Predicting Lung Disease Using Machine Learning: A Comparative Study of Logistic Regression, k-Nearest Neighbors, and Naive Bayes

Sridevi PC[1], A. Suphalakshmi[2], Vijayapuram Anusha [3], Manogna[4]
Assistant Professor[1], Dean Faculty of Engineering[2], B. tech[3,4]
Takshashil University, Ongur.
srideviPc@gmail.com[1]

ABSTRACT:

Machine Learning (ML) has emerged as a transformative technology across various domains, revolutionizing industries such as healthcare, finance, and manufacturing. This paper provides a comprehensive review of recent advancements in ML, focusing on key algorithms, applications, and challenges. Specifically, it explores the use of common ML algorithms for predicting lung disease for the given data set containing 1,000 instances. The study demonstrates how these algorithms are applied to classify lung disease cases, highlighting their effectiveness and comparing performance metrics. Despite the advancements, challenges such as the need for large labeled datasets, model interpretability, computational complexity, and biases in predictions remain. The paper also discusses future research directions, including few-shot learning, explainable AI, transfer learning, edge computing, and ethical AI, which aim to address current limitations and unlock new potentials for ML in healthcare and beyond.

Keywords: Machine Learning models, Artificial Intelligence, Deep Learning, Data Science, Algorithms

1. INTRODUCTION

Machine Learning (ML) application in disease prediction is seen as significant advancements in recent years, providing powerful tools for data-driven decision-making in various fields. In the healthcare sector, ML techniques are increasingly being adopted for early detection, diagnosis, and prediction of diseases. One such critical area is lung disease prediction, where timely identification can lead to more [1] effective treatments and improved patient outcomes. Traditionally, medical professionals have relied on diagnostic tools such as medical imaging and lab tests, but ML algorithms now offer the potential to enhance the accuracy and efficiency of these processes [2]. This paper explores the application of three widely used ML algorithms— Logistic Regression, k-Nearest Neighbors, and Naive Bayes—for predicting lung disease based on a dataset containing 1,000 instances. This research work assesses the performance of the traditional algorithm mentioned using the performance metrics like precision, Recall, Accuracy and F1 Score. this work provides the strengths and weaknesses of each algorithm; the result provides an insight into the most effective approaches for lung disease prediction. Additionally, the paper discusses the challenges associated with ML in healthcare, including data quality, model interpretability, and computational constraints. The paper is organized as introduction in chapter 1 and literature survey in chapter 2 and the 3rd chapter methodology, chapter 4 result and discussion and finally the conclusion.

2.LITERATURE SURVEY

Machine Learning techniques have gained considerable attention in medical research, particularly for disease prediction and classification tasks. Several studies have investigated the application of various algorithms for lung disease diagnosis. For binary classification problem the Logistic Regression model is used. This LR is a statistical method [3], has been extensively applied in medical research due to its simplicity and interpretability. In the context of lung disease prediction, LR has been employed to identify patterns in factors such as smoking history, age, and respiratory function, enabling early detection of diseases like lung cancer.

k-Nearest Neighbors (KNN) is another widely used algorithm in medical applications. KNN is a non-parametric method [4] that classifies a sample based on its proximity to labeled instances in the feature space. Numerous studies have utilized KNN to predict lung disease by analyzing patient data such as lung function tests, genetic information, and environmental exposure. KNN is simple and it has the ability to handle non-linear data made it a popular choice in healthcare applications.
Naive Bayes (NB), based on Bayes' Theorem, has also been employed for disease prediction tasks, including lung disease classification [5]. NB is particularly effective when the features are conditionally independent, making it suitable for problems where this assumption holds. Research has shown that Naive Bayes can efficiently classify lung disease instances with high accuracy, particularly when working with large datasets.

Several studies have compared these algorithms in the context of lung disease prediction. In one study, LR outperformed both KNN and NB in terms of classification accuracy when using a dataset of 1,000 medical instances. However, KNN showed better performance in handling non-linear relationships, while Naive Bayes demonstrated robustness in smaller datasets with fewer variables. These findings underline the need of picking the proper algorithm grounded on the environment of the data and the specific application.

