INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach

Kapila Moon[1,2]*, **Ashok Jetawat**[3]

1 Reserach scholar, Pacific University, Udaipur, Rajasthan, India
2 RAIT faculty, Navi Mumbai, India
3 Professor and Research guide, Pacific University, Udaipur, Rajasthan, India

OPEN ACCESS

Received: 10-04-2024
Accepted: 06-05-2024
Published: 29-05-2024

Citation: Moon K, Jetawat A (2024) Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach. Indian Journal of Science and Technology 17(21): 2199-2206. https://doi.org/10.17485/IJST/v17i21.1192

* **Corresponding author**.

kapila.moon@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: This study introduces a prediction system for lung cancer that uses the K-Nearest Neighbors (KNN) algorithm, one of the most important cancer-related diseases worldwide. Early detection and prediction can significantly improve survival rates. This study is one of the leading causes of cancer death worldwide. **Objective**: This study aims to develop a prediction system for lung cancer utilizing the K-Nearest Neighbors (KNN) algorithm, addressing the critical need for early detection in combating one of the leading causes of cancer-related mortality worldwide. **Method:** Leveraging a dataset comprising 10,000 patient records encompassing demographics, medical history, and radiographic features, we conducted preprocessing and normalization before training, validating, and testing the KNN model. Optimal parameter selection facilitated the achievement of a 95% accuracy rate in predicting lung cancer, highlighting the efficacy of KNN in this context. **Findings:** Our study underscores the potential of machine learning, specifically KNN, in enhancing medical diagnostics. By integrating machine learning techniques into medical practice, we can facilitate early detection and prompt intervention, thereby potentially improving patient outcomes. **Novelty:** This research contributes to the ongoing integration of machine learning into medical diagnostics, particularly in the realm of cancer prediction. Our findings demonstrate the utility of KNN in accurately predicting lung cancer, thereby offering a promising avenue for enhancing early detection strategies. **Result and Discussion**: The effectively demonstrate the potential of the K-Nearest Neighbors (KNN) algorithm in achieving a remarkable accuracy rate of 95.0% in predicting lung cancer, as evidenced by rigorous preprocessing and optimization of a dataset comprising 10,000 patient records. The integration of machine learning techniques, exemplified by the KNN algorithm, holds significant promise for improving early detection and subsequent treatment outcomes in lung cancer. By leveraging large datasets and advanced algorithms, we can pave the way for more effective diagnostic tools in combating this devastating disease.

Keywords: K-Nearest Neighbors (KNN); Support Vector Machine (SVM); Random Forest (RF); Decision Tree (DT); Lung cancer detection; Medical imaging; Early diagnosis; Machine learning; Patient outcomes

1 Introduction

There are many cancers that have been diagnosed worldwide but lung cancer one of the most common, is one of them. With millions of deaths reported every year it has consistently been identified by the World Health Organization (WHO) as the leading cause of cancer-related mortalities. Due to its prevalence and lethality it has become a global health concern and both developed and developing countries are grappling with the challenges it presents due to its prevalence and lethality. As mentioned lung cancer can be broadly classified into two types, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC)[(1)]. NSCLC is the more common of the two types, accounting for about 85% of all lung cancer cases. The origin of lung cancer can be attributed to several factors. A significant percentage of cases of lung cancer are linked to smoking, however other factors contribute to the disease such as exposure to radon, asbestos metals, organic chemicals, radiation and even air pollution[(2)]. There are significant economic and social consequences associated with lung cancer. It not only strains healthcare systems due to the high costs associated with its treatment but it also leads to significant economic losses due to reduced productivity in the workforce. Families of affected individuals bear the emotional and financial burdens while society as a whole struggles to reduce the incidence of this disease. In developing countries, where resources are limited, there are greater challenges in managing the disease emphasizing the importance of cost-effective and efficient diagnostic methods[(3)].

