2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI) | 979-8-3503-3091-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIIHI57871.2023.10489654

Validation & Diagnosis of Cystic Fibrosis Prognosis Using Gradient Boost Decision Trees

Prerna Reddy Ganga[1*]
Department of Computer Science and Engineering
Stanley College of Engineering & Technology for Women, Hyderabad, Telangana, India
Prernastan77@gmail.com

Kishor Kumar Reddy C[2]
Department of Computer Science and Engineering
Stanley College of Engineering & Technology for Women, Hyderabad, Telangana, India
Kishoar23@gmail.com

Anisha P R[3]
Department of Computer Science and Engineering
Stanley College of Engineering & Technology for Women, Hyderabad, Telangana, India
anishanaidu.pushpala@gmail.com

Srinath Doss[4]
Research Associate, The Independent Institute of Education, IIEMSA, South Africa
Faculty of Engineering & Technology, Botswana Botho University, Botswana
srinath.doss@bothouniversity.ac.bw

***Abstract*—Cystic Fibrosis (CF) is a recessive disorder which damages lungs, digestive track and any other organs. It is an inherited disease caused by a defective gene called cystic fibrosis transmembrane conductance regulator(CTRF). Various techniques have been proposed to detect CF at an early stage which helps prevent serious, lifelong problems related to CF. The present research provides a study and complete analysis on the available approaches; machine learning, deep learning and others. The performance measures in terms of accuracy, sensitivity, specificity and others found using various approaches are: K- Nearest Neighbours (KNN), Bayesian Optimization, Random Forest, Convolutional Neural Network (CNN), Support Vector Machine (SVM), Logistic Regression have compared with our proposed model Gradient Boost Decision trees (GBDT). The results depicts that the proposed model is outperforming in measures of accuracy, error rate, specificity, sensitivity and others.**

Keywords-cystic fibrosis, cystic fibrosis transmembrane conductance, machine learning, inheritance, performance measures, detection, problems.

I. INTRODUCTION

Cystic Fibrosis is a life-threatening, rare and inherited disease which can be passed from generation to generation. It is most commonly occurring in the North European ancestry. Its symptoms could be respiratory or digestive. CF affects the cells that produce mucus and sweat glands. Instead of acting as lubricants they block the duct and airways in the body. There is no cure, but treatments on daily basis can help reduce the problem and ease the way of life. CF occurs in the child when both the parents carry a CTRF gene.

A single parent carrying a gene does not risk the child being prone to the disease. Although, it does affect the risk of passing it down to their next generation. The metrics used for monitoring CF are Sweat chloride test, genetic testing, pulmonary function tests, imaging studies, respiratory culture, blood tests to study the inflammation, infection and pancreatic function. There are specific ranges to meet the conditions for the detection. In this model, dataset will be fed into different models to train them and predict the outcome with a good accuracy with the help of their respective estimators. The table 1 shows some selected features listed for the diagnosis which can be used for training the models. Globally CF is estimated to affect approximately 1 in 2,500 to 1 in 3,500 newborns [1].

The parents must carry the mutated gene and this carrier frequency varies among populations. Advancements in genetic testing and newborn screening have improved early detection and management of the disease. The survival rate for the disease has improved significantly over the years. Many individuals with CF now live their 40s and beyond. However, survival can vary, and it is influenced by factors such as early diagnosis, access to specialized care and the individual's overall health. For this purpose, machine learning techniques can be used to predict to what extent the disease persists.

The paper is organised as follows: Literature Survey in section 2 mentions the existing papers and how this study is producing a better accuracy compared to others. The next part, section 3 in this paper is the proposal method for which Gradient boosting is the chosen algorithm to perform CF diagnosis. Later in section 4 comes the results and discussion where all the research along with its observations are noted, showing why GBDT performs better which provided set of features. To conclude in the last section of the paper, it has a gist of everything done in the study and provides a summary of how it is analysed.

II. LITERATURE SURVEY

A study of Martha Rachel et_al[16] shows the cystic fibrosis using CNN, Random forest, KNN, SVM, Logistic Regression, Bayesian optimization, Decision trees. The authors through their thorough research found that the value of accuracy oscillated depending on the loss function.

