Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 3 Iss. 1 July 2025, pp: 70-79
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Lung Disease Risk Prediction Using Machine Learning Algorithms

Ananda Putri Aulia[1*], Qaula Adelia[2], Haykal Alya Mubarak[3], Mohd. Adzka Fatan[4], Sudarno[5]

[1,2,3]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]Department of Dirasat Islamiyah, Faculty of Al Qur'an Al Karim,
University of the Holy Quran and Islamic Sciences, Mukalla, Yamen
[5]Department of 'Ulumul Qur'an, Faculty of Al Qur'an Al Karim,
University of the Holy Quran and Islamic Sciences, Mukalla, Yamen

E-Mail: [1]12250321407@students.uin-suska.ac.id, [2]12250324177@students.uin-suska.ac.id,
[3]12150313837@students.uin-suska.ac.id, [4]akafatan05@gmail.com, [5]soedarno891@gmail.com

Received Dec 27th 2024; Revised May 30th 2025; Accepted Jun 09th 2025; Available Online Jul 06th 2025, Published Jul 31th 2025
Corresponding Author: Ananda Putri Aulia

Abstract

Lung diseases, including lung cancer, are one of the leading causes of death in the world. Early detection is essential to increase patients' chances of recovery and reduce healthcare costs. The utilization of machine learning algorithms can be used to solve this problem. This study evaluates five machine learning algorithms, namely K-Nearest Neighbors (K-NN), Naïve Bayes Classifier (NBC), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), for lung disease prediction using a dataset of 30,000 data with 11 attributes from Kaggle. The dataset was processed through data preprocessing and divided into training and test data with a ratio of 70%:30% and 80%:20%. The algorithm performance was evaluated using precision, recall, F1-score, and accuracy metrics. The results show that RF, SVM, and DT algorithms have the highest performance, with accuracy reaching 94.72% at 70%:30% ratio. The DT algorithm, which previously showed low performance in heart disease classification, provided competitive results in lung disease prediction. This research focuses on the importance of proper algorithm selection and data organization to improve the effectiveness of disease prediction. The findings contribute to the development of artificial intelligence technology for medical applications, particularly in supporting early diagnosis of lung diseases.

Keyword: Classification, Decision Tree, Lung Diseases, Machine Learning, Prediction

1. INTRODUCTION

Lung diseases include a wide range of disorders that affect the function of the respiratory system and are one of the leading causes of morbidity and mortality worldwide, including in Indonesia [1]. One of the most dangerous forms of lung disease is lung cancer, which results from the growth of cancer cells in the respiratory tract. The disease is often difficult to detect in its early stages, leading to delays in diagnosis and treatment that can reduce a patient's chances of recovery. Therefore, prediction and early detection of lung diseases, including lung cancer, is crucial to improve treatment outcomes and reduce healthcare costs [2]. However, the traditional diagnostic methods often face challenges, such as treatment delays and potential misdiagnosis, which can hinder treatment and increase healthcare costs.

The use of Artificial Intelligence (AI) technology, particularly Machine Learning (ML) algorithms, provides a promising solution to these challenges [3]. Machine Learning algorithms, such as K-Nearest Neighbors (K-NN), Naïve Bayes Classifier (NBC), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), are proven to have a great potential in disease prediction based on historical patient health data [4]. Each algorithm has specific strengths and weaknesses when it comes to analyzing complex health datasets.

The selection of these five algorithms is based on their popularity, ease of implementation, and proven effectiveness in medical prediction tasks, especially in binary classification problems. These algorithms represent a diverse group of models: instance-based (K-NN), probabilistic (NBC), tree-based (DT, RF), and margin-based (SVM), allowing comprehensive comparison. The K-NN algorithm is often chosen because it can produce accurate classifications when applied to large datasets [5]. Research conducted by Assegie, 2021 [6] found that K-NN is very effective in detecting heart disease with an accuracy rate of 91.99%. Another

study by Salama Abd Elminaam et al. 2023 [7] suggested using the RF algorithm to predict heart disease. This study found that the Random Forest algorithm obtained a higher accuracy rate of 98.6%, compared to other algorithms namely, K-NN, NBC, and DT algorithms.

Based on previous research, this study aims to compare five classification algorithms, K-NN, NBC, SVM, DT, and RF, on a lung disease dataset. Some previous studies [6], [7], [8], [9] showed that RF, SVM, K-NN, and NBC algorithms have superior performance in heart disease classification. In contrast, research by Arifuddin et al. 2024 [10] showed that DT tends to have lower performance than other algorithms. The novelty of this research is to re-examine the performance of these algorithms, including the DT, but with a focus on lung disease prediction. This study aims to evaluate whether algorithms that have previously shown superior performance in heart disease classification can also provide good results in lung disease prediction. In addition, this research will determine if the Decision Tree remains the lowest-performing algorithm or if it shows different results in the case of lung disease prediction.

2. MATERIAL AND METHOD

This paper adopts the methodology shown in Figure 1.

