Using Clustering for Pulmonary Disease Detection through Respiratory Sound Analysis

Luka Žontar
Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
Email: lz3057@student.uni-lj.si

***Abstract*—Curing pulmonary diseases in later stages of diseases can be very tough, which is why preventive healthcare is essential. Due to overcrowded hospitals, it is very hard to get a doctor's appointment for no reason. On the other hand, when symptoms occur, the disease may have already advanced to dangerous stages. This is why automatic preventive healthcare is very important and in this article we tried to determine, which features are the most important to include in our research. Signal features extracted from audio recordings of respiratory sounds proved to be very efficient in pulmonary disease detection. We examined demographic and three audio based feature subsets: time domain, frequency domain and cepstral domain features. Using PCA analysis and distributions of diagnoses in clusters, we tried to determine the most important features exploiting different clustering algorithms by checking how homogeneous the generated clusters are. To generate results we used five different algorithms: K-means, Spectral clustering, Hierarchical clustering, DBSCAN and Gaussian Mixture Model. Our results show how different features corresponds with diseases such as COPD, which demographic features are important and which are not. Moreover, time domain features proved to be less efficient in clustering homogeneous clusters that other audio based features. Essentially, we prove that unsupervised learning algorithms can be a powerful tool in automatization of preventive healthcare and learning more information about pulmonary diseases.**

***Index Terms*—Clustering, feature selection, pulmonary disease detection, audio analysis.**

I. Introduction

Auscultation, the process of listening to a person's lungs is one of the oldest medical procedures that is still in use today. With healthcare getting overcrowded, routine preventive analysis of respiratory sounds seems to be out of the question. Machine learning provides an alternative approach by automating this procedure. Using deep learning, convolutional neural networks and support vector machine algorithms, scientists have already developed models that are quite successful in pulmonary disease detection.

Unlike heartbeat sounds, lung sounds are much more irregular and vary from patient to patient. Moreover, a sound that a doctor might connect with a pulmonary disease does not necessarily occur in every respiratory cycle. Such variability, makes it very hard to efficiently predict whether or not a patient has a pulmonary disease. Another challenge is working with audio signals, since we have to first transform them into a discrete form and into lower dimensions, so that computational complexity is not excessive. To solve this, several features can be extracted from an audio signal that we will use in this research.

However, our main focus will be to explain, how can we analyze audio signals using clustering algorithms and determine which subset of the given features generate the most homogeneous clusters. In this article, we will first present the known approaches for pulmonary disease detection. Then, we will explain the features extracted from respiratory sound signals. Since clustering might not be the most common approach to audio signal processing, we will also try to explain, how different clustering algorithms can contribute to our results. Lastly, we will evaluate the clustering results of different subsets of both demographic features of a patient and audio features of the respiratory sounds.

II. Related work

An efficient preventive pulmonary disease classifier could help the doctors substantially in decreasing the number of deaths due to such conditions. Chamberlain and his colleauges [1] tried to detect crackles and wheezes by first cleaning the data using a denoising autoencoder and then building two separate support vector machine models (SVM) based on Mel-frequency cepstrum coefficients (MFCC), where one was taught to classify cracles and the other one to classify wheezes. They achieved 0.86 accurracy in wheeze detection and 0.74 in crackle detection.

Another approach was proposed by Aykanat et al. [2], who achieved similar results exploiting spectrogram images. Since spectrograms are visualizations of signal strength or loudness, convolutional neural networks (CNN) were used to model a classifier that can detect crackles and wheezes from images as efficiently as SVM models do in combination with MFCC.

Dumas et al. [3] tried to identify different profiles of bronchiolitis using clustering approach. The research targets children with different profiles of bronchiolitis, where the researchers clustered data into 4 different groups using latent class analysis. They gathered data such as whether a child had a history of wheeze or eczema. Profile A was described with eczema, wheeze, wheezing at emergency department presentation and rhinovirus infection. Profile B had the largest

[1]The implementation of the methodology that we developed in this article is licensed under MIT License and can be accessed on GitHub: https://github.com/lzontar/Clustering-Respiratory-Sounds

probability of respiratory syncytial virus infection, but in contrast with profile A, was only wheezing at emergency department presentation. Profile C was described with longer hospital stay and with more severe retractions. The last group was denoted with shorter stays and had the least severe illness.

III. Audio features

The goal of audio processing algorithms and their applications are to make machines learn as effortlesly from hearing as people do. In order to do that, we have to be able to transform an audio signal to a discrete, machine-friendly form. Thus, researchers have come up with several features that can describe audio signals. In this section we will present, which features were included in our research. Similarly as Sharma with colleauges [4], we separate features into four domains. In respiratory sound analysis we can detect five distinguishable sounds [5] that are relevant for us:

- **Stationary noise** of a recording - a steady murmuring noise in the background that is present in every recording, even though in high quality audio signals stationary noise is much less obvious. Since recordings with different devices have different stationary noises, we will try to take them into account in evaluation.
- **Normal breathing sounds** - can differ with different retrieval locations in auscultation. While at the chest wall, low noise can be detected during inspiration and is hardly audible during expiration, in trachea normal breathing sounds are characterized by a spectrum of noises that contains higher frequency components.
- **Crackles** - short and explosive lung sounds caused by abnormally closed airways that rapidly open, when a crackle occurs. We can furthermore distinguish between fine (short-duration) and coarse (long-duration) crackles. Location in respiratory cycle of different types of crackles is highly related with different pulmonary diseases.
- **Wheezes** - harmonic, musical abnormal lung sounds caused by obstructed airways that vibrate as air passes through them.