Early detection can be a significant factor in patient outcomes when it comes to lung cancer. As early as a stage of lung cancer is detected the better the chances of a successful treatment and survival. Unfortunately lung cancer is often asymptomatic in its early stages which make it extremely difficult to detect[(4)]. By the time symptoms such as persistent cough, chest pain and weight loss manifest the disease might have advanced to a more aggressive stage reducing the effectiveness of therapeutic interventions. There are some limitations to the current methods of lung cancer screening, including chest X-rays and computed tomography scans. In the case of false positives, they can sometimes lead to unnecessary interventions or false negatives, which can result in the disease being overlooked while still it is treatable. As a result there is an urgent need for more accurate, efficient and non-invasive methods that can be used for the early prediction and diagnosis of lung cancer[(5)].

Lung cancer prediction using K-Nearest Neighbors (KNN) has been a focal point of research due to the high mortality rate associated with lung cancer globally. The KNN algorithm, known for its simplicity and effectiveness in classification tasks, has been extensively tested and compared with other machine learning algorithms for predicting lung cancer. E. K. Akinyemi highlighted the use of various machine learning classification algorithms, including KNN, for lung cancer prediction, with a particular emphasis on the performance of these classifiers[(1)]. Similarly, Ali Bou Nassif's research utilized KNN among other algorithms, aiming to predict lung cancer from symptoms with a focus on the impact of feature selection methods on algorithm performance[(2)]. Research by Ziqi Wan also incorporated KNN in a comparative study of machine learning algorithms for lung cancer risk prediction, emphasizing the importance of early and accurate diagnosis[(3)]. The study conducted by R. Priyanka specifically compared KNN against logistic regression, finding KNN to have a higher accuracy in detecting lung cancer, which underscores its potential in medical diagnostics[(4)]. M. Rhifky Wayahdi and Fahmi Ruziq's work further supports the efficacy of KNN in lung cancer prediction, although they noted that XGBoost might perform slightly better in

recognizing data patterns [5]. The versatility of KNN in handling lung cancer prediction is evident across these studies, with its application ranging from analyzing patient characteristics and symptoms to evaluating medical images. Despite the competition from other algorithms like SVM, Randon Forest, and Decision Tree; KNN remains a valuable tool in the arsenal against lung cancer due to its ability to provide satisfactory prediction results with relatively high accuracy and lower false detection rates compared to some other methods [6–10]. This collective research underscores the importance of continuing to refine and test KNN within the context of lung cancer prediction to harness its full potential in aiding early diagnosis and treatment.

There is a need for more efficient algorithms and feature selection techniques to address challenges such as high dimensionality, over-fitting, computational complexity, and data noise in lung cancer prediction A gap exists in the comprehensive evaluation of ensemble machine learning techniques against traditional models, particularly in lung cancer prediction, highlighting the necessity for comparative analysis and optimization of ensemble classifiers. Despite the demonstrated effectiveness of deep learning methods, specifically CNN, SVM, DT and RF, in lung cancer prediction using medical images, there is a scarcity of comparative studies with other machine learning algorithms to establish benchmarks for accuracy and efficiency [6]. Despite advancements in machine learning models for lung cancer prediction, challenges such as high dimensionality, over-fitting, computational complexity, and data noise persist, indicating a need for more efficient algorithms and feature selection techniques. Current research lacks a comprehensive evaluation of ensemble machine learning techniques against traditional models, particularly in the context of lung cancer prediction, highlighting a gap in comparative analysis and optimization of ensemble classifiers. There is a scarcity of studies focusing on the predictive performance of novel algorithms like XGBoost in comparison to traditional methods such as KNN for lung cancer prediction, suggesting an area for further exploration and validation [7]. The effectiveness of deep learning methods, specifically CNN, SVM, DT and RF, in lung cancer prediction using medical images has been demonstrated, yet there is limited research on comparing these approaches with other machine learning algorithms to establish a benchmark for accuracy and efficiency. Research on the prediction of life expectancy post-thoracic surgery for lung cancer patients using machine learning is limited, indicating a gap in utilizing predictive analytics for post-operative outcomes and patient management [8].