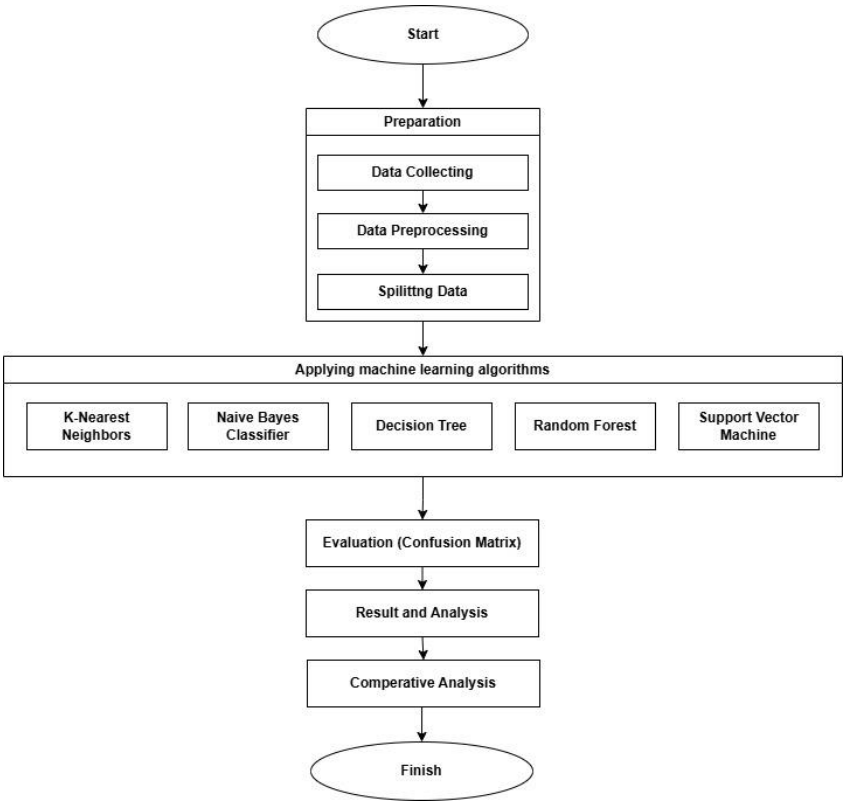

Figure 1. Research Methodology

At the early stage, a literature review was conducted to identify relevant literature discussing lung diseases and implementation of machine learning algorithms such as K-Nearest Neighbors (K-NN), Naive Bayes Classifier (NBC), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). In the preparation stage, several important steps are taken to prepare the dataset to be used. In the Data Collecting step, a dataset consisting of 30,000 data with 11 attributes was collected from the Kaggle platform as a credible data source. After the data is collected, Data Preprocessing is performed, which involves transforming data from text to numerical form. Next, the dataset is divided at the Data Splitting stage into training data and testing data with a ratio of 70%:30% and 80%:20%.The next stage is Applying Machine Learning Algorithms, where several machine learning algorithms are applied. The algorithms used include K-NN, NBC, DT, RF, and SVM. In the next stage, the performance is evaluated at the Evaluation stage using a confusion matrix. This matrix provides an evaluation metric such as accuracy, precision, recall, and F1-score, which is used to assess the performance of each algorithm. The results of this evaluation are then thoroughly analyzed at the Result and Analysis stage to determine which algorithm is most effective in predicting lung disease. Finally, Comparative Analysis is conducted to compare the performance of the various algorithms used, so that the best algorithm that has the highest accuracy and efficiency in lung disease prediction can be determined.

2.1. Lung Disease

Lung diseases cover a wide range of disorders that affect lung function, including obstructive diseases such as COPD (Chronic Obstructive Pulmonary Disease) and asthma, as well as restrictive diseases such as pulmonary fibrosis [11]. The disease is often characterized by difficulty breathing, decreased lung function, and lung tissue damage [12]. Data from the World Health Organization (WHO) shows that Chronic Obstructive Pulmonary Disease (COPD) is one of the leading causes of death worldwide. In 2019, COPD became the third leading cause of death globally, accounting for more than 3.23 million deaths. The main risk factors include smoking, exposure to air pollution, and recurrent respiratory infections from childhood [13]. In addition, the condition is often associated with comorbidities such as cardiovascular disease and lung cancer. Chronic Obstructive Pulmonary Disease (COPD) causes impaired lung function due to narrowing of the airways, chronic inflammation, or damage to the alveolus. Diagnosis of COPD is often done through spirometry, which measures airflow and lung capacity [14], [15].

2.2. Preparation

The dataset used in this research consists of 30,000 data with 11 attributes, which are taken from the Kaggle platform as one of the trusted data sources for scientific research. The attributes in this dataset include No, Age, Gender, Smoking, Work, Household, Activity_Labor, Activity_Exercise, Insurance, Disease_Congenital, and Result. However, the No and Insurance attributes are not included in the analysis process because they are considered less relevant in influencing the results of predicting patient health. After the data was compiled, a data preprocessing stage was performed to improve the quality and consistency of the data. Data that is in text form is converted into numeric form. After going through the data preprocessing stage, the dataset consisting of 30,000 data with 11 attributes is divided into two subsets. In this study, two data division scenarios were conducted. In the first scenario, the dataset is divided with a ratio of 70%:30%, where 70% of the data is used to train the algorithm and the remaining 30% is used to test the performance of the algorithm. The second scenario uses a ratio of 80%:20%, where 80% of the data is used as training data and 20% as test data.

