International Journal of Informatics and Computation (IJICOM)
Vol.7, No.2, 2025
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Lung Diseases Classification Using the Naïve Bayes Algorithm

Ulumuddin[1*], Rousyati[2], Rizal Nurzuli[3]

Abstract
Lung disease is one of the diseases with a high rate of spread and mortality, especially in developing countries. Early detection is very important to increase the chances of recovery. This study aims to classify the types of lung disease using Naive Bayes, a probability-based statistical classification method. We gathered the dataset that includes common symptoms of lung disease, such as chronic cough, shortness of breath, chest pain, and others. The results of the study showed that Naive Bayes can achieve a fairly high classification accuracy of 87%. These results indicate that Naive Bayes can be an effective approach to support medical decisions.

Keywords:
Lung Diseases, Classification, Naïve Bayes, Machine Learning

1. Introduction

Lung diseases such as tuberculosis, asthma, and chronic obstructive pulmonary disease (COPD) are major challenges in the world of health. Early diagnosis can help doctors provide appropriate and rapid treatment. However, the diagnostic process often takes a lot of time and money. Therefore, an automated classification system based on artificial intelligence is a promising alternative. Early detection is crucial to minimizing the risk of complications and improving patient outcomes. According to Dachi et al., public health awareness programs and early intervention strategies are essential in managing the spread and impact of pulmonary diseases in clinical settings. Effective classification of symptoms through intelligent systems is thus necessary to support timely diagnosis and treatment [1].

The development of information technology has opened up great opportunities in utilizing health data for deeper analysis. With the abundance of medical data available, data mining is one of the effective methods for finding hidden patterns that can be used to support the diagnosis and decision-making process. One of the data mining methods that is widely used for classification is the Naive Bayes algorithm. The use of data mining techniques has become increasingly common in the healthcare sector to identify patterns in complex datasets. Han et al. emphasized that machine learning algorithms, particularly those in data mining, can extract meaningful information from medical data, aiding clinical decision-making processes [2]. In the context of lung disease classification, algorithms must handle categorical features like symptoms and test results, making probabilistic models especially suitable for such tasks.

The Naïve Bayes algorithm is a widely adopted probabilistic classifier known for its

1. Ulumuddin, Universitas Bina Sarana Informatika, Indonesia (Corresponding Author Email: ulumuddin.udn@bsi.ac.id)
2. Rousyati, Universitas Bina Sarana Informatika, Indonesia
3. STKIP NU Slawi, Indonesia

robustness, simplicity, and ability to perform well even with relatively small datasets. Another research demonstrated that the Naïve Bayes algorithm could produce accurate classification results, especially when combined with appropriate feature selection techniques like Information Gain. Naive Bayes is a classification method based on Bayes' Theorem with the assumption of independence between features. Although simple, this algorithm has proven effective in various fields, including document classification, spam filtering, and the medical field. The main advantages of Naive Bayes are its efficiency in performing classification and its ability to work well even when the independence assumption is not fully met [3].

More studies have adapted Naïve Bayes for discretizing symptom data to classify tuberculosis patients effectively. In the context of lung disease classification, the use of Naive Bayes allows the system to provide predictions of disease types based on a combination of symptoms reported by the patient. This is certainly very helpful in the initial screening process and determining further medical steps. In other words, this classification system can function as a tool in making faster and more accurate medical decisions [4]. In addition, comparisons with other classification models indicate that Naïve Bayes, while basic, performs competitively. Kurniawan found that Naïve Bayes could yield results comparable to more complex algorithms like C4.5 in several health-related classification tasks [10]. These advantages make Naïve Bayes an ideal baseline model for medical diagnosis applications, particularly in resource-constrained environments where computational efficiency and interpretability are critical.

2. Related Works

Several papers present a comprehensive comparison of various classification techniques used in medical diagnosis. The researchers evaluated methods such as Decision Trees, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Naïve Bayes to identify the most effective algorithm for medical datasets. Their experiments utilized publicly available datasets related to health conditions, including heart disease and diabetes, to assess each algorithm's performance in terms of accuracy, precision, and computational efficiency [5][13][14][15][16]. Several studies have explored the application of the Naïve Bayes algorithm in the medical domain due to its simplicity, efficiency, and relatively high accuracy in classifying health-related datasets. Ridwan applied this algorithm to classify Diabetes Mellitus cases and demonstrated satisfactory classification accuracy, suggesting its potential for practical health screening systems [9]. Similarly, Imandasari et al. implemented Naïve Bayes for classifying the suitability of water source development locations, further confirming its versatility in various classification tasks within public health contexts [8].