A. Time domain features

Time-domain features evolve through time, which is why we window our signal, that is that only a certain part of a signal is considered in feature extraction. This windows slides from the beginning towards the end of the signal.

1) **Zero crossing rate (ZCR)** - denotes the number of times the signal crosses y axis. Since ZCR is higher in voiced segments, it is often used to detect voice. In our case we expect respiratory sounds with multiple crackles to have higher ZCR, since this metric is often used to detect percussive sounds such as the aforementioned one.
2) **Root mean square (RMS)** - calculates the energy of a signal that corresponds to the total magnitude of the signal. Rough approximation of energy is a signal's loudness. A research has been made [6], where the intensity of inspiration and expiration sounds was compared. Scientists confirmed that inspiration part of a respiratory cycle is louder than expiration. By providing this feature our model should be able to distinguish between inspiration and expiration phases of a cycle.

B. Frequency domain features

Using Fast Fourier Transform (FFT), we transform the signal from time domain to frequency domain. Frequency is the number of occurences of a repeating event in a unit of time and we use frequency domain representation, because it simplifies mathematical analysis of the signal. Spectrum is the frequency-domain representation of a signal and it provides a more intuitive understanding of the system.

1) **Chroma Short-time Fourier transform (STFT)** - the entire spectrum is mapped into 12 bins that represent the 12 semitones also denoted as chroma. Consequently, this feature can very well represent music. In our research it helps us understand the harmonic information about our audio signals, i.e. wheezes.
2) **Chroma Energy Normalized Statistics (CENS)** - is another chroma based feature. Here, we take statistics over large windows, which smooths local deviations in tempo and articulations.
3) **Spectral centroid** - indicates where the center of mass of the spectrum is located and thus describes brightness of the signal. Essentially, it gives us the idea of how loud and how high the sound is. Higher sounds can in our case indicate the presence of wheezes.
4) **Spectral bandwidth** - is the difference between the highest and lowest frequency in a window. In our research we will use it, because it will help us in determining whether a signal contains a crackle or not. High differences in highest in lowest frequencies will most likely correspond with crackles in our signals.
5) **Spectral contrast** - considers the spectral peaks and valleys and their differences in each frequency subband. Crackles are explosive and short sounds and thus produce clear valleys and peaks in the signal representation. This feature could help us detect the presence of crackles.
6) **Spectral flatness** - is the measure of uniformity in the frequency distribution of the power spectrum, calculated as the ratio between geometric and arithmetic mean. and can be used to distinguish between noise and harmonic sounds, which can be very important in respiratory sound analysis, since breathing tends to be less harmonic sound, if wheezes are not present. Moreover, it could be used to determine crackles, because flatness measures uniformity, which is a property that a respiratory sound containg crackles does not possess.
7) **Tonnetz** - computes the tonal centroid features and can be used to show harmonic relationships in an audio signal. Tonal features are often used for harmonic sounds classification. Similarly as chroma features, it will help us determine whether wheezes are present in our signal.
8) **Spectral rolloff** - specifies the frequency below which a certain percentage of the total spectral energy of the

signal lies. In our case, it will be used to determine uniformity of a signal and thus whether or not crackles are present.

9) **Spectral entropy** - we first divide windows into sub-frames and evaluate the probability of the energy in a sub-frame occuring. It tries to determine peakiness, which is just the opposite to flatness function.

C. Cepstral domain features

Using the inverse of a Fourier transfrom, we generate a cepstrum. These features are often used in pitch detection and are a very important tool in machine learning with audio signals.

1) **Mel Frequency Cepstral Coefficients (MFCC)** - are derived from the cepstral representation of the audio signal and describe spectral envelope. Frequency bands are uniformly spread on mel-scale, which tries to mimic human auditory system. In connection with SVM, MFCC were used to achieve benchmarking results in respiratory sound analysis and pulmonary disease detection. These coefficients reduce the dimensionality of the data. We will be interested in how crackles relate with MFCC.

IV. Unsupervised learning

Clustering algorithms try to identify different groups of objects and help us classify these data groups into structures that are easier to understand. In this section, we will present different algorithms that we will use in our research and explain how each could contribute to our results. Since we have an idea of the number of clusters, we will mostly use algorithms that take as a parameter the number of clusters. These are: K-means, Hierarchical clustering, Spectral clustering and Gaussian Mixture Model. To see whether our dataset forms clusters with different densities, we also included DBSCAN algorithm in our research. In this section, we will also explain how algorithm parameters were chosen.

A. Number of clusters

All the algorithms, expect DBSCAN, accept the number of clusters as a parameter. Intuitively, we can separate three different values for this parameter.

1) Number of clusters: 2 - essentially, we could group all diagnoses into two groups - being healthy or not.
2) Number of clusters: 3 - we have 3 different sounds. Firstly, we have normal lung sounds and then we also have crackles and wheezes that strongly relate with different pulmonary diseases.
3) Number of clusters: 7 - we have 7 different diagnoses, where 6 of those represent different pulmonary diseases and the other one represents healthy respiratory cycles.

We have to be aware that our hypothesis might be wrong, which is why we perform several different evaluation techniques on the number of clusters. As an example, we will perform this evaluation using K-means algorithm and all the available features.

