PULMONARY SOUNDS ANALYSIS FOR COMPARISON OF SMOKERS WITH HEALTHY NON-SMOKERS AND COPD PATIENTS

A THESIS SUBMITTED TO
THE INSTITUTE OF GRADUATE PROGRAMS
OF
İSTANBUL BİLGİ UNIVERSITY

AMROU WALI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

2024

PULMONARY SOUNDS ANALYSIS FOR COMPARISON OF SMOKERS WITH HEALTHY NON-SMOKERS AND COPD PATIENTS

Amrou Wali
122815005

Assist. Prof. Dr. İpek Şen
İstanbul Bilgi Üniversitesi

Prof. Dr. Z. Yasemin Kahya
Boğaziçi Üniversitesi

Assist. Prof. Dr. Aysa Jafari Farmand
İstanbul Bilgi Üniversitesi

Date of Approval: 6/10/2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Amrou Wali

Signature:

ABSTRACT

This study explores the classification of patients into three distinct categories—Chronic Obstructive Pulmonary Disease (COPD), smokers, and non-smokers—utilizing signal processing techniques and machine learning algorithms. By analyzing acoustic signals, forty-five features are extracted and evaluated for their efficacy in classification tasks. Notably, features such as AR1, AR2, AR3, and MFCC3 consistently demonstrate high accuracy across different classifiers, suggesting their importance in patient classification.

The research also investigates the impact of microphone placement on classification accuracy. Particularly, the fifth microphone emerges as significant, highlighting its influence on improving classification outcomes. These findings offer practical insights for healthcare practitioners, facilitating early interventions and personalized treatment plans. Moreover, they underscore the importance of considering microphone positioning strategies to enhance classification accuracy in clinical settings.

Looking ahead, future research could explore predictive modeling techniques to anticipate patient conditions over time. Additionally, optimizing microphone placement and refining algorithmic approaches could further improve classification accuracy. These endeavors hold promise for advancing healthcare practices, enabling more effective patient monitoring and intervention strategies.

In summary, this study establishes a foundation for enhancing patient classification methodologies, paving the way for improved healthcare delivery and patient outcomes. By leveraging signal processing and machine learning techniques, healthcare providers can make more informed decisions, leading to better patient care and management of respiratory conditions like COPD.

Keywords: COPD; Smokers; Non-smokers; Acoustic Features; Signal Interpretation

ÖZ

Bu çalışmada, sinyal işleme teknikleri ve makine öğrenimi algoritmaları kullanılarak hastaların üç farklı kategoriye (Kronik Obstrüktif Akciğer Hastalığı (KOAH), sigara içenler ve sigara içmeyenler) sınıflandırılması araştırılmaktadır. Akustik sinyaller analiz edilerek kırk beş özellik çıkarılmış ve sınıflandırma görevlerindeki etkinlikleri açısından değerlendirilmiştir. Özellikle AR1, AR2, AR3 ve MFCC3 gibi özellikler, farklı sınıflandırıcılar arasında tutarlı bir şekilde yüksek doğruluk göstererek hasta sınıflandırmasındaki önemlerini ortaya koymaktadır. Araştırma ayrıca mikrofon yerleşiminin sınıflandırma doğruluğu üzerindeki etkisini de incelemektedir. Özellikle beşinci mikrofon, sınıflandırma sonuçlarını iyileştirme üzerindeki etkisini göstererek önemli bir unsur olarak ortaya çıkmaktadır. Bu bulgular, sağlık hizmeti uygulayıcıları için erken müdahaleleri ve kişiselleştirilmiş tedavi planlarını kolaylaştıran pratik bilgiler sunmaktadır. Ayrıca, klinik ortamlarda sınıflandırma doğruluğunu artırmak için mikrofon konumlandırma stratejilerini dikkate almanın önemini vurgulamaktadır. İleriye dönük olarak, gelecekteki araştırmalar, zaman içinde hasta koşullarını tahmin etmek için öngörücü modelleme tekniklerini keşfedebilir. Ayrıca, mikrofon yerleşiminin optimize edilmesi ve algoritmik yaklaşımların iyileştirilmesi sınıflandırma doğruluğunu daha da artırabilir. Bu çabalar, daha etkili hasta izleme ve müdahale stratejileri sağlayarak sağlık uygulamalarını ilerletmek için umut vaat etmektedir. Özetle, bu çalışma hasta sınıflandırma metodolojilerini geliştirmek için bir temel oluşturarak sağlık hizmeti sunumunun ve hasta sonuçlarının iyileştirilmesinin önünü açmaktadır. Sinyal işleme ve makine öğrenimi tekniklerinden yararlanarak, sağlık hizmeti sağlayıcıları daha bilinçli kararlar verebilir, bu da daha iyi hasta bakımı ve KOAH gibi solunum rahatsızlıklarının yönetimine yol açabilir

Anahtar Kelimeler: KOAH; Sigara İçenler; Sigara İçmeyenler; Akustik Özelikler; Sinyal Yorumlama

I would like to thank my mother Rasha for always praying for me and giving me the emotional stability that I need, my father Ahmed Azzam for pushing me to the limit to be a man, my brother Mohammed for being the person to watch out for me and hold me strong, and my sister Lana for being the crazy boost of energy that was there whenever I would need it.

My best friend's Leen Fattah, Basel Zbedeh, Abdullah Hamwi, and the whole squad who were my family while living in Turkey where we have picked each other up and supported each other in our difficulties all throughout the years and had the best of times together.

I would also like to dedicate this thesis to myself. Go out there and chase your dreams. If you put your mind to it, no one can stop you.

ACKNOWLEDGMENTS

I would like to thank my amazing advisor Assoc. Prof. Ipek Şen, for her kindness and endless support throughout the past few years. Her mentorship and guidance were always a boost of motivation to keep pushing through and working hard whenever it got hard. To you I am always thankful when you understood when I was feeling down or unmotivated and you helped pick me back up. A student under your guidance and mentorship is a lucky student.

I would also like to thank my dear friend Serhat Ismet Sargin, who I have had a chance to be friends with during the years of our graduate studies and learned a lot from him. You were an amazing friend and helped me so much throughout my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WHO:	World Health Organization
COPD:	Chronic Obstructive Pulmonary Disease
PSD:	Power Spectral Density
NB:	Naive Bayes
DT:	Decision Tree
k-NN:	k-nearest Neighbor
ML:	Machine Learning
DL:	Deep Learning
NS:	Non-smokers
S:	Smokers
ACC:	Accuracy
TP:	True Positive
TN:	True Negative
FP:	False Positive
FN:	False Negative

INTRODUCTION

Smoking is a widespread problem in the twenty-first century. People are starting to smoke at a younger age because of the rising demand for cigarettes [1]. This is resulting in serious diseases and health issues. COPD could be one of them. COPD, or chronic obstructive pulmonary disease, is a disease or a group of disorders that can impede airflow and cause breathing issues [2]. Their condition will be determined by the recordings, which will make their assessment easier. This study aims to give awareness and insights on the effects of smoking on people and how it can lead to COPD. This project hopes to one day have a device implemented in clinics that would help show how fast a person is developing COPD acoustic features the more he smokes and how this would lead to him to get the disease. If we can create an early alarm to show the correlation and find an explanation of how lungs deteriorate through lung sounds using digital signal processing and to find how they deviate from being healthy and slowly progress over time towards the COPD group of people in the feature space and how it would help alarm the patient more. For example, a patient would come into the clinic once a year and would give statistically how many cigarettes he smokes in a day, then some implementation can be done to prove that over time, with this number of cigarettes, this is how fast a certain subject can develop the disease. According to WHO [3] (World Health Organization), COPD is the third leading cause of death worldwide. It caused over 3.23 million deaths in 2023 and smoking is the main risk factor for COPD [4].

In the literature people did studies to show similarity between lung sounds which will be discussed. An earlier study [5] had a purpose of investigating the differences between computerized respiratory sounds of healthy smokers and non-smokers. The study was done on forty-four participants, eighteen of which were smokers, and the rest were nonsmokers. In that study, in order to show the difference between them, the following features were used; maximum intensity(Imax), frequency at max intensity(Fmax), mean intensity over the whole frequency range and mean intensities over the following frequency intervals: 100-300, 300-600, 600-1200, and 1200-2000Hz. Their results were the following. For Fmax, the main difference in the sound that is taken from the posterior of the chest is in the inspiration phase. However, for Imax it is in expiration phase. For mean intensity, the difference between smokers and nonsmokers was observed at the expiration phase and the sound that is taken from the posterior side.

In another study [6] the purpose was to find a correlation between smokers and non-smokers while looking at age and gender as well. 162 subjects were evaluated and any patient with a lung disease was excluded. They found out that there is significant difference between man and women, but not in smokers versus nonsmokers.

Another study [7] is unclear whether crackle characteristics can be used to identify COPD. Twenty-six people were used for this study, 9 healthy non-smokers, 9 healthy smokers and 8 COPD subjects. Both crackle 2-cycle duration and crackle number showed some significant correlation with airway parameters at some branch generations, but due to the large number of correlations performed, these were consistent with chance findings. This means that there were no statistically significant differences in crackle characteristics, which is the number of crackles per breathing cycle and crackle 2-cycle duration between the three subject groups.

In another study [8], the authors developed a classifier to distinguish between smokers and non-smokers by using k-NN classifier, 30 subjects were used, 15 of which were smokers and 15 non-smokers. In that study, they state that the accuracy for non-smokers was 89.33% and 78.76 for smokers.

Most studies mentioned above talked about the statistical differences, however none implemented a classifier to find three different classes of subjects. In the study mentioned earlier [8] they implemented only a k-NN classifier, and they did not actually give out their overall accuracy scores. As far as the earlier studies mentioned, no study tried to classify three classes of subjects and implement different classifiers to find the best one out of them.

In [9], a preliminary study investigates the impact of smoking on pulmonary acoustic features, with a specific focus on distinguishing between smokers, non-smokers, and individuals with chronic obstructive pulmonary disease (COPD). Three novel measures are introduced: the first two aim to quantify the differences between the groups and rank the features based on their ability to discriminate, while the third measure assesses the linear shift of smokers away from non-smokers towards COPD within a multivariate feature space. The key features identified include the percentile frequencies f_{25} and f_{75}, along with the eighth Mel-frequency cepstral coefficient (MFCC8), which were consistently selected by the first two measures. Findings indicate that acoustic features captured during early-inspiration, early-expiration, and from the seventh microphone location are particularly effective in detecting early signs of COPD. The study emphasizes the necessity for larger datasets, the development of additional measures, further testing, and clinical interpretation to establish a reliable methodology for early COPD detection.

With the help of this, the aim was to find some ground of control in the feature space by finding that the classifier can differentiate between the 3 classes, therefore a next project step is to find that correlation and distance in the feature space to successfully show closeness between COPD and smokers with respect to the non-smoker class.

DATA

The database referred to in this study has been acquired using the 14-channel pulmonary sound data acquisition and processing system that was designed and implemented in the Boğaziçi University Lung Acoustics Laboratory [10,11]. The data acquisition procedure of this study has been approved by the Second Ethical Committee on Clinical Research of Istanbul which follows the Declaration of Helsinki.

