A Novel Intelligent Model For COVID-19 Detection Using Cough Auscultations and Hjorth Descriptors

Misha Urooj Khan*[1], Syeda Zuriat-e-Zehra Ali[2], Kanwal Habib[2], Hareem Khan[3], Faizan Tariq[2], Sannia Bibi[2]
[1]Department of Electronics Engineering, University of Engineering and Technology Taxila, Pakistan
[2] Department of Electrical Engineering, University of Engineering and Technology Taxila, Pakistan
[3] Department of Computer Engineering, University of Engineering and Technology Taxila, Pakistan
*misha.urooj@students.uettaxila.edu.pk

Abstract
COVID-19 has spread over the whole world gradually and has affected life, public health, and financial systems of various countries on a daily basis. Pathogenic laboratory tests, such as polymerase chain reaction (PCR), which take a long time and provide false-negative results, are considered to be the gold standard for its detection. World Health Organization (WHO) has declared COVID-19 an epidemic, as it affected more than 50 million people and killed 14 million globally. In this study, we proposed an efficient and novel algorithm for the diagnosis of COVID-19 disease from cough auscultations. We used a self-collected dataset of 1579 cough auscultations (CA) which were collected from different local hospitals. Dataset is processed by removing dc components and amplitude normalization. Then Region of interest (ROI) i.e., the part which contains low-frequency components is extracted by Empirical Mode Decomposition. Hjorth descriptor is applied on pre-processed and segmented signals to get activity, mobility, and complexity features. These extracted features are given as an input to Medium-KNN and Fine Decision Tree classifiers resulting in a cumulative accuracy of 99.8% and 94.9 % respectively. This developed system will help Pakistani doctors in non-invasive detection of COVID-19 and classification of coughs accurately.

Keywords
COVID-19, Cough auscultation, Machine learning, Hjorth descriptor, Intelligent System.

I. Introduction

Coronavirus was detected in the Chinese city of Wuhan and was reported to the World Health Organization (W.H.O.) on December 31, 2019. COVID-19 is a virus family that includes SARS (Severe Acute Respiratory Syndrome) and ARDS (Acute Respiratory Distress Syndrome). This outbreak has been considered a public health concern by the WHO [2] and stated that the virus is transmitted through the respiratory tract when an uninfected person comes into contact with an infected one. Within 2–14 days of this contact, the newly infected person develops symptoms. Dry cough, weakness, and fever are physical symptoms in mild to medium cases, while dyspnea (shortness of breath), tiredness, and fever are signs and symptoms of extreme cases. People who have other illnesses, such as diabetes, heart disease, or asthma, are more susceptible to the virus and can become seriously ill as a result. SARS is an infectious disease that first surfaced in China in 2003 and spread to 26 countries, with 8000 cases reported in the same year. It is passed on from one person to the next and is characterized by fever, chills, diarrhea, dyspnea, malaise, myalgia, and shivering. Acute respiratory distress syndrome (ARDS) is described by a sudden onset of lung inflammation followed by respiratory failure. PaO2/FiO2 ratios lower than 300 mm Hg are diagnostic of ARDS. As of April 28, 2021, the novel coronavirus-2019 (COVID-19) disease had already claimed over 2.53 million lives and infected over 114 million people globally (**Figure 1**) [3].

COVID-19 is currently diagnosed using time-consuming, costly, and inaccessible RT-PCR (reverse transcription polymer chain reaction) testing. Owing to a lack of medical equipment, medical personnel, and healthcare facilities in certain areas, such kit is not readily accessible. Machine learning, in addition to clinical techniques, greatly aids in disease detection using an image, sound, and textual data. Novel coronaviruses can be identified using machine learning techniques. Additionally, it can forecast the virus's varying nature around the world. But machine learning requires a massive amount of data to classify or predict this disease. Symptoms of COVID-19 vary considerably but commonly include coughing, loss of taste or smell, sore throat, headache, dyspnea, and fever. As a result, several research papers have been published recently to develop sound-based solutions for automatic diagnosis. Although developments in testing have increased the prevalence of these methods in recent days, the need for inexpensive, rapid, and scalable COVID-19 screening technology is critical.

In this work, we propose the use of cough auscultation which are basically the cough sounds heard within the lungs to diagnose COVID-19. This classification algorithm can be used as a pre-screening tool to lessen the pressure on health centers and to provide a more rapid and reliable testing mechanism for preventing this virus spread.

