quantity of positive samples that were accurately identified as such [40]. The false negative (FN) category records the number of positive samples that were incorrectly predicted as negative [40]. The false positive (FP) category, on the other hand, represents the number of negative samples that were incorrectly identified as positive by the model [40].

To assess the efficacy of the XGBoost classification model in detecting GERD disease, this study employed a set of performance metrics, namely accuracy, precision, recall, and F1-score.

2.8.1. Accuracy

Accuracy is a frequently used and easily understood index in machine learning. Accuracy is used to measure the ratio between the correct predictions generated by the model, and the overall total predictions. Accuracy is calculated using the formula in Equation 2.

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

2.8.2. Precision

Precision is used to calculate the proportion of true positive samples out of all predicted positive samples. The Precision metric is determined using the formula in Equation 3.

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

2.8.3. Recall

Recall, also known as the true positive rate, is a performance metric used to assess ability of the model to accurately identify positive samples. It is calculated as the percentage of correctly classified positive samples out of the total number of positive samples. Recall is computed by applying the formula in Equation 4.

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

2.8.4. F1-Score

F1-score is a metric that integrates precision and recall into a single value, providing a measure of balance between the two. A higher F1-score suggests that the test method is more effective in identifying positive samples and reducing the occurrence of misclassification. The F1-Score is obtained by applying the formula in Equation 5.

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (5)$$

3. RESULT

This study aims to perform early detection of GERD by utilizing machine learning techniques, namely feature extraction, feature selection, and data sampling. In this research, feature extraction is done with PCA, feature selection uses Pearson Correlation, and the sampling technique used is SMOTE. The machine learning model applied in this research is XGBoost.

3.1. Data Exploration

3.1.1. Descriptive Statistics

This initial phase involved an analysis of descriptive statistics to provide insight into the overall structure of the data. Descriptive statistical analysis includes calculating the mean and median values to understand the basic characteristics of each feature. Additionally, data distribution is analyzed, such as dividing the data into quartiles (Q1, Q2, and Q3), to ascertain the distribution of the data. The standard deviation is also calculated in order to understand the variance of the data from the mean value. The descriptive statistics of the data in this study can be seen in Table 1.

Table 1. Descriptive Statistics.

Feature	Count	Mean	Std	Min	Q1 (25%)	Median (50%)	Q3 (75%)	Max
IndexNo	1200.00	600.50	346.55	1.00	300.75	600.50	900.25	1200.00
HeartBurn	1200.00	1.74	1.23	1.00	1.00	1.00	2.00	5.00
DisabilityHB	1200.00	1.49	0.87	1.00	1.00	1.00	2.00	4.00
HBorNot	1200.00	1.66	0.47	1.00	1.00	2.00	2.00	2.00
HBfrequency	1200.00	0.61	0.95	0.00	0.00	0.00	1.00	3.00
…	…	…	…	…	…	…	…	…
frequentalcoholnew	1200.00	0.04	0.20	0.00	0.00	0.00	0.00	1.00
COFFEEINTAKEPERMONTH	1200.00	0.36	0.48	0.00	0.00	0.00	1.00	1.00
BELOWLEVEL	1200.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00
Incomelessthan5000	1200.00	0.61	0.49	0.00	0.00	1.00	1.00	1.00

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 855-870
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

3.1.2. Data Visualization

Data visualization is used to identify patterns and trends in the dataset. The distribution comparison between the two classes in the target column can be seen on the visualization. The data set reveals that Class 1 (Non-probable GERD) has a larger population than Class 2 (Probable GERD). The significant difference in the number of the two classes indicates an imbalance in the data. To overcome this problem, this study uses the SMOTE (Synthetic Minority Over-sampling Technique) technique on the training data to handle class imbalance and ensure the model is not biased towards either class when training. Figure 3 shows the visualization of the target class distribution in the data.

