JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 25 Issue 04, 2022 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

IMPLEMENTATION OF MACHINE LEARNING TECHNIQUES TO DETECT A LUNG CANCER

Madhavi Aluka#[1], Sumathi Ganesan#[2], Vijay Pal Reddy P*[3]

#Department of Computer Science Engineering,
Annamalai University, Chidambaram,
Tamilnadu, India.
[1]alukamadhavi2000@gmail.com
[2]summi.ganesan@gmail.com
[3]drpvijayapalreddy@gmail.com
*Matrusri Engineering College, Hyderabad, Telangana, India

Abstract- Lung cancer is a dangerous disease that is seen in different age groups people. Different types of lung cancer are detected at the advanced stage. For early detection of cancer, some methods are used for the prediction of lung cancers. This paperwork discusses the Machine Learning (ML) models that are implemented for detecting lung cancer. Here implementing the Logistic Regression, Support Vector Machine and Random Forest Classifier algorithms are studied. Analyzing the accuracy performance of all these algorithms. The results obtained from the proposed model were found that the SVM is more reliable and produces enhanced accuracy when compared with other models.

Keywords – Machine Learning, Lung Cancer, SVM, Precision, Recall.

I. INTRODUCTION

Nowadays a lot of people are facing lung cancer diseases and late prediction of this disease leads to death sometimes. Due to the bacteria growth of tissues across the lung leads to cancer. Different types of cancer are observed. Some conventional methods are used for detecting lung cancer. This work uses different machine learning algorithms and finds a better model which detects cancer. Because nowadays machine learning is massively used and had a great influence on the health sector because of its state of predicting and detecting diseases correctly and accurately.

1. ADENO Carcinoma

Adenocarcinoma is a type of cancer that starts in the glands of the organs. This cancer is obtained when the glands provide the fluid to the organs, if that fluid crosses the organs and which is overflows then this type of cancer is observed [1]. This cancer is found in the outer part of the lung and starts increasing slowly. this type of cancer is mostly affected by the people who do smoking every day.

2. Large Cell Carcinoma (LCLC)

LCLC is another type of cancer that mostly starts at the corner of the outer part of the lung and rapidly starts to grow by spreading the cancer cells to the lung cancer [2]. There will be a severe problem while taking a breath is the first symptom that is observed in this cancer type.

Submitted: 24/11/2022 **Accepted**: 14/12/2022

3. Small Cell Carcinoma

Small cell lung carcinoma is also known as oat cell cancer which is similar to other cancer but this cancer spreads faster and people don't know until they get diagnosed [3]. This cancer is very quick because of this growth it will not respond to the basic treatment including chemotherapy.

4. Squamous Cell Carcinoma

This cancer starts in the squamous cell which is flat that lies inside the air pores of the lungs and tries to block the airways [4]. This cancer is seen in people who smoke a lot. There are different scan ways for analyzing cancer which include CT scan images, MRI, and so on.

The author states that machine learning algorithms can be used for predicting lung cancer pulmonary nodules. In this proposed work the author obtained the dataset using the support vector machine, random forest for detecting lung cancer. Here the strengths and the drawbacks, challenges are observed while developing these techniques [5]. Cancer occurs when abnormal cell tissues are generated at the lung. In this paper, the author is trying to predict lung cancer by using a machine learning algorithm which is SVM (Support Vector Machine). From this model, it is observed that the SVM and boosting model provides good results when compared to the other models [6]. Lung cancer is a very dangerous disease that is observed in different age groups people and a lot of people are dying due to late detection of cancer. The author discusses that lung cancer can be detected by using CT (Computed Tomography) scan images. In the medical CT scan images is a good technique that can be used for analyzing lung cancer. This technique helps doctors to find out the cancer cells accurately. This research is carried out for evaluating the different computer-aided techniques and finding the best technique by analyzing the drawbacks and disadvantages that are seen in the other techniques. Here the author proposed the machine learning model by using the CT scan images for detection. Support Vector Machine algorithm is used for detecting lung cancer bypassing the input data as CT scan images. This model provides 100% accuracy when compared with the other techniques [7]. From the past years, it is seen that lung cancer growth is increasing a lot and it became difficult for detecting lung cancer in the early stages. This paper provides the information related to image processing techniques along with the ML algorithms together. Here the researcher also used SVM, decision tree, and CNN, algorithms are used for detecting the lung CT cancer images. From all these algorithms it is observed that SVM produces more accuracy compared with the other remaining algorithms [8].