In the case of using CNN for prediction of CF, the data used is purely X-rays of the patients. Data augmentation and knowledge transfer from the previously trained neural networks on large datasets was used in predicting the disease. Random Forest used the data on stool fat tests of different CF patients and trained in the model which produced an error of up to 27% percent only. The gradient boosting decision tree gave up to an accuracy of 92+- 2. We used the sweat chloride test to determine the outcome of having the disease.

The table 1 contains the features and their ranges are the normal values which should stabilize the symptoms of CF. For the given features if the values are less than the lower range, it is highly possible for the person to have CF.

TABLE 1 FEATURES WITH THEIR RANGES.

S.No	**Features**			
	Featured	***Lower range***	***Upper range***	***Units***
1	Sweat Chloride test	40	60	mEq/L
2	BMI	18.5	31	-
3	Lung infection	0	1	-
4	Stool fat test	1	7	gm of fat/ 24 hrs
5	Fecal Elastase test	1	200	Mcg/g
6	Vitamin A	30	60	mcg/mL
7	Vitamin D	30	100	Ng/mL
8	Vitamin E	1	5	mcg/mL
9	Vitamin K	0.2	3.2	ng/mL

This paper aims to detect cystic fibrosis using different techniques for various age groups in different stages of their life to reduce complications and ease the symptoms. The diagnosis for CF starts with the review of family history and performing some genetic tests. Machine learning plays a role after this process. The sweat tests determine the increase in levels of salts which tells there is a possibility of having the disease. However, it is not the only factor. Feature selection is a process where machine learning focuses to choose a collective set of attributes which can be helped in determining the presence of the disease and to what extent.

Features such as age, gender, ancestry are some examples which are considered but not the complete basis used in training the model. In this particular rare disease, sweat chloride test results, presence or absence of lung infection, pancreatic infection detection using SFT and FET values, reduction in essential vitamins A, E, D, K are the selective features used in analyzing the accuracy rate of the model after training and testing the dataset. Another table 2 depicts the algorithm against their accuracy scores. It shows for the value of its estimators chosen due to which the accuracy to predict the disease is increased.The occurrence of multiple features in the above table is to show that CF can vary in different patients according to each feature values.

In general, to detect CF the above features are tested manually and diagnosed by a doctor. Through Machine learning our aim is to find out which feature influences the most in the diagnosis. The table 2 in the results column contains accuracy scores with different parameters. They do not necessarily contain all the features. Using one or two of the listed features can also help to a certain extent to train the model to predict the presence of the disease.

Different researchers used AI models to examine the technology and microbiology of CF to predict the condition or stage of lung infections in the future to guide the doctors with their treatment plans. As the patient data is available less to the public all the training and testing is done to the machine learning models using many different types of algorithms. Using trial and error methods the probability of various factors is calculated. Graphs like AUC-ROC are plotted and decided on about which area to improve to decrease the error rates etc., [2] Some other papers focused on calculating and predicting using mean slope values. Meaning, the differences between the mean predicted levels of the features were taken from different age groups, severe and young groups being most least severe and the older mild group having intermediate values from other groups [1].

Datasets from different registers controlled by trusts performed experiments using auto-prognosis. It encouraged an agnostic, data driven approach for identifying risks. Mutations are one of the mid-complex levels to construct and identify through machine learning. Certain number of binary features can be used to encode the information from the genes and can be used to identify information of the patient, if it belongs to that sub category.

While using neural networks for the study, a large number of layers are separated via activation layers, convolutional layers and pooling layer which reduces the size of all the layers in the set. The loss function is calculated for the predicted and actual values. It does not focus on the accuracy and behaves like a negative logarithmic function for error values. When the loss function increases the model tends to overfitting, which means the proportions of negatives must be correctly identified. Using the structures of neural networks extensively, increases the chances of perfecting the accuracy to predict and the usefulness of this tool in medical streams can be helpful.