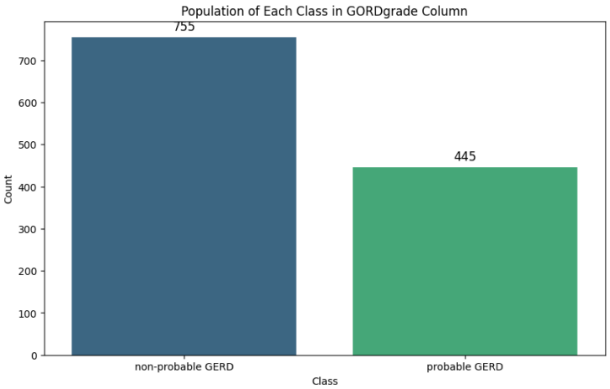

Figure 3. Target Column Class Distribution.

Furthermore, each feature was analyzed based on the histogram graph in Figure 4. The histogram graph shows the overall distribution of values in each feature to identify any concentration points or distribution patterns in each feature. From the histograms, it was found that most of the features were centered around certain values. The imbalance in the values of the various features also supports the finding that the data does not represent the overall character of the class well. However, some features have an even distribution across various value intervals, such as Age and PSSScore.

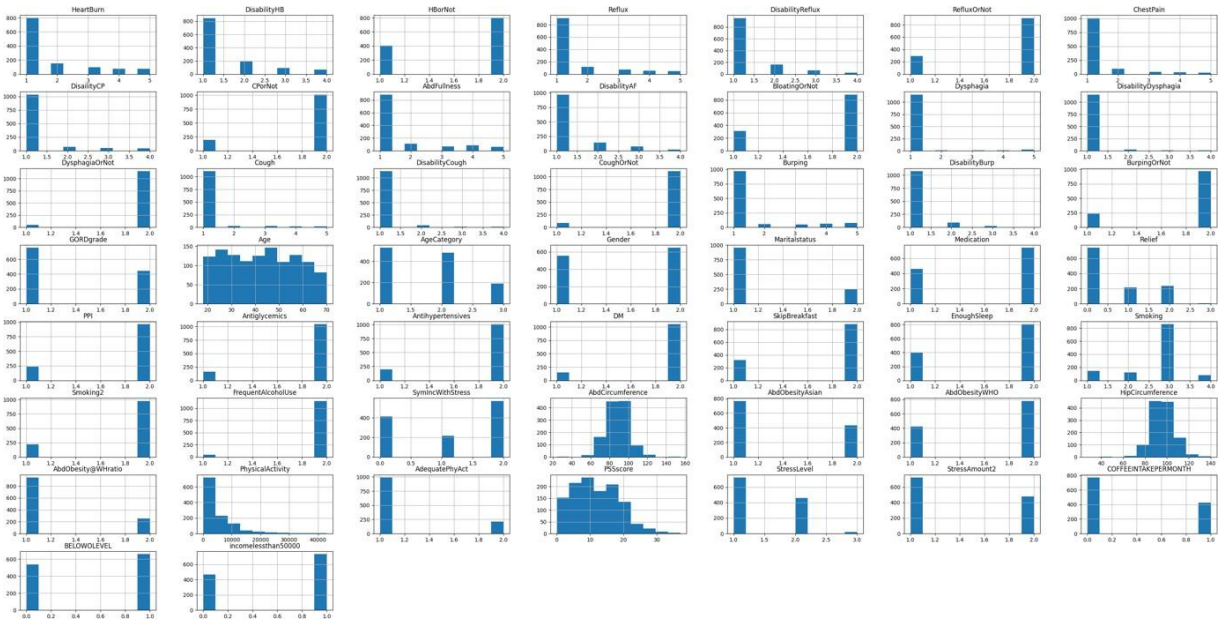

Figure 4. Features Values Distribution.