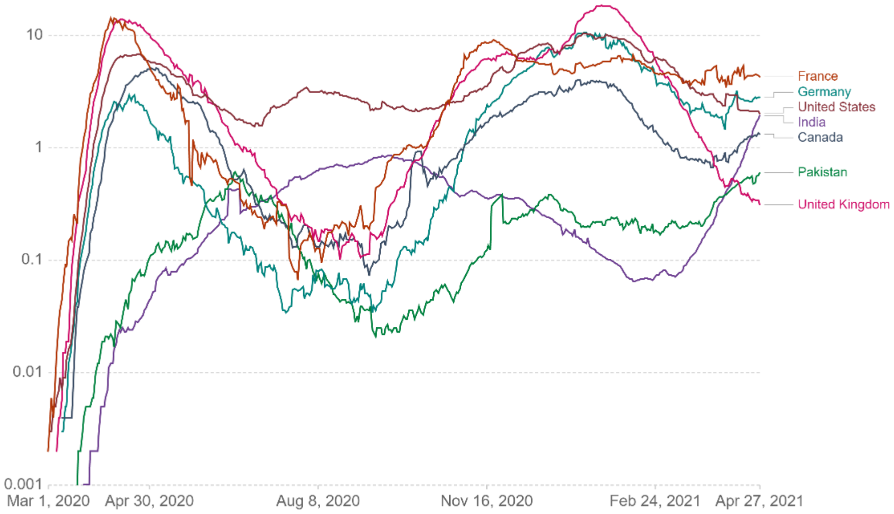

Figure 1. COVID-19 Deaths per million in different countries [3]

II. Literature Review

A computer system was developed in [1] to detect cough in two stages. For calculation of high-level data frames, mean, and standard deviation were computed. The results had 92.7% sensitivity, 88.58% precision, and 90.69% region under the curve. For automatic cough detection in smartphone-acquired audio signals, a novel feature Hu moment was proposed in [2]. After feature extraction, ML classifier (KNN) were used for classification. The sensitivity and specificity of this system were 88.51% and 99.7% respectively. To distinguish pneumonia and other lungs diseases, a detailed cough analysis was held by authors in [3]. Datasets of cough sounds from 91 patients were obtained and a wavelet-based algorithm is applied to extract features. For classification, Logistic Regression was used. The results based on sensitivity and specificity were 94% and 63% respectively. Cough events were identified using acoustic signals in [4]. To distinguish cough and non-cough cases, the detection algorithm used LPC coefficient, tonality index, spectral flatness, and spectral centroid. This system's sensitivity, accuracy, and F1-score were 86.78%, 99.42%, and 88.74% respectively. A speech recognition algorithm to identify cough events was presented in [5]. Gaussian Mixture Model-based Hidden Markov Model (GMM-HMM) hybrid method was used, as well as Perceptual Linear Prediction and Mel Frequency Cepstral Coefficients. In [6], authors presented CNN and transfer learning based automated detection from datasets of normal, pneumonia, and COVID-19 X-ray images. The results of this study showed 96.78% accuracy, 96.46% specificity, and 98.66% sensitivity.

In an experimental study on 2500 computerized tomography (CT) scans from normal, tumour, COVID-19 lungs, Zhou Tao et al. presented a new technique for detection of COVID-19. Evaluation of algorithm using ELD-COVID performance, sensitivity, and other parameters which shows accuracy of 99.054% and 342.92 sec detection speed [7]. A full automated machine learning (ML) and IOT based diagnostic system for COVID-19 in smart hospitals is discussed in [8]. Three ML models were applied on normal, normalized, featured selected laboratory findings and best ML model was used for comparison with benchmark studies. The study’s findings shows that SVM is best model with 95% accuracy and 94% recall rate [8]. Convolutional Neural Network is a fast and efficient detection tool for COVID-19. This model is also being used in Radiology dept. in Bahwal Victoria Hospital, Pakistan. This proposed model presents best accuracy of 96.68% [9]. For COVID-19 detection, Turker [10] introduced a novel fuzzy tree-based classification model. On 3-level F-transformed images, Exemplar division (ML model) was applied. Multi-Kernel-Binary-Pattern (MKBP) and Iterative Neighbourhood Component Algorithm (INCA) were used for feature extraction and feature selection, respectively. These feature vectors were applied on ML algorithms with SVM providing best accuracy of 97.01% among all for the data set of COVID-19, Normal, and Pneumonia. Publicly accessible deep learning CT diagnosis-based model was developed data and classification was performed on images collected from normal and bacterial infected patients. This model was successful in obtaining AUC of 95% and accuracy of 86% [11]. A comprehensive comparative investigation was studied in [12] consisting of 15 audio and less known features to aid the creation of classification models in Machine Learning. The system improved the classification accuracy of Cambridge dataset by 17% and on Coswara dataset it increased by 10%. Tong Xia et al. proposed an ensemble method employing various deep learning frameworks using balanced datasets to detect COVID-19 from sounds. The system achieved an AUC value of 0.74, 0.68 sensitivity and 0.69 specificity [13].

III. Material and Methods

1. Architecture of Proposed Methodology

Figure 2 shows the proposed methodology for the detection of COVID-19 subjects. A stethoscope fitted with a microphone records the cough auscultations (CA). The recorded raw CA signal is pre-processed by dc removal and amplitude normalization. The pre-processed signal is segmented by using the filtration technique. Now CA contains no wasteful or unwanted or distorted information/data. After that, three Hjorth descriptor features (complexity, mobility, activity) are derived and merged to create a concise and accurate feature depiction of covid and non-covid coughs. These attributes are then used to train the Weighted K-nearest neighbor (KNN) and Fine Decision Tree classifiers with accuracy of 99.8% and 94.4% respectively.

