IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Evaluating Cancer Prediction Machine Learning Models

[1] Ms.Jataboina Hoyala, [2] Mohammed Ayman Riaz, [3] Syed Abdul Ahmed, [4] Syed Danish Ali,
[1] Assistant Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology.
[234] Student Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract—

A disease known as "Cancer" develops when cells' genes undergo alterations that cause them to proliferate uncontrollably; this, in turn, causes tumors to grow, which invade and harm healthy bodily components. In lung cancer, the diseased cells in the lungs proliferate at an alarming pace. Using contemporary data analysis, we can identify this aberrant proliferation of cells that ultimately leads to cancer. Patients who may suffer later if the signs of cancer are not discovered at an early stage are the ones who benefit the most from early detection. The rising popularity of cigarette use among young people is one of the main issues. A number of factors, including industrial air pollution that people breathe in, contribute to the alarming rise in lung cancer cases in India. Machine learning (ML) techniques such RFC, KNN, K-means, SVM, and DTC are the primary focus of this work, which aims to predict lung cancer in various individuals. The primary goal of this study is to compare and contrast several machine learning algorithms using various performance measures. Search terms: cancer, support vector machine, k-nearest neighbor, random forest, machine learning.

INTRODUCTION

Patients having a history of lung illness, such as emphysema or chronic pulmonary disease, are at a higher risk of developing cancer. Tobacco, cigarette, and beedi overuse are major contributors to cancer in Indian society [1]. On the other hand, smoking isn't very widespread among Indian women, which suggests that other variables cause carcinoma and, ultimately, lung cancer in women. Radon gas, air contaminants, and chemicals used in the workplace are among additional risk factors that may lead to carcinoma. Looking at the big picture, India was responsible for almost 8% of all cancer-related fatalities worldwide in 2008[14]. Some rare forms of cancer still have no treatment, despite the fact that there are many strategies to avoid it. Medical image processing is one of the many important areas of study that makes use of Machine Learning (ML)[12]. The molecular level of cancer research stands to benefit greatly from the vast databases of biological and chemical information that have recently been accessible because to technological advancements in genomics, proteomics, and combinatorial chemistry [13]. A tumor's size and the extent to which it has spread throughout the body and the pulmonary system are indicators of its stage, both of which are produced by the unchecked development rate of cells. In some cases, we can detect it at an early stage, when the tumor is still small and treatable. In other cases, we can detect it at a later stage, when the tumor has already spread to nearby tissues, and the results can be used to better understand risk factors for carcinoma disease and how to prevent it. Machine learning, which makes use of algorithms and visuals to depict continuous health information, is the key to early detection of this significant disease. Radiologists and oncologists may use these patient statistics to make more accurate cancer diagnoses. The objective of better early diagnosis of lung cancer is an important one, and this approach has the potential to be a game-changer.

LITERATURE SURVEY

For the purpose of lung cancer prediction, several machine learning methods were used in [1]. Decision trees, logistic regression, support vector machines, and naive bayes were the techniques used. Examining and analyzing algorithms to identify cancer as early as feasible was the major objective of this work.
A novel method with excellent cancer detection accuracy is introduced in [2]. There were two primary components to this approach. In the first part, we identified patterns and features. In the second part, we utilized the retrieved features as input to two supervised machine learning models: Logistic Regression (LR) and Back Propagation Neural Network (BPNN). Next, we compared and contrasted the two models to see which one was more accurate. The primary goal of comparing several ML algorithms in [3] is to look at their accuracy, precision, and limits. Additionally, oncologists do extensive studies to better categorize tumors as malignant or benign.
In [4], the goal is to compare different cancer detection methods and machine learning techniques. Three machine learning algorithms—Naive Bayes, Random Forest, and KNN—were used. In order to assess the performance of