3.2. Data Pre-processing

3.2.1. Data Cleaning and Handling Missing Value

After analyzing the data, it was found that some values in the PhysicalActivity feature were outside the appropriate range. To address this issue, the data type in the column was converted to numeric, resulting in 19 rows with null values (NaN). These rows were then imputed using the mean value in order to maintain data completeness. Furthermore, some redundant features that have similar meanings with other features are removed to reduce the computational burden, such features include HBfrequency, RefluxFrequency, CPfrequency, BloatingFrequency, DysphagiaFrequency, CoughFrequency, BurpingFrequency, antiglycemicnew, antihypertensivenew, skipbreakfastnew, inadequatesleepnew, currentsmokernew, and frequentalcoholnew. After data was cleaned and missing values were handled, total number of columns or features used in this study became 51 columns and 1,200 rows.

3.2.2. Data Scaling

Data scaling is conducted in order to ensure that the range of values within each feature is consistent and can be utilized optimally at the feature extraction stage through the application of Principal Component Analysis (PCA). The results of data scaling process can be observed in Table 2.

Table 2. Data Scaling Results.

Feature	Sample Data before Scaling	Sample Data after Scaling
HeartBurn	2	0.207437
DisabilityHB	2	0.605915
HBorNot	1	-1.391323
…	…	…
COFFEINTAKEPERMONTH	1	1.332129
BELOWLEVEL	0	-1.096282
Incomelessthan50000	0	-1.240839

3.3. Data Splitting

The data set is divided into two portions, comprising 80% for training and 20% for testing, respectively, based on the entire data set utilized in the study. The training data is employed for model training and development so that the algorithms can learn and comprehend patterns within the data. Subsequently, the testing data is employed to evaluate the efficacy of trained models and assess their performance. The outcomes of data splitting are illustrated in Table 3.

Table 3. Data Splitting Results.

Data	Non-probable gerd	Probable gerd	Total
Training Data	591	369	960
Testing Data	164	76	240

3.4. Feature Selection and Feature Extraction

In this stage, a feature selection technique based on Pearson correlation coefficients will be employed, and Principal Component Analysis (PCA) will be utilized for the extraction of features.

3.4.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is used for feature extraction with the aim of reducing the dimensionality of the data while retaining the variance of the information. At this stage, the cumulative

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

variance was used to determine the number of principal components to be retained. In order to select the optimal number of components with a computationally efficient load while still representing the data comprehensively, a series of tests were conducted, retaining various levels of variance, specifically 95%, 75%, and 50%. The results of the principal component analysis (PCA) indicate that 32 principal components can be retained to maintain 95% of the variance, while 16 and 7 components can be retained to maintain 75% and 50% of the variance, respectively. The cumulative variance curves for the PCA components are presented in Figure 5.

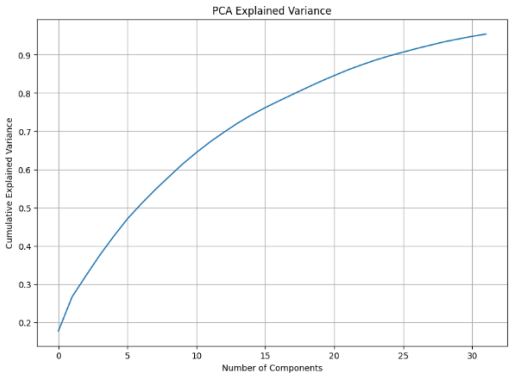

Figure 5. Cumulative Explained Variance Curve for PCA Components.

3.4.2. Pearson Correlation

In the feature selection stage, Pearson Correlation Coefficient (r) was used to calculate the correlation between each feature and the target column, GORDgrade. The objective of the correlation calculation is to identify the features that exhibit a strong linear relationship with the target column. The correlation value of each feature is sorted in descending order of the absolute value of the correlation coefficient. Based on the resulting rankings, the top 32, 16, and 7 features with the highest correlation values were selected for use in model training based on the number of components retained in PCA. The correlation of features to GORD grade based on the Pearson correlation calculation is illustrated in Figure 6.