The data consisted of 44 total subjects. Twenty of which were examined and identified as COPD patients, 14 healthy non-smokers, and 10 healthy smokers. Key information such as age, gender, height, weight, cigarette packs smoked per year, and years since being a smoker was noted and can be observed in Table 2.1.

Table 2.1. Data Information

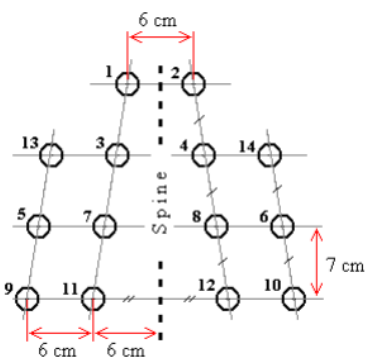
	# of Subjects	Gender	Height	Weight	Packs / Year	Age	Years since smoking
COPD	20	4 F; 16 M	1.67 ± 0.058	69.8 ± 13.029	407 ± 349.9	61 ± 7.65	37.47 ± 12.85
Smokers	10	3 F; 7 M	1.75 ± 0.076	68.86 ± 8.227	103.1 ± 48.33	26.3 ± 3.74	8.00 ± 6.27
Non-Smokers	14	4 F; 10 M	1.75 ± 0.058	75.95 ± 12.944	0	26.2 ± 3.40	0

The COPD recordings were taken from a hospital in Istanbul called Yedikule Hastanesi while the smokers and non-smokers recordings were taken from volunteers in Boğaziçi University including students and personnel. The system that recorded the data consists of 14 air-coupled electret microphones (SONY ECM-44 BPT) attached to the posterior chest wall, an analog amplifier-filter unit with a gain of 100 and a pass-band of 80 to 4000 Hz, a Fleisch type pneumotachograph (Validyne CD379) for simultaneous flow-cycle measurement and synchronization, a data acquisition card (NI DAQCard-6024E, 12-bit) for digitization, and a laptop computer that controls the process via a LabVIEW interface program and stores the data[10]. The data is sampled at a rate of 9600 samples per second over a 15-second acquisition session. During recording sessions, subjects sat upright, wore a nose clip, and breathed through the flowmeter's mouthpiece.

This thesis aims to use these recordings and help develop a classifier that can successfully classify 3 different classes of subjects and demonstrate that the classifier can successfully find the difference between them by extracting certain acoustic features. The aim is to help implement it into clinics that can help the health industry inform certain patients that they are now classified as a COPD patient. This is both time and cost efficient.

The dataset consisted of fifteen-second pulmonary sound recordings acquired from lung sounds off the subject’s posterior chest wall. The microphones were placed strategically to acquire sensitive and clear recordings without interference from outside noises by using a capsule on top of the microphone to isolate the outside noises and to provide the air coupling needed. The subject was then asked to breathe in and out over this fifteen second cycle. The fourteen different microphones/channels were placed on the back of the patient as shown in Figure 2.1:

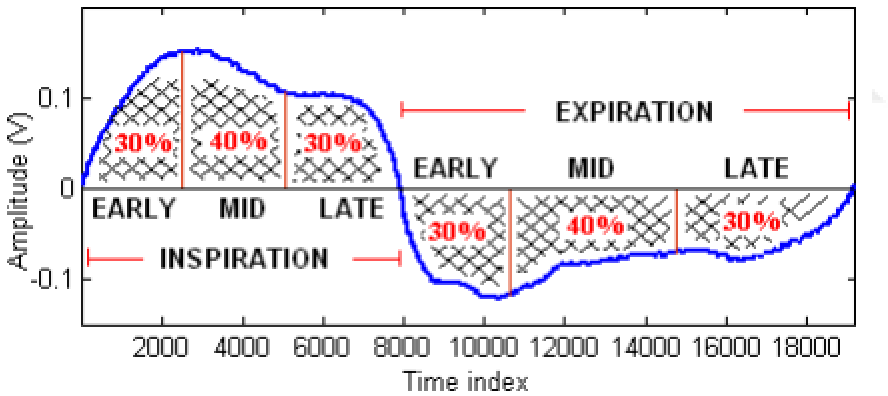
Figure 2.1. Microphone Placement



(Source: *Sen, I., Saraçlar, M. & Kahya, Y.P. (2021). Differential diagnosis of asthma and COPD based on multivariate pulmonary sounds analysis*)

Through fourteen microphones/channels, this study also divided the recording over six different sub-phases using the flow signal [11] these sub-phases were labeled as Early Inspiration, Mid Inspiration, Late Inspiration, Early Expiration, Mid Expiration, and Late Expiration. Early and late phases represent 30 percent each of the total air volume inhaled or exhaled during one period, while the mid phase represents the remaining 40 percent. Inspiration is marked by the positive portions of the flow rate signal, and expiration by the negative portions. Figure 2.2. represents a visual of it. To minimize noise-related errors, a dead band of one percent of the maximum amplitude deviation is set around the zero level. The start and end indices of the sub-phases are calculated using the area under each inspiration and expiration curve. Once determined, these indices are used to label the 14-channel pulmonary sound data for each subject [11].

Figure 2.2. Division of The Flow Cycle into Early/Mid/Late Inspiration/Expiration



(Source*: I. Şen, "A Multi-Channel Device for Respiratory Sound Data Acquisition and Transient Detection," M.S. Thesis, Boğaziçi University, 2005* [11])

METHODOLOGY

In this section, the features and classifiers with the performance parameters used in this project will be described.

3.1. Features

The fifteen second recording of every person was segmented into flow cycles first, then into six sub-phases and the features were calculated for each single sub-phase. Through this segmentation, full cycles would be observed. A full cycle is when the subject inhales and exhales fully during that period of segmentation. The study looks at full cycles and not missing cycles. Missing cycles are when the subject breathes in and out, however, it is not included in the fifteen second recordings. E.g., when the subject starts breathing before the recording starts or when the subject does not finish the expiration cycle and the 15 seconds are up. Because every subject has a different breathing duration, subjects have varying number of breathing cycles for a fifteen second recording. The next step was how these features were calculated and obtained to help with our purpose of classifying these subjects into three classes.

Instead of feeding the recorded data into the classifier, acoustic features were extracted from this data. By selecting features that are specifically tailored to distinguish between non-smokers, smokers, and COPD patients, the classifier can more effectively learn the patterns and differences between these groups. This targeted approach leads to better accuracy and reliability compared to using raw data. In Table 3.1., all features used in this study were listed.

Table 3.1. Features List

'F25'	'FMAX'	'MFCC1'	'MFCC12'
'F50'	'EBIN1'	'MFCC2'	'MFCC13'
'F75'	'EBIN2'	'MFCC3'	'MFCC14'
'F90'	'EBIN3'	'MFCC4'	'MFCC15'
'F95'	'EBIN4'	'MFCC5'	'MFCC16'
'AR1'	'EBIN5'	'MFCC6'	'MFCC17'
'AR2'	'EBIN6'	'MFCC7'	'MFCC18'
'AR3'	'EBIN7'	'MFCC8'	'MFCC19'
'AR4'	'STDDEV'	'MFCC9'	'MFCC20'
'AR5'	'KURTOSIS'	'MFCC10'	'MFCC21'
‘AR6’	'SKEWNESS'	'MFCC11'	
'PMAX'	'ENTROPY'		

3.1.1. Percentile Frequencies (Fx)

Fx is the frequency where x percent of the total power is reached, in the power spectral density of the signal. Percentile frequencies are used to examine how power is distributed across a signal's frequency spectrum. By dividing the spectrum into percentiles, one can identify the frequency at which a specific percentage of the total power is contained or exceeded. This technique can be useful for differentiating among various signal characteristics. Given a power spectral density (PSD) estimate, the 95th percentile frequency, for instance, represents the frequency below which 95% of the total power resides [12].

In this study, Welch’s method with 250-point Hamming windows and a 1024-point Fast Fourier Transform (FFT) has been applied; by discerning the frequencies at which significant portions of the signal's power reside, percentile frequencies enable the identification of spectral patterns unique to smokers, non-smokers, and COPD patients. Welch's method involves dividing a signal into overlapping segments, applying a windowing function (like Hamming), and then computing the FFT of each segment. The

PSD is obtained by averaging the squared magnitudes of these FFT results. The Welch PSD estimate for segment k is:

The percentile frequencies can then be derived by integrating the PSD to find the frequencies at which specific percentiles of power are accumulated.

$$P_{xx,k}(f) = \frac{1}{L}\sum\nolimits_{K\to k}^{L-1}\left|\frac{1}{M}\sum_{n=0}^{N-1} x_k(n)\cdot w(n)\cdot e^{-\frac{j2\pi fn}{M}}\right|^2 \tag{3.1}$$

Where N is the length of the signal segment, M is the FFT length (typically larger than N), L is the Number of overlapping segments, and $w(n)$ is the window function applied to each segment.

3.1.2. Mel-frequency Cepstral Coefficients (MFCCs)

MFCCs are commonly used in speech and sound analysis because they mimic the human auditory system's nonlinear response to sound [6,12,13]. To compute MFCCs, you typically convert a signal's spectrum to the Mel scale (which emphasizes lower frequencies) and then apply the discrete cosine transform (DCT) to extract the coefficients. The key steps for MFCC calculation are first calculation of FFT by obtaining the frequency spectrum of a signal. Then to apply the Mel filter bank by applying a set of triangular filters spaced according to the Mel scale to the spectrum and then taking the logarithm of the filtered outputs and finally applying DCT to the logarithmic outputs to get the MFCCs.

If there are K filters in the Mel filter bank and N FFT bins, the MFCC formula can be represented as

$$c_m = \sum\nolimits_{k=0}^{K-1}\log[\![S_k\cdot cos\left(\pi m\ \cdot\frac{k+0.5}{K}]\!]\right) \tag{3.2}$$

Where S_k is the filtered output from the Mel filters, and πm represents the coefficient index. In this study, k has been chosen to be 21, i.e. MFCC features used were from MFCC 1 to MFCC 21 [13].

3.1.3. The Auto-Regressive (AR) model

The AR model is a fundamental concept in time-series analysis and signal processing. It is used to describe a system where each observation in a time series is regressed against a certain number of prior observations, along with a stochastic error term. AR models are often applied to data that exhibit temporal dependencies, making them valuable for analyzing signals that may contain underlying patterns or dynamics over time. An Auto-regressive model of order p, denoted as AR(p), is a linear model where each value in a time series depends on a combination of its previous p values and a noise term. Mathematically, it is defined as:

$$x(n) = \sum_{i=1}^{p} a_i \cdot x(n-i) + e(n) \tag{3.3}$$

Where $x(n)$ is the current value of the time series, a_i represents the AR model coefficients, $e(n)$ is the noise or error term, typically assumed to be white noise with zero mean and constant variance, and p is the order of the AR model, indicating how many prior values are used to predict the current value.

