Application of multi-algorithm approach for lung cancer prediction

Original Scientific Paper

Zulkifli Zulkifli*
Department of Informatics Engineering,
Faculty of Tehcnology and Informatics,
Aisyah University, Indonesia
zulkifli@aisyahuniversity.ac.id

Vira Weldimira
Department of Medicine, Faculty of Medicine,
Aisyah University, Indonesia
viraweldimira@aisyahuniversity.ac.id

Kraugusteeliana Kraugusteeliana
Information system Departement,
Universitas Pembangunan Nasional Veteran,
Jakarta Indonesia
kraugusteeliana@upnvj.ac.id

Fitriana Fitriana
Department of Midwifery,
Faculty of Health,
Aisyah University, Indonesia
fitriana@aisyahuniversity.ac.id

Ferly Ardhy
Department of Informatics Engineering,
Faculty of Tehcnology and Informatics,
Aisyah University, Indonesia
ferly@aisyahuniversity.ac.id

*Corresponding author

Abstract *– Lung cancer is one of the leading causes of cancer-related mortality worldwide, with most cases diagnosed at an advanced stage. Accurate and cost-effective early detection remains a major challenge due to the heterogeneity of imaging and histopathological features. Therefore, this study aimed to develop diagnostic software for lung cancer prediction using a multi-algorithm method. Patient data, including 16 clinical and lifestyle variables, were processed and analyzed with five machine learning algorithms, namely Neural Network (NN), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Random Forest (RF), and Naïve Bayes (NB). Model performance was evaluated based on accuracy, precision, recall, and F1-score. The results showed that RF, NB, SVM, and NN achieved perfect predictive performance (100% across all metrics), while k-NN obtained slightly lower but still high performance (99%). These findings signified that multi-algorithm predictive modeling could provide robust diagnostic support for lung cancer detection. The proposed software offered potential as an accessible, low-cost decision-support tool to assist clinicians in early diagnosis and improve patient outcomes.*

Keywords*: lung cancer, multi-algorithm, prediction, accuracy level*

Received: July 21, 2025; Received in revised form: October 20, 2025; Accepted: November 3, 2025

1. INTRODUCTION

Lung cancer, one of the deadliest forms of the disease, claims the lives of approximately one million people annually [1]. Lung cancer is the most prevalent form of cancer, after prostate cancer in men and breast cancer in women [2]. Lung cancer remains the leading cause of cancer-related deaths globally, with 2.09 million new cases and 1.76 million deaths reported in 2018 [3]. Lung cancer affects both smokers and non-smokers and is medically known as carcinoma. It originates in the epithelial cells, and when these cells mutate or grow uncontrollably, lung cancer can develop. Lungs, which are essential for respiration, are located on either side of the chest. The left lung is slightly smaller than the right to make room for the heart. During breathing, the chest rises and falls as lungs expand when inhaling and contract after exhaling [4]. Following the discussion, lungs are crucial in oxygenating the blood. The heart pumps oxygen-poor, carbon dioxide-rich blood to lungs, where it is "cleansed" by releasing carbon dioxide and absorbing oxygen. Carbon dioxide is expelled during exhalation, while oxygen is drawn into lungs during inhalation [5]. Cancer is one of the leading causes of death worldwide and is considered one of the most dangerous diseases known to humans. A major challenge in treating cancer is that it is often diagnosed at an advanced stage, making it difficult to cure. Among the various types, lung cancer accounts for a large proportion of cancer-related fatalities. As a result, extensive study has been undertaken to develop systems capable of detecting lung cancer at an early stage [6].

The significance of early lung cancer screening is increasingly acknowledged, as it greatly improves the possibility of early detection and treatment. However, even patient diagnosed at an early stage are still at risk of recurrence, which often leads to disease progression to advanced stages, significantly worsening the survival outlook [7]. The complexity of this cancer is due to pathogenesis and progression, characterized by intricate regulatory networks. Therefore, the initial cause of lung cancer, progression, chemotherapy resistance, resistance to targeted therapies, such as Epidermal Growth Factor Receptor (EGFR) and Anaplastic Lymphoma Kinase (ALK), and immune resistance need to be investigated [8]. For instance, early screening can be facilitated with the use of detection software with multi-algorithm approach. This software can provide diagnostic information that will aid in early detection and also save costs.

Several studies have explored the application of artificial intelligence in lung cancer diagnosis, including a 2021 study by Tafadzwa L. Chaunzwa, which employed deep learning to classify lung cancer histology using CT images. The study involved training and validating a Convolutional Neural Network (CNN) on a dataset of 311 early-stage NSCLC patient who underwent surgical treatment at Massachusetts General Hospital (MGH), focusing on Adenocarcinoma (ADC) and Squamous Cell Carcinoma (SCC), the two most prevalent histological types [9]. In developing software for early lung cancer screening, a comprehensive literature review will be conducted to examine studies involving multi-algorithm approaches, data mining, and machine learning applications. For instance, a 2024 study by Feda Anisah Makkiyah detailed the development of an application utilizing a multi-algorithm approach to predict diabetes status [10]. Similarly, a 2022 study by Nafseh Ghafar Nia evaluated artificial intelligence techniques in disease diagnosis and prediction, aiming to reduce diagnostic errors, minimize doctors' workload, and improve overall diagnostic accuracy [11].

Most previous analyses focused on single-algorithm implementations or study prototypes that lacked clinical usability despite these advances. Few studies have explored the incorporation of multiple algorithms into a practical, user-friendly diagnostic tool. Furthermore, many models have been evaluated on limited datasets or in highly controlled environments, which raises concerns about generalizability to real-world settings.

In addressing the gaps, this study develops and evaluates a multi-algorithm software system for lung cancer prediction. The system incorporates five different ML methods, namely NN, SVM, k-NN, RF, and NB, to provide comparative performance perceptions. By incorporating diverse patient variables comprising clinical as well as lifestyle factors, this study aims to develop a cost-effective, and accessible decision-support tool to assist clinicians in the early detection of lung cancer.

The remainder of this paper is structured as follows Section 2: Related Works, where we discuss previous study relevant to our study. Section 3: Materials and Methodology, detailing the methods used and the workflow of the study. Section 4: Results, where we present the findings from our experiments. Section 5: Conclusion, summarizing the key insights and contributions of the study.

2. RELATED WORKS

This study explores the use of deep learning radiomics to classify lung cancer histology from standard CT images, focusing on adenocarcinoma (ADC) and squamous cell carcinoma (SCC). Different from traditional biopsy-based methods or earlier radiomics relying on hand-crafted features, the process applies CNNs and makes a comparison with machine learning models trained on CNN-derived features. The best-performing model achieves an AUC of 0.71, with higher specificity than sensitivity, and external validation shows modest accuracy. Strengths of the study include showing non-invasive histology prediction, effective use of transfer learning, and interpretable visual outputs. However, limitations such as small sample size, class imbalance, low sensitivity, and reduced performance on heterogeneous data limit clinical applicability. The work provides proof-of-concept that deep learning radiomics can complement pathology in lung cancer diagnosis, and larger datasets are needed for validation [9]. Previous studies in lung cancer screening showed that while low-dose CT reduced mortality, it suffered from high false positives and variability among radiologists. Earlier CADe systems improved sensitivity and were limited by false alarms. More recent work with deep learning has advanced nodule detection, classification, malignancy prediction, and prognosis, often matching or surpassing radiologists. This study shows the novelty of the broader role of AI in detecting nodules, standardizing reporting, and predicting outcomes. Results signify improved accuracy, efficiency, and reproducibility, though challenges remain, including false positives, difficulty with complex nodules, limited generalizability, as well as incorporation into clinical practice. In general, AI shows strong potential to complement radiologists in lung cancer screening and requires further validation with larger datasets [4].