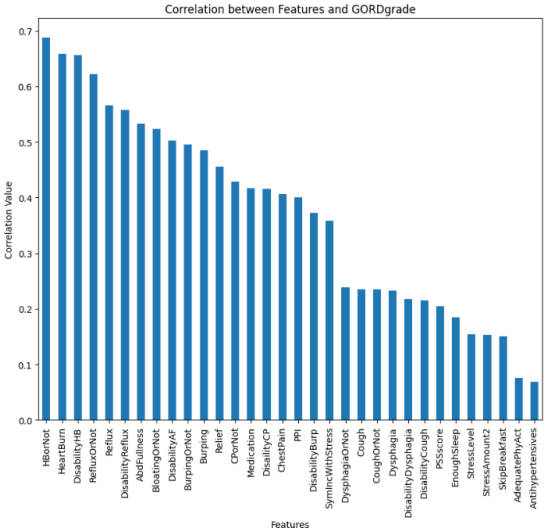

Figure 6. Feature Correlation with GORDgrade Using Pearson Correlation.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4143

3.5. Synthetic Minority Oversampling Technique (SMOTE)

This study requires the handling of unbalanced data, given that the target column in the dataset, designated as 'GORDgrade', exhibits an imbalance in the number of class values present. The "GORDgrade" column consists of 755 samples in class 1 (non-probable gastroesophageal reflux disease, or GERD) and 445 samples in class 2 (probable GERD). To address this issue, this research employs the Synthetic Minority Oversampling Technique (SMOTE) on the training data. In the training data, the class distribution in the target column consists of 591 samples for class 1 and 369 samples for class 2. Following the application of SMOTE, the class distribution in the target column becomes balanced, with 591 samples each for class 1 and class 2. A comparison of the class distribution in the target column before and after the application of SMOTE can be seen in Table 4.

Table 4. Class Distribution of Training Data Comparison Before and After SMOTE.

Total Training Data	Before SMOTE	After SMOTE
Non-probable gerd	591	591
Probable gerd	369	591

3.6. Machine Learning Model Performance

Based on the stages that have been previously completed, the results of the machine learning model, utilizing XGBoost for the detection of GERD on three data types, are obtained: original dataset, dataset with principal component analysis (PCA), and dataset consisting results from feature selection process using Pearson correlation. XGBoost model produces a macro average F1-score value of 0.9568 on the original dataset. These results are obtained from evaluation conducted on test dataset. Evaluation metrics for the performance of the XGBoost model with the original dataset are presented in Table 5.

Table 5. Original Dataset Model Performance.

Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
0.9625	0.9553	0.9584	0.9568

As illustrated in Table 6, XGBoost model demonstrates variability in results contingent upon the selection of components employed in the feature extraction technique utilizing PCA. The model produces a macro average F1-score of 0.9525 when 32 components, or 95% of the data variance, are used. Subsequently, when the number of components was reduced to 16 or 75% of the data variance, the macro average F1-score increased to 0.9615. For the number of components 7 or 50% of the data variance, the macro average F1-score obtained was 0.9476.

Table 6. PCA Model Evaluation Metrics Comparison.

Total Variance / Component	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
95% / 32	0.9583	0.9469	0.9589	0.9525
75% / 16	0.9667	0.9615	0.9615	0.9615
50% / 7	0.9542	0.9433	0.9523	0.9476

The results of the XGBoost model for early detection of GERD with feature selection techniques using Pearson correlation are shown in Table 7. By selecting the 32 most correlated features using Pearson correlation, the model achieved a macro average F1-score of 0.9711. When the number of features was reduced to 16 and 7, the macro average F1-score values were 0.9809 and 0.9095, respectively.

Table 7. Pearson Correlation Model Evaluation Metrics Comparison.

