Prediction of Lung Cancer Using Machine Learning

2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-7433-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRITO56286.2022.9965052

Shubhada Agarwal[1], Sanjeev Thakur[2], Alka Chaudhary[3]
[1,2] Amity School of Engineering & Technology
Amity University, Noida, Uttar Pradesh, India
[3]Amity Institute of Information Technology,
Amity University, Noida, Uttar Pradesh, India
[1]*shubhada.agarwal@s.amity.edu,* [2] sthakur3@amity.edu, [3]achaudhary4@amity.edu

Abstract

Cancer detection is done with the aid of the led expert docs and earlier tiers it may helpful. The opportunity for human error must be there. It produces the probability of error in lung cr detection which necessitates an automatic manner. Afterward, the report aims at early cancer detection through an automatic procedure to decrease human error and make the system more accurate and error-free free this system, use machine learning algorithms such as random forest, logistic regression, support vector machine and, decision tree algorithms to detect lung cancer Th research is conducted on COLAB. With COLAB or "Collaboratory" we can write and run Python in our browser, which requires a zero-configuration, free access to GPU, and is easy to share. We have implemented four algorithms on lung cancer dataset to check the performance based on diagnosis of the four parameters i.e. accuracy, Recall, Harmonic Mean, and Precision, and also presented the comparison of the four algorithms.

Keywords: Machine Learning, Lung Cancer Detection, Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine.

INTRODUCTION

Adenocarcinoma (AC) and squamous cell carcinoma (SCC) are the most frequent types of non-small cell lung cancer (NSCLC) and are responsible for the majority of cancer fatalities globally. The purpose of this study is to look at the possibility of histologically classifying NSCLC into AC and SCC using various feature extraction and classification approaches using pretreatment CT scans. The picture collection (102 patients) was received from the Cancer Imaging Archive Collection, which is freely accessible to the public (TCIA). We investigated four distinct techniques: (a) radionics with two classifiers (kNN and SVM), and (b) four cutting-edge convolutional neural networks (CNN) with transfer learning. and fine-tuning (Alexnet, ResNet101, Inceptionv3, InceptionResnetv2)), (c) CNN combined with a long short-term memory (LSTM) network to fuse information about spatial coherence of tumor CT slices, and

(d) combined model (LSTM + CNN + radio) Mix). In addition, CT images were independently evaluated by two professional radiologists. Our results show that the highest is CNN Inception (accuracy = 0.67, AUC = 0.74). LSTM + Inception performed better than all other methods (accuracy = 0.74, AUC = 0.78). In addition, LSTM + Inception outperformed experts by 7-25% ($p < 0.05$). The proposed method does not require detailed segmentation of the tumor area and can be used in combination with radiological findings to improve clinical decision-making [1].

A potential new tool for non-invasive, simple-to-use, and affordable lung cancer screening is e-nose-based breath analysis. Predictive models created using machine learning methods are the foundation of lung cancer screening with e-nose [2]. However, there are several disadvantages to utilising a single machine learning algorithm to identify lung cancer, such as low detection accuracy and high false-negative rates. Choose a group of individual learning models with great performance from more established models like support vector machines, decision trees, random forests, and logistic regression to handle these difficulties., and regression using K-nearest neighbours. In this paper, we have implemented four algorithms on lung cancer dataset to check the performance based on diagnosis of the four parameters i.e. accuracy, Recall, Harmonic Mean, and Precision, and also presented the comparison of the four algorithms.

REVIEW OF LITERATURE

Lung Cancer is mostly found in females and males due to reason of uncontrollable growth of cells in the lung. This is a reason it causes serious breathing problems in the inhaling and exhaling part of the chest. This is caused primarily due to cigarette and passive smoking which is analyzed by the world health organization. The cases of death are also increasing day by day especially in youths and old people as compared to other cancers. As there is a good facility of medical Is present but unable to control the deaths. The best way is to take precautions early before increasing from the initial stage so that if the symptoms are visible can be found as early as possible [3].

