Biological and Clinical Sciences Research Journal
eISSN: 2708-2261; pISSN: 2958-4728
www.bcsrj.com
DOI: https://doi.org/10.54112/bcsrj.v2023i1.187
*Biol. Clin. Sci. Res. J., Volume, **2023**: 187*
Original Research Article

EARLY DETECTION OF LUNGS CANCER USING MACHINE LEARNING ALGORITHMS

ANWAR MR[1], *BAKAR MA[1], AWAIS HM[2], DIN MU[1], MOHSIN M[1], NAZIR MA[1], SAQIB I[1], KHALID MM[1]

[1]*Ripah International University Faisalabad, Pakistan*
[2]*Faisalabad Medical University, Allied Hospital Faisalabad, Pakistan*
**Correspondence author email address: mabubakar1122@pgmail.com*

(Received, 7th September 2022, Revised 21st January 2023, Published 23rd January 2023)

Abstract: *Medical healthcare systems store a large amount of clinical data about patients related to their biographies and disease information. Doctors use clinical data for the early detection of diseases that helps with proper patients' treatments to save their lives. These clinical systems are helpful in detecting cancer diseases at early stages to save people's lives. Lung cancer is the third largely spreading disease in human beings all over the globe, which may lead so many people to death because of inaccurate detection of their disease at the initial stages. Therefore, this study will help doctors and radiologists in the detection of lung cancerous and non-cancerous patients at early stages with a random forest algorithm to save patients' lives. In this research work, a new and novel model based on random forest algorithm was employed to detect lung cancer from the Wisconsin data set. Lung cancer was detected at early stages, and it was decided whether targeted patient was cancerous or non-cancerous. This experimental outcome showed that the proposed methodology achieved an accuracy rate that was batter compared to previous studies for early detection of lung cancer.*

Keywords: lung cancer, stage, detection, accuracy

Introduction

It is well said, "Health is Wealth". However, a huge number of human beings die yearly because of different diseases; cancer is also prominent among those diseases. There are many cancers like skin cancer, breast cancer, Lungs Cancer etc. Cancer is a widely and rapidly spreading disease in human beings. A huge number of human beings die every year because of this disease. Lung cancer is the third largely spreading disease in humans all over the globe, which may lead many people to death because of inaccurate detection of their disease at the initial stages. It is also estimated that nearly 10,000 new lung cancer patients are detected annually in Pakistan.

Table 1: Problem Statement and this Research was organized on following bases.

Problem Statement	Purposed Methodology	Data Set	Parameters
Early detection of Lungs Cancer is not manually possible	Modal is based on random forest algorithm, an algorithm of Machine learning.	Wisconsin data set of UCI repository will be used.	Four Comparison Parameters will be used. Accuracy, Precision Rate, F1-Score & Recall.

A. Problem Background: There are many reasons of spreading cancer in the world, like un-healthy food, and it is a common disease in chain smokers and glass smokers. Lung's cancer is a very dangerous disease in humans, and the mortality rate of lung cancer patients is increasing. Early detection of disease is very important to reduce the mortality rate. However, the main problems are initially ignorance about this disease by patients' side; secondly, traditional manual approaches for the detection of lung cancer are being used by doctors' side owing to a lack of facilities & advance equipment.

These are major reasons for the increasing mortality rate of lung cancer patients worldwide, especially in poor and developing countries. It is an alarming situation for the world, and the question is how this situation can be handled and reduce the mortality rate of Lungs Cancer patients. Now a days, the world is transforming itself rapidly, and methods of detecting diseases and treatments are changing rapidly. Now early detection of diseases is necessary for increasing human beings' survival rate. Advance machinery have been used like CT scan, MRI machine, X-Ray and ultrasound machines for pre-detection of diseases, generating image-based data for doctors. Mostly, doctors use traditional manual methods for disease detection and somehow, they fail in exactly detecting

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.*, **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]

lung cancer disease. Suppose they detect that the patient is suffering from lung cancer but are not sure about the stage of lung cancer, such as a first, second or third stage.

