Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 5, No. 3, July 2023, pp: 150-158; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received May 27, 2023; revised June 20, 2023; accepted June 21, 2023; date of publication July 30, 2023
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeemi.v5i3.305

How to cite: Shalehah , Muhammad Itqan Mazdadi , Andi Farmadi , Dwi Kartini , and Muliadi, "Implementation of Particle Swarm Optimization Feature Selection on Naïve Bayes for Thoracic Surgery Classification", vol. 5, no. 3, pp. 150–158, July 2023.

Implementation of Particle Swarm Optimization Feature Selection on Naïve Bayes for Thoracic Surgery Classification

Shalehah , Muhammad Itqan Mazdadi , Andi Farmadi , Dwi Kartini , and Muliadi

Fakultas Matematika dan Ilmu Pengetahuan Alam , Universitas Lambung Mangkurat , Indonesia

Corresponding author: shlhhleha@gmail.com

ABSTRACT The use of the Naïve Bayes method alone in thoracic surgery classification often does not yield optimal results due to the complexity of the dataset and the numerous attributes that must be considered. As such, an additional method is required to enhance the accuracy and efficiency of the classification process. This study aims to compare the accuracy of all research models using Naïve Bayes with and without the PSO technique. The contribution is to enrich an understanding of the application of classification techniques and feature selection in health datasets, particularly in the context of thoracic surgery. The research method encompasses the dataset used, the theory of the Naive Bayes algorithm, the PSO algorithm, validation testing using separate validation, and performance assessment with the confusion matrix and AUC evaluation approach. Secondary data for this investigation was procured via the UCI Repository website. The accuracy was augmented using the PSO technique for thoracic surgery weight optimization. The sample in the study consisted of 470 data items, with 70 sample data from the class that died within one year and 400 samples relating to the surviving class. The testing outcomes of the Naive Bayes method using the thoracic surgery dataset yielded the highest accuracy of 81.91% with an 80:20 ratio and an AUC value of 0.620. The highest accuracy score was 93.62%, with an AUC value of 0.773 with a 90:10 ratio. Three features, PRE6, PRE14, and PRE17, had zero weight. This accuracy score was achieved when PSO was employed to refine feature selection for attribute weighting. Hence, the accuracy of Naïve Bayes in thoracic surgery improved with attribute weighting in feature selection using PSO. Consequently, this research enhances the precision and efficiency of thoracic surgery data processing, aiding lung cancer diagnosis speed and accuracy.

INDEX TERMS Naïve Bayes, Particle Swarm Optimization, Thoracic Surgery

I. INTRODUCTION

Lung cancer therapy and interventions must be rapid and focused. Surgery, radiation, chemotherapy, immunotherapy, hormone therapy, and gene therapy are options for treating lung cancer. One of the most common procedures performed on lung cancer patients is thoracic surgery. Thoracic surgery has risks and benefits for patients in both the short and long term, making it a significant issue in lung cancer patient management [1], [2]. Surgical care affecting the chest, often referred to as the thorax, is the focus of the surgical specialization known as thoracic surgery [3]–[5]. The life expectancy of patients one year after thoracic surgery is one of the issues in thoracic surgery research; therefore, in this study, classification is used to determine whether patients survived or died.

Naïve Bayes is one of the data mining classification methods that can handle thoracic surgery data [6], [7]. The Naïve Bayes algorithm uses the past to estimate potential possibilities in the future. Another benefit of naïve Bayes is its straightforward method that can provide highly accurate results [8]–[10]. Before implementing a classification model, data validation is required at the data mining stage. This can be done using separate validation approaches, which divide data into training and testing sets. This approach helps verify that the developed data model is correct and can be used in subsequent operations. To identify whether the accuracy value in the separate validation approach is more significant, a study explained that the distribution of data into four ratio variances, achieving a maximum accuracy of 93.00% [11].

In a study which classified thoracic surgery datasets using several methods, the highest accuracy rate was 85%, and the average AUC value was 0.787 [5]. This study shows that classification techniques can help process data within datasets. In thoracic surgery was classified using several methods, obtaining an accuracy rate of 84.51% and a ROC value of 0.738. This research applied classification methods with lower performance compared to classification methods applying other methods; thus, combining the Naïve Bayes classification method with feature selection techniques is necessary to achieve more effective and accurate results (Santoso, 2021).

Feature selection is a technique for reducing attribute dimensions. This dimension reduction is performed to obtain relevant and non-excessive attributes to speed up the classification process and increase the accuracy of classification algorithms [13]. Research by Geetha Pavani who performed feature selection and classification on thoracic surgery datasets [14]. The feature selection technique used in this study is Particle Swarm Optimization; although this method is an optimization method, it is used for feature selection in this research. Since PSO is an algorithm used for decision-making in searching for the best solution in the data mining classification process. Feature selection is used to improve the effectiveness and efficiency of classification algorithm performance. In addition, other studies have also compared PSO and C4.5 in classifying blood sugar datasets. It was concluded that the addition of PSO feature selection can improve accuracy, which is more than 95% superior to C4.5 [15].