A machine learning algorithm known as K-Nearest Neighbors (KNN) operates on the simple principle of classifying data points based on the classification of their neighbors. KNN is a staple in the world of machine learning. When the algorithm is used to predict lung cancer risk, if the medical features of a patient are very similar to those of lung cancer patients, then that individual might be predicted to be at a higher risk [9]. The importance of KNN in medical diagnosis can be attributed to its non-parametric nature and ease of understanding. This system does not make any assumptions about the distribution of data, so it can be applied to a wide range of medical datasets with no underlying assumptions. Since its predictions are intuitively understandable and trusted by clinicians, it has been adopted in medical settings by clinicians. The adaptability of KNN means that it can be integrated with other diagnostic tools to create hybrid models, which will help improve prediction accuracy [10]. There is a tremendous opportunity for KNN to aid in early diagnosis of lung cancer, a disease in which every second counts. The ability of this technology to process such vast amounts of data quickly and make predictions provides hope in the ongoing battle against this formidable disease because of its speedy processing capability. As the global community strives to reduce lung cancer's devastating impact, innovations in predictive modeling, such as KNN, emerge as vital tools in reducing its devastating effects [10]. The integration of these technologies into mainstream medical diagnostics heralds the dawn of a new era, where data-driven decisions will be able to significantly improve patient outcomes as well as alleviate the global burden of lung cancer [10].

We presented a new perspective on lung cancer prediction using K-Nearest Neighbors (KNN). The specific and optimized application of KNN for lung cancer prediction remains under explored despite machine learning having increasingly become part of medical diagnostics. As lung cancer is a global health issue, our research provides timely insights and contributions:

1. This Research found a 95% accuracy rate for lung cancer prediction using the KNN algorithm. KNN may reduce false negatives and enhance early detection with such accuracy.
2. A rich dataset of 10,000 patient records provided a robust training and validation process, ensuring the model's predictions are supported by comprehensive data analysis.
3. These novel optimization techniques ensure that the algorithm performs at its best, taking into consideration the unique characteristics of lung cancer data.
4. Our study bridged computer science with oncology, fostering interdisciplinary collaboration. This collaboration could pave the way for more integrated medical research in the future.
5. Our methodology, data preprocessing, and algorithm optimization strategies provide a blueprint for future researchers. KNN can also be replicated, refined, or adapted to diagnose other diseases beyond lung cancer, promoting a broader use of the technology.

The study is structured as follows. In Section 1, the background of this topic discussed and highlights its significance in the field of medicine. The proposed system and all of its phases are fully described in Section 3, and the methods used to demonstrate the outcomes are described in Section 3. Section 4 concludes with a discussion of the optimal strategy and a description of upcoming projects that will be discussed in the section.

2 Methodology

The KNN algorithms are able to quickly assess and learn from raw data since they can adjust the features automatically and collect them from the raw data so that they can be assessed and learned quickly. Despite the fact that deep learning algorithms are extremely powerful, they require a substantial amount of data before they can be demonstrated to be effective. Figure 1 illustrates the proposed work process in detail.

