FEATURE SELECTION FOR LANGUAGE INDEPENDENT SPEECH EMOTION RECOGNITION

CANSU ÖZKAN

Master's Thesis

Graduate School
İzmir University of Economics
İzmir
2022

FEATURE SELECTION FOR LANGUAGE INDEPENDENT SPEECH EMOTION RECOGNITION

CANSU ÖZKAN

A Thesis Submitted to

The Graduate School of İzmir University of Economics

Master’s Program in Computer Engineering

İzmir

2022

ABSTRACT

FEATURE SELECTION FOR LANGUAGE INDEPENDENT SPEECH EMOTION RECOGNITION

Özkan, Cansu

Master's Program in Computer Engineering

Advisor: Asst. Prof. Dr. Kaya Oğuz

July, 2022

Speech is the primary way of expressing ourselves. It is desired to extend this communication to computers. With the new developments in computer applications, machines are pretty much involved in our daily lives in a way that via personal assistants like Cortana or Siri. We want them to detect our commands and respond accordingly. Speech emotion recognition (SER) is a very popular and ongoing trend that enables machines to detect the human emotions. SER processes and classifies the speech signals and detects the embedded emotions in the speech. In general, the studies of SER focus on individual languages. Since the studies that focus on single and different languages are not very successful yet, and there are problems with the different accents of even English, providing a language independent SER is almost a necessity. This study focuses on finding the most informational features of speech to obtain the best recognition rates in language independent speech emotion recognition by analyzing how much the performance of the system changes according to the relations between the languages. Two classifiers which are Artificial Neural Networks (ANN) and AdaBoost was used to compare their performances. Berlin Database of Emotional Speech (EMO-DB), Toronto Emotional Speech Set (TESS), An Italian Emotional Speech Database (EMOVO), URDU Dataset, and KEIO University Japanese Emotional Speech Database (KEIO-ESD) were used as different language datasets. With the ANN classifier, 90.65 % recognition rate, and with the AdaBoost classifier, 72.60 % was obtained by using all datasets.

Keywords: Speech Emotion Recognition, Language Independent Speech Emotion Recognition, Artificial Neural Networks, AdaBoost Ensemble Learning, Emotion Specific Features

ÖZET

DİLDEN BAĞIMSIZ SESTEN DUYGU ANALİZİ İÇİN ÖZNİTELİK SEÇİMİ

Özkan, Cansu

Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Kaya Oğuz

Temmuz, 2022

Dil, kendimizi ifade etmemizi sağlayan birincil ve en önemli faktördür. Bu iletişim şeklini bilgisayar alanına da uyarlamak arzu edilen bir durumdur. Hızla gelişen bilişim sektöründe bilgisayarlar, makineler hayatımızın içinde oldukça fazla yer almaya başlamıştır. Örneğin, Cortana ve Siri gibi kişisel asistanlar çokça kullanılmakta, bunların kullanımı gitgide yaygınlaşmakta ve biz de bu kolaylıklara alışmaktayız. Bu kişisel asistanların bizim komutlarımızı anlamasını ve onlara göre tepki vermesini isteriz. Sesten duygu analizi, makinelerin insan duygularını anlamasını sağlayan oldukça popüler ve gelişmekte olan bir çalışma alanıdır. Bu sistem, ses sinyallerini çeşitli işlemlere tabi tutarak sesin içerdiği duyguları sınıflandırır. Literatürde sesten duygu analizi çalışmaları genellikle tek dili eğitip tek dilin içerdiği duyguları sınıflandırma üzerine odaklanmakta. Dünya üzerinde konuşulan çok sayıda dil olduğundan dilden bağımsız bir sistem oluşturmak bir ihtiyaç sayılabilir. Bu çalışma, sesin duygu ile ilgili en çok bilgi içeren özelliklerini bulmaya ve onları kullanarak dilden bağımsız bir sistem oluşturmaya odaklanmaktadır. Aynı zamanda bu sistemi oluşturmaya çalışırken dillerin birbiri ile olan yakınlığı ve benzerliğinin duyguları sınıflandırmadaki başarı oranına etkisini incelemektedir. Duyguların sınıflandırılması için Yapay Sinir Ağı ve AdaBoost teknikleri kullanılmıştır. Ayrıca, farklı dillerde veri setleri olarak Berlin (EMO-DB), Toronto (TESS), Italyan (EMOVO), Urdu (URDU) ve Japon (KEIO-ESD) veri setleri kullanılmıştır.

Anahtar Kelimeler: Sesten Duygu Analizi, Dilden Bağımsız Sesten Duygu Analizi, Yapay Sinir Ağı, AdaBoost Toplu Öğrenme , Duyguya Özgü Öznitelikler

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my supervisor Asst. Prof. Dr. Kaya OĞUZ for his guidance, advice, and mentorship. He was always supportive and understanding in the process of the study. He was also a guide for me while I was taking the steps of my academic career. I have learned a lot from him and I believe I will continue to learn.

I would like to thank my family for their support and everything they have done in my life. I especially thank my mother for her support, patience, motivation and love in my whole life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

For many years a series of studies have been performed about understanding the expressions of human emotions not only in computer science but also in psychology. Blanton (1915) indicated that the effects of emotions on the human voice are recognized by all people, and even animals can recognize main emotions such as love and anger. The tone of the voice is the oldest and the most universal shape of communication (Blanton, 1915). Now it is the time for computing machinery to understand these tones as well (Schuller, 2018).

It is known that the tone of the voice also communicates the emotions of the speaker. Emotions provide an additional layer of communication since the meaning and interpretation of the spoken words depend on the emotion recognized by the listener. Emotions are challenging to work with because they are implicitly subjective, and therefore it is very difficult to define them objectively. There are a variety of definitions of emotions and one of the widely accepted ones is the emotion theory by Paul Ekman (Ekman, 1971). In that theory, Ekman lists six emotions that are basic and culturally independent. These emotions are sadness, happiness, fear, anger, disgust, and surprise.

The next step in human computer interaction is to include the emotion in both directions; the computer should be able to understand the emotion and act accordingly, but it should also be able to simulate the emotion during speech synthesis. Speech commands have become a common sight in mobile devices as voice-controlled personal assistants such as Siri and Cortana, and they are used successfully to run several commands. It is desirable for humans who use these personal assistants to interact with them fluently just as in communicating with another person. It is only logical to extend the interaction to include the emotional state of the users. Speech Emotion Recognition (SER) is a collection of techniques that commits and categorizes speech signals to reach that goal. Emotions can be perceived differently so emotion recognition is a challenging work in spite of the fact that SER has various applications (Akçay and Oğuz, 2020).

Studies in SER have been an increasing trend and it will continue in the future. A lot of studies have been performed in several different areas on that topic such as human-computer interaction (Covie et al., 2001), mobile services (Yoon, Cho and Park, 2007), computer games (Szwoch and Szwoch, 2015), solving traffic problems (Tan et al.,

2022) and robots (Chen et al., 2020). There are a lot of applications of SER but this is a very challenging task because emotions are so subjective that even humans can not perceive them correctly.

A lot of datasets in different languages and novel methods have been produced to come up with successful predictions of emotions. Furthermore, feature extraction techniques from these produced datasets have been performed and these extracted features were used in many machine-learning algorithms. Features are a very essential part of all these SER systems because they are the core parts that give information about the emotion of speech.

In the basic structure of SER studies, the main step should be understanding the emotions deeply to improve the classification process. The most important part of the SER systems is features. Features are extracted to reach the most characteristics of the data which increases the prediction of correct emotion. The SER system needs a classifier which is a supervised learning structure because the speech samples will be given to the system with their emotion labels to train. This labeled speech data needs to be preprocessed before the training in order to get rid of unrelated information and clean the data. After all, the features are given to the classifier to train and predict the results.

There are some preprocessing techniques that are applied to data in SER systems such as framing, windowing, voice activity detection, normalization, noise reduction, and lastly feature selection and dimension reduction.

When we focus on the features, they can be categorized into four groups named prosodic, spectral, voice quality, and features based on the Teager energy operator. These features can also be supported with visual or linguistic features.

There are also several modalities that are not sufficient to recognize emotions on their own, but they support SER systems in order to increase classifier performance. Some of them are visual signals, word recognition, brain signals, and physiological signals.

At the end of all these preprocessing steps, utterances of emotions are given to the classifiers to classify emotions. Many machine learning algorithms can be used for classifying emotions in a way that gives the best results. There are traditional and deep learning-based algorithms used in SER although there is not any accepted algorithm to obtain the best results. Hidden Markov Models, Artificial Neural Networks, Support Vector Machines, Gaussian Mixture Model, Recurrent and Convolutional Neural Networks are some classifiers that are used in SER.

Most studies of SER use Ekman’s emotions and the datasets are labeled with them. The reason for using these emotions is, that the emotions need to be listened to, recognized, and labeled. So, it is very helpful to have a set of discrete emotions to select one of them in the labeling.

There is also another approach which is a dimensional emotional model. This model uses a small number of hidden dimensions for defining the character of emotions such as valence, control, power, and arousal. In that approach, the emotions are considered to be analogous to each other in a systematic way which means they are not completely independent (Russell and Mehrabian, 1977); (Watson, Clark and Tellegen, 1988). Generally, the two dimensional model is preferred. In that model, activation, arousal, or excitation are used in one dimension, while valence, evaluation, or appraisal are used in the other dimension. Arousal gives information about the strength of the emotion and valence gives if emotion is positive or negative. This method provides a quantitative approach which is good but we cannot be sure about where the values are going to be placed. This provides a metric but there is no criterion to measure which leads us to trust Ekman’s emotions theory.

One of the important components that the success of the emotion recognition studies depends on is the language. Since there are a lot of languages in the world, the SER system that works for only a couple of languages is not sufficient. Emotions can have some different expressions in different languages but the main emotions such as anger, sadness, happiness, and neutral are mostly understandable in every language. The success of the SER system depends on the language which it was trained. A system that is constructed to make good predictions for multiple languages would be a useful approach. Even the single language studies are not very successful yet and accents can also be problematic in SER studies. In order to perform such a system, it is necessary to extract common features that all languages have and also look into the similarities of languages.

1.1. An Overview of Language and Linguistics

Language is the primary way of expressing ourselves. As there are a lot of countries in the world, there are also a lot of languages that are spoken. All these languages are formed over time with the effects of culture, tradition, history, and rules of the area that they are spoken in. Consequently, every language has some major or minor differences. These differences can be seen as expressions, idioms, words, and emotions as well. From the emergence of languages in time, humanity, and languages themselves went through a lot of transformations. As a result of these transformations, languages have changed and are now separated from each other. Even today, as long as there are political and cultural changes, new languages become formal which come from the same language but with different specifics to different regions. These factors and the languages affect each other mutually and we can say that there is a two-way interaction between them. There are various language families in the world that some languages come from and have similarities. Comrie (1987) analyzes language families and tries to express which specifications separate languages from each other. It is emphasized in the study that deciding whether the varieties in speech should be considered as a different language or not is an important and difficult question in linguistics. Linguistics is a scientific study that studies human language. One advanced technique that is considered in the separation of languages is mutual intelligibility, which is a term used in linguistics. If two people who speak different but related languages can communicate and understand each other easily without making any preparation or working for this, these two languages are said to be mutually intelligible. However, this mutually intelligible term is not always successful because it depends on the traditional and geographical factors (Comrie, 1987).

It can be seen by anyone who can speak more than one language that some languages have clear similarities. Some words are very similar or even the same in writing but with minor pronunciation differences. For example, German is similar to English and Russian is similar to Polish. There are a lot of studies and hypotheses in history about the similarities and relatedness of languages. The important hypothesis about this situation is that the languages that have a common set of features are related to a common ancestor.

The study that examines the changes in languages over time and the relationships between them is called comparative linguistics or historical linguistics. Comparative linguistics studies the changes in languages and the genetic relationships among them. They try to answer the question of how the languages are classified into groups in the best way. They also introduce the memberships of languages to language families. The speakers who speak the languages that come from the same language family are members of the common speech community. With the language-shift process, speakers from different speech communities can also internalize languages from different language families (Dimmendaal, 2011).

1.2. The Relationships Among Languages

There are approximately 7151 living languages, which means at least one person speaks these languages as a first language. These living languages are divided into 142 different language families around the world (Ethnologue, 2022). A very important fact of historical linguistics is that the languages might have a considerable amount of relation, minimum amount of relation between them or they cannot be related at all. The languages which have relations are connected to language families according to that relation. Besides, the languages that come from a common language are also called daughter languages. If we give examples of daughter languages, Portuguese, Spanish, Catalan, French, Italian, and Romanian are all said to be daughter languages that come from Latin. On the other hand, Norwegian, Swedish, Danish, English, and, German are some other daughter languages called Germanic (Rowe and Levine, 2015).

1.3. The Family Tree Model

Language families can be described as a group of languages that comes from a common ancestral or parental language. It is called language family because family represents the tree form of languages in order to understand and compare relations of languages easily by simulating a family tree (Rowe and Levine, 2015).

Linguist August Schleicher worked on a family tree model and devised this relatedness of languages model in 1861. In that model, there is a main language on the top called proto language which is the parent language. There are many other ancestral and modern languages derived from this one. *Proto* means before and proto languages refer to ancestral languages. It is assumed that many languages were derived from these ancestral languages.

Proto-Indo-Europian is a big language family that 445 living languages are derived from according to the Ethnologue (2020). The most spoken languages belong to the Indo-European family. It contains the languages that are spoken in large areas in Europe. There are various subgroups in this family, two popular ones of them are Germanic and Italic. There are also other language families that are South-Asia and the Sino-Tibetan family.

The subgroups of the same language family are called daughter languages of the mother language. Also, the subtypes of the language family are sister languages. These family tree models of languages change over time, and this is called the regularity hypothesis. There is also another concept in this tree language family model which is the relatedness hypothesis. Relatedness hypothesis assumes that there are some similarities between languages that are derived from the same mother language.