The AR model is useful for capturing patterns and dependencies in time-series data. The order p determines the complexity of the model. A higher order means more previous values are used, which can capture more complex patterns, however, it can also lead to overfitting. There are different methods to estimate the coefficients in the literature. Yule Yule-Walker Equations solves a system of linear equations derived from the autocorrelation function. While Burg's Method minimizes forward and backward prediction errors, providing a stable solution. Another method is the Least Squares method which fits the AR model by minimizing the squared error between the observed and predicted values [14]. The method used in this thesis was the Burg’s Method.

The coefficients derived from an AR model can provide insights into the temporal dynamics of a signal, useful for analyzing complex signals like pulmonary sounds.

In this study, a sixth order AR model has been adapted. i.e. the AR Coefficients extracted were from AR1 to AR6.

3.1.4. Maximum Power (P_{max}) and Frequency of Maximum Power (f_{max})

Maximum power in the PSD refers to the highest value in the power spectral density, while the frequency of maximum power is the frequency at which this maximum occurs. Given a PSD estimate, you can find these by locating the index with the maximum power [15].

The maximum power is given by $P_{max} = max_f P_{xx}(f)$and the Frequency of maximum power is $f_{max} = \arg\max_f P_{xx}(f)$

The statistical descriptors provide insights into the signal's statistical properties where entropy measures the uncertainty or disorder in a signal. Shannon's entropy for a signal x with probability distribution $p(x)$ is given by:

$$H = -\sum_{i=1}^{N} p(x_i) \cdot \log_2\left(p(x_i)\right) \quad (3.4)$$

Where $p(x_i)$ represents the probability of each outcome x_i , and N is the number of outcomes.

The logarithm is typically taken to base two, representing information in bits. The key to understanding entropy is recognizing that it quantifies the average uncertainty or information in a system: higher entropy means more uncertainty, while lower entropy means less uncertainty or more predictability. In this study, a histogram based on the sound amplitude values has been used as an estimate of $p(x_i)$, as in [16], entropy has been used as a measure of uncertainty observed in the sound amplitude, hence, of the irregularity of the waveform.

3.1.5. Statistical Features

Three statistical features were used in this study.

Standard Deviation: is a measurement of how much a set of data can vary. A low standard deviation means the values are close to the mean, whereas a large standard deviation means the data are spread out over a greater range. Standard Deviation X: Each Value $\bar{X}$: Mean n: Total size.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad (3.5)$$

Kurtosis: quantifies the “peakedness” of the signal's distribution. Kurtosis’ formula is shown below:

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^4 \quad (3.6)$$

Skewness: was also used in this study, and it is the measure of the asymmetry of the distribution:

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3 \quad (3.7)$$

These statistical measures, combined with spectral and temporal features, can give a comprehensive view of the signal's characteristics.

3.1.6. Calculated Energy Values for Different Frequency Bands (BINx)

Energy values were calculated for each band after dividing the power spectral density (PSD) into distinct frequency bands. In signal processing, "Energy of Bins" often refers to the energy (or power) contained within specific frequency bins when analyzing signals in the frequency domain. This term is frequently used in the context of Fourier Transform analysis, particularly with the Fast Fourier Transform (FFT), where a signal is decomposed into its frequency components. In an FFT, a signal is transformed from the time domain to the frequency domain, breaking it into discrete frequency components. These components are typically organized into "bins," with each bin corresponding to a specific range of frequencies.

The energy (or power) of a frequency bin represents the squared magnitude of the FFT output for that bin. It provides a measure of how much energy is contained in that specific frequency range, reflecting the signal's spectral content. .100 Hz bands were targeted; however, it was more convenient to use bands of ten points each, corresponding to approximately 94 Hz (93.75 Hz). Note that the range 0-4800 Hz (Nyquist theory, 9600/2=4800 Hz) is expressed by 512 points (number of FFT points, 1024/2=512), therefore each point represents a 4800/512=9.375 Hz interval and 10 points cover 93.75 Hz. Thereby, a total of fifty-one bands were obtained (0-93.75 Hz, 93.75-187.5 Hz, 187.5-281.25 Hz, etc.), however only the seven ten bands were used.

The same PSD used in Pmax and Fmax calculations was segmented onto the frequency axis into fifty. Mathematically if $X(k)$ represents the fourier transform of a signal and N is the total number of bins, the energy of bins k is given by:

$$E(k) = |X(k)|^2 \tag{3.8}$$

Where $E(k)$ is the energy of the kth bin and $|X(k)|$ is the magnitude of the FFT output for the kth bin [17].

In this study, the energy of bins was calculated from the first bin to the seventh bin.

3.2. Feature Normalization

Z-score normalization, also called standardization, is a key step in preparing data for machine learning and digital signal processing (DSP). This technique adjusts your data so that it has a mean of zero and a standard deviation of one. Data features can come in all sorts of scales and units. Imagine one feature ranges from 1 to 10, while another ranges from 1,000 to 10,000. Without normalization, the feature with the larger range can end up having too much influence on the model, skewing the results. In DSP, signals often come with noise and distortions. Standardizing these signals makes the processing algorithms more robust by normalizing the variations and reducing the impact of outliers. First, you calculate the mean and standard deviation for each feature using the training data. Then, you apply the normalization formula:

$$Z = \frac{X - \mu}{\sigma} \tag{3.9}$$

X is the original feature value, and μ is the sample mean, and σ is the sample standard deviation, of the features, respectively.

By standardizing features, it aims to make sure that each feature contributes equally to the model. This is especially useful for algorithms that rely on distance calculations, like k-nearest neighbors where equal feature contribution ensures fair distance metrics. Moreover, standardized data helps the optimization process run smoother and faster, avoiding issues where one feature might dominate the learning process [18].

3.3. Feature Classification

Three different classifiers were used in this study, considering their diverse natures. In this study, machine learning was used over deep learning because traditional machine learning algorithms, such as decision trees, naive bayes, and k-nearest neighbors, often require less computational power and memory compared to deep learning models. Machine learning is more favorable over deep learning especially when the database is small. E.g. In applications where data collection is tough, since deep learning usually requires a bigger dataset. Also, deep learning requires GPUs, and it is costlier and would take a bigger size for a device to be implemented. Classical machine learning can compete with deep learning, Machine learning in our case makes it more suitable for applications with limited hardware resources, such as embedded systems or mobile devices, especially for our case, trying to implement a small and reliable device that can quickly detect and notify the patient on the spot.

3.3.1. k-Nearest Neighbors (k-NN) Classifier

The k-Nearest Neighbors (k-NN) classifier is a machine learning algorithm used for classification and regression tasks. The central idea of k-NN is to classify new data points based on the "neighborhood" formed by the closest data points in a feature space. In this study, I used the k-NN classifier with Euclidean distance as the distance metric to perform classification. Euclidean distance was used due to having 6-dimensional features. On higher dimensional data, calculation of covariance matrix can cause a problem. This is why Euclidean distance was preferred over Mahalanobis distance. The Euclidean distance is widely used due to its simplicity and geometric interpretation. Well, given a dataset with known class labels, the k-NN algorithm classifies a new data point by finding the k closest points (nearest neighbors) within the training dataset. The class of the new point is then determined by a majority vote among these neighbors. The choice of k plays a crucial role in the algorithm's performance, with smaller values of k leading to more localized decisions and larger values providing smoother classifications. The optimal value of k was determined through different selections and different runs that lead to the number being specified from 1 to 50 but only the odd numbers. To classify new data points, a majority vote among the k nearest neighbors to find the class with most of the members to predict the label was implemented [19].

The distance metric is a key component of the k-NN classifier. In this project, the Euclidean distance was used. It is a common and intuitive metric that calculates the straight-line distance between two points in a multi-dimensional space. Given two points $x_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $x_2 = (x_{21}, x_{22}, \dots, x_{2n})$, the Euclidean distance between is defined as:

$$d(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2} \tag{3.10}$$

3.3.2. Gaussian Naive Bayes Classifier

Gaussian Naive Bayes is a variant of the Naive Bayes classifier used for classification tasks, especially when the features are continuous and assumed to follow a Gaussian (normal) distribution. It is based on Bayes' Theorem, incorporating the "naive" assumption that each feature is conditionally independent given the class. This simplification allows Gaussian Naive Bayes to be computationally efficient while providing robust classification performance in many cases.

Bayes' Theorem is a foundational principle in probability theory, providing a way to update the probability of a hypothesis given new evidence. It is defined as:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{3.11}$$

Where $P(C|X)$ is the posterior probability of class C given the features x. $P(X|C)$Is the likelihood of observing the features x given the class C. $P(C)$ Is the prior probability of class C, and $P(X)$ Is the marginal probability of the features, also called as evidence.

The key assumption in Gaussian Naive Bayes is that each feature follows a Gaussian distribution within each class. This assumption allows the calculation of the likelihood as:

$$P(x|C_i) = \frac{1}{\sqrt{2\pi\sigma_i{}^2}} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma_i{}^2}\right. \tag{3.12}$$

where x is the feature value and μ_i and σ_i are the mean and standard deviation of the feature for class C_i, respectively.

To classify a new data point, the Gaussian Naive Bayes classifier computes the posterior probability for each class using Bayes' Theorem and selects the class with the highest probability. The prior probabilities $P(C_i)$ are class proportions in the population, and the means and standard deviations of the class likelihoods are sample averages and sample standard deviations calculated using the class samples. In this thesis, the method used Gaussian Naive Bayes as follows: Given a new data point, the posterior probabilities were calculated for each class using the estimated parameters and the class with the highest probability was chosen [20].

3.3.3. Decision Tree Classifier

Decision trees are a popular type of supervised learning algorithm used for classification and regression tasks. They create a model that predicts the target variable by learning simple decision rules inferred from the training data. A decision tree consists of nodes that represent decisions, branches that represent outcomes, and leaves that represent class labels or numerical values.

The core concept of a decision tree is to split the data (into branches) until the required acceptable success score is reached whether regression or classification into subsets based on the most crucial features at each step. The process continues recursively until a stopping criterion is met, resulting in a tree structure with decision nodes and terminal leaves. The decision nodes represent points where a decision is made based on a feature and a threshold value. The condition typically involves comparing a feature's value to a threshold. The branches connect nodes and indicate the direction of the decision. For binary trees, branches represent two outcomes (e.g., "true" or "false"), and the leaves are the terminal nodes that represent the output class or value. In classification trees, a leaf contains a class label, while in regression trees, it contains a numerical value [21].

A key aspect of building a decision tree is choosing the optimal feature and threshold for each split. One of the common splitting criteria includes the Gini Impurity splitting method which measures the impurity in a dataset. The goal is to minimize impurity at each split. The Gini impurity for a set S with k classes is:

$$G(S) = 1 - \sum_{i=1}^{k} \left(\frac{|S_i|}{|S|} \right)^2 \tag{3.13}$$

where S_i represents the number of elements in class i, and S represents total number of elements, being referenced to a particular node that had been reached.

In this thesis, all depths from 1 to 25 have been executed and compared to find the most ideal depth value.

CLASSIFIER DESIGN AND PERFORMANCE METRICS

The data set was split into an 80/20 split. The cycles of each subject were labeled from 1 to 3 representing each class and were fed into the classifier as training sets. The split would go over the classifier once and the confusion matrix with the performance parameters was recorded. To find better results and find how effective and dependable was the classifier, the algorithm would run this test five times and then take the average of the success scores while keeping all the data for each run-in case needed for more observations. Three success metrics were used in this study.

