KADİR HAS ÜNİVERSİTESİ

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

AUDIO DETECTION USING MACHINE LEARNING

&

TRANSFER LEARNING MODELS

MESUT ACAR

PROF. DR. HASAN DAĞ

MASTER'S DEGREE THESIS

İSTANBUL, MAY, 2021

MESUT ACAR M.S THESIS 2021

AUDIO DETECTION USING MACHINE LEARNING
&
TRANSFER LEARNING MODELS

MESUT ACAR
PROF. DR. HASAN DAĞ

MASTER’S DEGREE THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES
WITH THE AIM TO MEET THE PARTIAL REQUIREMENTS REQUIRED TO
RECEIVE A MASTER’S DEGREE IN THE DEPARTMENT OF MANAGEMENT
INFORMATION SYSTEMS

İSTANBUL, MAY , 2021

NOTICE ON RESEARCH ETHICS AND PUBLISHING METHODS

I, MESUT ACAR;

• hereby acknowledge, agree and undertake that this Master's Degree Thesis that I have prepared is entirely my own work and I have declared the citations from other studies in the bibliography in accordance with the rules;

• that this Master's Degree Thesis does not contain any material from any research submitted or accepted to obtain a degree or diploma at another educational institution;

• and that I commit and undertake to follow the "Kadir Has University Academic Codes of Conduct" prepared in accordance with the "Higher Education Council Codes of Conduct".

In addition, I acknowledge that any claim of irregularity that may arise in relation to this work will result in a disciplinary action in accordance with the university legislation.

MESUT ACAR

DATE AND SIGNATURE

ACCEPTANCE AND APPROVAL

This study, titled **AUDIO DETECTION USING MACHINE LEARNING & TRANSFER LEARNING MODELS**, prepared by the **MESUT ACAR**, was deemed successful with the **UNANIMOUS VOTING** as a result of the thesis defense examination held on the 12.05.2021 and approved as a **MASTER'S DEGREE THESIS** by our jury.

JURY: SIGNATURE:

Prof. Dr. Hasan Dağ (Advisor) (Kadir Has University)

____

Assist. Prof. Dr. E. Fatih Yetkin (Jury Member) (Kadir Has University)

____

Prof. Dr. Mustafa Bağrıyanık (Jury Member) (Istanbul Technical University)

____

I confirm that the signatures above belong to the aforementioned faculty members.

____

Prof. Dr. Mehmet Timur Aydemir
Director of the School of Graduate Studies
APPROVAL DATE:

TABLE of CONTENTS

LIST OF TABLES

LIST OF FIGURES

AUDIO DETECTION USING MACHINE LEARNING
&
TRANSFER LEARNING MODELS

ABSTRACT

In this paper, using datasets ESC-50 & ESC-10 of environmental sounds, machine learning algorithms, and feature extraction methods are used to develop recognition performance. K-NN, SVM, Random Forest are used for comparing the recognition results. The different feature extraction methods in the literature are used to get more meaningful attributes from these datasets and obtain a higher accuracy rate. This approach shows that SVM algorithm has a significantly good result with accuracy scores. The best accuracy scores obtained by classic machine learning algorithms are %42,15 for ESC-50 and %77,7 for ESC-10. In addition to this, the experiments have been done with a pre-trained ResNet neural network as a backbone, which achieves successful results despite the machine learning models. In this study, a higher accuracy rate is achieved from baseline machine learning algorithms in literature and using transfer learning with pre-trained Resnet backbones to reach some state of art results. The accuracy scores are %68,95 for ESC-50 and %87,25 for ESC-10.

Keywords: Sound Classification, Audio, Environmental Sound, MFCC, ML, Resnet, Neural Network, Transfer Learning.