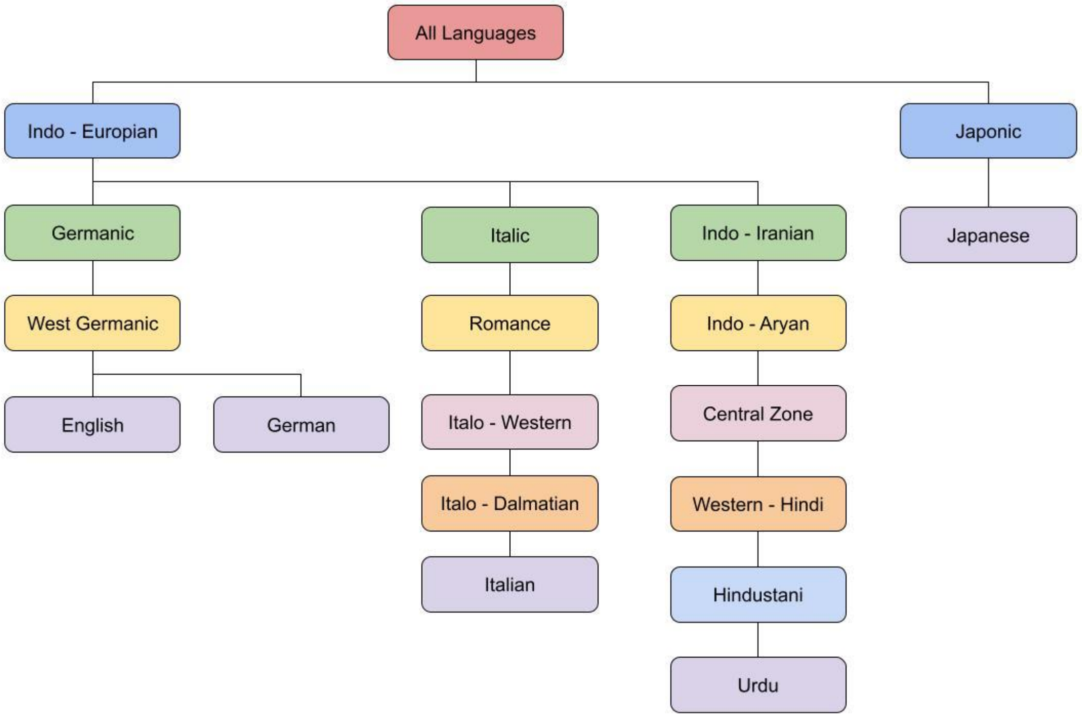

Figure 1. Language Families.

SER has been studied for years and has applications in many areas since it is desirable that machines understand human emotions and respond according to them. However, there are not too many studies about the language independent case of SER. It is important that a SER system works for language independent to provide universal usage without being restricted to one specific language. This study aims to construct a language independent SER system to decrease the effects of differences in languages while predicting embedded emotions from speech. This approach will eliminate the dependency on language, increase the recognition rates in cross language datasets and provide a wide area of usage. Since the SER studies that focus on single languages are not very successful yet and the accents can also be problematic, language independent SER aims to reduce these negative factors as much as possible.

The rest of the paper is as follows; related work is given under section 2, and section 3 gives a brief description of the data and the methods. After that, the proposed structure of the study is given. In section 5, the results of the study are placed and finally, section 6 is the conclusion part where the overall result is concluded.

CHAPTER 2: RELATED WORK

Speech emotion recognition (SER) is a methodology that tries to detect emotions from a speech by processing and classifying speech signals. SER systems classify emotions of given utterances to them. There are many machine learning algorithms in the studies that try to reach the best results in classifying emotions successfully. The classifiers used in SER studies include traditional classifiers and deep learning algorithms but there is no accepted algorithm that can be used. Common algorithms that are used in speech emotion recognition studies are Hidden Markov Models, Artificial Neural Networks, ensemble of classifiers, Support Vector Machines, Gaussian Mixture Model, Deep Learning based classifiers such as Recurrent and Convolutional Neural Networks, Machine learning techniques for classification enhancement and Multitask Learning. The general overview of these SER systems is described in the studies of Schuller (2018) and Akçay and Oğuz (2020).

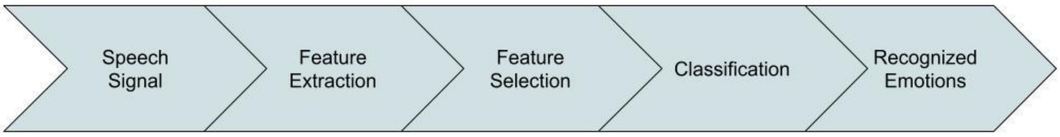

Figure 2. General structure of SER studies.

Schuller (2018), lists the traditional approach to speech emotion recognition under five titles. According to him, the way an engine recognizes emotion from speech can be investigated as modeling, annotation, audio features, textual features, and training and classification.

First of all, the recognition of emotion from speech needs to construct an appropriate model. After deciding on the model, the second point is to obtain the data to train and test the model for recognizing the emotion. Labeling the data correctly is a very important issue. The next step is to extract the necessary characteristic audio and textual features. This is an important step in SER systems because the features that best reflect the emotions yield the most successful results. While the sound features are important, they can be supported by what is being said to improve the success rate of recognition. Beyond the pronunciation of some words, what is said also has importance. Some SER systems make use of the words in the speech and relate them to specific emotions. When the SER systems are considered as a whole, the main characteristic part of SER systems can be seen as features. Finally, the last but not least part is the classifier. The speech features are given to a classifier to train and predict the emotions.

After the traditional approach, Schuller (2018) also describes the ongoing trends in SER systems under five topics. He defined the first trend as holistic speaker modeling which is about the robustness of voice production. A person whose voice is being recorded can be tired or have a cold besides being emotional. So, training a model that takes all this into account is difficult and new approaches provide more reliable ways for that. The second one is efficient data collection which is very problematic due to the lack of labeled useful data. He mentioned the third trend as follows. The semi-supervised learning algorithms can successfully label the data only if the engine was trained but the idea is the machine should label the previously unseen data. Data-learned features are the fourth topic which is the idea that the features can be learned from the data, and this yields better results for the classification. The last one is confidence measures which defend the necessity of providing the confidence measure of emotions in such a subjective task.

If we dive into more specific components of a SER system, there is a survey in which the datasets, features, and classifiers that are used in SER studies are described (Akçay and Oğuz, 2020).

The general structure of SER studies is start with a speech signal which is the part that will be given to a classifier to detect emotion. After that, the feature extraction part comes. It is a very important part because the core information will be given from these speech features. After obtaining the features, feature selection is performed to get the most informational features and reduce the dimension. The final but important part is the classifier. Selected features are given to a classifier to train and obtain the embedded emotions from speech.

Artificial Neural Networks are used commonly for many classification problems. In the basic structure of ANN, there is an input layer with one or more hidden layers and there is an output layer. There are nodes in the layers whose number changes according to data and the labeled classes, but the hidden layer can have many nodes as needed. Layers are connected to the weights which are initially randomly chosen. The chosen samples are loaded to the input layer and sent to the next layer and at the output layer, the update of weights is performed with a backpropagation algorithm. In the end, it is expected from the weight to classify the new data (Akçay and Oğuz, 2020).

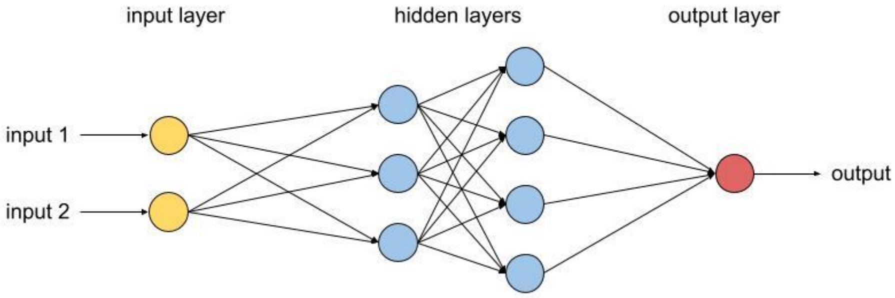

Figure 3. Layers of ANN classifier.

There are not many studies on language-independent or cross-corpus cases in SER systems. It is very important to increase recognition rates in not one language but also in different languages as well. There are several approaches to this case such as training classifiers with more than one language speech samples or training the classifier with a specific language and testing with different languages.

Latif et al. (2018) studied cross-corpus speech emotion recognition. They used German, English, and Italian for training the SVM, Logistic Regression, and Random Forests classifiers and they performed testing in the Urdu language. SVM outperformed other classifiers and they reached an 83.40% recognition rate from the SVM classifier in the Urdu language.

There is a study in which a cross-corpus language corpora consisting of six emotions was used, and Convolutional Neural Networks was used as a classifier. They trained the classifier in two ways; one is training with only an individual language and the other one is training with multiple languages. According to the results that they obtained, training in multiple languages was more successful than the other case. (Parry et al., 2019).

In the literature, ANN was used by Nicholson, Takahashi and Nakatsu (1999) as a classifier for a speaker and context-independent system. Speech power, pitch, LPC feature, and its delta were used as parameters. They created a subnetwork for each emotion and each neural network gave an output that yielded the likelihood of whether the utterance belonged to an emotion or not. They reached a 50% average recognition rate with this approach.

Another study that uses ANN is the study of Petrushin (1999) which was developed for call-centers. The study uses ANN, ensemble of ANN, and the k-nearest neighbor as classifiers. The features used were pitch, energy, the speaking rate, and the first and second formants. With the k-nearest neighbor classifier, they obtained 55% recognition rate, with the ANN classifier they obtained 65%, and finally with the ensemble of ANN they obtained 70% success rate.

Mao, Chen and Fu (2009) came up with a hybrid approach that combines Hidden Markov Models (HMM) and ANN to recognize emotions. First, they used HMM to get the likelihood probabilities of speech utterances. Then, they divided the speech into segments and the distortions and likelihood probabilities were given to the ANN as input. The recognition rate of isolated HMM was 75% while the hybrid classifier which included ANN scored 81.7%.

Rajisha, Sunija and Riyas (2016) used the Malayam language to compare the recognition rates of ANN and SVM classifiers. The MFCC, Short Time Energy, and pitch were used as features. They gave the extracted features to classifiers. The success of the SVM classifier was 78.2% and the success of ANN was 88.4%.

Ke et al. (2018) constructed SVM and ANN model classifiers to recognize emotions. They also investigated the effects of feature reduction in the system. Feature reduction improved the success of both classifiers and according to the results SVM was slightly better than the ANN with a rate of 46.67%, the success of ANN was also 45.83%.

Ensemble learning is a technique that combines different machine learning models to increase the recognition of emotions. It can be used as a classifier for SER studies too. There is a study that uses three different classifiers which are multiclass support vector machines, AdaBoost, and random forests to compare them. They used 14 features with seven basic emotions. The best results were achieved with the AdaBoost classifier which is 87.5% (Noroozi et al., 2017).

Another study that uses AdaBoost is that Pan, Tao and Li (2011). They compare SVM and AdaBoost algorithms as classifiers. They extracted features and selected them for dimensionality reduction. According to the results of their experiments, AdaBoost slightly outperformed the SVM.

It seems very useful to investigate each emotion individually in order to distinguish them from each other more successfully. With this determination, it was studied by focusing on each emotion individually and determining emotion-specific features (Özkan and Oğuz, 2021). The structure of speech emotion recognition studies is generally similar. However, the underlying structure shows some differences according to the characteristics of the problem and the classifier may change as well.

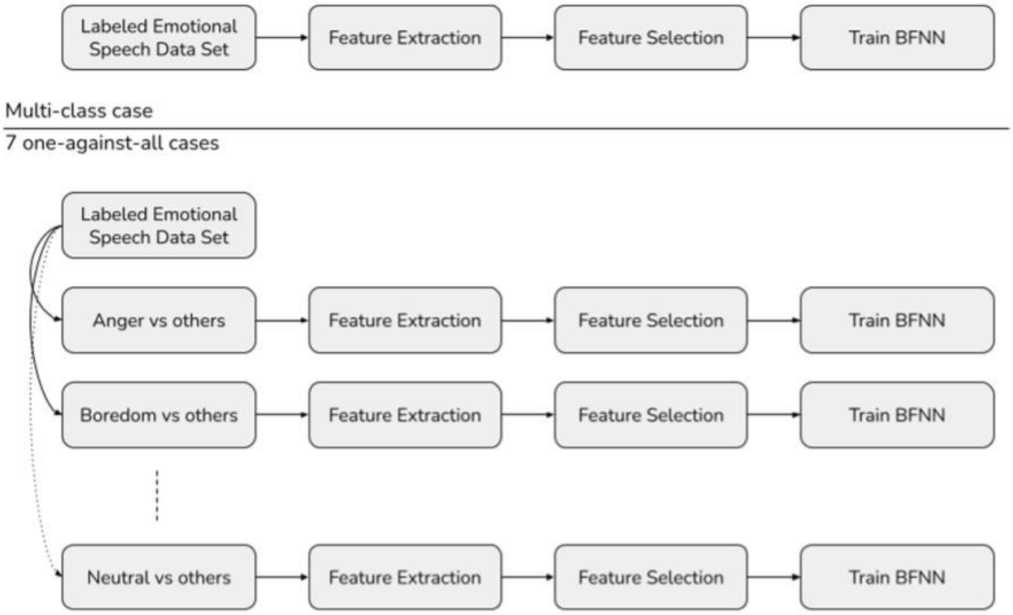

Figure 4. Detailed structure of emotion specific study (Source: Özkan and Oğuz, 2021).

In order to examine emotion specific features, it can either be examined how the one specific emotion is separated from other emotions or how the two emotions are separated from each other. There are two cases where multi-class and one-against-all classes are needed. When binary classification algorithms are used for multi-class problems that have N classes, there should be N number of one-against-all (OAA) or N(N-1)/2 number of one-against-one binary problems. So, to have different features for different emotions the binary one-against-all classification was used, and a standard backpropagation feedforward neural network was preferred as a classification to be able to compare the results of OAA with the multi-class case approach.

For the determination of emotion specific features, EMO-DB was used. Seven one-against-all cases for each emotion in the database and one multi-class case have the same processes. Fundamental frequency and Mel frequency cepstral coefficients (MFCC) which are the most commonly used features in speech emotion recognition studies were extracted as features. The MFCC coefficients were set to 20 because the best results were yielded with it. These two features were not used as they are, the mean, minimum, maximum values, standard deviation, kurtosis, skewness, first and second derivatives of MFCC were used in addition to the mean, minimum, maximum values, and standard deviation of the F0 feature.

Feature selection was performed as it reduces the number of dimensions and increases the classifier performance. For the feature selection process, mostly preferred standard forward selection was used. The structure is as follows; the process starts with the empty set of features. Each feature was tested one by one, and the error was calculated for each of them. The first feature that has the minimum cross-entropy value according to the neural net results was added to the feature set. After that addition, the remaining features were tried one by one with the selected one. Again, the feature that has the lowest cross entropy value when trained with the selected feature which is selected in the previous step was added to the feature set. This step was repeated. It continues in that fashion until there is no improvement in the cross-entropy value and finally the resulting set becomes the selected subset of features.