2. Cough Auscultation Collection

Here we have designed low-cost and efficient Auscultation Sensor (AS) to gather sound signals for the detection of viruses. The system includes a stethoscope, a sensitive microphone, an amplifier circuit, and a sound card. By placing AS near the left /right lung, the microphone starts sensing cough auscultations (CA). During the compilation of CA, each participant produced 25 seconds long signal. Later we divided these signals in to 5 seconds so pre-processing and segmentation can be done easily. These Cough auscultations are collected from different local hospitals. The collected database includes both men and women of various ages. The frequency range of the cough auscultation is 1-2 kHz and the sampling frequency of the experimental signal is set to 8kHz by using the Nyquist sampling theorem. **Table 1** shows the detailed statistics of the collected data and **Figure 3** shows the raw cough auscultations of the collected lung sounds.

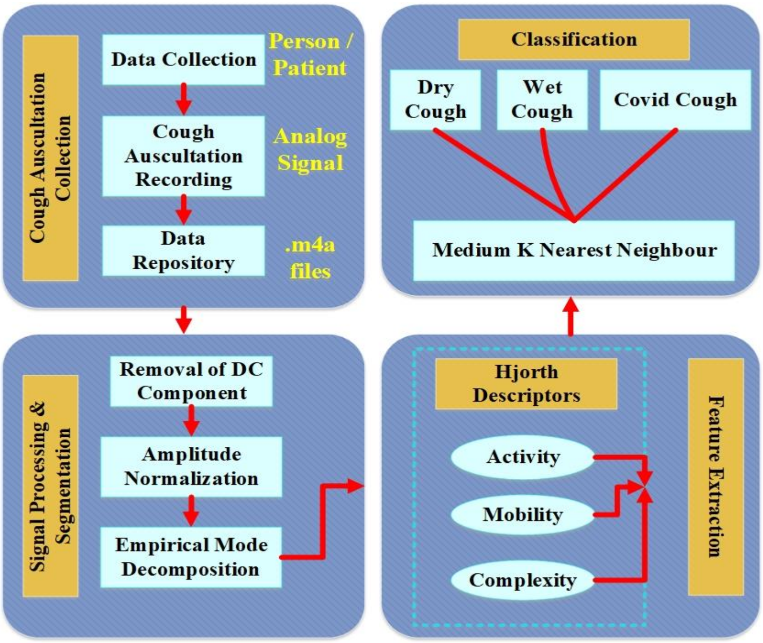

Figure 2. Block Diagram of Proposed Methodology

Table 1 Cough Auscultation Database Statistics

Cough Auscultations	Female Subjects			Male Subjects			Total
	Female	**Samples**	**Age**	**Male**	**Samples**	**Age**	
COVID-19	29	145	15-54	49	248	11-60	393
Wet Cough	62	313	18-52	51	257	21-64	570
Dry Cough	57	289	33-65	65	327	25-75	616

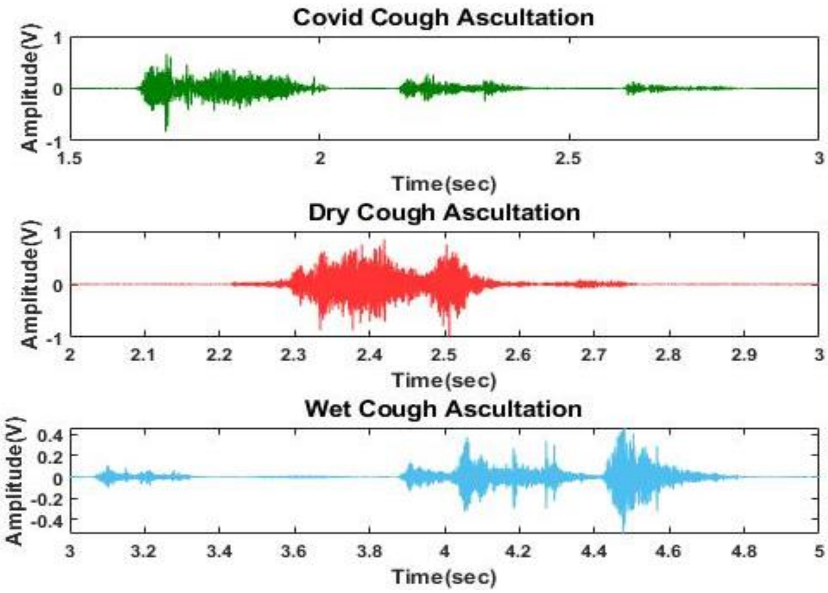

Figure 3. Raw Signal of Cough Auscultation (a) COVID Cough (b) Dry Cough (c) Wet Cough

3. Signal Pre-processing

The first step in signal processing is elimination of DC parts of signal. Filtering algorithms are essential to keep the fundamental component from input signals and eliminate DC component and harmonics. DC component is the non-periodic signal and has a large frequency spectrum. If *CA*(i) is input signal, then the DC component free signal CA_{dc} shown in **Figure 4** is,

$$CA_{dc} = CA(i) \; - \; \sum_{i=1}^{N} CA(\mathrm{i}) \qquad \text{i = 1, 2, 3........N} \qquad (1)$$

The second preprocessing step is normalization. *Normalization* is essential for algorithms to model data accurately. It maintains the ratios and general distributions in the data and keeps the values across all the columns within a certain limit. We can apply normalization to single or multiple columns in the same dataset. The formula for normalization is.

$$CA = \frac{CA-CAmin}{CAmax-CAmin} \qquad (2)$$

Here $CA_{\max}$ and $CA_{\min}$ are maximum and minimum values of data. Amplitude normalization is done with the operation as $CA(\mathrm{i}) \; = \; \frac{CA(\mathrm{i})}{\max|\mathrm{CA}|}$.The range of $CA(\mathrm{i})$ will be from -20 to 20.