1) Elbow method: Here, we map the total within-cluster sum of squares measures to the corresponding number of clusters. Wherever the greatest decrease in information gain is, i.e. where we see the elbow in the plot, is the optimal number of clusters. In Figure 1, we can see the elbow method being performed on our data.

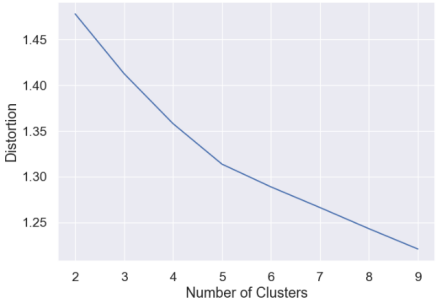

Fig. 1. Elbow method shows no clear elbow, which is why in this case it is difficult to determine the optimal number of clusters. Most likely, we would choose 5, since we notice the greatest decrease in information gain there.

2) Silhouette analysis: Silhouette score shows us how close each point is to its neighboring clusters. In Figure 2, we can see a visualization of Silhouette analysis. Mean value should be close to 1 and plotted Silhouette scores should be close to the mean Silhouette score. Moreover, widths of plots of each cluster should be uniformly distributed.

3) Calinski-Harabasz criterion: Calinski-Harabasz (CH) criterion evaluates the ratio between the inter-cluster dispersion and the between-clusters dispersion. Wherever this ratio is at its peak, the number of clusters is optimal. In Figure 3, we can see this evaluation being performed on our data.

B. Algorithms

1) K-means [7]: K-means is one of the most used algorithms in clustering analysis due to its speed and simplicity. As a parameter it takes the number of clusters and generates clusters with as equal variance as possible. We choose ”k-means++” algorithm to determine the optimal initial centroids.

2) Hirarchical clustering [8]: Using hierarchical clustering we merge entities until the desired number of clusters is achieved starting with each point as its own cluster. A popular visualization of hierarchical clustering is a dendrogram, which also helps us determine the number of clusters.

3) Gaussian Mixture Model [9]: Gaussian Mixture Models or GMM are often used in audio processing tasks due to its computational efficiency and ability to model arbitrary probability density functions. As the name suggests, it does that using a mixture of Gaussian probability distributions.

4) Spectral clustering [10]: Spectral clustering has its roots in graph theory, where the goal is to identify communities of nodes in graphs using eigenvalues (or spectrum) of matrices that are built from the graph that represents our dataset.

5) DBSCAN [11]: Density-Based Spatial Clustering of Applications with Noise or DBSCAN is an improvement of Mean-Shift algorithm that tries to find entities with common

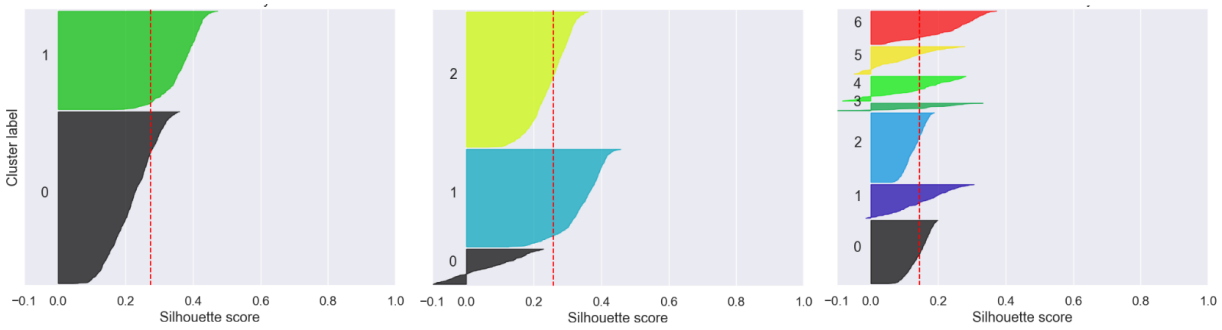

Fig. 2. We performed Silhouette score analysis on K-means algorithm, where the number of clusters were 2, 3 and 7. All the features were used in this evaluation. Mean Silhouette score is quite low, which means that clusters are more indifferent and taht the distance between them is indistinguishable. From the figure, we could most likely conclude that 2 is the optimal number of clusters, because cluster sizes are the most uniformly distributed and in both clusters, Silhouette scores are the closest to the mean in general, which indicates low variance and high density clusters.

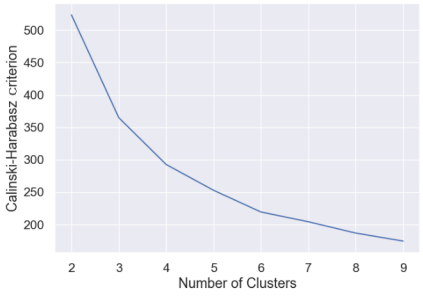

Fig. 3. CH criterion was compared with the number of clusters of K-means algorithm. We can see that CH criterion is monotonically decreasing, which means that the optimal number of clusters, considering this criterion, is the local maximum, which is 2. The CH criterion is the highest, when we generate 2 clusters for our data, which means that if we choose different number of clusters, the algorithm will generate less dense clusters that will be harder to distinguish from one another.

probability density function. All points within a distance ϵ are classified as neighborhood points. Furthermore, points that do not have the specified minimum number of neighbours are labeled as noise. We evaluate the optimal parameters ϵ and the minimum number of samples in a cluster by maximizing the average Silhouette score.

