**LUNG CANCER DETECTION**

**A PROJECT REPORT**

***Submitted by***

**PRATHEEPA R (2116210701192)**

***in partial fulfilment for the award of the degree***

***of***

**BACHELOR OF ENGINEERING**

**in**

**COMPUTER SCIENCE AND ENGINEERING**



**RAJALAKSHMI ENGINEERING COLLEGE**

**ANNA UNIVERSITY, CHENNAI**

**MAY 2024**

# RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

# BONAFIDE CERTIFICATE

Certified that this Thesis titled **“LUNG CANCER DETECTION USING AN ENSEMBLE OF DIFFERENT CLASSIFICATION MODELS AND REGULARIZATION TECHNIQUES”** is the bonafide work of **“PRATHEEPA R (2116210701192)”** who carried out the work under my supervision.Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**SIGNATURE**

Dr. S. Senthil Pandi M.E., Ph.D.,

**PROJECT COORDINATOR**

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai – 602 105

Submitted to Project Viva-Voce Examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Internal Examiner External Examiner**

# Lung Cancer Detection using an Ensemble of Different Classification Models and Regularization Techniques

*Pratheepa R*

*Department of Computer Science and Engineering*

*Rajalakshmi Engineering College*

**ABSTRACT**

Cancer is a situation where there is abnormal growth of cells in the body. It occurs due to changes in the DNA caused by various reasons, which causes uncontrollable division of cells that may spread to other tissues, causing various complications. Lung cancer is a type of cancer in which there is a growth of tumour in the lungs, severely affecting the organ. It may also spread to other parts of the body. Cancer may even lead to death, if left untreated. So, it becomes very important to diagnose cancer at an early stage and start treatment in order to avoid further complications. This study classifies a person as having cancer or not, based on the symptoms. The model uses an ensemble of the decision tree algorithm, the support vector machine algorithm, the random forest algorithm and gradient boosting algorithm. It also uses regularization techniques in order to prevent the data from overfitting or memorizing the training data, instead of learning the patterns in it. The proposed system has achieved an accuracy of 98.38%.

**INTRODUCTION**

Lung cancer is one of the most common types of cancer, which originates from a tumour growth in the lungs, and spreads to other parts of the body. The abnormal growth of cells causes severe problems and also leads to death, if not diagnosed early and treated properly. The World Health Organization (WHO) says that lung cancer is the leading cause of cancer death in recent years, accounting for 18% of all cancer deaths. It is more common in men than in women. Some common symptoms of lung cancer include fatigue, shortness of breath, coughing, wheezing and chest pain. Having these kinds of symptoms as the base factors or the features, the model is trained on a dataset that contains these symptoms as the features, and the target variable is a prediction on whether the person is having lung cancer or not. It is a binary classification model. The values for all the features are also values indicating yes or no. Based on this kind of data availability, a more appropriate algorithm for classification is the Decision Tree algorithm. It focuses on calculating the information gain for all the attributes, by computing the difference in entropy of states. But rather than relying on one particular model or algorithm, this system is trained on multiple suitable algorithms and predicts the final answer as a majority of the answers predicted by the individual models. Ensemble learning is a technique where multiple models are put to train, and are combined as one single model. This improves the accuracy of the system, rather than when relying on a single model. This proposed system makes use of Decision Tree algorithm, Support Vector Machine algorithm which has proven to do better on higher dimensions, Random Forest algorithm which itself is an ensemble of various decision trees, Gradient Boosting algorithm which is also an ensemble that combines multiple weak models and provides a strong model. This trained model which is an ensemble of multiple algorithms, showed to overfit the data, and achieved an accuracy of 87.34%. After applying regularization technique on the Gradient Boosting algorithm, the model ceased to overfit, and achieved an improved accuracy of 98.38%.