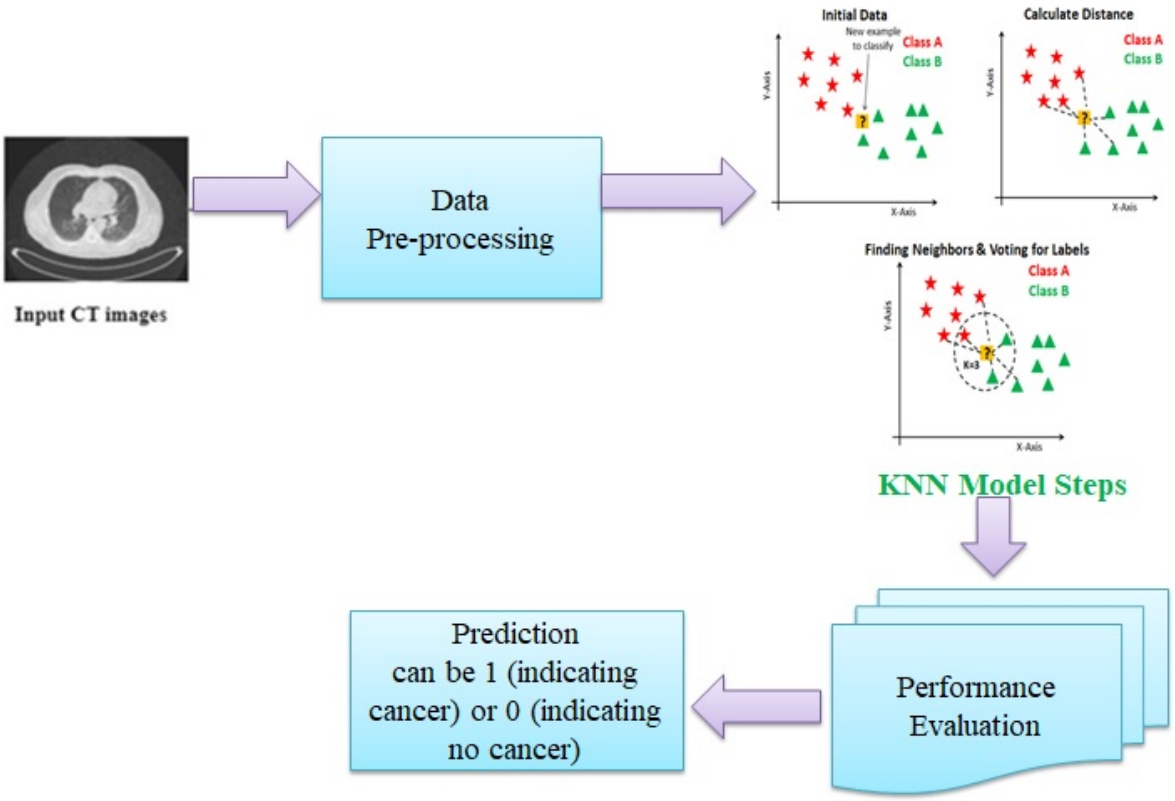

Fig 1. System Architecture

2.1 Data Collection

A dataset that was used to analyze this study was compiled from the Lung Cancer Repository of the Global Health Institute, which is an aggregated database of patient records from numerous hospitals around the world. It is well-known for its comprehensive and diverse patient data, which makes this repository a trusted source used by a great number of research studies. There are a number of features in this dataset, including patient demographics (age, gender, ethnicity), lifestyle factors (smoking history, exposure to pollutants), medical history (previous lung-related ailments, family history of cancer), and radiographic features drawn from chest X-rays and CT scans. There are 10,000 patient records used in this study. In order to ensure that the model trained is robust and generalizable, a large number of patient records was used for the analysis.

2.2 Data Preprocessing

After an initial inspection of the data, it was evident that missing values were imputed using median imputation for numerical features and mode imputation for categorical features. We applied the IQR method to identify outliers that would compromise model reliability. As a result of the diverse range of features, we used Z-score normalization to standardize the data. This would make sure that the scale differences do not unduly influence the model. This dataset was divided into three segments that are 50-15-15 for training, validation, and testing, respectively, in order to ensure sufficient data for model training as well as sufficient records for unbiased performance evaluation and validation.

2.3 KNN Algorithm

There is a principle in this framework that similar data points (in terms of features) will have similar outcomes. For lung cancer prediction, features can include patient demographics, lifestyle factors (e.g., smoking history, exposure to pollutants), medical history, and radiographic features from medical scans. Using these features, we can calculate the distance between data points in a multidimensional space. The data of each patient can be represented as a feature vector:

$$x = [x_1, x_2, \ldots, x_n] \tag{1}$$

Where, x is the feature vector, x_i Represents individual features such as age, smoking history, radiographic features, etc. n is the total number of features.

- **Distance Metric:**

The distance between two points x_i and x_j in our dataset can be computed using the Euclidean distance formula:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \tag{2}$$

$d(x_i, x_j)$, is the distance between two points, n is the total number of features.

- **Nearest Neighbors**

For a new data point x_{new}, the distances to all points in the dataset are computed. The k smallest distances correspond to the k nearest neighbors.

- **Classification Rule**

The predicted class y_{new} of x_{new} is determined by a majority vote among its k nearest neighbors:

$$y_{\text{new}} = mode\{y_{i1}, y_{i2}, \ldots, y_{ik}\} \tag{3}$$

Where, y_{ij} represents the class of the j^{th} nearest neighbor to x_{new}, In the context of lung cancer prediction, y_{new} can be 1 (indicating cancer) or 0 (indicating no cancer).