The study in [11] established low-dose CT (LDCT) as an effective tool for reducing lung cancer mortality, though its use was limited by high false-positive rates, heavy workloads, and variability among radiologists. Earlier computer-aided detection systems improved sensitivity and produced many false positives. Meanwhile, recent advances with AI and deep learning have shown superior performance in nodule detection, classification, malignancy prediction, as well as prognosis. The novelty of this study lies in the comprehensive review of the broader role of AI in LDCT screening, prioritizing the potential for detection, risk stratification, outcome prediction, and workflow optimization. Reported results show that AI can achieve high accuracy, reduce false positives, improve efficiency, and provide standardized reporting, in some cases matching or surpassing radiologists. The strengths of

this study are the demonstration of the wide applicability of AI and the potential to reduce diagnostic variability as well as workload. The weaknesses include persistent false positives, limited generalizability due to small or homogeneous datasets, a lack of large-scale clinical validation, and challenges in real-world integration. In general, the study shows that AI is a promising complement to radiologists in lung cancer screening, though further validation is needed. Previous study on cancer prediction explored a variety of machine and deep learning methods, such as logistic regression, artificial neural networks, support vector machines (SVM), decision trees, random forests (RF), and convolutional neural networks (CNN). These studies showed the potential of computational models in assisting early detection, though many faced challenges related to limited datasets, high computational costs, and reduced performance when scaled to larger data [12]. The novelty of this study was in the practical method of incorporating multiple cancers, including breast, lung, as well as prostate, into a single prediction framework, and applying different algorithms modified to each type. Specifically, SVM was used for breast cancer, and RF was applied to lung cancer as well as prostate cancer. The solutions tested included creating datasets with relevant attributes for each cancer type and developing a web-based interface where users could input personal data as well as receive prediction results. The study achieved promising results, where SVM effectively classified breast cancer as malignant or benign. RF provided reliable predictions for lung and prostate cancer based on symptoms as well as cell attributes [13]. The strengths of this study included the accessible interface, the use of established machine learning algorithms, and the focus on practical deployment for early detection. However, several weaknesses remained, including the restriction to offline use, lack of a database for storing patient information, and a limited scope that did not extend to other types of cancer or large-scale clinical validation. Future work should aim to expand the system with more cancer models, incorporate online databases, and test scalability in real-world healthcare settings [14].

A study by [15] proposed a hybrid optimization-based machine learning model for cancer prediction, addressing the limitations of earlier methods that lacked accuracy and generalizability. By testing SVM, RF, and deep learning models with optimization methods, the study showed improved accuracy and reduced error rates compared to baseline methods. Its strengths included innovation, versatility, and strong performance, while the weaknesses comprised limited datasets, high computational needs, as well as a lack of external validation. The study in general showed the potential of hybrid models for more reliable cancer prediction and required larger-scale validation for clinical use. [16] proposed a hybrid optimization-based machine learning model for cancer prediction, improving on earlier methods that suffered from low accuracy and poor generalization. By testing SVM, RF, and deep learning models, the study showed that the hybrid method achieved higher accuracy as well as fewer errors than traditional methods. Its strengths included innovation, flexibility across cancer types, and strong performance, while the weaknesses consisted of limited datasets, high computational costs, and a lack of clinical validation. The model, in general, showed promise for cancer prediction and required broader testing before practical use. Therefore, the current study aims to fill this gap through the use of software that incorporates 16 input variables. The variables include patient code, gender, age, smoking habits, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain.

Existing studies have shown the effectiveness of AI and machine learning in detecting or classifying lung cancer. However, most previous studies focused on single-algorithm implementations or on highly controlled experimental datasets, limiting real-world applicability. Many studies did not incorporate multiple algorithms in a unified diagnostic framework or provide comparative analysis across models. Several works also lacked scalability and clinical validation, restricting the use beyond study settings. This study addresses the gaps by developing a practical, multi-algorithm diagnostic software that simultaneously applies NN, SVM, k-Nearest Neighbor (k-NN), RF, and Naïve Bayes (NB) methods. By comparing predictive performance of the methods on clinical and lifestyle data, this study offers a more comprehensive and accessible decision-support tool for early lung cancer detection, bridging the gap between experimental models as well as real-world clinical implementation.

3. MATERIAL AND METHODOLOGY

The software for detecting lung cancer patient using a multi-algorithm method was developed through the following stages.

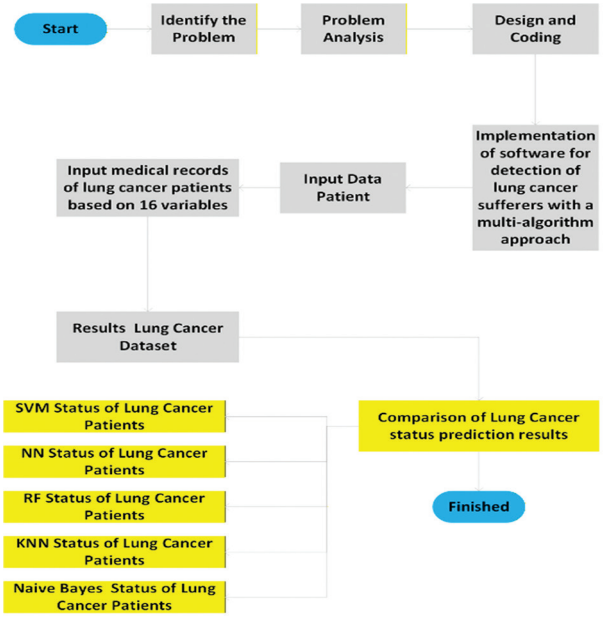

Fig. 1. Development stages of software formation for detecting lung cancer patient using multi-algorithm method

Fig. 1 showed the stages included in developing software for early lung cancer detection using a multi-algorithm method. The initial phases included problem identification, problem analysis, design, and coding. The implementation phase comprised two main processes, namely entering patient data and inputting medical records based on 16 variables related to lung cancer.

The output included lung cancer dataset and a comparative analysis of cancer status predictions using multiple algorithms, such as RF, NB, SVM, NN, as well as k-NN. Each algorithm provided a predicted lung cancer status for patient, allowing a comprehensive comparison of diagnostic accuracy across different models.

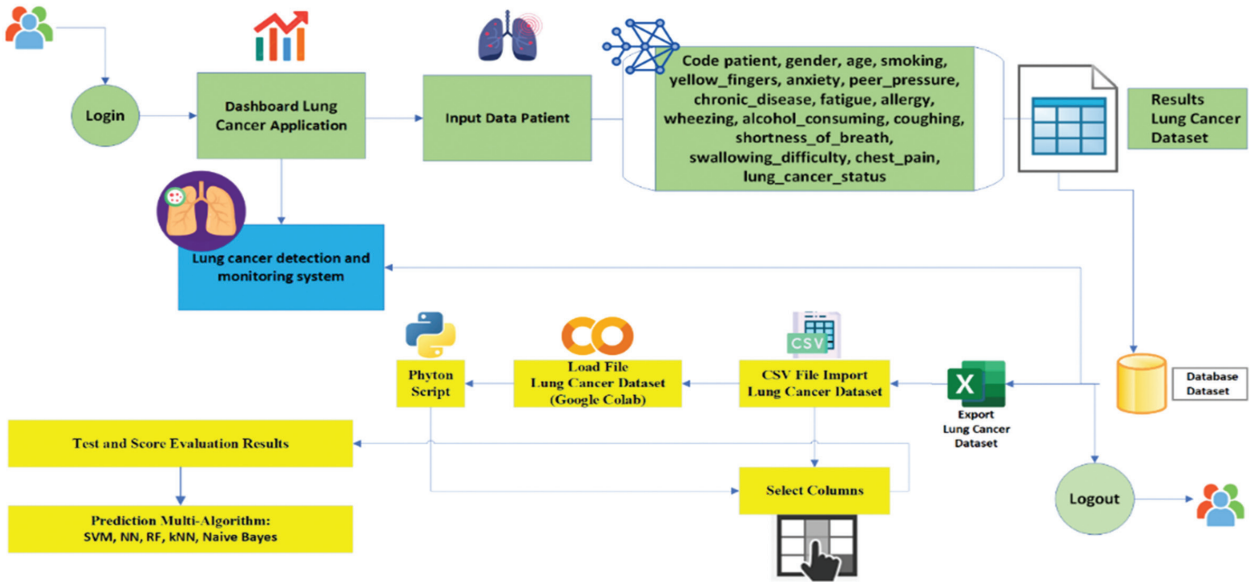

Fig. 2. Study framework for software development for lung cancer patient detection with multi-algorithm