MAKİNE ÖĞRENMESİ VE TRANSFER ÖĞRENİMİ KULLANILARAK SES TANIMA

ÖZET

Bu çalışmada çevre seslerinden oluşan ESC-50 ve ESC-10 veri seti, çeşitli makine öğrenmesi, transfer öğrenme altyapısı ve farklı öznitelik çıkarımı yöntemleri kullanarak sınıflandırma çalışmaları yapılmıştır. K-NN, SVM, Rastgele Orman makine öğrenimi algoritmaları kullanılmıştır. Farklı öznitelik çıkarım algoritmaları kullanılarak, bu veri seti için makine öğrenmesi algoritmalarında farklı sonuçlar elde edilmiştir. Bu yaklaşımda SVM algoritmasın gözle görülür bir şekilde performansının attığı gözlemlenmiştir. Klasik makine öğrenmesi algoritmaları ile elde edilen en iyi doğruluk puanları ESC-50 için %42,15 ve ESC-10 için %77,7'dir. Buna ek olarak, makine öğrenmesi modellerinden daha başarılı sonuçlar elde eden, omurga olarak önceden eğitilmiş bir ResNet sinir ağı ile deneyler yapılmıştır. Yapılan deneylerde, literatürdeki temel makine öğrenmesi algoritmalarından ve literatürdeki iyi sonuçlara ulaşmak için önceden eğitilmiş Resnet omurgaları ile transfer öğrenmesi kullanılarak daha yüksek bir doğruluk oranı elde edilmiştir. Resnet algoritması ile ESC-50 için %68,95, ESC-10 için ise %87,25 doğruluk oranı elde edilmiştir.

Anahtar Sözcükler: Ses Tespiti, Ses Sınıflandırma, Çevre Sesleri, MFCC, Makine Öğrenmesi, Yapay Sinir Ağı, Resnet, Transfer Öğrenimi

ACKNOWLEDGEMENT

I would like to express my deep thanks to my thesis advisor, Prof. Dr. Hasan DAĞ. This thesis was formed thanks to his valuable knowledge, guidance, and motivation.

I would like to thank Necmeddin ÇARKACI, who helped me in all circumstances and kept his doors open for all kinds of problems.

I would like to thank Dr.Aykut Çayır, who guided and encouraged me in every subject during my master's degree. I am grateful to Alperen Özlü for his help in this process.

Last, but not least, I would like to thank my family members, close friends, and colleagues who were with me in this process.

1. INTRODUCTION

Sound is one of the essential components of human perception in nature. The direction, intensity, and time of sound play an important role in defining environmental events, even if people cannot see physically. Thanks to this situation, most of the events faced by people can be recognized. Sound events differentiate with signals consisting of distinctive harmonic features. The signals of sound on-air come to the human ear and recognize it after applying neural processes in the brain. According to these approaches, sound detection or analysis has been developed.

The importance of Multimedia Systems as a part of Information Technologies is increasing day by day because of working with big data. Especially, using audio, video, image, and text in content-based information retrieval shows the importance of computer-based systems. In this retrieval process, converting multimedia items to the format which can be used in computer via feature extraction methods contribute to forming semantic results for the final decision.

Audio classification aims to predict the descriptive tags from a set of tags verified before analysis. In literature, audio classification occurred by three main subdomains: environmental sound classification, music classification, and speech classification. Environmental sound classification contributes to defining environmental conditions **to** achieve a detailed understanding of the acoustic scene itself.

Machine learning applications have had a steady stream of advances in recent years. Computer systems overcome difficult and complex tasks, at times even surpassing human limits. Most of the attractive achievements have come in with image processing, with successful deep learning methods incrementally. Especially using transfer learning or pre-trained model is a part of this achievement.

This study aims to obtain, analyze, and class environmental sounds using ESC- 50 and ESC-10 datasets and develop results for using sound detection in different areas.

Analyzing performance machine learning and deep learning algorithms and trying to different sub-datasets obtained from feature extraction methodology.

2. FEATURE EXTRACTION

2.1. Feature Extraction Methods for Audio Classification

Most of the events around us perceiving with five senses are related to hearing. Even though the sounds around us are very similar, they can have different acoustic properties. Audio analysis is compared to one audio with labeled sound. According to this aspect, Machine learning and Deep Neural networks are used for sound analysis and detection in many studies.

Data acquisition is essential to analyze audio successfully. Data acquisition consists of acoustic data type, record environment, metadata. After that and storing audio data, audio processing is done. In this step, using with Machine Learning and Deep Neural Network algorithms try to achieve good result for classification. Another critical part of the audio analysis is preprocessing because deleting noise or increasing sound level makes the target noise more understandable. The next step and most critical one is Audio Feature Extraction, which analyses the acoustic attribute of audio signals (Çakır, 2019).

In this study, two extraction methodologies are used. For machine learning algorithms, numerical values are used extracted from 11 different methods. Another one is consisting of mel-spectrograms images, which can be input for neural networks directly.