each ML algorithm, this research uses the Wisconsin Diagnosis Cancer data set and measures characteristics including accuracy and precision. This study[5] intends to investigate, investigate, and assess the most recent advances in cancer diagnosis using ML approaches for skin, bone, breast, and brain cancers. Cancer detection and treatment are both aided by supervised and unsupervised learning algorithms as well as deep learning. In [6], several ML approaches and algorithms, including as Support Vector Machine, K-Nearest Neighbors, and Naive Bayes, are used to directly communicate the extracted components to the separators, who then use them to discern between cancerous and benign cells in skin lesions. [7] demonstrates a method for accurately diagnosing lung cancer and its stages utilizing SVM methods and Image Processing, with the goal of achieving the most accurate findings possible. One example of a Machine Learning classification approach, SVMs, is applied in [8] on the Herlev pap-smear dataset. The active computational models were powered during the classification phase using Gaussian fitting energy. Expert cytologists carefully annotate these separated photos and then correlate them. In [9], the morphological features of healthy and cancer-affected bones were comparable in the dataset. The collection included many photos of bones. They created two sets of characters and then identified the optimal edge acquisition method to address the issue. We used two ML models—Random Forest and Support Vector Machine—to see how well these feature sets performed. A method that improved the parameters and improved the results of three separate classifiers—Subsequent Minor Improvement (SMO), Decision Tree (J48), and Naïve Bayes (NB)—was devised in [10]. It was with the Wisconsin Breast Cancer dataset that the three classifiers were compared.

III. Methodology

A. Block Diagram

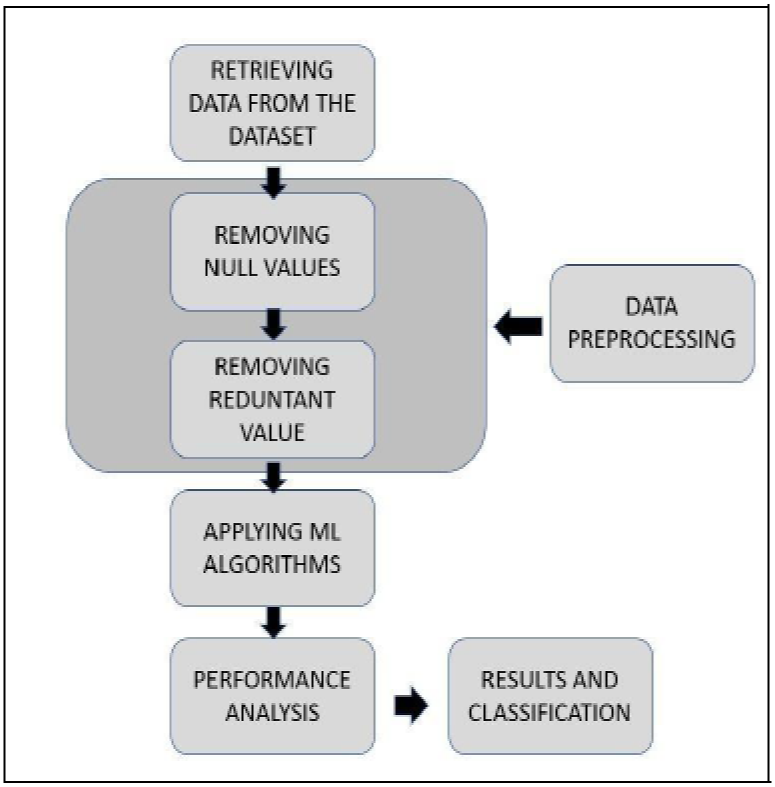
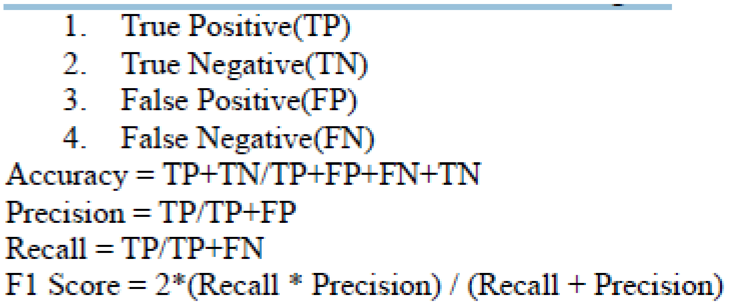


Fig.1: Process Classification

B. Algorithms used

Machine learning algorithms are an aspect of AI that allows the model to learn and respond based on the given data. Supervised, unsupervised, and reinforcement learning are the three main groups into which machine learning approaches fall. The following machine learning algorithms were used for this project: 1) A Classifier via Random Forest 2) The Most Proximate Neighbor 3. k-Means algorithms Fourthly, SVM 5. Classifier Based on Decision Trees Part C: Dataset Overview This research makes use of a dataset[16] sourced from data globe, a free and opensource website. The dataset has 25 columns and around 1000 items. There are 25 columns total; 24 of them are properties of the predictive kind and 1 is the class label. Lung cancer symptoms and risk factors (such as obesity, genetics, coughing, and exhaustion) are shown in each column. From 0 to 9, these columns display the values that correspond to the severity of the issue. By using these characteristics for training and testing, the prediction model is constructed.

