Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

# Machine Learning based Classification and Detection of Lung Cancer

**Trailokya Raj Ojha**

Department of Computer Science and Engineering, Nepal Engineering College, Pokhara University, Nepal

**Email:** [1]trailokyaro@nec.edu.np

**ORCID id:** https://orcid.org/0000-0001-7554-1731

## Abstract

Lung cancer has surpassed all other types of cancer as the most common cause of death worldwide. There is an increased mortality ratio and a poor diagnosis for lung cancer than any other types of cancer. Thus, forecasting rates becomes a difficult task for humans. Consequently, numerous machine learning algorithms have been suggested to offer efficient and speedy forecasting of ambiguous raw data with minimal inaccuracies. In this research, various machine learning algorithms including Support Vector Machine, Adaptive Boosting, k-Nearest Neighbor, Logistic Regression, J48, and Naïve Bayes have been implemented on medical history and physical activities of participants to identify and classify the lung cancer. Various physiological factors have been taken into account and applied to machine learning algorithms. The results indicate that all algorithms can predict incidence rates with high scores; however, Logistic Regression achieved better performance with an accuracy and f-measure of 94.7% compared to other algorithms.

**Keywords:** lung cancer, machine learning, classification, prediction, logistic regression

## 1. Introduction

The leading reason for cancer-related fatalities worldwide is lung cancer. Every year, many people die from lung cancer compared to other types of cancer. Because of prevalent smoking and polluted air, cancer affects the lung which is a fatal illness and a continuing global concern. Individuals with previous lung diseases have a high risk of diagnosing positive for lung cancer. Smoking and consumption of tobacco are treated as the main reasons for cancer


*DOI: https://doi.org/10.36548/jaicn.2023.2.003*
*Received: 09.04.2023, received in revised form: 09.05.2023, accepted: 23.05.2023, published: 02.06.2023*

caused in lung. Early diagnosis of lung cancer can help in the better treatment of the patient and can save the life.

As one of the main reason of fatalities worldwide, when a few cells in a body part start to expand uncontrollably, cancer begins. Although many types of tumour differ in how their cells divide and spread, each of these types develops as a result of "hereditary adjustments" and "epigenetic changes" to the DNA genome. Recent research has further supported the concept that epigenetic alterations has a fundamental responsibility in the development of human tumours [1].

An estimated 1.7 million people every year pass away from this illness [2]. Lung cancer has been linked to smoking as a main cause, with smoking responsible for over 80% of lung cancer occurrences worldwide. Lung cancer is difficult to find in the early stages. According to research, 25% of people with lung cancer who received an early diagnosis had no symptoms at all. As cancer caused in lung is not visible to the naked eye, it is often mistaken as those of other conditions such as bronchitis, asthma, and persistent coughing. [2].

Many risk factors are responsible for lung cancer. Different type of machine learning algorithms are helpful for predicting the co-relation between risk factors and lung cancer. Machine learning algorithms help give accurate analysis and make the correct prediction. Over the previous two decades, artificial intelligence and Machine Learning (ML) algorithms have become more and more important to help people analyse unstable data and come to stable conclusions about it. They are practically used in every aspect of human life. Algorithms for machine learning have been developed to categorize, forecast, or minimize the volume of raw data.

In this study, based on patients' statistical data and medical records, various machine learning algorithms are employed to estimate the likelihood of a lung cancer occurrence. Various methods of ML such as Support Vector Machine (SVM), Adaptive Boosting (AdaBoost) , k-Nearest Neighbor (k-NN), Logistic Regression, J48, and Naïve Bayes, are employed in the suggested work for lung cancer detection.

The Kaggle repository has been used as a data source for the study [3]. Dataset is pre-processed before applying to the machine learning algorithms. The uneven distribution of data between cancer and non-cancer classes is handled using the data balancing technique. To

determine the most accurate prediction model, the accuracy of each machine learning model is calculated and compared.

## 2. Related Works

Many approaches based on machine learning are based on neural networks. It entails categorizing data into different classes using the labels from the input training dataset. The task of classifying data can be accomplished using a variety of machine learning techniques.

Detecting cancers in lungs early can help to save the lives. Abdullah et al. [4] conducted research that looks at the precision ratios of three different machine learning classifiers. The experimental result showed that SVM achieves the best result. The accuracy for SVM was noted as 95.56%. Convolutional Neural Network (CNN) resulted with 92.11% accuracy and kNN resulted with of 88.40% accuracy.

Danjuma [5] conducted a study to analyze the machine learning algorithms' performance for patients with lung cancer. On datasets for thoracic surgery adopted from UCI repository, J48, multilayer perceptron, and the Naïve Bayes algorithms were utilized for evaluating the model. The performance measurement was done employing 10-fold cross validation. With an accuracy of 82.3%, comparative analysis revealed that the multilayer perceptron shows the best performance.