Total Feature	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
32	0.9750	0.9711	0.9711	0.9711
16	0.9833	0.9777	0.9843	0.9809
7	0.9208	0.9055	0.9138	0.9095

4. DISCUSSIONS

The results of the machine learning model performance indicate that the use of PCA as a feature extraction technique yields the most optimal results when 16 principal components or 75% of the data variance are retained, with a macro average F1-score of 0.9615. This indicates that the model attains optimal performance by retaining the majority of significant variance information within the data set. However, when 32 principal components or 95% of the data variance were employed, the model's performance actually decreased due to the model becoming more complex. Conversely, insufficient components were also identified when utilising seven principal components or 50% of the variance, which was inadequate for capturing the relevant data variation for GERD detection.

In the feature selection technique utilizing Pearson correlation, the selection of the optimal 16 features yielded a macro average F1-score of 0.9809. This demonstrates that the quality of the model can be enhanced even when the number of features is reduced. By reducing the number of features, the model's efficiency can be maintained, its complexity can be reduced, and overfitting can be avoided. Conversely, the use of an insufficient number of features, namely seven, resulted in a reduction in model performance, with a macro average F1-score of 0.9095. This is because the reduction in the number of features can lead to the elimination of some crucial information that is necessary for the model to function effectively.

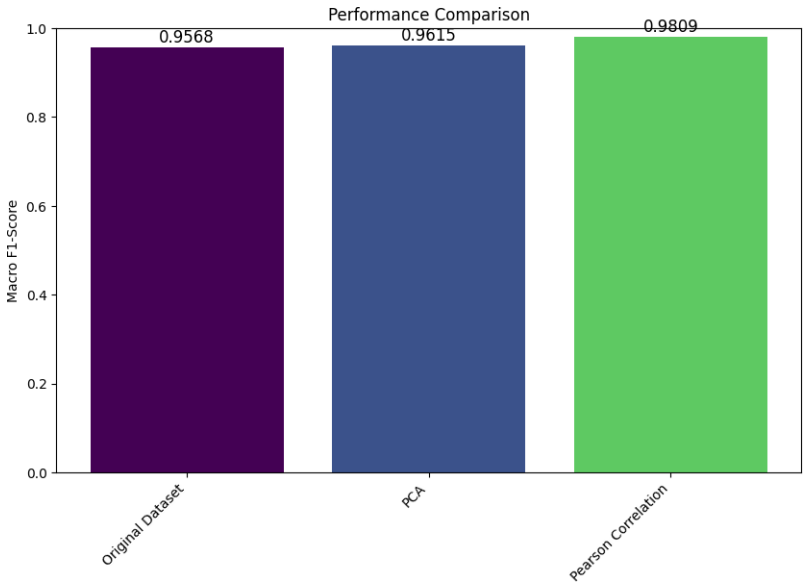

Figure 7. Models F1-Score Comparison.

Figure 7 illustrates the effectiveness of the XGBoost machine learning model in detecting GERD. The dataset that underwent feature selection with Pearson Correlation demonstrated the highest macro average F1-score, reaching 0.9809. In contrast, the original dataset yielded a macro average F1-score of 0.9568, while the dataset extracted through PCA achieved a macro average F1-score of 0.9615. These results demonstrate that feature selection using Pearson correlation is more effective in improving model performance than using the original dataset or PCA. The selection of features based on their correlation

with the target has a notable impact on the accuracy of the model, thereby enhancing the prediction quality. Conversely, the application of PCA still yields favorable performance, despite a slight decline in accuracy, which could be explained by the discrepancy in feature representation following transformation. This study illustrates that feature selection based on correlation is a more suitable approach for detecting GERD with the XGBoost. . The findings of this study are consistent with prior studies conducted by Pei et al., Raihan et al., and Zhang et al [11], [24], [25], which demonstrated model performance or accuracy above 90%. A comparison of the performance or accuracy of the XGBoost model in previous studies with this study can be seen in Table 8.