Lung cancer is a difficult and dangerous cancer. this cancer can lead to loa t of death in men and women. Various methods have been used to detect lung cancer in an early stage. Here in this paper, some algorithms have been applied first in principal compared analysis, KNN, SVM, Machine learning techniques to detect include naive Bayes, decision trees, and artificial neural networks. Comparison has been done among these algorithms ANN gives the best accuracy of 82,43%with image processing and the decision tree without image processing gives 93.24% without image processing [2].

lots of cases of deaths are of lung cancer globally with the result of 5 million cases annually Lung cancer has a higher mortality rate than breast and prostate cancer combined. It can become better if the detection is made at an early stage. the survival rate. Different algorithms have been used and CT images are used for detection here the main focus is **on** lung cancer detection and categorization Large cell carcinoma, squamous cell carcinoma, and adenocarcinoma

are all types of lung cancer that may be detected using a newly discovered method.

they are compared based on feature set, extraction LBP and DCT. Machine learning algorithms that are used here are SVM and KNN which have been evaluated. The performance of these two algorithms has been observed and the accuracy rate achieved by these two is 93% by SVM and 91% by KNN [4].

Lung cancer is considered to be dangerous and difficult to detect it leads to a lot of deaths in both genders. there are several tied that has been implemented to detect lung cancer at starting state. here analysis has been done on the techniques. Most of them are done with the help of CT images and some analysis has been done with the help of x-rays. multiple classification methods are used with numerous segmentation algorithms in which image recognition is done to identify lung cancer. marker controlled watershed segmentation gives more accurate results. The result has been obtained by methods of deep learning that are achieved higher than the hod implemented using classical machine learning techniques [5].

Because lung cancer patients have a higher survival percentage when their disease is detected early. The methods of blood-based screening can boost patient uptake for lung cancer detection. Here, an innovative multidisciplinary method that combines machine learning and metabolic analysis is used to identify early-stage lung cancer. algorithms Around 110 people are detected with lung cancer and 43 are healthy people. The level of 61 plasma metabolic for targeted metabolic studies. there are 5 top important metabolic biomarkers developed by FCBF . Native Bayes are recommended as an exploitable tool for early lung tumor prediction [6][7]. Our approach is to detect lung cancer from CT scans using deep residual learning. The UNet and ResNet models may be used to extract characteristics and identify lung areas that are more susceptible to cancer. Additionally, the XGBoost and Random forest classifiers use a feature set that is passed through a number of classifiers. We attain an accuracy of 84% based on LIDC - IRDI's prior attempts [8].

OUR PROPOSED WORK

A developing technique called machine learning enables computers to automatically learn from previous data. Machine learning creates mathematical models and forecasts based on knowledge and past data using a range of techniques. Various activities, including speech recognition and picture identification, are presently carried out using it., email filtering, Facebook auto-tagging, and recommender systems. In this research paper, we are going to a predict lung cancer dataset by using algorithmic learning processes. Four machine learning algorithms—Random Forest, Logistic Regression, Support Vector Machine, and Decision Tree—are used in this perspective.

Static Vector Machine With this method, split and reverse are possible. Utilizing this technique is mostly done to address issues with segregation.. This algorithm creates moving lines that can divide n-dimensional space into classes.

Benefits: SVM functions best when there is a large gap between classes. High-density settings are where SVM performs best. When the maximum size exceeds the number of samples, SVM can be used. SVM functions well in memory. Disadvantages: For huge data sets, the SVM method is not appropriate. When the target classes overlap and there is extra noise in the data set, SVM does not perform very well. SVM will perform worse than expected when the number of features for each data point is greater than the number of training data samples.

Logistic Regression is known as a popular algorithm in machine learning. It is part of the supervised technique. This is used for the categorical dependent variables using the independent variable set. The output is given by logistic regression for categorical dependent values in the form of discrete values or categorical values.