Based on the justification provided, this research suggests using the Particle Swarm Optimization method and the Naive Bayes algorithm to categorize datasets related to thoracic surgery. The thoracic surgery dataset has 470 data, including 16 predictive variables and one target attribute for classification. Combining these properties with additional classification techniques will yield the best results because they are too numerous. Naïve Bayes may produce poor accuracy rankings because its weakness is susceptibility to too many characteristics [8], [12]. As a result, PSO should be used in this research's feature selection process to increase the accuracy value of the layer dataset [16]. Then, PSO must be used in the feature selection process of this research to increase the accuracy value of the thoracic surgery dataset.

The purpose of this research is to analyze the accuracy comparison of all research models using Naïve Bayes with and without using Particle Swarm Optimization. By applying Particle Swarm Optimization, the Naïve Bayes algorithm is expected to be more efficient and effective in classifying thoracic surgery data, resulting in higher accuracy. The results of this research are expected to provide contributions such as:

a. Enhancing understanding of the application of classification techniques and feature selection in health datasets, particularly in thoracic surgery cases.
b. Assisting medical professionals in optimizing decision-making based on analysis.
c. More accurate data evaluation using the Naïve Bayes algorithm and Particle Swarm Optimization.

II. METHOD

This research method explains the dataset used, the theory of the Naïve Bayes algorithm, the Particle Swarm Optimization algorithm, testing validation using Split Validation, and performance measurement using the evaluation methods of Confusion Matrix and AUC. The distribution of training data and testing data. The data division in separate validation consists of various ratio variations, namely 70:30, 80:20, and 90:10. The following is the research procedure to be carried out. FIGURE 1 shows the flow of this research.

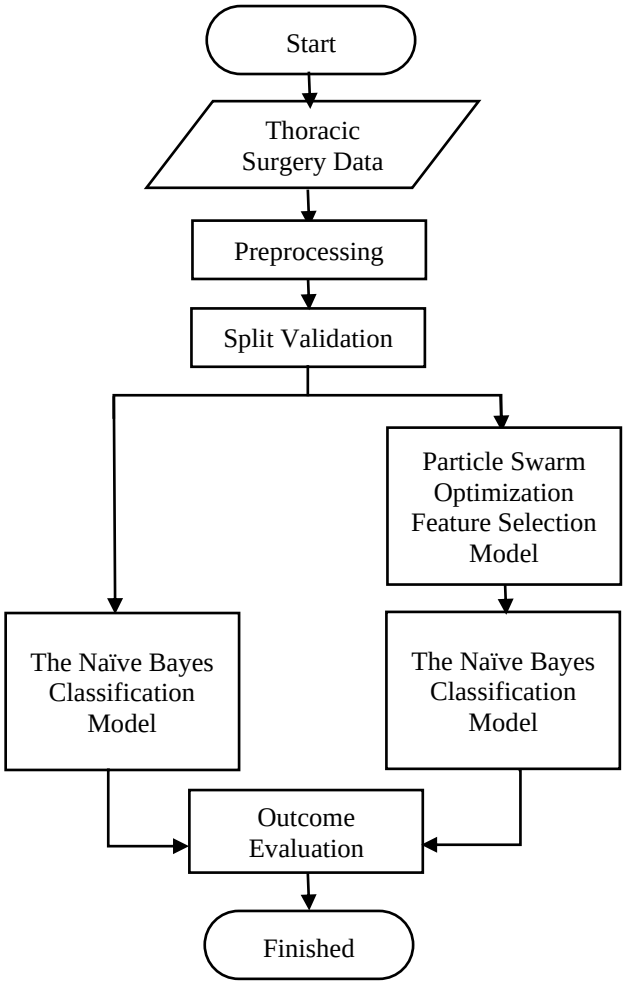


FIGURE 1.Research Flowchart

A. DATA COLLECTION

In this study, secondary data were obtained from the UCI Repository website. This dataset is about classification problems related to the life expectancy of patients with lung cancer after surgery, where death occurs within one year after the operation. There are 470 data points and 16 attributes as predictor variables and 1 attribute as the target variable. The

dataset contains information about each patient represented by 16 attributes, which are preoperative and postoperative conditions. The cancer surgery data contains patient data from patients who underwent cancer surgery between 2007-2011 [17].

The 16 attributes are a combination of nominal, binary, and numeric data. The thoracic surgery patient data has two classes: death within one year (True) and survival (False), with 70 samples for the true class and 400 samples for the false class. The following are the attributes and descriptions of the thoracic surgery dataset, as shown in TABLE 1.

TABLE 1
Surgery Data Attribute Description

No	Attribute	Description	Category	Range
1	DGN	Diagnosis - specific combination of ICD-10 codes for primary and secondary tumors and more than one tumor, if present	Nominal	{DGN1, DGN2, DGN3, DGN4, DGN5, DGN6, DGN8}
2	PRE4	The amount of air that can be forcibly exhaled from the lungs after taking the deepest possible breath (FVC)	Numeric	{1.44, 6.3}
3	PRE5	Amount of air exhaled at the end of the first second of FVC (FEV1)	Numeric	{0.96, 86.3}
4	PRE6	A measure of the general ability of cancer patients in daily activities (Zubrod Scale)	Nominal	{PRZ0, PRZ2, PRZ2}
5	PRE7	Pain before surgery	Binary	{T,F}
6	PRE8	Haemoptysis before surgery	Binary	{T,F}
7	PRE9	Dyspnoea before surgery	Binary	{T,F}
8	PRE10	Cough before surgery	Binary	{T,F}
9	PRE11	Weak condition before surgery	Binary	{T,F}
10	PRE14	Tumor size (TNM)	Nominal	{OC11, OC12, OC13, OC14}
11	PRE17	Diabetes mellitus	Binary	{T,F}
12	PRE19	Myocardial Infarction (MI) up to 6 months	Binary	{T,F}
13	PRE25	Diseases that affect the arteries/blood vessels	Binary	{T,F}
14	PRE30	Smoke	Binary	{T,F}
15	PRE32	Asthma	Binary	{T,F}
16	AGE	Age at surgery	Numeric	{21, 87}
17	Risk1Y	Survival period live 1 year - ()		

