How to Cite:
Devi, P., Sagar, S., & Rohil, H. (2022). Prediction of lung disease using machine and deep learning techniques: A review. *International Journal of Health Sciences*, *6*(S2), 7583–7601. https://doi.org/10.53730/ijhs.v6nS2.6833

Prediction of lung disease using machine and deep learning techniques: A review

Poonam Devi
Affiliation, City, Country/Territory
Corresponding author email: poonam.20scse3020009@galgotiasuniversity.edu.in

Dr. Shrddha Sagar
Affiliation, City, Country/Territory
Email: shrddha.sagar@galgotiasuniversity.edu.in

Prof. Harish Rohil
Department of Computer Science and engineering, Chaudhary Devi Lal University, Sirsa, India

***Abstract*---**Nowadays lung diseases are becoming a significant problem. In spite of this, Corona virus disease 2019 (COVID-19) has become a pandemic all over the world from last two years which effected lungs of few patients also. Many people are suffering from lungs diseases like Asthma, Allergies, lung cancer etc. The patients whose lung gets affected due to COVID-19 may face some lungs diseases in near future, so it very significant to early diagnosis of lungs diseases to save human life. Machine learning (ML) with feature selection techniques play significant role in the medical field by making diseases diagnoses accurate and early. The objective of this paper is to presents a review of recent ML algorithms and feature selection techniques used to predict lung diseases. As we cover the study between 2020 – 2021, some supervised (SVM, Logistic Regression, Random Forest, Logistic model tree, Bayesian Networks) machine learning techniques on 18,253 data instances and unsupervised (KNN,CNN) machine learning techniques on 8,761 data instances were used to detect accuracy, precision, recall, sensitivity and F1-score in order to predict lung diseases. The models are yet to be used to solve real world problems but imbalanced data, selection bias and variance is the limitation of some models.

***Keywords*---**lung disease, feature selection techniques, machine learning model, selection bias, variance.

Introduction

The pressure of health is increasing day by day all over the world. COVID-19 pandemic has shaken the whole world. Before COVID-19, postgraduate medical journal reports that in1918 Spanish Influenza world first pandemic came into consideration from 1918 to 1920 and consisted of four waves. It infected about 500 million people and resulted deaths of 50 million (W.K.Chen, 1993). As per World Health Organization (WHO) report "coronavirus disease2019 (COVID-19) is termed as illness due to a novel corona virus defined severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), that was first pinpoint among an outbreak of inhaling diseases cases in Wuhan City, China. COVID-19 firstly reported by WHO on December 31, 2019. On January 30, 2020, WHO declared the COVID-19 eruption a global health emergency. On March 11, 2020, WHO declared COVID-19 a global epidemic (https://www.medscape.com/answers/2500114-197402/how-did-the-coronavirus-outbreak-start.). "As per WHO report till date of 5 September 2021 total 222,406,582 are COVID-19 confirmed case and resulted deaths are 4,592,934 (https://covid19.who.int/). As we have seen that both of the pandemic are related to respiratory diseases and people all over the world are suffer from other lungs diseases like asthma, pneumonia, lung cancer and tuberculosis. Respiratory maladies are mainly the issue in the lungs that forestall lungs from proper functioning. Usually there are three types of lung maladies (https://medlineplus.gov/ency/article/000066.html).

Airway diseases

Such maladies affect the airways or tubes which carry oxygen and other gases into and out of body lungs. They usually cause an obstruction in the tubes. Airway maladies include asthma, COPD (chronic obstructive pulmonary disease) (Aykanat, M ,2020, https://www.medicalnewstoday.com/articles/types-of-lung-diseases#airway-diseases).

Lung tissue diseases

Such diseases affect the structure of lung carrier. Because of that it becomes hard for the lungs patients to intake oxygen and to free carbon-dioxide (CO2). People tolerating such type of lung maladies sense like they are exhausting too tight dress. "As a result they can't breathe deeply" (https://medlineplus.gov/ency/article/000066.html, https://www.medicalnewstoday.com/articles/types-of-lung-diseases#airway-diseases).

Lung circulation diseases

Such maladies affect the blood bearers in the lungs. They are caused by clotting, inflammation of the blood bearers. They affect the ability of the lungs to intake oxygen and free carbon dioxide (CO2). These maladies may also affect heart function (https://medlineplus.gov/ency/article/000066.html, https://www.medicalnewstoday.com/articles/types-of-lung-diseases#airway-diseases). Many lung maladies include a combination of these three types.

Cancer is the most serious lung disease but it can cure using early prediction. Asthma is another serious problem in which patient face recurrent attacks of breathlessness, in asthma airway become narrow and swollen, difficult for the person to respire properly and person sense that they cannot get ample air. Pneumonia is most common infection of lung which is caused by bacteria or virus, including the corona virus that causes COVID-19. Acute respiratory distress syndrome (ARDS), this is a critical, sudden injury to the lungs from a serious illness. COVID-19 is one example. Many people who have ARDS need help breathing from a machine called a ventilator until their lungs recover. Person having lungs diseases must do regular exercise, yoga and should avoid smoking (https://www.medicalnewstoday.com/articles/types-of-lung-diseases#airway-diseases).