Fig. 2 showed study framework for software development to detect lung cancer patient using multi-algorithm method. The framework offered a high level of study flexibility, as it could be developed according to specific needs, consisting of several stages. First, the user logged in, and upon successful login, the software dashboard would be entered. The user then input patient data, which included values for 16 lung cancer-related variables, namely patient code, gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, chest pain, and lung cancer status. Using these 16 data points, lung cancer status was generated through NB data processing. This output formed lung cancer dataset, which was stored in the database. After storage, the dataset was exported as a CSV file, which was then loaded into Google Colab for scripting with Python. This study used the Orange widget tools to assess accuracy during the analysis. Concerning the process mentioned earlier, multi-algorithm predictions were obtained using RF [17], NB [18], SVM [19], NN [20], and k-NN [21].

3.1. NEURAL NETWORK

Neural networks (NN) included several major formulas at different stages, such as initialization, forward propagation, activation functions, cost calculation, backward propagation, and parameter updates [22]. The following sentences comprised the essential formulas used during the analysis.

- Initialization weights $W^{[l]}$ and biases $b^{[l]}$ for layer l [23]:

$$W^{[l]} \sim \mathcal{N}(0, \sqrt{\frac{2}{n^{[l-1]}}}) \tag{1}$$

$$b^{[l]} = 0 \tag{2}$$

- Forward Propagation for a single layer l [24]:

$$A^{[l]} = \sigma(Z^{[l]}) \tag{3}$$

Where σ was the activation function ReLU [25]. For the output layer, the activation ay differed, e.g., Softmax for multi-class classification:

$$A_i^{[L]} = \frac{e^{Z_i^{[L]}}}{\sum_j e^{Z_j^{[L]}}} \tag{4}$$

- Activation Function is ReLU

$$\sigma(Z) = \max(0, Z) \tag{5}$$

- Cost Function for multi-class classification (Softmax), the cost was [26]:

$$J = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(A_k^{[L](i)}) \tag{6}$$

- Backward Propagation for output layer L [27]:

$$dZ^{[L]} = A^{[L]} - Y \tag{7}$$

- Parameter Update for each layer l:

$$W^{[l]} = W^{[l]} - \alpha . dW^{[l]} \quad (8)$$

$$b^{[l]} = b^{[l]} - \alpha . db^{[l]} \quad (9)$$

Where α is the learning rate.

3.2. SUPPORT VECTOR MACHINE

Concerning training vectors $xi \in Rp$, i=1,…,n, in two classes, and a vector $y \in$ {1,-1}n, this study aimed to find $w \in Rp$ and $b \in R$ considering the prediction given by sign ($wT\emptyset(x)+b$) was correct for most samples [28].

- SVM solved the following primal problem, where:

$$min_{w,b,\varsigma} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \quad (10)$$

$$\text{Subject to } y_i(w^T \emptyset(x_i) + b) \geq 1 - \varsigma_i, \quad (11)$$

$$\varsigma_i \geq 0, i = 1, \ldots, n \quad (12)$$

Intuitively, this study maximized the margin by minimizing, $||w||^2 = w^2 w$ while incurring a penalty when a sample was misclassified or in the margin boundary [29]. The value $y_i(wT\emptyset(x_i)+b)$ was ≥1 for all samples, which indicated a perfect prediction [29]. However, problems were often not perfectly separable with a hyperplane, allowing some samples to be at a distance ς_i from the correct margin boundary [30].

3.3. k-NEAREST NEIGHBOR

k-NN aimed to learn an optimal linear transformation matrix of size ($n_{components}$, $n_{features}$).

- Which maximized the sum over all samples i of the probability p_i that i was correctly classified [31], where:

$$\arg max_l \sum_{i=0}^{N-1} p_i \quad (13)$$

- With $N = n_{samples}$ and the probability of the sample being correctly classified according to a stochastic nearest neighbors rule in the learned embedded space [32].

$$p_i \sum_{j \in C_i} p_{ij} \quad (14)$$

- Where C_i was the set of points in the same class as sample i, and p_{ij} represented the softmax over Euclidean distances in the embedded space [33].

$$p_{ij} = \frac{\exp(-||L_{xi} - L_{xj}||^2)}{\sum_{k \neq i} \exp(-||L_{xi} - L_{xj}||^2)}, p_{ii} = 0 \quad (15)$$

3.4. RANDOM FOREST

RF algorithm was a combination of multiple decision trees, where each tree was constructed using a randomly selected subset of data and features.

- Each tree in the forest made its prediction, and the final output was determined through majority voting for classification tasks or by averaging for regression tasks [34], [17].

$$l(y) = \arg max_c \left(\sum_{n=1}^{N} I_{h_n(y)=c} \right) \quad (16)$$

Where I was the indicator function and H_N signified N Tree of RF. In addition, RF had an internal mechanism that provided an estimation of its generalization error called Out-of-Bag (OOB) Error Estimate.

3.5. NAIVE BAYES

NB is a supervised learning algorithm following Bayes' theorem under the naive assumption that all features were conditionally independent of the class variable [35].

- According to Bayes' theorem, the relationship between class variable y and dependent feature vector x_1 through x_n is as follows:

$$P(y | x_1, \ldots, x_n) = \frac{P(y) P(x_1, \ldots, x_n | y)}{P(x_1, \ldots, x_n)} \quad (17)$$

- Using the naive conditional independence assumption [36].

$$P(x_i | y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i | y), \quad (18)$$

- For all values of i, the relationship was simplified as follows [37].

$$P(y | x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i | y)}{P(x_1, \ldots, x_n)} \quad (19)$$

- Since $P(x_1, \ldots, x_n)$ was constant given the input, the following classification rule was used [38].

$$P(y | x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i | y)$$
$$\Downarrow \quad (20)$$
$$\hat{y} = \arg max_y P(y) \prod_{i=1}^{n} P(x_i | y)$$

$P(y)$ and $P(x_i | y)$ was estimated using Maximum A Posteriori (MAP). The $P(y)$ was determined based on the relative frequency of class y in the training set [39]. The various NB classifiers differed in the assumptions made regarding the distribution of $P(x_i | y)$ [40].

4. RESULT

The stages of software development for detecting lung cancer patient using a multi-algorithm method included generating a dataset by inputting patient data and medical records based on 16 variables. Lung cancer status predictions were measured using algorithms such as RF, NB, SVM, NN, and k-NN. Following the process, problem identification was conducted during this stage. The identified issue was the need to develop a system capable of detecting lung cancer patient using a multi-algorithm method that incorporated 16 input variables. These included patient code, gender, age,

smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain, along with one output variable, named lung cancer status.

4.1. PROBLEM IDENTIFICATION

Problem identification was conducted at this stage during the analysis including:

- The need for software to detect lung cancer patient using multi-algorithm method. This was to be used in decision-making or policy implementation for lung cancer patient services.
- The developed model should apply to health services.
- Measuring the accuracy of the cancer status prediction in lung cancer patient using multi-algorithm, namely RF, NB, SVM, NN, and k-NN.