2.3. K-Nearest Neighbors (K-NN)

The K-NN algorithm is a nearest neighbor algorithm. The algorithm is calculated from the distance value of the test data and the training data, from the smallest nearest neighbor value [6], [7]. The advantage of the KNN algorithm is that it can produce good classification when using large amounts of data [8]. Research by [9] found that KNN is very effective in detecting Parkinson's disease with an accuracy rate of 97.22%. The stages of the K-NN algorithm begin with determining the value of the K parameter, which is the number of nearest neighbors to be used in the classification process. After the K value is determined, the next step is to calculate the distance between the training data and the test data. The most commonly used distance calculation in the K-NN algorithm is the Euclidean distance, which measures the shortest linear distance between two points in multidimensional space. This distance is calculated for each training data against the test data, and the results are used to determine the K most relevant nearest neighbors [16]. The K-NN algorithm can be shown in equation 1 [17].

$$\mathrm{P(X, Cj)} = \sum \mathrm{Cosine(x, di).\, y(yid\,, Cj)} \tag{1}$$

y_i represents the test data to be classified, while C_j is the class or category of the nearest neighbor belonging to the class set. This function works by calculating the distance between the test data yi and a number of nearest training data using a distance metric such as Euclidean Distance. Once the nearest neighbors are identified, the function $y(y_i,C_j)$ will determine the class Cj that occurs most frequently among those neighbors and assign it as the prediction result for yi, thus allowing the K-NN algorithm to perform classification based on the fit of the data in a multidimensional space.

2.4. Naïve Bayes Classifier (NBC)

NBC algorithm is based on Bayes' theorem [18], which calculates the probability that each class will anticipate the data correctly [19]. Research by [20] found that NBC obtained an accuracy rate of 74.19% and an average error rate of 25.81%, higher than the K-NN algorithm which has an average value of 54.84% and an average error rate of 45.18% in the classification of Hepatitis disease. While based on research conducted by [21] found that the The NBC algorithm is able to function or work well for garbage classification in the design of a garbage classifier using Arduino. The NBC algorithm equation can be shown in equation 2 [16].

$$\mathrm{P(B)} = \mathrm{P(A)P(A)/P(B)} \tag{2}$$

P(B) is the probability of an event or class B occurring, better known as the prior probability for class B. It describes the likelihood of class B occurring without taking into account other conditions or features in the data. P(A) is the probability of event A, better known as the prior probability for feature A, which describes the likelihood of a feature or attribute occurring, regardless of class. P(B) is the conditional probability.

2.5. Decision Tree (DT)

DT is a data mining technique used to classify objects by dividing data into sets based on input variables, thus forming a hierarchical tree [22]. The main objective is to build a training algorithm that can be applied to predict the class or value of a target variable through learning decision rules inferred from training data. DT algorithms can be used to solve regression and classification problems. Research conducted by [23] found that the classification and identification of pests, diseases, and nutrient deficiencies in rice plants using the DT algorithm obtained an accuracy of 97.88%. Based on research conducted by [24] the DT algorithm can build a neural network topology. The general equation of the DT algorithm can be shown in equation 3 [22].

$$\text{Entropy (S)} = -\sum_{i=1}^{n} \text{Pi} \log_2 (\text{Pi}) \tag{3}$$

In a DT, S refers to the set of all cases or data used in training the algorithm. n is the number of classes in the classification problem to be predicted, while pi indicates the number of samples corresponding to class I, reflecting the frequency or amount of data in each class. The DT algorithm uses this information to divide the data based on its attributes, with the aim of separating the data into the correct classes by selecting the attributes that are most effective in separating the classes.

2.6. Random Forest (RF)

The RF algorithm is one of the most widely used machine learning algorithms, especially for analyzing remote sensing data [25], [26]. Research conducted by [27] found that the classification of medical treatment and immunotherapy using a RF algorithm is better than using a decision tree algorithm. Another study conducted by [28] found that the RF algorithm can be used to quickly assess satellite image-based bathymetry grids. The RF algorithm equation can be shown in equation 4 [29].

$$\text{P(C|X)} = \frac{\text{p(C|X)}.\,\text{p(C)}}{\text{P(X)}} \tag{4}$$

2.7. Support Vector Machine (SVM)

SVM algorithm is a method that utilizes data analysis and identifies patterns in classification. SVM has the advantage of being able to identify various hyperplanes that define the boundaries between two different classes [30]. SVM accuracy results are influenced by the parameters and kernel functions used [31]. Research by [32] found that SVM is very effective in the classification of heart disease, testing the SVM algorithm with normalization has better accuracy results compared to the K-NN algorithm either with or without normalization. Research that also uses the SVM algorithm was conducted by [33] which discusses location classification methodology using support vector machines, finding that SVM-based location classification algorithms that use strong motion data can provide more reliable classification results. The SVM algorithm equation can be shown in equation 5 [30].