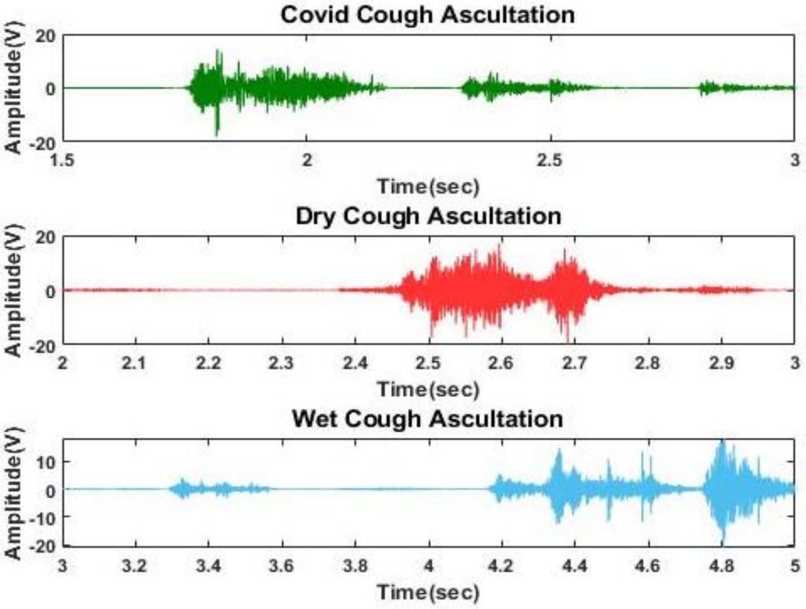

Figure 4. Pre-processed Signal of Cough Auscultation (a) COVID Cough (b) Dry Cough (c) Wet Cough

4. Signal Segmentation

Signal segmentation is done to segment out the region of interest (ROI) from a time domain signal. Cough auscultations (CA) have strong amplitudes in power spectrum where frequencies are less than 2kHz. So, here we will segment this region from our pre-processed CA whose frequency range is 8kHz with the help of Empirical Mode Decomposition (EMD).

The *Empirical Mode Decomposition (EMD) algorithm* is based on the premise that every non-stationary and non-linear signal is made up of a variety of simple intrinsic oscillation modes. The method's main idea is to empirically define these functions in the data based on their characteristic time scales, then decompose them accordingly. It takes signal oscillations into account at the most local level and divides them into non-overlapping time frame individually. *IMFs (Intrinsic mode functions)* can have similar frequencies at multiple time stamps (typically in less than 1% of cases) but global orthogonality cannot be guaranteed.

EMD decomposes a pre-processed cough auscultation $CA(t)$ into its constituent IMFs $CA_{IMF}(t)$ which are locally orthogonal, and a left-over signal known as residual signal $r(t)$. Mathematically,