Faisal et al. [6], using the data adapted from the UCI repository, study and evaluation of different models were carried out. The result shows that the Gradient Boosted Tree performed best compared to every other classifier. The Gradient-boosted performance evaluations showed the best precision of 90% compared to other classifiers.

In a study by Tuncal et al. [7], lung cancer for eleven European nations with records dating back to 1970 used LSTM, SVM and back propagation learning algorithm. Results indicated that all algorithms can estimate incidence rates with high scores; nevertheless, Support Vector Regression outperformed the other methods that were taken into consideration.

Dursun [8] developed prediction models for prostate cancer survivorship in an examination of cancer data obtained from SEER Program of the National Cancer Institute with ANN, SVM, logistic regression and decision tree. When creating, evaluating, and comparing

models with n-fold cross-validation technique, SVM performed better than the rest of the models.

Floyd [9] showed that machine learning techniques including Bayesian Networks, SVM and ANN are useful for the prediction of lung pancreatic cancer. In the study of predicted survival time of pancreatic cancer patients, these methods prove beneficial in enhancing prognostic predictions of patient survival when compared to relying solely on logistic regression.

A study of 349 patients conducted by Patrick et al. [10], developed machine learning model to determine a postoperative femoral nerve block in patients. These classifiers used SVM, BayesNet, multilayer perceptron, Alternating Decision Tree and Logistic Regression. In terms of Receiver Operating Curve (ROC), machine learning techniques, more notably the Alternating Decision Tree, outperformed conventional Logistic Regression, and vice versa in terms of kappa statistics and the percentage correctly classified.

Research conducted by Radhika et al. [11] primarily focused on the prediction and categorization of medical imaging data. The comparative research conducted by applying different ML algorithms, the Support Vector Machines had superior accuracy (99.2%). SVM was followed by Logistic Regression marking an accuracy of 66.7%, 90% by Decision Tree , and Naive Bayes (NB) provided 87.87% of accuracy.

To evaluate the ML techniques for predicting and prognosing cancer, Kourou et al. [12] looked at several ML algorithms. It was concluded that supervised models are the main focus of study for the creation of prediction algorithms. Bayesian models were used by Ribes et al. [13] to forecast both mortality rates and incidence in Catalonia, while Malvezzi et al. [14], studied European cancer mortality predictions for the year 2014. Cancer survival rates in the Gaza Strip were predicted with random forest and Rule Induction Algorithms by Alhaj and Maghari [15].

In study [5], predictive data mining techniques were used for comparing the performance of ANN, NB, and decision tree to predict postoperative life expectancy and criticality in lung cancer patients. A 10-fold cross validation technique was used in the study to evaluate the performance of different models. Comparative analysis of the classification system for the detection of brain tumors was covered in the research by [16]. By employing

volumetric and location information, the overall accuracy rate was determined using multiple classification classes.

In a study conducted by Karhan and Tunç [17], different outcomes on the lung cancer dataset were achieved for different classifiers. The implementation of different models like KNN, SVM, NN, and Logistic Regression led to the achievement of comparable accuracy rates. The accuracy of the support vector machine was 99.3%. The suggested strategy was used on a medical dataset to assist clinicians in making more informed decisions.

## 3. Methodology

This section is divided into sub-sections namely dataset description, data pre-processing, machine learning techniques, and implementation mechanism. Below is the detailed description of each sub section.

### 3.1 Dataset Description

The dataset “survey lung cancer” acquired from the Kaggle data repository is used in this study. The dataset contains 309 entries in total with 162 male and 147 female. The dataset contains 15 attributes and one target class as listed in the table 1 below. The attribute lung_cancer is used as the output variable. The output variable contains values 'YES' and 'NO'. The value 'YES' indicates the risk for lung cancer, while the value ‘NO’ indicates the absence of the risk for lung cancer. The dataset used in this study was not balanced, as it has 270 rows which have a value of 'YES' whereas only 39 rows contains ‘NO' in the lung_cancer column. For accurate prediction, the data pre-processing was performed to balance the dataset. The description of the attributes of the datasets are depicted in Table 1.