Accuracy was calculated by looking at the overall success of the classifier in classifying the correct label by looking at the correctly classified label/total label number.

The F1 score is a common metric used to evaluate the performance of a classification model, especially in scenarios where data is imbalanced. For a three-class classifier, calculating the F1 score involves determining the F1 score for each class individually and then aggregating these results into an overall metric. The F1 score is a performance metric that combines precision and recall evaluating the effectiveness of a classifier. It is particularly useful when dealing with imbalanced datasets, where a single measure such as accuracy may not adequately capture the model's performance.

Precision measures the proportion of true positive predictions among all positive predictions. For a given class C, precision is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4.1}$$

where TP is the number of true positives (correctly predicted instances of class C) and FP is the number of false positives (incorrectly predicted instances of class C).

Recall (also known as sensitivity) measures the proportion of true positive predictions among all actual positives. It is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4.2}$$

where FN is the number of false negatives (missed instances of class C).

F1 Score is the harmonic mean of precision and recall, providing a balance between them. It is defined as:

$$F1 = 2x \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.3}$$

To calculate F1 score for a Three-Class Classifier, one would need to find the TP (True Positive; cell 1), TN (True Negative; cell 5 + cell 6 + cell 8 + cell 9), FP (False Positive; cell 4 + cell 7), and FN (False Negative; cell 2 + cell 3).

In a 3x3 matrix. The true positive in our case was the COPD class so any false negative value for class will be the sum over the corresponding row except for the TP value. The false positive value for a class will be the sum over the related column except for the true positive value. The true negative value of a class will be the sum of the values across all columns and rows except for the values associated with the class. True positive occurs when the predicted value and actual value are the same. [22].

Table 4.1. Cell Numbering

	Predicted C	Predicted NS	Predicted S
Actual C	Cell 1	Cell 2	Cell 3
Actual Ns	Cell 4	Cell 5	Cell 6
Actual S	Cell 7	Cell 8	Cell 9

These are going to be useful for the purpose of finding another success metric summarized in the next paragraph.

The third metric score is called μ. It is a success metric used to summarize the confusion matrix for comparison in a neutral and symmetric way (μ corresponds to μ_3 described in the study [23], it is basically the harmonic mean of the class correct classification rates. That is

$$\mu = \sqrt[3]{CCR_{COPD} \times CCR_S \times CCR_{NS}} \tag{4.4}$$

Where CCR_{COPD}, CCR_S , CCR_{NS} , are the correct classification rates of COPD, healthy smokers, healthy non-smokers classes, respectively.

EXPERIMENTAL RESULTS

As previously detailed, forty-five features were extracted from the dataset following the methodology outlined above. These features were used to classify the data into three distinct classes using three different classifiers. Each classifier was evaluated based on the performance of these features, optimizing the hyper-parameters such as k in the k-nearest neighbors (k-NN) classifier and depth in the decision tree.

The first step involved running all three classifiers on the dataset with the forty-five features to observe and record the maximum accuracy scores achieved by each classifier. After careful analysis, it was found that the k-NN classifier achieved the highest accuracy among the three classifiers. The features that gained the highest accuracy were identified as AR1, MFCC3, AR2, AR3, and EBIN7 sorted with respect to their accuracies in descending order. The highest accuracy recorded by the k-NN classifier was an impressive 86.3% for AR1. The accuracy scores for each feature were then plotted, showing the features arranged from the highest to the lowest accuracy as derived from the k-NN classifier. The plot can be observed in Figure 5.1. As for the best k-NN value, a bar plot is listed in Figure 5.2. that shows that the 13th neighbor showed 5 different counts across the 45 features proposing the idea that it is the best neighbor value.

Next, the performance of the Decision Tree (DT) classifier was examined. The features were sorted based on the accuracy derived from the DT classifier, revealing that the highest accuracies were associated with AR1, AR2, MFCC3, AR3, and AR5. The highest accuracy achieved by the DT classifier was 82.9%, with AR1 being the top feature. This sorting and accuracy data were also plotted to provide a visual representation. The plot can be observed in Figure 5.3.. As for the depth values, Figure 5.4. shows the depth value corresponding to the maximum accuracy derived from each feature.

Figure 5.1. Accuracy Score with k-NN Indexing

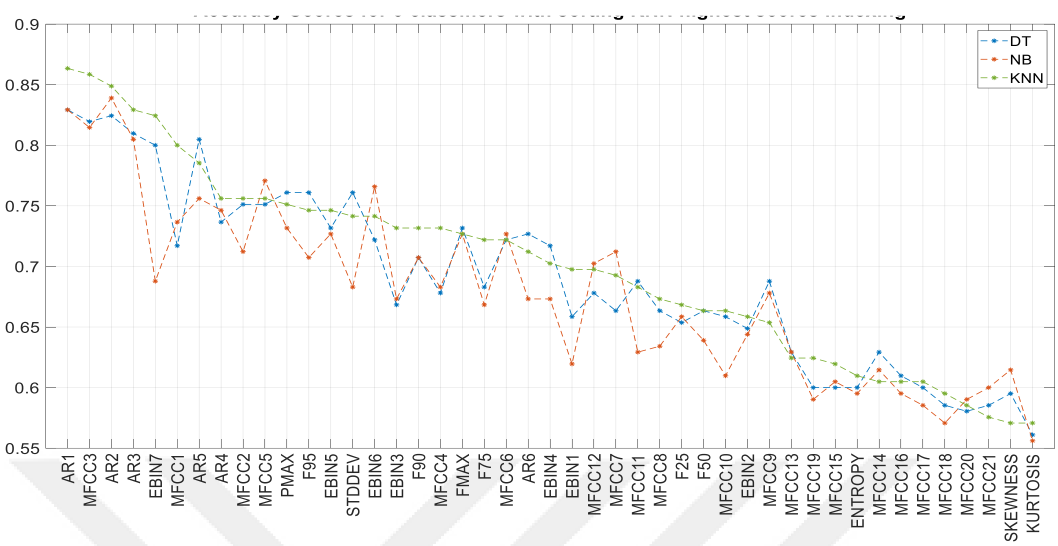


Figure 5.2. k-NN Value for Each Feature

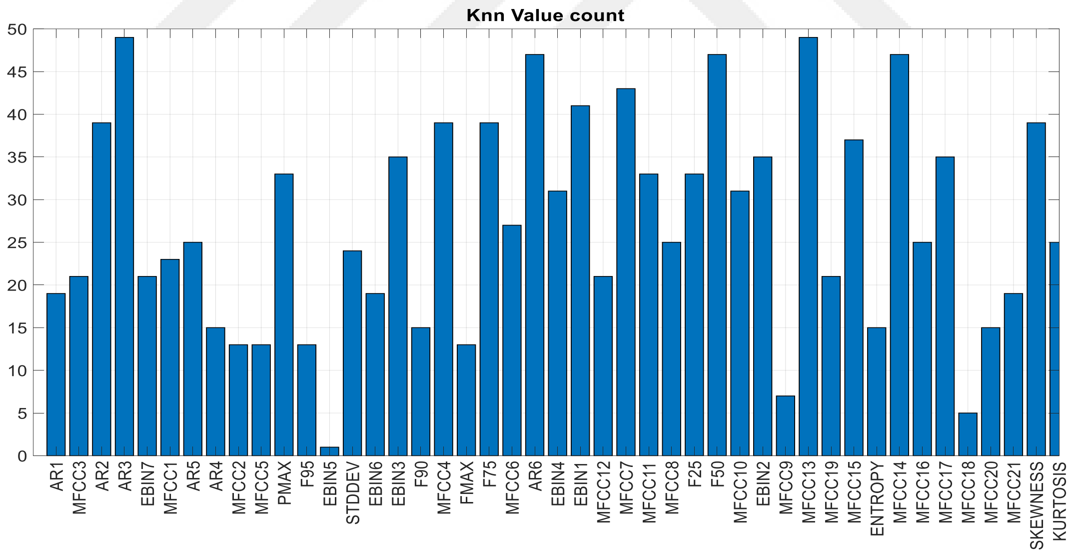


Figure 5.3. Accuracy Score with DT Indexing

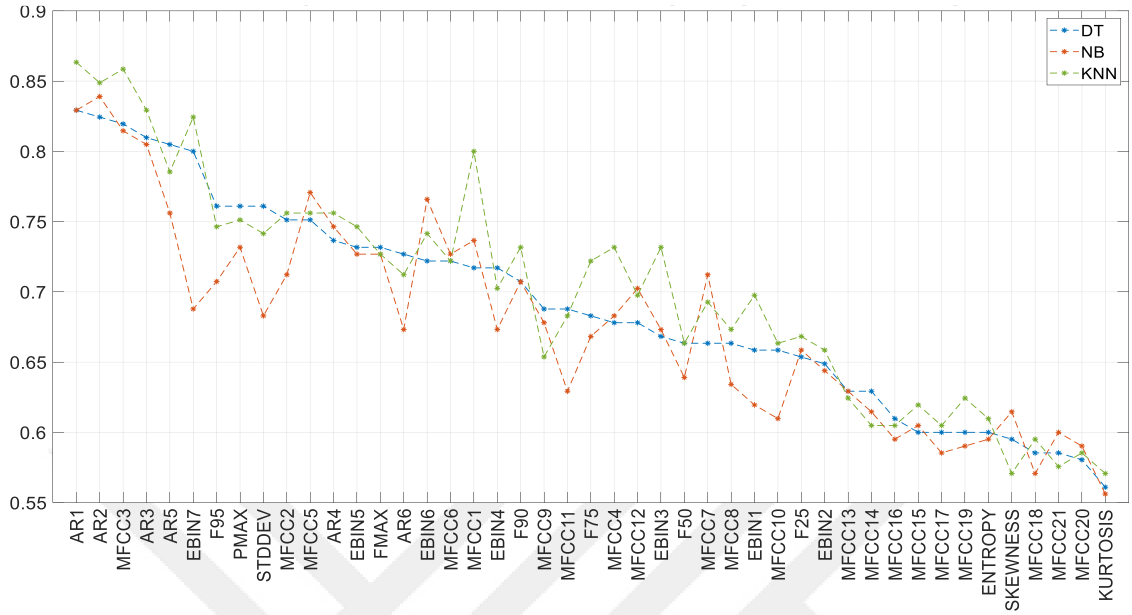
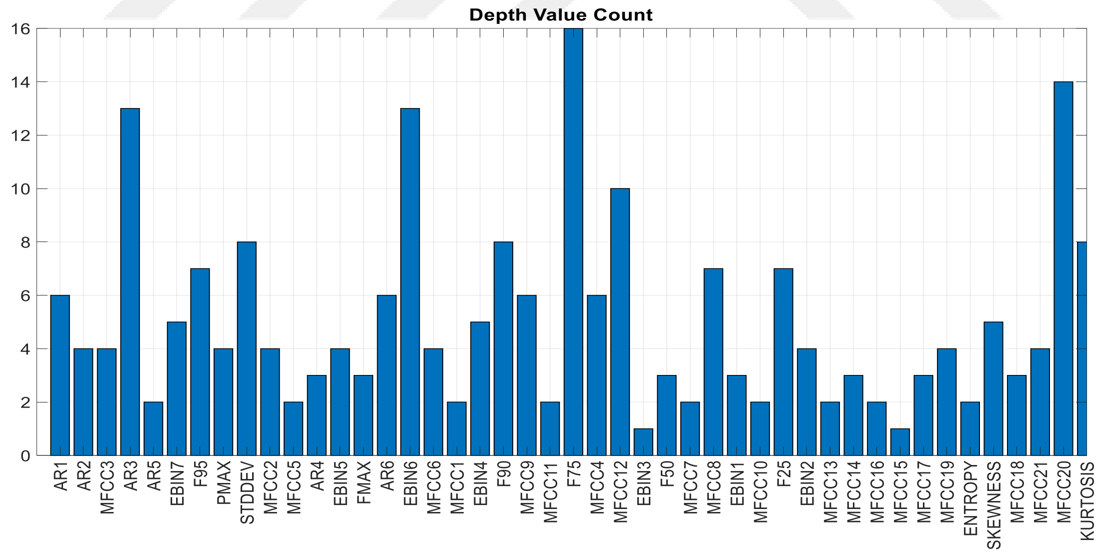