$$\text{f(x)} = \text{W}^{\text{t}}\phi(\text{x}) + \text{b} \tag{5}$$

3. RESULTS AND ANALYSIS

In this part, the performance of five classification algorithms, namely K-Nearest Neighbors (K-NN), Naive Bayes Classifier, Decision Tree, Random Forest, and Support Vector Machine (SVM), is evaluated to predict lung diseases using precision, recall, F1 score, and accuracy metrics. The evaluation was conducted on two data splits, 80%:20% and 70%:30%, to see the effect of data splits on algorithm performance. These ratios are commonly used in machine learning research because they provide a balanced trade-off between training and testing data. The 80:20 split allows the model to learn from a larger portion of the dataset, which is useful for capturing general patterns, while the 70:30 split provides a larger test set, helping to better evaluate the model's generalization ability on unseen data. Precision measures the accuracy of positive predictions, recall indicates the algorithm's ability to detect all positive cases, F1 score combines precision and recall, while accuracy describes the overall proportion of correct predictions. The results of this evaluation are used to compare the effectiveness of each algorithm in predicting lung disease on the dataset used.

3.1. K-Nearest Neighbors (K-NN)

Table 1 presents the evaluation results of applying the K-NN algorithm.

Table 1. K-NN Performance Evaluation

Split Data	Precision	Recall	F1-Score
80:20	100%	88.24%	93.75%
70:30	86.88%	86.82%	86.85%

The Table 1 shows the performance evaluation of the K-NN algorithm based on Precision, Recall, and F1-Score metrics for two data splits, namely 80:20 and 70:30. At 80:20 split, the algorithm achieved 100% Precision, 82.24% Recall, and 93.75% F1-Score, which shows an excellent balance of performance. Meanwhile, at the 70:30 split, the algorithm obtained Precision 86.88%, Recall 86.82%, and F1-Score 86.85%, with slightly lower performance than 80:20. These results indicate that the 80:20 split provides more optimal performance due to more training data. In Figure 2, the accuracy result of the K-NN algorithm with an 80:20 split obtained an accuracy of 94.27% higher than the 70:30 split which only obtained an accuracy of 87.17%.

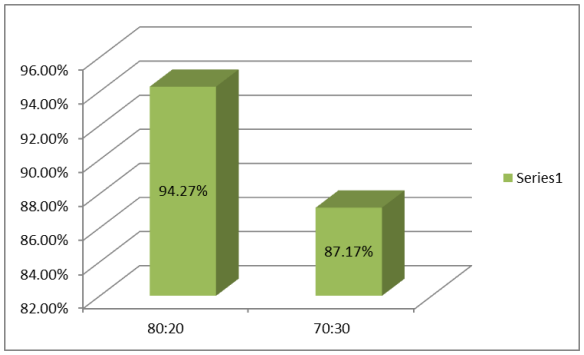

Figure 2. Accuracy K-NN

3.2. Naïve Bayes Classifier (NBC)

Table 2. presents the evaluation results of the application of the NBC algorithm.

Table 2. NBC Performance Evaluation

Split Data	Precision	Recall	F1-Score
80:20	100%	88.24%	93.75%
70:30	86.88%	86.82%	86.85%

The Table 2 shows the performance evaluation of the NBC algorithm based on Precision, Recall, and F1-Score metrics for two data splits, namely 80:20 and 70:30. At 80:20 split, the algorithm achieved 100% Precision, 88.24% Recall, and 93.75% F1-Score, which shows an excellent balance of performance. Meanwhile, at the 70:30 split, the algorithm obtained Precision 86.88%, Recall 86.82%, and F1-Score 86.85%, with slightly lower performance than 80:20. These results indicate that the 80:20 split provides more optimal performance due to more training data. In Figure 3, the accuracy result of the NBC algorithm with an 80:20 split obtained an accuracy of 94.27%, which is higher than the 70:30 split that only obtained an accuracy of 87.17%.

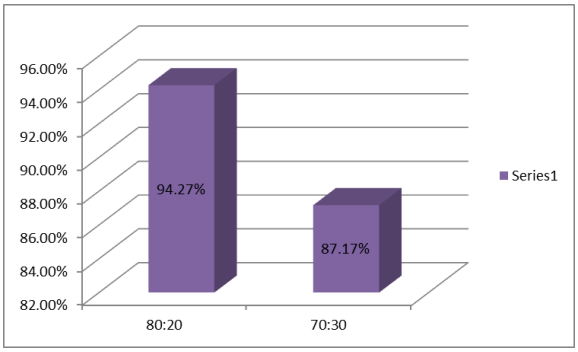

Figure 3. Accuracy Naïve Bayes Classifier

3.3. Decision Tree (DT)

Table 3. present the results of the evaluation of the application of the DT algorithm.