In the area of lung disease detection, Rahmadewi and Kurnia applied a segmentation method on X-ray images and successfully classified lung disease types, although they relied on image-based inputs rather than structured clinical attributes [6]. Meanwhile, Surono used equal-width discretization as a pre-processing step with Naïve Bayes for classifying tuberculosis patient data, which enhanced classification consistency by standardizing numerical inputs [4]. These works reinforce that Naïve Bayes, combined with suitable pre-processing, offers a solid foundation for medical classification systems.

Another paper extended multiple classification algorithms for medical diagnosis, including Naïve Bayes, Decision Trees, k-NN, and SVM. Their findings highlight that Naïve Bayes holds competitive performance, especially in handling high-dimensional data with limited training time. The study suggests that while Naïve Bayes may not always outperform complex models, its efficiency and interpretability make it suitable for real-time diagnosis support systems, especially when integrated with feature selection or ensemble learning

techniques [12].

An article explored Naïve Bayes to make decisions regarding the classification of determining the recipients of basic food assistance predictions for the recipients of basic food assistance, namely, eligible and not eligible. The Naive Bayes algorithm can produce accuracy for 135 training data with 40 testing data, and seven attributes used resulted in an accuracy of 86%, a recall of 85%, and a precision of 88% [7]. Another implementation of Naïve Bayes is to predict the feasibility of the location of the development of clean water sources in the Tirta Lihou PDAM. Based on the results, there are 8 feasible classes and 11 classes are not feasible with the number of accuracies obtained at 78,95%. From the results obtained, it can be determined that the location is feasible to develop water sources for the community [8].

Diabetes Mellitus, or diabetes, is a metabolic disease caused by high blood sugar levels. Blood sugar is stored or used for energy from the blood that is transferred to human cells by the hormone insulin. When attacked by Diabetes, the human body does not usually produce enough insulin, and the body cannot use the insulin properly as needed. Diabetes Mellitus is listed as the largest contributor to death in the world. Diabetes Mellitus can be classified based on the possibility of being affected by the symptom attributes in the early stages. This disease can be detected because many symptoms are detected. The data used in this analysis is data from the UCI Machine Learning dataset, namely Early-Stage Diabetes Risk in 2020, and consists of 17 attributes. The analysis carried out includes data preprocessing, models, and evaluation. The classification method test shows an accuracy of 90.20% and an AUC value of 0.95 [9].

Another work discussed Naive Bayes and C.45 for determining the acceptance of the Indonesian Credit Card application. For credit card submission cases, C.45 is better than Naive Bayes and when determining the age of birth, Naive Bayes is better than C.45. Whereas in the case of determining the eligibility of prospective credit members in the cooperative, Naive Bayes provides better value in precision, but for recall and accuracy, C.45 gives better results [10].

3. Proposed Method

Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management [11]. In this paper, we utilize Naive Bayes to calculate the probability of data falling into a certain class based on the values of lung disease features. This study adopts the Naive Bayes algorithm as a probabilistic classification model built upon Bayes' Theorem, which calculates the probability of a class given a set of input features. The fundamental formula is:

$$P(C_k \mid \mathrm{x}) = \frac{P(C_k) \cdot P(\mathrm{x} \mid C_k)}{P(\mathrm{x})} \tag{1}$$

This expression calculates the posterior probability $P(C_k \mid \mathrm{x})$, or the probability that an input vector $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ belongs to class C_k, using three components: the prior probability of the class $P(C_k)$, the likelihood of the features given the class $P(\mathrm{x} \mid C_k)$, and the marginal probability of the features $P(\mathrm{x})$. The algorithm assumes that each feature x_i is conditionally independent of the others given the class, which simplifies the likelihood term into a product of individual feature probabilities:

$$P(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k) \tag{2}$$

This naive independence assumption makes computation tractable even with high-

dimensional data. The classification is done by choosing the class with the highest posterior probability, resulting in the decision rule:

$$\hat{C} = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k) \quad (3)$$

In practice, to prevent computational underflow from multiplying many small probabilities, the model uses the logarithmic form of the equation:

$$\hat{C} = \arg\max_{C_k} [\log P(C_k) + \sum_{i=1}^{n} \log P(x_i|C_k)] \quad (4)$$

This transformation retains the ranking of probabilities while where the likelihood $P(x_i \mid C_k)$ is modeled using a normal distribution parameterized by class-specific means and variances.

4. Experimental Setup

1. Dataset

This study utilized patient data collected through field research conducted at the Jatibogor Village Community Health Center, located in the Suradadi District of Tegal Regency. The research focused on individuals exhibiting symptoms indicative of lung-related illnesses, intending to support diagnostic and classification efforts in pulmonary health assessment. The dataset comprises structured clinical attributes reflecting key respiratory symptoms and diagnostic outcomes. These attributes include the presence or absence of chronic cough, shortness of breath, chest pain, fever, and weight loss, that consist all recorded as binary variables (yes/no). Additionally, the dataset incorporates radiological findings (X-ray results categorized as normal or abnormal) and the final diagnosis label, which classifies each patient case into one of four categories: asthma, tuberculosis, chronic obstructive pulmonary disease (COPD), or pneumonia.