In spite of the fact that deep neural networks are very popular nowadays, a shallow backpropagation feedforward neural network (BFNN) was used in the proposed method because it is faster to train, needs fewer data, and is also suitable for speech emotion recognition studies in order to obtain high classification rates. The foundation of deep neural networks is that it is possible to design them using more than one hidden layer. It is also possible that the number of neurons placed in the hidden layer can also be changed. In the proposed structure the number of the hidden layers is one and the number of neurons in that layer changes to 5, 10, 15, and 20. The reason for these changes and variations is to find the best possible configuration of features for each emotion. During the feature selection process, different configurations of the number of neurons were tried. The whole process was repeated for each neuron number as 5, 10, 15, and 20. Cross-entropy value was again used to determine which neuron number was suitable for specific emotion and multiclass case. The error rate and cross-entropy values were used to construct a selected subset of features and obtain the success of classification (Özkan and Oğuz, 2020).

CHAPTER 3: DATA AND METHODS

3.1. Databases

Databases are primary components of speech emotion recognition studies. Whole classification processes depend on the emotion labels in these databases. The quality of the database used in the speech emotion recognition process has great importance as it directly affects recognition success. A low-quality, incomplete or incorrect data may lead the study to failure even if the proposed structure is well-designed. That's why the data should be collected carefully and the database which is going to be used should also be chosen meticulously.

In the literature, databases of speech emotion recognition studies can be classified into three categories that are known as acted (simulated), elicited (induced), and natural speech emotion databases (Akçay and Oğuz, 2020).

Acted speech databases are created in sound-proof environments. The utterances are vocalized by the professional actors. Creating speech databases with this method is easy to record and study. However, when acted databases are compared with natural human speech, they tend to sound more exaggerated. Even if the classification results of such studies have high scores, their performance falls when they are adapted to natural human speech (Akçay and Oğuz, 2020).

In the creation of elicited speech databases, speakers are in a simulated emotional situation and the utterances are recorded in a condition of these various stimulated emotions. When it is compared to acted databases, elicited ones are closer to natural human speech.

The third type of speech database is natural speech which is acquired from talk shows, call centers, radio programs, and other similar sources. In spite of the fact that natural speech databases seem preferable due to their convenience of adapting studies to daily life, sometimes it can be difficult to obtain data from them to process because of ethical issues.

The most used emotions in the speech emotion recognition studies in the literature are chosen from 'basic emotions' that are anger, happiness, sadness, fear, disgust, and surprise (Murray and Arnott, 1993).

3.1.1. Berlin Database of Emotional Speech (EMO-DB)

Berlin Database of Emotional Speech (EMO-DB) is one of the databases used in this study. Furthermore, it is one of the most used databases in speech emotion recognition studies. It is a publicly available acted database that is recorded by professional actors. The database contains five male and five female speakers reading ten sentences with the emotions of anger, boredom, disgust, fear, happiness, sadness, and neutral. 10 different daily conversation sentences are vocalized by these speakers. For the labeling part, sentences were listened to by the 20 people, and recognition rates were determined according to peoples' guesses. (Burghardt et al., 1999).

3.1.2. Toronto Emotional Speech Set (TESS)

This dataset is created in the Northwestern University Auditory Test No 6. Two actresses from Toronto, who speak English as their first language, vocalized the 200 sets of words with the emotions of anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The ages of the speakers are 26 and 64 and the recordings were named as older and younger talkers with the emotion that they vocalize (Dupuis and Pichora-Fuller, 2010).

3.1.3. An Italian Emotional Speech Database (EMOVO)

Another popular and commonly used dataset in speech emotion recognition studies is the EMOVO dataset. EMOVO is an acted dataset that is the first dataset in the Italian language. It was vocalized by 6 expert people with 14 sentences. The emotions are based on six basic states which are disgust, fear, anger, joy, surprise, sadness, and neutral. These emotions were chosen because they are considered well-known big-six emotions in most speech studies (Cowie et al., 2001). The speech samples were labeled by the people who create the dataset in the way that the first part indicates the emotional state, the second part indicates the actor/actress and the third one indicates the type of sentence. There are four types of sentences: long, short, nonsense, and questions. (Constantini et al., 2014).

3.1.4. URDU Dataset

URDU is the Urdu language dataset that is created with the speeches from URDU talk shows on YouTube. There are 27 male and 11 female speakers chosen randomly in order to create a dataset with emotions of anger, happiness, sadness, and neutral. The dataset contains 40 utterances. Labeling is performed by the people who create this dataset as speaker information, file information, and emotion respectively (Latif et al., 2018).

3.1.5. KEIO University Japanese Emotional Speech Database (KEIO-ESD)

It is the dataset that was created at KEIO University (now at Aichi University of Technology) which has two types of speech data. The first one is a set of words that are spoken by a male Japanese man and the second one is the synthesized speech designed by a system that is trained using human speech. The words were voiced by a Japanese man with 47 different emotions. The naming of the files was designed in a way that contains information about the number of files, emotion, and the duration of the speech. In the synthesized set, there are 12 emotions that the speeches are created with. The synthesized set was not used in that study because they cannot give the same intonations as humans give (Moriyama, 2011).

3.2. Methods

There is a classification system in speech emotion recognition studies. This classification system labels each speech utterance as a specific emotion. It sets each utterance to a proper emotion according to the features extracted from speech. There are different numbers of proper classifiers available for speech emotion recognition studies to obtain good results. Although many machine learning algorithms including traditional classifiers and deep learning algorithms are used, there is no generally accepted machine-learning algorithm just as in any other complex problem. In most cases, the classifiers are chosen considering the success of the previous studies.

Generally, classification algorithms are used in speech emotion recognition systems which require input, output, and a function that maps input to output. There is a need for labeled data for the learning algorithm to predict classes of inputs. This labeled data identifies the data samples and their classes. The data is divided into two parts, one part is used for training the learning algorithm, then the other part is used for testing.

3.2.1. Artificial Neural Networks

With the rapid developments in modern computers, machines are now capable of various tasks that are relatively easy for humans. In some problems where more experience allows a chance of getting more sensitive and successful results, there is a need for computer solutions. In this case, neural networks take the stage. Many scientists use neural networks for different reasons. For example, signal processing, artificial intelligence, pattern recognition, classifying patterns, constraint optimization problems, and many others (Fausett, 1993).

Artificial neural networks are information processing systems whose formation is inspired by biological neural networks. In the architecture of neural nets, the information is processed in neurons which are a large number of simple elements. The information processed in these neurons is passed between all neurons with connection links that have associated weights. Weights have required information to solve problems on the net. The weight of a link is multiplied by the transmitted signal to form the activation function of the neuron. After that, the sum of weighted input signals called net input is created to determine the output signal.

A neural network is simply composed of three main parts. The first one is architecture, how the connections between neurons are placed in the net. The second one is called training or learning, which is the mechanism that determines the weights of connections, and the last one is its activation function which is formed by the multiplication of weights and inputs of neurons. Each neuron has its own activation function that it sends as a signal to other neurons.

If we think about a scenario, there is a neuron B and neurons A1, A2, and A3. These A neurons send their input signal to B. The activation functions of these neurons are a_1, a_2, and a_3, and the weights of these connected links are w_1, w_2, and w_3 respectively. The net input of neuron B is going to be

$$b\_net = a_1w_1 + a_2w_2 + a_3w_3$$

and the activation function of neuron B is going to be a function of its net input.

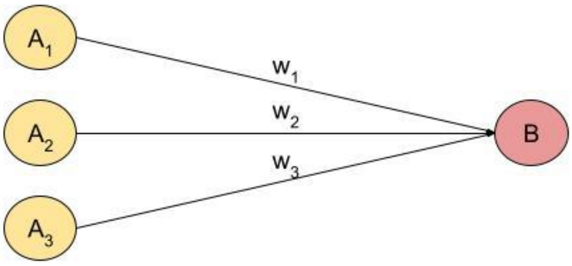

Figure 5. Simple artificial neuron.

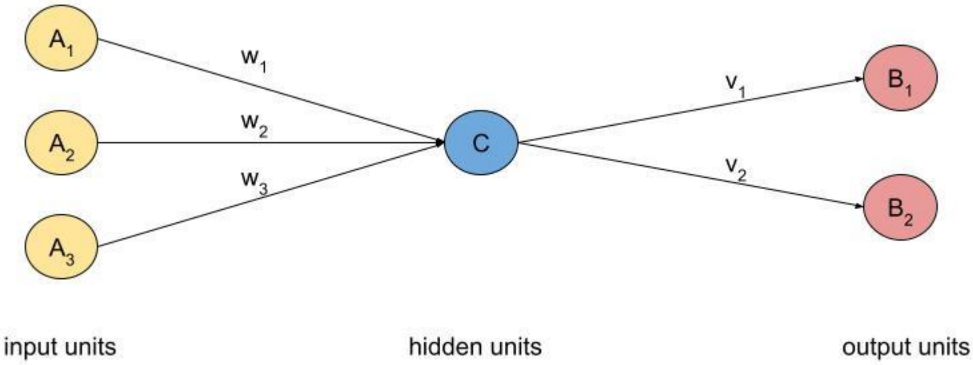

Figure 6. Simple neural network structure.

3.2.1.1. Architecture of Neural Networks

The net architecture can be identified as the placement of the neurons and the connections between the neurons in the net. In general, neurons in the nets are represented as layers in which all neurons in the same layer show the same behaviors. The behavior of these neurons is identified by the activation functions that are usually the same in each layer.

In the literature, neural nets are considered as a single layer or multi-layer. A single layer neural net has one input layer that gets signals from outside and one output layer where we can see the response of the net from. These layers are connected with one weighted connection link. On the other hand, in the multilayer net, there are multiple layers of nodes between the input and output layers, and this creates multiple weighted links between them. When these two types of layered architecture were compared, though a multilayer net is more suitable than a single layer net for complicated problems, training a multilayer net can be more difficult in some cases.

Pattern classification problem is one of the simplest problems that can be solved with neural networks. In the process of pattern classification, there are particular classes or categories that each pattern belongs to, and the neural net knows which input vector belongs to which class.

At the end of the training, the neural net predicts each input if it belongs to a class or not. In a simplest version of this classification, generally, a single class is used but in extensions of the method, there are several classes that the input vector can belong to. For this case, each class needs an output unit.

Pattern classification architecture consists of multiple input units and a single output unit as in Figure 7. This considers the membership of vectors in a class.

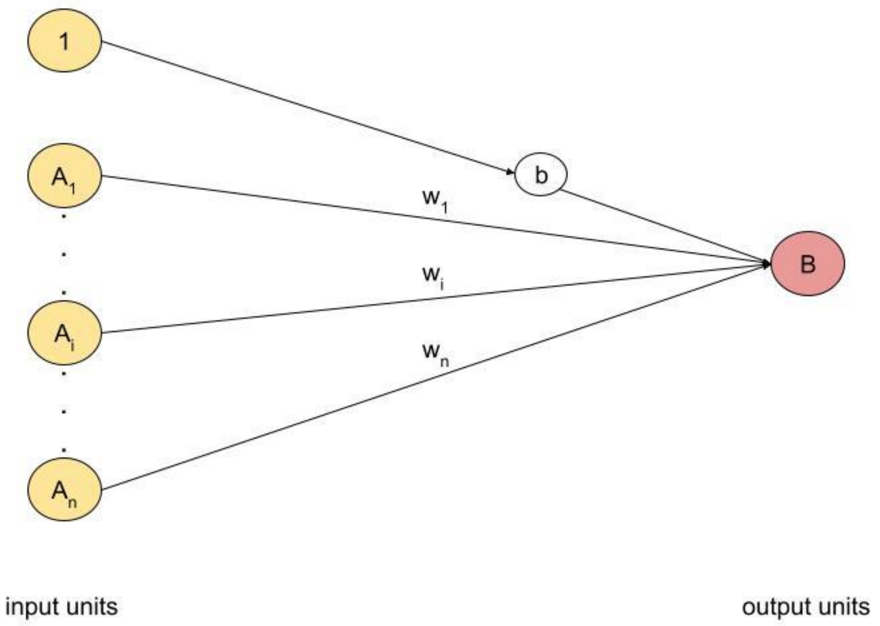

Figure 7. Single layer net for classification.

A bias can be included, and it behaves as a weight on a unit connection. The activation of this unit is always 1. In such a case, the activation function is

$$\text{f(net)} = \begin{cases} +1 & if\ net\ \geq 0; \\ -1 & if\ net\ < 0; \end{cases}$$

where

$$\text{net} = b + \sum_i x_i w_i\ .$$

If the bias is not used and fixed threshold θ is used,

$$\text{f(net)} = \begin{cases} +1 & if\ net\ \geq 0; \\ -1 & if\ net\ < 0; \end{cases}$$

where

$$\text{net} = \sum_i x_i w_i\ .$$

The aim of the neural nets is to train the net and get responses with the desired classification. There are some issues that are common in all neural nets. The response of the net when the input vector belongs to a class is ‘yes’ whose value is 1, when it does not belong to a class the response is ‘no’ whose value is -1. Because it is desired to get one of the two responses, the activation function can be considered as step function and the net input to the output is

$$y\ =\ b + \sum_i x_i w_i$$

The decision boundary between the region when $y\ > 0$ and $y < 0$ is determined as

$$b + \sum_i x_i w_i \;= 0.$$

The problem can be said to be linearly separable if the +1 responses lie on one side of the decision boundary and -1 responses lie on the other side of the boundary.

The Hebb rule is the simplest and the earliest rule for neural networks which suggests that the learning process occurs by modifying the weights between the neurons. When two neurons that are interconnected are both firing at the same time, the weight between these two neurons is increased and this provides learning. The method suggests that learning occurs when firing neurons at the same time, but stronger learning form also occurs when the condition that both neurons do not fire at the same time.

The most widely known version of neural nets is the perceptron learning rule which is stronger than Hebb rule. It has an iterative learning procedure. If the necessary conditions are provided, it converges the weights that the neural net produces the output values correctly with. Different types of perceptrons are explained in Rosenblatt (1958).

There are three layers in an original perceptron which are sensory units, associator units, and response units. This structure of the perceptron is similar to the retina. A single specific perceptron uses binary activation for sensor and associator units and it also uses for response units the activation of +1, 0, or -1.

A binary step function that has an arbitrary threshold is used as an activation function for each associator and a binary signal is sent from the associator units to the output unit. If the output of the perceptron is y = *f*(y), the activation function is

$$f(y) = \begin{cases} 1 & if\ y > \theta \\ 0 & if -\theta \leq \mathrm{y} \leq \theta \\ -1 & if\ y < -\theta \end{cases}$$

According to the perceptron learning rule, the weights that go from associator units to the response unit are set. A response of the output unit is calculated for each training input and then it is determined if there is an error in the pattern. If there is an error occurred, the formula of the weight is changed as

$$w_i(\text{new}) = w_i(\text{old}) + \alpha t x_i.$$

where t is the target value +1 or -1, α is the learning rate.