III. PROPOSAL METHOD

The methodology used in this paper is observational and experimental case study. As the study chosen is a topic on inherited disease, the accuracy scores depending on the feature selection is variable and important as it signifies on which feature to use as dominant. For example, the accuracy score in KNN is far less than the others. It has considered FEV1 scores and FVC scores and collected samples from cities. Therefore, it depends on the dataset used and how many samples trained. The KNN model gave different accuracy scores in different cities based on the population and the impact of the ratios of the diseased versus not infected. The common accuracy score found to be 55% was the total population combined in different places. Climate is one of the reasons why sweat glands secretion might be higher or lesser. In cold countries the normal range of sweat tests might not be the same as taken in good conditioned weathers. Similarly, taking other factors into consideration various algorithms are applied to other factors effecting CF and the accuracy scores are calculated. The aim is to find which factor is most relevant to the disease and can be used for its diagnosis.

In inherited diseases which do not have a cure, the early prediction of the disease helps to prevent the symptoms on a large scale and increases life expectancy. After gripping onto which feature and conditions match the most the machine learning model can be used on a daily basis in various places like hospitals, camps etc., Especially for the north European ancestry to detect the infections early.

The most effective feature for enhancing the model accuracy will be discussed in the results section. When single or 2 features are selected the model can be prone to under-fitting or overfitting. Cross validation techniques are put in use to obtain good accuracy scores for this reason.

The potential implications of findings for improving cystic fibrosis prediction models are more observational and experimental. The observations based on the learnings from the CF patients and using those findings as a basis for the prediction is one way. The other way is to predict in a "maybe" format where the patient might or might not have the disease by placing the threshold values as the factors to evaluate the presence of the disease. The below is the graph (1) based on table (b) to show the variation in accuracy levels using different prediction features.

The convolutional network is used for obtaining features from CT scan images with the goal of predicting the slope of FVC. The deep CT feature extractor network consists of two key components, (i) a CNN-based feature extractor network and (ii) a self-attention module which further refined the convolutional features extracted from the CNN. [6]

In the experimental study, we have considered the FEV1 scores, stool microbes and screening for DNA through CFTR gene mutations, from already existing samples to improve diagnosis for the newborns at hospitals. This procedure helps us to train the gradient boosting decision tree model to obtain the highest accuracy among the existing methods. In this the model learns from a number of weak models. It modifies the error obtained in the previous model and trains it for a better accuracy.

About 60 CF patients FEV1 scores, sputum counts were noted and the data was preprocessed and trained as a split data with 60% in the training 10% in validation and 30% in testing to obtain the results. Through this we could also study certain relationships between the infections, bacteria and other scores. It helped in creating a clear dataset to train the model with relative threshold value. In the dataset fed to the decision tree, faulty measures were fed into the model for a possibility when screening, due to a glitch it might produce a wrong value. The model in that case would detect the false measures which has improved our error rate to drop to 10%.

The accuracy improved to 92% when the correlation between features was detected and the model was tested to recognize the same.

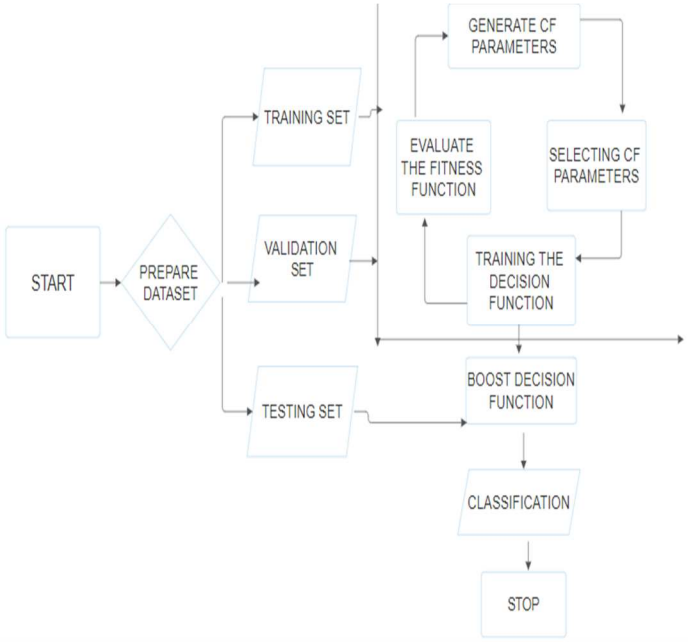

Figure 1 – Flow graph of Gradient Boost Decision Trees

During the advancement of the disease the lung infection increases, which results in the drop in FEV1 scores. The microbes causing stool infections increase which in turn cause the symptoms to rise. This progression shows the correlation between different features which makes our model optimized