Table 3. DT Performance Evaluation

Split Data	Precision	Recall	F1-Score
80:20	94.97%	94.12%	94.23%
70:30	95.33%	94.59%	94.69%

The Table 3 shows the performance evaluation of the DT algorithm based on Precision, Recall, and F1-Score metrics for two data splits, namely 80:20 and 70:30. At 80:20 split, the algorithm achieved Precision 94.97%, Recall 94.12%, and F1-Score 94.23%, which shows a good balance of performance. At the 70:30 split, the algorithm performed slightly higher with a Precision of 95.33%, Recall of 94.59%, and F1-Score 94.69%. These results indicate that the 70:30 split provides slightly more optimal performance than 80:20. In Figure 4, the accuracy result of the DT algorithm with 80:20 split obtained an accuracy of 94.27% which is slightly lower than the 70:30 split which obtained an accuracy of 94.72%.

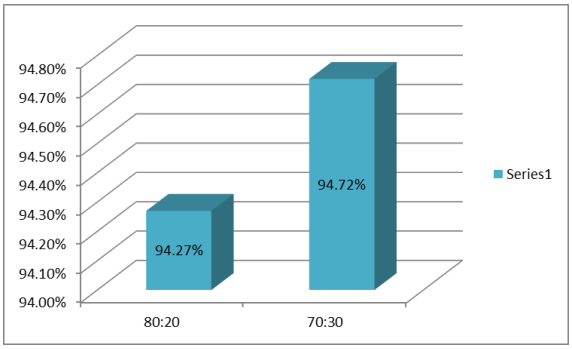

Figure 4. Accuracy Decision Tree

3.4. Random Forest (RF)

Table 4 presents the evaluation results of applying the RF algorithm.

Table 4. RF Performance Evaluation

Split Data	Precision	Recall	F1-Score
80:20	100%	88.24%	93.75%
70:30	100%	89.19%	94.29%

The Table 4 shows the performance evaluation of the RF algorithm based on Precision, Recall, and F1-Score metrics for two data splits, namely 80:20 and 70:30. In both scenarios, the algorithm achieves 100% Precision, indicating that all positive predictions made by the algorithm are actually relevant. At the 80:20 split, the algorithm's Recall was 88.24%, while at the 70:30 split, it increased to 89.19%. This resulted in an F1-Score of 93.75% for the 80:20 split and a slightly higher 94.29% for the 70:30 split. These results show that while the Precision performance remains perfect, the 70:30 split provides slightly better performance on Recall and F1-Score. In Figure 5, the accuracy result of the RF algorithm with an 80:20 split obtained an accuracy of 94.27% which is slightly lower than the 70:30 split which obtained an accuracy of 94.72%.

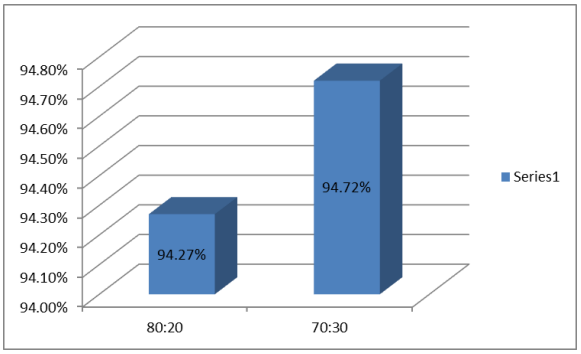

Figure 5. Accuracy Random Forest

3.5. Support Vector Machine (SVM)

Table 5 presents the evaluation results of applying the SVM algorithm.

Table 5. SVM Performance Evaluation

Split Data	Precision	Recall	F1-Score
80:20	100%	88.24%	93.75%
70:30	100%	89.19%	94.29%

The Table 5 shows the performance evaluation of the SVM algorithm based on Precision, Recall, and F1-Score metrics for two data splits, namely 80:20 and 70:30. In both scenarios, the algorithm achieves 100% Precision, indicating that all positive predictions made by the algorithm are relevant. At the 80:20 split, the algorithm's Recall was 88.24%, while at the 70:30 split, it increased to 89.19%. This resulted in an F1-Score of 93.75% for the 80:20 split and a slightly higher 94.29% for the 70:30 split. These results show that while the Precision performance remains perfect, the 70:30 split provides slightly better performance on Recall and F1-Score. In Figure 6, the accuracy result of the SVM algorithm with an 80:20 split obtained an accuracy of 94.27% which is slightly lower than the 70:30 split which obtained an accuracy of 94.72%.

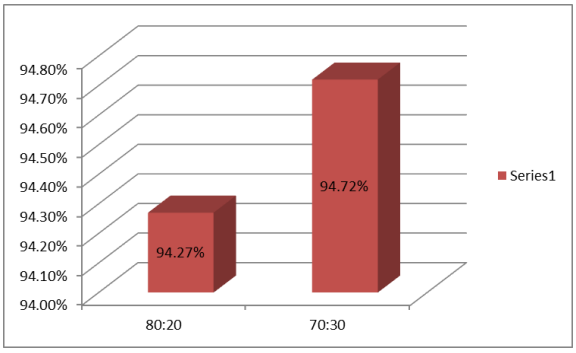

Figure 6. Accuracy Support Vector Machine

To identify the performance of the lung disease prediction algorithm, the confusion matrix is used, which is a commonly used tool in binary classification. The dataset used consists of 30,000 data, which is divided into two ratios of 80%:20% and 70%:30%, to assess how the different data splits affect the performance of the algorithm. The confusion matrix gives an overview of the algorithm's ability to classify the data by showing true positive, true negative, false positive and false negative.