In the simple perceptron, the associator unit’s output, which is a binary vector, is accepted as input signal to the output unit. However, according to the perceptron learning rule, binary input is not necessary but the weights in the associator unit are important. So, the consideration is limited to the single layer portion of the neural net. The aim of the neural net is to classify each input if it belongs to a particular class or not. If the input belongs to a particular class, the response of the net is +1, otherwise, the response is -1. The architecture is shown in Figure 8.

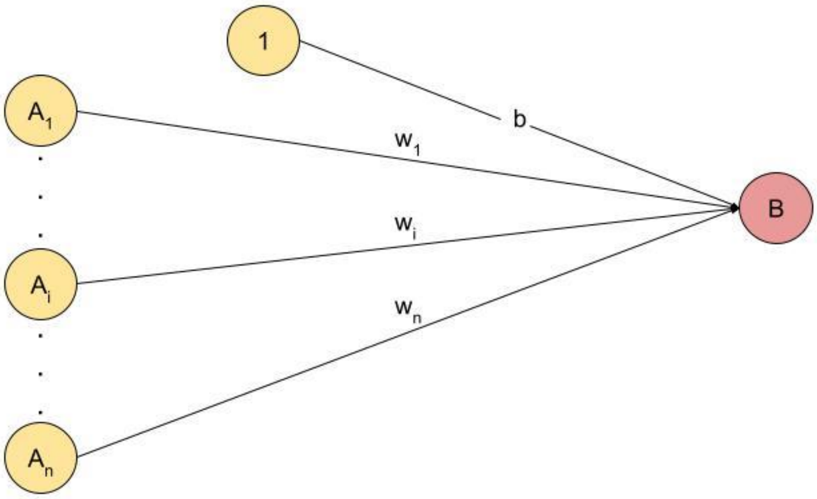

Figure 8. Perceptron for single classification.

ADALINE is another early version of neural nets. The bipolar activations which are 1 or -1 used in that version for the inputs and output. The weights in that method are adjustable and there is also a bias that is treated like adjustable weights too.

Generally, the delta rule which is also known as the least mean squares or Widrow-Hoff rule used while training the net. That rule can be used with a single layer net and several output units. ADALINE is an exclusive version that has one output unit. There is a mean squared error between the activation and the target value, and that error is minimized by the learning rule so that the net can continue to learn all training models. At the end of the training process, threshold function is applied to the neural net with the aim of obtaining the activation. If the input of the net is greater than or equal to 0, the activation is set to 1 but otherwise, it is set to -1. A problem that is linearly separable can be easily modeled by the ADALINE.

The basic architecture of the model is shown in Figure 9.

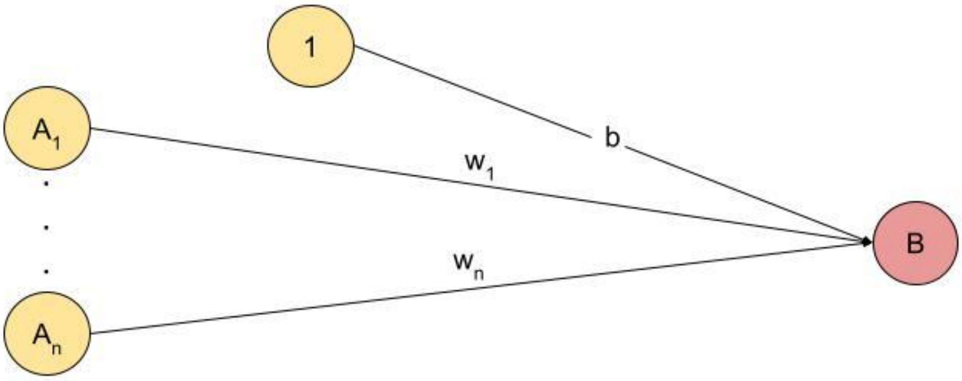

Figure 9. ADALINE architecture.

There is a single neuron, and it receives input from some other units. It receives a signal whose value is always +1 in order to use that signal for training other weights. If the model combines in a way that the outputs from some of the input units become input for other input units, the net is turned into a multilayer net which is called MADALINE.

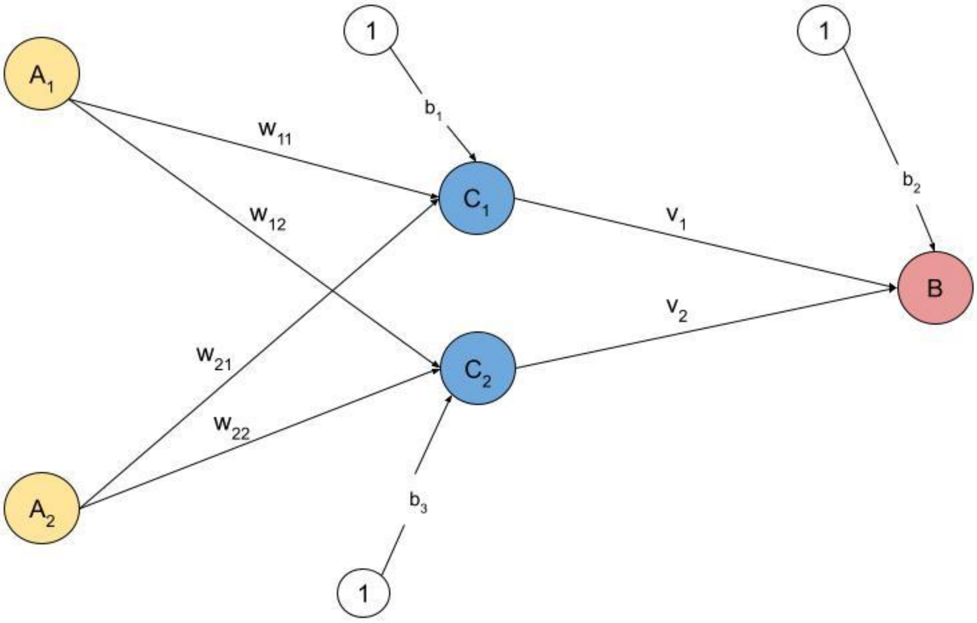

Figure 10. MADALINE architecture with two ADALINEs.

3.2.2. Backpropagation Feedforward Neural Network

With the limitations of single-layer neural networks, the popularity of neural nets declined in the early days and according to that situation, the popularity of neural nets decreased (Fausett, 1993). Furthermore, this led scientists to discover multilayer neural nets and use them in the solutions of various problems where the given set of inputs is mapped to a set of target outputs. The main purpose is to respond to inputs used for training and get answers for other inputs similar to that training ones.

The general structure of the backpropagation feedforward neural nets consists of three main parts. In the first part, the feedforward of the input training pattern is performed. As a second part, the calculation of the error and the backpropagation of it is done and finally, the last part is arranging the weights again.

In the problem of learning mappings, a single layer net may not be sufficient enough while a multilayer net can easily learn the complex mappings. One single hidden layer is enough for some applications but there is also a chance to gain an advantage from using more than one hidden layer in some problems.

The basic architecture of a multilayer neural net that has one layer of hidden units is shown in Figure 11 where C is the layer of hidden units and B is the output units. These biases behave like weights which are always 1. The information flow is shown in the figure, in the backpropagation phase, the signals are sent back in to reverse direction.

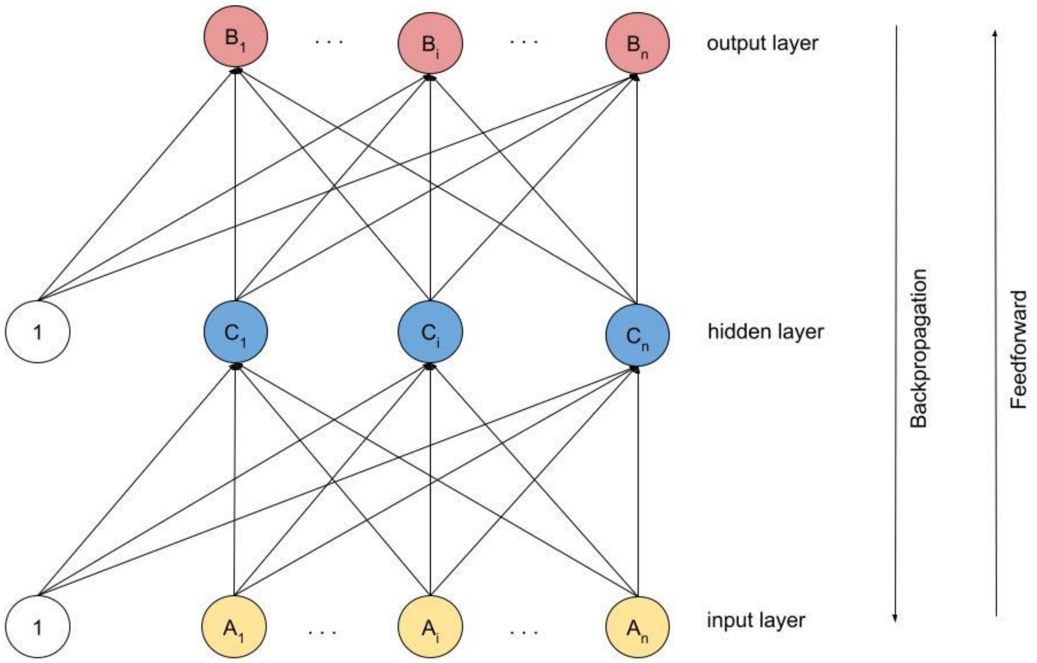

Figure 11. Structure of backpropagation feedforward neural network.

In the feedforward part, input signals are sent to each input unit (A_i) in the net. Then, input units broadcast these received signals to each hidden unit (C_1,..., C_p) in order to calculate their activations and send them to each output unit. After each output unit (B_k) receives signals from hidden units, they calculate their activation and create the response of the net to the given input pattern. The output unit compares the calculated activation and its target value to specify the error. According to that error, δ_k (k = 1,..., m) is computed to use the process of sending the error back to all units in the previous layer, and weights between the output and the hidden layer are updated.

3.3. Ensemble Learning

As the studies in machine learning increased rapidly over time and new techniques developed, the traditional single machine learning algorithms started to become insufficient and were not satisfactory. Consequently, different algorithms have been developed in order to increase the success of the machine learning algorithms and obtain higher recognition rates. Ensemble learning is a technique that combines different machine learning algorithms to improve the success of classification. These algorithms are performed together so that better performance can be achieved than a case in which a single algorithm is used.

The main structure of ensemble learning techniques is combining different machine learning algorithms and performing them together. However, the combining can be done in several ways which are mixing training data, mixing combinations, and mixing models.

3.3.1. Mixing Training Data

In biology, genetic diversity has a big importance for species to survive. If the genetic diversity of species is less, this species is more likely to disappear. Even if they are experts in one area, when they are faced with difficult conditions in another area that they are not used to, the chance of adaptation is very low. They are vulnerable to unexpected conditions such as natural disasters or some diseases. In order to avoid this kind of situation, the species are divided into groups, and these groups are left to different environments that have different conditions so that they get used to difficult situations and try to survive in different areas. As a result, they have genetic diversity and become more durable (Kumar and Jain, 2020).

If we modify the relationship of this evolution and genetic diversity to machine learning algorithms, we divide the training data into subgroups, and instead of trying classifiers on one single and large training data, we can train separate classifiers on each subgroup of that. In this way, at least some subgroups of the data can perform well in case the distribution of the whole data may not look like the real-world testing data. After getting the output of each classifier, the results are combined at the end. This technique is called mixing training data.

In this approach, the decision tree is used. It is a top-down approach in which a decision is represented by each node based on one or more parameters. According to the parameters that are used in the decision tree, the nodes are traversed through until a suitable depth. When the depth of a decision tree is increased, the performance of the training data set will be better. However, the greater depth of a decision tree is an important factor that needs to be considered. In order to get more accurate results, it is necessary to increase the depth of the tree which causes an increase in overfitting in the training dataset, and in the result of this, the success of the test dataset will decrease.

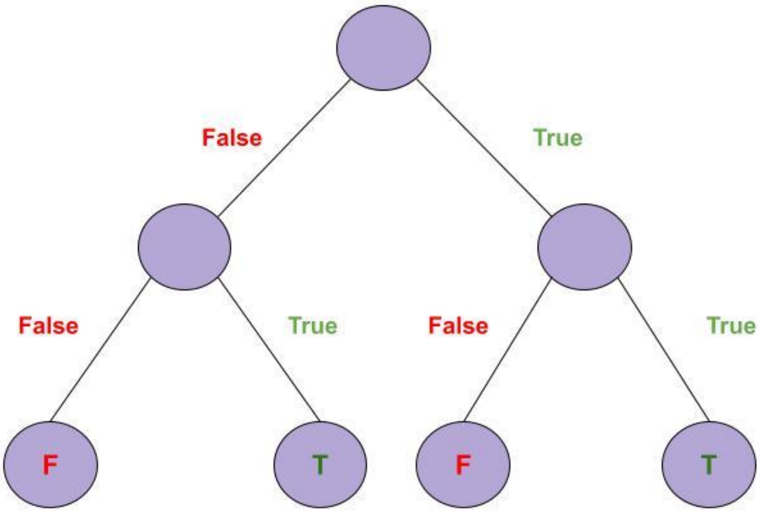

Figure 12. A simple decision tree.

One way to solve the problem between increasing depth and overfitting relation is using more than one decision tree instead of one single tree in which each tree will have a different subset of training data. This leads to the random forest which consists of multiple decision trees, and different sets of training data are used for training each tree. With using random forest, both shallow decision trees and better accuracy are reached at the same time.

As it is mentioned, in the mixing training data we divide the dataset into several groups. If we look at how we do this process, there are two different sampling methods that we faced which are sampling with replacement and sampling without replacement.

In sampling with replacement, the data is divided into groups in a way that the groups do not need to be disjointed which means some elements of the data can appear in more than one subgroup. There is no distinctly separated data. On the other hand, sampling without replacement is the opposite of this. The elements in the dataset are divided into two or more subgroups that each element can only be placed in exactly one group. More than one group cannot contain the same element. There is a distinct separation between the elements of subgroups.