2.1.1. Feature Extraction for Machine Learning

Audio analysis is based on feature extraction from audio signals. The cost of machine learning and deep neural networks is directly proportional to features. The features of audio are represented numerical values, which can be based on signal energy, frequency distributions, and time-frequency. Most of the audio features are obtained from dividing audio signals into windows and analyzing spectrogram or statistical information of signals, as shown in Figure 2.1 and Figure 2.2.

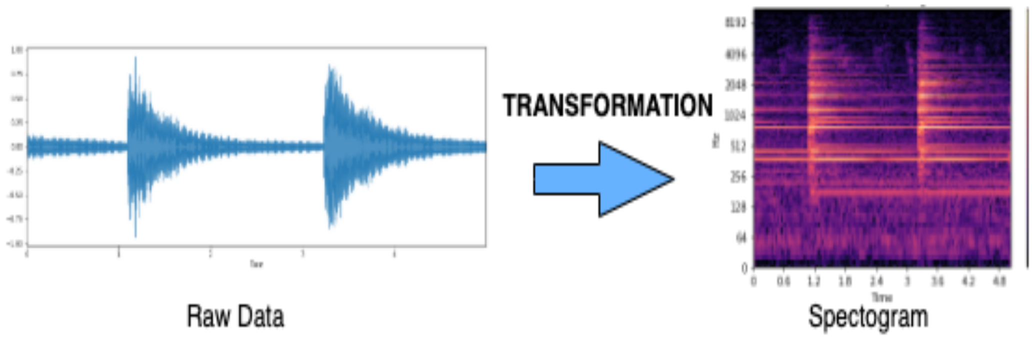


Figure 2.1: Transformation of Raw Audio Data

The processes of feature extraction of audio are dividing frame, dividing window, creating spectrum, creating spectogram, which is described in below:

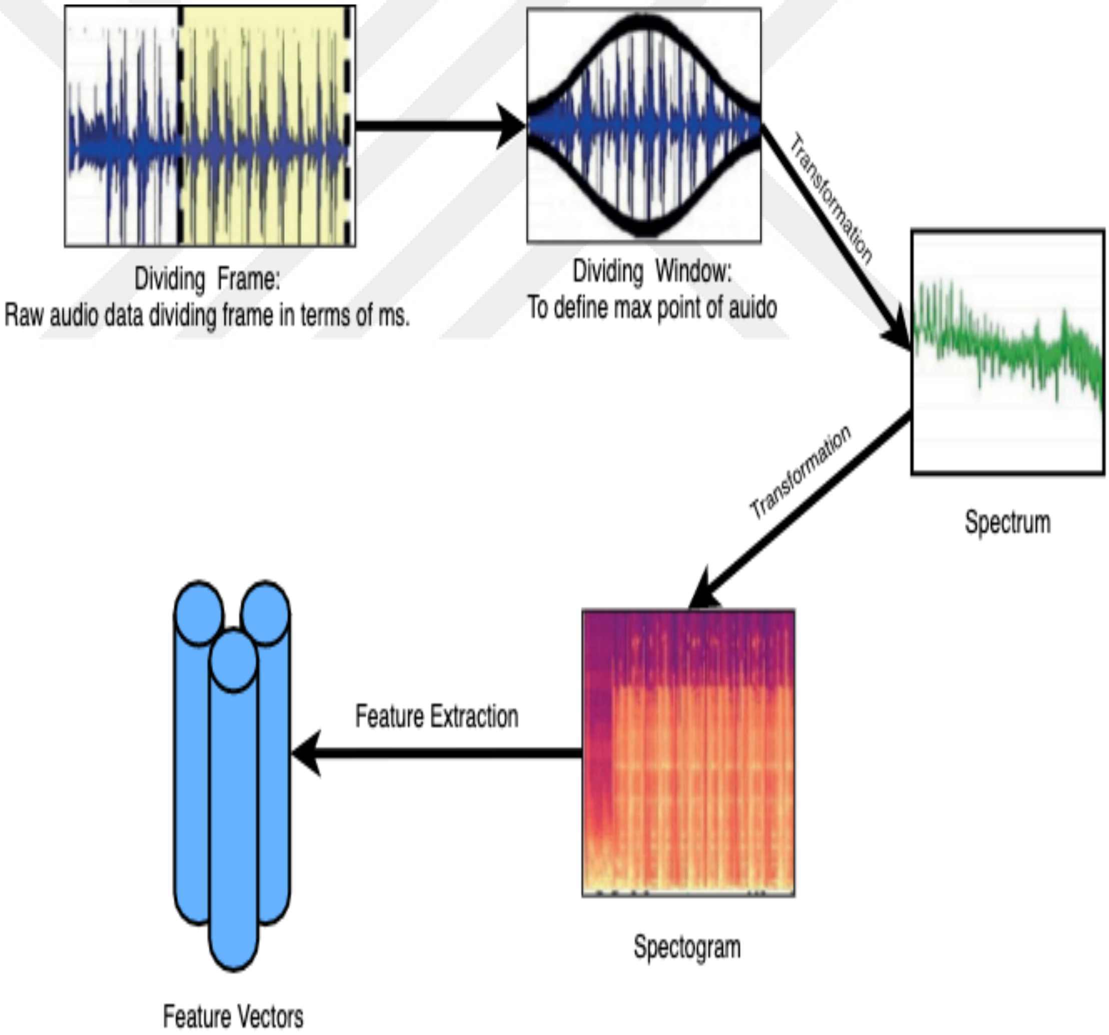


Figure 2.2: The Feature Extraction Methodology (Akustik Öznitelik Çıkarımı-Karasulu, B.,2019)

2.1.1.1. Short Time Fourier Transform

The Short-time Fourier transform uses to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time (Chengjin Xu, Junjun Guan, Ming Bao, Jiangang Lu, Wei Ye,2018). In real life, the procedure for computing STFT s is to divide a longer time signal into smaller segments of equal length. After that, Fourier Transform is calculated on each shorter segment. Plots the changing spectra as a function of time is created, known as a spectrogram or waterfall plot (E. Sejdic ´ et al, 2009).

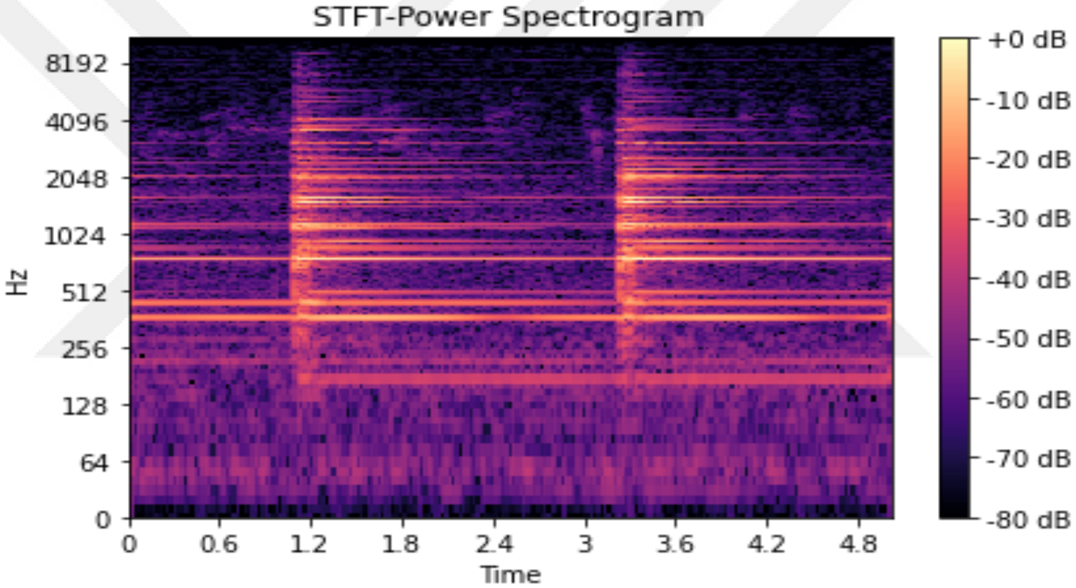


Figure 2.3: Short-Time Fourier Transform

2.1.1.2 Constant Q Transform

The Constant Q transform is a technique that converts a signal from the time domain into the frequency domain as it shown in Figure 2.4. In contradistinction to the Fourier transform, the center frequencies of the frequency-bins are spaced geometrically. These properties make it more suitable for the analysis of music than the Fourier transform. In the Fourier transform, the frequency bins are linearly spaced, which is not perfect for the examination of musical data. Low frequencies can have a resolution much lower than needed, or the opposite of this situation can be happened (Brüder, 2013).