ML is a branch of artificial intelligence (AI) that focus on producing systems that are able to learn from examples and improve without being explicitly programmed. (https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/) ML has been applied effectively in many fields including health care(https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529). One important research direction leverages ML to understand and identify lung diseases in patients. A number of review and research paper has been published on the use of ML in lung diseases research. (IshanSen,2020) discussed Bagging, Logistics Regression, Random Forest, Logistic model tree and Bayesian Networks model to find accuracy by using K-Fold Cross Validation Technique for training of datasets. Various AI techniques have been investigated in the literature review including deep learning, machine learning and Convolutional Neural Network (CNN). (Anuradha D. Gunasinghe, 2019) survey the application of deep learning to investigate lung diseases using patient chest X-rays. The aim of this paper is to highlight ongoing efforts to use state-of-the-art ML evolutionary models and deep learning to predict lung diseases.

Feature Selection

To get the good result feature selection is the most critical task in any machine learning model because sometime increases the number of features may decrease the accuracy of model. Some of the attributes value may be missing may have duplicacy or not co-related to final output. So instead of selecting all the features of any particular dataset we need to choose some genuine features in dataset by applying feature selection techniques on it. (Ali,A.A.,Hassan 2020, https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/) Three feature selection approaches used are filter, wrapper and embedded (V.Durgadevi, 2021)

Filter Method

Filter method is one of the easiest and fastest feature selection methods where we use statistical methods like ANOVA test, Chi square test and correlation coefficient to select important features/attributes and that selected features are very much correlated with targeted output. It has a low chance of over-fitting problem. (https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/)

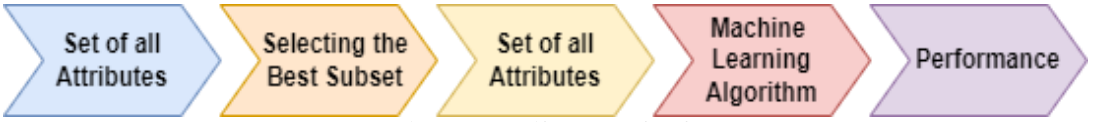

Figure 1. Filter Method

Wrapper Method

In this approach the best important feature are selected form the dataset with respect to the target output using induction algorithm like forward selection, backward elimination or recursive feature elimination. It is less fast than filter method due to high computational time and high chance of over - fitting problem because it involves training of machine learning model with different combinations. (https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/)

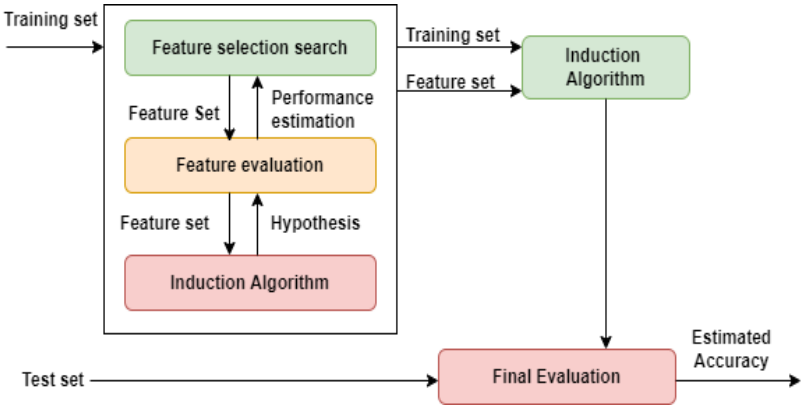

Figure 2. Wrapper Method

Embedded Method

This technique is similar to wrapper method and use induction algorithm to evaluate the important features form the dataset. The only difference is that it create lot of subsets by using all possible permutation and combination on attributes and then select the subset which give best accuracy. (https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/)

Figure 3. Embedded Method

Performance Measure

The performance measure is discussed in this section. The evaluation criteria are used to measure performance of trained model by using testing dataset. Confusion matrix is used to measure performance of model, confusion matrix is a two dimensional table labeled as “Actual” and “Predicated” both the dimension have “True Positive(TP)” , “False Positive (FP)”, “True Negative (TN)” and “False Negative(FN)” (Ali,A.A.,Hassan 2020, https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b)

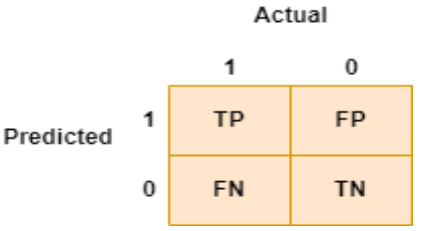

Figure 4. Confusion Matrix

Accuracy: Accuracy is one of the classification performances measures that is most commonly define as total number of correct predictions divided by all predictions made.