**Table 1.** Description of the Attributes

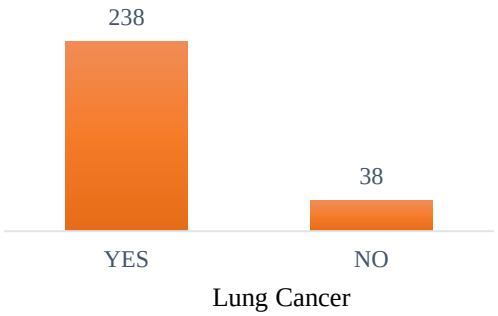
| **Name of Attribute** | **Type and Values** | **Description of the attribute** |
|---|---|---|
| Gender | Predictor Variable (Male, Female) | Participant's gender |
| Age | Predictor Variable (21 to 87) | Age of the participant |
| Smoking | Predictor Variable (Yes, No) | Describes the smoking habit |
| Yellow_finger | Predictor Variable (Yes, No) | Describes whether the participant has yellow fingers or not. |
| Anxiety | Predictor Variable (Yes, No) | Tells whether the participant has an anxiety problem or not. |
| Peer_ pressure | Predictor Variable (Yes, No) | Describes the status of peer pressure on the participants. |
| Chronic diseases | Predictor Variable (Yes, No) | Tells whether the participant has chronic diseases or not. |
| Fatigue | Predictor Variable (Yes, No) | Shows the participant's fatigue status. |
| Allergy | Predictor Variable (Yes, No) | Describes allergic condition of participant. |
| Wheezing | Predictor Variable (Yes, No) | Describes whether the participant has a wheezing problem or not. |
| Alcohol consuming | Predictor Variable (Yes, No) | Describes the alcohol consumption status of participants. |
| Coughing | Predictor Variable (Yes, No) | Shows the state of participants coughing problem. |
| Shortness_of_breath | Predictor Variable (Yes, No) | Describes whether the participant has a breathing problem. |
| Swallowing difficulty | Predictor Variable (Yes, No) | Tells if there is a difficulty faced while swallowing. |
| Chest pain | Predictor Variable (Yes, No) | Describes if the participant is suffering from chest pain or not. |
| Lung_cancer | Response Variable (Yes, No) | Describes the state of lung cancer. This is the response variable. |

### 3.2 Data Pre-processing

Final prediction could be negatively affected by noise and/or missing values in the raw data. Sometimes the secondary source dataset isn't ready to apply in the machine learning algorithm to process. To make the dataset ready to be processed by algorithms, it is pre-

processed. The null values are checked, adjusted, and data balance is done at this point. In this stage, anything that affects the performance of machine learning models is more effectively dealt with. The dataset contains sixteen attributes, including the output variable. Firstly, the null values in dataset is checked and if it occurs, it is filled by using supervised filter in WEKA. In the next step, the numeric values are converted to categorical values. For this, the value '2' was replaced with 'Yes' and '1' with 'No.' To make it more understandable, the values of age attribute 'M' was replaced by 'male', and 'F' was replaced with 'Female'.

The dataset used in the study had a significant amount of imbalance, with 270 rows having a value of "YES" and only 39 rows having a value of "NO" in the target variable (lung cancer) column. Predictions and results are ineffectual if such uneven data is not controlled. The dataset contained 33 duplicate entries, which were also removed. After removing the duplicate entries, the dataset contains 276 entries 238 entries with cancer, and 38 entries without cancer. To address the unequal distribution of data between cancer and non-cancer groups, the SMOTE [18] technique was applied in this study. The participants were distributed equitably because the "non-cancer" (the minority class) was oversampled. The dataset did not have any null or missing values, so neither data imputation nor dropping was used. Figure 1 shows the distribution based on the risk of lung cancer before and after balancing the dataset.



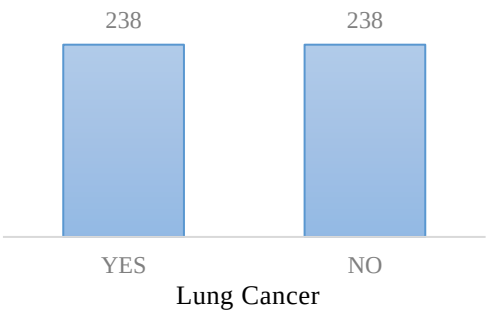




**Figure 1(a).** Data set before oversampling **Figure 1(b).** Data set after oversampling

The next step after balancing the dataset is to develop a model. After pre-processing the dataset, it is divided into training and test sets. In this study, the split technique with 80% training and 20% test data has been used. The randomly selected 80% data is used as training set, and the remaining 20% of the data is used as a test set. Furthermore, the training and test sets are applied to the machine learning model.

### 3.3 Machine Learning Techniques

#### Naive Bayes

The Naive Bayes classifier assumes that characteristics are independent of class, which considerably simplifies learning. Even though independence is often a bad assumption, in practice, Naïve Bayes frequently outperforms more advanced classifiers. The Bayes theorem offers the conditional probability of an event occurring in comparison to an already occurring event. The Naive Bayes classifier's conditional probability can be calculated as follows:

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} \quad (1)$$

The dataset with several attributes does not lend itself to a Naive Bayes classifier. Instead, it excels at working with tiny datasets that require less training data, is highly scalable, skilled at working with both continuous and discrete data, and is insensitive to superfluous features [19]. The precision of the Naive Bayes classifier does not have direct relation to the feature dependency rate obtained as a class-conditional similar data between different characteristics [20].