Advantages of Logistic regression is easy to implement, makes no assumptions about distributions, and easily extends to multiple classes but disadvantages: if there are less number of observations than the number of features in that case we will not use logistic. regression constructs linear boundaries and the major A restriction is the supposition that a dependent variable and independent variables are linearly related. *Random Forest* This algorithm is the art of supervised learning, using collaborative reading. A technique called ensemble learning combines predictions from many machine learning algorithms to get predictions that are more accurate than those from a singlemachine.

Advantages of the Random Forest Algorithm are

It is known to be powerful and very accurate and gives good results in many problems including non-linear relationships.

Decision Tree this tree is considered the most powerful and famous tool for category and prediction. The selection tree structure is just like the shape of a tree in which the inner node affords an attribute, the branch offers a look, and the leaf ode affords a category label. in a selection, the tree source is split into subsets based on attribute fee. This system takes vicinity on every derived subset in a recursive way that's why it's far known as recursive partitioning [9][10].

In this research paper, we have used the lung cancer dataset which we had downloaded from Kaggle [11]. We have 13 parameters in the dataset through which we made the analysis and prediction by using machine learning algorithms. This research is conducted at COLAB. With Colab or "Collaboratory" we can write and run Python in our browser, which requires a zero-configuration, free access to GPU, and easy to share.

RESULTS

In the implementation, we have used four algorithms namely random forest, support vector machine, decision tree, and logistic regression. Tables 1,2,3 &4,

Presented the results on the parameters such as accuracy , recalls, harmonic mean, and precision in respect of each machine learning algorithm.

Accuracy is a measure for determining model relationships and patterns between variables in a data record, best based on input and training data.

The proportion of accurately categorized positive samples (true positives) to all classified positive samples is known as precision (true or false).

Match rate = True Positive / True Positive + False

Recalls are determined by dividing the total number of positive samples by the proportion of positive samples that were correctly identified as positive. The capacity of a model to identify positive samples is measured by recall. More positive samples were found the greater therecall.

Harmonic Mean is used in machine learning to calculate what is called an F-number or F-measure. The F-score is a test to evaluate the performance of an algorithm in information retrieval.

We have presented the random forest algorithm results in table 1. As per the results, we achieved an accuracy of 92.3% on the lung cancer dataset.

TABLE I. SHOWS RESULTS OF RANDOM FOREST

Accuracy	Precision	Recall	Harmonic Mean
92.3%	97%	94.1%	95%

Table 2 presented the support vector machine algorithm results with an accuracy 88.5% on the lung cancer database.

TABLE II. SHOWS THE RESULTS OF THE SUPPORT VECTOR MACHINE

Accuracy	Precision	Recall	Harmonic Mean
88.5%	91.5%	95.6%	93.5%

We have presented the decision tree algorithm results in table 3. As per the results, we achieved an accuracy of 91% on the lung cancer dataset.

TABLE III. SHOWS THE RESULTS OF THE DECISION TREE

Accuracy	Precision	Recall	Harmonic Mean
91%	96.9%	92.6%	94.7%

Table 4 presented the logistic regression algorithm results with an accuraof cy 89.the 7% on lung cancer dataset.

TABLE IV. SHOWS THE RESULTS OF LOGISTIC REGRESSION

Accuracy	Precision	Recall	Harmonic Mean
89.7%	94.1%	94.1%	94.1%

I. COMPARISON

A comparison table that summarizes results of all implemented algorithms are given below in Table 5.