B. NAÏVE BAYES

Thomas Bayes, a British physicist, developed the Naive Bayes algorithm. The Bayes theorem is an algorithm that forecasts opportunities based on past performance. The theorem is paired with Naive, which makes the assumption that the criteria governing the characteristics are independent of one another. Therefore, the Nave Bayes method makes the assumption that the existence or absence of certain class features has no effect on the traits of other classes[16]. The equation for Bayes' theorem in Eq. (1) is often as follows.

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (1)$$

Several variables need to be explained in the Bayes theorem Eq. (1). The variable y represents data with an unknown class, which we want to predict based on the Naïve Bayes model. Meanwhile, the variable x represents the hypothesis stating that the data y belongs to a certain class. Furthermore, this study has P(x|y), the probability of hypothesis x gave the condition of y, known as the posterior probability. This probability describes the extent to which hypothesis x is expected to be true, given the condition of y. Then, there is P(x), the probability of hypothesis x before considering the data y, known as the prior probability. This probability is used to assess the likelihood of hypothesis x occurring without considering the information provided by data y. In addition, P(y|x) is the probability of data y being observed based on hypothesis x's condition. This provides information about the relationship between data y and the hypothesis x being evaluated and the extent to which data y is expected within the context of x. Finally, P(y) is the overall probability of data y, which explains how often data y occurs in a given dataset. Understanding each variable in the Bayes theorem equation can be correctly applied to develop an efficient and accurate Naïve Bayes model, with or without using feature selection methods such as Particle Swarm Optimization.

C. PARTICLE SWARM OPTIMIZATION (PSO)

The best solution to a given problem is found using the artificial intelligence-based technique Particle Swarm Optimization to resolve optimization problems. This approach draws its inspiration from the movement patterns of fish, herbivores, and birds, where each animal thing is broken down into a particle. J. Kennedy and R.C. Eberhart were the ones who first suggested the PSO algorithm. A population-based iterative algorithm is PSO. Numerous particles make up the population, which is utilized to solve optimization problems after being started with a population of random solutions. As a result, throughout the search process, particles tend to fly towards more effective search locations. [17]. The following formula may be used to get the position displacement and particle velocity in Eq. (2) and (3):

$$Vi\,(t) = Vi\,(t-1) + c1r1\,[XPbest\,i - Xi(t)] + c2r2\,[XGbest\,i - Xi\,(t)] \quad (2)$$

$$Xi\,(t) = Xi\,(t-1) + Vi\,(t) \quad (3)$$

In Eq. (2) and (3), the variables used to calculate the displacement and velocity of particles are as follows. Vi(t) denotes the velocity of particle i at iteration t, which is used to determine how fast the particle moves in searching for the optimal solution. Xi(t) is the position of particle i at iteration t, reflecting the solution achieved by the particle then. Then, c1 and c2 are learning rate factors that indicate the extent to which individual particle (cognitive) and social (group) influences affect the change in particle velocity. Here, c1 describes the extent to which a particle considers its success in finding a solution, while c2 indicates the influence of its group members in searching for a better solution. Next, r1 and r2 are random numbers uniformly distributed between intervals 0 and 1. These random numbers add a stochastic component to the solution search process so particles can achieve more optimal solutions by involving some random exploration in the search space.

XPbest i is the best position of particle i that has been achieved so far in the search for the optimal solution. This reflects the individual particle's achievements so far and is used to direct particle movement toward a better solution. Meanwhile, XGbest i is the best global position achieved among all particles in the group. This indicates the best solution the entire group finds and provides direction for particles to achieve more optimal solutions collectively. By incorporating all these variables in equations 2 and 3, the displacement and velocity of particles in the Particle Swarm Optimization algorithm can be calculated. This process is repeated until the desired solution or iteration limit is reached, resulting in an optimal solution for the problem.

D. *CONFUSION MATRIX*

One technique for evaluating a categorization method's effectiveness is the confusion matrix. The confusion matrix comprises data that contrasts the classification outcomes produced by the system with the expected classification outcomes [18]. Based on the computation of the testing object, the classification model is evaluated using the confusion matrix. It is collated into a table where its prognosis for correctness and incorrectness is provided [19], [20]. The matrix is explained in TABLE 2.

TABLE 2
Confusion Matrix

Classification	Predicted Class	
	Class = Yes	Class = No
Class = Yes	True Positif (TP)	False Negatif (FN)
Class = No	False Positif (FP)	True Negatif (TN)

The following is a formula for measuring the level of accuracy in the confusion matrix in equation 4, namely:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}\ x\ 100 \qquad (4)$$

In the accuracy equation (4), each variable has a significant meaning within the research data classification context. Accuracy is the metric that measures how well the classification model identifies the correct results. The accuracy value is obtained by calculating the proportion of the number of correctly classified observations compared to the total number of observations.