$$CA_{emd}(t) = \sum_{N} CA_{IMF}(t) + r(t) \qquad (3)$$

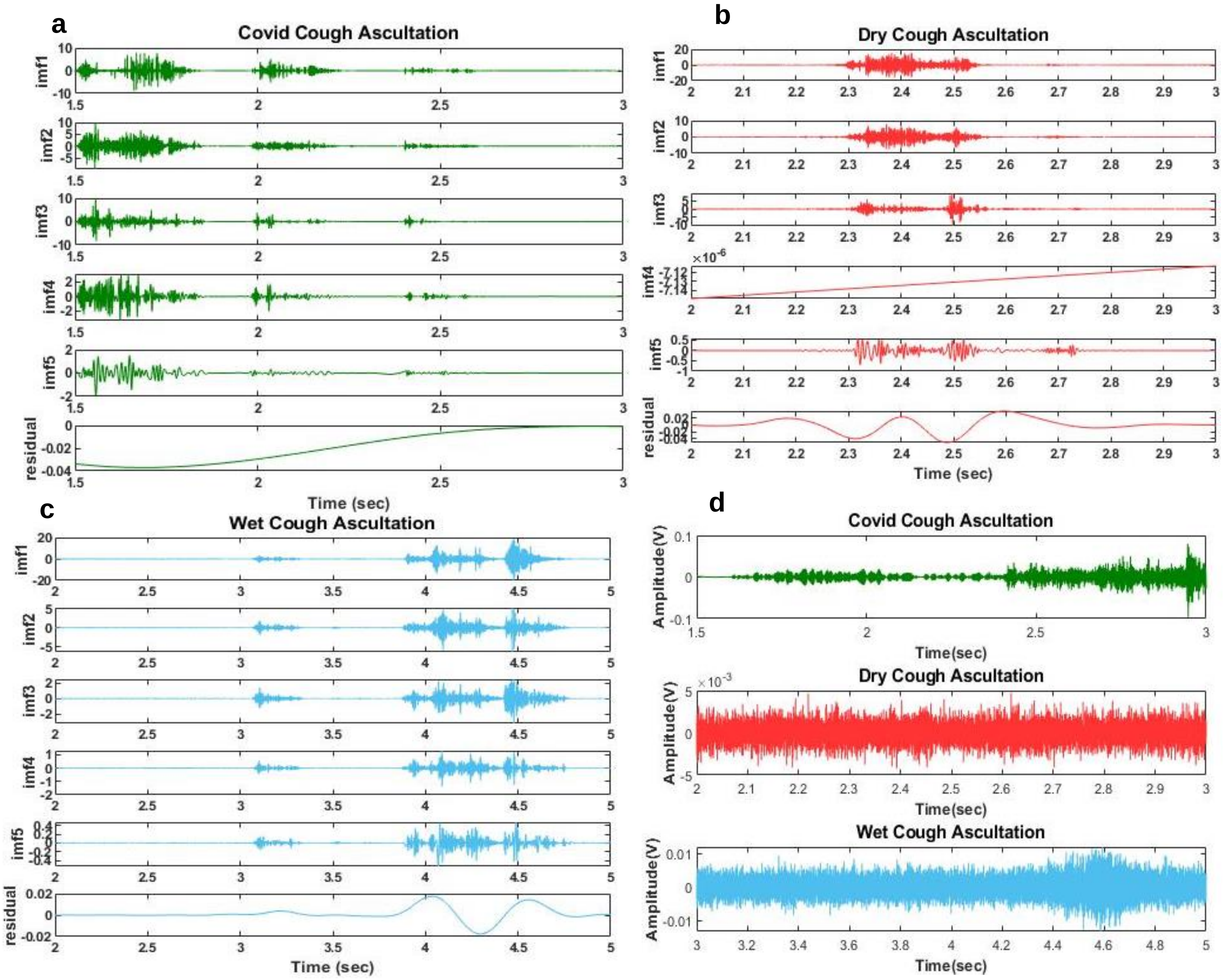

Figure 5. Empirical Mode Decomposition of (a) COVID-19Cough (b) Dry Cough (c) Wet Cough (d) Segmented Cough Auscultations

Figure 5 shows the decomposition of pre-processed CA signals into their respective IMFs. We summed IMF2-IMF3 to preserve the low frequency part while segmenting out the ambient noise, heart sounds and large frequency components. **Figure 6** shows the sum of IMF2-IMF3 of cough auscultations. **Table 2** shows IMFs and their respective frequency range.

Table 2 Cough Auscultations and their IMF's Frequency range

Parameters	Covid Cough	Wet Cough	Dry Cough
IMF1	0-1600Hz	0-2600Hz	0-2600Hz
IMF2	0-700Hz	0-1250Hz	0-1200Hz
IMF3	0-200Hz	0-600Hz	0-400Hz
IMF4	0-100Hz	0-250Hz	0-180Hz
IMF5	0-50Hz	0-10Hz	0-80Hz

5. Feature Extraction

Hjorth descriptors also called normalized shaped detectors (NSD) which consists of parameters that are used to define the complexity of biomedical signals. These parameters include complexity, mobility, and activity [1] as shown in **Figure 6**. These descriptors are widely used in previous studies as a measuring tool for human health using EEG time domain signals, EMG signals and lungs sounds [2][3]. Consider a cough auscultation signal represented as $x_{CA}(r)$ where Range r = 0,1,2…., R-1. $x_{(CA)}(n)'$ and $x_{CA}(n)''$ are respective first and second derivatives of signal elaborated as:

$$x_{CA}(r)' = x_{CA}(r) - x_{CA}(r-1) \qquad (4)$$

$$x_{CA}(r)'' = x_{CA}(r) - 2x_{CA}(r-1) - x_{CA}(r-2) \qquad (5)$$

σ_x is considered as standard deviation of x_{CA}(n). It is used for expression of three parameters of Hjorth descriptors names as activity, mobility and complexity as:

$$ACT_{CA} = \sigma_x^2 = \frac{\sum_{n=0}^{N-1} x_{CA}(r) - \bar{x}_{CA}}{R} \qquad (6)$$

$$MOB_{CA} = \frac{\sigma_x{}'}{\sigma_x} \qquad (7)$$

$$COM_{CA} = FF = \frac{MOB_{CA}{}'}{MOB_{CA}} = \frac{\sigma_x{}''/\sigma_x{}'}{\sigma_x{}'/\sigma_x} \qquad (8)$$

ACT_{CA} represent activity, MOB_{CA} is mobility and COM_{CA} is complexity.