**LITERATURE REVIEW**

In a study conducted by Qurina Firdaus, Riyanto Sigit, Tri Harsono and Anwar titled “Lung Cancer Detection Based On CT-Scan Images With Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods”, it has been discussed on the various techniques used by them for preprocessing the images to improve the quality of the image, the segmentation methods used to specifically capture the expected cancer part in the image, and the prediction methods used to finally classify the image as being benign or malignant. The accuracy level recorded was 83.33%. In another study by “Nusraat Nawreen”, “Umma Hany”, “Tahmina Islam” titled “Lung Cancer Detection and Classification using CT Scan Image Processing”, it has been discussed on the various preprocessing techniques used by the authors to improve the quality of the image, the methods used to detect the edges in the image in order to find the expected cancer affected part, and prediction methods used to finally classify the image based on the severity levels using a support vector machine classifier. In a study by “Mattakoyya Aharonu”, “R Lokesh Kumar” titled “Convolutional Neural Network based Framework for Automatic Lung Cancer Detection from Lung CT Images”, they have discussed about using a convolutional neural network for classifying the images based on the severity levels. They have mentioned about the various preprocessing techniques adopted by them to smooth the image to make it suitable for correct prediction, the layers present in the convolutional neural network created by them, the methods or techniques used by them so as to train the model accurately, and also about the prediction methods used to classify the image. In a study by “Shubhada Agarwal”, “Sanjeev Thakur”, “Alka Chaudary” titled “Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis”, it is discussed on using various machine learning algorithms like the decision tree algorithm, support vector machine algorithm, logistic regression algorithm and the random forest algorithm. Comparing all the accuracy scores obtained by each of the models, the best algorithm to be used is concluded based on which individual model has secured the maximum accuracy. In a study by “Devananda Rao”, “Mahammad Arshad” titled “Early Detection of lung Cancer Using Machine Learning Technique”, it is discussed on the steps adopted by them in collecting the data required for their research, preprocessing the data to make it suitable for training, the methods used for enhancing the quality of the image, and techniques used for detecting whether a particular situation is cancerous or not. This would actually help in detecting lung cancer at an early stage, and further help in timely treatment. In a study by “Amita Dessai”, “Moffy Vas” titled “Lung cancer detection system using lung CT image processing”, it is discussed on using an artificial neural network to use certain morphological methods to separate the expected cancer part from the rest of the image, and further using some prediction techniques to classify what type of cancer it is, mainly focusing on lung cancer. In a study by “Akash Vishwakarma”, “Aditya Saini”, “Kalpana Guleria”, “Shagun Sharma” titled “An Early Prognosis of Lung Cancer using Machine Intelligence”, is discussed on using various kinds of machine learning algorithms such as the naive bayes algorithm, the decision tree algorithm, logistic regression algorithm, the random forest algorithm and the multilayer perceptron in detecting lung cancer well before it aggravates and causes danger. The study is about finding which algorithm is the most suitable one for this use case, by comparing the accuracy values achieved during their study.

**PROPOSED METHODOLOGY**

The steps performed in the process of developing the proposed methodology include data collection which consists of the input part and the target value predicted as the output, preprocessing the data which includes filling up null values, correcting incorrect values, and converting string values to integer for proper computation, training on different algorithms both individually and collectively like using an ensemble, comparing their accuracy rates and finding out the best model which could predict whether the person has lung cancer or not based on the symptoms recorded.

DATA COLLECTION:

The data required for this study is collected in such a way that it contains the various symptoms of people including their age and gender. The various kinds of symptoms included in the dataset are swallowing difficulty, fatigue, shortness of breath, chest pain, and details regarding their smoking and alcohol consumption habits. The values for each symptom are also a binary answer indicating whether the particular symptom is present or not. Given the symptoms of people, the target value is a binary answer indicating whether the person has lung cancer to not. Based on this data collected, the model is trained on various algorithms after preprocessing the content.

PREPROCESSING:

The values for the gender attribute are collected as male or female, which is actually present as a string. The model needs only numeric values during the training part. So, the string values are converted or mapped into an integer value. Similarly, for the target variable, the value is present as a “yes” or “no”, which is also converted into an integer variable with each integer representing a particular category. The integer chosen to map the string values can any random ones as the model is only going to classify into one particular category and not going to predict a continuous value which might depend on the integers chosen. So, it just needs a label to identify the corresponding string. After the preprocessing part is done which mainly includes converting strings into integer values, suitable algorithms are chosen to train the model on.

ALGORITHMS:

Algorithms which are suitable for classification problems are chosen in order to test their accuracy against each other, and also try different kinds of ensembles for their accuracy. The best kind of model was chosen based on the accuracy achieved.

SYSTEM:

The dataset was initially trained against the individual algorithms, and had their accuracies recorded. Later, the proposed system was built which showed the maximum accuracy. The proposed system is an ensemble of different classification models which deploys four different algorithms such as the Decision Tree algorithm, the Support Vector Machine algorithm, the Random Forest algorithm and the Gradient Boosting algorithm.

DECISION TREE:

Decision tree is a supervised classification machine learning algorithm that constructs a tree like structure based on the attributes. It decides the attribute to split on, at every stage and creates the tree. The entropy of every attribute before splitting, and the entropy of the state after splitting based on that attribute is calculated. The difference between the entropy values is calculated as the information gain. The information gain of all the attributes is computed. The attribute with the highest information gain is chosen to be split on. In this manner, the decision tree is constructed until all the data has been split into its corresponding classifications. When given an input, it traverses down the tree based on the values of the attributes, and whichever leaf node it reaches, is the final classification of the input.

