Covid 19 Cough Classification Using Machine Learning

1st Abhijeet Deshmukh
Instrumentation and Control Engineering
College of Engineering, Pune
Pune,India
deshmukhar19.instru@coep.ac.in

2nd Anisha Raut
Instrumentation and Control Engineering
College of Engineering, Pune
Pune, India
rautab19.instru@coep.ac.in

3rd Dr. Amruta Deshpande
Dept. of Instrumentation and Control
College of Engineering, Pune
Pune, India

***Abstract*—This document is a LaTex report for the Mini-Project which aims at classifying Covid-19 cough using various Machine Learning Models.**

***Index Terms*—COVID-19 Cough classification Machine learning SMOTE K-nearest neighbour (KNN) Catboost Classifier Spectral features**

I. Introduction

Cough and respiratory sound processing can assist in the early diagnosis of infections such as Covid-19. Even asymptomatic Covid-19 patients can be diagnosed early enough if appropriate speech modeling and signal-processing is applied. Covid-19 affects various speech subsystems that are involved in respiration, phonation and articulation.Clinically, coughs are identified by an underlying cause which can be due to common cold, bacterial infection, a hereditary disease like cystic fibrosis or viral infection such as influenza or coronaviruses. However, few coughs are also idiopathic like an acid reflux cough. These are presented in various forms like dry cough. SARS-CoV-2, a new coronavirus, has a trademark dry cough which is an important marker for analysis of cough to classify the disease correctly. Dry cough has a characteristic sound that comes from the upper respiratory tract while wet or chesty cough involves lower respiratory tract with mucus activity.

II. Data

The Research Paper we referred to has made use of two datasets: the Coswara Dataset and the Sarcos Dataset. Coswara Dataset is publicly open dataset hence was easily available. The pre-processed Coswara dataset, used for feature extraction and classifier training contains total 2278 audio samples out of which 482 are positive.The Subjects in dataset are from five continents: Asia (Bahrain, Bangladesh, China, India, Indonesia, Iran, Japan, Malaysia, Oman, Philippines, Qatar, Saudi Arabia, Singapore, Sri Lanka, United Arab Emirates), Australia, Europe (Belgium, Finland, France, Germany, Ireland, Netherlands, Norway, Romania, Spain, Sweden, Switzerland, Ukraine, United Kingdom), North America (Canada, United States), and South America (Argentina, Mexico). Age, gender, geographical location, current health status and pre-existing medical conditions are also recorded. Health status includes 'healthy', 'exposed', 'cured' or 'infected'. Audio recordings were sampled at 44.1 KHz

III. Data Extraction And Pre-processing

The raw cough audio recordings from dataset have the sampling rate of 44.1 KHz and is subjected to some simple pre-processing steps, described below. We aim at finding the mfcc coefficients of each sample. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