3.6. Confusion Matrix

The evaluation of five classification algorithms K-NN, NBC, DT, RF, and SVM was conducted using a dataset of 30,000 entries with two data split scenarios (80:20 and 70:30). The performance of each algorithm was summarized through confusion matrices, which provide insight into the number of correctly classified instances (True Negatives and True Positives) and misclassified cases (False Positives and False Negatives). Presenting these results in a comparative form allows for a clearer understanding of the strengths and weaknesses of each algorithm. Figures 7 and 8 illustrate the confusion matrix visualisation of the K-NN algorithm, while the other algorithms are not shown. Overall, the confusion matrix results of this study are shown in Table 6.

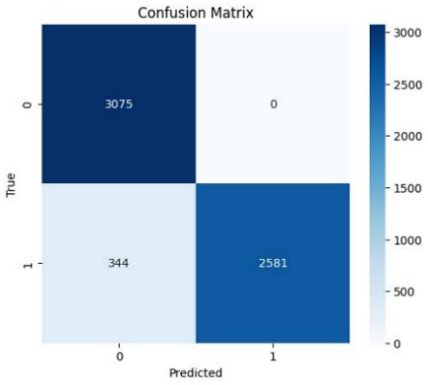

Figure 7. Confusion Matrix 80:20

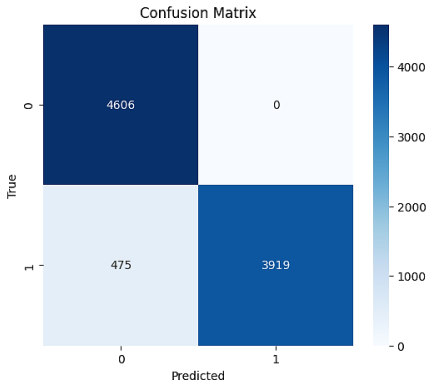

Figure 8. Confusion Matrix 70:30

Table 6. Confusion Matrix

Algoritma	Split Data	TN	TP	FP	FN
K-NN	80:20:00	3,075	2,581	0	344
	70:30:00	4,606	3,919	0	475
NBC	80:20:00	2,705	2,545	370	380
	70:30:00	4,03	3,815	576	579
DT	80:20:00	3,075	2,581	0	344
	70:30:00	4,606	3,919	0	475
RF	80:20:00	3,075	2,581	0	344
	70:30:00	4,606	3,919	0	475
SVM	80:20:00	3,075	2,581	0	344
	70:30:00	4,606	3,919	0	475

The results reveal that K-NN, DT, RF, and SVM performed in a highly consistent manner across both split scenarios. These algorithms produced high numbers of True Negatives (TN) and True Positives (TP) while recording no False Positives (FP). This consistency demonstrates their strong ability to correctly identify negative instances without misclassifying them as positive. However, all four algorithms still experienced a notable number of False Negatives (FN), indicating a limitation in accurately detecting all positive cases.

In contrast, the Naïve Bayes Classifier (NBC) displayed a different performance pattern. Compared to the other algorithms, NBC generated substantially higher numbers of both False Positives and False Negatives. For example, in the 80:20 split, NBC recorded 370 FP and 380 FN, while in the 70:30 split, these errors increased to 576 FP and 579 FN. This suggests that although NBC is able to capture a considerable portion of the correct classifications, its overall reliability in distinguishing between the two classes is weaker, leading to higher misclassification rates.

When comparing the two data split scenarios, all algorithms showed an increase in False Negatives when the proportion of test data was larger (70:30). This is expected since a larger test set introduces more data variability, making classification more challenging. Nevertheless, K-NN, DT, RF, and SVM maintained relatively stable and consistent results, whereas NBC was more sensitive to the change in test size. Overall, the findings highlight that K-NN, DT, RF, and SVM outperform NBC in terms of stability and error reduction, particularly in avoiding False Positives, though reducing False Negatives remains a common challenge across all models.

3.7. Comparative Analysis

A comparative analysis was conducted to comparing the performance of five classification algorithms, namely K-Nearest Neighbors (K-NN), Naive Bayes Classifier (NBC), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). This analysis includes accuracy evaluation on two data sharing scenarios: 80:20 and 70:30. The comparison results are summarized in Figure 9.