V. Feature subsets evaluation

In clustering, we essentially try to split data into groups based on feature similarity of entities. As we mentioned before, we could evaluate our data based on diagnosis, the main respiratory sounds or the fact whether a patient is healthy or not. In this section we will evaluate our results by splitting our dataset based on four different feature subsets:

1) time domain features,
2) frequency domain features,
3) cepstral domain features,
4) demographic features: age and BMI.

In each subsection, we evaluate our results based on different classes. Even though homogeneous clusters are often smaller in size in our results, we will mostly interpret those due to simplicity and effectiveness of our results. Firstly, we will take a look, how different feature subsets affect determining whether a patient is healthy or not. We continue with evaluating clusters with different respiratory sounds: normal sounds, wheezes and crackles. We will conclude our results by trying to cluster data into 7 different clusters, where each cluster represents one diagnosis.

A. Preprocessing

Firstly, we have to preprocess and clean our dataset. Our dataset originally consists of 122 patients, each containing demographic information, patient diagnosis and an audio recording of his breathing. Each recording also contains the information about all the respiratory cycles. It contains the information of when the respiratory cycle begins and ends and furthermore, if it contains crackles of wheezes. Rows that contain missing values are filtered. Moreover, we filter all the rows where the breathing was not recorded with Meditron stethoscope and where acquisition mode is not sequential/single chanell. This vastly decreases the number of patients, however, usage of different stethoscopes and exploiting different acquisition modes may provide results that are highly dependent on these features. We continue preprocessing our data by normalizing it to the $[0, 1]$ interval. To do that, we exclude all non-numerical features.

After preprocessing, we are left with 55 patients, where diagnoses are not uniformly distributed. In Figure 4, we can see the distribution of diagnoses. We resample our data using Synthetic Minority Over-sampling TEchnique (SMOTE) [12] for easier interpretation of results. SMOTE improves random oversampling by creating synthetic minority class samples.

B. Healthy or unhealthy

In this subsection, we try to determine, which subsets of features are more descriptive of whether a patient has a pulmonary disease or not and how the results differ in clustering algorithms and why. In Figures 5, 6, we provide an example of DBSCAN in combination with cepstral domain features.

We found that K-means and Hierarchical clustering often result in the same clusters distribution. K-means and Hierar-

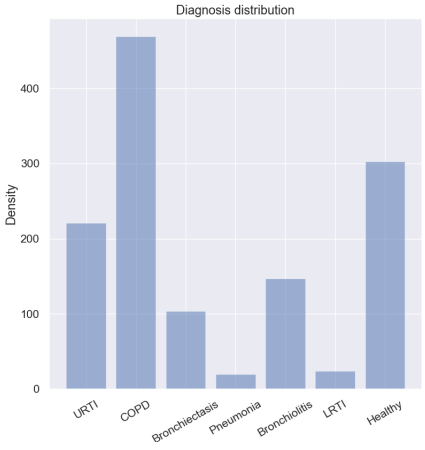

Fig. 4. In the histogram representing the diagnosis distribution, we see that some classes have much more entities than others. For easier interpretation, a uniform distribution would be more appropriate.

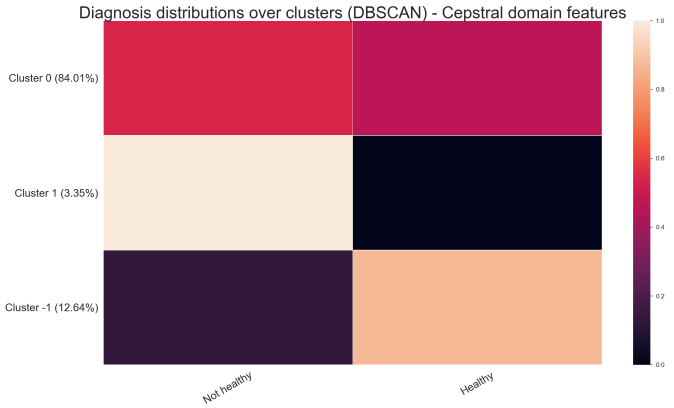

Fig. 5. This heatmap represents the relationship between clusters and diagnoses, whether a patient is healthy or not. Each cell represents While the majority of entities belongs to a cluster that has diagnoses almost uniformly distributed, we will focus on the smaller clusters that are highly homogeneous. While the smallest clusters that take into account about 15% of all entities, mostly describe healthy patients, we find it very interesting that outliers found by DBSCAN algorithm that we used in this case mostly correspond to unhealthy people.

chical clustering are very similar, since they both try to find clusters with as equal variance as possible, however, while K-means is regarded as a top-down approach, Hierarchical clustering is just the opposite, a bottom-up approach. Figure 7 corresponds with a PCA analysis of clustering healthy and unhealthy patients using K-means algorithm and frequency domain features.