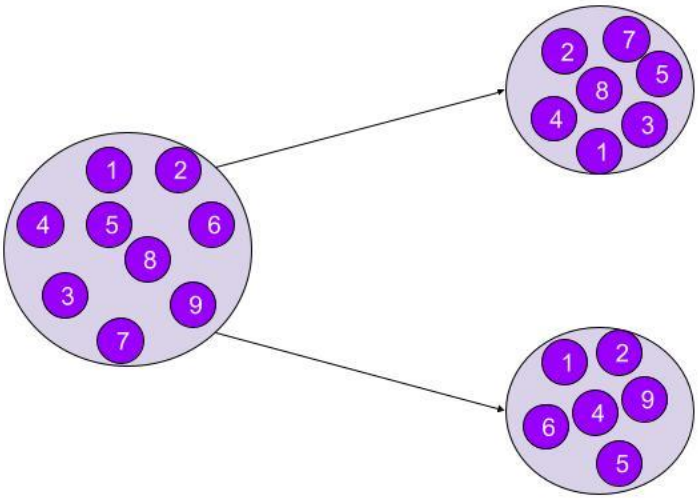

Figure 13. Sampling with replacement.

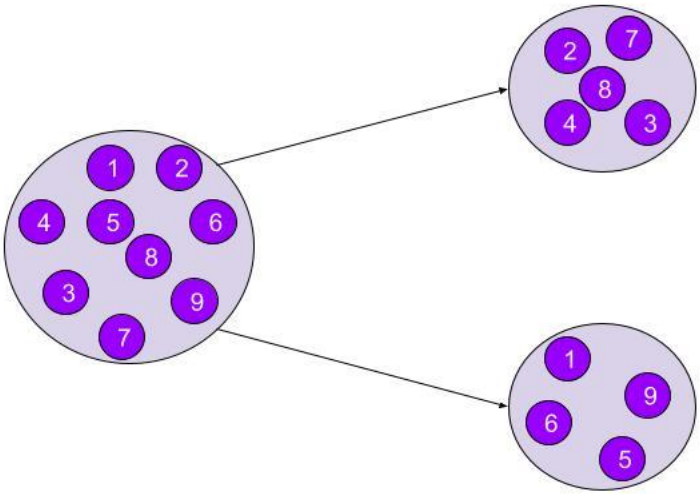

Figure 14. Sampling without replacement.

One of the most important techniques of ensemble learning is bagging which stands for bootstrap aggregation. In this type of ensemble learning technique, the dataset is divided into n subgroups with replacement. Then, the divided subgroups of complete data are trained individually with relatively weak different machine learning classifiers. At the end of these steps, the output of the models is brought together with a voting mechanism and the final result is obtained. The reason for using sampling with replacement technique is that it is provided that each machine learning classifier has random samples. This gives a chance to get better results than individual models.

The division of the data samples into subgroups is not so trivial. In some cases, good results can be obtained in training and the test datasets in the experiments but that is not how it works when it comes to the real world. The reason is distribution of samples that are used in the experiments may not have all different kinds of samples that can be faced in the real world. That’s why the division into subgroups should be performed effectively and that leads us to cross-validation. k-fold cross-validation is the technique that is mostly used and very popular in machine learning studies. The main structure of this technique is iteratively dividing validation and training data using sampling without replacement. k is the number of portions that data is divided into. The dataset is divided into k different parts and k-1 / k of the data is used as the training part and 1 / k of the data is used as testing. This process continues k times so that each testing and training data changes and finally the overall success is calculated.

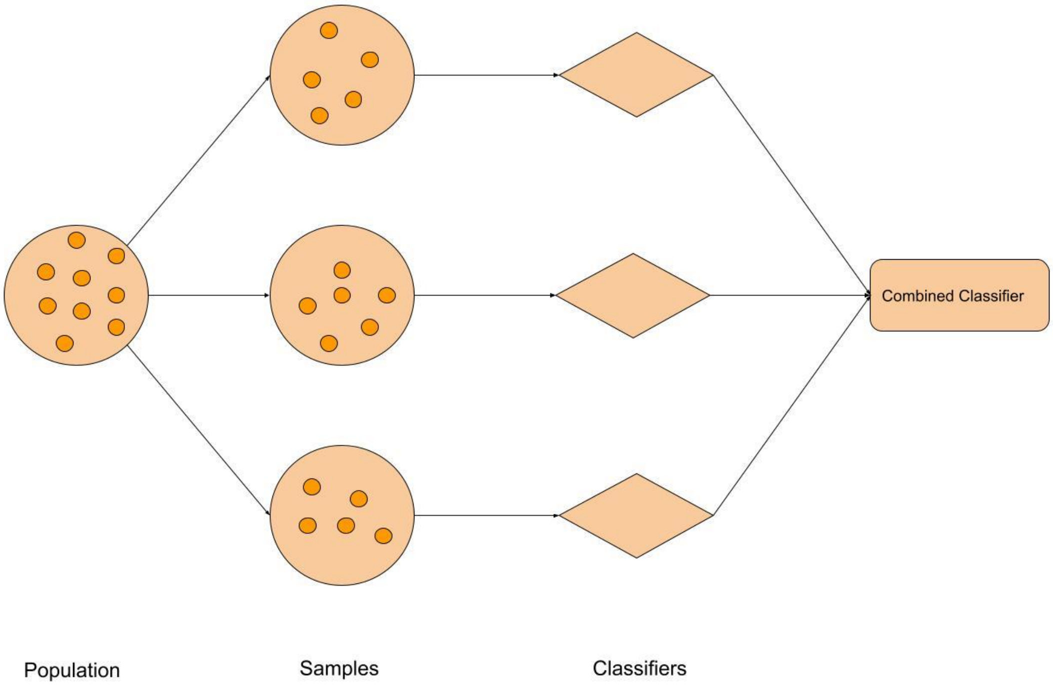

Figure 15. Mixing training data with bagging.

3.3.2. Mixing Combinations

Mixing combinations is another technique of ensemble learning which combines the machine learning models in order to make a better model. It has two powerful techniques which are boosting and stacking.

Boosting method starts with multiple collections of learners. The training dataset is divided into subgroups. Each learner is trained with a specific subgroup. If the performance of the learner is not good enough, more importance can be given to that learner. We can make it clear with an example. Let us assume that there is a class with students. They take tests about all subjects and some of them are successful in all subjects while some of them are not. In such a case, the students who are not successful enough are determined, and giving more attention to them would make their performance better (Kumar and Jain, 2020).

AdaBoost is the simplest but very important technique of boosting. This is a method that generally outperforms very complicated techniques that are used in machine learning studies. In the first step, data is trained with a classifier and the resulting observations are separated as correctly and incorrectly classified elements. Then, the misclassified elements are given a high weight, and training is run again. The weights are increased so that the model gives more importance to these misclassified elements and in the next step they have a higher chance to get correctly classified. In the next iteration, the misclassified elements in the previous iteration whose weight is increased will be correctly classified with a high probability. This provides a high-quality classifier using very weak classifiers. In the end, all outputs of the classifiers are combined with a voting mechanism.

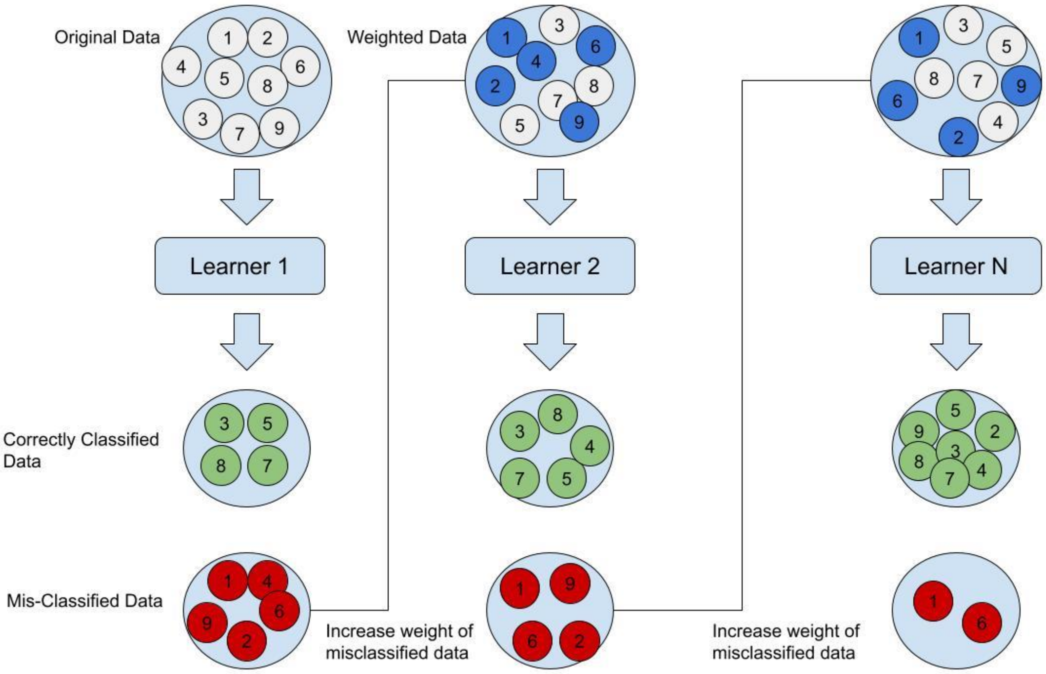
Original Data
1
2
4
5
8
6
3
7
9
Weighted Data
Learner 1
Learner 2
Learner N
Correctly Classified Data
Mis-Classified Data
Increase weight of misclassified data
Increase weight of misclassified data

Figure 16. AdaBoost structure.

There is another version of boosting methods which is Gradient Boosting. In the method, the weak learners are iteratively increased. The difference of the gradient boosting from AdaBoost is that in AdaBoost, a new learner is added after increasing the badly classified observations weight. However, in Gradient Boosting, the new model is trained on residual errors from the previous predictor.

XGBoost is another boosting algorithm. It is an expert algorithm in gradient boosting techniques. It adds some parameters and advances vanilla gradient boosting techniques. It determines the depth of weak learners, and it also determines the penalization parameters that are added to high depth prevention trees. The proportional shrinking of leaf nodes and Newton's tree boosting is used for the algorithm. It also uses the random parameters for optimal learning.

Stacking is another mixing combination technique that is very different from boosting. In that technique, the training data is trained with multiple machine learning algorithms which are called base learners. The prediction results of these learners are considered as training data and added to another learner called meta learner. It is called stacking because the logic of that method can be thought of as stacking one machine learning learner on top of the other one.

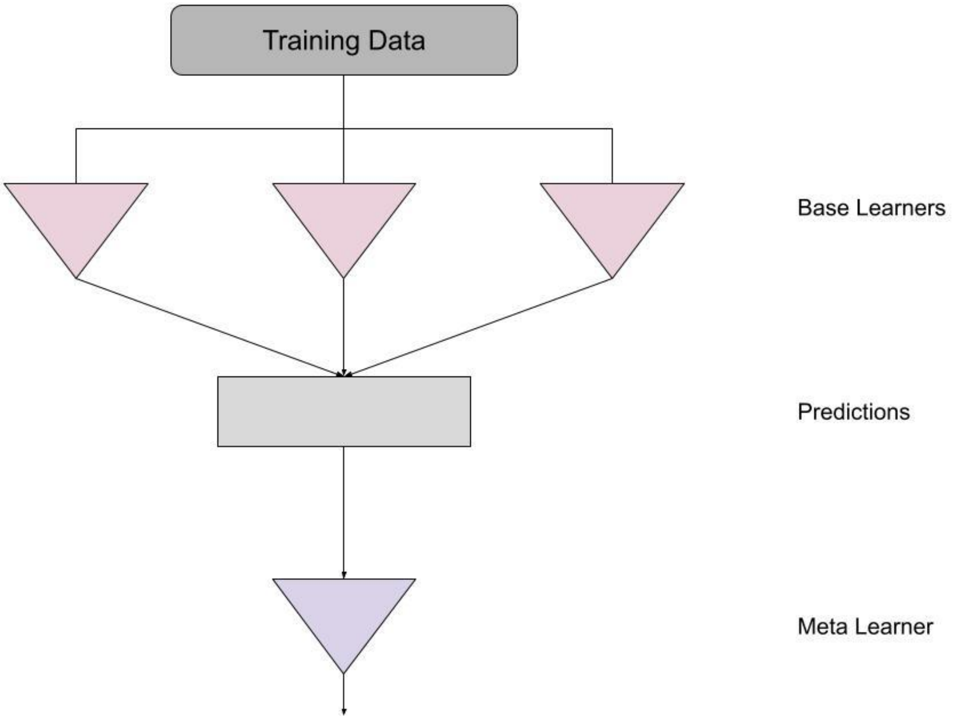

Figure 17. Stacking structure.

3.3.3. Mixing Models

Mixing training data and mixing combination techniques are related to how to divide training data into subgroups and mix them in different ways to construct ensemble models. Mixing models is such a technique that mixes the machine learning models using the same training dataset in each of them. Then, the results of these classifiers are combined together in different ways to get better performance. Mixing model techniques can be investigated under four concepts which are voting ensembles, hyperparameter tuning ensembles, horizontal ensembles, and snapshot ensembles.

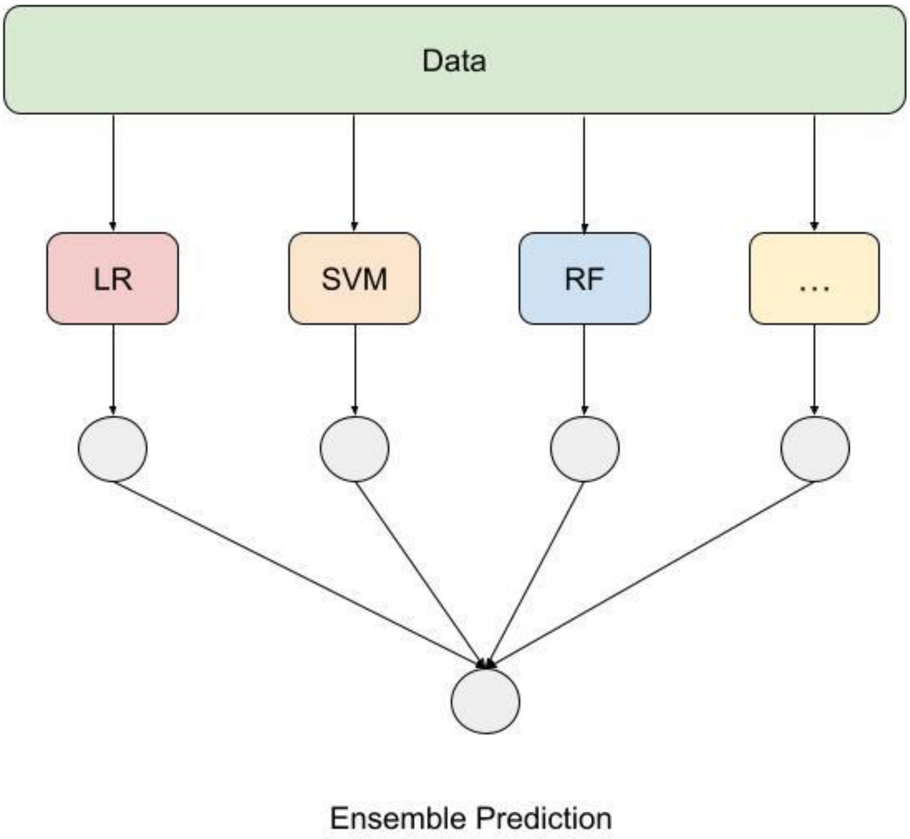

Figure 18. Mixing different models.