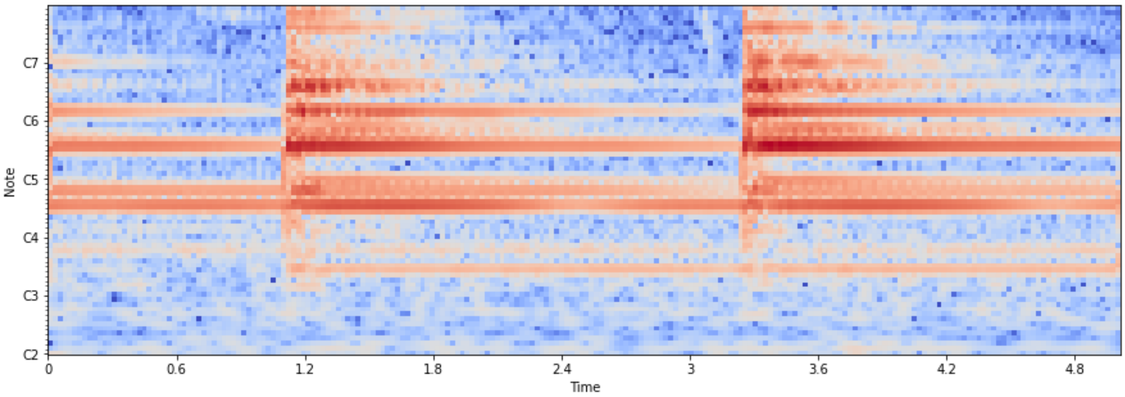


Figure 2.4: Constant Q Transform

2.1.1.3 Spectral Features

The spectral features are based on frequency, which is obtained by transforming the time-based signal into the frequency domain using the Fourier Transform, like spectral centroid, spectral bandwidth, spectral density, spectral roll-off, etc. The spectral features used in this work are explained below.

2.1.1.3.1 Spectral Centroid

The spectral centroid is the feature used in signal processing to define the character of the spectrum. It indicates where the center of mass of the spectrum is placed as it shown in Figure 2.5 (Grey, 1998).

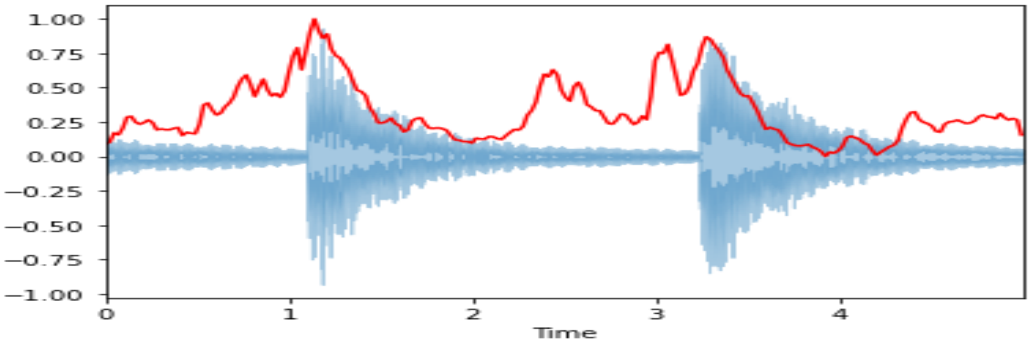


Figure 2.5: Spectral Centroid

2.1.1.3.2 Spectral Bandwidth

The spectral bandwidth is defined as the width of the band of light at one-half the peak maximum and is represented by the two lines and wavelength axis as it shown in Figure 2.6. (Keppy,2008)