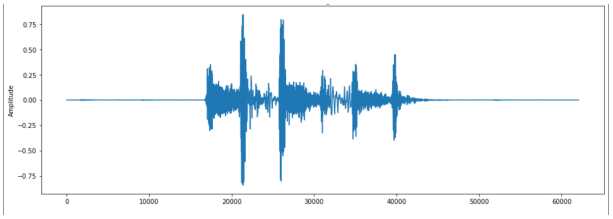

Fig. 1. Raw wave of cough audio sample

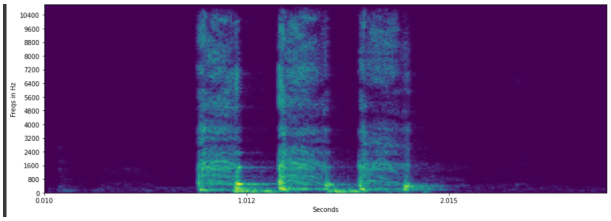

Fig. 2. Spectrogram of cough audio sample

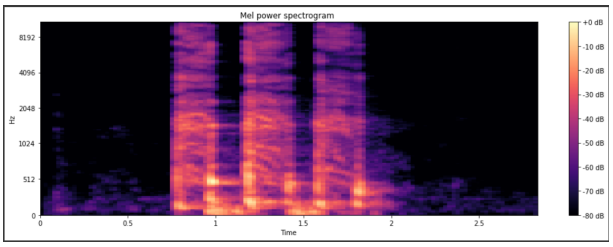

Fig. 3. Mel Power Spectrogram of cough audio sample

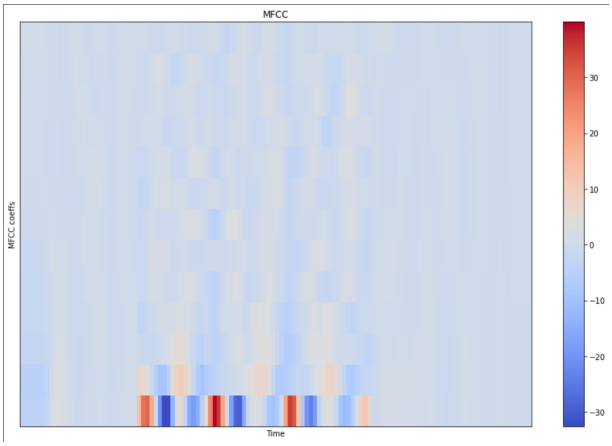

Fig. 4. mfcc coefficients of cough audio sample

A. Spectrogram of Cough audio sample

A spectrogram is a visual way of representing the signal strength, or "loudness", of a signal over time at various frequencies present in a particular waveform. Not only can one see whether there is more or less energy at, but one can also see how energy levels vary over time.

B. Mel power spectrogram

The mel spectrogram remaps the values in hertz to the mel scale.The mel scale is a scale of pitches that human hearing generally perceives to be equidistant from each other. As frequency increases, the interval, in hertz, between mel scale values (or simply mels) increases. The linear audio spectrogram is ideally suited for applications where all frequencies have equal importance, while mel spectrograms are better suited for applications that need to model human hearing perception. Mel spectrogram data is also suited for use in audio classification applications.

C. Feature Extraction

The features extracted from all the audio files of 'heavy cough' include:

- chroma stft :Computing a chromagram from a waveform or power spectrogram. Chromagram is defined as the whole spectral audio information mapped into one octave. Each octave is divided into 12 bins representing each one semitone. The chroma feature is a descriptor, which represents the tonal content of audio signal in a condensed form.
- RMS: Compute root-mean-square (RMS) value for each frame, from a spectrogram S.
- Spectral Centroid: The spectral centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the center of mass of the spectrum is located.
- Spectral Bandwidth: The spectral bandwidth or spectral spread is derived from the spectral centroid. It is the spectral range of interest around the centroid, that is, the variance from the spectral centroid.
- Roll off: Spectral roll off is the frequency below which a specified percentage of the total spectral energy lies.
- Zero Crossing Rate: The zero-crossing rate (ZCR) is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.
- MFCC coefficients: MFCCs have been used very successfully as features in audio analysis and especially in automatic speech recognition [42,43]. They have also been found to be useful in differentiating dry coughs from wet coughs and classifying tuberculosis coughs Mel Frequency Cepstral Coefficients has 39 features.The 39 MFCC features parameters are 12 Cepstrum coefficients plus the energy term. Here we considered 20 MFCC features.

D. Dataset Balancing

COVID-19 positive subjects are under represented in both datasets. To compensate for this imbalance, which can detrimentally affect machine learning, we have applied SMOTE (Synthetic Minority Oversampling Technique) data balancing to create equal number of COVID-19 positive coughs during training. This technique has previously been successfully applied to cough detection and classification based on audio recordings. SMOTE oversamples the minor class by generating synthetic examples.

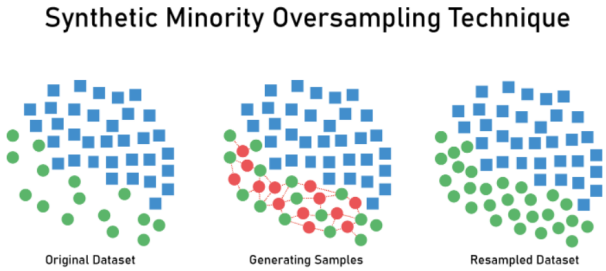

Fig. 5. Dataset balancing using SMOTE

E. Machine Learning Models

1) Catboost Classifier: CatBoost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters. To prevent overfitting, the overfitting detector is used. When it is triggered, trees stop being built.
2) Gradient Boosting Classifier: Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.
3) Light Gradient Boosting Machine: Light Gradient Boosted Machine, or LightGBM for short, is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. LightGBM extends the gradient boosting algorithm by adding

a type of automatic feature selection as well as focusing on boosting examples with larger gradients.

4) Extreme Gradient Boosting: Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.
5) Random Forest Classifier: It is suppervised learning algorithm and is widely used for classification and regression problems. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
6) Extra Trees Classifier: Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a forest to output it's classification result.
7) K Neighbors Classifier: The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to.
8) Naive Bayes: Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
9) Decision Tree Classifier: It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.
10) Dummy Classifier: The dummy classifier gives you a measure of "baseline" performance–i.e. the success rate one should expect to achieve even if simply guessing.