There are hard voting and soft voting/averaging techniques under the title of mixing models. Different machine learning algorithms are used to train voting ensemble models. For example, if there is a case in which the same data is trained with three different machine learning models, the combination of the outputs of these models is used to create ensemble predictions. The main issue in that kind of technique is how to combine the outputs of the models. The simplest and most well-known way is voting. In a simple classification problem, the machine learning algorithm chooses a specific category and the category that gets the most votes is the final chosen class. By following these steps, mostly the chosen class will have higher accuracy than any other single model.

Soft voting differs from hard voting in taking the probabilities of each class separately and combining these probabilities by taking the average of the predictions. Two methods have the same initial steps. However, in the hard voting, the voting classifier is used while in the soft voting there is an inference of the model from the test dataset. In the process of soft voting, the average output result is calculated with equally weighted models. Besides, there is an option that if a specific model has more importance than others the weight of that model can be increased, and the weights of other models can be decreased. This is a technique called weighted averaging.

Hyperparameter tuning ensemble is another technique that is used in ensemble learning. Instead of the other two early mentioned techniques which use different machine learning models to obtain ensemble models, this hyperparameter ensemble gives a chance to use a same good model with different hyperparameters in each training and combine their results to make an ensemble model.

In classical machine learning, it is useful and efficient to use hard voting, soft voting, and hyperparameter tuning. However, the things are not the same as in some special situations such as deep learning. In deep learning, training data and model size are very large. So, using these models might cause high computation and time. Computations can continue for days, and this would not be an effective situation. Horizontal voting is an option that provides saving models after a minimum number of epochs and then combining with voting to get an accuracy improvement.

Snapshot ensemble is a changed version of horizontal voting ensemble. As it is mentioned earlier, in horizontal voting the models are saved after a minimum epoch but in the snapshot, ensemble model the learning rate of the model is modified. It is used especially in deep learning to start initial higher learning and decrease the learning rate.

CHAPTER 4: PROPOSED METHOD

The general structure in the speech emotion recognition studies in the literature follows steps like extracting features from speech, training them, and obtaining predictions for multi-class cases. Features have a considerable amount of importance in speech emotion recognition studies because they give the specific information that provides distinguish emotions from each other.

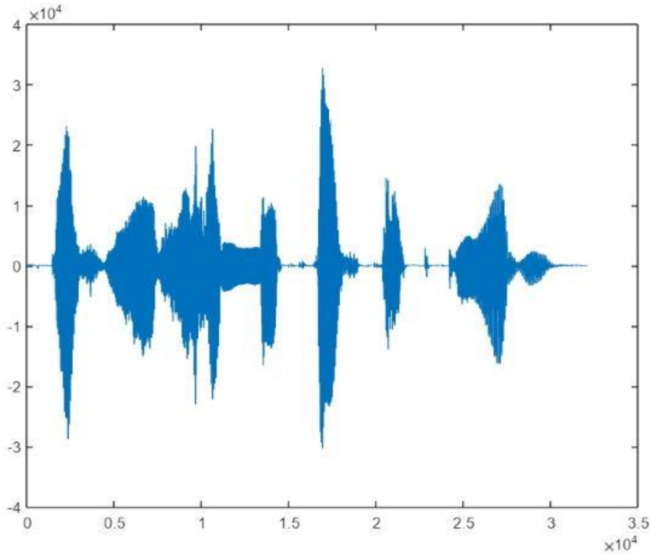

Figure 19. Reading sample of an audio speech.

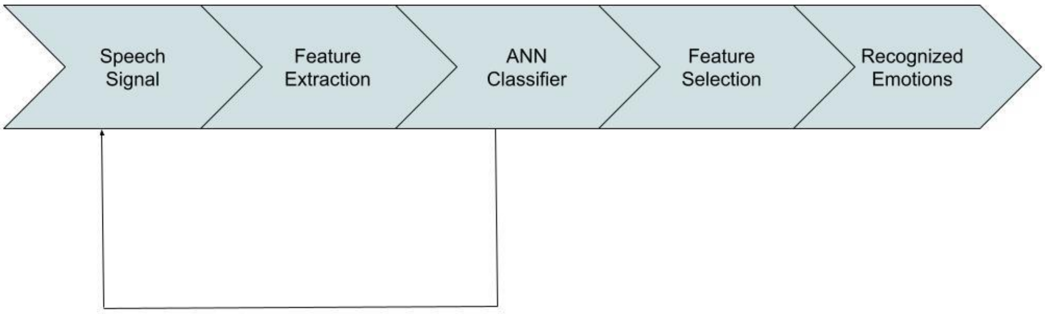

Figure 20. Proposed structure for ANN classifier.

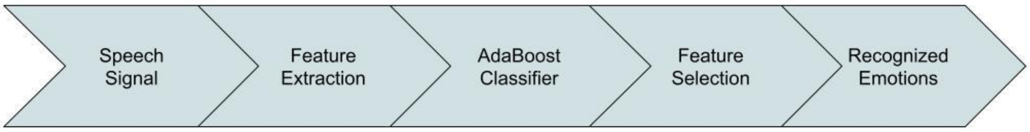

Figure 21. Proposed structure for AdaBoost classifier.

In the previous study (Özkan and Oğuz, 2020), the selection of emotion-specific speech features was performed using only the EMO-DB which is the German language. After the selection process of the emotion-specific features, that method becomes a guide for taking one more step further which is inspecting language independent speech emotion recognition study. The feature selection part of the study was not aimed as emotion specific but language independent this time.

In the proposed method there are two types of classifiers that are used in order to compare which procedure is going to be more successful for making a prediction for emotions. First classifier uses the utterances as they are in the databases. Original length of the utterances was preserved, and they were trained with artificial neural networks to predict emotions. In the second classifier, AdaBoost which is an ensemble learning technique was used. Both classifiers were trained with datasets in German, English, Italian, Urdu, and Japanese, and the predictions were obtained. The number of speech samples in datasets also differs. The TESS dataset has a considerably large number of samples when compared to other datasets. So, in order to make speech sample numbers closer to each other and provide the reliability of training, 300 random samples were chosen from the TESS dataset in both classifiers.

In the first classifier which is an artificial neural network, the utterances in the databases were used as a whole, their original length was protected. The MFCC and pitch features of these utterances from all datasets were extracted first. Then the statistical values of these features were calculated. The mean, maximum, minimum values, standard deviation, skewness, kurtosis, the first and the second derivatives of MFCC were used as spectral features while the mean, maximum, minimum values, and the standard deviation of the pitch were used as prosodic features. There are also other features that were added to improve the accuracy of the classification. One feature that was added is the band power of the utterances which gives the average power of the speech signal. It measures the power and power spectral density of a channel band. The difference feature which is the difference between the maximum and the minimum values of the pitch was also added. It gives information about how much the pitch values are changed in each utterance.

All these features were extracted and calculated for utterances in all language datasets. Feature selection was also necessary for reducing the dimensionality and choosing the most useful features for prediction. Backward selection was used for that purpose. All features were used for training at the beginning and each feature was dropped from the set of features one by one. The one that decreases the cross-entropy value the most was deleted permanently from the feature set and this process was repeated. While repeating these steps until there is no improvement in the cross-entropy value, the selected subset of features was constructed. The resulting feature set contains mean, maximum, minimum values, skewness, the first and the second derivatives of MFCC and, mean, maximum, minimum values, standard deviation, difference of pitch, and the band power feature. After that, backpropagation feedforward neural net was trained with the features, and predictions for emotions were obtained.

Various emotions are used by the people who created datasets for speech emotion recognition studies. For the reason that emotions are changed from one dataset to another, it is difficult to use all speech samples placed in all datasets. Since each dataset has different emotions, it is needed to choose common emotions that all datasets have. That’s why four basic emotions that all datasets contain were selected for the proposed method. Ekman listed six basic, inborn, and culturally independent emotions which are sadness, happiness, fear, anger, disgust, surprise, and neutral (Ekman, and Oster, 1979). These selected four emotions were happiness, sadness, anger, and neutral from Ekman’s emotions.

For the first step, EMO-DB which is a dataset in German and contains seven emotions that are happiness, sadness, fear, anger, disgust, surprise, and neutral was used in order to obtain predictions. Because the emotions that each dataset contains are different, speech samples which are happy, sad, angry, and neutral were separated from other emotions and labeled as it is. Their MFCC and pitch features were extracted, and the calculation of their statistical values was followed by the extraction. Mean, minimum, maximum values, standard deviation, skewness, kurtosis, delta (first derivative), and delta-delta (second derivative) features of MFCC and, mean, minimum, maximum values, and standard deviation of pitch feature were used. Difference and band power were also added to the feature set.

After performing these feature procedures, backpropagation feedforward neural network was trained with all feature sets.

After getting predictions using only EMO-DB, the TESS dataset, which is in English, was also added. Exactly the same procedures for separating four selected emotions samples and feature extraction as in EMO-DB were applied to the TESS dataset as well. Two different language speech samples were trained together, and predictions were obtained.

All different language datasets were added in the same fashion in the order of EMO-DB, TESS, EMOVO, URDU, and KEIO-ESD. In each new addition of the dataset, the neural net was trained with the added datasets in the previous steps and the newly added dataset. The recognition rates for each combination were calculated.

Since the emotions are different in different datasets, the utterances, number of samples also change from one dataset to another. This affects the neural net recognition rate. Training the net with the different number of samples that come from different languages may not lead to objective results. If the numbers have a considerable number of differences and one or more languages dominate the training set, the net will be trained with the dominant language and learn it better. This can make it difficult to get completely reliable predictions from the neural net. So, in order to make the number of samples that come from different languages closer, a process was applied. The number of samples for each different dataset and their distribution to the emotions are listed in the table. As it is seen from the table, the TESS dataset has the largest sample of utterances. Other ones have nearly the same number of samples or are close to each other. In order to make the number of samples of the TESS dataset closer to other ones, 300 random speech samples were chosen from the separated four emotion samples. In each combination that languages were added, 300 TESS speech samples were used in training and testing.

Table 1. Information of different language databases.

Database	**Language**	**Sample**	**Emotions**							
EMO-DB	German	535	angry	bored	disgust	fear	happy	sad	notr	
TESS	English	2800	angry		disgust	fear	happy	sad	notr	surprised
EMOVO	Italian	588	angry		disgust	fear	happy	sad	notr	surprised
URDU	Urdu	400	angry				happy	sad	notr	
KEIO-ESD	Japanese	1025	KEIO-ESD has 47 different emotions							

While adding the languages one by one, it is expected from the results to reflect the distance between the language families. It is predicted that the results in languages coming from closer families should not be changed much but the performance is predicted to decline as the similarity between languages decreases.

Since the emotion and feature sets were changed, the feature selection was needed again for the new construction of the method. For this purpose, this time backward feature selection method was used. In all language combinations, training and obtaining predictions were performed with all features that were extracted and calculated. After training with all features, features were taken out one by one from the feature set. The process was performed as follows. First, the net was trained with all features and the cross-entropy value was noted. After that, each feature was taken out one by one, and the cross-entropy values when they were excluded were determined. The feature that decreases the cross-entropy value the most when removed from the feature set was taken out. It is understood that this specific feature decreases the recognition rate. That’s why it will not be a proper feature for the dataset combination. After removing the first feature from the feature set, the same process was applied. It continues in that fashion until there is no improvement in the cross-entropy value.

Second classifier is AdaBoost. As in the first classifier, different languages were used in that method. The EMO-DB, TESS, EMOVO, URDU, and KEIO-ESD datasets were combined with the four emotions which are happy, sad, neutral, and angry. Keeping the number of speech samples that are going to be trained closer was aimed. That’s why random 300 speech samples were chosen from the TESS dataset which has the maximum number of speech samples when compared to other datasets.

The MFCC and pitch features were extracted as in the first classifier to be used in training and testing. The mean, minimum, maximum values, standard deviation, skewness, kurtosis, the first and the second derivatives of MFCC were used. The mean, minimum, maximum values and standard deviation of pitch features were also used. The difference feature which gives the difference between maximum and minimum values of pitch and band power was also added as in the first classifier.

The features and emotions of the dataset were extracted and given to the algorithm to be trained and tested. In order to keep data distribution homogeneous, k-fold cross validation was used. As a weak learner, the tree was used with 4 numbers of splits.

The feature selection was also applied for this classifier. Using the backward selection, training was started with all features and each feature was deleted from the feature set one by one and the success of the classifier was obtained. After selecting features, the remaining feature set consisted of mean, maximum, minimum values, standard deviation, skewness, and the first derivative of MFCC and, maximum value, standard deviation, difference feature of pitch, and the band power feature.

4.1. Features

Features are a very essential part of speech emotion recognition systems that are directly related to the success of the system. When the features are chosen carefully, the recognition rate can increase considerably. Although there are a lot of studies that are investigating which features are proper to increase emotion recognition rate, there is no certain rule for that. Appropriate features change for different purposes and studies.

Speech signal is a continuous signal that contains information and emotion about the speech. There are global and local features that are extracted from the speech according to the main purpose. Global features represent statistical values such as mean, standard deviation, minimum and maximum, while local features are related to temporal dynamics of a sound. In the speech signal, the emotional features are not placed in a specific order. That’s why the stationary state of the signal that is represented by local features is very important.

Global and local features of a sound are examined in categories as spectral and prosodic features.

4.1.1. Mel Frequency Cepstral Coefficients

In this study, Mel Frequency Cepstral Coefficients (MFCC) is used as a spectral feature to obtain the short-term power spectrum of a signal. MFCC is the most commonly used spectral feature in speech emotion recognition studies. It is a very popular and efficient feature that imitates the human ear using cepstral analysis. MFCC is computed from the speech frames and speech lengths are different from each other which causes extracting different numbers of coefficients from each speech. However, all coefficients are not used in speech emotion recognition since the most informational ones are placed in the beginning because of the property of cosine transform (Krishna Kishore and Krishna Satish, 2013). Generally, the number of coefficients that are used in speech emotion recognition studies is between eight and fourteen. In the proposed method 14 coefficients are used.