Furthermore, we found that using only time domain features, we cannot successfully distinguish between healthy and unhealthy patients. The most homogeneous clusters were generated with Spectral clustering used with time domain feature, which is shown in Figure 8. We are left with a question, why does Spectral clustering provide the most optimal results

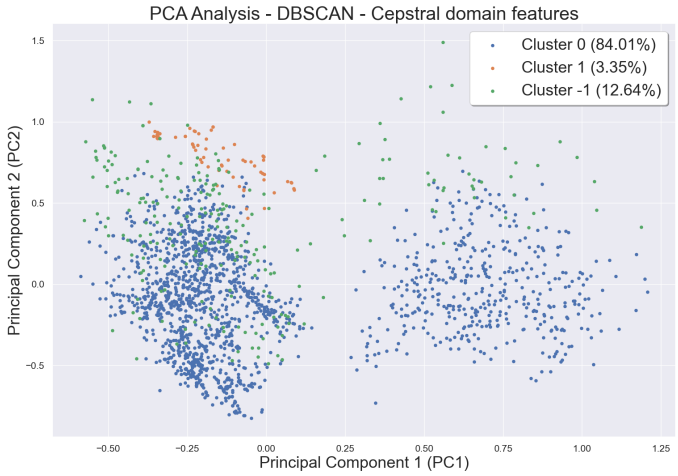

Fig. 6. Using PCA analysis we try to define the homogeneous clusters. From the figure, we can see that the majority of outliers are not very distinctive from other clusters. Interestingly, outliers highly relate to healthy patients. As we can also see in Figure 9, most outliers correspond to normal sounds.

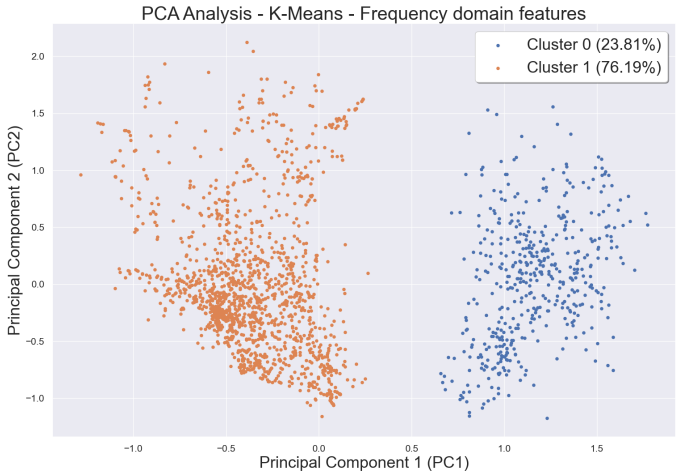

Fig. 7. PCA analysis in this case generates very well separated clusters. The minority cluster (Cluster 1) is homogeneous and represents healthy patients, while in the other cluster diagnoses are much more uniformly distributed. We can see that clusters are separated by PC1, which corresponds with higher spectral rolloff and centroids. Which indicates the the presence of both wheezes and crackles. Judging by chroma vector that indicates that G and G# semitones are present in the signal, we assume that these recordings contain wheezes, since the aforementioned semitones are quite high and could imply on harmonic sounds such as wheezes.

for us? It differs from other algorithms, as it tends to avoid local minima solutions and works more efficiently with data containing transitive connections.

C. *Respiratory sounds*

Continuing with describing respiratory sounds with different feature subsets and using the aforementioned algorithms provides some interesting results as well. In Figures 9, 10, we

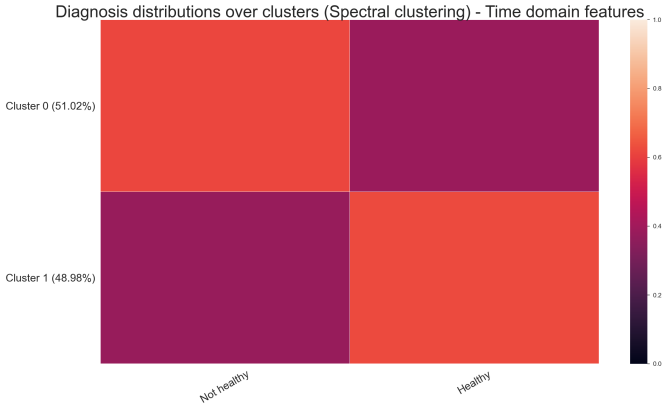

Fig. 8. In this heatmap, we see how different diagnoses are distributed over clusters. While in both clusters one of the diagnoses prevails, it is not as clear as we would like it to be, which is why, we cannot draw any conclusions using time domain features.

continue the interpretation of clustering with DBSCAN.

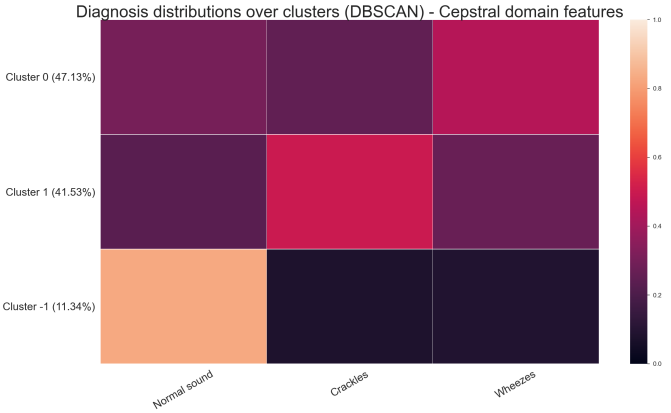

Fig. 9. Here we see the relation between different respiratory sounds, MFCC and clusters created with DBSCAN. Obviously, most of the patients that were denoted as healthy in Figure 6, produce normal breathing sounds.

In general, clustering algorithms seem to perform quite poorly when trying to distinguish between types of respiratory sounds. For example, if we take into account demographic features in combination with GMM, the clusters are not very homogeneous, as we can see in Figure 11

D. Diagnosis

To conclude results section, we try to determine which features describe different diagnoses with the most detail. We begin with DBSCAN interpretation with frequency domain features that is shown in Figures 12.