The TP (True Positive) variable describes the number of cases where the classification model correctly identifies a positive outcome. In the context of this research, this means the number of cases where the model identifies patients who indeed needed surgery as needing surgery. TN (True Negative) is the number of cases where the classification model correctly identifies a negative result, meaning the number of cases where the model identifies patients who do not need surgery as not requiring surgery.

FP (False Positive) and FN (False Negative) describe errors in the model's predictions. FP is the number of cases where the classification model incorrectly identifies a positive outcome, that is, the number of cases where the model identifies patients who do not need surgery as needing surgery. Meanwhile, FN is the number of cases where the classification model incorrectly identifies a negative outcome, meaning the number of cases where the model identifies patients requiring surgery as not needing surgery.

By understanding the meaning of each variable, one can see how the accuracy equation (4) reflects the performance of the classification model in identifying the correct outcomes, both in terms of positive and negative results. Through this research, the aim is to analyze the accuracy comparison of all models using Naïve Bayes with and without using Particle Swarm Optimization to determine whether implementing such optimization techniques can improve classification accuracy in thoracic surgery cases.

E. *AREA UNDER THE ROC (RECEIVER OPERATING CHARACTERISTIC) CURVE*

Calculating under the ROC curve often involves using the area under the curve formula. AUC may be thought of as a likelihood. The categorization approach will give the positive example a higher score than the negative example if one chooses a positive and negative example at random. As a result, a higher AUC value denotes a more effective classification approach, making the AUC value a maximization objective [20]. Because the unit square area's x- and y-axes have values ranging from 0 to 1, the Area under Curve (AUC) value will always fall within the 0–1 range. For values larger than 0.5, random guesses result in diagonal lines between (0.0) and (1.1) that have an area of 0.5. TABLE 3 shows many categories into which AUC values for data mining categorization may be classified [20].

TABLE 3
Accuracy Of Classification Results Based On Auc Value

AUC value	Category
0.90-1.00	Excellent Classification
0.80-0.90	Good Classification
0.70-0.80	Fair Classification
0.60-070	Poor Classification
0.50-0.60	Failure

1) DATA COLLECTION

The use of secondary data in study. Data acquired by third parties rather than directly coming from the subject of the study is known as secondary data. The thoracic surgery data used in this research may be obtained from the UCI Repository or viewed online at https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data. This dataset includes the postoperative survival rate of lung cancer patients, when death occurs within a year following surgery. The total number of data is 470, with 16 characteristics serving as predictor variables and 1 attribute serving as a target variable.

2) PREPROCESSING

The pre-processing stage aims to transform raw data into quality data, which can then be further processed. This stage is critical in data analysis as it ensures that the data used in the classification process do not contain issues or inconsistencies. The pre-processing process involves several steps, which will be detailed below. The first step in pre-processing is to identify and handle problematic data, such as empty data or errors [21]–[23]. In this study, after collecting the data, it was found that the data does not contain missing values or duplicates. Thus no further action needs to be taken. The second step is to transform the data to match the data type required by the Naive Bayes and Particle Swarm Optimization algorithms. The attributes used in the thoracic surgery dataset consist of nominal, binary, and numerical data.

Therefore, this stage involves converting nominal and binary attributes into numerical ones. This process is important as the Naive Bayes and Particle Swarm Optimization algorithms work with numerical data, so a conversion is needed for the data to be processed by the algorithms used. Once the conversion process is completed, the third step is to divide the dataset into two classes: patients who died within one year (True) and patients who survived (False) after thoracic surgery. The number of samples for the true class is 70 data, while the number of samples for the false class is 400 data. The data division is crucial for training the classification model later and for evaluating its performance against real cases in trials in the next stage. By completing this pre-processing stage, the thoracic surgery dataset has been collected, checked for accuracy, transformed into the appropriate data type, and divided into the required classes. Next, this dataset can be processed using the Naive Bayes and Particle Swarm Optimization algorithms to train the classification model and evaluate its performance in classifying the life expectancy figures of lung cancer patients after surgery.

3) PARTICLE SWARM OPTIMIZATION

At this point, the Naive Bayes classification is optimized using the Particle Swarm Optimization technique to assist raise the accuracy value of the suggested model. RapidMiner is the program used to put the Particle Swarm Optimization and Naive Bayes methods into action. The study's particle value is 10, the number of iterations is set to 30, and all other parameter settings are left at their default levels.

4) DATA SHARING

The data is first separated into training data and evaluated using split validation prior to classification. In this experiment, shuffled sampling was employed to distribute the data. This study used three different ratios—70:30, 80:20, and 90:10—in the distribution of training and testing data.

5) EVALUATION OF RESULTS

Using Particle Swarm Optimization along with a feature selection optimization approach, weights on each attribute will be optimized to maximize accuracy value. The confusion matrix and Area Under Curve (AUC) will be used during the evaluation phase. With the use of a connection matrix intended to measure the effectiveness of the model used, the assessment of the model will be carried out during the evaluation stage. AUC, or Area Under Curve, is a performance metric. This research will measure the effectiveness of Naïve Bayes and Naïve Bayes with the Particle Swarm Optimization approach before comparing them.

This study's features were taken from the thoracic surgery dataset obtained from the UCI repository. These features represent the patient's condition before and after the thoracic surgery, including nominal, binary, and numeric attributes. As part of the preprocessing process, features with nominal and binary data types were converted to numeric so the Naïve Bayes algorithm and Particle Swarm Optimization could process them.