#### Support Vector Machine

SVM is a supervised learning method which can be used in combination with classification methods to examine the data for regression and classification. The primary goal of SVM is to create a precise linear or deterministic partition that divides the n-proportional space into groups to simplify the combination of newly created data into the relevant modules for additional references [21]. The method of accurate linear data sorting is known as a hyper-plane.

The linear SVM is necessary to linearly separate the data which signifies separating the dataset into two distinct classes by a certain linear separation. The data is referred to as a linear differential as shown below.

High risk class= $(H*M+b) \leq -1, \forall M$

Low risk class= $(H*M+b) \geq 1, \forall M$

Here, H is the hyper-plane vector, M is the data set matrix, and b is bias.

In deterministic data division, the non-linear SVM itself is used to refer to a dataset that cannot be divided by the shortest path; this particular information is referred to as non-linear data. It can be represented as follows.

$$F(X, Y) = (1 + XY)^{d} \quad (2)$$

Here, X and Y are the low and high risk, respectively, and d is the degree of the polynomial.

The default kernel Radial Basis Function (RBF) with gamma kernel parameter provided by WEKA is used as hyper-parameter in SVM. The default values of gamma hyper-parameter range from $2^{-15}$ to $2^{3}$.

**k-Nearest Neighbor**

k-NN is regarded as a lazy-learning strategy and is a nonparametric supervised learning method. It maintains the dataset and acts at the moment of classification rather than instantly training the given dataset [22]. Finding similarities between new and existing data is the fundamental tenet of k-NN, which then assigns fresh samples to the category that resembles the existing categories the most. The complete training data set is scanned for the k similar examples, also known as the neighbors, to produce predictions for a new instance. The forecast of the output is determined by these k instances. In this study, the value of k is used as 1, 3 and 5.

K-NN can be used with a variety of distance metrics, including Manhattan, Minkowski, and Euclidean distances. In this study, the Euclidean distance based model is used. The Euclidean distance is the shortest path between two points. The formula below can be used to obtain the Euclidean distance between two points [23].

$$d(P1, P2) = \sqrt{(m1 - m2)^2 + (n1 - n2)^2} \quad (3)$$

**AdaBoost**

The most typical and widely used ensemble learning method is the AdaBoost algorithm. Boosting is a technique that combines all ineffective classifiers into one strong classifier. It generates ‘n’ different multiple decision trees throughout the instruction session. The data record that was first identified erroneously is given top priority after the final decision tree is

built. The output of these is collected and sent to the subsequent decision model based on this. Until the necessary base of learners that were suggested to be created at the beginning is reached, this process iterates and repeats itself [21]. The default values of hyper-parameters such as base_estimator, n_estimators, learning_rate and algorithm were used.

**J48 Classifier**

A decision tree has many benefits for data mining, making it easy for users to understand and execute. Even with inaccurate or inadequate datasets, it can nevertheless yield more accurate forecasts. The decision tree is developed by J48 using the notion of knowledge acquisition. Using the selection of each attribute as a function of information gain, the J48 algorithm splits the dataset into smaller subsets. When every instance in each subset belongs to the same class, which is determined by the attribute with the greatest information gain, the splitting procedure is complete. When features are not given any information gain, J48, on the other hand, can manage both continuous and discrete properties [24]. Adjusting the missing values, classifying discrete, continual value of attribute, scaling decision trees, and continuous data utilizing limits, deriving rules, and other features are among the characteristics of the J48 method [25].

**Logistic Regression**

It is a regression technique that anticipates a category wise dependent data. The logistic regression equation incorporates the highest possible ratio to assess the statistically significant nature of the variables [26].

Based on the outcomes of a collection of variables known as predictors, logistic regression is effective at estimating the existence or missing a characteristic or result. It is suitable for models if the variable of interest is categorical, despite looking similar to a linear regression model [27]. In this case, logistic regression is chosen because the target variable in the dataset has only two values.

### 3.4 Implementation Mechanism

The machine learning approach is used in this research to create a prediction model for participants' risk of developing lung cancer. The WEKA toolkit [28] is used to analyse ML models by applying the datasets to find the strongest and most likely correlation between them. The main steps involved in the designed model are data acquisition, pre-processing, and data

manipulation using data mining techniques. Different data mining approaches like SVM, Naïve Bayes, k-NN, Adaptive Boosting, J48, and Logistic Regression classifiers are implemented for knowledge representation. Figure 2 depicts an outline of the working mechanism.