Section D. General Design The processes outlined in Fig.1 have been followed to process the data. A dataset on lung cancer was used, sourced from data world [16]. Python is being used to implement the sklearn library of machine learning algorithms. Part one of Figure 1 shows data pre-processing. This includes cleansing the dataset by removing null and unnecessary values. The next step is to use python packages such as matplotlib and seaborn to visualize the data. Data visualization allows for a more thorough comprehension and analysis of the data. The implementation of prediction models follows pre-processing. A quarter of the data is used for testing purposes, whereas three quarters are utilized for training the model. Next, the models are assessed using performance criteria such as F1 score, recall, accuracy, and precision. A confusion matrix is used to compute these. There are four different terms in the confusion matrix:

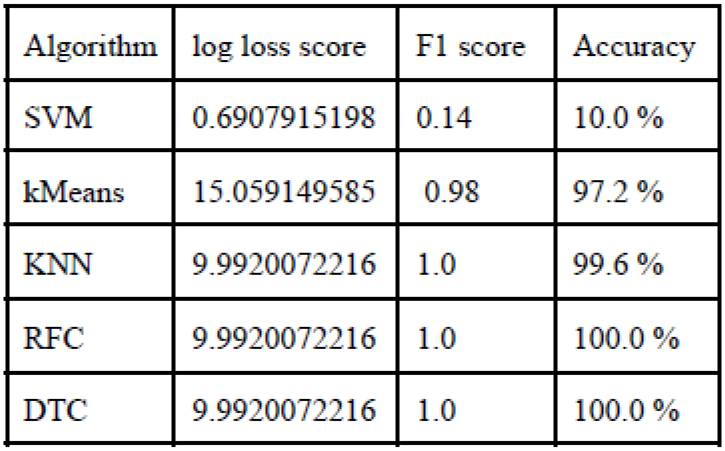


1. True Positive(TP)
2. True Negative(TN)
3. False Positive(FP)
4. False Negative(FN)

Accuracy = TP+TN/TP+FP+FN+TN
Precision = TP/TP+FP
Recall = TP/TP+FN
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

IV. RESULTS AND DISCUSSIONS

A total of 750 observations were used to create the training set, whereas 250 were used to create the testing set. Each prediction model's confusion matrix was used for deployment. We computed many performance measures. The accuracy and f1 score that various ML systems achieved are graphically shown in Fig. 2. Based on the data in Table.1, it is clear that the models with the highest accuracy and F1 score were Randomforest, decision tree, and KNN.



Algorithm	log loss score	F1 score	Accuracy
SVM	0.6907915198	0.14	10.0 %
kMeans	15.059149585	0.98	97.2 %
KNN	9.9920072216	1.0	99.6 %
RFC	9.9920072216	1.0	100.0 %
DTC	9.9920072216	1.0	100.0 %

Table.1: Performance metrics

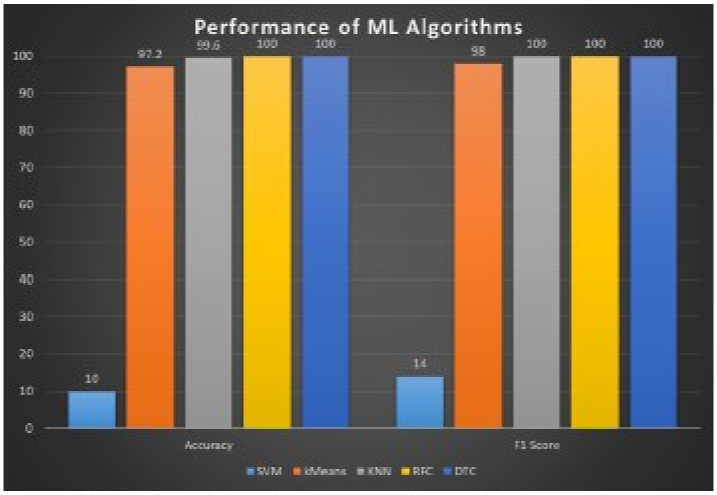


Fig.2 Graphical Representation of Performance metrics

In [1], there is a quantitative comparison of classifiers' predicting ability. The support vector machine method produces the best results when considering accurate classification and other criteria. They found that compared to the BPNN, the LR model used a much larger amount of characteristics in [2]. Based on the results shown in [6], SVM outperforms all other classifiers when compared to them, with an AC of 97.8% and an AUC of 0.94. The results obtained by KNN in terms of sensitivity are 86.2% and specificity are 85%.