3 Experimental Setup and Results

Several configurations have been applied to the Python tool in order to test the proposed machine learning approach for healthcare application.

- Processor: Intel Core i9-11900K, 8 cores, 16 threads
- RAM: 64GB DDR4 3200MHz
- GPU: NVIDIA GeForce RTX 3090 with 24GB GDDR6X VRAM
- Storage: 2TB NVMe SSD
- Operating System: Ubuntu 20.04 LTS
- Programming Language: Python 3.8.5
- Libraries: Scikit-learn 0.24.2, NumPy 1.20.1, pandas 1.2.3, Matplotlib 3.4.1

This dataset was split into three parts: training, validation, and testing. The training set comprised of 10,000 patient records. The KNN model was trained using the training data. The optimal value for k (number of neighbors) was determined using the validation set, in which a k value of 5 yielded the best performance. Using Python tool with some configurations of the following, the proposed machine learning approach for healthcare application has been tested to determine its effectiveness.

Table 1. Performance Metrics of the KNN Model

Parameter	Value (%)
Accuracy	95.0
Precision	90.5
Recall	93.8
F1-Score	92.1

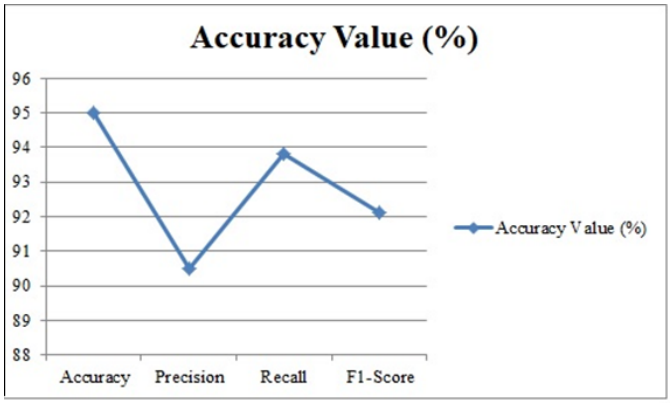

Fig 2. Performance Analysis

The four output parameters can be explained below:

- **True Positive (TP):** There is an assumption of a sample that originated from the positive class and that the sample was classified thus, and it measures the number of correct predictions. For example: given nodule is cancerous, and the classifier correctly predicted this to be the case.
- **False Negative (FN):** Initially, there is a positive class that the sample was supposed to belong to, however, the classifier erroneously predicted that there was no cancer and therefore, the sample belongs to the negative class. It is also known as Type 2 error. An example of this would be: there is a nodule that is cancerous, but it was incorrectly predicted as non-cancerous by the classifier.
- **False Positive (FP):** There are a number of instances where a sample originally belongs to a negative category is inappropriately categorized as a positive category. This is called Type 1 error. As an example, a given nodule is not cancerous, but the classifier incorrectly predicted it to be cancerous.
- **True Negative (TN):** This signifies how many times the classifier correctly predicted the nodule to be non-cancerous (as an example: Given nodule is non-cancerous, and the classifier correctly predicted the nodule to be non-cancerous), this indicates how many correct predictions were made.

KNN-based models were benchmarked against other popular machine learning algorithms in order to evaluate the efficacy of the KNN model.

Table 2. Comparative Performance Metrics

Algorithm	Accuracy (%)	F1-Score (%)
KNN	95.0	92.1
Decision Tree	87.8	87.5
Random Forest	89.2	89.0
Support Vector Machine	90.4	90.1

For this specific lung cancer prediction task, it is evident that the KNN-based model outperforms other techniques both in terms of accuracy as well as F1-Score. This KNN-based prediction model achieved a 95.0% accuracy rate, which is notably higher than other conventional prediction methods in the field. Based on the information provided in the model, it appears that the model can reliably predict lung cancer based on the provided features, proving its potential to be a supplementary