Table 8. Comparison of XGBoost Model Performance with Other Studies.

Author	Accuracy
Pei et al.	94.15%
Raihan et al.	99.16%
Zhang et al.	94.7%
Our Proposed Method	98.33%

5. CONCLUSION

This study implements XGBoost machine learning model for the early detection of GERD disease, leveraging GERD disease dataset procured from the Harvard Dataverse online data repository. The dataset comprises 69 attributes, including clinical symptoms of GERD disease and patient habits. Additionally, data cleaning was conducted to address the presence of attributes with identical or redundant meanings. Furthermore, the target attribute exhibits a significantly higher prevalence of non-probable GERD labels compared to probable GERD labels. Consequently, this study employed the SMOTE technique to address the issue of data imbalance. Furthermore, this study employed principal component analysis (PCA) for feature extraction and Pearson correlation for feature selection, with the objective of identifying the most significant attributes and enhancing computational efficiency. Test results indicated that the optimal model performance was achieved when PCA retained 75% of the data variance or utilized 16 principal components, and when Pearson Correlation selected 16 features with the highest correlation to the target. The application of these techniques yielded a macro average F1-score of 0.9615 for PCA and 0.9809 for Pearson Correlation. The use of the original dataset resulted in a macro average F1-score of 0.9568. Experimental results demonstrate that feature selection through Pearson Correlation has a better f1-score value than the feature extraction method with PCA or based on the original dataset in enhancing the performance of XGBoost model for the early detection of GERD in this study. The research conducted with the XGBoost classification model has the potential to be utilized in a decision support system for the early detection of GERD. This system could reduce the number of undetected GERD cases in a timely manner and facilitate the implementation of preventive measures. Further research should be conducted with larger and more diverse datasets in order to improve the overall generalization of the model. Moreover, further research could be conducted to explore the use of alternative deep learning and machine learning models.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] K. H. A. Boulton and P. W. Dettmar, "A narrative review of the prevalence of gastroesophageal reflux disease (GERD)," *Ann Esophagus*, vol. 5, no. 57–66, pp. 7–7, Mar. 2022, doi: 10.21037/aoe-20-80.

[2] M. Durazzo *et al.*, "Extra-Esophageal Presentation of Gastroesophageal Reflux Disease: 2020 Update," *JCM*, vol. 9, no. 8, p. 2559, Aug. 2020, doi: 10.3390/jcm9082559.

[3] J. Maret-Ouda, S. R. Markar, and J. Lagergren, "Gastroesophageal Reflux Disease: A Review," *JAMA*, vol. 324, no. 24, p. 2536, Dec. 2020, doi: 10.1001/jama.2020.21360.

[4] A. Ravindran and P. G. Iyer, "Gastroesophageal Reflux Disease and Complications," in *Geriatric Gastroenterology*, C. S. Pitchumoni and T. S. Dharmarajan, Eds., Cham: Springer International Publishing, 2020, pp. 1–17. doi: 10.1007/978-3-319-90761-1_42-1.

[5] R. Fass, "Gastroesophageal Reflux Disease," *N Engl J Med*, vol. 387, no. 13, pp. 1207–1216, Sep. 2022, doi: 10.1056/NEJMcp2114026.

[6] R. Fass, G. E. Boeckxstaens, H. El-Serag, R. Rosen, D. Sifrim, and M. F. Vaezi, "Gastro-oesophageal reflux disease," *Nat Rev Dis Primers*, vol. 7, no. 1, p. 55, Jul. 2021, doi: 10.1038/s41572-021-00287-w.

[7] D. A. Katzka and P. J. Kahrilas, "Advances in the diagnosis and management of gastroesophageal reflux disease," *BMJ*, p. m3786, Nov. 2020, doi: 10.1136/bmj.m3786.