The cancer symptoms are seen when they are in the advantage stage, so it becomes very tough to analyze cancer. Because of this reason, there is a need for early detection of lung cancer. Histopathology lung cancer images are considered and predicting lung cancer by using Stochastic Diffusion Search (SDS). Through these algorithms, the features are extracted and the ML algorithms are used. Here decision tree, neural network, naive Bayes. From these algorithms, it is observed that NN provides more accuracy while dealing with other algorithms [9]. Another researcher carried out work on early detection of lung cancer diagnosis using machine learning algorithms. Here the author used the pioneering interdisciplinary mechanism and six ML algorithms are considered. Analyzing all these algorithms and finding the best

algorithms. From these algorithms, Naive Bayes algorithms provide good results [10]. Kourou et al. describe the prediction of the lung cancer prognosis using machine learning applications. This paper deals with the Neural Networks (NN), SVM, a Decision tree that is widely used in predicting lung cancer. When training the lung cancer data to all these models it is observed that NN has more accuracy whereas SVM provides less accuracy among the used ML algorithms [11].

Due to the increase of cancer cases, there is a necessity of implementing the model which tries to detect cancer in the early stages. This paper explores the usage of algorithms which includes SVM and CNN. This building model tries to classify lung cancer in the starting stages. Through this, a lot of people's lives can be saved. The author collected the UCI lung dataset which consists of both cancerous and non-cancerous images. The main motto of this paper is to classify lung cancer by using the WEKA tool. The results show that the SVM model produced an accuracy of 95.46% whereas CNN (92.11%) and KNN (88.40%).

According to Abdullah et al. [12], lung cancer is seen in both females when the cells are uncontrollable. Many medical facilities are available for predicting cancer but these medical facilities are not proving effective results. Here the author used Linear Regression, SVM, and Random Forest. The detection model of lung cancer is shown in Fig. 1. Initially, the author collected the lung cancer dataset from the UCI machine learning repository which consists of benign and malignant tumors. Firstly, the input data is processed and then the data is converted into a cancerous and non-cancerous set by using the Weka tool. After pre-processing the data then use the different ML algorithms by applying the processed data to that model. Observing the training and testing accuracies, precision, recall, and F1 score measures for all four models. The author discusses that after building the proposed model only RBF provided good accuracy with 81.25% while compared with other models. This model is considered an effective algorithm that is good for detecting lung cancer as shown in Fig. 1.

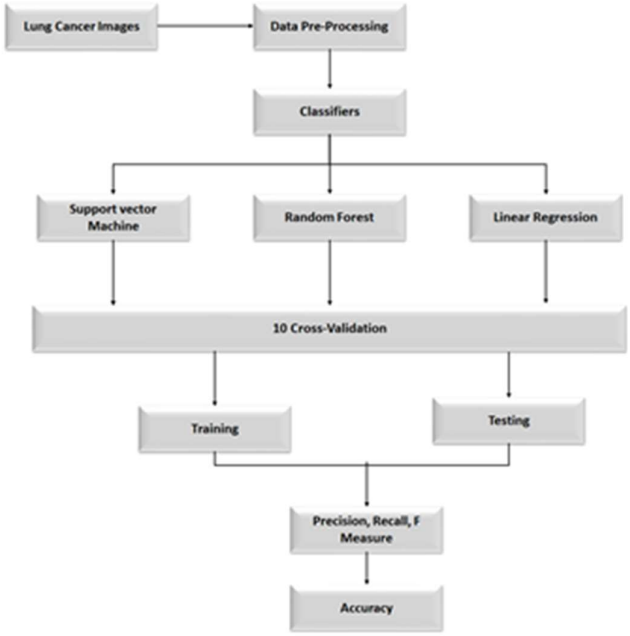

Fig. 1 Detection of Lung Cancer
Source: Adapted from [13]