4.2. PROBLEM ANALYSIS

Tool user communicated to understand the software expected by user, both doctors and hospitals.

4.3. SOFTWARE DESIGN FOR DETECTING LUNG CANCER PATIENT USING MULTI-ALGORITHM

Software for detecting lung cancer patient with multi-algorithm method was developed on Android. Fig. 3 shown some representation of the software on Android.

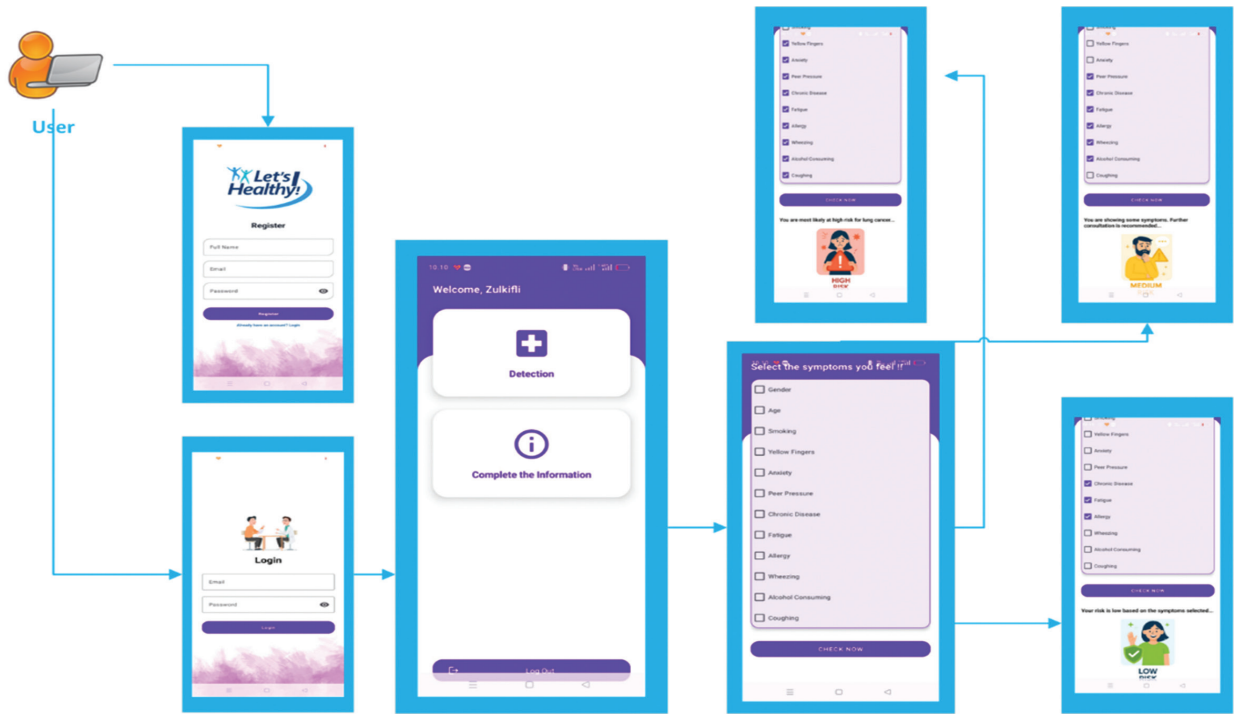

Fig. 3. Software presentation for lung cancer patient detection with multi-algorithm method

4.4. DATASET DATABASE

The details and explanation of the dataset shown in Table 1 [41] were used to measure the accuracy of lung cancer detection through various algorithms, including RF, NB, SVM, NN, as well as k-NN. This data was sourced from Kaggle.com and originally contained 3,000 data points, with only 10 data entries shown in Table 2.

4.5. APPLICATION OF NEURAL NETWORK ALGORITHM

Table 3 shows an example of a dataset that has been normalized during analysis, we used the Standard Scaler technique when normalizing the data. The formula used to convert the original test data ranged from 0.1 to 0.9 because the activation function used was sigmoid with a value above 0 [42].

4.6 MULTI-ALGORITHM PERFORMANCE

The dataset was divided into two parts, containing 90% training and 10% testing, with stratification as well as random states equaling 2. The performance of the RF, NB, SVM, NN, and k-NN algorithms was shown in Table 4.

The confusion matrix in the earlier image showed that the classification model performed well in detecting lung cancer. Among the total test data, 148 cases were correctly identified as not having lung cancer (true negatives), and 152 cases were correctly detected with it (true positives). There were no prediction errors during the process, in the form of false positives or negatives. This signified that the model achieved 100% accuracy, precision, recall, and F1-score. In other words, the model perfectly distinguished patients with lung cancer from those without in the test data used.

Table 4 showed the performance of five ML algorithms, namely RF, NB, SVM, NN, k-NN evaluated using accuracy, precision, recall, and F1-score metrics. The results signified that RF, NB, SVM, and NN each achieved a perfect score of 1.00 across all metrics. This indicated that the four algorithms classified the data without any error. Meanwhile, k-NN showed a slight decrease in performance with accuracy, precision, recall, and F1-score values of 0.99. Despite the minor decrease, k-NN still showed a very high and nearly perfect performance. The performance evaluation results were shown by the confusion matrix in Fig. 4. The results of the performance evaluation were visualized using three-dimensional (3D) TSNe.

Table 1. Details of the dataset

Variable	Association with Lung Cancer	Explanation
Smoking	Very Strong	The primary cause of lung cancer. Around 85–90% of all lung cancer cases were connected to smoking. Tobacco smoke contains carcinogens that damage lung cells over time
Age	Moderate to Strong	Risk increased significantly with age, since DNA damage accumulated over time. Most lung cancer cases occurred in people aged 55 and older
Gender	Weak to Moderate	Men historically had a higher risk due to smoking rates, but now gender differences are narrowing. Biological differences might play a minimal role
Yellow fingers	Indirect indicator	Often a sign of heavy smoking, not a direct cause. It was a proxy variable that signaled nicotine exposure
Chronic disease	Moderate	Chronic lung inflammation and damage increase susceptibility to cancer. Often comorbid with smoking
Coughing, wheezing, shortness of breath, chest pain, and swallowing difficulty	Symptoms, not causes	These showed possible existing lung damage or disease, not risk factors
Anxiety, peer pressure, fatigue, allergy, and alcohol consumption	Weak or indirect	Minimal or no direct causal relationship with lung cancer. Peer pressure could indirectly lead to smoking behavior

Table 2. Dataset of patient with lung cancer

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
P-0001	1	65	1	1	1	2	2	1	2	2	2	2	2	2	1	Y
P-0002	0	55	1	2	2	1	1	2	2	2	1	1	1	2	2	Y
P-0050	1	38	2	2	1	1	2	2	1	2	1	2	2	2	1	Y
P-0117	0	52	2	2	1	2	1	2	1	1	2	1	2	1	2	Y
P-0118	1	34	2	2	2	2	1	1	2	2	2	2	1	1	1	N
P-0158	1	54	2	1	2	2	2	2	2	2	2	1	1	1	1	N
P-0159	1	78	2	2	1	1	2	1	2	1	2	2	1	2	1	Y
P-0211	1	30	1	1	2	2	1	2	2	1	2	2	2	1	2	N
P-2999	1	40	1	2	2	1	2	2	2	2	2	1	2	1	2	Y
P-3000	1	54	2	1	2	2	1	1	2	2	2	1	1	1	1	Y