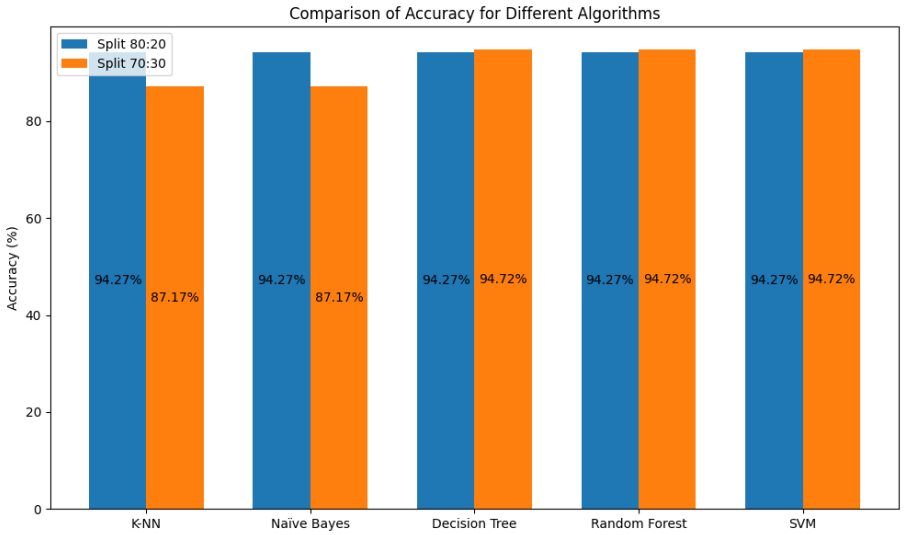

Figure 9. Comparison of Accuracy for Different Algorithms

The diagram in Figure 9 shows the accuracy comparison of the five algorithms for two data splits (80:20 and 70:30). It can be seen that the Decision Tree, Random Forest, and Support Vector Machine

algorithms achieve the highest accuracy in the 70:30 data division, which is 94.72%. This result is in line with research conducted by Salama Abd Elminaam et al. (2023) [7] who found that Random Forest has a high accuracy rate of 98.6%. On the other hand, K-NN and Naive Bayes showed a significant decrease in accuracy at a 70:30 split compared to 80:20. This shows that algorithm selection and data partitioning have a great influence on model performance.

4. CONCLUSION

This study evaluates the performance of five classification algorithms namely K-Nearest Neighbors (K-NN), Naive Bayes Classifier (NBC), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) to assess their ability to predict lung diseases. The results showed that algorithms previously known to excel in heart disease classification also performed well in lung disease prediction, with DT, RF, and SVM consistently achieving high accuracy in both data sharing scenarios. In contrast to previous results showing lower performance, DT provided competitive results this time, suggesting that the effectiveness of algorithms can vary significantly depending on the characteristics of the data set and the classification problem at hand. These findings confirm the importance of contextual evaluation in selecting algorithms for classification tasks in the medical field.

REFERENCES

[1] Aditya Ingole, Yuvraj Patil, Yashraj Wawkar, and Aboli Deole, "Review on Deep Learning for Pulmonary Diseases Detection Using Chest X-Ray," International Journal of Advanced Research in Science, Communication and Technology, pp. 542–547, May 2024, doi: 10.48175/ijarsct-18577.

[2] S. Kamran Hussain et al., "Machine Learning Approaches for Early Detection of Lung Cancer," Journal of Computing & Biomedical Informatics, 2023, doi: 10.56979/601/2023.

[3] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges," Feb. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/a17020078.

[4] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in IOP Conference Series: Materials Science and Engineering, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012072.

[5] A. F. Lubis et al., "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Naïve Bayes and Decission Tree," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 44–51, Apr. 2024, doi: 10.57152/predatecs.v2i1.1103.

[6] T. A. Assegie, "Heart disease prediction model with k-nearest neighbor algorithm," International Journal of Informatics and Communication Technology (IJ-ICT), vol. 10, no. 3, p. 225, Dec. 2021, doi: 10.11591/ijict.v10i3.pp225-230.

[7] D. Salama AbdElminaam, N. Mohamed, H. Wael, A. Khaled, and A. Moataz, "MLHeartDisPrediction: Heart Disease Prediction using Machine Learning," 2023. doi: 10.21608/jocc.2023.282098.

[8] A. A. Ahmad and H. Polat, "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm," Diagnostics, vol. 13, no. 14, Jul. 2023, doi: 10.3390/diagnostics13142392.

[9] S. Hadijah Hasanah, "Application of Machine Learning for Heart Disease Classification Using Naive Bayes," Jurnal Matematika MANTIK, vol. 8, no. 1, pp. 68–77, Jun. 2022, doi: 10.15642/mantik.2022.8.1.68-77.

[10] A. Arifuddin, G. S. Buana, R. A. Vinarti, and A. Djunaidy, "Performance Comparison of Decision Tree and Support Vector Machine Algorithms for Heart Failure Prediction," in Procedia Computer Science, Elsevier B.V., 2024, pp. 628–636. doi: 10.1016/j.procs.2024.03.048.

[11] O. Wijaya et al., "Correlation of Sputum Macrophage and Neutrophil with COPD Assessment Test (CAT)," vol. Vol. 32, No 4, 2012.

[12] D. Anwar, Y. Chan, and M. Basyar, "Correlation Between The Degree of Breathlessness According to Modified Medical Research Council Scale (MMRC scale) with The Degree of Chronic Obstructive Pulmonary Disease," vol. Vol. 32, No 4.

[13] G. R. Macklin et al., "Evolving epidemiology of poliovirus serotype 2 following withdrawal of the serotype 2 oral poliovirus vaccine," Science (1979), vol. 368, no. 6489, pp. 401–405, Apr. 2020, doi: 10.1126/science.aba1238.