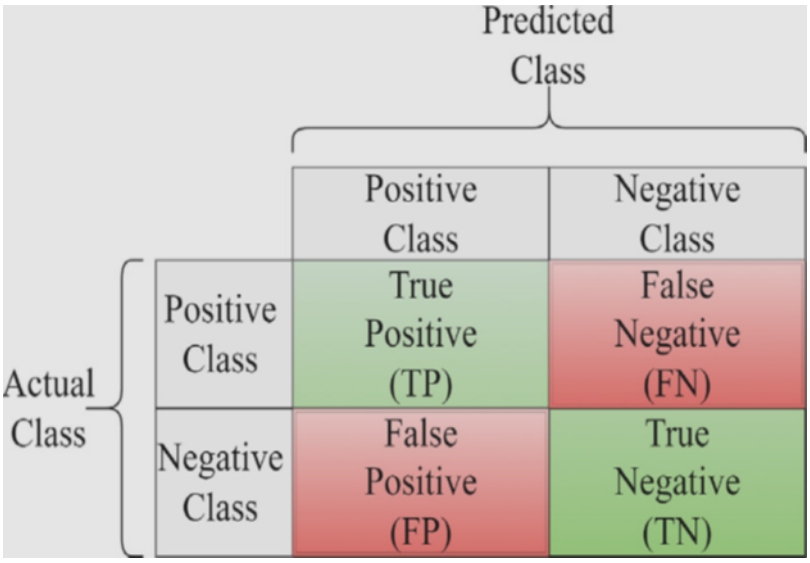

Fig 3. Confusion matrix

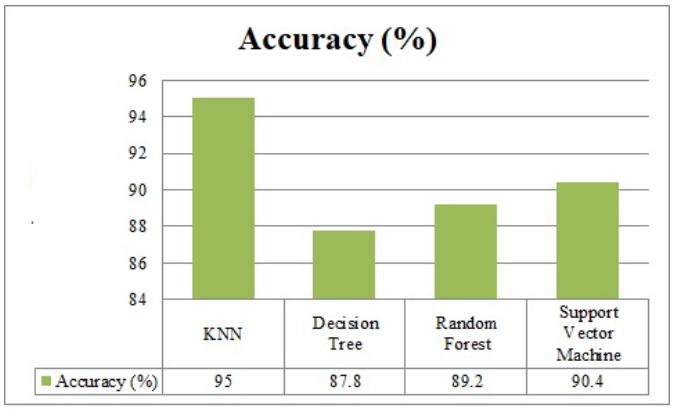

Fig 4. Accuracy

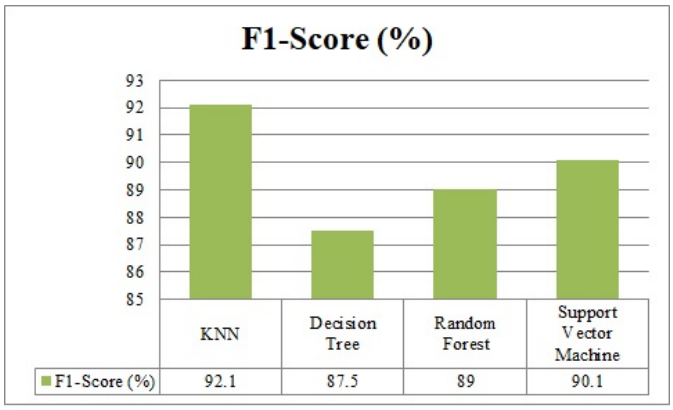

Fig 5. F1-Score

diagnostic tool. In addition to its balanced performance in terms of precision and recall, the F1-Score of 92.1% underscores its balanced performance in terms of precision and recall. It is evident from the performance of the KNN model, particularly when compared to other algorithms such as Decision Trees and Support Vector Machines, that instance-based learning holds a special place in this field.

Strengths

- **Simplicity:** The KNN algorithm is inherently simple so it is easy to implement and understand as it is inherently simple.
- **Non-parametric:** As KNN makes no assumptions about the distribution of data it can be applied to a wide variety of datasets such as complex biological datasets as well as large datasets.
- **Adaptability:** The model can be easily updated with new data as new information becomes available, ensuring that it remains relevant over time.