2. Data Pre-Processing

In this study, the preprocessing stage played a critical role in preparing the raw clinical data for effective model training and classification. The researchers began by performing data cleaning to remove inconsistencies, such as missing values or duplicate records, ensuring the dataset's integrity. They then standardized categorical variables, including symptom indicators (e.g., chronic cough, fever). We convert all binary responses into uniform numerical representations (e.g., yes = 1, no = 0) to facilitate computational processing. This transformation allowed the Naïve Bayes algorithm to efficiently handle categorical input features during probability estimation.

Following normalization, the dataset underwent label encoding for the target variable, where disease classifications (asthma, tuberculosis, COPD, and pneumonia) were mapped to numerical identifiers. The pre-processing phase also involved evaluating class distribution to identify potential imbalance, which could bias model performance. In cases where class imbalance was detected, the researchers applied sampling techniques to ensure a more equitable representation across disease categories. Overall, the pre-processing procedures established a clean and structured dataset, suitable for probabilistic modeling and accurate classification using the Naïve Bayes algorithm.

3. Classification and Evaluation

In the classification process, we implemented the Naïve Bayes model to categorize patients into one of four lung disease classes: asthma, tuberculosis, COPD, or pneumonia. They trained the model using a structured dataset containing symptom attributes such as chronic cough, chest pain, fever, weight loss, shortness of breath, and X-ray findings. Each record in the dataset represented a patient case, with binary or categorical features acting as predictors and the final diagnosis as the target class. The algorithm computed the posterior probability for each class based on the conditional likelihood of the input features, ultimately assigning the class label with the highest probability.

To optimize classification performance, the dataset was split into training and testing sets using stratified sampling, preserving class distribution. The training data was used to fit the Naïve Bayes model, while the testing set served to evaluate the model's generalization capabilities. During classification, the model relied on the assumption of conditional independence between features, which is a core principle of Naïve Bayes. Despite this simplifying assumption, the model was chosen due to its robustness in handling categorical medical data and its computational efficiency for multi-class classification tasks.

The evaluation of model performance employed several standard classification metrics, including accuracy, precision, recall (sensitivity), and the F1-score. Accuracy measures the overall correctness of the model by calculating the proportion of true predictions over the total cases. Precision evaluated the correctness of positive predictions for each disease class, while recall assessed the model's ability to correctly identify actual disease cases. The F1-score, as the harmonic mean of precision and recall, provided a balanced measure for imbalanced datasets. Additionally, a confusion matrix was used to analyze misclassification patterns across the four classes, offering insight into where the model performed well and where improvements were necessary. These evaluation metrics collectively validated the reliability and diagnostic potential of the Naïve Bayes classifier in identifying lung diseases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **TP**: True Positives display correctly predicted positive cases
- **TN**: True Negatives display correctly predicted negative cases
- **FP**: False Positives display incorrectly predicted positive cases
- **FN**: False Negatives display incorrectly predicted negative cases

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures how many of the predicted positive cases are actually positive.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall indicates how many actual positive cases the model successfully identified.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score provides a balance between precision and recall, especially useful in cases of class imbalance. These metrics allow for a comprehensive assessment of how well the Naïve Bayes classifier identifies and distinguishes between various lung disease categories.

5. Results and Analysis

In this part, we present the evaluation metrics from the classification results in Table 1:
Table 1: Evaluation metrics result of the study with Naïve Bayes algorithm

Metric	Score
Accuracy	87%
Precision	85%
Recall	84%
F1-Score	84.5%

According to the experimental result, the evaluation of the Naïve Bayes algorithm in classifying lung diseases yielded an accuracy of 87%, indicating that the model correctly predicted the disease class in the majority of instances. This high accuracy reflects the algorithm's overall effectiveness in handling the multi-class classification task, especially with input features such as symptoms and X-ray outcomes. Precision, at 85%, suggests that among all instances classified as positive for a specific lung disease (e.g., asthma or tuberculosis), 85% were truly correct. This high precision value demonstrates that the model made relatively few false positive predictions, which is crucial in medical diagnostics to avoid misdiagnosing healthy individuals or those with other conditions.

Meanwhile, the recall rate of 84% indicates that the model successfully identified 84% of all actual positive cases. When combined with precision, the F1-Score of 84.5% confirms the model's balanced performance, especially in scenarios with class imbalance. These results affirm the Naïve Bayes algorithm's viability for supporting clinical decision-making in the early detection of lung diseases based on patient symptoms. The model shows good performance in distinguishing between different lung diseases based on symptoms. The assumption of independence between features is not fully met in medical data, but the results are still satisfactory. Naive Bayes is also proven to be fast and efficient in the training and prediction process.