Here the author takes the best selective methods which can detect and predict the lung cancer information from the dataset. The data consists of MRI scan images of lungs this data is firstly image processes where this module tries to reduce the noises and tries to prepare the image for segmentation. Based on the extraction of the image now the segmentation process will be taken place by enhancing the image by using some methods.IN feature extraction the image's pixel will be clearly extracted for analyzing eth image whether the obtained image is cancerous or non-cancerous. Then these data will be sent to the SVM model which tries to identify the lung cancer cell based on the extracted feature and provides the diagnosis results as lung cancer or non-cancerous [14]. The block diagram of lung detection is shown in Fig. 2.

Fig. 2 Lung Cancer detection Block diagram
Source: Adapted from [14]

Different existing methods have been analyzed and the different authors suggested their research work along with the models that they have used for detecting lung cancer. Based upon this research the proposed work is obtained by considering all the drawbacks that are observed in the existing methods.

In recent years lung, cancer is a severe problem faced by people. For early detection of lung cancer, there needs to be a model that detects cancer in the starting stage. This model helps the medical and health sector effectively which decreases the death rates. This paper addresses the Random Forest classifier model which is implemented for the detection of lung cancer. The author considered two lung cancer datasets and these two datasets are applied to the Random Forest model. The model tries to learn the data from the two datasets. Where for the first dataset the model produces an accuracy of 100% while coming to the second dataset the obtained accuracy is 96.31%. From this conclusion, the random forest classifier provided effective accuracy when compared with other methods [15].

The structure of this paper is provided as described. Section 2 discusses the proposed work that includes the algorithms that are implemented using ML. Section 3 presents the details about the results that are obtained for the ML model when the lung cancer dataset is applied to the model are analyzed. However, section 4 discusses the conclusion about the overall paperwork that is carried out for this project.

II. PROPOSED WORK

This section describes the methods that are used for carrying out the work. In machine learning, various algorithms can be used for detecting lung cancer. In this work, three machine learning algorithms are used for detecting lung cancer. Here Logistic regression, Support Vector Machine, and Random Forest Classifier are used.

Logistic regression is supervised machine learning which is used for predicting a certain class. This algorithm tries to find the dependent variables. This model predicts the output in the form of a categorical way. Here applying the gathered dataset to the logistic regression model when the model is trained with this data the model predicts the output as lung cancer or non-cancer. This algorithm is tough for calculating and analyzing complex regression problems. However, this model can only be used for discrete values and it is unable to predict the continuous changes in the values. But for this proposed work this model is suitable because of analyzing the classification problems.

Support Vector Machine (SVM) is a type of supervised learning which is mostly used for solving classification and regression problems. This model solves linear and non-linear problems. However, SVM takes the data points and the hyperplane line separates the data points. Initially, the model is trained by the labeled data, based on the classified labeled data the output will be predicted.

Random Forest Classifier is designed with multiple decision trees where these trees try to extract every feature from the lung cancer images. This algorithm works by selecting the sample data randomly from the given dataset. Now, this model builds each decision tree for extracting the features. Now predicting the output based on the provided features generated by the model.

Apart from these three algorithms, there are many more algorithms that are available in machine learning. But the other algorithms working and their analyzing step are different for the unsupervised learning methods. In the supervised learning model, these used models are very suitable because it deals with the classification issue.