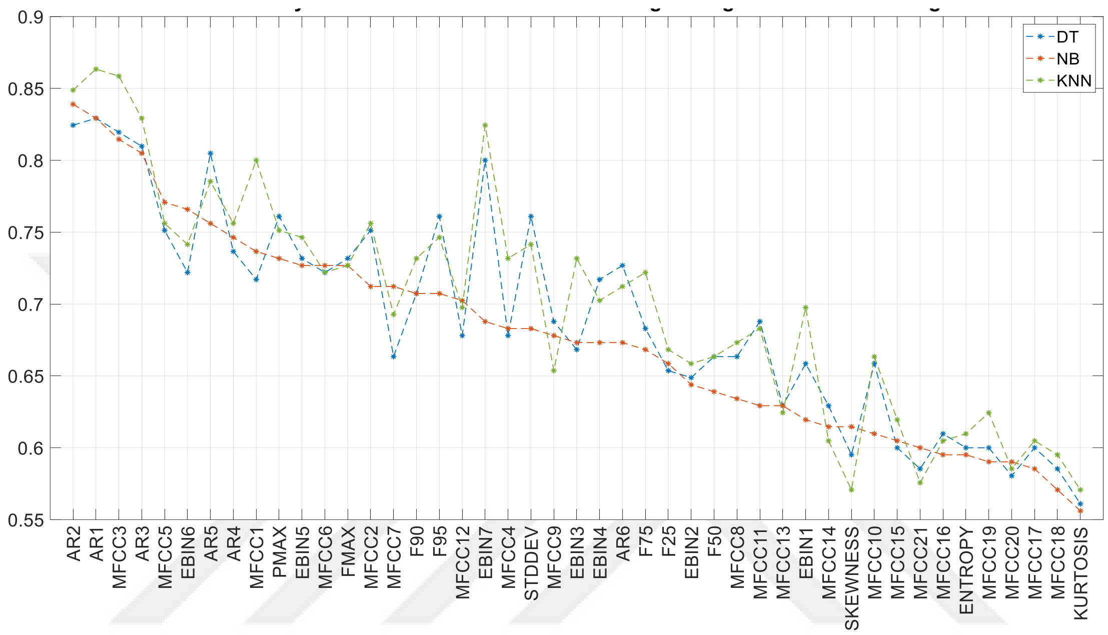
Figure 5.4. Depth Value Count



Similarly, the Naive Bayes (NB) classifier was evaluated, with the features sorted according to the highest accuracy derived from this classifier. The top features for the NB classifier were AR2, AR1, MFCC3, AR3, and MFCC5. The highest accuracy recorded by the NB classifier was 83.9%, with AR2 being the most significant feature.

This sorting and accuracy data were visualized in a plot as well. The plot can be observed in Figure 5.5.

Figure 5.5. Accuracy with NB Indexing



Each point in these plots is associated with the maximum accuracy taken from a subphase and microphone pair. These are also noted so that we can extract a future statistical measure. E.g. how many times a certain mic has given the max accuracy. By examining Figures 5.1, 5.3, 5.5, it is evident that certain features consistently rank among the top in terms of accuracy across all classifiers. Specifically, AR1, AR2, AR3, and MFCC3 repeatedly appear in the top four features with the highest accuracy scores for each classifier. This consistency suggests that these features are highly effective in classifying subjects into different categories, regardless of the classifier used. Therefore, it can be concluded that AR1, AR2, AR3, and MFCC3 play a crucial role in the classification process, significantly contributing to the overall accuracy of the classifiers.

In Table 5.1 below, it sums up the four highest scoring features and their responding Accuracy, F1, and μ scores with the addition of the microphone and sub-phase for the responding accuracy derived from these features for the k-NN classifier. Table 5.2

presents the four confusion matrices associated with the best four features, yielded by the k-NN classifier.

Table 5.1. Classifier Results for Top 4 Features (Values are in %)

	DTT			NB			k-NN					
	ACC	F1	µ	ACC	F1	µ	ACC	F1	µ	Mic	SP	k
AR1	0.83	0.98	0.7	0.83	0.98	0.7	0.86	0.98	0.81	5	5	19
AR2	0.82	0.97	0.7	0.83	0.96	0.7	0.84	0.96	0.78	5	5	39
MFCC 3	0.82	0.96	0.7	0.81	0.96	0.7	0.85	0.97	0.78	2	4	21
AR3	0.81	0.97	0.7	0.81	0.99	0.7	0,82	0.98	0.74	6	6	49

When examining Table 5.2 below, it becomes apparent that the ability to accurately identify and classify COPD subjects is exceptionally high. This heightened accuracy can be attributed to the fact that 50% of the dataset comprises COPD subjects, which likely enhances the classifier's ability to learn and distinguish the characteristics specific to this group. Despite this significant representation of COPD subjects in the dataset, the classifier also demonstrates impressive performance in correctly classifying the other two categories. This indicates a robust overall capability of the classifier, not only in recognizing COPD patients but also in accurately distinguishing between the remaining classes, ensuring reliable classification across the board.

Table 5.2. Confusion Matrix for Top 4 Features

		AR1				AR2	
	ACTUAL C	ACTUAL NS	ACTUAL S		ACTUAL C	ACTUAL NS	ACTUAL S
PREDICTED C	100	0	0	PREDICTED C	98.1	1.8	0

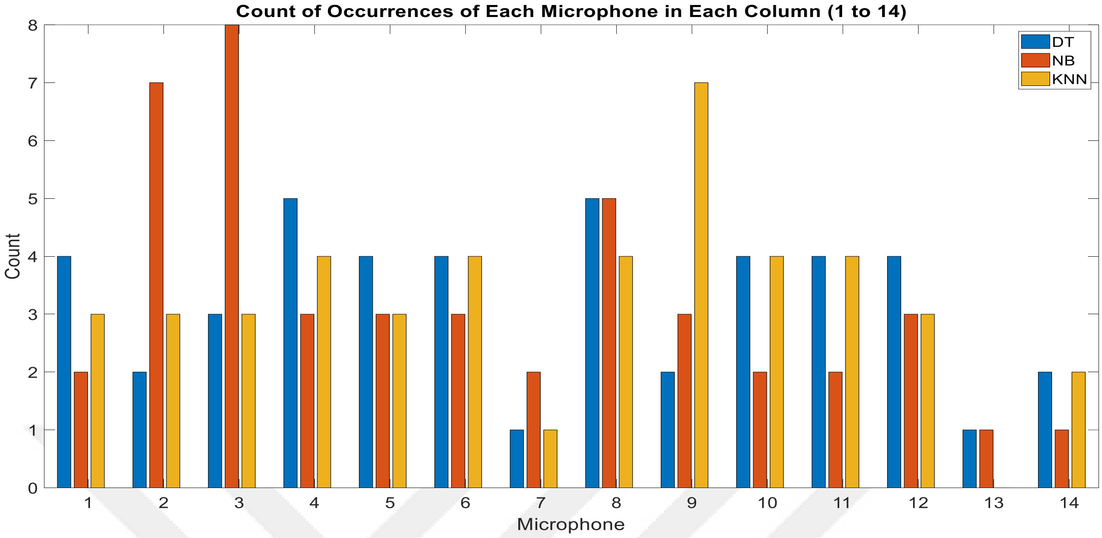
PREDICTED NS	5.6	79.1	15.2	PREDICTED NS	16.5	58.6	24.8
PREDICTED S	0	34.8	65.1	PREDICTED S	2.8	24.6	72.5
	ACTUAL C	AR3 ACTUAL NS	ACTUAL S		ACTUAL C	MFCC3 ACTUAL NS	ACTUAL S
PREDICTED C	100	0	0	PREDICTED C	99	1	0
PREDICTED NS	20.6	68	11.3	PREDICTED NS	10.5	69.6	19.7
PREDICTED S	1.5	36.1	62.2	PREDICTED S	0	28.7	71.2

An observation to be noted is that the classifier thinks a non-smoker is actually COPD classified while it almost never thinks that a smoker is COPD. COPD is associated with smoking, but the classifier doesn't say so in terms of acoustic sounds surprisingly. One would expect smokers to be confused with COPD but that is not the case. This can be further investigated in future studies.

In Appendix A, three F1 plots were created to illustrate the F1 scores for the three classifiers, specifically focusing on their performance in identifying the COPD class. These plots provide a clear comparison of how effectively each classifier distinguishes COPD patients from the other classes. Upon reviewing the three plots in Appendix A it is evident that all three classifiers exhibit a high success rate in accurately identifying the COPD class. This consistent performance across different classifiers underscores their reliability and effectiveness in detecting COPD, highlighting the robustness of the classification approach used in this study.

This study also analyzed how frequently each microphone was associated with the maximum accuracy of each classifier. Since there are forty-five features in total, the sum of observations should equal forty-five. In Figure 5.6 below the results can be observed.