To handle these types of situations, artificial intelligence has a great role. Machine learning algorithms are abundantly used in medical image processing worldwide. In previous research, it is observed that various classification techniques (QUEST, CHAID, CART, C5.0, NN, LR, SVM, DA and NB) have been used to measure the accuracy of algorithms by using 9 datasets (Adult, House, Credit, Segment, Car, White Wine, Red Wine NHANES & Vehicle). They analyzed which algorithm produced the best results with which dataset. Too much data is available regarding medical images, and is increasing daily. So many difficulties are being faced with finding meaningful patterns and relevant information from huge amounts of ungrouped and unlabeled data. Data mining is also a vast field of research. Data mining has many techniques for mining relevant data. Important data mining techniques are classification, regression, clustering, association rules, sequential patterns and prediction. There are also R-language, outer detection and oracle data mining techniques etc. Data mining techniques greatly help researchers and companies get knowledge-based information from junk data.

Table 2: Comparative Analysis of Existing Approaches

Author & Year	Title	Proposed Model & Algorithms	Limitations	Data Set	Accuracy
Tripathi, S., (2022**)**	**Radgennets: Deep learning-based radiogenomics model for gene mutation prediction in lungs cancer**	**Convolutional Neural Networks & Dense Neural Networks.**	**Prediction of gene mutation is the crucial problem in the previous work**	**130 patients PET/CT scans**	**Area under curve (AUC) score 94%**
Boddu, R. S. K., (2021)	**Analyzing the impact of Machine Learning and Artificial Intelligence and its effect on the management of Lung Cancer detection in covid-19 pandemic**	**Machine Learning & Deep Learning Algorithms**	**Fusion of ML&DL not used in old Methodologies**	**Chest Computer Tomography (CT) scan**	**General detection Is done with this fusion**
Ubaldi, L., (2021)	**Strategies to develop Radiomics & machine learning models for Lung Cancer stage & Histology prediction using small data samples**	**Fusion of Radiomics & Machine Learning Algorithms (Random Forest)**	**Radiomics fusion is not implemented in previous work**	**MAASTRO NSCLC collection (130 patient's data)**	**Lung1 & Testing of L-RT dataset (AUC = 0.72 ± 0.04 for Random Forest & AUC = 0.84 ± 0.03**
Savitha, G., & P. Jidesh (2020)	**A holistic deep learning approach for identification and classification of sub-solid Lung Nodules in computed tomographic scans**	**CRF+DCCN with SoftMax Classifier**	**CRF is not used in old work**	**LIDC/IDRI**	**89.48%**
Author &	**Title**	**Proposed Model**	**Limitations**		

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.,* **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]

Year		& Algorithms		Data Set	Accuracy
Bhaskar,N., & Ganashree, T. S. (2020)	**Lungs Cancer Detection with FPCM and Watershed Segmentation Algorithms**	**SVM**	**FPC & Watershed Transform calculations are not used in previous work**	**50 CT pictures**	**60%**
Shakeel, P. M., et al, (2019)	**Lungs cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks**	**DITNN**	**Profuse clustering is unique technique and not used previously**	**Cancer imaging Archive (CIA)**	**98.42**
Xie, H. (2019)	**Automated pulmonary nodule detection in CT images using deep Convolutional neural networks**	**R-CNN 2D CNN CAD systems**	**Three networks are not used in previously**	**LIDC-IDRI LUNA16 dataset**	**CPM score of 0.790**

B. Contribution of this study

This study proposed the best model to process lung images by using classification modalities of random forest algorithm, an algorithm of machine learning for detecting cancer at early stages regarding accuracy and sensitivity rate. Moreover, this research used feature extraction techniques such as thresholding and image histograms to extract features from images. Before this research, these two techniques had not been used together. Pradeep K R et al., (2018) evaluated classification techniques for the prediction of survivability chances of lung cancer patients by using healthcare analytics. He wants to facilitate doctors for better detection of cancer survivability rate and to predict how many years a patient will survive so that the next treatment can be planned. This research implemented classification modalities to generate a hidden information pattern for predicting cancerous and non-cancerous images. This study evaluated accuracies, sensitivity and area under the curve with the random forest model and then compared it with other existing models. This study will help doctors and radiologists for the detection of lungs cancer in CT scan images at early stages with better accuracy. A lot of work has already been done, and still, working is continued. It will be very beneficial for patients to save their lives and for doctors to pre-detect disease at an early stage of the disease. The aim of this research is that doctors could decide on further treatments according to the severity of the disease in the future.

Problem Statement

Early detection of lung cancer is very hard for doctors; they use traditional manual approaches and methods for the detection of lung cancer among patients. Human error chances are included in this way.

Research Questions

How to Early detect lungs Cancer using a random forest machine learning algorithm?

How to compare the random forest algorithm model with other existing models?