Despite the promising results of ML algorithms in healthcare, challenges remain, particularly regarding the interpretability of complex models, the need for large labeled datasets, and the potential for bias in predictions. The growing interest in Explainable AI (XAI) [6] aims to address these issues by providing transparent models that healthcare professionals can trust and use effectively. Additionally, recent advancements in deep learning and transfer learning are opening new avenues for improving disease prediction systems by leveraging large-scale, diverse datasets.
In conclusion, this paper builds upon existing research by comparing LR, KNN, and NB algorithms for lung disease prediction and offers insights into their applicability, strengths, and limitations in healthcare. The results in this research work contribute to the growing body of knowledge on ML applications in medical diagnosis, providing valuable recommendations for future research and practical implementations in the healthcare sector.

3.METHODOLOGY

The research work employs methods at different levels which includes the data collection, processing and implementing the models like NB, KNN and LR. This work evaluates the models using the metrics like accuracy, precision, recall all of them are discussed in this section.

3.1 Data set

The lung diseases dataset, taken from Kaggle [11], consists of various patient-related features and medical indicators aimed at understanding the factors contributing to lung health. Key attributes in the dataset typically include the age of the patient, which may influence susceptibility to lung diseases; gender, which can affect the risk of lung diseases; smoking history, a categorical variable indicating whether the patient has a history of smoking (e.g., Yes, No); breathing patterns, which include measures like forced vital capacity (FVC) or forced expiratory volume (FEV) to assess lung function; medical history, which includes records of other chronic diseases (e.g., asthma, tuberculosis) or prior lung conditions that could contribute to the development of lung diseases; oxygen levels, which measure oxygen saturation or partial oxygen pressure, indicating lung efficiency; chest X-ray/CT imaging data, which shows abnormalities or structural issues in the lungs; and symptoms, which include information on symptoms related to lung diseases.

The lung diseases dataset, taken from Kaggle, consists of various patient-related features and medical indicators aimed at understanding the factors contributing to lung health. Key attributes in the dataset typically include. The lung disease dataset, consisting of 5,200 instances, includes 26 attributes that capture a wide range of patient demographics, lifestyle factors, environmental exposures, symptoms, and the target classification label. Among these, most features are of integer type (indicating binary or ordinal encoding), while a few are object type, likely representing categorical or identifier data.

The Patient Id is an object-type field used to uniquely identify each individual, while the index column appears to be a positional identifier. Age and Gender are demographic features, with gender numerically encoded (e.g., 0 for female, 1 for male). Pollution from air, usage of Alcohol, Allergic to dust, Hazards caused due to Occupational, Risk by Genetic, Un Balanced Diet, Obesity, Smoking, and Passive Smoker, each likely encoded as 0 (no risk/exposure) or 1 (risk/exposure). Medical history is captured through chronic lung disease, and a rich set of symptoms related to lung conditions. These symptom attributes are also binary or ordinal in nature. The target variable, Level, is of object type and likely represents the severity or presence of lung disease (e.g., "Low", "Medium", "High" or "Yes"/"No"). This well-structured dataset enables the use of supervised machine learning techniques to build predictive models for lung disease classification. The inclusion of diverse input features allows for capturing multifactorial contributors to lung health, making the dataset suitable for comparative analysis using algorithms like Logistic Regression, KNN, and Naive Bayes the figur1 contains sample data set.

Table 1: Sample Data set

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occu Pational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occu Pational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring
mean	499.500000	37.174000	1.402000	3.84400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	4.859000	3.856000	3.855000	4.240000	3.777000	3.746000	3.923000	3.536000	3.853000	2.926000
std	288.819436	12.005493	0.490054	2.03034	2.626047	1.980833	2.107805	2.126999	1.848518	2.135528	2.427965	2.244616	2.206546	2.285087	2.041921	2.270383	2.388048	1.832502	2.039007	1.474686
min	0.000000	14.000000	1.000000	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	249.750000	27.750000	1.000000	2.00000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	3.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occu Pational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring
	00	00																		
50%	499.500000000	36.000000000	1.000000000	3.000000	5.000000000	6.000000000	5.000000000	5.000000000	4.000000000	4.000000000	4.000000000	3.000000000	3.000000000	4.000000000	4.000000000	4.000000000	4.000000000	3.000000000	4.000000000	3.000000000
75%	749.250000000	45.000000000	2.000000000	6.000000	7.000000000	7.000000000	7.000000000	7.000000000	6.000000000	7.000000000	7.000000000	5.000000000	6.000000000	6.000000000	5.000000000	5.000000000	5.000000000	5.000000000	6.000000000	4.000000000
max	999.000000000	73.000000000	2.000000000	8.000000	8.000000000	8.000000000	8.000000000	7.000000000	7.000000000	7.000000000	9.000000000	9.000000000	8.000000000	9.000000000	8.000000000	8.000000000	9.000000000	7.000000000	7.000000000	7.000000000

DATASET STATISTICS

The information of the data is analyzed to get the insight in to the data , to ensure smooth further processing. The statics analysis includes analysis of numerical future, variable distribution analysis [12].