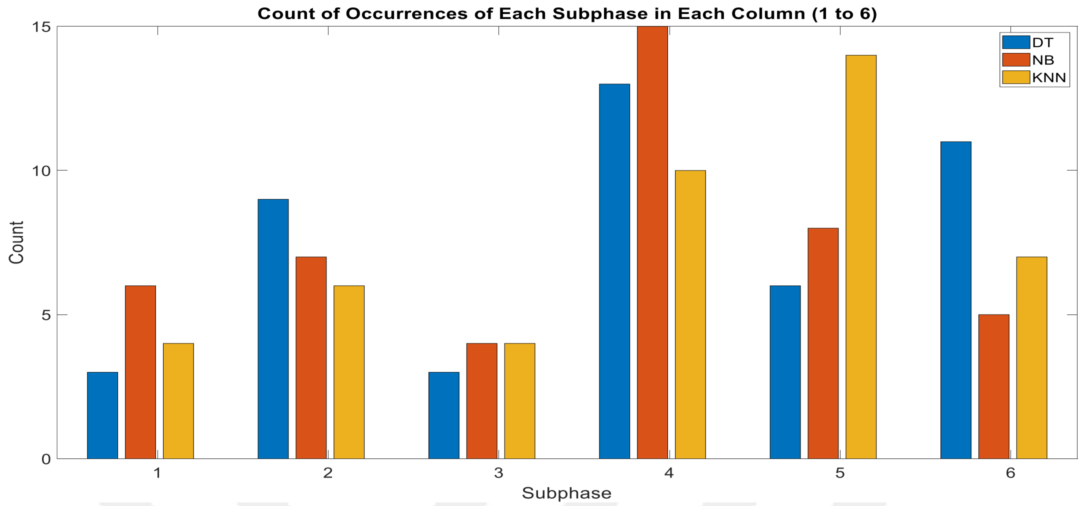
Figure 5.6. Count of Occurrences of Each Microphone for All Features



Upon examining the data, it is evident that for the Naïve Bayes classifier, the third microphone achieved maximum accuracy eight times. Naïve Bayes is more robust with respect to unequal sample counts of the classes, So, maybe Microphone 2 and 3 with respect to naïve bayes are worth attention in future studies. However, the main focus is on the yellow bar, which represents the k-NN classifier, as it yielded the highest overall accuracy in this study. The Figure indicates that the ninth microphone, which is positioned at the bottom of the back (as shown in Figure 2.1), was associated with the maximum accuracy for seven different features. This underscores the significant contribution of the ninth microphone to the k-NN classifier's performance in achieving the best results.

This study also scrutinized the frequency of occurrence for each sub-phase when examining the maximum accuracy achieved by each classifier across all features. With a total of forty-five features under consideration, the sum of observations should indeed equal forty-five. In Figure 5.7 below, the results are plotted.

Figure 5.7. Count of Occurrences of Each Subphase

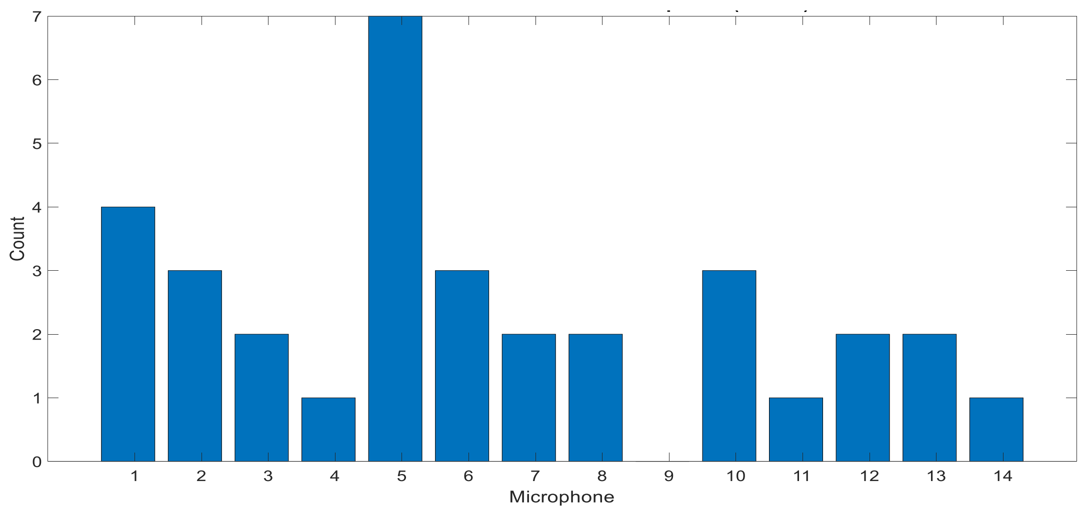


Upon careful examination, it is notable that for the Naïve Bayes classifier, the 4th sub-phase, characterized as early expiration, was associated with maximum accuracy on fifteen separate occasions. This observation underscores the noteworthy prevalence of the 4th sub-phase, specifically early expiration, within the context of the Naïve Bayes classifier's performance across various features.

To look more in depth at what microphone or sub-phase was most crucial in finding said high accuracy, the study narrowed down the features used to only the ones who gained an accuracy score of 75% and higher. The features that were found were the following: (AR1, AR2, AR3, AR4, AR5, MFCC 1, MFCC 2, MFCC 3, MFCC5, EBIN 7, and PMAX).

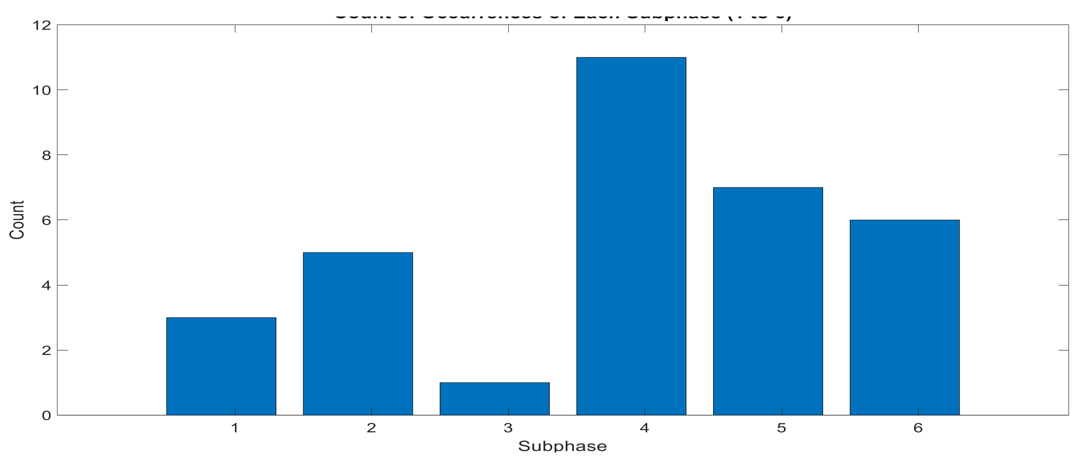
Upon analysis, it was observed that the fifth microphone was recurrently highlighted, appearing a total of seven times across all three classifiers, when considering features with accuracy scores surpassing the 75% threshold. This observation suggests a potential significance of the fifth microphone in contributing to accurate classification outcomes. This can be observed in Figure 5.8.

Figure 5.8. Count of Occurrences of Each Microphone for Above 75% Accuracy



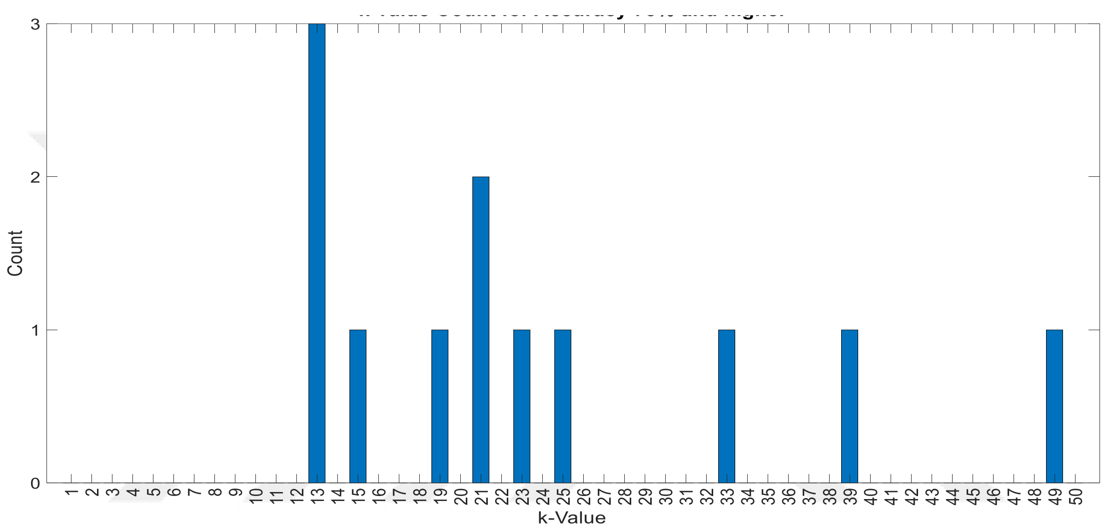
In contrast, in Figure 5.9 below, it specifically examines sub-phase counts for features with accuracy scores exceeding 75%. Among these high-scoring features, a total of eleven instances were recorded for the 4th sub-phase. Notably, this represents a subset of thirty-three total sub-phase counts, as only features with accuracy rates surpassing 75% were considered. This observation highlights the prevalence of the 4th sub-phase within the subset of features that achieved high accuracy scores, indicating its significance in contributing to accurate classification outcomes.

Figure 5.9. Count of Occurrences of Each Sub-phase for Above 75% Accuracy



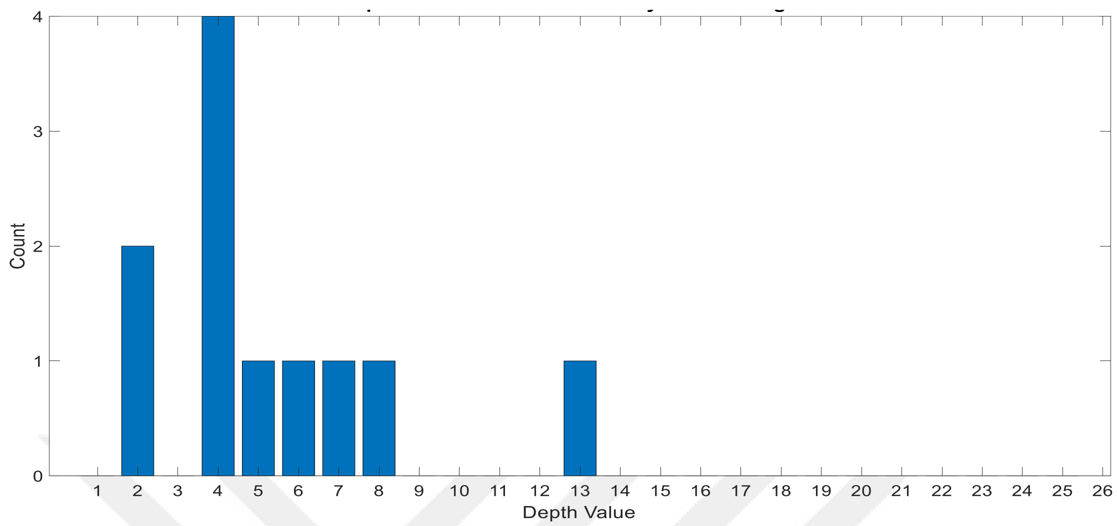
Also, the same thing was done to find the best k-NN value for any feature scoring a maximum accuracy of 75% and above and this can be observed through Figure 5.10. It was noted that the 13th neighbor was again the most counted amongst the highest scoring features.

Figure 5.10. k-Value for Accuracy 75% and Higher



Similarly, the count for the depth values repeated in the decision tree classifier for the 75% accuracy and higher scoring features were plotted and can be observed in Figure 5.11. The figure shows that the 4th depth value shows 4 counts amongst the top scoring features where the accuracy is 75% and higher.