[8] J. S. Nirwan, S. S. Hasan, Z.-U.-D. Babar, B. R. Conway, and M. U. Ghori, "Global Prevalence and Risk Factors of Gastro-oesophageal Reflux Disease (GORD): Systematic Review with Meta-analysis," *Sci Rep*, vol. 10, no. 1, p. 5814, Apr. 2020, doi: 10.1038/s41598-020-62795-1.

[9] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022, doi: 10.3390/healthcare10030541.

[10] A. Maydeo *et al.*, "Impact of Mobile Endoscopy Unit for Rendering Gastrointestinal Endoscopy Services at Two Community Health Centers in Western India," *Journal of Digestive Endoscopy*, vol. 12, no. 04, pp. 190–195, Dec. 2021, doi: 10.1055/s-0041-1741387.

[11] X. Pei, Q. Deng, Z. Liu, X. Yan, and W. Sun, "Machine Learning Algorithms for Predicting Fatty Liver Disease," *Ann Nutr Metab*, vol. 77, no. 1, pp. 38–45, 2021, doi: 10.1159/000513654.

[12] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, p. baaa010, Jan. 2020, doi: 10.1093/database/baaa010.

[13] S. Khandakar, "Unveiling Early Detection And Prevention Of Cancer: Machine Learning And Deep Learning Approaches:," *EATP*, vol. 30, no. 5, pp. 14614–14628, May 2024, doi: 10.53555/kuey.v30i5.7014.

[14] M. Shehab *et al.*, "Machine learning in medical applications: A review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, p. 105458, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105458.

[15] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny, "A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues," *Journal of Biomedical Informatics*, vol. 113, p. 103627, Jan. 2021, doi: 10.1016/j.jbi.2020.103627.

[16] F. Mazhar, M. Sajid, N. Aslam, M. Imran, and H. Ahmad, "Boosting Early Diabetes Detection: An Ensemble Learning Approach with XGBoost and LightGBM," *JCBI*, vol. 6, no. 02, Mar. 2024, doi: https://doi.org/10.56979/602/2024.

[17] S. Kabiraj *et al.*, "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India: IEEE, Jul. 2020, pp. 1–4. doi: 10.1109/ICCCNT49239.2020.9225451.

[18] Srividya B. V. and S. Sasi, "Early Detection of Gastroesophageal Reflux Disease Using Logistic Regression and Support Vector Machine:," *International Journal of Organizational and Collective Intelligence*, vol. 11, no. 2, pp. 75–90, Apr. 2021, doi: 10.4018/IJOCI.2021040104.