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

Precision: Precision defined as the number of positive samples classifier correctly to the total number of positive samples predicted.

$$Precision = \frac{TP}{TP+FP}$$

Recall: Recall is defined as number of accurately classifier positive samples to total number of positive samples (https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b)

$$Recall = \frac{TP}{TP+FN}$$

Specificity: Specificity is defined as number of negatives return by our machine learning model.

$$Specificity = \frac{TN}{TN+FP}$$

F1- Score: F1-score give the harmonic mean of precision and recall. For best case F1 would be 1 and for worst case it would be 0.

$$F1 = \frac{2*(Precision*Recall)}{(Precision + Recall)}$$

ML Models and Techniques used to Predict Lung Diseases

Machine learning (ML) is a subset of AI that that utilizes different feature selection and optimization techniques and can produce a system that are able to learn from examples; improve model result time to time without being explicitly programmed. ML techniques can be divided into three main categories: "*supervised learning, unsupervised learning and reinforcement learning."* (https://medium.com/@MohammedS/ performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b) In supervised learning, the algorithm learns from labeled dataset where input and output is known. Classification and Regression are the two type of supervised learning. Classification is about predicting discrete or class values. e.g. male or female, true or false. Regression is about predicting continuous or quantity values. e.g. salary, age, price. Apart from this, in unsupervised learning, the algorithm learns from unlabeled datasets where input is known but output is not available. Clustering and Association are two types of unsupervised learning in which clustering is the grouping of similar type of data points, where association is used to find important relationship between data points. (https://www.javatpoint.com/machine-learning -random-forest-algorithm) In reinforcement learning, there are four main aspects; Environment, Agent, Action, and Reward. Here intelligent agent learns through trial and error and take an action in an environment to maximize its reward. So agent gets a reward for every correct action and a punishment for every wrong action.

ML Model

The various steps involved in machine learning model is explained as follows.

- Gathering data: Deciding quantity and qualitative data is the part of this problem. You may collect primary data or may use pre-collected data sets from Kaggle or UCI repository (https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html).
- Preparing data: Raw data is not useful, while preparing data should ensure that there is no bias, missing value, delicacy and error in collected data. Now split the data into training and evaluation sets.
- Choosing a model/Algorithm: The third step consists of choosing the right model. There are many different algorithm for different task, but you need to choose the model which meets the goal.
- Training: Training to model is the main part of machine learning. The objective is to use training data to answer a question or incrementally improve the model prediction. Each iteration in model is a training step

(https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html).

- Evaluation: After completing training of model, the fifth step of machine learning is to evaluate the model. In evaluation testing of machine learning is done against unused control datasets to see how well did you do.
- Hyperparameter tuning: After evaluating model, you should test the originally set parameters to improve AI performance. Because increasing the number of test cycle may lead to improve performance (https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d).
- Prediction: Once you have gone through the process of gathering data, preparing the data, selecting the algorithm, training, evaluating the model and tuning the parameters, it is time to answer questions using predictions. There can be all kinds of predictions, ranging from image recognition to semantics to predictive analytics (https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html).

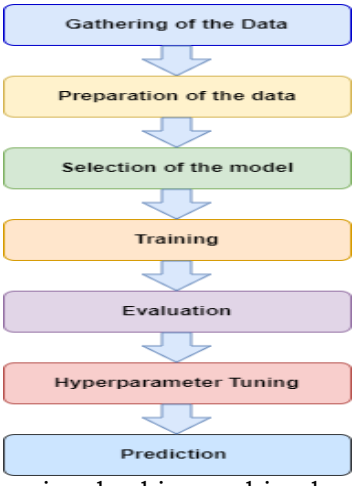

Figure 5. Steps involved in machine learning models

Process for diagnosing the Lung using ML model

As we know lung infection is increasing day by day (https://www.medscape.com/answers/2500114-197402/how-did-the-coronavirus-outbreak-start.), even after Covid-19 the cases are increasing and if timely treatment is not provided to patients, it may causes big health issue like lung cancer etc. and may causes loss of file. So it is extremely important to identify such patient's diseases to save human life. The current techniques for detecting lung diseases are CT scan, Chest X-ray or symptoms examination (https://medlineplus.gov/ency/article/000066.html).. Although these all are also good but processing is slow, even some time patient may avoid some regular checkup and without consent of doctor we may not identify exact problem timely. In ML terms, the task is usually formulated as a classification problem. ML is a pre-trained model which is able to identify weather the patient has any lung diseases or not.

The model presented in this review employed various ML techniques to identify lung diseases. Some studies use single technique to find its accuracy, precision, recall, sensitivity and F1-score whereas others use group of techniques to find accuracy or to do comparisons between those multiple approaches. (Ishan Sen, 2020)

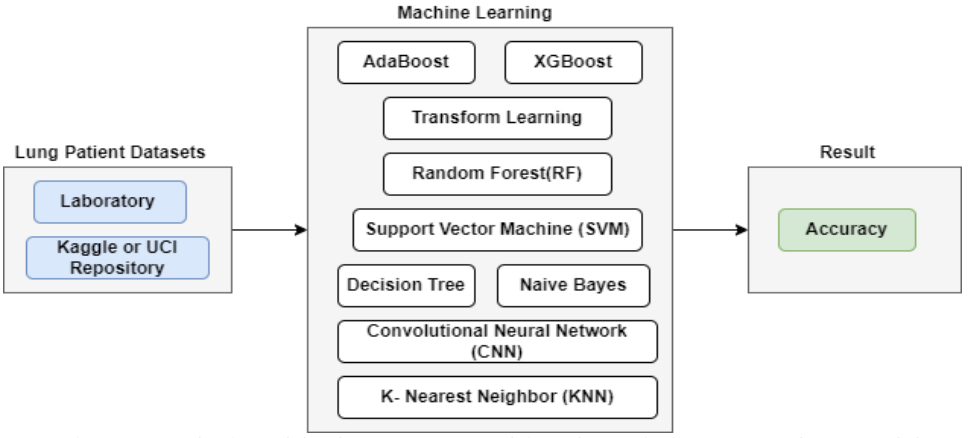

Figure 6. Review on relationship between machine learning approaches and lung application