Figure 5.11. Depth Value for Accuracy 75% and Higher



Also, the scores for μ were plotted and can be observed in Appendix.

CONCLUSION

In conclusion, the primary aim of this study was to identify a reliable and accurate method for classifying patients into three distinct categories using the k-Nearest Neighbors (k-NN) classifier. The study successfully achieved this goal, demonstrating that the k-NN classifier could accurately categorize patients with a high degree of accuracy. The results of this study provide a solid foundation for future research, which can build upon these findings to explore more nuanced aspects of patient classification. The k-NN classifier was found to be the most successful among the 3 classifiers with a high accuracy of 86.3%. The microphone that yielded the highest accuracies among the features that gave an accuracy of 75% and higher was the 5th microphone. The sub-phase that yielded the highest accuracy among the features that gave an accuracy of 75% and higher was the 4th sub-phase or early expiration. The best k -value for the k-NN classifier was found to be the 13th neighbor amongst the 75% and higher scoring maximum accuracy. Lastly, the best features found to classify the 3 classes were AR1, AR2, AR3, and MFCC3.

One intriguing direction for future research is to investigate the relative distances between the classified categories. This involves determining how closely or distantly patients in one category are from those in another category. For example, once a patient is accurately classified into a specific category, researchers could explore how far they are from the characteristics of another category. This could be particularly useful in clinical scenarios where patient conditions are dynamic and can change over time.

Consider a healthy smoker visiting the clinic and recording the acoustic sounds and then the algorithm classifies him a healthy smoker very efficiently. However, this classification is just the beginning. Additional patient information, such as the amount of cigarette smoking, can be integrated into the model. With this data, the algorithm can predict how the patient's condition might evolve. Projecting how the patient's acoustic characteristics might change over time and begin to resemble those of patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD). This predictive capability is invaluable as it can provide early warnings to patients, helping them understand the potential long-term consequences of their habits and encouraging preventive measures.

The study's results indicated that the highest accuracy in classification was obtained using the fifth microphone and the fourth sub-phase. This specific finding opens another avenue for future research: the investigation of microphone placement and its impact on classification accuracy. By understanding why this microphone and sub-phase yielded the best results, researchers can optimize data collection methods to further enhance the precision of patient classification.

The classifiers employed in this study showed remarkable performance, particularly in accurately identifying patients with COPD. This level of accuracy is crucial as it underscores the potential clinical applications of the classifier. With reliable classification, healthcare providers can make more informed decisions, leading to better patient outcomes. Early and accurate identification of COPD patients, for instance, allows for timely intervention, which can significantly improve the quality of life for these patients.

In summary, this study not only achieved its primary goal of accurately classifying patients into three categories but also laid the groundwork for future research that could further refine and expand these findings. The insights gained from this research have the potential to transform patient care by enabling early diagnosis and personalized treatment plans. The promising results, particularly regarding the optimal microphone placement and the classifier's accuracy, highlight the importance of continued research in this area. Future studies can build on these findings to develop even more sophisticated tools for patient classification and health monitoring.

REFERENCES

[1] Centers for Disease Control and Prevention, "Youth and Tobacco Use," [Online]. Available: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/youth_data/tobacco_use/index.htm.

[2] Mayo Clinic, "COPD - Symptoms and causes," [Online]. Available: https://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679

[3] World Health Organization, "Chronic obstructive pulmonary disease (COPD)," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[4] National Heart, Lung, and Blood Institute, "Causes and Risk Factors of COPD," [Online]. Available: https://www.nhlbi.nih.gov/health/copd/causes.

[5] A. Oliveira, I. Sen, Y. P. Kahya, V. Afreixo, and A. Marques, "Computerised respiratory sounds can differentiate smokers and non-smokers," Journal of Clinical Monitoring and Computing, vol. 31, no. 3, pp. 571–580, 2017. DOI: 10.1007/s10877-016-9887-8.

[6] V. Gross, A. Dittmar, T. Penzel, F. Schüttler, and P. von Wichert, "The relationship between normal lung sounds, age, and gender," American Journal of Respiratory and Critical Care Medicine, vol. 162, no. 3 Pt 1, pp. 905–909, 2000. DOI: 10.1164/ajrccm.162.3.9905104.

[7] S. Bennett, A. Bruton, A. Barney, T. Havelock, and M. Bennett, "The Relationship Between Crackle Characteristics and Airway Morphology in COPD," *Journal/Conference Name*, vol. x, no. x, pp. x-x, Year.

[8] N. S. Abdullah, C. K. Lam, K. Sundaraj, and R. Palaniappan, "Classification Of Respiratory Sounds in Smokers and Non-Smokers Using k-NN Classifier," in Springer, Singapore, 2017, pp. 73-78. DOI: 10.1007/978-981-10-3737-5_15.

[9] I. Sen, A. Wali, and S. A. Hassan, "Effect of Smoking on Pulmonary Acoustic Parameters in Terms of Displacement Away From Non-smokers Towards COPD:

A Preliminary Study," in 2022 Medical Technologies Congress (TIPTEKNO), Antalya, Turkey, 2022, pp. 1-4, doi: 10.1109/TIPTEKNO56568.2022.9960198.

[10] I. Sen and Y. P. Kahya, “A multi-channel device for respiratory sound data acquisition and transient detection”, In Proc. of the 27th Ann. Int. Conf. of the IEEE Eng. in Med. and Biol. Soc. (EMBC), Shanghai, China, Sept. 2005, pp. 6658-6661.

[11] I. Şen, "A Multi-channel Device for Respiratory Sound Data Acquisition and Transient Detection," M.S. Thesis, Boğaziçi University, 2005

[12] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70-73, Jun. 1967.

[13] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. New York, NY: John Wiley & Sons, 2000.

[14] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. New York, NY: Springer, 2002.

[15] M. B. Priestley, *Spectral Analysis and Time Series*. London, UK: Academic Press, 1981.

[16] S. Aydore, I. Sen, Y. P. Kahya, and M. K. Mihcak, "Classification of respiratory signals by linear analysis," in 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 2009, pp. 2617-2620, doi: 10.1109/IEMBS.2009.5335395.

[17] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2009.

[18] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," Applied Soft Computing, vol. 105524, 2019. DOI: 10.1016/j.asoc.2019.105524.

[19] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21-27, Jan. 1967.

[20] H. Zhang, "The Optimality of Naive Bayes," in *Proc. 17th Int. FLAIRS Conf.*, 2004, pp. 562-567.

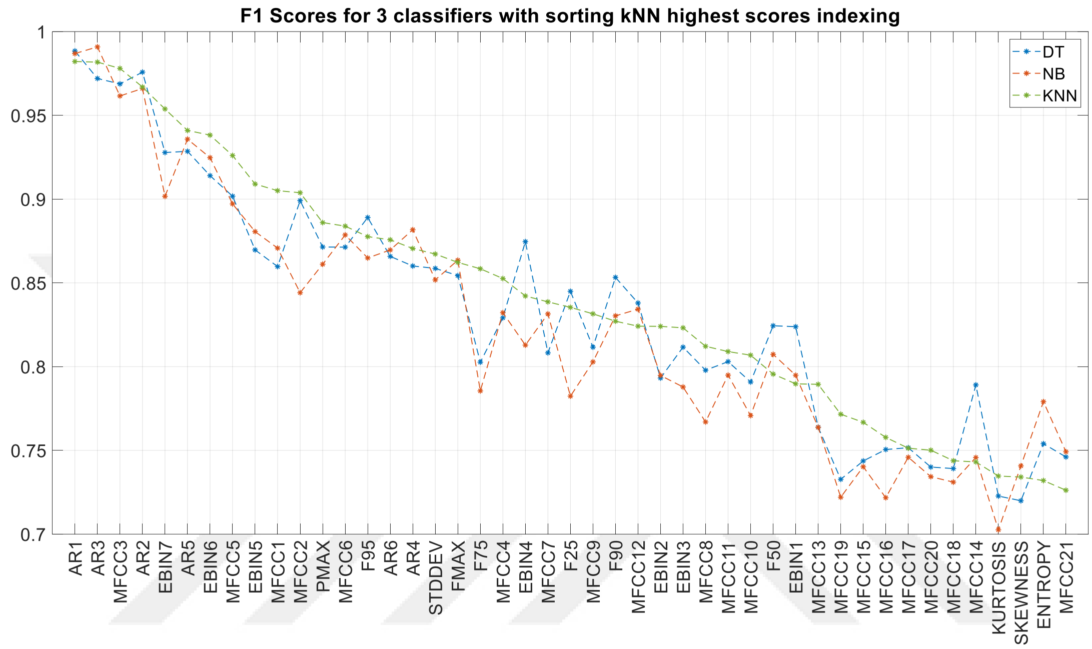
[21] L. Rokach and O. Maimon, "Decision Trees," in *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Boston, MA: Springer, 2010, pp. 165-192.

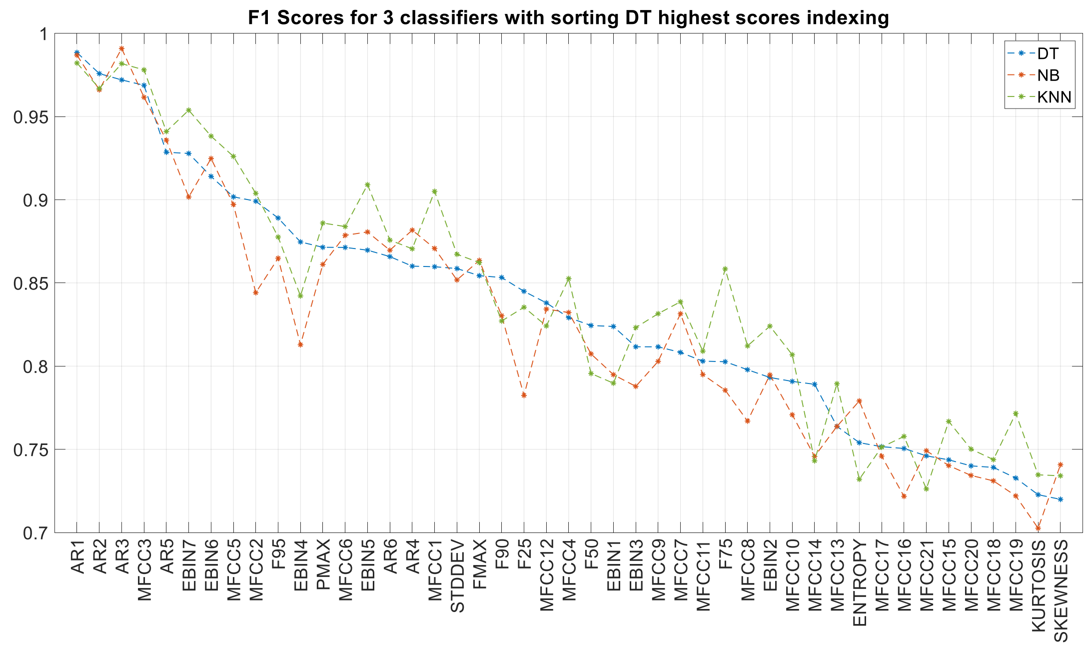
[22] Bharathi, "Latest Guide on Confusion Matrix for Multi-Class Classification," Analytics Vidhya, Feb. 13, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/