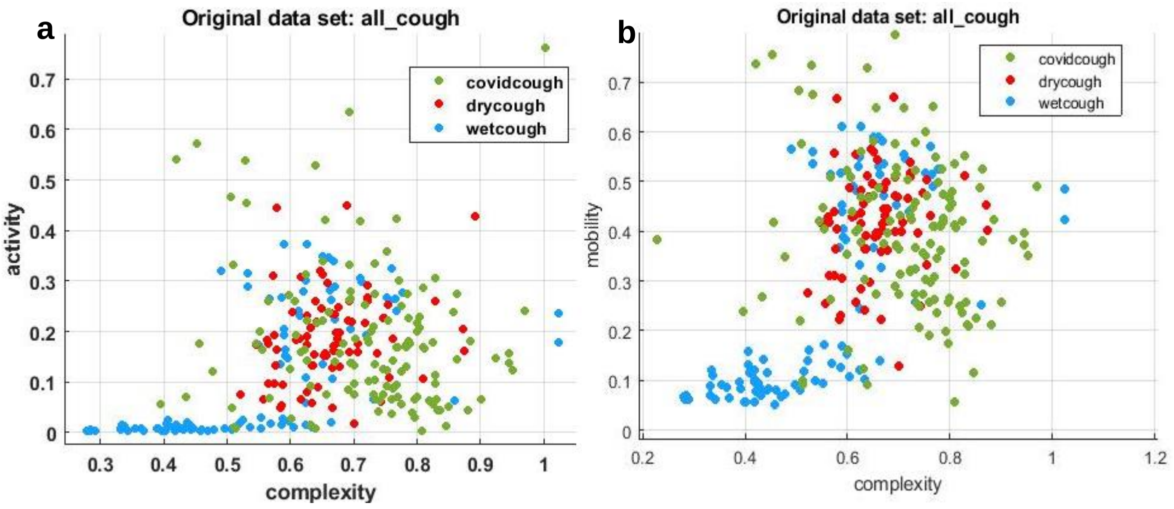

Figure 6. Extracted features (Hjorth Descriptors) (a) ACT_{CA} vs COM_{CA} (b) MOB_{CA} vs COM_{CA}

6. Classification

Classification is described as machine learning and statistics based supervised learning approach. In it, computer program learns from data and generates new observations or classifications. This is the process of classifying a collection of data. It is possible to apply it to both structured and unstructured data. The first step in the process is to predict the data point class. In terms of modelling, classification necessitates a training dataset containing a large number of examples of inputs and outputs from which to learn.

The *K Nearest Neighbour (KNN) algorithm* is based on the Supervised Learning approach. KNN assumes new and existing cases which are identical and assigns new case to the category nearest to existing labels.

The *Decision Tree (DT) algorithm* can be used to solve regression and classification problems. It learns basic decision rule from training data and creates model using this data for prediction of variables' class or significance.