Once again, DBSCAN provides interesting results. In evaluation of how strong in pulmonary disease detection are demographic features, we find that DBSCAN and demographic features separate COPD and bronchiectasis from the other diagnoses. The results are shown in Figures 13, 14.

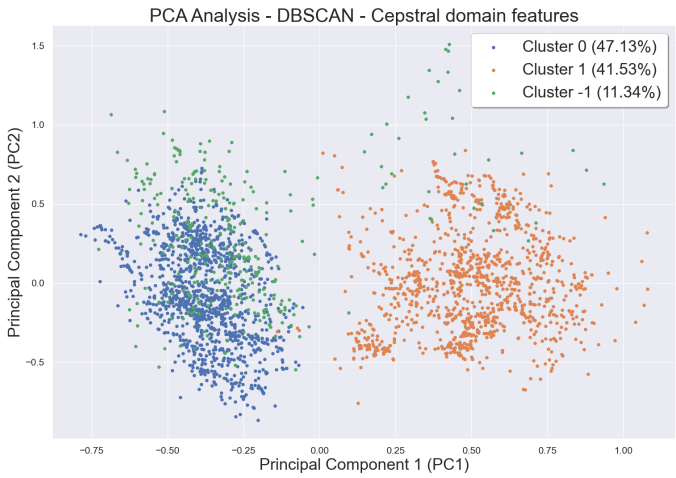

Fig. 10. From the figure, we can see that entities denoted as outliers by DBSCAN have higher PC1 than other two clusters. PC1 has higher values in lower order coefficients, which give us the most information about the overall shape of the spectral envelope. In PC1, the most important is the first order coefficient, which represents the distribution of spectral energy between low and high frequencies [13]. It relates with spectral centroid, which indicates that in our case, normal sounds have higher scores in spectral centroids. Considering the ground rules of DBSCAN and how outliers are generated, we can also assume that normal sounds vary much more than abnormal respiratory sounds such as crackles and wheezes.

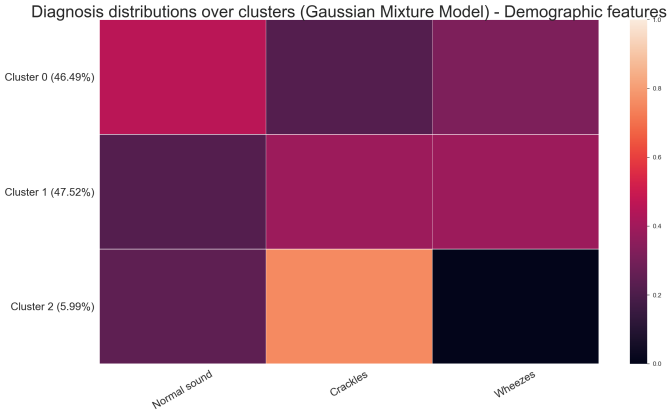

Fig. 11. In the figure we can see, how different respiratory sounds are distributed over clusters generated with GMM using demographic features. About 6% of entities form a fairly homogeneous cluster and with PCA analysis, we conclude that this cluster corresponds to older and heavier people. In reality these are the people that have a higher probability of crackles occurring.

Interestingly, Gaussian Mixture Model produces clustering with the most homogeneous clusters. GMM assumes that there exist k subpopulations of data with normal distribution. From Figures 15, we can assume that certain diseases have normally distributed values of frequency domain features.

Lastly, we interpret the relation of cepstral domain features and different diagnoses. In Figures 17 and 18, we can see a heatmap of clusters related with diagnoses and PCA analysis

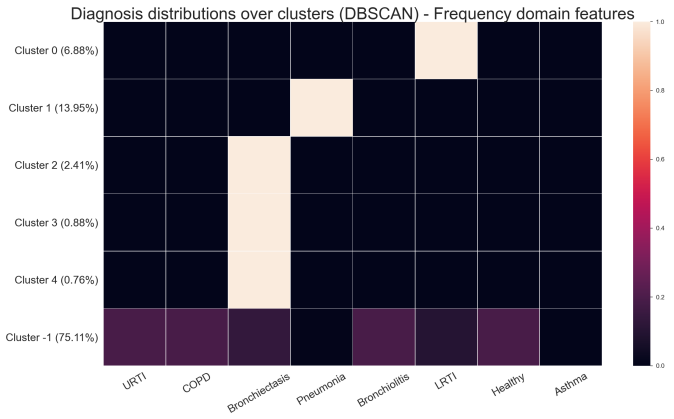

Fig. 12. In this figure we can see, how different diagnoses are distributed over clusters generated with DBSCAN using frequency domain features. 75% of data points are interpreted as outliers. While Silhouette score is maximized, when searching for the optimal ϵ and the minimum number of samples, we can still generate irrational results by setting ϵ too low and the minimum number of samples too high and thus classifying the majority of data points as outliers. Nevertheless, we find two clusters that are still statistically significant, homogeneous and are defined by different probability density functions detected by DBSCAN.