ML Techniques

A large number of dataset in form of attributes and x-ray images of COVID-19 patient lungs and lung diseases patient are available on Kaggle and UCI repository. In my literature study, it was found that a large number of supervised and unsupervised machine learning algorithms and deep learning techniques are available to do early prediction of diseases (https://www.medicalnewstoday.com/articles/types-of-lung-diseases#airway-diseases).

Ensemble learning: Ensemble learning is a collection of multiple supervised machine learning models through which we pass our dataset and get output accordingly. There are basically two techniques in ensemble learning *Bagging/ Bootstrap aggregation and Boosting*. Where random forest is a bagging technique and AdaBoost, Gradient boosting and XGBoost are boosting technique.

AdaBoost: AdaBoost stands for *Adaptive Boosting*, it is a classification technique in which boosting ensemble learning algorithm is used. In AdaBoost technique weight (W) is assigned to each record and the value of initial W = 1/N, where N is the total number of record. We create sequential base learner for each and every feature with the help of decision tree (DT) of depth =1, which we called *Stomps*, then we create another stomps and so on. The successive model is boosted by the re-weighting of records in accordance with previous model outputs.

XGBoost: XGBoost stands for *Extreme Gradient Boosting*, it is a classification technique in which boosting ensemble learning algorithm is used. In XGBoost base model is constructed using binary decision tree on the basis of classification

of any attribute, then we compute similarity weight and on the basis of similarity weight we find information gain of each binary DT. The successive model is boosted by classifying again on basis previous model output. The equation of similarity weight is:

$$\textit{Similarity weight} = \frac{\sum (\text{Residual})^2}{\sum (P(1-P))}$$

$$\textit{Info\_gain} = \sum (\text{Similarity weight})$$

Where P is the probability of base model and Residual is the Output – base model probability. It is a classification technique so output will be either 0 or 1

Transfer Learning. Transfer learning is an optimize deep learning model that is used on unlabeled dataset. It is a method to utilize pre-trained model for one task to solve related ones. It saves lots of computation power. Here, we use pre-trained weights for new problem and make change in last few layers of model and will get high accuracy results due to use of hyper-parameter tuning. For example, knowledge gained in early prediction of lung diseases from lung patients could apply to covid-19 patient's chest x-ray whose lungs get effected due to covid; to predict their respiratory diseases (https://ruder.io/transfer-learning/.).
As per Panand Yang theory transfer learning consist the hypothesis of domain and task. A domain, D consist of feature space X and marginal probability distribution P(X) over the feature space, where

$$X = x_1, x_2, \ldots x_n \in X$$

$$\textit{Domain } d = \{X, P(X)\}$$

Task T consists of label space Y and conditional probability distribution P(Y/X) that is learned from training data consisting of pairs $x_i \square X$ and $y_i \square Y$ where Y is set of all labels, i.e True Or false and y_i is either true or false[11]. Source domain d_s, source task Ts, target domain d_T
and target task T_T. Objective of transfer learning is accredit us to learn the target conditional probability distribution given as

$$P(Y_T|X_T), \textit{ where } d_S \neq d_T \textit{ or } T_S \neq T_t$$

Random Forest. Random forest is a supervised machine learning technique that is based on the concept of ensemble learning. It is used for both classical and regression problems in machine learning. In this technique we combine multiple decision tree classifiers to solve complex problems and to improve the performance of model. Because this model is not dependent on one decision

tree(DT) output; nested of it, it take majority of votes out of all DT, so this model have low bias and low variance.

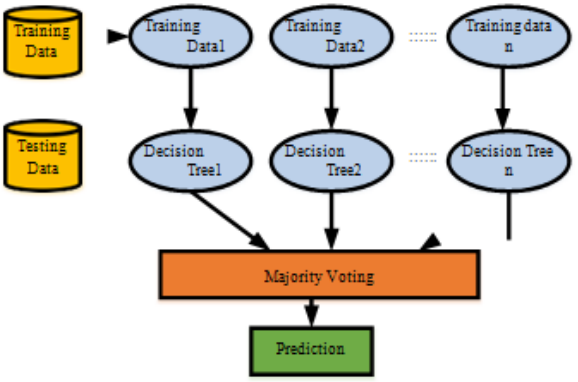

Figure 6. Random Forest Model

Fig 6 presents the Random Forest model. This algorithm contain multiple decision trees on different subset of given dataset and consider average to improve the predictive accuracy of that particular dataset and to predict the final output (https://www.javatpoint.com/machine-learning -random-forest-algorithm).