To carry on this work for analyzing the performance of the model while detecting lung cancer are considered. Firstly, the dataset related to lung cancer needs to be considered. Here the dataset consists of 900 images with different cancer classes. Then the dataset needs to be processed and dividing the data into training and testing sets. The training set is applied to the ML models which includes Logistic Regression, SVM, and Random Forest classifier and finding the accuracy, loss of both training and testing sets. Analyzing the classification report for all three models.

III. RESULTS OF ML

In this section different machine learning models are used for analyzing the performance of the model when the dataset is passed through the ML algorithms. Here Logistic regression, SVM, and Random Forest algorithms are used for the detection of Lung Cancer stages which were identified in the image's dataset of a total of 900 images. The lung images deal with the different types they are ADENO Carcinoma (images), Large Cell Carcinoma (images), Small Cell Carcinoma (images), and Normal CT scans (images). Firstly, the dataset needs to be loaded and then pre-processing data for extracting each and every feature from the dataset. Then the processed data need to be randomly split into two tests which are training and testing sets. Now the training set is applied to the three ML algorithms for obtaining the actuaries and loss when the model is completely trained by the dataset.

```
lg = LogisticRegression()
lg.fit(X_train, y_train)
y_pred = lg.predict(X_test)
cr = classification_report(y_test,y_pred)
print("Classification report for - \n{}\n".format(cr))
```

```
Classification report for -
              precision    recall  f1-score   support

           0       0.76      0.74      0.75       100
           1       0.80      0.78      0.79        50
           2       0.93      0.98      0.95        42
           3       0.73      0.74      0.74        78

    accuracy                           0.79       270
   macro avg       0.80      0.81      0.81       270
weighted avg       0.78      0.79      0.78       270
```

Fig. 3 Plotting the Classification Report of Logistic Regression

The classification report shows the report values of the logistic regression model that is generated when the dataset is passed to the model. The report of the logistic regression is shown in Fig. 3. Here the X_train and y_train are fitted into the model for training the data and the X_test is used for testing the data. The classification report shows the values of the scores that the model generated are Precision, Recall, F1 score, and Support for the different types of Lung cancer which are represented as 0,1,2, and 3. Now analyzing the classification report for Random Forest Classifier, as shown in Fig. 4. Estimators are used for selecting the best and most accurate value based upon the model-generated observations.

```
: rfc = RandomForestClassifier(n_estimators=500)
  rfc.fit(X_train, y_train)
  y_pred = rfc.predict(X_test)
  cr1 = classification_report(y_test,y_pred)
  print("Classification report for - \n{}\n".format(cr1))
```

```
Classification report for -
              precision    recall  f1-score   support

           0       0.77      0.88      0.82       100
           1       0.90      0.72      0.80        50
           2       0.89      0.98      0.93        42
           3       0.89      0.79      0.84        78

    accuracy                           0.84       270
   macro avg       0.86      0.84      0.85       270
weighted avg       0.85      0.84      0.84       270
```

Fig. 4 Plotting the Classification Report of Random Forest Classifier

Similarly, the SVM model is built and analyzes the classification report when the data is sent to the model as shown in Fig. 5.

```
2]: print("Classification report for - \n{}:\n{}\n".format(
        clf, metrics.classification_report(y_test, y_pred)))

Classification report for -
GridSearchCV(estimator=SVC(),
             param_grid=[{'C': [1, 10, 100, 1000], 'kernel': ['linear']},
                         {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001],
                          'kernel': ['rbf']}]):
              precision    recall  f1-score   support

           0       0.76      0.74      0.75       101
           1       0.86      0.76      0.81        50
           2       0.94      0.98      0.96        45
           3       0.72      0.78      0.75        74

    accuracy                           0.80       270
   macro avg       0.82      0.82      0.82       270
weighted avg       0.80      0.80      0.80       270
```

Fig. 5 Plotting the Classification Report of Support Vector Machine

Finally based on the generated accuracies for the three algorithms concluding that the logistic regression (78%), Random Forest (80%), and SVM (83%). Also analyzing the Precision and Recall for all these algorithms. From the accuracies obtained from the model, it is observed that the SVM model provides good accuracy when compared with the other models. Table I shows the comparison of the three ML algorithms accuracies and their classification report.