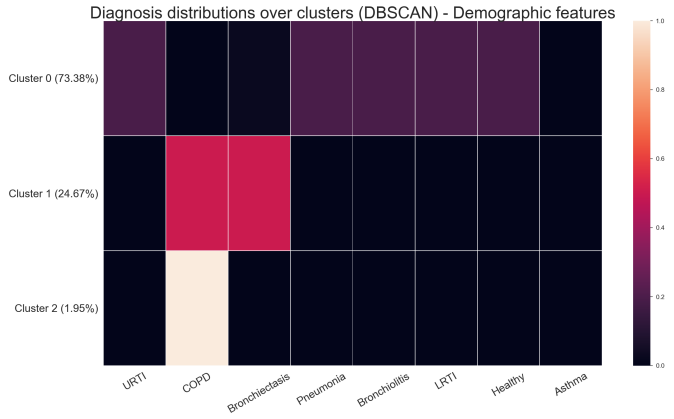

Fig. 13. Here, we can see how DBSCAN in combination with age and BMI can efficiently separate COPD and bronchiectasis diagnoses from the rest. While our primary goal might be to detect whether a pulmonary disease is even present, the results seem promising as we successfully separated two of the diagnoses.

performed on the aforementioned data clustered with Spectral clustering algorithm.

Our results have shown that clustering can be very useful in automatization of preventive healthcare and moreover in learning about different diseases. While time domain features have proven to be the least important in pulmonary disease detection, we found that frequency domain features are very efficient in distinguishing COPD from other diseases. Furthermore, we have confirmed that cepstral features are an important tool in audio signal processing and that clustering algorithms could be used in pulmonary disease detection as well. It seems as though BMI does not significantly affect the

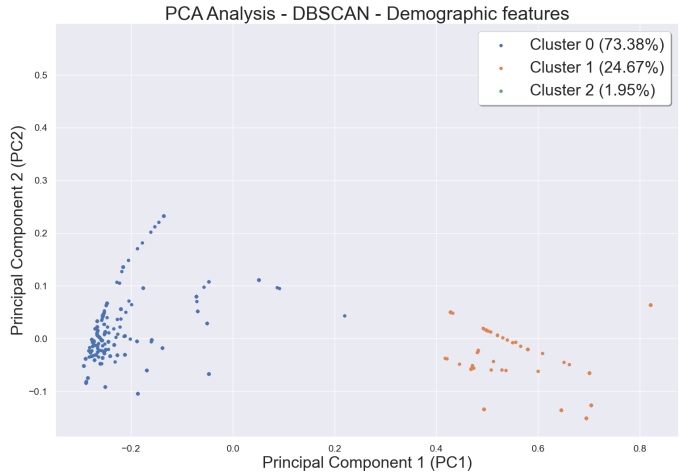

Fig. 14. Using PCA analysis, we can see that bronchiectasis and COPD are clearly separated by PC1. Higher PC1 increases the chances of being diagnosed with either COPD or bronchiectasis. PC1 is highly dependent on age of the patient, which confirms that elderly are more exposed to such diseases. Interestingly, BMI does not seem to affect the presence of the aforementioned diseases.

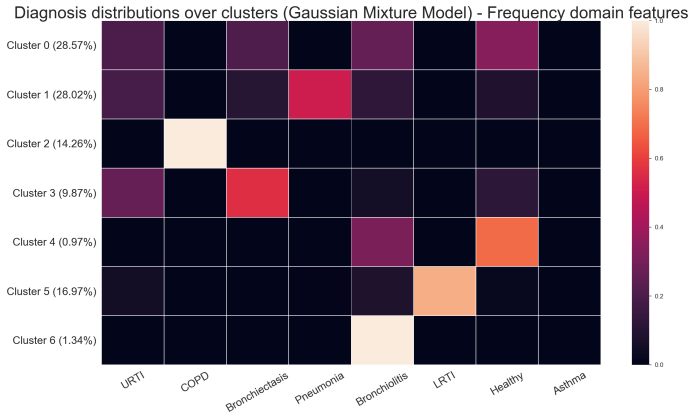

Fig. 15. In this figure we can see, how different diagnoses are distributed over clusters generated with GMM using frequency domain features. We can see that Cluster 2 contains all the patients diagnosed with COPD and that Cluster 5 contains the vast majority of patients diagnosed with LRTI. Moreover, we notice that Cluster 6 only contains bronchiolitis cases and that also Cluster 1 and 3 have entities, where a certain diagnosis, bronchiectasis or pneumonia, obviously prevails. GMM algorithm has proved to be very efficient in clustering pulmonary diseases.

presence of the considered diseases. Lastly, we also showed, how the chroma vector coincides with the presence of wheezes or diseases, where one of the symptoms is wheezing.

VI. Conclusion

In this article, we tried to determine which features are more important and which can be discarded in pulmonary disease detection. Our satisfactory and interesting results have shown that clustering can successfully cope with the aforementioned problem. Additionally, we have shown, how different audio

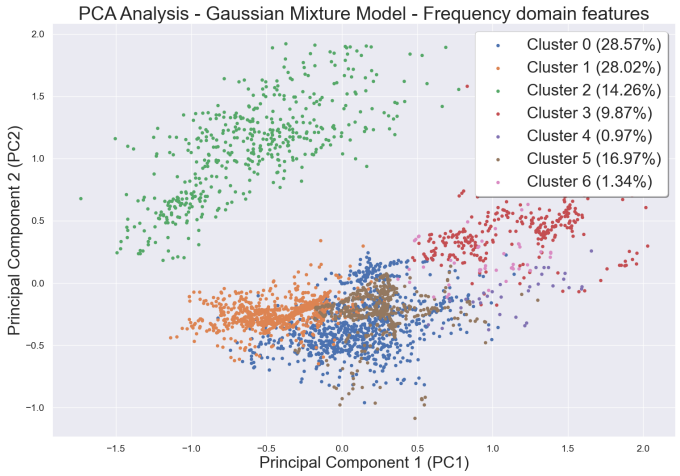

Fig. 16. In this figure, PC1 very nicely divides Cluster 2 that corresponds with COPD patients from the other clusters. PC1 is highly correlated with spectral rolloff and bandwidth and moreover, dependent on the chrom vector. These features correspond with the presence of both crackles and wheezes that are well-known symptoms of COPD.