IV. RESULTS AND DISCUSSION

The accuracy scores using each machine learning technique has been displayed in the following table. Table 2 contains the algorithm with their accuracy scores compared with our GBDT model. Accuracy of gradient boosting decision tree model is 92%. Second highest and closest to our model is the model made of logistic regression with an accuracy of 89%. The formula used to measure accuracy (1) is:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

TABLE 2 - ALGORITHMS WITH THEIR ACCURACY SCORES

Algorithm	Accuracy Score (%)
Convolutional Neural Networks (CNN)	74
Random Forest (RF)	73
K- Nearest Neighbors (KNN)	55
Support Vector Machine (SVM)	78
Logistic Regression (LR)	89
Bayesian Optimization (BO)	63
Gradient Boosting Decision trees (GBDT)	92

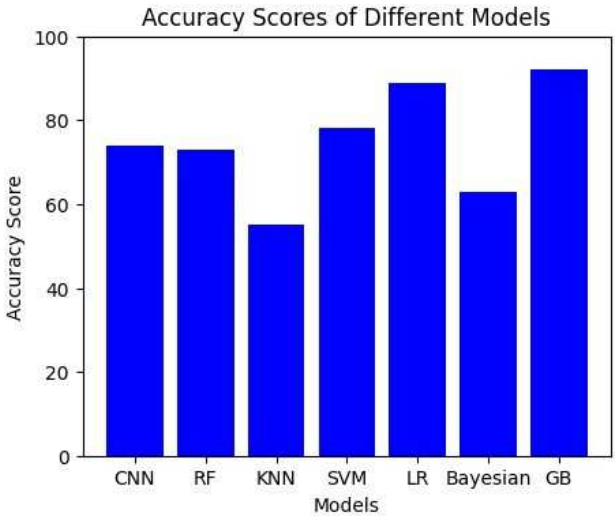

Figure 1- Accuracy score of different models

Table 3 contains the algorithms against their error rate in a comparison with GBDT measures, whose value is 8%. Multiple trial and error methods, using weak decision trees and repeatedly feeding that result into the developing tree has made this outcome successful. The figure 2 also shows the relative errors. The least of those errors being of the gradient boosting.

TABLE 3 - ALGORITHMS WITH THEIR ERROR RATES

Algorithm	Error Rate (%)
Convolutional Neural Networks (CNN)	26
Random Forest	27
K- Nearest Neighbors (KNN)	45
Support Vector Machine (SVM)	22
Logistic Regression	11
Bayesian Optimization	37
Gradient Boosting Decision trees	8

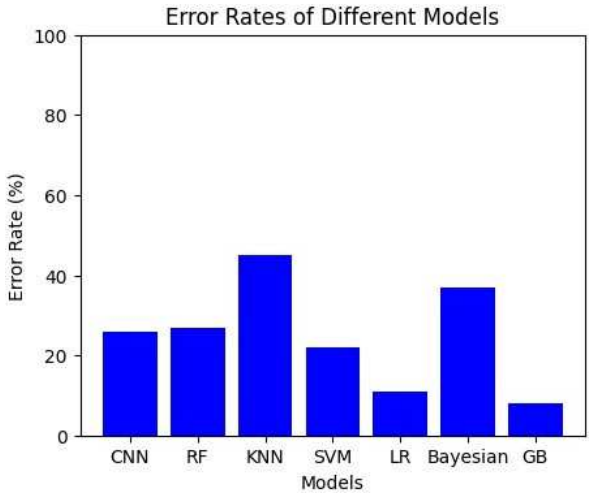

Figure 2 - Error Rates of Different Models

Table 4 contains the algorithms against the scores of their performance measures showing the gradient boosting

model's performance to be better than the existing research. Among the sensitivity score of the performed algorithms GBDT has the highest sensitivity of 95. All these values have been calculated depending on the elements from the confusion matrix.