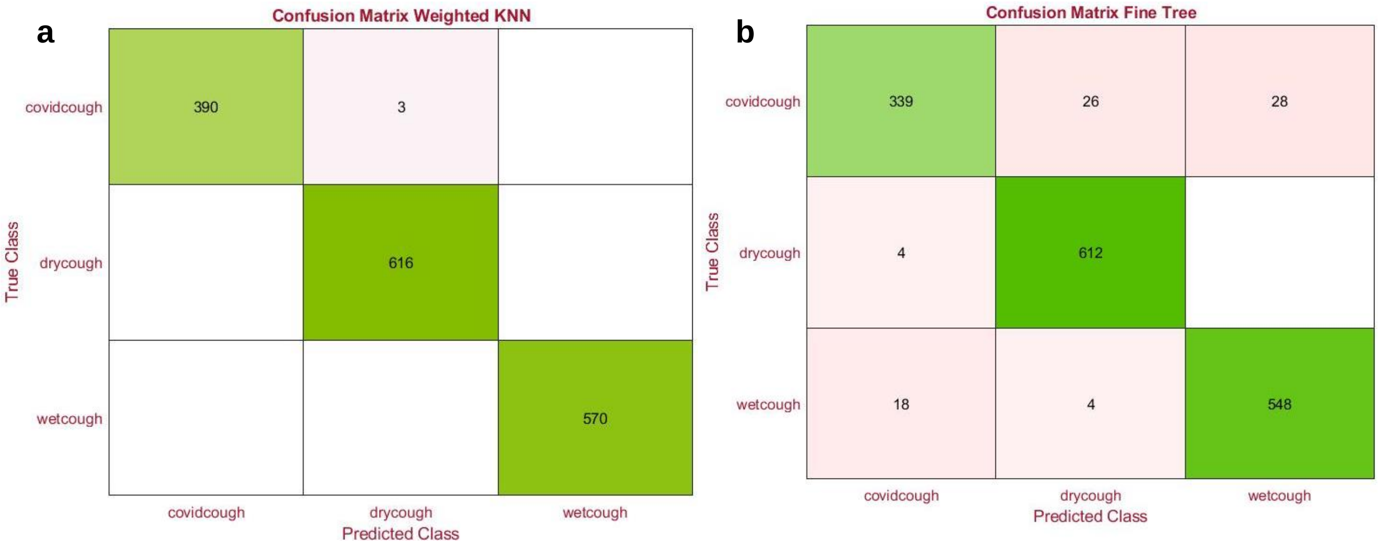

Figure 7. Confusion Matrix (a) Weighted KNN (b) Fine Tree

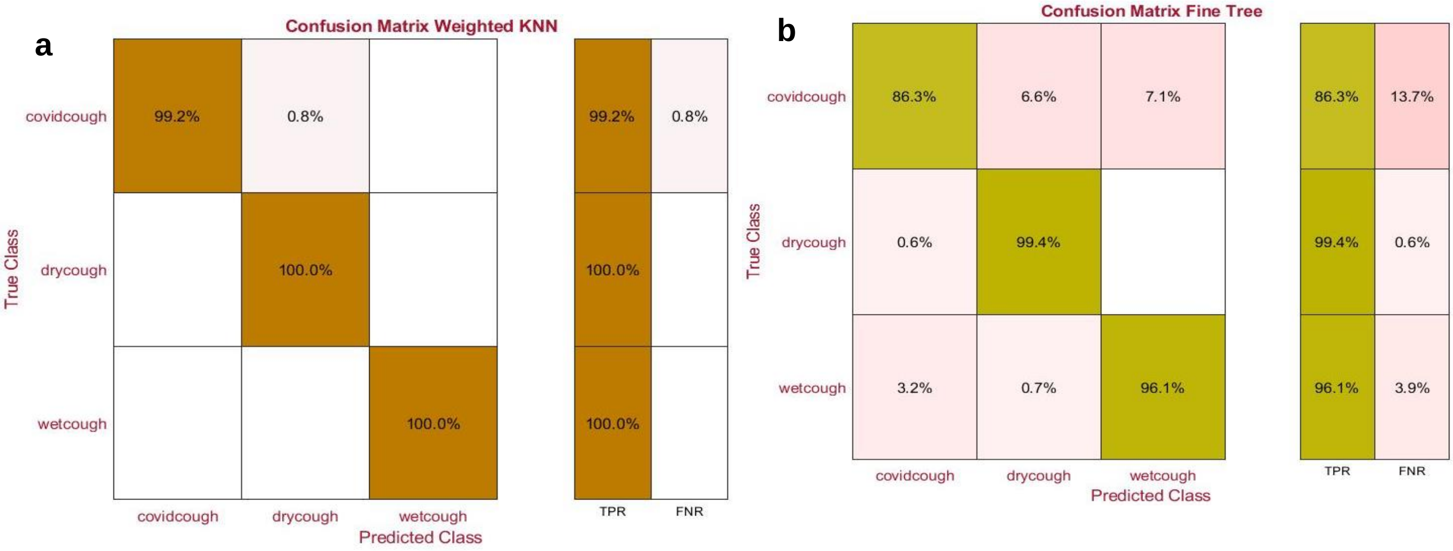

Figure 8. Confusion Matrix w.r.t True Positive Rate (TPR) and False Negative Rate (FNR) (a) Weighted KNN (b) Fine Tree

IV. Results and Discussion

i. Performance evaluation parameters

We used several performance metrics derived from the testing dataset by measuring sensitivity (S_e), specificity (S_c), false value (F_n), true value (T_p), accuracy (A_c), and Precision (P_p), for monitoring the effectiveness of parameters. Mathematically,

$$S_e = \frac{TP}{TP+FN} \times 100 \qquad (9)$$

$$S_c = \frac{TN}{FP+TN} \times 100 \qquad (10)$$

$$T_p = \frac{TP}{TP+FP} \times 100 \qquad (11)$$

$$F_n = \frac{TN}{TN+FN} \times 100 \qquad (12)$$

$$A_c = \frac{\sum TP+TN}{\sum TP+FP+FN+TN} \times 100 \qquad (13) \quad P_p = \frac{\sum TP}{\sum TP+FP} \times 100 \qquad (14)$$

Where TP denotes the quantity of accurately observed breath sounds, TN denotes that they actually do not have the disease and predicted as yes, FN signifies that they do have the disease but predicted as no, FP denotes the number of incorrectly predicted breath sounds chosen by the W-KNN classifier [10].

ii. Classifiers Performance

Weighted KNN (W-KNN) classifier took 472.22 seconds to train the Hjorth descriptors with a prediction rate of 29000 observations per second. The total miss-classification cost of the prediction value is 3. W-KNN used 10 nearest neighbors per prediction value. This classifier scored 99.8% sensitivity (S_e), 97.82% specificity (S_c), 99.8% accuracy (A_c), and Precision (P_p) is 99% as shown in **Figure 7**. The false value (F_n) for covid cough (CC) is 0.8% while true value (T_p) for CC is 99.2%, wet cough (WC) and dry cough (DC) is 100% as shown in **Figure 8**.

Fine Decision Tree (F-DT) classifier took 10.145 seconds to train the Hjorth descriptors with a prediction rate of 10000 observations per second. The total miss-classification cost of the prediction value is 80. F-DT uses 100 splits while splits criteria was Gini's diversity index. This classifier scored 93.93% sensitivity (S_e), 92.02% specificity (S_c), 94.4% accuracy (A_c), and 93.03% Precision (P_p). **Figure 7** and **8** shows confusion matrix of W-KNN and F-DT.

iii. Performance evaluation of Hjorth Descriptors

Activity shows the magnitude variations in the signal. Its small value signifies a small peak intensity difference [8]. Signal mobility is the first derivative that shows the variance of the signal gradient, while frequency variations are shown by complexity. The results of the CA Hjorth descriptor measurements shown in **Table 3** depicts that dry cough have the largest activity and mobility, while COVID-19 Cough has the highest complexity. The activity parameter of dry cough shows uniformity in time domain while COVID-19 Cough mobility value shows the most spread and variation. Due to the amplitude normalization phase of pre-processing, all reported parameter values are very small. Hjorth descriptor benefits from reduced cost computing and fast feature extraction as it has 3 parameters in comparison to other approaches. Thus, the time domain Hjorth descriptor values are smaller but much more consistent.