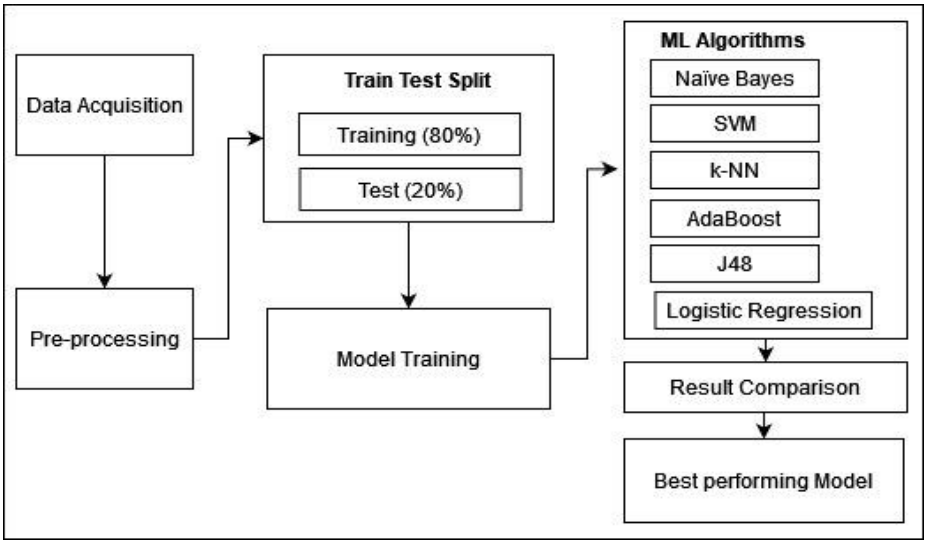



**Figure 2.** Proposed Implementation Procedure

The first step is to acquire data. The data used in this study is acquired from Kaggle repository. After data adoption, data are pre-processed to make ready to apply to machine learning algorithms. In next step, training and test strategy is developed. The randomly selected 80% data is used as training set, and the remaining 20% of the data is used as a test set. Then the data is deployed to ML model and different performance parameters are noted. And based on the outcome, the best performing model is concluded.

After the models have been developed, they are evaluated using various accuracy matrices, including the accuracy, recall, precision, F measure, and ROC curve. A measure of precision is the proportion of forecasts in the positive class that are positive. Recall counts how many correctly classified predictions are made using all positive cases. F-Measure gives a value to balance recall and precision in a single number. With the formula provided in the following, these parameters are determined.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{4}$$

$$Precision = \left(\frac{TP}{TP+FP}\right) \tag{5}$$

$$Recall = \left(\frac{TP}{TP+FN}\right) \tag{6}$$

$$F - Measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \quad (7)$$

Where, TP indicates True Positive, FP indicates False Positive, TN denotes True Negative, and FN indicates False Negative.

## 4. Results and Discussion

In this section, the performance evaluation of machine learning models is done using the WEKA 3.8.6 [28] environment. It offers a wide range of data preparation, association, classification, clustering, visualization, and other functions. The model is trained using the classification methods. Table 2 summarizes the effectiveness of various ML classifiers used to train the model.

**Table 2.** Machine Learning Algorithms' Performance Comparison

| Classifier | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|
| Naïve Bayes | 0.916 | 0.917 | 0.916 | 0.916 |
| SVM | 0.926 | 0.931 | 0.926 | 0.926 |
| k-NN | 0.905 | 0.905 | 0.905 | 0.905 |
| AdaBoost | 0.905 | 0.906 | 0.905 | 0.905 |
| J48 | 0.905 | 0.910 | 0.905 | 0.905 |
| Logistic Regression | 0.947 | 0.948 | 0.947 | 0.947 |

From Table 2, it can be seen that the logistic regression classifies best with the 94.7% accuracy. The precision, recall, and F-measure are measured as 94.8%, 94.7%, and 94.7% respectively for the Logistic Regression classifier. The k-NN, AdaBoost, and J48 classifiers have the same accuracy, recall, and F-measure, that is 90.5%. The precision values for the k-NN, AdaBoost, and J48 classifier are noted as 90.5%, 90.6%, and 91.0%, respectively. Similarly, the Naive Bayes classifier has accuracy, recall, and F-measure of 91.6% whereas these metrics for SVM is 92.6%. The precision for Naïve Bayes and SVM are noted as 91.7% and 93.1% respectively.

The logistic regression classifier works in the manner of predicting the probability of occurrence of some class based on other dependent variables. In this study, the probability of occurrence of lung cancer is dependent on other parameters such as age, smoking habit, alcohol

consumption, chest pain and others; as a result, logistic regression classifier performed best compared to other classifiers.

From the result in table 2, it can be concluded that logistic regression classifier performs better on the dataset used in the research compared to another classifier. From the result, it is also seen that the k-NN, AdaBoost and J48 classifiers have the same accuracy of 90.5%. The graphical representation of the accuracy of different classification models is depicted in Figure 3.