LIMITATIONS

When dealing with a high number of observations, support vector machines (SVMs) become computationally inefficient. Overfitting the training data is a problem with the Decision Tree, despite its improved accuracy. If you want more precise results, you need to do more preprocessing.

FUTURE SCOPE

The focus of this research is only on lung cancer. Additional adjustments may be made to categorize different forms of cancer. Considering the study's limitations, we might utilize a more precise dataset to improve cancer categorization.

CONCLUSION

The purpose of this study was to evaluate five different ML algorithms for use in lung cancer diagnosis. The dataset utilized for this investigation was related to lung cancer. A variety of performance indicators were computed, including accuracy, the F1 score, and the log loss score. We created a visual representation based on these metrics. Table 1 shows that compared to other ML algorithms, RFC, DTC, and KNN perform the best. Therefore, these models may be made more accurate, which will help with cancer diagnosis, with a few more tweaks to the implementation aspect.

REFERENCES

[1] Radhika P R, Rakhi.A.S.Nair, Veena G, "*A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms*", Department of Computer Science and Engineering,Amrita VishwaVidyapeetham ,Amritapuri ,India, 2019.2.20.

[2] Moh'd Rasoul Al-hadidi, Abdulsalam Alarabeyyat, Mohannad Alhanahnah, "*Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm*", Computer Science Department ,AlBalqa Applied University, University of Kent ,Salt, Jordan,3 Kent, UK, 2016.8.31.

[3] Eali Stephen Neal Joshua, Midhun Chakkravarthy, Debnath Bhattacharyya, *"An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study"*, Department of Computer Science and Multimedia, Lincoln University College, Kuala Lumpur 47301, Malaysia, 2020.5.7

[4] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury ," *Breast Cancer Detection Using Machine Learning Algorithms*", University of Petroleum & Energy Studies, Amity University Uttar Pradesh,2018.12.21.

[5] Tanzila Saba, "*Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges*", College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia, 2020.6.033.

[6] Vidya M, Dr. Maya V Karki, *"Skin Cancer Detection using Machine Learning Techniques",* Department of Electronics and Communication Ramaiah Institute of Technology, Bangalore, India, 2020.

[7] Wasudeo Rahane, Himali Dalvi, Yamini Magar, Anjali Kalane*, "Lung Cancer Detection Using Image Processing and Machine Learning HealthCar"*, Information Technology Department, NBN Sinhgad School of Engineering, Pune, India, 2018.

[8] Aditya Arora, Anurag Tripathi, Anupama Bhan, *"Classification of Cervical Cancer Detection using Machine Learning Algorithm"*, Amity School of Engineering and Technology, Amity University, Sector 125, Naida, Uttar Pradesh 201313, 2021.

[9] Ashish Sharma, Dhirendra P. Yadav, Hitendra Garg, Mukesh Kumar, Bhisham Sharma and Deepika Koundal, *"Bone Cancer Detection Using Feature Extraction Based Machine Learning Model"*, Department of Computer Engineering & Applications, GLA University, NH#2, Delhi Mathura Highway, Post Ajhai, Mathura, (UP), India, 2021.

[10] Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, and Gunter Saake, *"Analysis of Breast Cancer Detection Using Different Machine Learning Techniques",* University of Magdeburg, Magdeburg, Germany, pp. 108–117, 2020.

[11] Muhammet Fatih Ak, *"A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications"*, Industrial Engineering Department, Antalya Bilim University, 07190 Antalya, Turkey, 2020.4.26.

[12] Deepshikha Shrivastava , Sugata Sanyal , Arnab Kumar Maji and Debdatta Kanda, "*Bone cancer detection using machine learning techniques"*, Department of Information Technology, North Eastern Hill University, Shillong, India, 2020.

[13] John F. McCarthy,kenneth a. marx,patrick e. hoffman,alexander g. gee,philip o'neil,m l. ujwal,john hotchkiss, *"Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis,andManagement"*,AnVil,Incorporated,Burlington,Massachus hetts,USA, 2006.1.12.

[14] Rajesh Dikshit, Prakash C Gupta, Chinthanie Ramasundarahettige, *"Cancer mortality in India: a nationally representative survey"*, Tata Memorial Hospital, Mumbai, India, March 28, 2012.

[15] Dataset used: https://data.world/cancerdatahp/lung-cancer-data