Support vector machine (SVM). SVM is a supervised machine learning algorithm that is mainly used for classification problems in machine learning. In the SVM model, we plot each data item as a point in n-dimensional space with the value of every feature being the value of a particular coordinate.

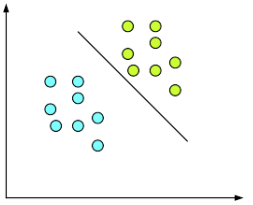

Figure 7. SVM Model

Fig7 illustrates the working of SVM. The hyperplane performs the classification and finds the difference or margin between the two classes. (https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html). The equation of hyperplane is :

$$W^T . x + b = 0$$

Where W^T is the slop of hyperplane and b is the constant. (https://medium.com/@MohammedS/ performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b) The equation of optimize function that will help to reduce over-fitting problem is :

$$(W^*, b^*) = \min \frac{||W||}{2} + C \sum_{i=1}^{n} £_i$$

where||W|| is norm of W, C is total number of errors, $£_i$ is value of the i^{th} error (https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/)

Decision tree (DT). DT is a supervised machine learning algorithm that is mainly used for classification problems in machine learning. It is basically a tree like structure that shows relationship between classes and attributes. To get better results in DT our aim is to reach leaf node quickly and for that we need to choose write parameters/attributes. Just because of single DT this model have low bias and high variance. We need entropy in DT to measure purity of split, so the equation of entropy is:

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

Where P_+ is the positive sample, P_- is the negative sample. (https://machinelearningmastery.com/bayes-theorem-for-machine-learning/)
In DT we use information gain method to get the average of all the entropy. So equation of *information gain is* :

$$Info\_Gain(S, A) = H(S) - \sum \frac{|S_v|}{|S|} * H(S_v)$$

Where H(S)is the entropy of sample/attribute, |S| is the sum of all sample, S_v is the value of any particular y^{th} sample and $H(S_v)$ is the entropy of y^{th} sample. (https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/)

Naive Bayes (NB). NBis a simple probabilistic classifier based on Bayes' theorem. Given a record X and a number of m classes C1, C2, ..Cm, NB classification maximizes P(Ci |X) using Bayes' theorem, as follows:

$$P(Ci|X) = P(X|Ci)P(Ci) / P(X)$$

where, P(Ci), P(X|Ci), and P(X) may be estimated from the given data. The word naive refers to the main assumption of conditional independence. NB assumes independence among class attributes. This is not necessarily true in real-world applications. (https://machinelearningmastery.com/bayes-theorem-for-machine-learning/)

Convolutional neural network(CNN):CNN is a deep learning neural network basically designed for structured arrays of data. Here in convolution operation we use a n*n matrix image as an input, apply filter 'f' which work as a detector; on this matrix and get output image as m*m. where:

$$m = n-f+1$$

CNN may contain many convolutional layers and many filters assembled on top of each other in horizontal manner to recognized image clearly. CNN also use padding and pooling feature, where pooling is use to find location invariant, and padding is used to extend the size of input and output. (https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529)

K Nearest Neighbor (KNN):KNN is a supervised machine learning technique which is used for classification as well as regression problems. Here *K* denote the number of nearest similar point with respect to any particular predicted point, referred to as the KNN. A distance method, such as *Euclidean distance*, is used to calculate the distance between predicted points to K selected points. The algorithm then finds the most common class among its KNNs and assigns it to the given data point.(https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/)

Results and Discussions

Comparison of Accuracy of various Models

From study of literature, we find that lots of supervised and unsupervised machine learning techniques were used to identify lung diseases where different technique give different accuracy measure. Table1 present the models and their accuracy used in lung disease detection.

(Shimpy Goyal, 2021) proposed a machine deep learning model to predict lung diseases like pneumonia and Covid-19 using patient's chest X-ray. Soft computing methods like artificial neural network (ANN), support vector machine (SVM), K-nearest neighbor (KNN), ensemble learning and deep learning classifiers are used for classification. Feature extraction techniques are used for image quality enhancements and accurate region of interest to enhance the image quality and the detection and classification result. Recurrent neural network – Long short - term memory (RNN-LSTM) deep learning architecture with normalized and fused features of test samples used to predicts the lung diseases. The RNN-LSTM proposed design help to reduce both training and testing time in comparison to the CNN based model.

(Ishan Sen, 2020) predicts lung diseases accuracy using different machine learning algorithms. For that he uses 323 instances along with 19 attributes, preprocess it on WEKA tool using unsupervised filters and Randomize filter to replace missing values. Then use K-Fold Cross Validation technique to evaluate ML models. In this study five ML algorithms – Bagging, Logistics Regression, Random Forest, Logistics model tree and Bayesian networks are used to find accuracy, where the Random Forest model give the highest accuracy 90.15%.

(Murat Aykanat, 2020) compared different ML models by classifying of respiratory diseases with audio and text data. To record patients lung sound an electronic stethoscope and its software is used. In this study the ML classification algorithms - support vector machine (SVM), k-nearest neighbor (k-NN) and Gaussian Bayes (GB) are used to find accuracy, where the SVM give most

accurate result with text data, k-NN gives most accurate result with audio data and using both text and audio data classification SVM was most accurate.