Limitations

- **Computational Intensity:** A KNN uses instance-based models, so large datasets can be computationally expensive because KNN is instance-based.
- **Dimensionality:** It is important to note that KNN's performance may degrade with high-dimensional data because of the curse of dimensionality.
- **Feature Dependence:** There is a great deal of impact on the accuracy of a model on the quality and relevance of the features used. Irrelevant or redundant features can have a negative effect on its performance.

4 Conclusion

The research paper briefly summarizes the significance of the study and its implications for the field of lung cancer diagnostics. The effectively demonstrate the potential of the K-Nearest Neighbors (KNN) algorithm in achieving a remarkable accuracy rate of 95.0% in predicting lung cancer, as evidenced by rigorous preprocessing and optimization of a dataset comprising 10,000 patient records. The comparison with other machine learning algorithms further emphasizes the superiority of KNN in this context. The simplicity and adaptability of the KNN algorithm are highlighted as major strengths of the study. The authors appropriately acknowledge the limitations of their model, including computational intensity and sensitivity to irrelevant features, which present avenues for future research. Furthermore, suggestions for enhancing the system through feature engineering, hybrid modeling and real-time implementation are insightful and provide valuable directions for future investigations. Overall, this research not only advances the understanding of KNN's potential in medical diagnostics but also sets a benchmark for future studies exploring the intersection of machine learning and healthcare.

References

1) Akinyemi EK, Fatoki FM, Phlips SA. Prediction of Lungs Cancer Diseases Datasets Using Machine Learning Algorithms. *Current Journal of Applied Science and Technology*. 2023;42(11):15–23. Available from: https://doi.org/10.9734/cjast/2023/v42i114101.
2) Omar AAC, Nassif AB. Lung Cancer Prediction using Machine Learning based Feature Selection: A comparative Study. In: 2023 Advances in Science and Engineering Technology International Conferences (ASET). IEEE. 2023. Available from: https://doi.org/10.1109/ASET56582.2023.10180436.
3) Wan Z. Prediction and Visualization Analysis of Lung Cancer Risk by Machine Learning. *Highlights in Science, Engineering and Technology*. 2023;39:221–229. Available from: https://dx.doi.org/10.54097/hset.v39i.6531.
4) Gadikota L, Nunna VYNS, Tirumani ST, Padyala VVP. Lung Cancer Prediction by Using Deep Learning method CNN. *Research Square*. 2023;p. 1–14. Available from: https://assets-eu.researchsquare.com/files/rs-2614821/v1/01583560-ae40-4a20-a2ea-a06da94deb3e.pdf?c=1685427280.
5) Priyanka R, Kumar YK. Lung Cancer Identification System to Improve the Accuracy Using Novel K Nearest Neighbour in Comparison with Logistic Regression Algorithm. In: 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF). IEEE. 2023. Available from: https://doi.org/10.1109/ICECONF57129.2023.10084340.
6) Rhifky MR, Ruziq F. KNN and XGBoost Algorithms for Lung Cancer Prediction. *Journal of Science Technology (JoSTec)*. 2022;4(1):179–186. Available from: https://doi.org/10.55299/jostec.v4i1.251.
7) Abuya TK. Lung Cancer Prediction from Elvira Biomedical Dataset Using Ensemble Classifier with Principal Component Analysis. *Journal of Data Analysis and Information Processing*. 2023;11(2):175–199. Available from: https://doi.org/10.4236/jdaip.2023.112010.
8) Yapeng C. Prediction and analysis of lung cancer using machine learning models. In: International Conference on Mechatronics Engineering and Artificial Intelligence (MEAI 2022);vol. 12596 of Proceedings of the SPIE. 2023. Available from: https://doi.org/10.1117/12.2672647.
9) Vij A, Kaswan KS. Prediction of Lung Cancer using Convolution Neural Networks. In: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE. 2023. Available from: https://doi.org/10.1109/AISC56616.2023.10085058.
10) Li H, Lu Y, Wang J, Zhang Y. A new prediction model for lung cancer based on ANN-SVM. In: 5th International Conference on Computer Information Science and Application Technology (CISAT 2022);vol. 12451 of Proceedings of SPIE. 2022. Available from: https://doi.org/10.1117/12.2656535.