Research Objective

The first objective is to provide an early detection system to doctors for detecting cancerous and non-cancerous patients. The second objective is to improve the detection accuracy of lung cancer with proposed model, which will be comparatively better than other existing models. This proposed model will help doctors define stage (severity) of lung cancer disease so that doctors decide further treatments according to the severity (stage) of lung cancer disease in the future.

Methodology

This research work fell under the supervised random forest machine learning algorithm and data set. In this research work, the Wisconsin data set of the UCI repository has been taken. This data set is publicly available at the UCI repository. In this data set, features were extracted from data in which some noise is present in the form of irrelevant data types. This is because of some reasons as to make the dataset unknown. This research work has been divided into phases for good management andbetter execution of

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.,* **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]

processes in experiments. These phases consist of Data Processing, Features Selection, Classifications & Evaluation and Comparisons of Accuracies with other models. Below, there have been represented all the phases in graphical form.

Phase 1: Data Processing

Data Set carries data type constraints and missing values. Firstly, there was to find data type constraints and change these data type constraints according to the required data type. In the second stage of data processing, missing values were dealt with in python coding.

Phase 2: Features Selection from Raw Data

In 2nd phase, firstly, all missing values were filled with the mean function. It is because the mean function fills null value with the appropriate mean value and manually selects features. Most of the time, the model is responsible for features extraction, but four features were selected were manually 1. Shape 2. Texture 3. Geometry 4. GLCM.

Phase 3: Classifications and Evaluations

In 3rd phase, the data was divided into 80:20 ratios, which means 80% data is for classification and 20% for testing classified results. In this phase, the 10K-fold cross validation method was used, which means that 80% data set was divided into similar 10 folds, and 20% data was divided concerning training data. All models will be trained in 10 iterations and tested on every iteration. Consequently, mean accuracy was used for the results. Evaluations of models were performed through training accuracies. In this dissertation Model, the proposed random forest algorithm was used for classifying labeled data. A refine and agile model was built using a random forest machine learning algorithm for classification & evaluation regarding benign and malignant patients.

Phase 4: Comparison of Accuracy with Other Models

In the 4th phase, after implementing a random forest machine learning algorithm, there was a stage of collecting data by applying four parameters to results. These four proposed parameters are accuracy, precision rate, F1-score and recall. These parameters were calculated from the confusion matrix of the random-access algorithm. After gaining results, a comparison of results was performed with other models. The accuracy of the random forest model was better in comparison with other models for the early detection of lung cancer. Based on this accuracy for early detection of lung cancer, there was detected lungs cancer at an early stage, and it was decided that the target patient was cancerous or non-cancerous.

Results and Discussion

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{F1-score} = 2\,\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

This research is based on supervised machine learning methods and uses the UCI repository for fetching the lung cancer dataset; all data description is presented in the the data description. All research is based on improving the accuracy of state of the art models in this research, four algorithms are used logistic regression, decision tree, random forest and support vector machine and we already have discussed these algorithms in methodology. This research study worked on the following parameters as accuracy, precision, recall, F1-score and Precision, recall and F1-score is calculated from confusion matrix as,

Here, TP=True Positive, TN= True Negative, FP= False Positive and FN= False Negative.

Parameters

In this study, the research has found the best-suited model with maximum accuracy and measured the parameters as mentioned earlier as Recall, Precision and F1-score. The necessary description of all used parameters is the below section.

Accuracy

The research-based definition of accuracy is "Accuracy is the degree of closeness between measurement results and the true or reference value". Whereas precision is "a measurement is a measure of the reproducibility of a set of measurements".

Confusion Matrix

The confusion matrix is a technique used to summarize the performance of a classification algorithm. If the number of cases in each category is not the same, or if your data set has more than two categories, only the accuracy of the one can be incorrect. A confusion matrix is an important matrix to find out the values of TP, TN, FP, FN and these values are used for defining the parameter according to the above formulas. In the experimental result, all applied models to draw the confusion matrix as presented in table 3.

Table 3 Represents the Confusion Matrix of Applied Models

Algorithms	Arrays of all predictive values
Logistic Regression	[[0 1 0] [0 1 1] [0 0 3]]
Decision Tree	[[0 1 0] [0 1 1] [1 1 1]]
Random Forest	[[1 0 0] [0 1 1] [0 0 3]]
Support Vector Machine	[[0 1 0] [0 2 0] [0 2 1]]

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.*, **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]

Recall

Recall is defined as in research point of view "What are the results of images in computer research Retrieval on retrieval of information is a subset of relevant documents that have been successfully retrieved. For example, for a text search in a set of documents, the retrieval speed is the number of results divided by the number of results that should return the correct number of results."