[14] "World Health Organization, 'Chronic obstructive pulmonary disease (COPD),' 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd). [Accessed: 8 Dec. 2024]."

[15] G. R. Macklin et al., "Evolving epidemiology of poliovirus serotype 2 following withdrawal of the serotype 2 oral poliovirus vaccine," Science (1979), vol. 368, no. 6489, pp. 401–405, Apr. 2020, doi: 10.1126/science.aba1238.

[16] A. I. Putri et al., "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/predatecs.v2i1.1110.

[17] M. Muta'alimah, C. K. Zarry, A. Kurniawan, H. Hasysya, M. F. Firas, and N. Nadhirah, "Classifications of Offline Shopping Trends and Patterns with Machine Learning Algorithms," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 18–25, Apr. 2024, doi: 10.57152/predatecs.v2i1.1099.

[18] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," Eng Appl Artif Intell, vol. 136, p. 108972, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108972.

[19] R. Syahputra, G. J. Yanris, and D. Irmayani, "SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter," Sinkron, vol. 7, no. 2, pp. 671–678, May 2022, doi: 10.33395/sinkron.v7i2.11430.

[20] R. Alfyani and Muljono, "Comparison of Naïve Bayes and KNN Algorithms to understand Hepatitis," International Seminar on Application for Technology of Information and Communication (ISemantic), 2020, doi: 10.1109/iSemantic50169.2020.9234299.

[21] I. Fadil, M. A. Helmiawan, F. Supriadi, A. Saeppani, Y. Sofiyan, and A. Guntara, "Waste Classifier using Naive Bayes Algorithm," in 2022 10th International Conference on Cyber and IT Service Management (CITSM), 2022, pp. 1–5. doi: 10.1109/CITSM56380.2022.9935894.

[22] M. R. Anugrah, N. Nazira, N. A. Al-Qadr, and N. Ihza, "Implementation of C4.5 and Support Vector Machine (SVM) Algorithm for Classification of Coronary Heart Disease," vol. 1, no. 1, pp. 20–25, 2023, doi: 10.7910/DVN/76SIQD.

[23] A. Pushpa Athisaya Sakila Rani and N. Suresh Singh, "Classification and identification of pest, diseases and nutrient deficiency in paddy using layer based EMD phase features with decision tree," Information Processing in Agriculture, 2024, doi: https://doi.org/10.1016/j.inpa.2024.09.003.

[24] G. Pagliarini, S. Scaboro, G. Serra, G. Sciavicco, and I. E. Stan, "Neural-symbolic temporal decision trees for multivariate time series classification," Inf Comput, vol. 301, p. 105209, 2024, doi: https://doi.org/10.1016/j.ic.2024.105209.

[25] S. Talukdar et al., "Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review," Remote Sens (Basel), vol. 12, no. 7, 2020, doi: 10.3390/rs12071135.

[26] G. Aziz, N. Minallah, A. Saeed, J. Frnda, and W. Khan, "Remote sensing based forest cover classification using machine learning," Sci Rep, vol. 14, no. 1, p. 69, 2024, doi: 10.1038/s41598-023-50863-1.

[27] A. Y. Mahmoud, "Novel efficient feature selection: Classification of medical and immunotherapy treatments utilising Random Forest and Decision Trees," Intell Based Med, vol. 10, p. 100151, 2024, doi: https://doi.org/10.1016/j.ibmed.2024.100151.

[28] M. B. Sharr, C. E. Parrish, and J. Jung, "Automated classification of valid and invalid satellite derived bathymetry with random forest," International Journal of Applied Earth Observation and Geoinformation, vol. 129, p. 103796, 2024, doi: https://doi.org/10.1016/j.jag.2024.103796.

[29] M. Fauzi Fayyad, D. Takratama Savra, V. Kurniawan, and B. Hilmi Estanto, "Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm," vol. 1, no. 1, pp. 1–9, 2023, doi: 10.57152/predatecs.v1i1.814.

[30] A. Rahmah, N. Sepriyanti, M. H. Zikri, I. Ambarani, and M. Yusuf Bin Shahar, "Implementation of Support Vector Machine and Random Forest for Heart Failure Disease Classification," vol. 1, no. 1, pp. 34–40, 2023, doi: 10.57152/predatecs.v1i1.816.

[31] N. W. Azani, C. P. Trisya, L. M. Sari, H. Handayani, and M. R. M. Alhamid, "Performance Comparison of ARIMA, LSTM and SVM Models for Electric Energy Consumption Analysis," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 1, no. 2, Feb. 2024, doi: 10.57152/predatecs.v1i2.869.

[32] D. A. Anggoro, "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 5, pp. 1689–1694, May 2020, doi: 10.30534/ijeter/2020/32852020.

[33] J. Cai and N. Xi, "Site classification methodology using support vector machine: A study," Earthquake Research Advances, vol. 4, no. 4, p. 100294, 2024, doi: https://doi.org/10.1016/j.eqrea.2024.100294.