Note: a=code patient, b=gender, c=age, d=smoking, e=yellow fingers, f=anxiety, g=peer pressure, h=chronic disease, i=fatigue, j=allergy, k=wheezing, l=alcohol consuming, m=coughing, n=shortness of breath, o=swallowing difficulty, p=chest pain, q= lung cancer status

Table 3. Detailed confidence levels of the proposed model against the actual stunting status

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
P-0001	0,8	0,56	0	0	0	0,8	0,8	0	0,8	0,8	0,8	0,8	0,8	0,8	0	Y
P-0002	0	0,4	0	0,8	0,8	0	0	0,8	0,8	0,8	0	0	0	0,8	0,8	Y
P-0050	0,8	0,128	0,8	0,8	0	0	0,8	0,8	0	0,8	0	0,8	0,8	0,8	0	Y
P-0117	0	0,352	0,8	0,8	0	0,8	0	0,8	0	0	0,8	0	0,8	0	0,8	Y
P-0118	0,8	0,064	0,8	0,8	0,8	0,8	0	0	0,8	0,8	0,8	0,8	0	0	0	N
P-0158	0,8	0,048	0	0	0	0,8	0,8	0,8	0,8	0,8	0,8	0	0	0	0	N
P-0159	0,8	0,384	0,8	0	0,8	0,8	0,8	0	0,8	0	0,8	0,8	0	0,8	0	Y
P-0211	0,8	0,768	0,8	0,8	0	0	0	0,8	0,8	0	0,8	0,8	0,8	0	0,8	N
P-2999	0,8	0	0	0	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0	0,8	0	0,8	Y
P-3000	0,8	0,16	0	0,8	0,8	0	0	0	0,8	0,8	0,8	0	0	0	0	Y

Note: a=code patient, b=gender, c=age, d=smoking, e=yellow fingers, f=anxiety, g=peer pressure, h=chronic disease, i=fatigue, j=allergy, k=wheezing, l=alcohol consuming, m=coughing, n=shortness of breath, o=swallowing difficulty, p=chest pain, q= lung cancer status

Table 4. Multi-algorithm performance

1	2	3	4	5
RF	1.00	1.00	1.00	1.00
NB	1.00	1.00	1.00	1.00
SVM	1.00	1.00	1.00	1.00
NN	1.00	1.00	1.00	1.00
k-NN	0.99	0.99	0.99	0.99

Note: 1=Algorithm, 2= Precision, 3= F1-Score, 4= Recall, 5= Accuracy

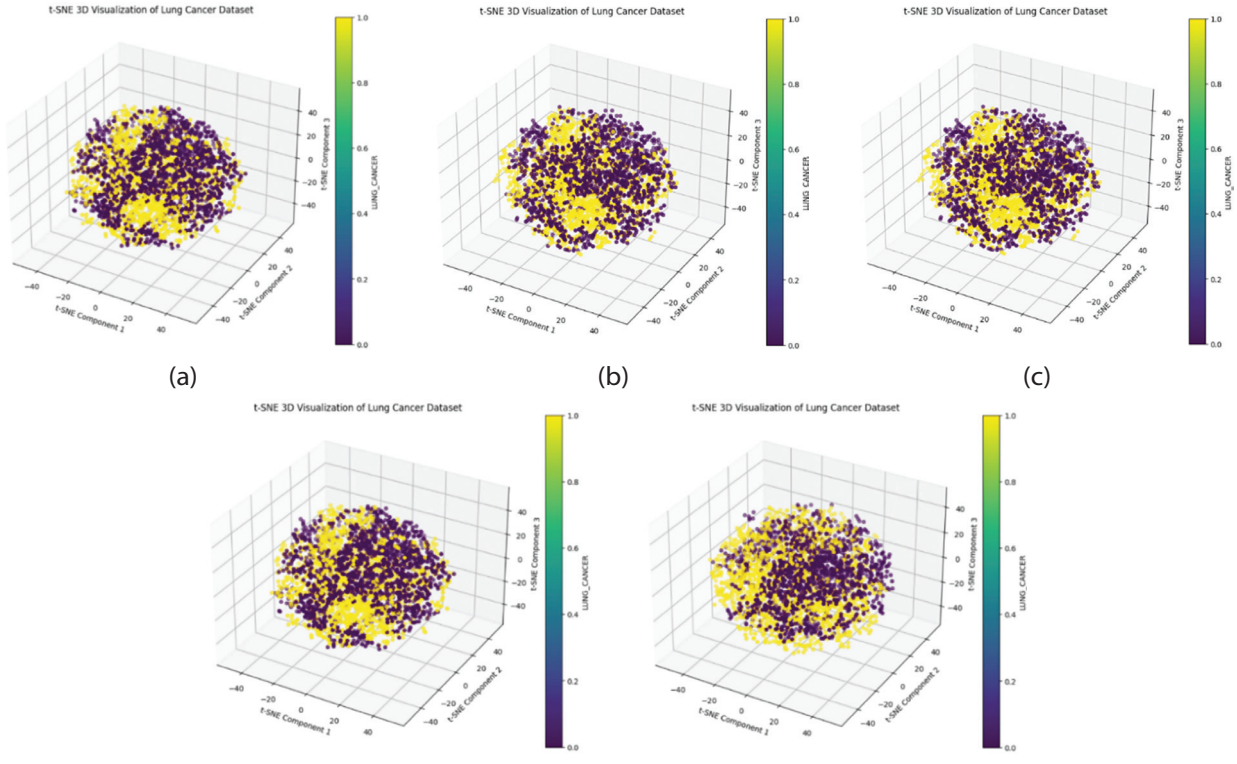

Fig. 5. 3D visualization of TSNe models (a) RF, (b) NB, (c) SVM, (d) NN, and (e) k-NN

Fig. 5 was a 3D TSNe visualization of lung cancer dataset, showing the distribution of data in 3D space based on the three principal components generated from TSNe method.

The data was represented as dots in two colors, namely purple and yellow. Purple dots (value 0) represented patient without lung cancer, while yellow dots (value 1) indicated patient diagnosed with lung cancer, respectively.

4.7. RECEIVER OPERATING CHARACTERISTICS (ROC) ANALYSIS

ROC was used to explain, assess, and categorize the performance of RF, NB, SVM, NN, and k-NN algorithms [43]. Fig. 6 showed ROC curves for these algorithms with each curve representing performance across the target classes "Survived" and "Died". The curves were shown using different colors, namely cyan, orange, blue, purple, and green lines, respectively.