Feature selection is an essential stage in data analysis, as the right features will improve the accuracy of the classification model. In this research, feature selection was carried out using the Particle Swarm Optimization (PSO) technique to enhance feature selection efficiency and improve the accuracy of the Naïve Bayes classification model. PSO was used to find the best feature combination that provides higher accuracy from the Naïve Bayes model. In optimizing using PSO, the number of particles used was 10, the number of iterations was set to 30, and using various ratio variations in the distribution of training and testing data were 70:30, 80:20, and 90:10.

After applying PSO in feature selection and choosing the best attributes, the confusion matrix and Area Under Curve (AUC) evaluation methods were used to assess the performance of the Naïve Bayes classification model with and without PSO optimization. The final results show a comparison of accuracy between the model that only uses Naïve Bayes and then the model that uses a combination of Naïve Bayes and Particle Swarm Optimization, so it could be seen whether the application of PSO provides an increase in accuracy in the classification of life expectancy numbers for lung cancer patients post-surgery.

III. RESULTS

A. THE RESULTS OF THE NAÏVE BAYES RESEARCH METHOD

The results of this study will present the results of experiments using the Naïve Bayes method with evaluation using Split Validation. After testing the model, the accuracy and AUC values will be obtained in TABLE 4 below:

TABLE 4
Naïve bayes accuracy results

Split Validation	Accuracy	AUC
70:30	81.56%	0.642
80:20	81.91%	0.620
90:10	76.60%	0.580

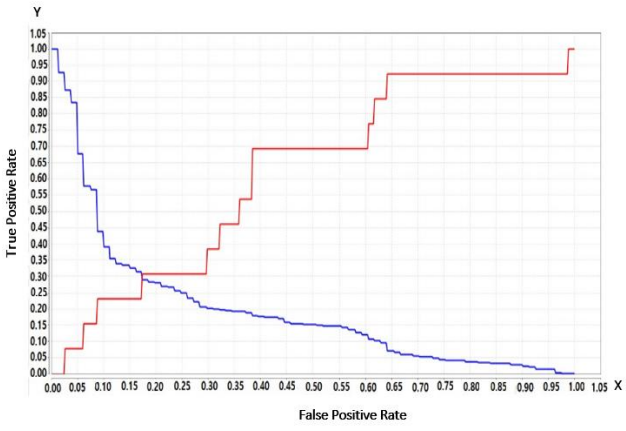
In TABLE 4, it is found that the Naïve Bayes model with a comparison of training data and test data of 80:20 has the highest accuracy value of 81.91% with an AUC value of 0.620.

TABLE 5
Confusion matrix naïve Bayes

Classification	Predicted Class	
	Class = F	Class = T
Class = F	75	11
Class = T	6	2

From the results of the confusion matrix in TABLE 5, the accuracy value is 81.91%

FIGURE 2. Naïve Bayes ROC Curve On 80:20 Split Validation



In testing the Naïve Bayes method, the ROC curve is also obtained, as shown in FIGURE 2 above, producing an AUC of 0.620. The AUC value is categorized as Poor Classification (poor) because it is in the value range of 0.60-0.70.

B. THE RESULTS OF THE RESEARCH USING THE NAÏVE BAYES METHOD WITH PARTICLE SWARM OPTIMIZATION

In order to choose features, Particle Swarm Optimization is used to optimize the weight of each characteristic in the dataset for thoracic surgery. The Particle Swarm Optimization method will be employed in RapidMiner calculations to assess the accuracy of Naive Bayes. The model will be put to the test, with the outcomes shown in TABLE 5 below:

TABLE 6
Naïve Bayes accuracy results with Particle Swarm optimization

Particle	Iteration	Split Rasio	Accuracy (%)	AUC
10	30	70:30	89.36	0.576
10	30	80:20	93.62	0.655
10	30	90:10	93.62	0.773

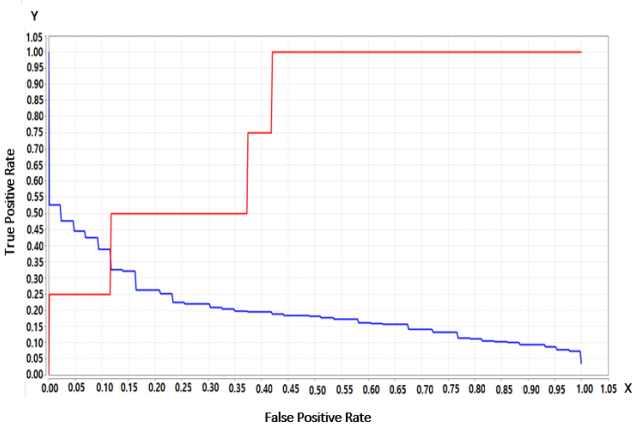
With a comparison of training data and test data of 80:20 and 90:10, TABLE 5 above demonstrates that the Nave Bayes and Particle Swarm Optimization models have the same high accuracy value of 93.62%. The AUC value for a split ratio of 90:10, however, is 0.773 greater than the split ratio of 80:20, which is 0.655.