TABLE 4 - ALGORITHMS WITH THEIR PERFORMANCE

Precision	Specificity	Sensitivity	F1 Score	
93	60	95	94	Gradient Boosting Decision trees
89	58	90	91	Logistic Regression
76	62	73	71	Support Vector Machine (SVM)
54	60	55	58	K- Nearest Neighbors (KNN)
74	71	69	76	Random Forest
70	69	68	71	Convolutional Neural Networks (CNN)
65	64	59	66	Bayesian Optimization

.

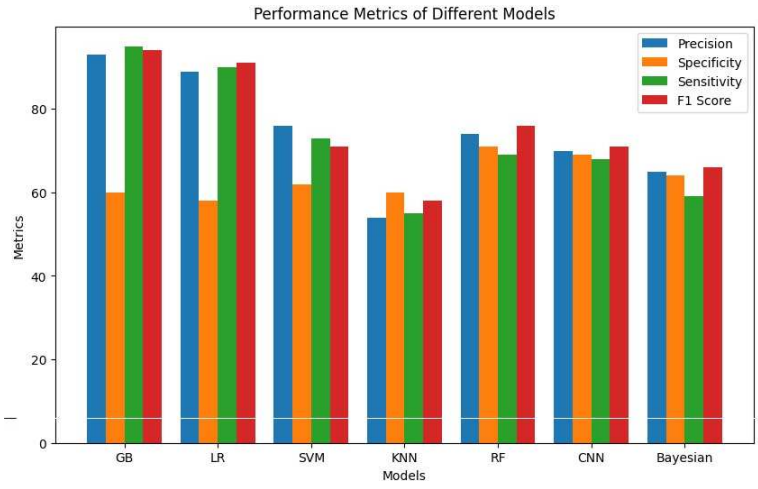

Figure 3 - Performance Measures of Different Algorithms

V. CONCLUSION

This study focuses on improving the overall performance of the model to enhance learning and practical usage. The research starts with obtaining data of the CF patients and processing it to understand the features and their importance. The paper later slides through feature selection and which features are optimal for the better performance of the model. In this case they were, sweat chloride tests, stool microbes, lung infections and correlations between them.

The accuracy as derived from the repeated training of the dataset has produced 92%. Lowest accuracy observed in the model is that of the KNN. After identifying the correlation the optimality reduced the error rate to 8% only. The specificity of GBDT model is 60 and has a precision of over 90. To prove that GBDT performs better that any other model the F1 score is calculated to be 94, which shows that even if the data remains unbalanced it still helps in the classification for gradient boosting to be a dominating algorithm.

To optimize our parameters 30% of the data was held for testing. It also helped in performing least errors on the overall dataset. Comparing the performance measures held in table 4, the F1 score is the highest for gradient boosting model and lowest for KNN. It shows that the classification for our chosen model of decision trees is better in many ways. To conclude, gradient boosting model not only gives a better performance but also helps in understanding that determining the correlation between the features is as important as the feature selection. Decision trees of this boosting method also provide with tuning options to optimise the loss functions for the betterment of the model.

With an advantage of having a flexibility the requirement of data pre-processing is often negligible. Instead numerical and categorical data can be fed into the model as they are and that itself makes the model great compared to other models. Gradient boosting is also faster compared to any other boosting models like XGBoosting which makes it more reliable for tasks like diagnosis having time constraints. Although, it may require the time and resources to put the model to a wide scale usage, it is worth it given the performance of the model. The predictors in this model provide a lesser error rate as they train the faulty predictors which had failed previously.
This is the reason why we chose the Gradient Boost decision trees in a study of diagnosis of a rare disease called Cystic Fibrosis.

VI. REFERENCES

[1] Zhenpeng Zhou, Daniel Alvarez. Carlos Milla and Richard N.Zare, "Proof of concept for identifying cystic fibrosis from perspiration samples" in journal list of national library of medicine - 2019.