MFCC extraction from speech signals is basically divided into some steps. The speech signal is divided into segments, segments are converted into the frequency domain using Discrete Fourier Transform (DFT), number of sub-band energies is calculated using Mel Filter-bank, logarithm of these sub-bands are calculated and finally, Inverse Fourier Transform is applied to obtain coefficients.

4.1.1.1. Framing and Windowing

Speech signals need to be processed in short time intervals called frames due to their slow time-varying nature. In these short time intervals, the speech signal is assumed to be stationary. While investigating short-term spectral features, generally 20 ms window length is used and window is advanced every 10 ms in order to overlap frames for the purpose of smoothing the transition between the frames (Deller Jr, 1993), (Benesty, Sondhi and Huang, 2007). Advancing frames every 10 ms allows us to observe the temporal characteristics of a sound. Furthermore, 20 ms window length is good enough to analyze spectral resolution, and also short enough to observe temporal changes of the sound (Rao and Vempada, 2013). The main aim of this overlapping structure is to center every speech sound on a frame. After that, a window is applied to these frames for shrinking the signal against frame boundaries. In general, to perform these operations, Hanning or Hamming windows are used (Picone, 1993).

4.1.1.2. Discrete Fourier Transform (DFT)

After frame blocking and windowing, each windowed frame which is a signal is converted to the frequency domain by applying DFT.

4.1.1.3. Mel Spectrum

After the two steps explained above, Mel Spectrum is computed with a Mel Filter-bank which is a set of band pass filters. Fourier transformed signals are passed to these filters to compute the Mel Spectrum. Mel can be explained as a unit that is based on frequency that comes to human ears. Mel filter banks can be applied both in the time domain and frequency domain even though the frequency domain is preferred in MFCC calculation. Mostly the triangular and Hanning filters are used in these operations.

4.1.1.4. Discrete Cosine Transform (DCT)

As a last step, the speech signals that are converted to the frequency domain are converted back to the time domain again. Before the conversion, the Mel spectrum is represented on a log scale. Then, DCT is applied in order to obtain cepstral coefficients. The zeroth coefficient is not included in general because it represents the speaker-specific information.

4.1.1.5. Dynamic MFCC Features

The cepstral coefficients that are calculated by following all these operations are accepted as static features because they just have information about the frames that are extracted from. There are other informational features about the MFCC that can be calculated by taking the first and second derivatives called delta and delta-delta coefficients (Furui, 1981). These features give information about speech rate and acceleration of speech respectively.

In the proposed structure, MFCC features have been extracted and the statistical values which are mean, minimum, maximum, and standard deviation are calculated. In addition, skewness, kurtosis, first derivative, and second derivative of MFCC are calculated as well but not all these features have been used, feature selection is performed to find the most informational features.

The mean value was calculated for each of 14 coefficients of MFCC and there are 14 columns mean value for each speech sample. Minimum, maximum values, standard deviation, skewness, and kurtosis were also obtained in the same way. Each statistical value was investigated column-wise for each of 14 columns of MFCC. Skewness measures the asymmetry of the probability distribution of a random variable with a real value about its mean. On the other hand, kurtosis measures the tailedness of the probability distribution of a random variable with a real value about its mean. The MFCC derivatives give information about the variations between the frames. First derivative of MFCC gives information about speech rate and the second derivative of MFCC represents the acceleration of speech (Rao and Manjunath, 2017).

4.1.2. Fundamental Frequency (F0)

In speech emotion recognition studies, besides spectral features, there are also prosodic features that are mostly used as features. Prosodic features are related to rhythm and intonation that are perceived by the human ear. In a typical speech, there are some intonations that we use in order to express our feelings or ask a question to the person that we are dealing with. Prosodic features help to determine all these intonations that we use in the speech.

The most used and popular prosodic feature in speech emotion recognition studies is the fundamental frequency. Fundamental frequency is the vibration rate of the vocal cord, which vibrates in a quasiperiodic way in voiced speeches, over time. This vibration rate is related to the air pressure that is felt on the vocal cords and also the tension of facial muscles. That’s why fundamental frequency is a very important feature for the detection of emotion.

Fundamental frequency is directly related to the pitch and the way that we understand fundamental frequency can be described as pitch. In other words, F0 is the physical concept while pitch is the perception of a signal with our ears and interpretation of it with our brain.

Pitch is mostly preferred in speech emotion recognition and gives us information about the emotion of speech since it has some differences in different emotions. For example, in the emotion of fear, the value of pitch is high but in the emotion of disgust, the value is low (Koolagudi, Murthy and Bhaskar, 2018). In addition to that, the statistical values of prosodic features are very informative for the purpose of emotion detection. Range, minimum, maximum, mean, standard deviation, slope, median, skewness, and kurtosis are also used in the studies (Wu and Liang, 2010), (Rao, Koolagudi and Vempada, 2013). The mean, maximum, minimum values, and standard deviation of F0 were extracted in the proposed structure as features, and feature selection was performed to choose the most efficient ones. There is another feature that was used related to the pitch. This is the feature that was added differently from the studies in the literature which is the difference between the maximum and the minimum value of pitch for each speech.

The mean of fundamental frequency was also obtained the same way as MFCC but because the pitch feature has 1 column value for each speech sample, there was one mean, maximum, minimum, and the standard deviation for each speech sample. Difference and band power features were also obtained as one value for each sample.

Band power of the speech samples is another feature that was added.

4.1.3. Sequential Feature Selection

Sequential feature selection methods are mostly used techniques for choosing the proper and useful features for the algorithms. The features are added or removed from the dataset sequentially for reducing the number of features in order to ensure that the model meets the optimal performance and result. Sequential feature selection methods trace only one direction which is either increasing or reducing the number of features in the dataset.

In sequential forward selection, the features are added sequentially to the empty feature set until the added features do not reduce the criterion that is determined for the algorithm.

On the other hand, in backward selection, the features are removed from the dataset sequentially as opposed to the forward selection until the removal of a feature does not improve the criterion.

4.2. Parameters

For both classifiers, the speech samples and the speech rate of signals were obtained with the MATLAB audioread function. After obtaining these samples, the MFCC and pitch features were extracted with MFCC and pitch functions by giving the samples and speech rate, which changes one dataset from the other one. The default parameters were not changed in these functions. The returned MFCC coefficients were 14 because it yielded the best results in the experiments. For each speech feature there were 14 column values while the pitch feature returned 1 column values for each speech sample.

Statistical values of MFCC were also extracted with mean, min, max, std, skewness, and kurtosis functions with default parameters. For the first and second derivatives, the diff function of MATLAB was used with the parameters n equal to 1 and 2 respectively. The parameter n refers to the nth derivative.

Pitch statistical values were extracted with mean, min, max, and std functions and they were used as they are. The difference feature which gives the difference between the maximum and the minimum values of pitch was calculated by simple subtraction. Finally, the band power feature was extracted with the bandpower function with no parameters.

In the first classifier which is ANN, the number of neurons in the hidden layer was set to 15 which yielded the best results in the previous study (Özkan and Oğuz, 2021). The training was repeated 20 times and the average error rate and cross-entropy values of the net were considered. 80% of the data was left for training, 10% for testing, and 10% for validation.

For the second classifier which is AdaBoost, the k-fold cross validation was performed with k equals 4. The algorithm was trained, and results were obtained for 4 cases and their average success was considered. The MaxNumSplits parameter which is the depth of the weak learner tree was set to 4 because there were 4 emotions that had to be separated from each other.

CHAPTER 5: RESULTS

In the first classifier, four emotion (happy, sad, angry, neutral) speech samples of the EMO-DB were extracted from other emotion speech samples. After that, their MFCC and pitch features were extracted, and statistical values were calculated. Emotion labels of speech samples and the features were given to the neural net. In the beginning, all features were used to train the net and obtain the results. The neural net was trained with all features 20 times and their average cross-entropy values were considered for the prediction. With the 339 happy, angry, sad, and neutral speech samples of EMO-DB, the obtained net success was 86.43%.

After training the neural net only with a German database, the TESS dataset was added. This dataset is in English and comes from the same family as German. Happy, angry, sad, and neutral speech samples of the dataset were extracted from other emotion samples. Since the number of speech samples in the TESS dataset was considerably greater than the other datasets, random 300 samples of TESS were chosen in order to make speech numbers closer and provide reliable results. All calculated features of EMO-DB and TESS were used to train the net. With 639 samples and training 20 times, the results were obtained as 93.11%.

The third dataset that was added to the German and English datasets was EMOVO which is in Italian and comes from another language family from German and English. 336 happy, angry, sad, and neutral speech samples of EMOVO and their features were extracted in order to train. After the addition of EMOVO, 975 speech sample features in German, English, and Italian and their emotion labels were given to the neural net for training. The net success was 86.05%.

After the EMOVO, the URDU dataset was added. URDU language comes from Indo-Iranian language family. The same processes were applied as in the previous language additions. URDU has 400 speech samples of happy, angry, sad, and neutral emotions. With the addition of URDU, 1375 features of speech samples and their emotion labels were given to the neural net for training 20 times. The success of the net was 82.76%.

The last dataset added to the language set in order to train was KEIO-ESD which is in Japanese. Japanese comes from the Japonic language family. The number of happy, angry, sad, and neutral speech samples of KEIO were 165. It is less than all other datasets. The features of these 165 speech samples were extracted and they were given to the neural net with their emotion labels. 1540 speech samples were trained 20 times again and the success was 83.12%.

As we can see from the table, adding the TESS dataset which comes from the same language family after the EMO-DB increased the results. When EMOVO was added, the recognition rate decreased since it comes from a different language family. The Urdu dataset also decreased the recognition rate. Japanese did not decrease the results but it has a relatively less speech sample than all other datasets.

Table 2. Recognition rates of dataset combinations.

DB	#samples	true guesses	cross entropy	%success
EMO (German)	339	293	0.0341	**86.43**
EMO+TESS (English)	639	595	0.0580	**93.11**
EMO+TESS+EMOVO (Italian)	975	839	0.1101	**86.05**
EMO+TESS+EMOVO+URDU (Urdu)	1375	1138	0.1303	**82.76**
EMO+TESS+EMOVO+URDU+KEIO (Japanese)	1540	1280	0.1282	**83.12**

After all, different language datasets were added and the successes of nets were obtained, the backward selection was applied in the case when all languages were added in. When all features were used, the success rate was 83.12% and the cross-entropy was 0.1282. Cross-entropy values were used in order to determine the success of the trained net. In the backward selection process, each feature was deleted from the dataset. For all these cases, the nets were trained 20 times and their cross-entropy values were noted. The decrease in the cross-entropy means that the specific feature does not give much information about the emotion of its speech sample. The feature whose deletion from the dataset gave the lowest cross-entropy value was deleted from the dataset permanently. This process was repeated until there was no improvement in the cross-entropy value. After two steps, the improvement of cross-entropy was stopped and the remaining features were mean, maximum, minimum values, skewness, first and the second derivatives of MFCC and mean, maximum, minimum values, standard deviation, difference features of pitch, and the band power feature. The cross-entropy value of this feature combination was 0.1115 and the success of the net was 90.65% which is shown in Table 3.

Table 3. ANN recognition rate after backward feature selection.

DB	#of samples	true guesses	cross entropy	%success
EMO+TESS+EMOVO+URDU+KEIO	1540	1396	0.1115	**90.65**

Table 4. Remaining features after the first backward feature selection in ANN classifier.

mfcc	mean	✓
	max	✓
	min	✓
	std	✓
	skewness	✓
	kurtosis	
	delta	✓
	ddelta	✓
f0	mean	✓
	max	✓
	min	✓
	std	✓
	difference	✓
	band power	✓
cross entropy		**0.1117**

For the comparison of studies in the literature which focuse on a single language, the ANN classifier was trained with all language datasets (EMO-DB, TESS, EMOVO, URDU and KEIO-ESD). This multilingual trained net was tested with each language indipendently and the results were compared with other studies in the Table 5.

Table 5. Comparison of studies in the literature that focus on single languages.

Paper	**Language**	**Classifier**	**Recognition Rate**
Petrushin (1999)	English	ANN	65%
Mao, Chen, and Fu (2009)	German, Mandarin	Hybrid (ANN and HMM)	81.7%
Nicholson, Takahashi, and Nakatsu (1999)	Japanese	ANN	50%
Latif et al. (2018)	Urdu	SVM	83.40%
Latif et al. (2018)	Urdu (Multi-lingual training)	SVM	70.98%
Latif et al. (2018)	Italian	SVM	74.01%
This study	German	ANN	85.84%
	English	ANN	85.90%
	Italian	ANN	79.46%
	Urdu	ANN	86.75%
	Japanese	ANN	93.94%

If we look into the second classifier which is AdaBoost, the case that had all features were used to train the algorithm. The extracted features and their emotion labels were given to the AdaBoost algorithm in order to train and predict the results. As in the ANN classifier, the 1540 speech samples were trained, and the predictions were obtained. Classification trees were used as weak learners within the algorithm and the k-fold cross-validation was performed. The performance of the algorithm with all features and all datasets was determined as 71.75%.

Table 6. Remaining features after the second backward feature selection in ANN classifier.

mfcc	mean	✓
	max	✓
	min	✓
	std	
	skewness	✓
	kurtosis	
	delta	✓
	ddelta	✓
f0	mean	✓
	max	✓
	min	✓
	std	✓
	difference	✓
	band power	✓
cross entropy		**0.1115**

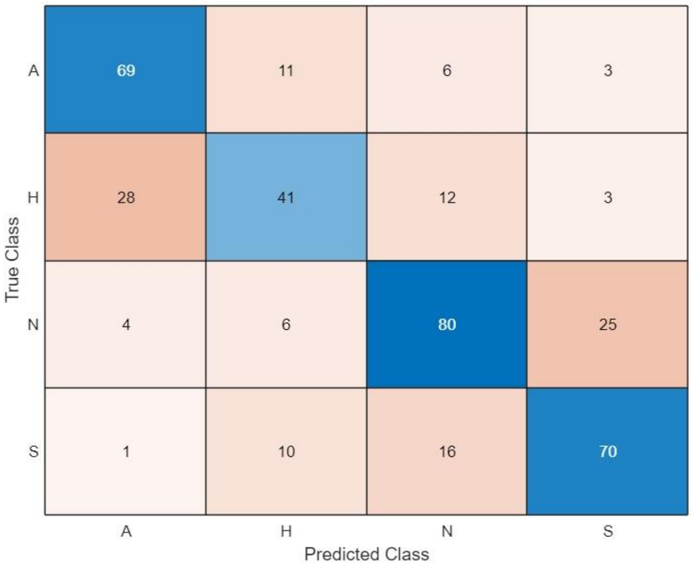

Figure 22. Confusion table of AdaBoost classifier before backward feature selection.