TABLE I

COMPARISON OF MACHINE LEARNING ALGORITHMS BY APPLYING ON LUNG IMAGES FOR ACCURACY, PRECISION, AND RECALL

Algorithm	Accuracy	Precision	Recall
Linear Regression	**78%**	**78%**	**77%**
Random Forest	**80%**	**79%**	**80%**
SVM	**83%**	**83%**	**82%**

IV. CONCLUSIONS

The different Machine learning algorithms are analyzed from these algorithms it is observed that they could not achieve respective optimal accuracy from these ML algorithms as expected. Training complexity was also a drawback due to the consumption of the maximum amount of time to train the model. Because of this reason trying to implement the model by using deep learning techniques.

ACKNOWLEDGMENT

The authors would like to thank the management of Annamalai University for their support and for allowing us to carry out this scientific research.

REFERENCES

[1] L. Dell'Atti, and A. Benedetto Galosi. "Female urethra adenocarcinoma." Clinical genitourinary cancer 16, no. 2 (2018): e263-e267.

[2] M. Saadat, M. K. Manshadi, M. Mohammadi, M. J. Zare, M. Zarei, R. Kamali, and A. Sanati-Nezhad. "Magnetic particle targeting for diagnosis and therapy of lung cancers." Journal of Controlled Release (2020).

[3] J. Ko, M. M. Winslow, and J. Sage. "Mechanisms of small cell lung cancer metastasis." EMBO Molecular Medicine 13, no. 1 (2021): e13122.
[4] N. H. Patel. "The Role of Autophagy and Senescence in the Responses of Non-Small Cell Lung Cancer Cells to Chemotherapy and Radiation." (2021).
[5] C.M. Lynch, B. Abdollahi, J. D. Fuqua, R. Alexandra, J A. Bartholomai, R. N. Balgemann, V. H. van Berkel, and H. B. Frieboes. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." International journal of medical informatics 108 (2017): 1-8.
[6] A. Chauhan. "Detection of lung cancer using machine learning techniques based on routine blood indices." In 2020 IEEE international conference for innovation in technology (NEOCON), pp. 1-6. IEEE, 2020.
[7] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi. "Lung cancer detection using CT scan images." Procedia Computer Science 125 (2018): 107-114.
[8] V.J. Pawar, K. D. Kharat, S. R. Pardeshi, and P. D. Pathak. "Lung Cancer Detection System Using Image Processing and Machine Learning Techniques." Cancer 3 (2020): 4.
[9] S. Shanthi, and N. Rajkumar. "Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods." Neural Processing Letters 53, no. 4 (2021): 2617-2630.
[10] Y. Xie, W. Y. Meng, RZ.Li, Y.W. Wang, X. Qian, C. Chang. Z. F. Yu et al. "Early lung cancer diagnostic biomarker discovery by machine learning methods." Translational oncology 14, no. 1 (2021): 100907.
[11] 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 13 (2015): 8-17.
[12] 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 1, no. 2 (2021): 141-149.
[13] R. Patra. "Prediction of Lung Cancer Using Machine Learning Classifier." In International Conference on Computing Science, Communication and Security, pp. 132-142. Springer, Singapore, 2020.
[14] A. Asuntha, A. Brindha, S. Indirani, and A. Srinivasan. "Lung cancer detection using SVM algorithm and optimization techniques." J. Chem. Pharm. Sci 9, no. 4 (2016): 3198-3203.
[15] A. Rajini, and M. A. Jabbar. "Lung Cancer Prediction Using Random Forest." Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 14, no. 5 (2021): 1650-1657.