TABLE V. SHOWS THE COMPARISON RESULTS OF MACHINE LEARNING ALGORITHMS

Parameters-➔ Algorithms	Accuracy	Precision	Recall	HarmonicMean
Random Forest	92.3%	97%	94.1 %	95%
Support Vector machine	88.5%	91.5%	95.6 %	93.5%
Logistic Regression	89.7%	94.1%	94.1 %	94.1%
Decision tree	91%	96.9%	92.6 %	94.7%

CONCLUSIONS

In this project, we have implemented the machine learning algorithms lung cancer dataset with 13 parameters to train the system. Specifically, we have used random forest, logistic regression, support vector machine and decision tree algorithm. This research is conducted at COLAB. With Colab or "Collaboratory" we can write and run Python in our browser, which requires a zero-configuration, free access to GPU, and is easy to share. As we have implemented four algorithms on the lung cancer dataset to check their perform based one of the four parameters i.e. accuracy, precision harmonic mean, recalled also presented the comparison of the four algorithms based on their results on four parameters. As per the results, the random forest algorithm provides an accuracy rate of 92.3 % on the lung cancer dataset. Subsequentially, the decision tree algorithm, support vector ad logistic regression performed with accuracy rates of 91%, 89.7%, and 88.5% respectively. In a comparison of the results, the random forest algorithm trained the cyst with a high accuracy rate.

REFERENCES

[1] M., P., Karaiskos, P., Kouloulias, V. et al. Lung cancer histology classification from CT images based on radionics and deep learning models. Med Biol Eng Comput 59, 215–226 (2021). https://doi.org/10.1007/s11517-020-02302-w

[2] B. Subrato, et al. "Comparative performance analysis of different classification algorithm for prediction of lung cancer." International Conference on Intelligent Systems Design and Applications. Springer, Cham, 2018.

[3] Günaydin, Özge, M. Günay, and Ö. Şengel. "Comparison of lung cancer detection algorithms." 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). IEEE, 2019.

[4] Xie, Ying, et al. "Early lung cancer diagnostic biomarker discovery by machine learning methods." Translational oncology 14.1 (2021): 100907.

[5] Amini, Mehdi, et al. "Overall survival prognostic modeling of non-small cell lung cancer patients using positron emission tomography/computed tomography harmonized radionics features: the quest for the optimal machine learning algorithm." Clinical Oncology 34.2 (2022): 114-127.

[6] Rehman, Amjad, et al. "Lung cancer detection and classification from chest CT scans using machine learning techniques." 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA).

[7] Chaudhary, A., V. N. Tiwari, and Anil Kumar. "Design an anomaly based fuzzy intrusion detection system for packet dropping attack in mobile ad hoc networks." 2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014.

[8] Abdullah, D. Mustafa, and N. Sadiq Ahmed. "A review of most recent lung cancer detection techniques using machine learning." International Journal of Science and Business 5.3 (2021): 159-173.

[9] Chaudhary, A., V. N. Tiwari, and A. Kumar. "A new intrusion detection system based on soft computing techniques using neuro-fuzzy classifier for packet dropping attack in manets." International Journal of Network Security 18.3 (2016): 514-522.

[10] Joshua, E. Stephen Neal, M. Chakravarthy, and D. Bhattacharyya. "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study." Rev. d'Intelligence Artif. 34.3 (2020): 351-359.

[11] https://www.kaggle.com/

[12] Chaudhary, A., V. N. Tiwari, and A. Kumar. "A cooperative intrusion detection system for sleep deprivation attack using neuro-fuzzy classifier in mobile ad hoc networks." Computational Intelligence in Data Mining-Volume 2. Springer, New Delhi, 2015. 345-353.

[13] Chaudhary, A., A. Kumar, and V. N. Tiwari. "A reliable solution against packet dropping attack due to malicious nodes using fuzzy logic in MANETs." 2014 International Conference on Reliability Optimization and Information Technology (ICROIT). IEEE, 2014.

[14] Chaudhary, A. "Mamdani and sugeno fuzzy inference systems' comparison for detection of packet dropping attack in mobile ad hoc networks." Emerging technologies in data mining and information security. Springer, Singapore, 2019. 805-811.

[15] Yadav, H., A. Chaudhary, and A. Rana. "Ultra Low power SRAM Cell for High Speed Applications using 90nm CMOS Technology." 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2020.