Preprocessing

Categorical features such as Gender and Smoking status were label-encoded. Numerical features were scaled using Min-Max normalization [13] to enhance the performance of distance-based algorithms like KNN

Models

This study employs three popular and interpretable machine learning algorithms—**NB [14]**, **k-NN [15]**, and **LR [16]**—to predict lung disease based on patient data. These models are chosen for their simplicity, effectiveness, and widespread use in classification problems, particularly in medical diagnostics.

Naive Bayes (NB)

Naive Bayes is based on Bayes' Theorem and it is a **probabilistic classifier**, the concept behind this theorem is that it assumes each predictor are independence [7]. This model can be used well with high dimensional data and it will require only less amount of data set.

Bayes' Theorem:

$$P(M|X)= \frac{P(X|M)\cdot P(M)}{P(X)} \quad (1)$$

Where:

- P(M|X) is the class M posterior probability (e.g., presence of lung disease) given feature vector X
- P(X|M) class M is the likelihood of features.
- P(M) class M's prior probability
- P(X) is the probability of the features (acts as a normalization constant)

2. k-Nearest Neighbors (KNN)

KNN is classify the data based on majority of classes and it is a instance based algorithm and it is non-parametric [8] label among the k-nearest data points in the feature space. It is intuitive and effective for multi-class problems and no assumption is made for the distribution of data.

Distance Metric:

The most commonly used distance function is **Euclidean distance**:

$$\mathbf{d}(r, \mathbf{q}) = \mathbf{i} = \sqrt{\sum_{i=1}^{n}(\mathbf{ri} - \mathbf{qi})\mathbf{2}} \quad (2)$$

Where:

- r and q are the feature vectors of two data points
- n is the number of features

After computing distances of all training instances and test instances difference, the class label is assigned based on a **majority vote** of the kkk nearest neighbors.

3. Logistic Regression (LR)

Logistic Regression is a **binary classification linear model**, which estimates category of particular given input using probability [9]. It is suitable for modelling the likelihood of disease presence or absence.

Logistic Function (Sigmoid):

$$\mathbf{P(y = 1 \mid X) = 1 + e - (\beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn)1} \quad (3)$$

Where:

- β0 is the intercept
- βi are the coefficients for the features xi
- The output is the predicted probability that y=1 (e.g., lung disease present)

Training and evaluation

The research work uses 80-20 data split for testing and training [17]. Python 3.9 and Scikit-learn library are used to implement the classification model [20]. For KNN, the number of neighbours (k) was set to 5. Logistic Regression used L2 regularization with the 'liblinear' solver. Naive Bayes was implemented using GaussianNB. No advanced hyperparameter tuning was conducted as the focus was on baseline comparisons.

All three models are evaluated using performance metrics such as accuracy, precision, recall, F1 score, and confusion matrix [10].

4.Result and Discussion

The result obtained from this research work is in this comparative analysis which include statistics of the data set and model and evaluation process.

Statistics

The data info revels that the research work uses 5200 data set and 8 attribute among which 5 data are categorical data and 3 data are numeric data. Among which 5.8 % data are missing. That is in total 301 data is missing. Figure 1 is the shows a histogram of the 'Age' variable, Distributions. The bin width is set to 50, meaning each bar represents a range of 50 years. The bin width is set to 50, meaning each bar represents a range of 50 years.