SUPPORT VECTOR MACHINE:

The support vector machine algorithm is used for classification of the input features using a hyperplane. It does well on higher dimensional data. Initially, it randomly chooses a line or a hyperplane, and finds the support vectors. The support vectors are the closest points to the hyperplane in each classification. The distance between the hyperplane and the support vector is termed as margin. The algorithm aims to maximise this margin, and finds the optimal hyperplane. When an input is provided, the model finds on which side of the hyperplane is the data point, and determines the classification.

RANDOM FOREST:

Random forest is an ensemble of multiple decision trees, which shows improved accuracy rather than when relying on a single decision tree. Each decision tree is trained on a different subset of the dataset, such that one model captures the features that the other model fails to capture. The final prediction is decided by voting. When an input is provided, the predictions of each decision tree is computed. Whichever is the prediction by majority of the decision trees, is chosen as the final prediction of the random forest model.

GRADIENT BOOSTING:

Gradient boosting combines multiple weak learners to provide a strong model. It typically trains multiple decision trees sequentially. Each model is trained on the errors of the previous model. After successive training of multiple weak learners, a strong model is generated which is used for predicting the category. When an input is provided, the predictions of each model are carried onto the subsequent models with a scaling factor, which is known as the learning rate. The final prediction of the model is the sum of the prediction of the first model and weighted sum of the subsequent models. The weights of the subsequent models is the learning rate considered.

REGULARIZATION:

Regularization is a technique used to prevent overfitting. Overfitting occurs when the model memorizes the training data, instead of finding the patterns in it. Regularization can be done by adding a penalty term to the loss function, which makes the loss function non zero even when there is no error. This non zero part makes the loss function move towards zero, thus preventing the model from overfitting the training data. This proposed system has inculcated the regularization technique in the form of adjusting the learning rate. If the learning rate is more, the algorithm makes huge updates to the model parameters, and this may make the model completely fit the data. If the learning rate is less, it makes the algorithm do less updates to the model parameters. In such a case, the model learns slowly and develops as a less complex model. In that way, the model does not overfit the training data. But the right threshold needs to be found out experimentally in order to achieve better accuracy, and the proposed system has a learning rate of 0.4 in the gradient boosting algorithm.

SYSTEM DESIGN:

INPUT

FINAL PREDICTION

PREDICTION

PREDICTION

PREDICTION

PREDICTION

GRADIENT BOOSTING

RANDOM FOREST

SUPPORT VECTOR MACHINE

DECISION TREE

The proposed system takes an input and passes through all the individual models. The input is a set of binary values indicating either yes or no for all the symptoms which are considered as the features. The string values are mapped to integers because the training has been done based on the converted integer values. The age and gender of the person is also collected. The gender which is also a string value is converted to an integer value which represents either male or female. Converting or using any functions on the strings is not necessary as it is a classification problem, and not any kind of regression problem. After mapping the strings to the appropriate integer values which needs to be the same as the one used during the training part, the input is created. This input is passed to the all the individual algorithms. The predictions of the individual models are computed and the majority of those predictions is chosen as the final prediction of the model. Each model captures different features and gives an output, thus making the overall model being effective in making the right classification.

ACCURACY:

The accuracy of the model is calculated by the proportion of the number of correct classifications done by the model to the total number of classifications done by the model.

Accuracy = Correct Predictions

Total Predictions

For a classification problem that classifies into either of two classes, the model’s accuracy can also be calculated as the sum of model true positive and model true negative, divided by the sum of model true positive, model false positive, model true negative and model false negative.