6. Conclusion

This study demonstrates that the Naïve Bayes algorithm performs effectively in classifying lung diseases by utilizing clinical symptom data, such as chronic cough, chest pain, and X-ray results. The model achieved an accuracy of 87%, indicating a high level of reliability in predicting disease categories such as asthma, tuberculosis, COPD, and pneumonia. These findings support the algorithm’s potential use as a decision support tool in aiding medical personnel with early-stage diagnosis. For future research, expanding the dataset to include more diverse and comprehensive patient records is highly recommended to enhance model generalizability. Additionally, integrating ensemble learning techniques such as Random Forest or Gradient Boosting may improve classification performance further by reducing error and increasing robustness. This development could contribute to more accurate and scalable diagnostic tools in clinical practice.

Acknowledgment

We are grateful to Allah SWT who has given many blessings to all of us so that we can complete this research, and we offer our deepest gratitude we offer to all parties who have helped researchers in completing this research. Hopefully, the results of this research will be useful for all of us

References

[1] R. A. Dachi, L. Hakim, and T. Wandra, “Dissemination on Pulmonary Tuberculosis at Putri Hijau Hospital Medan,” *Jurnal Abdimas Mutiara*, vol. 3, pp. 367–374, 2022. [Online]. Available: http://e-journal.sari-mutiara.ac.id/index.php/JAM/article/view/3268

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[3] L. D. Utami et al., “Integration of Information Gain for Feature Selection and Adaboost to Reduce Bias in Restaurant Review Sentiment Analysis Using Naïve Bayes Algorithm,” *Jurnal Intelligent System*, vol. 1, no. 2, 2015. [Online]. Available: http://journal.ilmukomputer.org

[4] S. Surono, “Equal-Width Interval Discretization in Naïve Bayes (Case Study: Classification of Tuberculosis Patients),” *Program Studi Matematika FAST UAD Jl Ringroad Selatan*.

[5] D. Fitriati and I. Gibran, “Expert System for Meningitis Diagnosis Using Forward Chaining Method,” *Jurnal UMJ*, vol. 12, no. 1, pp. 46–50, 2021. [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it/index

[6] R. Rahmadewi and R. Kurnia, “Lung Disease Classification Based on X-ray Images Using Sobel Segmentation Method,” *Jurnal Nasional Teknik Elektro*, vol. 5, no. 1, p. 7, 2016, doi: 10.25077/jnte.v5n1.174.2016.

[7] Damuri et al., “Implementation of Data Mining with Naïve Bayes Algorithm for Eligibility Classification of Food Aid Recipients,” *JURIKOM (Jurnal Riset Komputer)*, vol. 8, no. 6, p. 219, 2021, doi: 10.30865/jurikom.v8i6.3655.

[8] T. Imandasari et al., “Naïve Bayes Algorithm in Classifying Locations for Water Source Development,” *Prosiding Seminar Nasional Riset Informatika dan Sains*, vol. 1, no. September, p. 750, 2019, doi: 10.30645/senaris.v1i0.81.

[9] Ridwan, “Application of Naïve Bayes Algorithm for Diabetes Mellitus Classification,” *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.

[10] Y. I. Kurniawan, “Comparison of Naïve Bayes and C4.5 Algorithms in Data Mining Classification,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 4, pp. 455–464, 2018, doi: 10.25126/jtiik.201854803.

[11] Rish, “An Empirical Study of the Naïve Bayes Classifier.”

[12] T. T. Tran, T. M. Pham, and N. M. Nguyen, “A comparative study of classification techniques in medical diagnosis,” *Procedia Computer Science*, vol. 132, pp. 1013–1020, 2018, doi: 10.1016/j.procs.2018.05.212.

[13] P. Barandela, R. Valdovinos, and J. S. Sánchez, “New applications of ensembles of classifiers,” *Pattern Analysis & Applications*, vol. 6, no. 3, pp. 245–256, 2003, doi: 10.1007/s10044-003-0204-1.

[14] R. M. Haraty and A. A. Zantout, “A survey of data mining techniques applied to medical data,” in *Proceedings of the International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2014, pp. 222–227, doi: 10.1109/ICTER.2014.7083915.

[15] L. Rokach and O. Maimon, “Decision Trees,” in *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 2005, pp. 165–192, doi: 10.1007/0-387-25465-X_9.

[16] A. Patel, A. Doshi, and A. Patel, “Diagnosis of asthma using Naïve Bayes classification: A comparative study,” *International Journal of Computer Applications*, vol. 66, no. 23, pp. 6–10, 2013, doi: 10.5120/11055-6374.