IV. Hyperparameter Optimization

For optimizing the the models , High end Optuna Hyperparameter optimization library is used on lightgbm model, the study is created for the optimization with 100 iterators, and the weight optimization over given features is as featured in the Parallel Coordinate plot Fig 6. After performing the study of hyperparameter optimization the final product of optimum parameters is observed using the Optimization plot history Fig 7. The importance of each feature is determined using the feature plot Fig 8

The best parameters that are obtained after tunning of LightGBM :

- boosting type: 'gbdt'.
- lambda l1: 2.996043971443949.
- lambda l2: 0.002792588987601735.
- colsample bytree: 0.5,
- bagging fraction: 0.8,

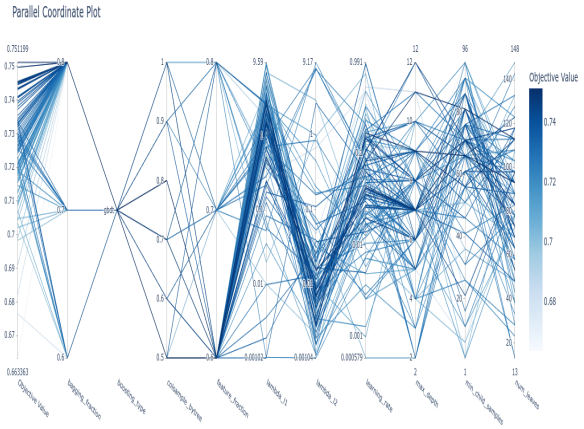

Fig. 6. Parallel Cooardinate Plot

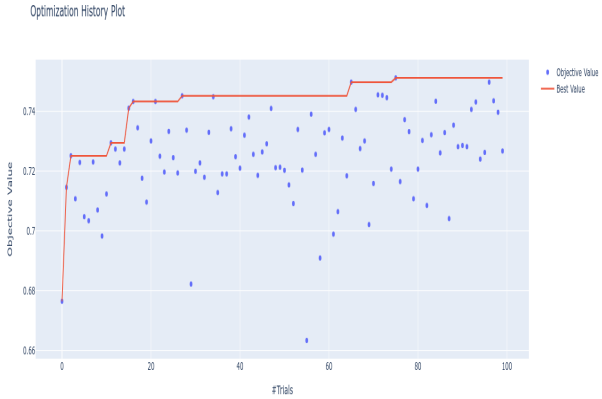

Fig. 7. Optimization History Plot

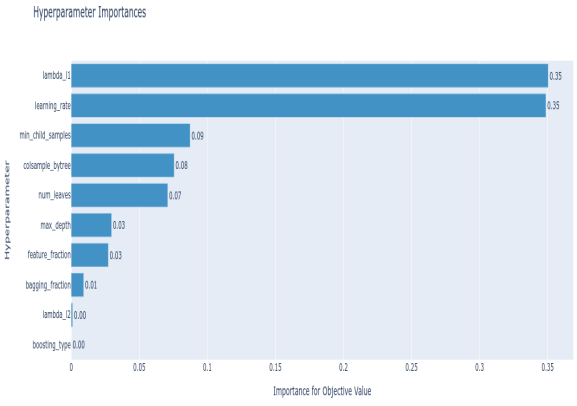

Fig. 8. Hyperparameter Importance : feature plot

- feature fraction: 0.6,
- learning rate: 0.16714392495530678,
- max depth: 9,
- num leaves: 93,
- min child samples: 66.

  the hyperparameter tunning improves the result of the model , the latency of the model to fail on test case data is highly optimized by applying the stratified K fold on validation set, same is done on Catboost and we got best result as discussed in result section .

V. Results

The Accuracy obtained by Catboost classifier Model is the highest which is 89.75with AUC score of 0.9256 The AUC score and Accuracy obtained using each machine learning model is as follows:

Sr.no	Model	Accuracy(%)	AUC
1	Catboost classifier	87.95	0.9256
2	Gradient Boosting Classifier	87.51	0.9231
3	Light Gradient Boosting Machine	87.52	0.9226
4	Extreme Gradinet Boosting	87.01	0.9182
5	Random Forest Classifier	85.69	9043
6	Extra trees classifier	85.94	0.895
7	K-neigbors Classifier	82.86	0.8467
8	Naïve Bayes	79.91	0.7959
9	Decision Tree Classifier	84.8	0.7723
10	Dummy Classifier	78.53	0.5