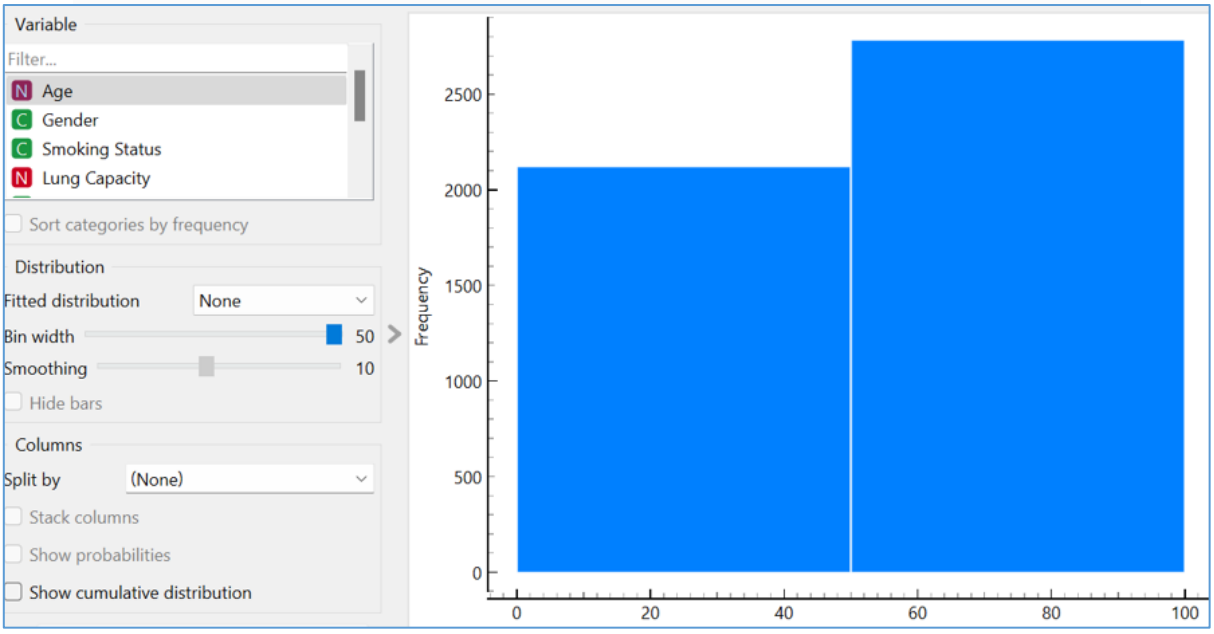
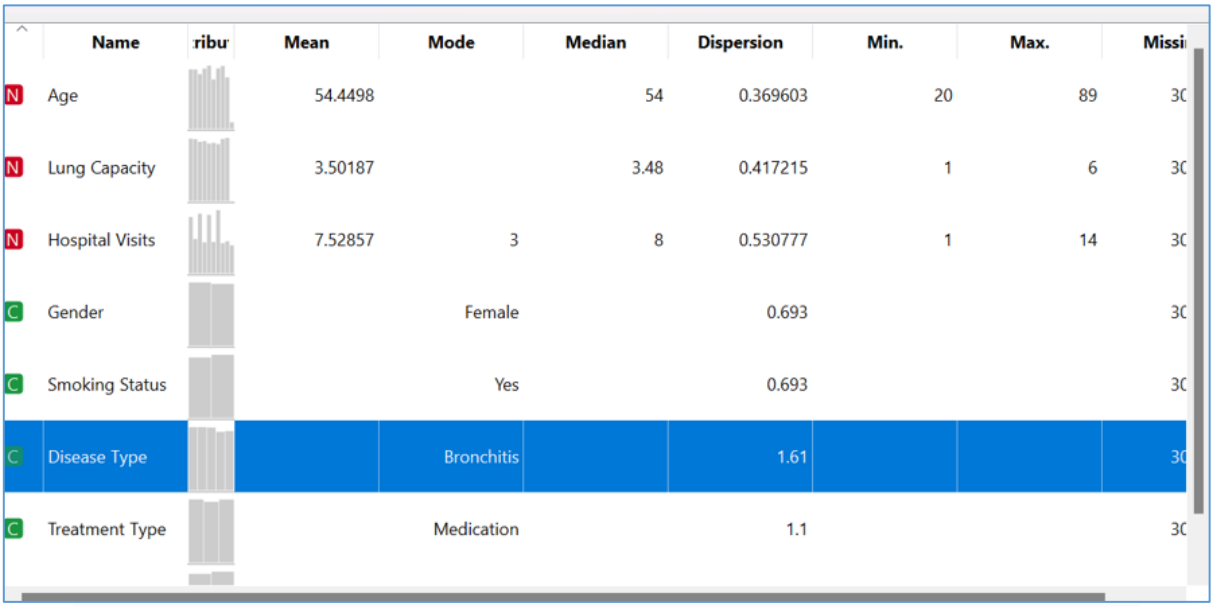


Figure 1: Age Distribution

The dataset in figure 2 includes features like Age, Lung Capacity, and Disease Type. Notably, there are 30 missing values across all features. "Disease Type" is selected, with "Bronchitis" as the mode and a dispersion of 1.61. Addressing the missing data is a key next step.