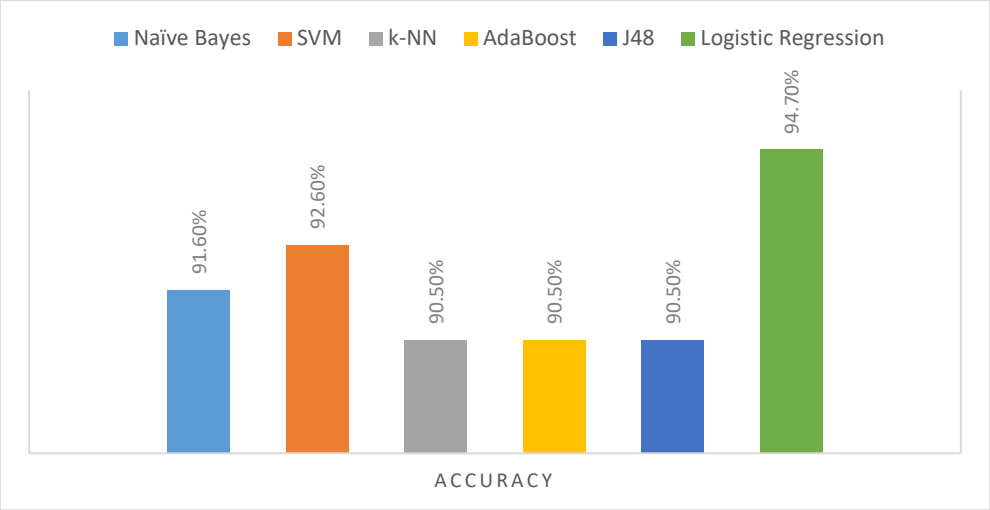



**Figure 3.** Performance Comparison of ML Algorithms

The comparison of different evaluation metrics (precision, recall, and F-measure) of classification models is shown in Figure 4.

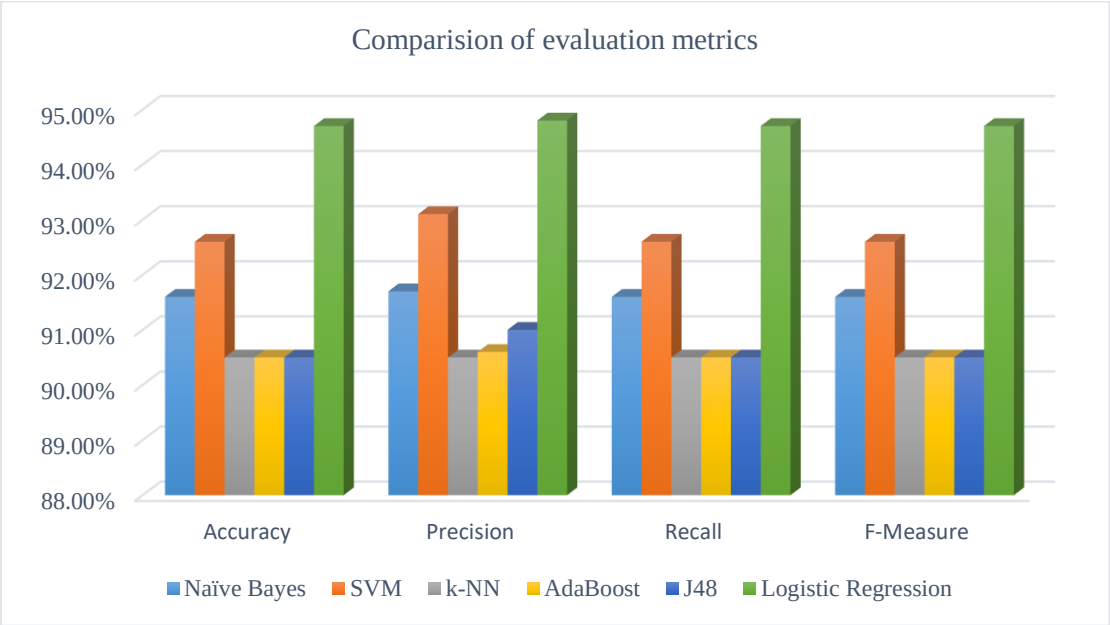



**Figure 4.** Comparison of Evaluation Metrics of ML Algorithms

From Figure 4, it can be observed that Logistic Regression performs the best in every aspect compared to other classification models. The Receiver Operator Characteristic curve is a measurement tool commonly used for binary classification problems. Its purpose is to distinguish the "signal" from the "noise" by plotting the true positive rate against the false positive rate at various threshold values. Figure 5 depicts the ROC curve for the logistic regression classifier and Figure 6 depicts the combined ROC curve for different ML models.