F1-score

F1- score combines the precision and recall of the classifier into a single metric using the harmonic mean. It is mainly used to compare the performance of two classifications.

Logistic Regression

Logistic regression is a statistical model, a basic form for modeling binary dependent variables using logistic functions, although there are more complex extensions. In regression analysis, logistic regression (or logit regression) estimates the parameters of a logistic model (a form of binary regression). Logistic Regression has two types as Linear Regression and Logistic Regression. Logistic Regression was used in this research work and achieved the results as in Model 0: logistic Regression's accuracy of 67%, precision of 41.67%, recall of 50% and F1-score 45.34%.

Decision Tree

Decision Tree is a flowchart-like structure where each internal node represents a "test" for a property (for example, whether a coin run goes up or down), each branch represents the test result, and each leaf node represents the result of the test. Class Label (determined after calculating all properties). In our research, the results of Decision Tree as Model one represents the accuracy of 33%, precision of 27.66%, recall 27.67% and F1-score 26.67% as in the figure below. This study achieved the lowest results among all applied models.

```
model 1
[[0 1 0]
 [0 1 1]
 [1 1 1]]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         1
           2       0.33      0.50      0.40         2
           3       0.50      0.33      0.40         3

    accuracy                           0.33         6
   macro avg       0.28      0.28      0.27         6
weighted avg       0.36      0.33      0.33         6
```

Figure 1: Results of Decision Tree Model

Random Forest

Random Forest or Random Decision Forest works by building multiple decision trees during training as a holistic learning method for classification, regression and other tasks and by producing classes as class models or predictions means / averages of a single tree. Our research's best Model 2 as Random Forest represents the highest accuracy of 83%, precision. 91.67%, recall 83.34% and F1-score 84.34%.

Support Vector Machine

```
model 2
[[1 0 0]
 [0 1 1]
 [0 0 3]]
              precision    recall  f1-score   support

           1       1.00      1.00      1.00         1
           2       1.00      0.50      0.67         2
           3       0.75      1.00      0.86         3

    accuracy                           0.83         6
   macro avg       0.92      0.83      0.84         6
weighted avg       0.88      0.83      0.82         6
```

SVM (Support Vector Machine) is a special linear classifier based on the principle of maximizing margins. This increases the complexity of the classifier to achieve good generalization performance by minimizing structural risks. The Second lowest results of Model 3 as a Support Vector Machine represent an accuracy of 50%, precision of 46.6 7%, recall of 44.34% and F1-score of 35.67%.

```
model 3
[[0 1 0]
 [0 2 0]
 [0 2 1]]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         1
           2       0.40      1.00      0.57         2
           3       1.00      0.33      0.50         3

    accuracy                           0.50         6
   macro avg       0.47      0.44      0.36         6
weighted avg       0.63      0.50      0.44         6
```

Figure 2: Results of Random Forest Model

Figure 3: Results of Random Forest Model

Comparison of the Experimental Research

This research work implemented classification modalities to classify cancerous and non-cancerous patients. This study evaluated our model's accuracy, precision, recall and area F1-score and then compared it with other existing models to find the accurate model. This study will help doctors and radiologists predict lung cancer in CT scan images at early stages with better accuracy as detailed discussion presented in our dissertation. The all-comparative results are mentioned in table number 4 below. In this research, it is noticed that on the image dataset the direct implementation for classification Support Vector Machine achieved the maximum classification results, but, in our case, SVM gained the second lowest results. Regression algorithms perform very well on unlabeled data most of the time. Even though this research work has made the 10,000 iterations but not achieved the maximum results as compared to Random Forest Model.

Table 4 Comparison with Previous Research work

Modals	Accuracy	Precision	Recall	F1-Score

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.*, **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]

Random Forest	83%	91.67%	83.34%	84.34%	**KNN**	77%	25 %	16.66%	30%

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	67 %	41.67%	50%	45.34%
Decision Tree	33%	27.66%	27.67%	26.67%
Random Forest	**83%**	**91.67%**	**83.34%**	**84.34%**
Support vector Machine	50%	46.67%	44.34%	35.67%

Table 1 Represents the All Models Accuracy, Precision, Recall and F1-Score

In table 4, it is clear about the performance of Random Forest as highlighted in table. Here the important thing is that the values of precision, recall and F1-score are label-wise counted in all models results, but in the above table, we took the average value of three labels.