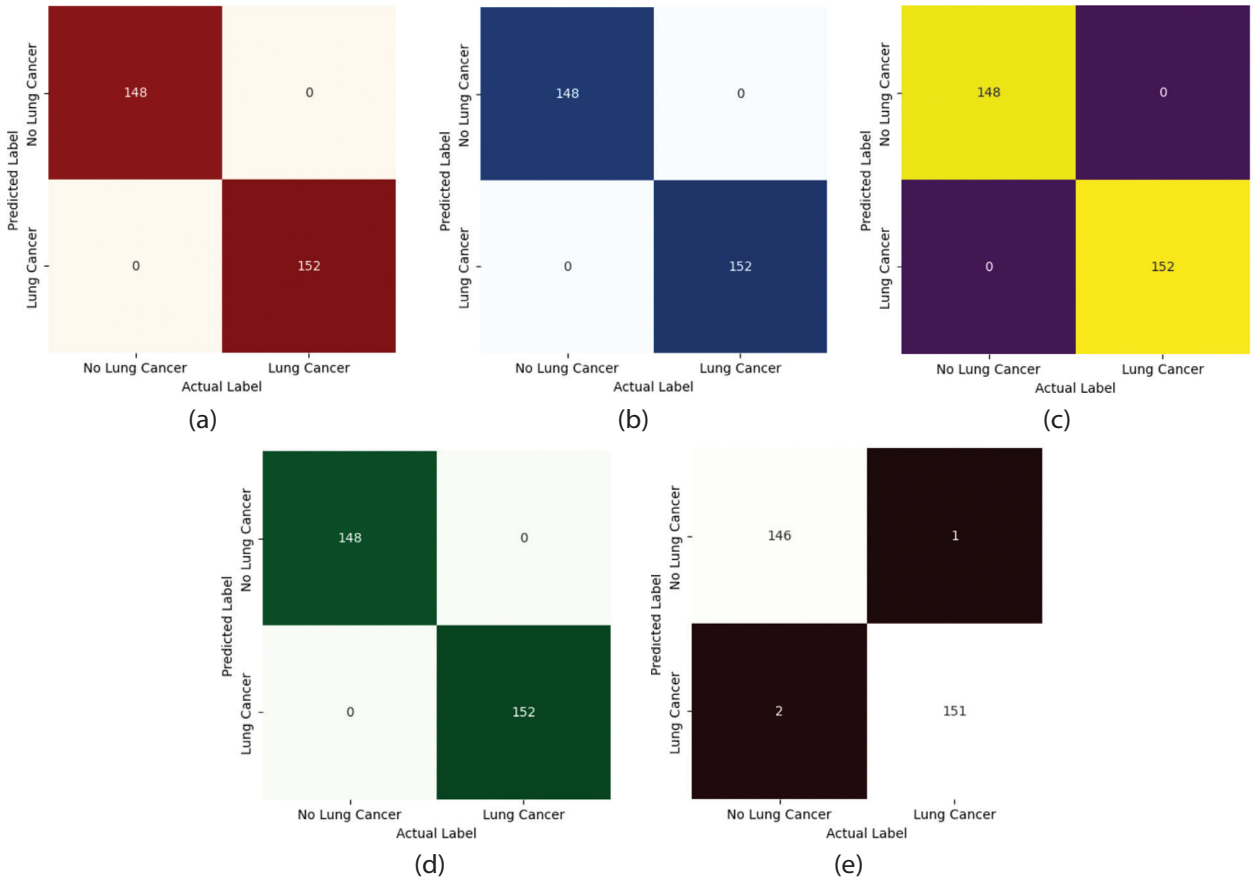

Fig. 4. Confusion matrix visualization (a) RF, (b) NB, (c) SVM, (d) NN, and (e) k-NN

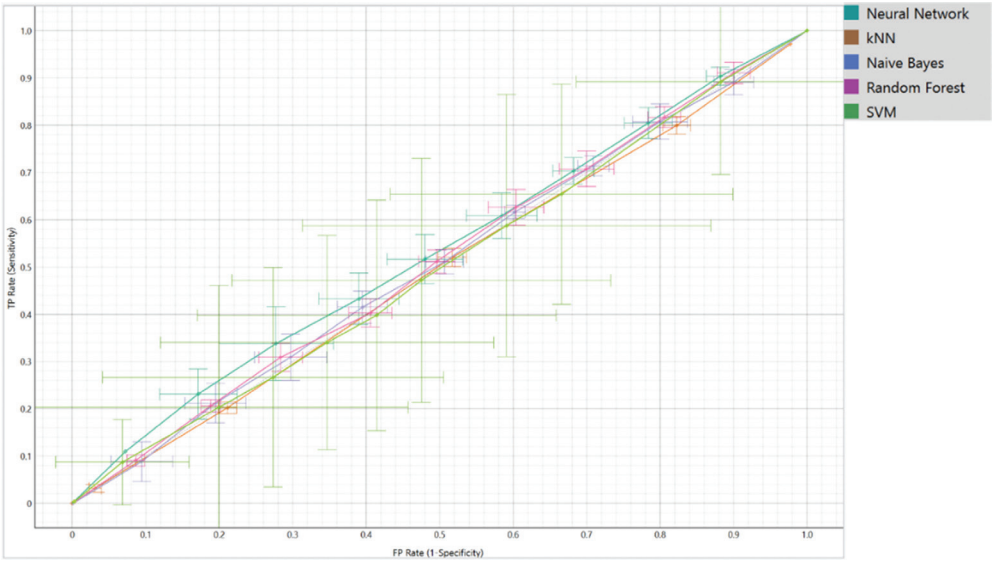

Fig. 6. Performance Curves of RF, NB, SVM, NN, and k-NN Algorithms

Table 5. Validation results of lung cancer patient detection predictions between neural RF, NB, SVM, NN, and k-NN algorithm

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P-0001	0	0	0	1	0	0	0.580	N	0.600	Y	0.506	Y	0.527	Y	0.500	Y
P-0002	0	0	1	1	1	0	0.668	N	0.600	Y	0.506	Y	0.654	Y	0.500	Y
P-0050	1	1	0	1	1	0	0.505	Y	0.600	N	0.585	Y	0.606	Y	0.500	Y
P-0117	1	0	1	1	1	0	0.618	N	0.600	Y	0.511	Y	0.572	Y	0.500	Y
P-0118	0	0	0	1	0	0	0.757	N	0.600	N	0.506	Y	0.712	N	0.500	Y
P-0158	0	1	1	0	1	1	0.701	Y	0.800	Y	0.531	N	0.565	Y	0.507	Y
P-0159	1	1	0	0	1	1	0.755	Y	0.800	N	0.516	Y	0.643	Y	0.500	Y
P-0211	0	1	1	1	1	1	0.810	Y	0.800	Y	0.533	Y	0.616	Y	0.532	Y
P-2999	1	0	1	0	1	1	0.539	N	0.600	Y	0.546	N	0.655	Y	0.500	Y
P-3000	1	0	1	0	0	1	0.735	N	0.600	Y	0.521	Y	0.681	N	0.536	Y

Note: 1=code patien, 2=actual, 3=NN, 4= k-NN, 5= NB, 6= RF, 7=svm, 8= NN numerical, 9=NN validation against actual, 10= k-NN numerical, 11= validation of k-NN against actual, 12=NB numerical, 13=validation of NB against actual, 14=RF numerical, 15=RF validation against actual, 16=SVM numerical, 17=SVM validation against actual Note: 1=code patien, 2=actual, 3=NN, 4= k-NN, 5= NB, 6= RF, 7=svm, 8= NN numerical, 9=NN validation against actual, 10= k-NN numerical, 11= validation of k-NN against actual, 12=NB numerical, 13=validation of NB against actual, 14=RF numerical, 15=RF validation against actual, 16=SVM numerical, 17=SVM validation against actual

4.8. COMPARISON OF NN ALGORITHMS, RF, NB, SVM, NN, AND K-NN ON PREDICTING LUNG CANCER

Fig. 7 showed a comparison of RF, NB, SVM, NN, and k-NN algorithms for detecting lung cancer patient.