(Prateek Chhikara, 2020) Compared machine learning model with ResNet, ImageNet, Xception and Inception in term of precision, recall, accuracy and ROC accuracy. In this study he used chest X-ray dataset and CT scan image to detect pneumonia using deep Convolutional neural network (CNN) and transfer learning. The proposed model precision is 0.907, recall 0.957, F1-score 0.931 and validation accuracy 90.1%

(Subhrato Bharti, 2020) propose a hybrid deep learning model using different machine learning techniques to detect lung diseases using patient chest X-ray. In this study author use visual geometry group based neural network (VGG), data augmentation and spatial transformer network (STN) with CNN and named this hybrid model as VGG Data STN with CNN (VDSNet). For implementation author use Jupyter Notebook , Tensor flow and Keras tool. Here author find accuracy of different model like- VDSNet, vanilla grey, vanilla RGB, hybrid CNN and modified capsule network, out of these VDSNet exhibits the highest accuracy of 73% .

(Rachna jain 2020) compared the performance of deep learning based CNN models to analysis or detect covid-19 in humans using chest X-ray of covid-19 infected patients and normal persons. Author compared Inception V3, Xception and ResNet models. To check the model performance author collect 6432 chest x-ray sample form kaggle repository , out of which 5467 used for training purpose and 965 for validation. After finding the accuracy of all given three models it was concluded that the Xception model gives the highest accuracy 97.97%.

(Anuradha D. Gunasinghe, 2019) proposed a profound learning improved CNN model on patient chest X-ray for early forecast of lung illnesses and found that to recognize and analyze the lung infections as soon as conceivable will assist the specialist with saving the patient's life. In this study she used chest X-ray datasets from Kaggle and UCI repository that is passed through VGG16 architecture to extract features.

(Naman Gupta, 2019) use feature extraction, feature selection and machine learning techniques to predict lung diseases like Chronic Obstructive Pulmonary Disease (COPD) and Fibrosis using patients Computerized Tomography (CT) scan image. In first step author extracts features like Haralick texture using gray level co-occurrence matrix, Zernike's moments, Gabor and Tamura texture. Now in second step apply feature selection techniques like Improvised Crow Search Algorithm (ICSA), Improvised Grey Wolf Algorithm (IGWA) and Improvised Cuttlefish Algorithm (ICFA) to select optimal features, where ICSA eliminated the highest number of insignificant features of near 71% and IGWA gave the best accuracy of 99.4%. In final step, classification algorithms like Support Vector Machine (SVM), k-nearest neighbor (k-NN), Decision tree and Random Forest were applied to predict lung diseases.

Table 1
Accuracy of various models used in lung disease detection

Author Name	Number of features used	Feature Extraction Method	Classification Techniques	Accuracy
Barua 2005	12(text)	IOS	MLP	95.01%
Er,O. 2008	38(text)	Manual selection	MLNN	96.08%
Er. O. Sertkaya 2008	38(text)	Manual selection	PNN,AIS, LVQ, MLNN	94%
Er. 2010	38(text)	Manual selection	PNN,GRNN,RBF ,MLNN, LVQ	93.75%
Yamashita 2011	Not specified	Segmentation	HMM	88.7%
Amaral2012	7(text)	Manual selection	KNN,SVM,MLP	95%
Naman Gupta 2019	Not specified	Segmentation	IGWA	99.4%
Jayalakshmy2020	Automatic	Segmentation	CNN	83.78%
SubratoBharati 2020	Kaggle repository	Manual selection	VDS Net	73%
PragyaChaturvedi 2020	Not specified	Segmentation	2DCNN	97.20%
RachnaJain2020	Kaggle repository	Data augmentation	Xception	97.97%
Ishansen 2020	Kaggle repository	Manual selection	RF	90.15%
PrateekChhikara	Not specified	Manual selection	CNN	90%