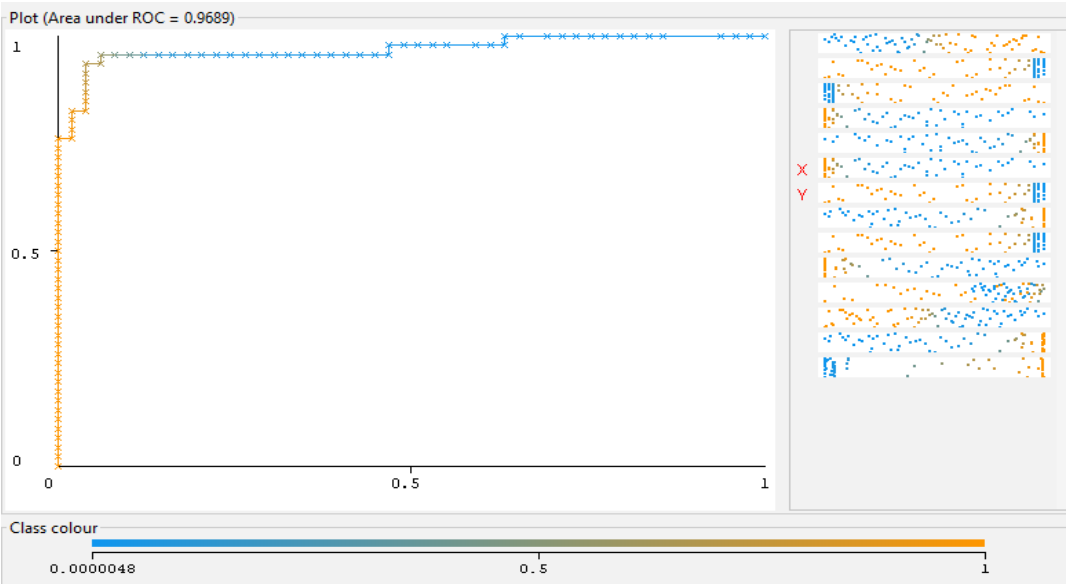



**Figure 5.** ROC Curve for the Logistic Regression Classifier

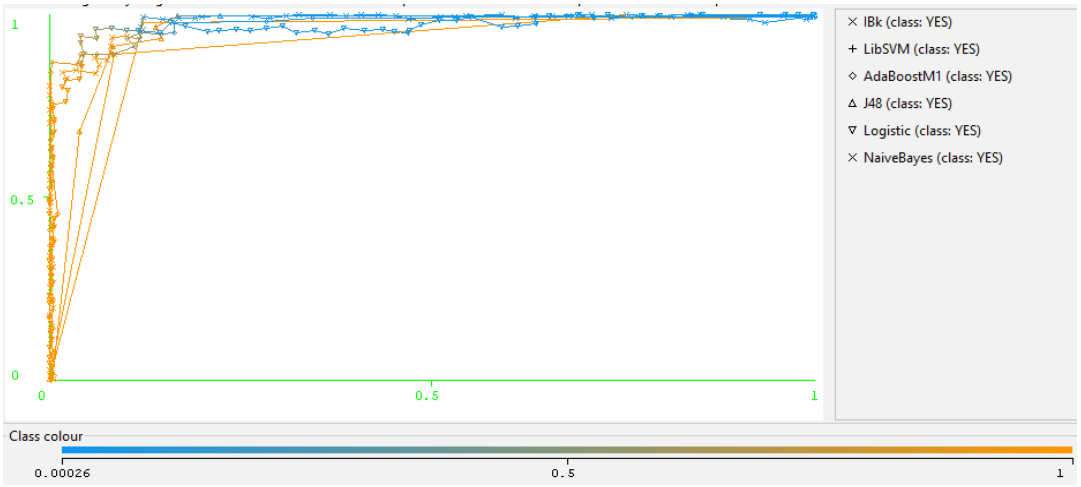



**Figure 6.** The Combined ROC Curve for Different ML Models

To assess the classifier's ability to differentiate between classes, the Area Under the Curve (AUC) is utilized as a summary of the ROC curve. From Figure 5 and Figure 6, it is evident that the AUC falls between 0.5 and 1. This indicates that the classifier has exhibited improved performance in accurately predicting true positives and true negatives.

## 5. Conclusion

In this study, machine learning -based classification algorithms are used for identification and categorization of lung cancer. The technique involves six different steps, such as data acquisition, data preprocessing, splitting the training and test set, model training, and result comparison. The various machine learning algorithms used in this study are Naïve Bayes, Support Vector Machine, k-Nearest Neighbor, Adaptive Boosting, J48, and Logistic Regression. Different performance parameters namely accuracy, recall, precision, and F-measure are considered as the simulation result. The simulation result is also depicted in the combined ROC curve for all the algorithms used in this study. This makes it easy to compare the result in a graphical form. After completing the training of all models, it is found that the logistic regression model outperforms the other models with an accuracy and f-measure of 94.7%. Thus, this study may aid in improving the clinical prediction of lung cancer.

## References

[1] M. S. Kumar and K. V. Rao, "Prediction of Lung Cancer Using Machine Learning Technique: A Survey," in 2021 IEEE International Conference on Computer Communication and Informatics (ICCCI), Jan. 2021, pp. 1–5.