TABLE 7
Naïve Bayes confusion matrix with PSO

Classification	Predicted Class	
	Class = F	Class = T
Class = F	43	3
Class = T	0	1

From the results of the confusion matrix in TABLE 5, the accuracy value is 93.62%

FIGURE 3. Naïve Bayes ROC Curve with PSO On Split Validation 90:10



In testing the Naïve Bayes method with Particle Swarm Optimization, the ROC curve was also obtained, as shown in FIGURE 3 above, with an AUC value of 0.773 and categorized as a Fair Classification because it is in the 0.70-0.80.

IV. DISCUSSION

Two experiments were run for the investigation, one without feature selection and the other using the Naive Bayes algorithm combined with the Particle Swarm Optimization feature selection technique. split validation was done using three distinct ratios. Three ratios—70:30, 80:20, and 90:10—are used in the thoracic surgery dataset to divide the training and test data. The accuracy value displays the signal for understanding the ideal outcomes for each experiment. experiments with classification using the Naive Bayes

algorithm on datasets related to thoracic surgery. Additionally, to enhance classification performance in relation to the postoperative life expectancy of lung cancer patients, where mortality occurs within one year after surgery, feature selection using particle swarm optimization was done in this work on features in the thoracic surgery dataset. An initialization experiment was performed in PSO with up to 10 particles, with a 30-iteration limit. The performance of the model in categorizing the dataset for thoracic surgery will also be appropriately assessed based on the confusion matrix and AUC findings. The acquired model performance is used to contrast the basic Nave Bayes model with the Nave Bayes model that incorporates PSO.

In Naïve Bayes, evaluating the accuracy of the thoracic surgery dataset obtained the best accuracy value: split validation with a ratio of 80% training data and 20% testing data with an accuracy value of 81.91% and an AUC value of 0.642. In the Naïve Bayes model with PSO feature selection, the best accuracy value is obtained in split validation with a ratio of 90:10 and 80:20 with an accuracy value of 93.62% using ten particles, and the best AUC value is obtained in split validation with 90% training data and 10% data testing that is equal to 0.773. Based on the general guidelines for the classification of AUC values, the results of AUC evaluation in thoracic surgery datasets are included in the appropriate classification. So in naïve Bayes with PSO, the best accuracy and AUC value are in split validation with a ratio of 90:10.

The weighting results for each attribute are obtained from the experimental results of the Naïve Bayes method and PSO feature selection. In each split validation ratio, several attributes have a zero weight, which means that these attributes do not influence the research being conducted. In the Naïve Bayes and PSO models in Split validation 70:30, nine attributes have zero weight, namely the attributes DGN, PRE5, PRE6, PRE7, PRE8, PRE9, PRE14, PRE17, and PRE32. In the Naïve Bayes and PSO models in Split validation 80:20, six attributes have zero weight: DGN, PRE7, PRE11, PRE14, PRE17, and PRE19. Whereas in the Naïve Bayes and PSO models in Split validation 90:10, three attributes have zero weight, namely PRE6, PRE14, and PRE17.

The Naive Bayes test results improved following feature selection and weighting using Particle Swarm Optimization, as seen in FIGURE 4. It may be inferred that Particle Swarm Optimization can improve the classification value for the classification of lung cancer patients' life expectancy after thoracic surgery. From the interpretation of the research results presented, it can be seen that the implementation of Particle Swarm Optimization (PSO) in the Naïve Bayes method for thoracic surgery classification results in a significant increase in accuracy. In the Naïve Bayes experiment without feature selection, the best validation was achieved at an 80:20 split ratio with an accuracy of 81.91%. However, by using PSO for feature selection, the accuracy increased to 93.62% in separate validation with split ratios of 80:20 and 90:10. This indicates that PSO contributed positively to improving classification quality in the context of thoracic surgery. The selection of features using PSO also affected the relevant and important attributes in the classification process. Some attributes had zero weight in separate validation, indicating that these attributes did not affect the research results. Therefore, implementing PSO helped identify the most relevant features and reduce data dimensionality, speeding up the training process and producing a more efficient model.

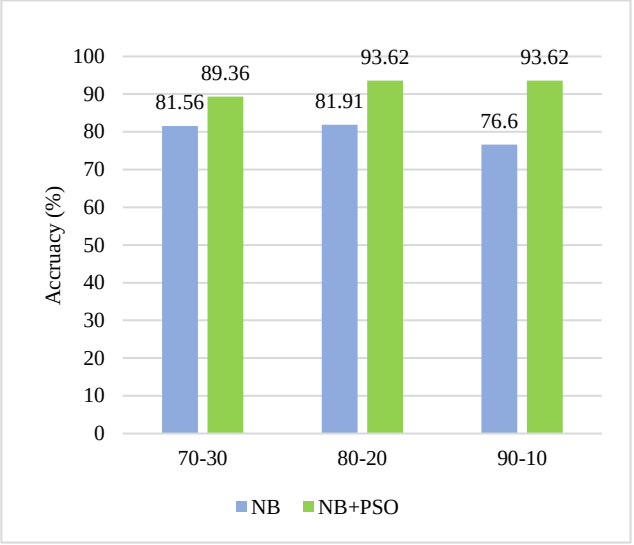


FIGURE 4. Comparison of Accuracy Values in Thoracic Surgery Datasets with difference training and testing percentage

The AUC (Area Under Curve) value was also considered in this research. In the Naïve Bayes experiment without PSO, the best AUC value achieved was 0.642, indicating poor classification. Meanwhile, in the PSO experiment, the best AUC value increased to 0.773, categorized as fair classification. Therefore, the application of PSO increased the accuracy and improved the classification quality based on the obtained AUC value. Overall, the research results showed that integrating Particle Swarm Optimization with the Naïve Bayes method for thoracic surgery classification significantly improved accuracy and classification quality. This confirms that this approach effectively enhances the model's performance in identifying the survival rate of lung cancer patients after thoracic surgery. It can be concluded that Particle Swarm Optimization plays a crucial role as a feature selection technique in improving the efficiency and effectiveness of the Naïve Bayes classification model.