Name	ribu	Mean	Mode	Median	Dispersion	Min.	Max.	Missi
Age		54.4498		54	0.369603	20	89	30
Lung Capacity		3.50187		3.48	0.417215	1	6	30
Hospital Visits		7.52857	3	8	0.530777	1	14	30
Gender			Female		0.693			30
Smoking Status			Yes		0.693			30
Disease Type			Bronchitis		1.61			30
Treatment Type			Medication		1.1			30

Figure 2 : future Statistics

k-Nearest Neighbors (kNN)

This study evaluates the predictive performance of three supervised machine learning models on a lung disease dataset comprising 1,000 patient records and 24 predictor variables. The kNN model in table 2 and figure 3 yielded an **accuracy of 48.3%**, **F1-score of 48.2%**, **precision of 48.4%**, and **recall of 48.3%**. The **Area Under the Curve (AUC)** was recorded at **46.6%**, this model lack in differentiating the classes propely. The relatively low performance across all metrics may be attributed to the algorithm's sensitivity to the curse of dimensionality and lack of feature scaling. Additionally, kNN is a lazy learner and does not generalize from the training data, which can hinder performance in datasets with mixed or noisy features.

Table 2: k-Nearest Neighbors (kNN) - Performance Table

Metric	Percentage
AUC	46.6
Accuracy (CA)	48.3
F1 Score	48.2
Precision	48.4
Recall	48.3

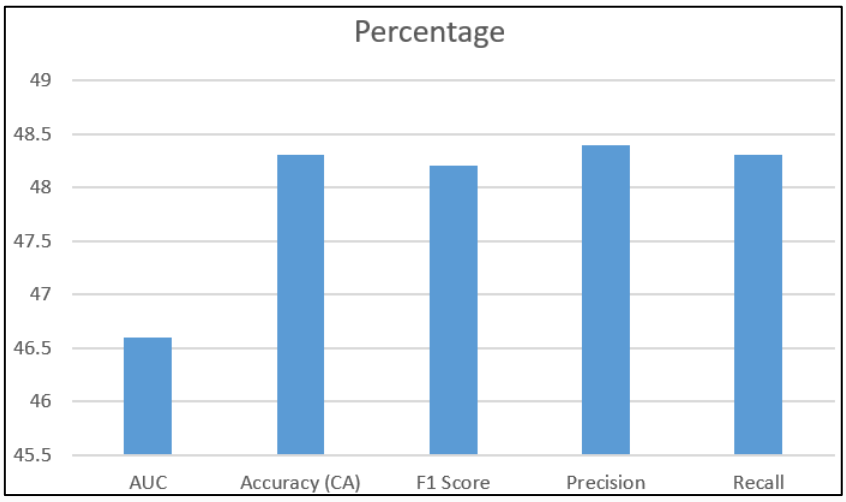


Figure 3: k-Nearest Neighbors (kNN) - Performance

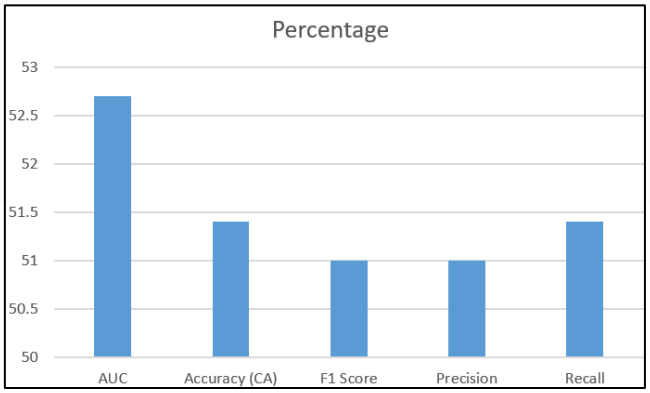
Logistic Regression (LR)

Logistic Regression performed marginally better, achieving an **51.4% accuracy**, **F1-score of 51.0%**, and **AUC of 52.7%**. The **precision and recall** were both at **51.0%** and **51.4%**, respectively table 3 figure 4. These results suggest that logistic regression slightly improves the prediction of lung disease but still falls short of clinical utility. The performance indicates that either the connection between the predictors and the target variable is only mildly linear, or the features themselves aren't highly effective at distinguishing between outcomes.