After getting the predictions, the backward feature selection was also applied to the second classifier. The steps were the same as in the first classifier. Each feature was removed from the dataset one by one, and their results were noted. A feature whose deletion improved the results the most was deleted from the dataset permanently. This feature selection process continued until there was no improvement in the results and it stopped after four steps. The remaining features were mean, maximum, minimum values, standard deviation, skewness, and the first derivative of MFCC and, maximum value, standard deviation, the difference of pitch feature, and the band power feature. The success of the algorithm with these features was 72.60%.

Table 7. Remaining features after the first backward feature selection in AdaBoost classifier.

mfcc	mean	√
	max	√
	min	√
	std	√
	skewness	√
	kurtosis	√
	delta	√
	ddelta	√
f0	mean	
	max	√
	min	√
	std	√
	difference	√
	band power	√
total score		**71.753 %**

Table 8. Remaining features after the second backward feature selection in AdaBoost classifier.

mfcc	mean	✓
	max	✓
	min	✓
	std	✓
	skewness	✓
	kurtosis	✓
	delta	✓
	ddelta	✓
f0	mean	
	max	✓
	min	
	std	✓
	difference	✓
	band power	✓
total score		**72.208 %**

Table 9. Remaining features after the third backward feature selection in AdaBoost classifier.

mfcc	mean	√
	max	√
	min	√
	std	√
	skewness	√
	kurtosis	√
	delta	√
	ddelta	
f0	mean	
	max	√
	min	
	std	√
	difference	√
	band power	√
total score		**72.533 %**

Table 10. Remaining features after the fourth backward feature selection in AdaBoost classifier.

mfcc	mean	√
	max	√
	min	√
	std	√
	skewness	√
	kurtosis	
	delta	√
	ddelta	
f0	mean	
	max	√
	min	
	std	√
	difference	√
	band power	√
total score		72.597 %

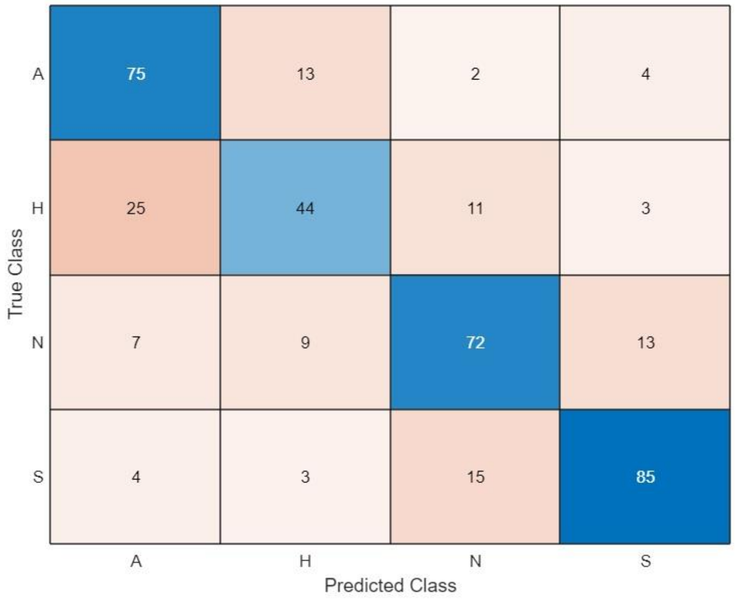

Figure 23. Confusion chart of AdaBoost classifier after backward feature selection.

CHAPTER 6: CONCLUSION

This thesis focuses on language independent speech emotion recognition. Five different language datasets which are EMO-DB (German), TESS (English), EMOVO (Italian), URDU (Urdu), and KEIO-ESD (Japanese) were used to obtain this aim. Since the emotions of the datasets are different from each other, the four common basic ones were selected for the SER system. Happy, angry, sad, and neutral were the utilized emotions. The MFCC and pitch speech features were extracted, and their statistical values were calculated. Difference feature which is the difference between the maximum and the minimum values of pitch and the band power feature was also added to the feature set. ANN and AdaBoost were used as classifiers to train and predict the emotions. The training started with one dataset which is EMO-DB and TESS, EMOVO, URDU, and KEIO-ESD were added respectively to see how the relation between the languages affects the recognition rates. Starting with EMO-DB, after all, added datasets the ANN was trained with the available data and the recognition rates were noted. It was seen from the results that languages that came from the same language family increased the recognition rate or did not decrease. However, if a language that came from another language family than the available dataset was added, the recognition rate of the classifier was decreased. It can be said from this situation that training the classifier with similar languages was more successful than training with languages that came from different language families. The relation of languages affects the success of the classifier.

All extracted features and the emotion labels of all datasets were given to two classifiers which are ANN and AdaBoost. The backward sequential feature selection was applied to two classifiers to reduce the dimensionality and use the most informational features. In the ANN classifier, the remaining features after feature selection were mean, maximum, minimum values, skewness, first and the second derivatives of MFCC feature and mean, maximum, minimum values, standard deviation, difference of pitch features, and the band power feature. On the other hand, the remaining features after the feature selection in AdaBoost were mean, maximum, minimum values, standard deviation, skewness, the first derivative of MFCC and, maximum values, standard deviation, difference of pitch, and the band power feature. The success rate of the ANN classifier with feature selection was 90.65% while the success rate of AdaBoost classifier with feature selection was 72.60%. It was seen that the ANN classifier obtained more successful results than AdaBoost classifier.

In the literature, there are many studies that use single and multiple languages in SER. Language independent SER provides more universal systems that can be used in a wide area. Also, it provides the capability that the system can be used by many people who speak different languages, not restricted to a just few ones. Even the studies that focus on a single language are not very successful and even in English, some accent problems can occur. Language independent SER will help to reduce these problems by making the system independent from language. It is explained that which speech features are independent from language and these features can be used with different language datasets to obtain successful recognition rates independent from language that is used.

REFERENCES

Akçay, M.B. and Oğuz, K. (2020) *Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers*, Speech Communication, Vol. 116, pp. 56–76.

Benesty, J., Sondhi, M.M. and Huang, Y. (2007) *Springer Handbook of Speech Processing*. 1st edition. Berlin: Springer.

Blanton, S. (1915) *The voice and the emotions*, Quarterly Journal of Speech, Vol. 1(2), pp.154-172.

Burghardt, F., Kienast, M., Paeschke, A., and Weiss, B. (1999). *Berlin Database of Emotional Speech* [Online]. Available at: http://emodb.bilderbar.info/docu/ (Accessed: 12 January 2020).

Chen, L., Su, W., Feng, Y., Wu, M., She, J., and Hirota, K. (2020) *Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction*, Information Sciences, Vol. 509, pp. 150–163.

Comrie, B. (1987) *The world's major languages*. 3rd edition. London: Routledge.

Constantini, G., Iaderola, I., Paolini, A., and Todisco, M. *EMOVO corpus: an Italian emotional speech database, International Conference on Language Resources and Evaluation*. Rejkjavic. May, 2014.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J.G. (2001) *Emotion recognition in human-computer interaction*, IEEE Signal processing magazine, Vol. 18.1, pp. 32-80.

Deller Jr, J. R. (1993) *Discrete-time processing of speech signals*. 1st edition. Chippenham: Wiley-IEEE Press.

Dimmendaal, G.J. (2011) *Historical Linguistics and the Comparative Study of African Languages*, 1st edition. New York: John Benjamins Publishing Company.

Dupuis, K., and Pichora-Fuller, M. K. (2010). *Toronto Emotional Speech Set (TESS)* [Online]. Available at: https://tspace.library.utoronto.ca/handle/1807/24487 (Accessed: 12 January 2020).

Ekman, P. (1971) *Universals and cultural differences in facial expressions of emotion*, Nebraska Symposium on Motivation, Vol. 19, pp. 207-282.

Ekman, P., and Oster, H. (1979) *Facial expressions of emotion*, Annual Review of Psychology, Vol. 30, pp. 527–554.

Ekman, P., Friesen, W. V., and Ellsworth, P. (2013) *Emotion in the human face: Guidelines for research and an integration of findings*. 1st edition. Burlington:

Elsevier.

Ethnologue. (2022). *How many languages are there in the world?* [Online]. Available at: https://www.ethnologue.com/guides/how-many-languages. (Accessed: 20 May 2022).

Fausett, L. (1993) *Fundamental of Neural Networks: Architectures, Algorithms and Applications.* 1st edition. Hoboken: Prentice Hall.

Furui, S. (1981) *Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 29(3), pp. 342–350.

Garton, S., and Copland, F. (2019) *The Routledge Handbook of Teaching English to Young Learners.* 1st edition. New York: Routledge.

Katsaggelos, A.K. (2002) *IEEE Signal Processing Magazine: Farewell*, IEEE Signal Processing Magazine, Vol. 19(6), pp. 2–4.

Ke, X., Zhu, Y., Wen, L., and Zhang, W. (2018) *Speech emotion recognition based on SVM and ANN*, International Journal of Machine Learning and Computing, Vol. 8(3), pp. 198–202.

Koolagudi, S.G., Murthy, Y.V.S. and Bhaskar, S.P. (2018) *Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition*, International Journal of Speech Technology, Vol. 21(1), pp. 167–183.

Krishna Kishore, K. V. and Krishna Satish, P. *Emotion recognition in speech using MFCC and wavelet features*, *Proceedings of the 2013 3rd IEEE International Advance Computing Conference, IACC 2013,* February 2013.

Kumar, A. and Jain, M. (2020) *Ensemble Learning for AI Developers*, *Ensemble Learning for AI Developers*. 1st edition. Berkeley: BApress.

Latif, S., Qayyum, A., Usman, M., and Qadir, J. *Cross lingual speech emotion recognition: Urdu vs. western languages, 2018 International Conference on Frontiers of Information Technology (FIT)*, IEEE, December 2018.

Mao, X., Chen, L. and Fu, L. (2009) *Multi-level speech emotion recognition based on HMM and ANN, 2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009,* March 2009.

Moriyama, T. *Keio University Japanese Emotional Speech Database (Keio-ESD), Speech Resources Consortium, National Institute of Informatics*. Tokyo Polytechnic University, Tokyo. 2011.

Mourad, T. (2022). *Arabic Speech Recognition by Stationary Bionic Wavelet*

Transform and MFCC Using a Multi-layer Perceptron for Voice Control, The Stationary Bionic Wavelet Transform and its Applications for ECG and Speech Processing. Signals and Communication Technology. 1st edition. Cham: Springer.

Murray, I.R. and Arnott, J.L. (1993) *Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion*, Journal of the Acoustical Society of America, Vol. 93(2), pp. 1097–1108.

Nicholson, J., Takahashi, K. and Nakatsu, R. (1999) *Emotion recognition in speech using neural networks, ICONIP 1999, 6th International Conference on Neural Information Processing - Proceedings*, 2000.

Nievergelt, J. (1969) *R69-13 Perceptrons: An Introduction to Computational Geometry*, IEEE Transactions on Computers, Vol. C–18(6), pp. 572.

Noroozi, F., Kaminska, D., Sapinski, T., and Anbarjafari, G. (2017) *Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and adaboost*, Journal of the Audio Engineering Society, Vol. 65(7/8), pp. 562-572.

Özkan, C. and Oguz, K. *Selecting emotion specific speech features to distinguish one emotion from others*, *2021 International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2021 - Proceedings*. August 2021.

Pan, S., Tao, J. and Li, Y. (2011) *The CASIA audio emotion recognition method for audio/visual emotion challenge 2011*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 6975 LNCS, pp. 388–395.

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., and Hofer, G. *Analysis of deep learning architectures for cross-corpus speech emotion recognition*, *Proceedings of the Annual Conference of the International Speech Communication Association*, *INTERSPEECH,* September 2019.

Petrushin, V.A. (1999) *Emotion in speech: Recognition and application to call centers,* Intelligent Engineering Systems Through Artificial Neural Networks, Vol. 9, pp. 1085–1092.

Picone, J.W. *Signal Modeling Techniques in Speech Recognition, Proceedings of the IEEE.* Tsukuba. 3 June 1993.

Rajisha, T.M., Sunija, A.P. and Riyas, K.S. (2016) *Performance Analysis of Malayalam Language Speech Emotion Recognition System Using ANN/SVM*, Procedia Technology, Vol. 24, pp. 1097–1104.

Rao, K.S., Koolagudi, S.G. and Vempada, R.R. (2013) *Emotion recognition from* *speech using global and local prosodic features*, International Journal of Speech Technology, Vol. 16(2), pp. 143–160.

Rao, K. S., and Manjunath, K. E. (2017). *Speech recognition using articulatory and excitation source features*. 1st edition. Cham: Springer.

Rosenblatt, F. (1958) *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychological Review, Vol. *65*(6), pp. 386.

Rowe, B.M. and Levine, D.P. (2015) *A concise introduction to linguistics: A Concise Introduction to Linguistics*. 4th edition. Oxfordshire: Routledge.

Russell, J.A., Mehrabian, A., (1977) *Evidence for a three-factor theory of emotions,* Journal of research in Personality Res, Vol. 11(3), pp. 273–294.

Schuller, B.W. (2018) *Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends*, Communications of the ACM, Vol. 61(5), pp. 90–99.

Szwoch, M. and Szwoch, W. (2015) *Emotion recognition for affect aware video games, Choraś, R. (eds) Image Processing and Communications Challenges 6. Advances in Intelligent Systems and Computing*, 2015th edition. Cham: Springer.

Tan, L., Yu, K., Lin, L., Cheng, X., Srivastava, G., Lin, C.W., and Wei, W. (2022) *Efficiency Solution for Autonomous Vehicles in a 5G-Enabled Space – Air – Ground Integrated Intelligent Transportation System,* IEEE Transactions on Intelligent Transportation Systems, Vol. 23(3), pp. 2830–2842.

Watson, D., Clark, L.A., Tellegen, A., (1988) *Development and validation of brief measures of positive and negative affect: the panas scales*, Journal of personality and social psychology, Vol. 54 (6), pp. 1063.

Wu, C.H. and Liang, W. B. *Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels, International Conference on Affective Computing and Intelligent Interaction*, 2010.

Yoon, W. J., Cho, Y. H., and Park, K. S. *A study of speech emotion recognition and its application to mobile services, International Conference on Ubiquitous Intelligence and Computing*, Berlin, Heidelberg. July 2007.