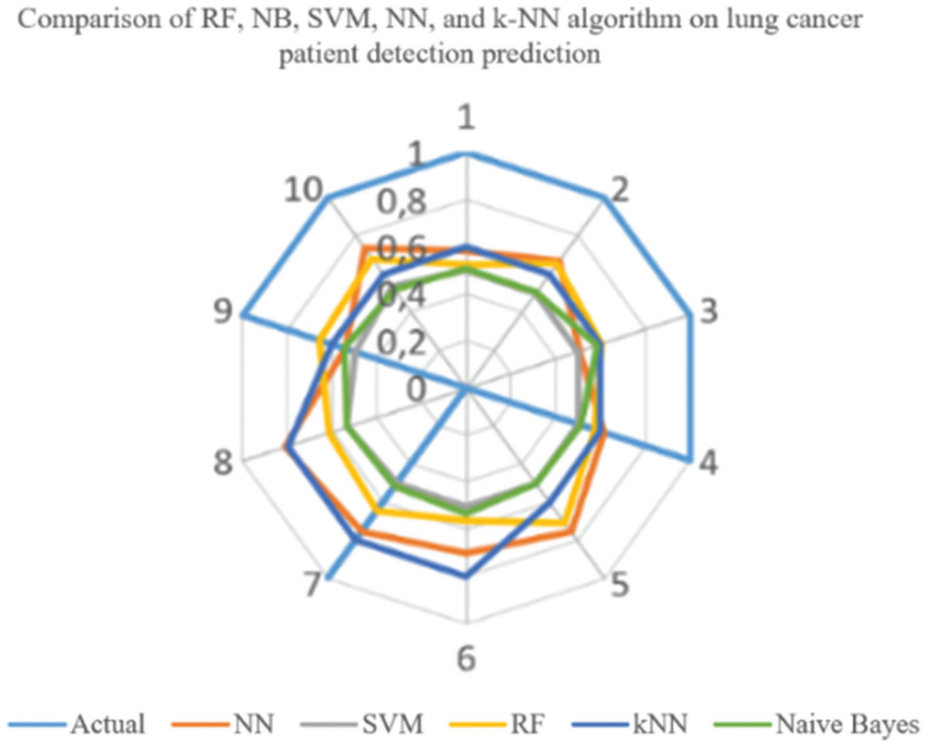

Fig. 7. Comparison of the Prediction Accuracy Performance for Lung Cancer Patient Detection using RF, NB, SVM, NN, and k-NN

5. DISCUSSION

This study evaluated the performance of five machine learning algorithms, namely, RF, NB, SVM, NN, and k-NN, for lung cancer prediction. The results showed that RF, NB, SVM, and NN achieved a perfect accuracy of 1.00, while k-NN obtained slightly lower accuracy of 0.99.

The high performance reported supported a previous study showing the suitability of ensemble and kernel-based methods for cancer diagnosis. For instance, [9] achieved strong classification accuracy using CNNs for lung cancer histology. [12] reported that incorporating machine learning classifiers with multi-attribute decision-making produced high performance in distinguishing benign from malignant lung X-rays. In many cases, tree-based methods such as RF excelled because the models captured nonlinear feature interactions and handled categorical as well as numerical variables effectively. The outcome might explain the perfect accuracy of RF in this study.

NB achieved near-perfect results, which were unexpected given its simplifying assumption of feature independence. However, previous studies [11] showed that NB performed competitively in medical diagnosis when datasets were relatively small or structured with limited feature interdependencies. SVM and NN achieved comparable performance, consistent with previous studies on lung and breast cancer prediction, where the ability to model complex nonlinear decision boundaries proved beneficial [14].

The comparatively lower performance of k-NN (though still at 99%) could be explained by its sensitivity to feature scaling and the curse of dimensionality. Having 16 input variables, the distances between samples might become less discriminative, leading to reduced performance compared to algorithms that modeled feature interactions more explicitly. Previous studies in medical prediction tasks observed strong baseline performance of k-NN but lower robustness compared to ensemble and kernel methods.

Achieving near-perfect accuracy across multiple algorithms was unusual in real-world biomedical datasets and raised concerns about potential overfitting, limited dataset diversity, or issues in ground-truth labeling. Different from previous analyses, such as [13], which validated the lung cancer classification models on independent test cohorts, this study did not describe the validation procedure. While the comparative trends between algorithms were consistent with the literature, the absolute performance values might be inflated.

As the results of this study showed promising potential for a multi-algorithm method in lung cancer prediction, several limitations should be acknowledged. First, the dataset used was obtained from a publicly available source (Kaggle) and contained 3,000 records. Although the size was sufficient for proof-of-concept analysis, it might not adequately represent the variability observed in real-world clinical populations. The dataset might not capture diverse demographic, genetic, and environmental factors that influence lung cancer incidence, limiting the generalizability of the findings to broader patient populations.

Second, the study lacked a clear description of how data were partitioned for training and testing. Without independent test sets or cross-validation procedures, the reported near-perfect accuracy might be inflated due to overfitting or data leakage. This concern was heightened by the unusually high performance across multiple algorithms, which was rarely observed in complex medical prediction tasks.

Third, the ground-truth labels were not independently validated against clinical or pathological standards. In some cases, labels appeared to be generated through algorithmic preprocessing (e.g., NB), which might bias performance metrics. Additionally, issues such as duplicate patient identifiers and inconsistent labeling in the dataset raised further questions about data quality as well as reliability.

Fourth, the study focused only on structured tabular data derived from 16 predefined variables. This method did not incorporate imaging, genomic, or longitudinal clinical data, even though the features were often crucial for accurate lung cancer diagnosis. Future studies should validate the software on larger, more diverse, and clinically verified datasets, while also incorporating multimodal data sources to improve predictive robustness as well as generalizability.

6. CONCLUSION

In conclusion, software was developed to detect lung cancer patient by predicting lung cancer status using multiple algorithms. The algorithms used to measure the accuracy of the predictions included RF, NB, SVM, NN, and k-NN. This software was designed for use by doctors, requiring the input of 16 patient variables related to lung cancer. Based on the dataset, the software predicted lung cancer status of each patient. The results showed that RF, NB, SVM, and NN each achieved an accuracy of 100%, while k-NN closely followed with 99%. Among the five algorithms, k-NN showed the lowest performance, despite the fact that it was still highly accurate.

This study had limitations related to dataset size, potential overfitting, and restricted clinical scope despite the promising results. Future studies should validate the model on larger and more diverse datasets, incorporate additional clinical and imaging variables, as well as apply rigorous cross-validation to ensure reliability. Expanding the system into a clinician-friendly decision-support tool with explainable outputs would further improve its practical value in real-world healthcare settings.

REFERENCES:

[1] S. Nageswaran et al. "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing", BioMed Research International, Vol. 2022, 2022.

[2] P. Sathe, A. Mahajan, D. Patkar, M. Verma, "End-to-End Fully Automated Lung Cancer Volume Estimation System", International Journal of Electrical and Computer Engineering Systems, Vol. 15, No. 8, 2024, pp. 651-661.

[3] A. Shimazaki et al. "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method", Scientific Reports, Vol. 12, No. 1, 2022, pp. 1-10.

[4] S. M. Lee, C. M. Park, "Application of artificial intelligence in lung cancer screening", Journal of the Korean Society of Radiology, Vol. 80, No. 5, 2022, pp. 872-879.

[5] L. Karunakaran, "Machine Learning Approach Of Lung Cancer Prediction Using Multi Machine Learning Models", StudyGate Data Science Track ComURS2025 Computing Undergraduate Study Symposium, Sabaragamuwa University of Sri Lanka, Belihuloya, March 2025.

[6] A. A. Shah, H. A. M. Malik, A. H. Muhammad, A. Alourani, Z. A. Butt, "Deep learning ensemble 2D CNN approach towards the detection of lung cancer", Scientific Reports, Vol. 13, No. 1, 2023, pp. 1-15.

[7] H. Sharma, J. S. Jain, P. Bansal, S. Gupta, "Feature extraction and classification of chest X-ray images using CNN to detect pneumonia", Proceedings of the 10th International Conference on Cloud Computing, Data Science & Engineering, Noida, India, 29-31 January 2020, pp. 227-231.

[8] D. Huang, Z. Li, T. Jiang, C. Yang, N. Li, "Artificial intelligence in lung cancer: current applications, future perspectives, and challenges", Frontiers in Oncology, Vol. 14, 2024, pp. 1-19.

[9] T. L. Chaunzwa et al. "Deep learning classification of lung cancer histology using CT images", Scientific Reports, Vol. 11, No. 1, 2021, pp. 1-13.