Conclusion

The experimental outcome shows that the proposed methodology achieved an 83% accuracy rate as compared to other models and previous studies. Furthermore, this study will help doctors and radiologists predict lung cancer at early stages to save patients' lives. This research used the UCI repository dataset, which contains the missing values, and those missing values create the noise in the dataset. This work removed these values with the mean function. This model used four algorithms: Logistic Regression, Decision Tree, Random Forest and Support Vector Machine. These are the frequently used algorithms in research of labeled based image data. The previous work shows that using the KNN algorithm achieved an accuracy of 77%, which is sufficient but not reliable. This study work chooses the other four algorithms, get the maximum accuracy and compares the results with each other. This research work calculated the average of three labels value in one parameter. Respectively, all models calculated the values as average values and made the table 2 to find out the best-suited model. This research provides the Random Forest algorithm as an efficient algorithm not only accuracy but as for precision, recall and F1-score. It achieved the accuracy, precision, recall and F1-score 83%, 91.67%, 83.34%, and 84.34%, respectively. These are the highest results among the other three algorithms and logistic Regression is the 2nd algorithm among remaining two algorithms with the results as accuracy, precision, recall and F1-score, 67 %, 41.67%, 50%, 45.34% respectively. Remaining two models are below average of the results. In the future, there is the plan to use another remaining algorithm with this dataset and the image-based dataset without already extracted features. There will use some optimizer for enhancement of the model accuracy.

Conflict of interest

The authors declared absence of conflict of interest.

References

Boddu, R. S. K., Karmakar, P., Bhaumik, A., Nassa, V. K., & Bhattacharya, S. (2022). Analyzing The Impact of Machine Learning and Artificial Intelligence and its Effect on Management of Lungs Cancer Detection in covid-19 Pandemic. *Materials Today: Proceedings*, **56**, 2213-2216.

Dabeer, S., Khan, M. M., & Islam, S. (2019). Cancer diagnosis in histopathological image: CNN based approach. *Informatics in Medicine Unlocked*, **16**, 100231.

de Oliveira Torres, W., de Carvalho Filho, A. O., Rabêlo, R. D. A. L., & e Silva, R. R. V. (2020). Texture analysis of lung nodules in computerized tomography images using functional diversity. *Computers & Electrical Engineering*, **84**, 106618.

Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., ... & Frieboes, H. B. (2017). Prediction of lungs cancer patient survival via supervised machine learning classification techniques. *International journal of medical informatics*, **108**, 1-8.

Moitra, D., & Mandal, R. K. (2020). Classification of Non-Small Cell Lungs cancer using One-Dimensional Convolutional Neural Network. *Expert Systems with Applications*, 113564.

Nasser, I. M., & Abu-Naser, S. S. (2019). Lungs cancer Detection Using Artificial Neural Network. *International Journal of Engineering and Information Systems (IJEAIS)*, **3**(3), 17-23.

Tripathi, S., Augustin, A. I., Moyer, E. J., Zavalny, A., Dheer, S., Sukumaran, R., ... & Kim, E. (2022). Radgennets: Deep learning-based radiogenomics model for gene mutation prediction in lungs cancer. bioRxiv.

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.*, **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]

Tripathi, S., Augustin, A. I., Moyer, E. J., Zavalny, A., Dheer, S., Sukumaran, R., ... & Kim, E. (2022). Radgennets: Deep learning-based radiogenomics model for gene mutation prediction in lungs cancer. bioRxiv.

Ubaldi, L., Valenti, V., Borgese, R. F., Collura, G., Fantacci, M. E., Ferrera, G., ... & Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lungs cancer stage and histology prediction using small data samples. *Physica Medica*, **90**, 13-22.

Xie, Y., Zhang, J., Xia, Y., Fulham, M., & Zhang, Y. (2018). Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Information Fusion*, **42**, 102-110.

Zhang, Q., Wang, H., Yoon, S. W., Won, D., & Srihari, K. (2019). Lung nodule diagnosis on 3D computed tomography images using deep Convolutional neural networks. *Procedia Manufacturing*, **39**, 363-370.

[Citation: Anwar, M.R., Bakar, M.A., Awais, H.M, Din, M.U., Mohsin, M., Nazir, M.A., Saqib, I., Khalid, M.M. (2023). Early detection of lungs cancer using machine learning algorithms. *Biol. Clin. Sci. Res. J.,* **2023***: 187*. doi: https://doi.org/10.54112/bcsrj.v2023i1.187]