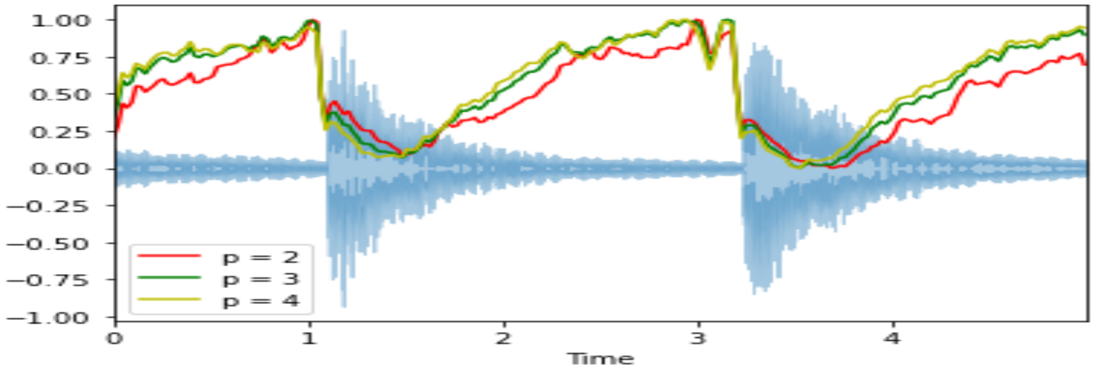


Figure 2.6: Specteral Bandwith

2.1.1.3.3 Spectral Contrast

Octave-based Spectral Contrast considers the spectral peak, spectral valley, and difference in each sub-band. Most music's strong spectral peaks roughly correspond with harmonic components, while non-harmonic components, or noises, often appear at spectral valleys. Thus, the Spectral Contrast feature could roughly reflect the relative distribution of the harmonic and non-harmonic components in the spectrum as it shown in Figure 2.7.

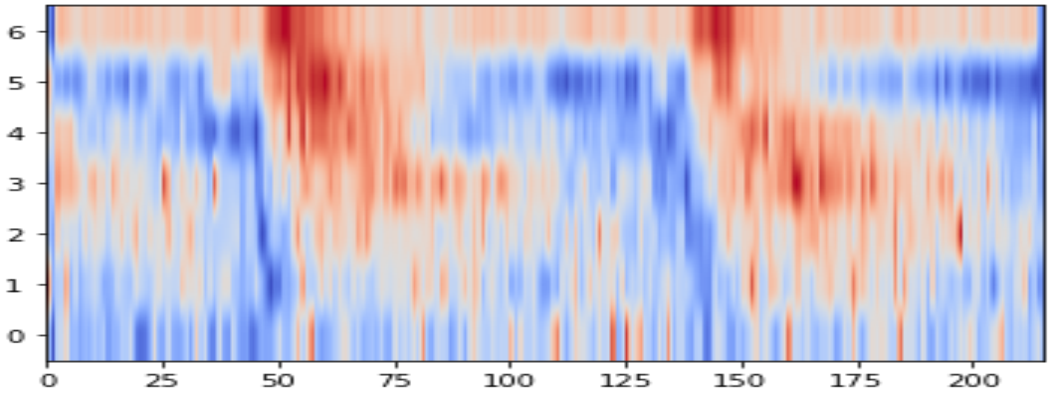


Figure 2.7: Spectral Contrast

2.1.1.3.4 Spectral Roll off

The distribution of spectral energy to low and high frequencies can be characterized by the spectral roll off feature as it is shown in Figure 2.8.

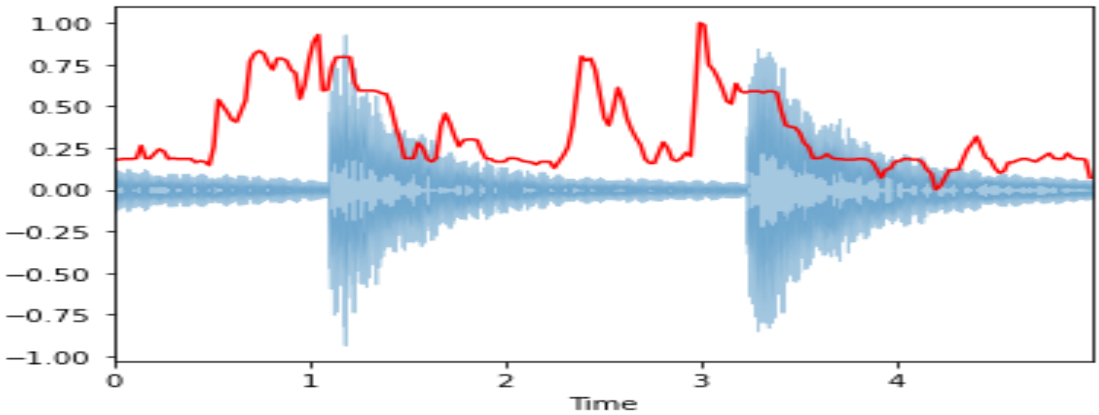


Figure 2.8: Spectral Roll Off

2.1.1.4 Root Mean Square Energy (RMSE)

The root-mean-square energy is most often used to characterize a sound wave because it is directly related to the energy carried by the sound wave, which is called the intensity. The intensity of a sound wave is the average amount of energy transmitted per unit of time through a unit area in a specified direction as it is shown in Figure 2.9.