Then, comparing the results of this research with previous studies showed that the Naïve Bayes method combined with PSO feature selection performed better in classifying thoracic surgery datasets. This is because the combined methods increased the accuracy and AUC values compared to previous studies that used different classification methods or without PSO feature selection. The research results [12], despite also using different classification and feature selection methods, did not achieve the accuracy and AUC values comparable to the results of this study. This study showed that using Naïve Bayes with PSO feature selection could enhance performance in classifying thoracic surgery datasets due to the increased accuracy and AUC values. This indicates that combining methods and optimization effectively improved the quality of classification of survival rates of lung cancer patients post-thoracic surgery. Besides, this comparison also helps understand the potential of the methods and algorithms used in this research to produce better results than previous approaches.

However, the limitation of this study was the use of a relatively limited dataset in terms of the number of patients and available features. This could affect the generalization of the results obtained. Therefore, for future research, to get more accurate and extensive results, it is recommended to use a larger and more diverse dataset that encompasses more patients and features relevant to thoracic surgery. Nevertheless, despite these limitations, this research successfully demonstrated that combining Naïve Bayes and Particle Swarm Optimization could improve classification performance in thoracic surgery.

Therefore, this research implies using Particle Swarm Optimization in feature selection can enhance the classification accuracy of survival rates of lung cancer patients post-thoracic surgery using the Naïve Bayes method. This indicates that this combined method could effectively improve predictive performance in clinical systems, thus helping doctors make better and more accurate decisions regarding patient care after undergoing thoracic surgery. With the increased accuracy of the predictive model based on the medical dataset, these research results could also contribute significantly to developing artificial intelligence technology in the medical field to improve the quality of health services.

V. CONCLUSION

Based on this research, the Naive Bayes classification algorithm and the Particle Swarm Optimization feature selection method have been applied to the thoracic surgery dataset. Optimal weight using particle swarm optimization increased accuracy in thoracic operations. The test results using the Naive Bayes algorithm and the thoracic surgery dataset showed a maximum accuracy of 81.91% at a ratio of 80:20, with an AUC value of 0.620. Particle Swarm Optimization was used to enhance feature selection for attribute weighting, and the highest accuracy value was 93.62% with an AUC value of 0.773 at a ratio of 90:10, where three attributes—specifically, PRE6, PRE14, and PRE17—had zero weights. The accuracy value in thoracic surgery using Naïve Bayes increased due to attribute weighting in feature selection using Particle Swarm Optimization. The thoracic surgery dataset might have a higher accuracy value when using the Naïve Bayes and Particle Swarm Optimization approach than just the Naïve Bayes classification method. The accuracy of the Naive Bayes technique on the dataset from thoracic operations can be improved by using particle swarm optimization (PSO).

Therefore, these research findings proved that the combination of the Naïve Bayes classification algorithm and the Particle Swarm Optimization feature selection method successfully improved accuracy in classifying thoracic surgery data. However, to further optimize the performance of this method, future research should focus on several important aspects. One aspect that needs attention in further research is the use of larger and more diverse datasets. Then, combining other feature selection methods with Particle Swarm Optimization is also a good step to take in future research. This could help find the best features that significantly contribute to enhancing the accuracy of thoracic surgery classification. Lastly, future research should also focus on evaluating the performance of the proposed algorithm and method. Using diverse and comprehensive evaluation metrics will ensure that the improvement in classification accuracy results from applying the correct method. Thus, through improvements in these aspects, future research is expected to achieve more accurate and extensive results in classifying thoracic surgery data using a combination of Naïve Bayes and Particle Swarm Optimization.

REFERENCES

[1] E. Marret *et al.*, "Protective ventilation during anaesthesia reduces major postoperative complications after lung cancer surgery: A double-blind randomised controlled trial," *Eur. J. Anaesthesiol.*, vol. 35, no. 10, pp. 727–735, 2018, doi: 10.1097/EJA.0000000000000804.

[2] Y. Zhang, F. Fu, and H. Chen, "Management of Ground-Glass Opacities in the Lung Cancer Spectrum," *Ann. Thorac. Surg.*, vol. 110, no. 6, pp. 1796–1804, 2020, doi: 10.1016/j.athoracsur.2020.04.094.

[3] A. Bondzi-Simpson *et al.*, "Ethiopia's first minimally invasive surgery program: A novel approach in global surgical education," *JTCVS Open*, vol. 13, no. C, pp. 459–467, 2023, doi: 10.1016/j.xjon.2022.11.015.

[4] X. Li, Y. Liu, Y. Zhou, Y. Gao, C. Duan, and C. Zhang, "Day surgery unit robotics thoracic surgery: feasibility and management," *J. Cancer Res. Clin. Oncol.*, vol. 1, no. 1, pp. 1–6, 2023, doi: 10.1007/s00432-023-04731-0.