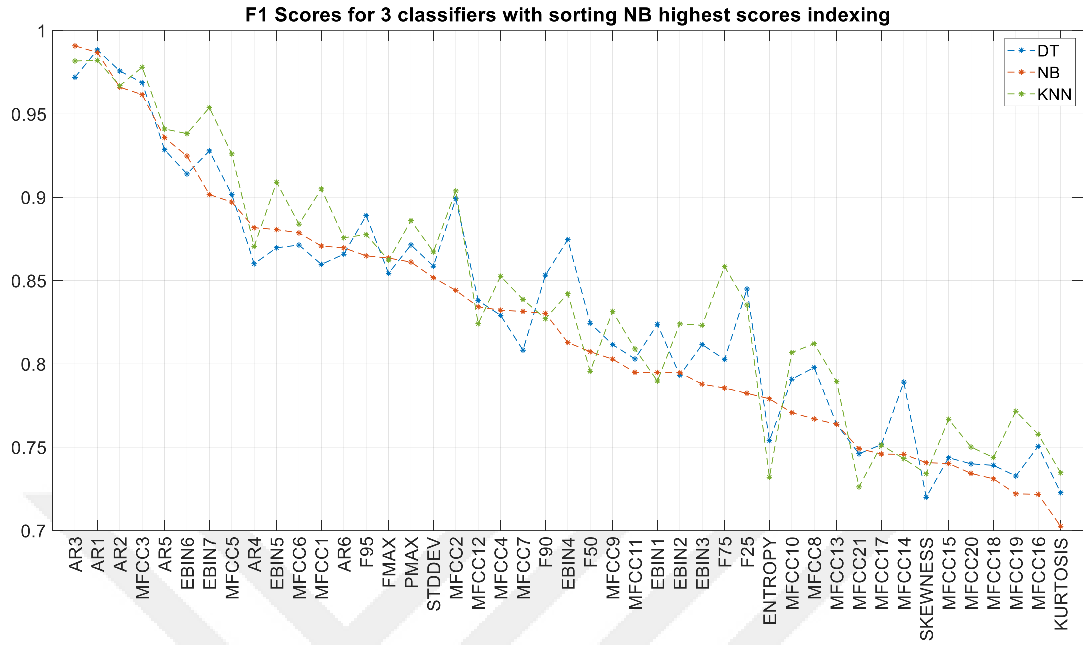
[23] I. Sen, M. Saraclar, and Y. P. Kahya, "Differential Diagnosis of Asthma and COPD Based on Multivariate Pulmonary Sounds Analysis," in IEEE Transactions on Biomedical Engineering, vol. 68, no. 5, pp. 1601-1610, May 2021, doi: 10.1109/TBME.2021.3049288.

APPENDICES

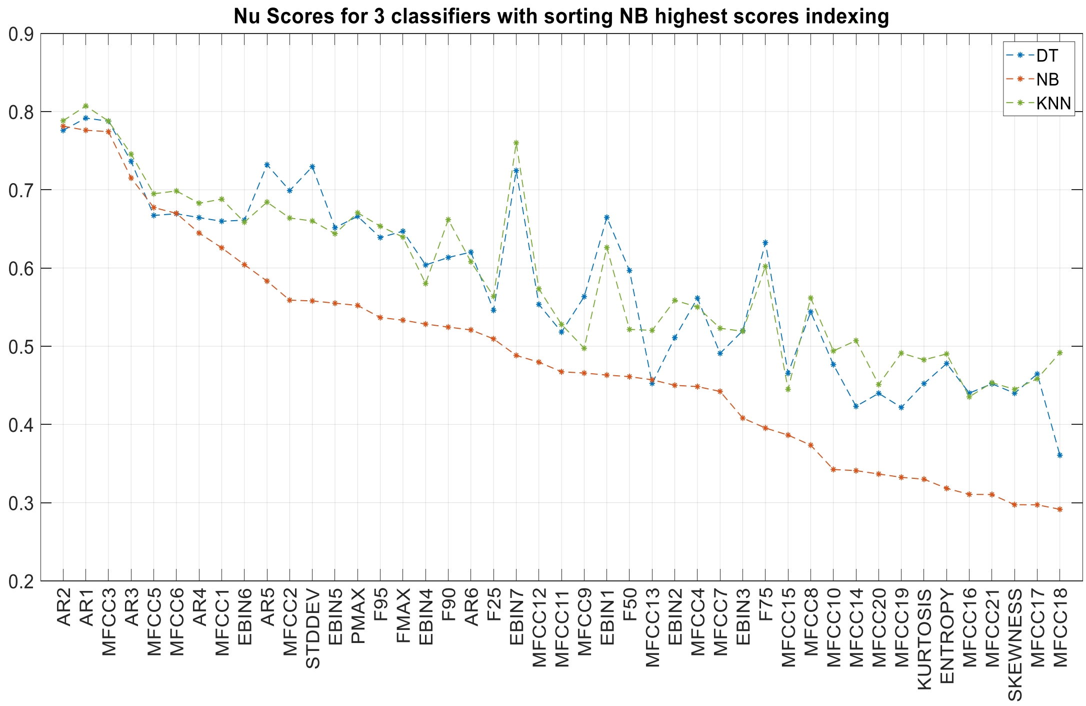
Appendix A. F1 Scores

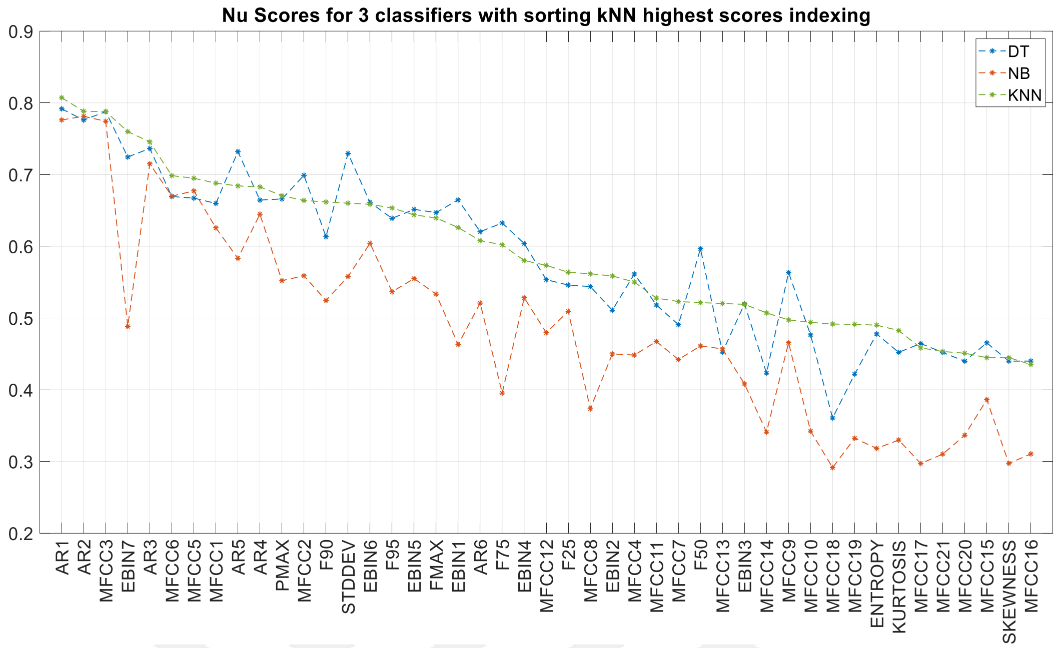


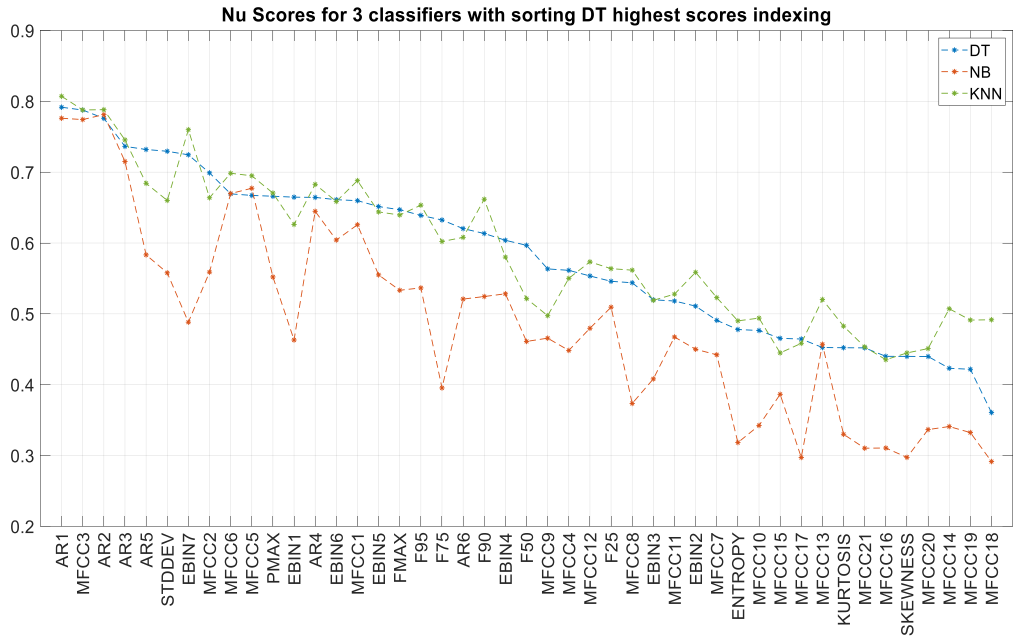




Appendix B. μ Scores




Nu Scores for 3 classifiers with sorting kNN highest scores indexing
DT
NB
KNN
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
AR1 AR2 MFCC3 EBIN7 AR3 MFCC6 MFCC5 MFCC1 AR5 AR4 PMAX MFCC2 F90 STDDEV EBIN6 F95 EBIN5 FMAX EBIN1 AR6 F75 EBIN4 MFCC12 F25 MFCC8 EBIN2 MFCC4 MFCC11 MFCC7 F50 MFCC13 EBIN3 MFCC14 MFCC9 MFCC10 MFCC18 MFCC19 ENTROPY KURTOSIS MFCC17 MFCC21 MFCC20 MFCC15 SKEWNESS MFCC16


Nu Scores for 3 classifiers with sorting DT highest scores indexing
DT
NB
KNN
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
AR1 MFCC3 AR2 AR3 AR5 STDDEV EBIN7 MFCC2 MFCC6 MFCC5 PMAX EBIN1 AR4 EBIN6 MFCC1 EBIN5 FMAX F95 F75 AR6 F90 EBIN4 F50 MFCC9 MFCC4 MFCC12 F25 MFCC8 EBIN3 MFCC11 EBIN2 MFCC7 ENTROPY MFCC10 MFCC15 MFCC17 MFCC13 KURTOSIS MFCC21 MFCC16 SKEWNESS MFCC20 MFCC14 MFCC19 MFCC18