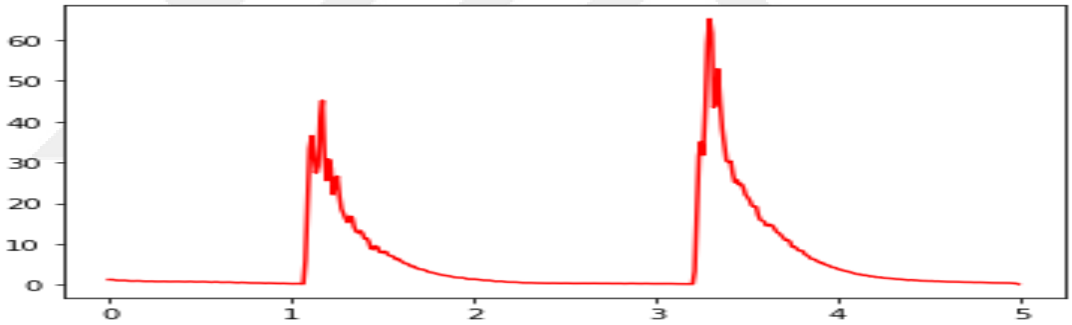


Figure 2.9: Root Mean Square

2.1.1.5 Zero Crossing Rate (ZCR)

The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to zero to negative or from negative to zero to positive (Chen, 1988). This feature has been used heavily in both speech recognition and music information retrieval, being a main feature to classify sounds. It shows that how many times audio signal pass zero signal level(Korkmaz and Boyacı, 2018; Babaee, Anuar, Wahab, Shamshirband and Chronopoulos, 2017). The plot of ZCR is shown in Figure 2.10.

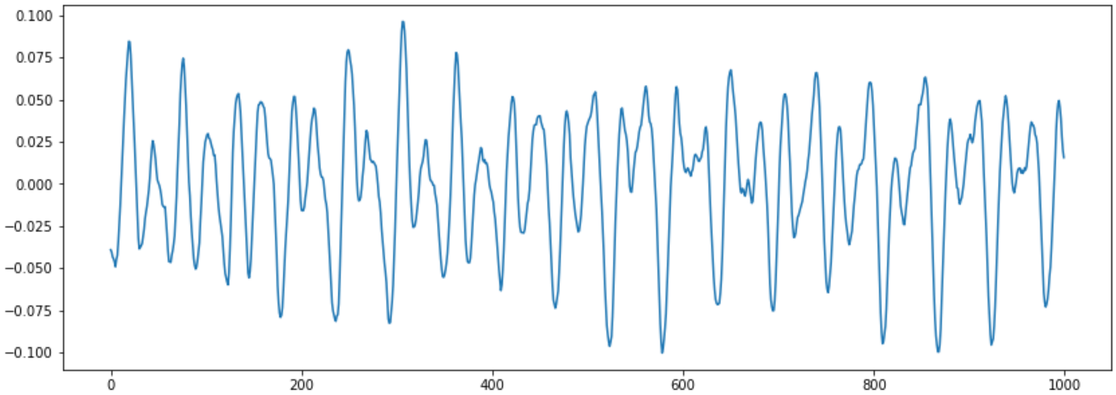


Figure 2.10: Zero Crossing Rate

2.1.1.6 Mel Frequency Cepstral Coefficient (MFCC)

The MFCC feature extraction technique basically includes windowing the signal, applying the Discreate Fourier Transform, taking the log of the magnitude, and then warping the frequencies on a Mel scale. It is computed by passing the Fourier transformed signal through a set of band-pass filters known as the Mel-filter bank. A Mel is a unit of measure based on the human ears' perceived frequency. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly (Karasulu, B. ,2019).