Table 3: Logistic Regression - Performance Table

Metric	Percentage
AUC	52.7
Accuracy (CA)	51.4
F1 Score	51.0
Precision	51.0
Recall	51.4

Figure 4: Logistic Regression - Performance



Naive Bayes (NB)

Among the three models, Naive Bayes showed the **best performance**, with an **accuracy of 53.6%**, **F1-score of 53.4%**, **precision and recall both at 53.4%**, and an **AUC of 53.9%**. Despite its simplifying assumption of feature independence, NB outperformed kNN and LR. This may be due to the algorithm's robustness to irrelevant features and its effectiveness on small-to-medium-sized

datasets. However, the overall performance still indicates that the features may not be sufficiently informative or separable to yield high classification accuracy.

Table 4: **Naive Bayes** - Performance Table

Metric	Percentage
AUC	53.9
Accuracy (CA)	53.6
F1 Score	53.4
Precision	53.4
Recall	53.4

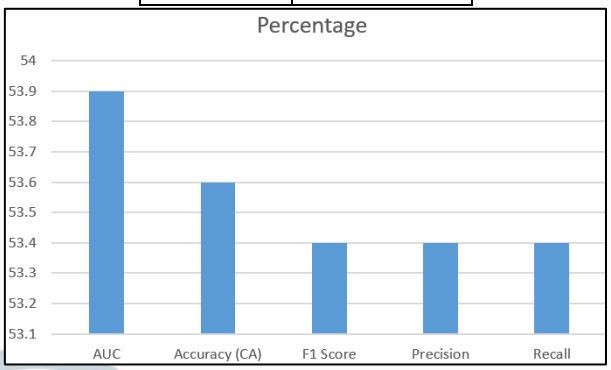


Figure 5: **Naive Bayes** – Performance

Table 5: Confusion matrix Naïve Bayes

	Detected as Positive	Detected as Negative
Actually Positive	267	233
Actually Negative	232	268

Table 6: Confusion matrix Knn

	Detected as Positive	Detected as Negative
Actually Positive	241	259
Actually Negative	256	244

Table 7: confusion Matrix LR

	Detected as Positive	Detected as Negative
Actually Positive	257	243
Actually Negative	246	254

Table 8: Comparative table

Model	Acc	Pre	Recall	F1-Score	AUC
LR	51.40%	51.00%	51.40%	51.00%	52.70%
KNN	48.30%	48.40%	48.30%	48.20%	46.60%
Naive Bayes	53.60%	53.40%	53.40%	53.40%	53.90%

The table through 5 to 7 provides the confusion matrix of the three models and table 8 gives the comparative analysis of each model with highest accuracy for model naïve bayes with 53 accuracies. Overall, while Naive Bayes provided slightly better results, **none of the models surpassed a 55% accuracy threshold**, pointing toward the dataset's limitations or the need for advanced feature engineering. The close clustering of metrics across models suggests:

- Limited variance in feature importance
- Possible class imbalance
- The need for dimensionality reduction or data transformation (e.g., PCA, normalization)

The result shows that all the model used here are having low accuracy and further refinement are needed. This may include hyper parameter tuning, ensambling and optimization to get better positive negative distinguishing of classes.

5.CONCLUSION

This research presents a comparative study of three classical machine learning algorithms—k-Nearest Neighbors (kNN), Logistic Regression (LR), and Naive Bayes (NB)—for lung disease prediction using a Kaggle dataset consisting of 1,000 patient records. Among the models evaluated, Naive Bayes achieved the highest accuracy (53.6%) and AUC (53.9%), followed closely by Logistic Regression and kNN. Despite these results, all three models demonstrated limited predictive capability, suggesting that the current feature set and model complexity are insufficient for high-confidence clinical predictions. To enhance performance in future studies, several improvements are recommended. First, advanced preprocessing techniques such as feature scaling, categorical encoding, and normalization can help optimize model inputs. Second, incorporating domain-specific medical features—such as lung function tests, biomarkers, and imaging-derived variables—may significantly improve the models' ability to capture underlying patterns. Additionally, the use of ensemble learning methods (e.g., Random Forests, Gradient Boosting) or deep learning architectures could offer improved generalization and feature extraction. Addressing data-related challenges like class imbalance and noise through data augmentation, resampling, or feature selection is also critical. In summary, while traditional ML models provide a useful starting point, more sophisticated and context-aware approaches are essential for achieving clinically viable accuracy in lung disease prediction tasks.