Accuracy = TP+TN

TP+FP+TN+FN

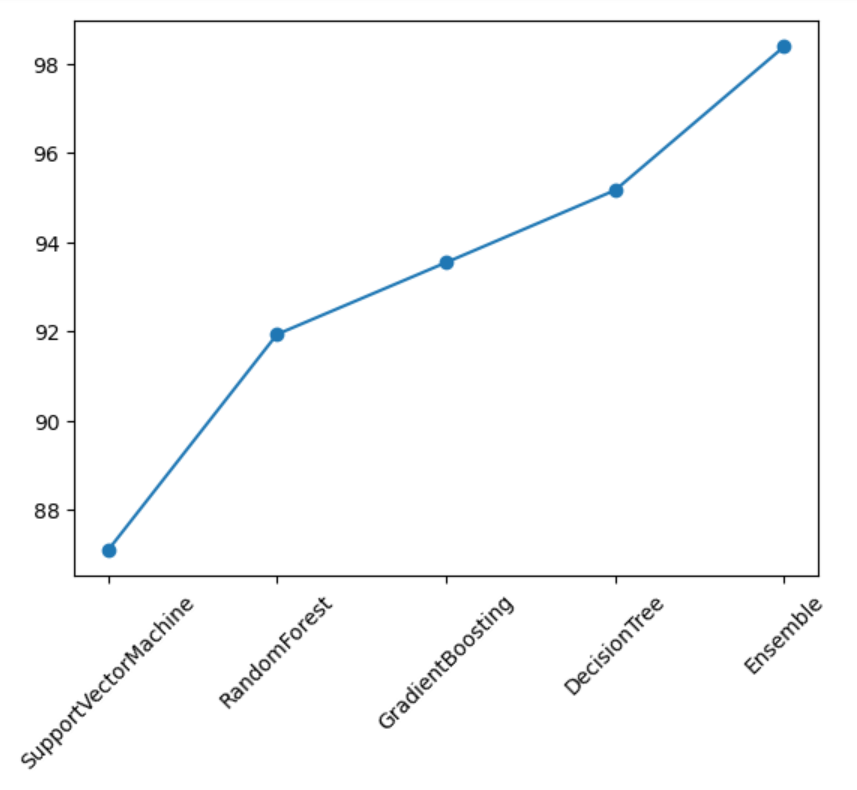
**EXPERIMENT RESULT AND ANALYSIS**

The model when trained as an individual decision tree classification algorithm gave an accuracy of 95.16%, when trained as an individual support vector machine algorithm gave an accuracy of 87.09%, when trained as an individual random forest algorithm gave an accuracy of 91.93% and when trained as an individual gradient boosting algorithm gave an accuracy of 93.54%. Thus, taking an ensemble of all the four algorithms, Decision Tree, Gradient Boosting, Random Forest, and Support Vector Machine, recorded an accuracy of 87.34%. It also recorded 100% accuracy on the training data, which means that the model has completely learnt the training data and has overfit the data. Hence, a regularization technique was adopted and the learning rate was adjusted to 0.4 in the gradient boosting algorithm. After performing regularization on the ensemble model of the four algorithms, the model achieved an accuracy of 98.38%, thus forming the proposed system.

The following table shows the accuracy levels recorded by the individual models, and the ensemble model of all the four algorithms after regularization techniques:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Decision Tree | 95.16% |
| Support Vector Machine | 87.09% |
| Random Forest | 91.93% |
| Gradient Boosting | 93.54% |
| Ensemble model with regularization techniques | 98.38% |

The graph which is depicted below is the line plot indicating the accuracy levels achieved by the individual algorithms, and the ensemble model after regularization, which is the proposed system.



From the table and the graph, it is concluded that the proposed system, “Ensemble Model with Regularization Techniques” has achieved the highest accuracy than any of the individual models.

**CONCLUSION**

The proposed system inputs the symptoms of the person regarding swallowing difficulty, shortness of breath, fatigue, chest pain, and smoking habit. Based on the values for these parameters, the system which is built as an ensemble of four different algorithms, Decision Tree, Support Vector Machine, Random Forest and Gradient Boosting, along with regularization techniques, predicts whether the person has lung cancer or not. Acting as a binary classification model, the system recorded an accuracy of 98.38%. This proposed system can be effectively used by hospitals and doctors for diagnosing a person’s condition regarding lung cancer, by just looking at the symptoms, instead of completely relying on scans.

**REFERENCE**

[1] Qurina Firdaus, Riyanto Sigit, Tri Harsono, Anwar – Lung Cancer Detection Based on CT-Scan Images With Detection Features Using Gray Level Co-Occurrence Matrix and Support Vector Machine Methods – 2020 International Electronics Symposium

[2] Nusraat Nawreen, Umma Hany, Tahmina Islam - Lung Cancer Detection and Classification using CT Scan Image Processing – 2021 International Conference

[3] Sanjana Narvekar, Mayur Shirodkar, Tanvi Raut, Purva Vaingankar – A Survey on Detection of Lung Cancer using Different Image Processing Techniques – 2022 3rd International Conference

[4] Mattakoyya Aharonu, R Lokesh Kumar – Convolutional Neural Network based Framework for Automatic Lung Cancer Detection from Lung CT Images – 2022 International Conference

[5] Shubhada Agarwal, Sanjeev Thakur, Alka Chaudary – Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis – 2022 10th International Conference

[6] Nidhi S. Nadkarni, Sangam Borkar – Detection of Lung Cancer in CT Image Processing – 2019 3rd International Conference

[7] Mohseena Thaseen, S.K.UmaMaheshwaran, Darshana A Naik, Mrunal Swapnil Aware, Prachi Pundhir, Bhasker Pant – A Review of Using CNN Approach for Lung Cancer Detection Through Machine Learning – 2022 2nd International Conference

[8] Qing Wu, Wenbing Zhao – Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm – 2017 International Symposium

[9] Naji Khosravan, Ulas Bagci – Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis – 2018 40th Annual International Conference

[10] B.Devananda Rao, Mahammad Arshad – Early Detection of lung cancer Using Machine Learning Technique – 2023 International Conference