[2] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," Neural Computing and Applications, vol. 31, no. 10, pp. 6863–6877, 2019.

[3] "Lung Cancer Dataset," https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection. [Accessed: August 09, 2022].

[4] D. M. Abdullah, A.M. Abdulazeez and A.B. Sallow, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques", Qubahan Academic Journal, Vol. 1, no. 2, pp. 141-149, 2021, doi: 10.48161.

[5] K. Joro Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients." arXiv preprint arXiv:1504.04646, 2015.

[6] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," in 2018 IEEE 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST), 2018, pp. 1–4.

[7] K. Tuncal, B. Sekeroglu, and C. Ozkan, "Lung cancer incidence prediction using machine learning algorithms," Journal of Advances in Information Technology, vol. 11, no. 2, pp. 91–96, May 2020, doi: 10.12720/jait.11.2.91-96.

[8] D. Delen, "Analysis of cancer data: a data mining approach," Expert Systems, vol. 26, no. 1, pp. 100–112, 2009.

[9] S. Floyd, "Data Mining Techniques for Prognosis in Pancreatic Cancer," Doctoral dissertation, Worcester Polytechnic Institute, 2007.

[10] P. Tighe, S. Laduzenski, D. Edwards, N. Ellis, A. P. Boezaart, and H. Aygtug, "Use of Machine Learning Theory to Predict the Need for Femoral Nerve Block Following." Pain Medicine, vol. 12, no. 10, pp. 1566-1575, 2011.

[11] P. R. Radhika, R. A. Nair, and G. Veena, " A comparative study of lung cancer detection using machine learning algorithms," in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Feb. 2019, pp. 1–4.

[12] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. v. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13. Elsevier, pp. 8–17, 2015. doi: 10.1016/j.csbj.2014.11.005.

[13] J. Ribes et al., "Cancer incidence and mortality projections up to 2020 in Catalonia by means of Bayesian models," Clinical and Translational Oncology, vol. 16, no. 8, pp. 714–724, 2014, doi: 10.1007/s12094-013-1140-z.

[14] M. Malvezzi, P. Bertuccio, F. Levi, C. la Vecchia, and E. Negri, "European cancer mortality predictions for the year 2014," Annals of Oncology, vol. 25, no. 8, pp. 1650–1656, 2014, doi: 10.1093/annonc/mdu138.

[15] M. A. , Alhaj and A. Y. Maghari, "Cancer survivability prediction using random forest and rule induction algorithms," in 2017 8th International Conference on Information Technology (ICIT), IEEE, May 2017, pp. 388–391.

[16] M. Kaur and R. Mittal, "Survey of Intelligent Methods for Brain Tumor Detection," International Journal of Computer Science Issues (IJCSI), vol. 11, no. 5, p. 108, 2014.

[17] Z. Karhan and T. Tunç, "Lung Cancer Detection and Classification with Classification Algorithms," IOSR Journal of Computer Engineering (IOSR-JCE), vol. 18, no. 6, p. 71, 2016.

[18] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," Applied Soft Computing Journal, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.

[19] T. I. Shoily et al., "Detection of Stroke Disease using Machine Learning Algorithms." in 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1-6, Jul. 2019.

[20] I. Rish, "An empirical study of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, pp. 41-46, Aug. 2001.

[21] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning," Revue d'Intelligence Artificielle, vol. 34, no. 6, pp. 753–761, Dec. 2020, doi: 10.18280/RIA.340609.

[22] G. Sailasya and G. L. Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms." International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.

[23] A. Pandey and A. Jain, "Comparative Analysis of KNN Algorithm using Various Normalization Techniques," International Journal of Computer Network and Information Security, vol. 9, no. 11, pp. 36–42, Nov. 2017, doi: 10.5815/ijcnis.2017.11.04.

[24] L. Alam, A. R. Onik, T. Dhaka, B. Nutan, F. Haq, and T. I. Mamun, "An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier," International Journal of Computer Applications, vol. 124, no. 13, 2015.

[25] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," International journal of computer applications, vol. 98, no. 22, 2014.

[26] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression. New York: John Wiley & Sons, 2000.

[27] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," Expert Syst Appl, vol. 34, no. 1, pp. 366–374, Jan. 2008, doi: 10.1016/j.eswa.2006.09.004.

[28] "WEKA Tool," Available Online: https://www.weka.io/. [Accessed: August 27, 2022].

## Author's Biography

**Trailokya Raj Ojha** is working as an Assistant Professor at the Department of Computer Science and Engineering at Nepal Engineering College, Changunarayan Bhaktapur, Nepal. He has received master's degree in Software Systems from Tampere University of Technology, Finland in 2014. He has been involved in software development and the academic profession since 2008. His research fields include software engineering, process improvement, software project management, Internet of Things, and machine learning.