6.REFERENCE

[1] Jung, T., & Vij, N. (2021). Early diagnosis and real-time monitoring of regional lung function changes to prevent chronic obstructive pulmonary disease progression to severe emphysema. *Journal of Clinical Medicine, 10*(24), 5811. https://doi.org/10.3390/jcm10245811

[2] Kaplan, A., Cao, H., FitzGerald, J. M., Iannotti, N., Yang, E., Kocks, J. W. H., Kostikas, K., et al. (2021). Artificial intelligence/machine learning in respiratory medicine and potential

role in asthma and COPD diagnosis. *The Journal of Allergy and Clinical Immunology: In Practice, 9*(6), 2255–2261. https://doi.org/10.1016/j.jaip.2021.01.059

[3] De Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. F. (2017). Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems with Applications, 69*, 62–73. https://doi.org/10.1016/j.eswa.2016.09.023

[4] Obid, S. J., Xudoyqulov, D. S. O., & Avazov, A. E. O. (2023). Non-parametric methods: K-nearest neighbors model. *International Journal of Advance Scientific Research, 3*(12), 18–25.

[5] Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies, 4*(1), 39–45.

[6] Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv*. https://arxiv.org/abs/2006.11371

[7] Indaryono, N. A. P., Saedudin, R. R., & Hamami, F. (2024). Comparison analysis of Random Forest and Naïve Bayes algorithms for rainfall classification based on climate in Indonesia. *SITEKNIK: Journal of Information Systems, Engineering and Applied Technology, 1*(2), 102–109.

[8] Prakash, S., Kalaiselvi, B., & Sivachandar, K. (2025). Recognizing fake documents by instance-based ML algorithm tuning with neighborhood size. *Journal of Applied Data Sciences, 6*(2), 1214–1228.

[9] Ramírez, C. A. M., & Graff Guerrero, M. (2024). Comparison of some logistic regression methodologies in supervised classification for functional data. In *2024 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)* (Vol. 8, pp. 1–9). IEEE. https://doi.org/10.1109/ROPEC56681.2024.000XX

[10] Sathyanarayanan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, 4023–4031.

[11] SyedAli110. (n.d.). *Lungs disease prediction - XGBoost, AdaBoost and RF*. Kaggle. https://www.kaggle.com/code/syedali110/lungs-disease-prediction-xgboost-adaboost-and-rf

[12] Miceli, E., Gino, D., & Castaldo, P. (2024). Approaches to estimate global safety factors for reliability assessment of RC structures using non-linear numerical analyses. *Engineering Structures, 311*, 118193. https://doi.org/10.1016/j.engstruct.2024.118193

[13] Shantal, M., Othman, Z., & Abu Bakar, A. (2025). Missing data imputation using correlation coefficient and min-max normalization weighting. *Intelligent Data Analysis, 29*(2), 372–384.

[14] Al-Haddad, L. A., & Mahdi, N. M. (2024). Efficient multidisciplinary modeling of aircraft undercarriage landing gear using data-driven Naïve Bayes and finite element analysis. *Multiscale and Multidisciplinary Modeling, Experiments and Design, 7*(4), 3187–3199.

[15] Sholekhah, F., Putri, A. D., Rahmaddeni, R., & Efrizoni, L. (2024). Perbandingan algoritma Naïve Bayes dan K-nearest neighbors untuk klasifikasi metabolik sindrom: Comparison of Naive Bayes and K-Nearest Neighbors algorithms for metabolic syndrome classification. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4*(2), 507–514.

[16] Chen, S., Ke, S., Han, S., Gupta, S., & Sivarajah, U. (2024). Which product description phrases affect sales forecasting? An explainable AI framework by integrating WaveNet neural network models with multiple regression. *Decision Support Systems, 176*, 114065.

[17] Jha, K. M., Velaga, V., Routhu, K. K., Sadaram, G., & Boppana, S. B. (2025). Evaluating the effectiveness of machine learning for heart disease prediction in healthcare sector. *Journal of Cardiobiology, 9*(1), 1–17.

[18] Adamczyk, J., & Ludynia, P. (2024). Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python. *SoftwareX, 28*, 101944.