[10] Z. Zulkifli, F. A. Makkiyah, D. Antoni, F. Fitriana, T. Jamaan, A. Taufik, "Multi-Algorithm to Measure the Accuracy Level of Diabetes Status Prediction", Journal of Applied Data Sciences, Vol. 5, No. 2, 2024, pp. 736-746.

[11] N. Ghaffar Nia, E. Kaplanoglu, A. Nasab, "Evaluation of artificial intelligence techniques in disease diagnosis and prediction", Discover Artificial Intelligence, Vol. 3, No. 1, 2023.

[12] T. Meeradevi, S. Sasikala, L. Murali, N. Manikandan, K. Ramaswamy, "Lung cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model", Scientific Reports, Vol. 15, No. 1, 2025, pp. 1-19.

[13] S. Alazwari, J. Alsamri, M. M. Asiri, M. Maashi, S. A. Asklany, and A. Mahmud, "Computer-aided diagnosis for lung cancer using waterwheel plant algorithm with deep learning", Scientific Reports, Vol. 14, No. 1, 2024, pp. 1-14.

[14] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra, N. Sharma, "Cancer Prediction using Machine Learning", Proceedings of the 2nd International Conference on Innovative Practices in Technology and Management, Gautam Buddha Nagar, India, 23-25 February 2022, pp. 217-221.

[15] S. R. Abdani, N. Jamil, S. M. Z. Syed Zainal Ariffin, S. Ibrahim, "3D-based Convolutional Neural Networks for Medical Image Segmentation: A Review", International Journal of Electrical and Computer Engineering Systems, Vol. 16, No. 5, 2025, pp. 347-363.

[16] M. Phogat, D. Kumar, "A Hybrid Metaheuristics based technique for Mutation Based Disease Classification", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 6, 2023, pp. 635-646.

[17] E. Dritsas, M. Trigka, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction", Sensors, Vol. 22, No. 14, 2022.

[18] S. A. Ali, N. R. Roy, D. Raj, "Software Defect Prediction using Machine Learning", Proceedings of the 10th International Conference on Computing for Sustainable Global Development, New Delhi, India, 15-17 March 2023, pp. 639-642.

[19] I. Aydin, M. Karaköse, E. Akin, "Artificial immune based support vector machine algorithm for fault diagnosis of induction motors", Proceedings of

the International Aegean Conference on Electrical Machines and Power Electronics, Bodrum, Turkey, 10-12 September 2007, pp. 217-221.

[20] S. S. Cross, R. F. Harrison, R. L. Kennedy, "Introduction to neural networks", Lancet, Vol. 346, No. 8982, 1995.

[21] S. Goyal, "Handling Class-Imbalance with KNN (Neighbourhood) Under-Sampling for Software Defect Prediction", Artificial Intelligence Review, Vol. 55, No. 3, 2022, pp. 2023-2064.

[22] J. Díaz-Ramírez, "Machine Learning and Deep Learning", Ingeniare, Vol. 29, No. 2, 2021, pp. 182-183.

[23] M. A. Ijaz, M. K. Abid, N. Aslam, A. Q. Mudaseer, "Detecting Monkeypox in humans using deep learning", VFAST Transactions on Software Engineering, Vol. 11, No. 2, pp. 265-272, 2023.

[24] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P. R. Pinheiro, "CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection", IEEE Access, Vol. 8, 2020, pp. 91916-91923.

[25] A. Bhat, U. Bhardwaj, M. Singla, K. Garg, "Prediction of COVID 19 Using Chest X-Ray Images through CNN optimised using Genetic Algorithm", Proceedings of the 2nd International Conference on Intelligent Technologies, Hubli, India, 24-26 June 2022, pp. 1-8.

[26] R. T. N. V. S. Chappa, M. El-Sharkawy, "Squeeze-and-Excitation SqueezeNext: An Efficient DNN for Hardware Deployment", Proceedings of the 10th Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 6-8 January 2020, pp. 691-697.

[27] S. M. Fati, E. M. Senan, N. ElHakim, "Deep and Hybrid Learning Technique for Early Detection of Tuberculosis Based on X-ray Images Using Feature Fusion", Applied Sciences, Vol. 12, No. 14, 2022, p. 7092.

[28] A. Thakkar, D. Mungra, A. Agrawal, K. Chaudhari, "Improving the Performance of Sentiment Analysis Using Improved Preprocessing Technique and Artificial Neural Network", IEEE Transactions on Affective Computing, Vol. 13, No. 4, 2022, pp. 1771-1782.

[29] A. K. Jean, M. Diarra, B. A. Bakary, G. Pierre, A. K. Jérôme, U. B. Franche-comté, "Application based on Hybrid CNN-SVM and PCA- SVM Approaches for Classification of Cocoa Beans", International Journal of Advanced Science and Computer Applications, Vol. 13, No. 9, 2022, pp. 231-238.

[30] A. Fadli, Y. Ramadhani, M. S. Aliim, "Purwarupa Sistem Deteksi COVID-19 Berbasis Website Menggunakan Algoritma CNN", Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), Vol. 5, No. 5, 2021, pp. 876-883.

[31] Ž. Vujović, "Classification Model Evaluation Metrics", International Journal of Advanced Science and Computer Applications, Vol. 12, No. 6, 2021, pp. 599-606.

[32] D. Wang, H. Yu, D. Wang, G. Li, "Face recognition system based on CNN", Proceedings of the International Conference on Computer Information and Big Data Applications, Guiyang, China, 17-19 April 2020, pp. 470-473.

[33] F. H. Bitew, C. S. Sparks, S. H. Nyarko, "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia", Public Health Nutrition, Vol. 25, No. 2, 2022, pp. 269-280.

[34] F. Aldi, I. Nozomi, R. B. Sentosa, A. Junaidi, "Machine Learning to Identify Monkey Pox Disease", Sinkron, Vol. 8, No. 3, 2023, pp. 1335-1347.

[35] P. Singh, S. Verma, "Multi-classifier model for software fault prediction", International Arab Journal of Information Technology, Vol. 15, No. 5, 2018, pp. 912-919.

[36] W. Tan et al. "Classification of COVID-19 pneumonia from chest CT images based on reconstructed super-resolution images and VGG neural network", Health Information Science and Systems, Vol. 9, No. 1, 2021, pp. 1-12.

[37] S. H. Basha, A. M. Anter, A. E. Hassanien, A. Abdalla, "Hybrid intelligent model for classifying chest X-ray images of COVID-19 patients using genetic algorithm and neutrosophic logic", Soft Computing, Vol. 27, 2023, pp. 3427-3442.

[38] S. S. Shreem, H. Turabieh, S. Al Azwari, F. Baothman, "Improved binary genetic algorithm as a feature selection to predict student performance", Soft Computing, Vol. 26, No. 4, 2022, pp. 1811-1823.

[39] A. Purbasari, F. R. Rinawan, A. Zulianto, A. I. Susanti, H. Komara, "CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers", Proceedings of the 8th International Conference on Advanced Informatics: Concepts, Theory and Applications, Bandung, Indonesia, 29-30 September 2021, pp. 1-6.

[40] H. El Massari, Z. Sabouri, S. Mhammedi, N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology", Journal of ICT Standardization, Vol. 10, No. 2, 2022, pp. 319-338.

[41] O. Esteban et al. "fMRIPrep: a robust preprocessing pipeline for functional MRI", Nature Methods, Vol. 16, 2019, pp. 111-116.

[42] T. Byun et al. "Input prioritization for testing neural networks", Proceedings of the IEEE International Conference On Artificial Intelligence Testing, Newark, CA, USA, 4-9 April 2019.

[43] F. Movahedi, R. Padman, J. F. Antaki, "Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores", Journal of Thoracic and Cardiovascular Surgery, Vol. 165, No. 4, 2023, pp. 1433-1442.