Table 3 - Hjorth descriptors for all Cough Auscultations.

Cough Auscultations	Activity (At)	Mobility (Mb)	Complexity (Cx)
COVID-19Cough	0.1820∓0.1366	0.3980∓0.1543	0.7147∓0.1269
Dry Cough	0.1838∓0.0907	0.4130∓0.1155	0.6766∓0.1386
Wet Cough	0.1122∓0.1257	0.2692∓0.2004	0.5479∓0.1509

V. Conclusions

Now, the COVID-19 virus is a terrifying label all over the globe. Several COVID-19 related experiments have been conducted, particularly the screening process to monitor the disease's transmission. Lung sounds and Cough auscultations can be used to test for lung abnormalities in patients. Here we have proposed an intelligent machine learning model for COVID-19 detection using cough auscultations collected from local Pakistani hospitals. We applied dc component removal and amplitude normalization for signal denoising then signal segmentation was performed to segment out low-frequency components (<2kHz). Hjorth descriptors are extracted and fed to classifiers such as Weighed-KNN and Fine decision tree resulting in cumulative accuracy of 99.8% and 99.4% respectively. Based on cough auscultation, our proposed approach will assist people in Asian countries in predicting the existence of COVID-19 in initial phases.

Acknowledgments

We would like to thank UET Taxila Dispensary for their cooperation during data collection.

References

[1] J. Monge-Alvarez, C. Hoyos-Barcelo, L. M. San-Jose-Revuelta, and P. Casaseca-De-La-Higuera, "A Machine Hearing System for Robust Cough Detection Based on a High-Level Representation of Band-Specific Audio Features," IEEE Trans. Biomed. Eng., vol. 66, no. 8, pp. 2319–2330, Aug. 2019, doi: 10.1109/TBME.2018.2888998.

[2] J. Monge-Alvarez, C. Hoyos-Barcelo, P. Lesso, and P. Casaseca-De-La-Higuera, "Robust Detection of Audio-Cough Events Using Local Hu Moments," IEEE J. Biomed. Heal. Informatics, vol. 23, no. 1, pp. 184–196, Jan. 2019, doi: 10.1109/JBHI.2018.2800741.

[3] K. Kosasih, U. R. Abeyratne, V. Swarnkar, and R. Triasih, "Wavelet Augmented Cough Analysis for Rapid Childhood Pneumonia Diagnosis," IEEE Trans. Biomed. Eng., vol. 62, no. 4, pp. 1185–1194, Apr. 2015, doi: 10.1109/TBME.2014.2381214.

[4] R. X. Adhi Pramono, S. Anas Imtiaz, and E. Rodriguez-Villegas, "Automatic Identification of Cough Events from Acoustic Signals," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Jul. 2019, pp. 217–220, doi: 10.1109/EMBC.2019.8856420.

[5] F. Barkani, H. Satori, and M. Hamidi, "Cough Detection System Based on ASR-HMM," Oct. 2020, doi: 10.1109/ICDS50568.2020.9268765.

[6] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," Phys. Eng. Sci. Med., vol. 43, no. 2, pp. 635–640, Jun. 2020, doi: 10.1007/s13246-020-00865-4.

[7] T. Zhou, H. Lu, Z. Yang, S. Qiu, B. Huo, and Y. Dong, "The ensemble deep learning model for novel COVID-19on CT images," Appl. Soft Comput., vol. 98, p. 106885, Jan. 2021.

[8] K. H. Abdulkareem et al., "Realizing an Effective COVID-19Diagnosis System Based on Machine Learning and IOT in Smart Hospital Environment," IEEE Internet Things J., 2021, doi: 10.1109/JIOT.2021.3050775.

[9] G. Gilanie et al., "Coronavirus (COVID-19) detection from chest radiology images using convolutional neural networks," Biomed. Signal Process. Control, vol. 66, p. 102490, Apr. 2021, doi: 10.1016/j.bspc.2021.102490.

[10] T. Tuncer, F. Ozyurt, S. Dogan, and A. Subasi, "A novel COVID-19and pneumonia classification method based on F-transform," Chemom. Intell. Lab. Syst., vol. 210, p. 104256, Mar. 2021, doi: 10.1016/j.chemolab.2021.104256.

[11] S. Ying et al., "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," medRxiv. medRxiv, p. 2020.02.23.20026930, Feb. 25, 2020, doi: 10.1101/2020.02.23.20026930.

[12] J. A. Meister, K. A. Nguyen, and Z. Luo, "Audio feature ranking for sound-based COVID-19patient detection," Apr. 2021, Accessed: May 01, 2021.

[13] T. Xia, J. Han, L. Qendro, T. Dang, and C. Mascolo, "Uncertainty-Aware COVID-19Detection from Imbalanced Sound Data," Apr. 2021, Accessed: May 01, 2021.