[19] J. H. Rubenstein *et al.*, "Predicting Incident Adenocarcinoma of the Esophagus or Gastric Cardia Using Machine Learning of Electronic Health Records," *Gastroenterology*, vol. 165, no. 6, pp. 1420-1429.e10, Dec. 2023, doi: 10.1053/j.gastro.2023.08.011.
[20] N. Wickramasinghe *et al.*, "The association between symptoms of gastroesophageal reflux disease and perceived stress: A countrywide study of Sri Lanka," *PLoS ONE*, vol. 18, no. 11, p. e0294135, Nov. 2023, doi: 10.1371/journal.pone.0294135.
[21] J. Wu, Y. Li, and Y. Ma, "Comparison of XGBoost and the Neural Network model on the class-balanced datasets," in *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, Greenville, SC, USA: IEEE, Nov. 2021, pp. 457–461. doi: 10.1109/ICFTIC54370.2021.9647373.
[22] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, "An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach," *MethodsX*, vol. 10, p. 102119, 2023, doi: 10.1016/j.mex.2023.102119.
[23] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, "eXtreme Gradient Boosting Algorithm with Machine Learning: a Review," *ACAD J NAWROZ UNIV*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.
[24] Md. J. Raihan, Md. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci Rep*, vol. 13, no. 1, p. 6263, Apr. 2023, doi: 10.1038/s41598-023-33525-0.
[25] S. Zhang, Y. Yuan, Z. Yao, X. Wang, and Z. Lei, "Improvement of the Performance of Models for Predicting Coronary Artery Disease Based on XGBoost Algorithm and Feature Processing Technology," *Electronics*, vol. 11, no. 3, p. 315, Jan. 2022, doi: 10.3390/electronics11030315.
[26] R. Magdum, "What is Data Exploration? and its Importance in Data Analytics," *IRJET*, vol. 09, no. 01, pp. 1482–1485, 2022.
[27] A. D. Gupta, K. Singh, K. D. Singh, P. Kushwaha, B. P. Lohani, and S. Kumar, "Unveiling Insights: Exploring Healthcare Data through Data Analysis," in *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, Gautam Buddha Nagar, India: IEEE, May 2024, pp. 575–581. doi: 10.1109/IC3SE62002.2024.10593333.
[28] O. Sami, Y. Elsheikh, and F. Almasalha, "The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease," *IJACSA*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120695.
[29] F. Xiong, C. Cao, M. Tang, Z. Wang, J. Tang, and J. Yi, "Fault Detection of UHV Converter Valve Based on Optimized Cost-Sensitive Extreme Random Forest," *Energies*, vol. 15, no. 21, p. 8059, Oct. 2022, doi: 10.3390/en15218059.
[30] S. Justin, W. Saleh, T. Al Ghamdi, and J. Shermina, "Hyperparameter Optimization Based Deep Belief Network for Clean Buses Using Solar Energy Model," *Intelligent Automation & Soft Computing*, vol. 37, no. 1, pp. 1091–1109, 2023, doi: 10.32604/iasc.2023.032589.
[31] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.
[32] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Minimizing features while maintaining performance in data classification problems," *PeerJ Computer Science*, vol. 8, p. e1081, Sep. 2022, doi: 10.7717/peerj-cs.1081.
[33] H. Mamdouh Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft Comput*, vol. 27, no. 16, pp. 11259–11274, Aug. 2023, doi: 10.1007/s00500-023-08587-x.
[34] Z. M. Zain *et al.*, "Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis," *Int. J. Adv. Intell. Informatics*, vol. 6, no. 3, p. 313, Nov. 2020, doi: 10.26555/ijain.v6i3.462.
[35] F. Bagherzadeh, "Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance," *Journal of Water Process Engineering*, vol. 41, p. 102033, 2021, doi: https://doi.org/10.1016/j.jwpe.2021.102033.

[36] F. Zinzendoff Okwonu, B. Laro Asaju, and F. Irimisose Arunaye, "Breakdown Analysis of Pearson Correlation Coefficient and Robust Correlation Methods," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 917, no. 1, p. 012065, Sep. 2020, doi: 10.1088/1757-899X/917/1/012065.
[37] U. N. Wisesty, T. A. B. Wirayuda, F. Sthevanie, and R. Rismala, "Analysis of Data and Feature Processing on Stroke Prediction using Wide Range Machine Learning Model," *join*, vol. 9, no. 1, pp. 29–40, Apr. 2024, doi: 10.15575/join.v9i1.1249.
[38] I. Świetlicka, W. Kuniszyk-Jóźkowiak, and M. Świetlicki, "Artificial Neural Networks Combined with the Principal Component Analysis for Non-Fluent Speech Recognition," *Sensors*, vol. 22, no. 1, p. 321, Jan. 2022, doi: 10.3390/s22010321.
[39] R. Wardoyo, I. M. A. Wirawan, and I. G. A. Pradipta, "Oversampling Approach Using Radius-SMOTE for Imbalance Electroencephalography Datasets," *Emerg Sci J*, vol. 6, no. 2, pp. 382–398, Mar. 2022, doi: 10.28991/ESJ-2022-06-02-013.
[40] S. Yang and G. Berdine, "Confusion matrix," *The Chronicles*, vol. 12, no. 53, pp. 75–79, Oct. 2024, doi: 10.12746/swrccc.v12i53.1391.