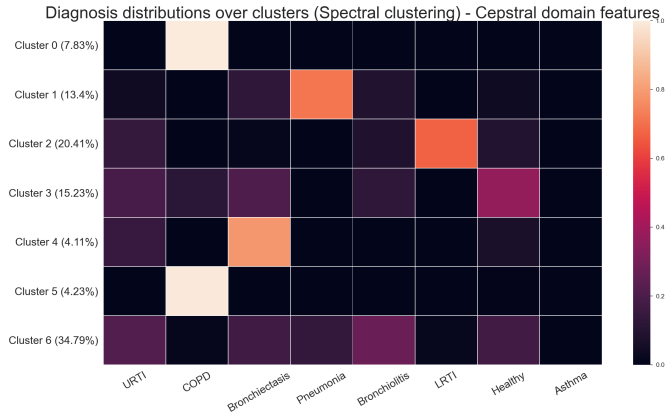

Fig. 17. The heatmap shows the efficiency of Spectral clustering with cepstral features. Clusters 0 and 5 are completely homogeneous. Using Spectral clustering, we separated the majority of patients diagnosed with COPD into these two clusters. Furthermore, the majority of cases of pneumonia are in Cluster 1. Most of patients that suffer from LRTI are assigned Cluster 2 and the majority of cases in Cluster 4 are diagnosed with bronchiectasis. Figure 18 explains how clusters are separated.

features corresponds with different pulmonary diseases and pointed out the most important connections. In future work, we could use the aforementioned connections to build a more accurrate classification model. All in all, we proved that unsupervised learning algorithms can be a strong tool in automatization of preventive healthcare and learning more information about pulmonary diseases.

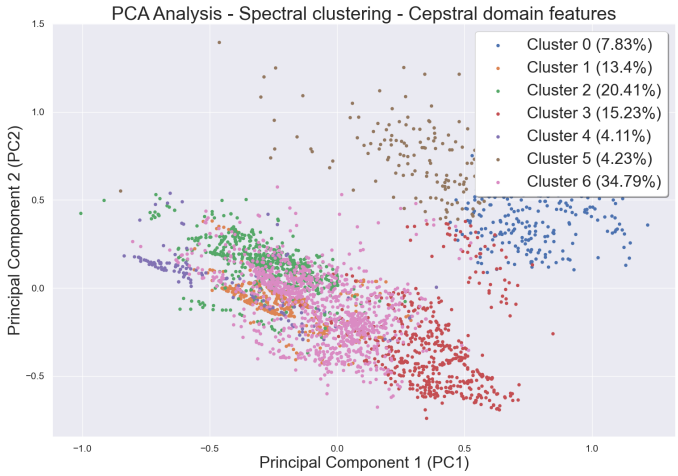

Fig. 18. In this case it is much harder to distinguish clusters and interpret them effectively. While PC2 takes into account coefficients of lower order, PC1 does exactly the opposite. We see that both COPD clusters have higher PC1 and PC2 scores. In PC2, the most important is the first order coefficient, which, as we already mentioned, describes spectral centroid and indicates the presence of wheezes. On the other hand PC1 is linearly related with zero-order coefficients, which describes the average power of the input signal. From this we conclude that COPD disease is represented with higher average power of the input signal and spectral centroid. Amongst the most common symptoms of COPD are wheezes and crackles. The presence of both highly increases the average power of the signal as shown in the PCA analysis.

References

[1] D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani, and R. R. Fletcher, "Application of semi-supervised deep learning to lung sound analysis," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 804–807.

[2] M. Aykanat, O. Kilic, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, 09 2017.

[3] O. Dumas, J. M. Mansbach, T. Jartti, K. Hasegawa, A. F. Sullivan, P. A. Piedra, and C. A. Camargo, "A clustering approach to identify severe bronchiolitis profiles in children," *Thorax*, vol. 71, no. 8, pp. 712–718, 2016.

[4] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[5] A. Oliveira and A. Marques, "Respiratory sounds in healthy people: A systematic review," *Respiratory Medicine*, 01 2014.

[6] H. Kiyokawa and H. Pasterkamp, "Volume-dependent variations of regional lung sound, amplitude, and phase," *Journal of applied physiology (Bethesda, Md. : 1985)*, vol. 93, pp. 1030–8, 09 2002.

[7] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[8] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[9] C. Chuan, S. Vasana, and A. Asaithambi, "Using wavelets and gaussian mixture models for audio classification," in *2012 IEEE International Symposium on Multimedia*, 2012, pp. 421–426.

[10] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Adv. Neural Inf. Process. Syst*, vol. 14, 04 2002.

[11] M. Ester, H. P. Kriegel, J. Sander, and X. Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," 12 1996.

[12] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC bioinformatics*, vol. 14, p. 106, 03 2013.

[13] L. Cen, F. Wu, Z. Yu, and F. Hu, *A Real-Time Speech Emotion Recognition System and its Application in Online Learning*, 12 2016, pp. 27–46.