Fig. 9. Results of various ML models

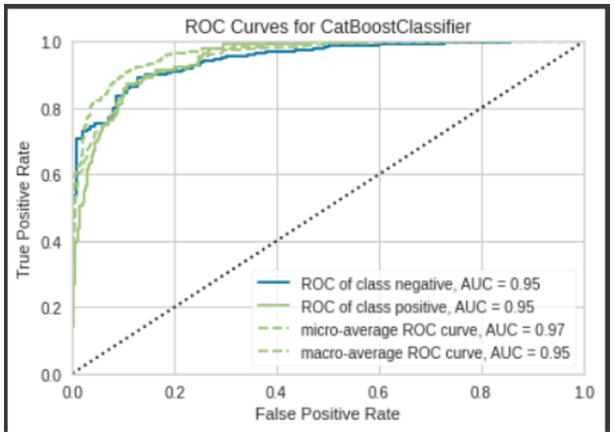

Fig. 10. ROC curves for Catboost Classifier

VI. Conclusion

We have developed COVID-19 cough classifiers using smartphone audio recordings and ten machine learning models. To train and evaluate these models, we have used the Coswara dataset.Our bestperforming classifier is the catboost classifier architecture and is able to discriminate between COVID-19 coughs and healthy coughs with an AUC of 0.9256 on the Coswara dataset. The dataset is publicly available and contains data from 2277 subjects (483 COVID-19 postive and 1794 sujects healthy) residing on all five continents except Africa.

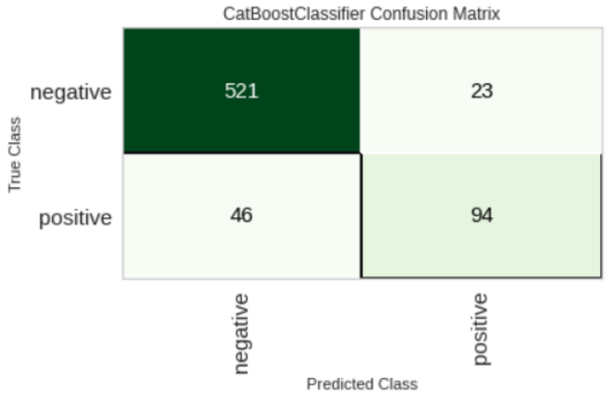

Fig. 11. ROC curves for Catboost Classifier

since better performance is achieved using a larger number of MFCCs than is required to mimic the human auditory system, we also conclude that at least some of the information used by the classifiers to discriminate the COVID-19 coughs and the non-COVID coughs may not be perceivable to the human ear. Although the systems we describe require more stringent validation on a larger dataset, the results we have presented are very promising and indicate that COVID-19 screening based on automatic classification of coughing sounds is viable. Since the data has been captured on smartphones, and since the classifier can in principle also be implemented on such device, such cough classification is cost-efficient, easy to apply and deploy. Furthermore, it could be applied remotely, thus avoiding contact with medical personnel.

VII. Future Scope

We have developed machine learning models for classifying Covid-19 positive and healthy samples. Further the dataset can be used to train CNN, DNN, ANN models such as Resnet50, vg16 etc to get more accurate results. Also Hyperparameter tuning remains an important aspect while designing these models which can be implemented to get better results.

References

[1] 1. COVID-19 cough classification using machine learning and global smartphone recordings Computers in Biology and Medicine 135 (2021) 104572 (ELSEVIER) Madhurananda Pahar , Marisa Klopper , Robin Warren , Thomas Niesler

[2] 2. Cough Classification for COVID-19 based on audio mfcc features using Convolutional Neural Networks. 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Oct 2-4, 2020 Vipin Bansal, Gaurav Pahwa ,Nirmal Kannan

[3] 3. Cough Sound Classification Based on Similarity Metrics 2021. 44th International Conference on Telecommunications and Signal Processing. IEEE Nikos Petrellis ,George K. Adam

[4] https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

[5] https://link.springer.com/content/pdf/bbm

[6] https://www.sciencedirect.com/science/article/pii/S0010482521003668

[7] https://medium.com/analytics-vidhya/bank-data-smote-b5cb01a5e0a2

[8] S. Bahl, R.P. Singh, M. Javaid, I.H. Khan, R. Vaishya, R. Suman. "Telemedicine technologies for confronting COVID-19 pandemic: A review," Journal of Industrial Integration and Management, 2020, vol. 5(4), pp. 547–561, doi: https://doi.org/10.1142/S2424862220300057

[9] D. Dong et al., "The role of imaging in the detection and management of COVID-19: a review," IEEE Reviews in Biomedical Engineering, doi: 10.1109/RBME.2020.2990959.

[10] M. Pahar, I. Miranda, A. Diacon, T. Niesler, Deep neural network based cough detection using bed-mounted accelerometer measurements, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 8002–8006, https://doi.org/10.1109/ ICASSP39728.2021.9414744.

[11] L.L. Blagus, R, SMOTE for high-dimensional class-imbalanced data, BMC Bioinf. 14 (2013) 106, https://doi.org/10.1186/1471-2105-14-106.