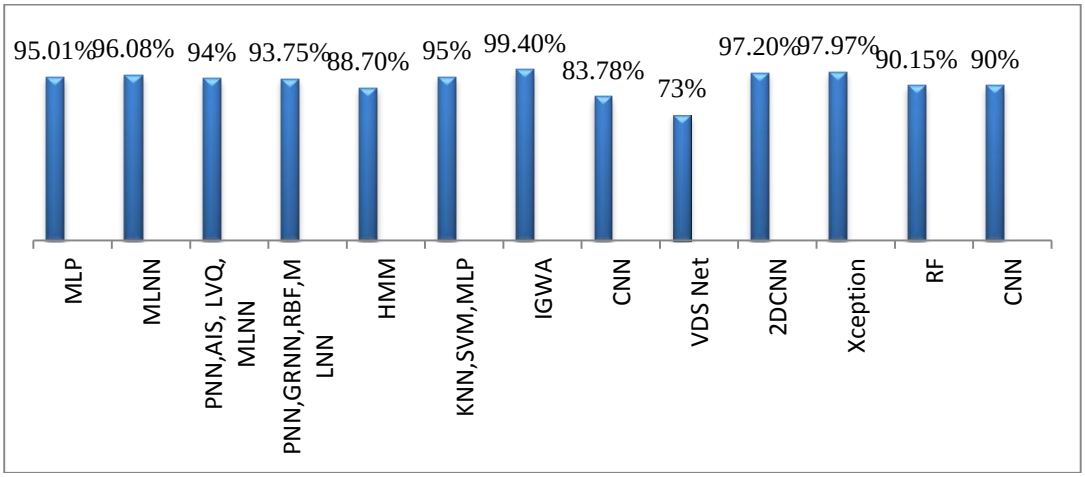

Figure 8. Model v/s Accuracy

Above Figure 8 shows the accuracy of machine learning techniques used in past. In which, (Naman Gupta, 2019) Improvised Grey Wolf Algorithm (IGWA) model give the highest accuracy of 99.4%. (Subhrato Bharti 2020) As we all know machine learning and deep learning is the area that is need to explore more and have lot of techniques like transform learning , random forest, Hyperparameter tuning etc. that may be used for unlabeled and labeled dataset to get more accurate result

Table 2
Accuracy, precision, recall, F-measure and specificity of lung prediction

Author Name	Classification Technique	Accuracy in %	Precision in %	Recall in %	F-measure in %	Specificity in %	F1-measure in %
Ishan sen 2020	Bagging	88	85.60	88	86.10	-	-
	Logistic Regression	88.92	87.30	88.90	87.70	-	-
	Random Forest	90.15	89.40	90.20	89.70	-	-
	Logistic model tree	89.23	87.80	89.20	88.10	-	-
	Bayesian network	83.69	84.90	83.70	84.20	-	-
Shimpy Goyal2021	KNN	85.72	84.09	91.07	87.44	85.77	87.44
	SVM	87.88	84.78	87.73	86.23	88.02	86.23
	ANN	91.22	88.28	93.63	90.88	89.07	90.88
	Ensemble	87.93	85.76	88.43	87.07	89.47	87.07
	RNN-LSTM	95.04	93.65	96.78	95.19	94.21	95.19

Murat Aykanat 2020	SVM	73	73	93	91	-	-
	k-NN	67	67	100	100	-	-
	Gaussian Bayes	58	58	64	64	-	-
Subhrato Bharti 2020	Vanilla gray	67.80	68	58	-	-	66
	Vanilla RGB	69	68	61	-	-	66
	Hybrid CNN VGG	69.50	68	62	-	-	67
	VDSNet	73	69	63	-	-	68
	Modified CapsNet	63.80	61	48	-	-	58
	Basic CapsNet	60.50	64	51	-	-	61
V.Durgadevi 2021	SMMLCA	85.40	84.20	74.20	-	-	71.40
	CNN	78.20	79.40	69.40			66
	RML	81.20	80.34	71.54			68.6

Note - SMMLCA – Spark framework with Multiple Machine Learning Classifier Algorithm
RML – Recurrent Machine Learning
RNN-LSTM – Recurrent Neural Network with Long-Short Term Memory

Conclusion

In this study, we explore a survey of different machine learning techniques to predict respiratory diseases and also discuss trends of future research. We have discussed that there are plethora of machine learning techniques to accurately predict lung diseases, but it can be said that transform learning and CNN may be best suited for patient's chest x-ray analysis. The available models are yet to be used to solve real world problems but imbalanced data, selection bias and variance is the limitation of some models. So first we need to make balanced dataset using feature selection and feature optimization techniques and then after we have to pass our training or test dataset through ML or deep learning model to accurately predict lung diseases. So we need to choose –

- Need to choose best data set tool form where we may get genuine data.
- Number of features to be selected for best result.
- Feature selection method used to select features.
- Best machine learning or deep learning technique.

References

A Gentle Introduction to Bayes Theorem for Machine Learning Retrieved March 25, 2022, https://machinelearningmastery.com/bayes-theorem-for-machine-learning/

Ali,A.A.,Hassan, H.S.,&Anwar, E.M. (2020). Improve the Accuracy of Heart Disease Predictions Using Machine Learning and Feature Selection Techniques. Machine Learning, Image Processing, Network Security and Data Sciences, 214–228. doi:10.1007/978-981-15-6318-8_19"

Amaral,J.L.M.,Lopes,A.J.,Jansen,J.M.,Faria,A.C.D.,&Melo, P.L.(2012). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. Computer Methods and Programs inBiomedicine,105(3),183–193.doi:10.1016/j.cmpb.2011.09.009

Anuradha D. Gunasinghe, Early prediction of lung diseases , International conference for convergence in technology , 978-1-5386-8075-9/19/$31.00@2019 IEEE. (2022).Retrieved 26 February 2022,from http://www.cs.columbia.edu/~jebara/4771/notes/class6x.pdf

Aykanat, M., Kiliç, O., Kurt, B., &Saryal, S. B. (2020). Lung disease classification using machine learning algorithms. International Journal of Applied Mathematics Electronics and Computers, 8(4), 125–132.doi:10.18100/ijamec.799363