[2] PR Anisha, CKK Reddy, K Apoorva, CM Mangipudi, "Early Diagnosis of breast cancer Prediction using Random Forest Classifier" in IDP conference Series Materials Science and Engineering 1116(1).012187.

[3] V Reddy, M Allugunti, "Internet of things based on early detection of diabetes using machine learning algorithms": Dpa in International Journal of Innovative Technology and Exploring Engineering.Article in a conference proceedings:

[4] PR Anisha and Narsimha Prasad "LVA Pragmatic Approach for detecting liver cancer".

[5] CKK Reddy, RG Reddy, P.Ramesh, PS Nidhi, LL Gayatri, B Subbarayudu, "Comparitive analysis on sorting and searching algorithms " in UCIET.

[6] Marta Rachel, Stanislaw Topolewicz and Sabina Galiniak , "Detection of Cystic Fibrosis Symptoms Based on X-Ray Images Using Machine Learning – Pilot Study" in biomedical journal of scientific and technical research - 2019-2020.

[7] Ahmed M.Alaa and Mihaela van der Schaar, "Prognostication and risk factors for cystic fibrosis via automated learning" in scientific reports - 2018.

[8] Mark D.Schluchter, Michael W.Konstan, Mitchell L. Drumm, James R.Yankaskas and Michael R. knowles, "Classifying Severity of Cystic fibrosis Lung disease using longitudinal Pulmonary Function Data" in American Journal of Respiratory and Critical Care Medicine.

[9] Nicole Filipow, Gwyneth Davies, Eleanor Main, Neil J Sebire, Colin Walis, Felix Ratjen, Sanja Stanojevic, "Unsupervised phenotypic clustering for determining clinical status in children with cystic fibrosis".

[10] P.R Anisha, C. Kishore Kumar Reddy, NG Nguyem, "Blockchain Technology: A Boon at the pandemic Times - A solution for Global Economy Upliftment with AI and IoT", 227-252 in Springer innovations in communication and computing - 2021.

[11]. C. Kishore Kumar Reddy, PR Anisha, R. Shastry, BV Ramana Murthy, "Comparitive study on internet of things: enablers and constraints"- 3rd ICDECT - 2019, 2k20- Springer.

[12] Mark D. Schluchter, Michael W. Konstan, Mitchell L.Drumm, James R. Yankassas and Michael R. Knowles, "Classifying Severity of Cystic Fibrosis Lung Disease Using Longitudinal Pulmonary Function Data.

[13]. Janelle Wells, Marjorie Rosenberg, Gary Hoffman, Michael Anstead, Philip M.Farrell, "A Decision- Tree Approach to Cost Comparison of Newborn Screening Strategies for Cystic Fibrosis".

[14] Kris De Boeck "Cystic fibrosis in the year 2020: A disease with a new face", published in Acta paediatrica nurturing the child, 03 January 2020.

[15] Lucy Allen, Lorna Allen, Siobhan B. Carr, Gwyneth Davies, Damian Downey, Marie Egan, Julian T. Forton, Robert Gray, Charles Haworth, Alexander Horsley, Alan R. Smyth, Kevin W. Southern & Jane C. Davies in Nature Communications volume 14, Article number: 693 (2023) " Future therapies for cystic fibrosis" 08 February 2023.

[16] Elaine Yu; Sandeep Sharma "Cystic Fibrosis" in The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. August 8, 2022.

[17] Simon Y Graeber, Marcus A Mall, on "The future of cystic fibrosis treatment: from disease mechanisms to novel therapeutic approaches" in The Lancet Journal on September 30, 2023.

[18] Jennifer L Taylor-Cousar, Paul D Robinson, Michal Shteinberg, Damian G Downey on "CFTR modulator therapy: transforming the landscape of clinical care in cystic fibrosis" in The Lancet Journal on September 30, 2023.

[19] Pierre-Régis Burgel, Espérie Burnet, Lucile Regard, Clémence Martin on "The Changing Epidemiology of Cystic Fibrosis: The Implications for Adult Care" in Chest on January 2023.