[5] E. Williams and J. Agzarian, "A narrative review of traumatic mediastinal injuries and their management: The thoracic surgeon perspective," *Mediastinum*, vol. 5, no. 3, pp. 1–9, 2021, doi: 10.21037/MED-21-13.

[6] M. A. Nematollahi *et al.*, "Body composition predicts hypertension using machine learning methods: a cohort study," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-34127-6.

[7] B. Shen, G. Coruzzi, and D. Shasha, "EnsInfer: a simple ensemble approach to network inference outperforms any single method," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–13, 2023, doi: 10.1186/s12859-023-05231-1.

[8] Y. J. Lee, S. H. O, and J. E. Eck, "Improving Recidivism Forecasting With a Relaxed Naïve Bayes Classifier," *Crime Delinq.*, vol. 1, no. 1, pp. 1–29, 2023, doi: 10.1177/00111287231186093.

[9] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
[10] M. Ismail, N. Hassan, and S. S. Bafjaish, "Journal of Soft Computing and Data Mining Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task," *J. Soft Comput. Data Min.*, vol. 1, no. 2, pp. 1–10, 2020, [Online]. Available: http://penerbit.uthm.edu.my/ojs/index.php/jscdm
[11] J. P. M. Miguel, L. A. Neves, A. S. Martins, M. Z. do Nascimento, and T. A. A. Tosta, "Analysis of neural networks trained with evolutionary algorithms for the classification of breast cancer histological images," *Expert Syst. Appl.*, vol. 231, no. 30, p. 120609, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120609.
[12] Roshan S and Rohini V, "Prediction of Post-Surgical Survival of Lung Cancer Patients After Thoracic Surgery Using Data Mining Techniques.," *Int. J. Adv. Res.*, vol. 5, no. 4, pp. 596–600, 2017, doi: 10.21474/ijar01/3852.
[13] W. Kanyongo and A. E. Ezugwu, "Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives," *Informatics Med. Unlocked*, vol. 38, no. March, p. 101232, 2023, doi: 10.1016/j.imu.2023.101232.
[14] P. Geetha Pavani, B. Biswal, M. V. S. Sairam, and N. Bala Subrahmanyam, "A semantic contour based segmentation of lungs from chest x-rays for the classification of tuberculosis using Naïve Bayes classifier," *Int. J. Imaging Syst. Technol.*, vol. 31, no. 4, pp. 2189–2203, 2021, doi: 10.1002/ima.22556.
[15] M. Qois Syafi, "Increasing Accuracy of Heart Disease Classification on C4.5 Algorithm Based on Information Gain Ratio and Particle Swarm Optimization Using Adaboost Ensemble," *J. Adv. Inf. Syst. Technol.*, vol. 4, no. 1, pp. 100–112, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jaist
[16] A. Mustopa, Hermanto, Anna, E. B. Pratama, A. Hendini, and D. Risdiansyah, "Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine and Naive Bayes Algorithm Based on Particle Swarm Optimization," in *International Conference on Informatics and Computing (ICIC)*, Gorontalo, 2020, pp. 1–7. doi: 10.1109/ICIC50835.2020.9288655.
[17] Y. J. Zhang, H. Zhang, and R. Gupta, "A new hybrid method with data - characteristic - driven analysis for artificial intelligence and robotics index return forecasting," *Financ. Innov.*, vol. 10, no. 4, pp. 1–23, 2023, doi: 10.1186/s40854-023-00483-5.
[18] D. Valero-carreras, J. Alcaraz, and M. Landete, "Computers and Operations Research Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, no. December 2022, p. 106131, 2023, doi: 10.1016/j.cor.2022.106131.
[19] S. S. Zakariaee, A. I. Abdi, N. Naderi, and M. Babashahi, "Prognostic significance of chest CT severity score in mortality prediction of COVID - 19 patients , a machine learning study," *Egypt. J. Radiol. Nucl. Med.*, 2023, doi: 10.1186/s43055-023-01022-z.
[20] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. Heidelberg: Springer Berlin, 2011. [Online]. Available: https://doi.org/10.1007/978-3-642-19721-5
[21] A. T. Tunkiel, D. Sui, and T. Wiktorski, "Impact of data pre-processing techniques on recurrent neural network performance in context of real-time drilling logs in an automated prediction framework," *J. Pet. Sci. Eng.*, vol. 208, no. 1, p. 109760, 2022, doi: 10.1016/j.petrol.2021.109760.
[22] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.
[23] P. Mishra *et al.*, "MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing," *Chemom. Intell. Lab. Syst.*, vol. 205, no. August, p. 104139, 2020, doi: 10.1016/j.chemolab.2020.104139.

BIOGRAPHY

Shalehah is an undergraduate student in the Department of Computer Science, Lambung Mangkurat University. Her research interest is sentred on Data Mining.

Muhammad Itqan Mazdadi is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking.

Andi Farmadi is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science.

Dwi Kartini is a lecturer in Department of Computer Science, Lambung Mangkurat University. Her research interest is centered on Data Science.

Muliadi is a lecturer in Department of Computer Science, Lambung Mangkurat University. His highest education is S2, research focus on Artificial Intelligence, Decision Support System, Data Science. Relevant skills Data Science Start-up Business Digital Entrepreneurship Academy Data