Barua, M., Nazeran, H., Nava, P., Diong, B., & Goldman, M. (2005).Classification of Impulse Oscillometric Patterns of Lung Function in Asthmatic Children using Artificial Neural Networks. 2005 IEEEEngineeringinMedicineandBiology27thAnnualConference.doi:10.1109/iembs.2005.1616411

Chaturvedi, P., Jhamb, A., Vanani, M., &Nemade, V. (2021, March).Prediction and Classification of Lung Cancer Using Machine LearningTechniques.InIOPConferenceSeries:MaterialsScienceandEngineering(Vol.1099,No. 1,p.012059).IOP Publishing

Convolutional Neural Network Retrieved March 25, 2022, https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529

Decision Tree Algorithm - A Complete Guide. Retrieved March 25, 2022 https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

Er, O., Sertkaya, C., Temurtas, F., &Tanrikulu, A. C. (2008). A Comparative Study on Chronic Obstructive Pulmonaryand Pneumonia Diseases Diagnosis using Neural Networks and ArtificialImmuneSystem.JournalofMedicalSystems,33(6),485–492.doi:10.1007/s10916-008-9209-x

Er,O.,&Temurtas,F.(2008).A Study on Chronic Obstructive Pulmonary Disease Diagnosis Using `Multilayer Neural Networks. Journal of Medical Systems,32(5),429–432.doi:10.1007/s109-1008-9148-6

Feature selection techniques in machine learning. GeeksforGeeks. Retrieved March 25, 2022, from https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/

Gunasinghe, A. D., Aponso, A. C., &Thirimanna, H. (2019). Early Prediction of Lung Diseases. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT).doi:10.1109/i2ct45611.2019.9033668

IshanSen (2020), In Depth Analysis of Lung Disease Prediction Using MAchien Learning Algorithms , Springer Nature Singapore Pte. Ltd., MIND 2020. CCIS 1241, pp 204-213.2020

Jayalakshmy, S., &Sudha, G. F. (2020). Scalogram based predictionmodel for respiratory disorders using optimized convolutional

neuralnetworks.ArtificialIntelligenceinMedicine,103,101809.doi:10.1016/j.artmed.2020.101809

Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machinelearning and advanced imaging techniques. Translational Lung CancerResearch, 7(3),304–312.doi:10.21037/tlcr.2018.05.15

Kahya, Y. P., Guler, E. C., &Sahin, S. (n.d.). Respiratory disease diagnosis using lung sounds. Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine andBiologySociety."MagnificentMilestonesandEmergingOpportunitiesinMedical Engineering"(Cat.No.97CH36136).doi:10.1109/iembs.1997.758751

Murat Aykanat (2020) lung diseases classification using machine learning algorithms, e-ISSN: 2147-8228 , doi- 10.18100/ijamec.799363

Naman Gupta, Gupta, D. ,Khanna, A. ,RebouçasFilho, P. P. ,& de Albuquerque, V. H. C. (2019). Evolutionary algorithms for automatic lung disease detection.Measurement,140,590–608.doi:10.1016/j.measurement.2019.02.042

Prateek Chhikara, Deep Convolutional Neural Network with Transfer Learning for Detecting Pneumonia on Chest X-Rays, ISBN : 978-981-15-0338-2

Rachna Jain, Gupta,M.,Taneja,S.,&Hemanth,D.J.(2021). Deep learning based detection and analysis of COVID-19 on chest X-rayimages. Applied Intelligence,51(3),1690-1700.

ShimpyGoyal (2021) Detection and classification of lung diseases for pneumonia and covid-19 using machine and deep learning techniques , Springer nature , Journal of Ambient Intelligence and Humanized Computing. Doi-10.1007/s12652-021-03464-7

Subrato Bharati, Podder, P., &Mondal, M. R. H. (2020). Hybrid deep learning for detecting lung diseases from X-ray images.Informatics In Medicine Unlocked,20,100391.doi:10.1016/j.imu.2020.100391

V.Durgadevi Multiple Classifier System Based Lung Disease Prediction with Spark Framework in Big Data, Vol.12 No.9 (2021), 2266–2276, https://turcomat.org/index.php/turkbilmat/article/view/3701/3168

W.K.Chen,LinearNetworksandSystems.Belmont,CA,USA:Wadsworth,1993,pp.123–135.

Website: https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d

Website: https://www.analyticsvidhya.com/blog/2017/09/understating-support-vector-machine-example-code/

Website: https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html

Website:https://medium.com/@MohammedS/ performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Website:https://ruder.io/transfer-learning/.

Website:https://www.javatpoint.com/machine-learning -random-forest-algorithm

Website:https://www.medicalnewstoday.com/articles/types-of-lung-diseases#airway-diseases

Website:https://www.medscape.com/answers/2500114-197402/how-did-the-coronavirus-outbreak-start.

Website:MedicalEncyclopediahttps://medlineplus.gov/ency/article/000066.html

Website:WHOdashboard-https://covid19.who.int/

Yamashita,M.,Matsunaga,S.,&Miyahara,S.(2011).Discriminationbetween healthy subjects and patients with pulmonary emphysema

bydetectionofabnormalrespiration.2011IEEEInternationalConference on Acoustics, Speech and Signal Processing (ICASSP).doi:10.1109/icassp.2011.5946498