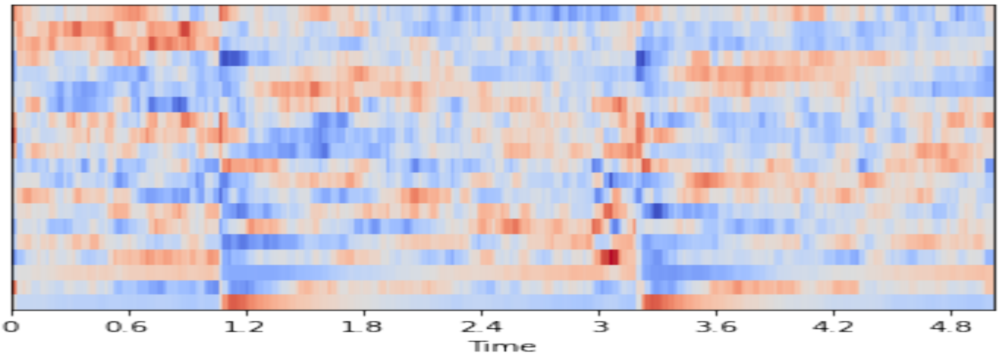


Figure 2.11: MFCC Feature

2.1.2. Image Based Spectrograms for Neural Networks

Raw audio as input is rarely taken by deep learning models. The common approach is to transform the audio into a spectrogram. The spectrogram is an image that describes to audio wave. Therefore, it is very suitable to be input neural network architectures to classify audios.

Spectrograms are output from sound signals using Fourier Transform. As shown in section 2.1.1, Fourier Transform extracts frequencies of the signal and shows the amplitude of each frequency.

A Spectrogram slides the duration of the sound signal into smaller time segments. After that, Fourier Transform is applied to each segment. The reason for this situation is to identify the frequencies related segment. Finally, a single plot that contains the A Spectrogram chops up the duration of the sound signal into smaller time segments and then applies the Fourier Transform to each segment to determine the frequencies contained in that segment. It then combines the Fourier Transforms for all those segments into a single plot.

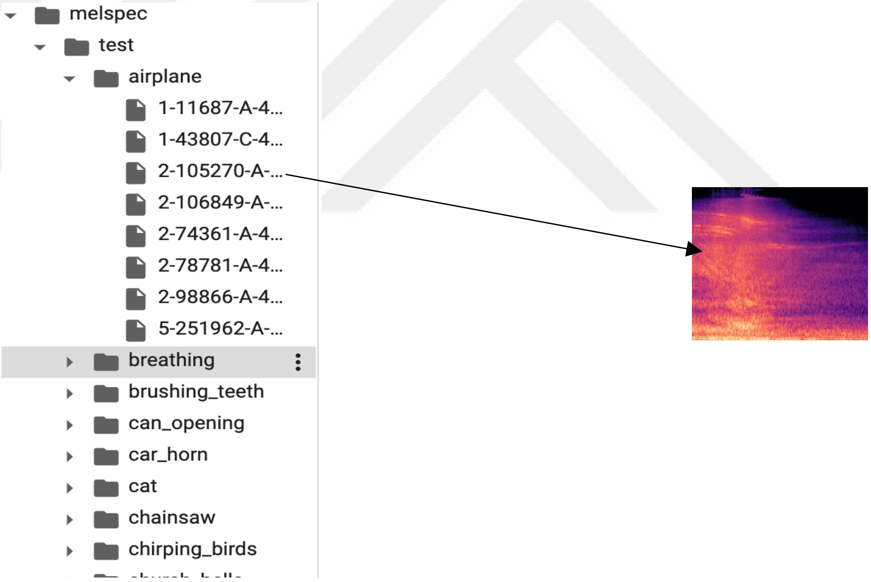


Figure 2.12: Spectogram Image Based Data Set

Image based spectrograms are collected in folders according to class hierarchy as it is shown on Figure 2.12.

3. DATA SET

In this study, two datasets are used ESC-50 and ESC-10. ESC-10 is a subset of the ESC-50 dataset. ESC-50 exists labeled set comprising 50 classes of various environmental sounds. ESC-10 consists of 10 classes. Datasets consist of sound clips extracted from recordings available publicly through the Freesound Project (Karol J. Piczak,2015).

Since available public datasets of environmental recordings are still very limited, otherwise obtaining sound manually or crating new sound data set is very costly. That's why ESC- 50 dataset is used for this study. It is a labeled set of 2000 environmental audio records, which is suitable for sound classification. It consists of 5 sec-long records classified 50 classes (40 samples for each class) that are divided into five major categories, as shown in Table 3.1. ESC-10 Dataset is a subset of ESC-50, which consists of 10 classes; chain saw, clock tick, crackling fire, crying baby, dog, helicopter, rain, rooster, sea waves, sneezing shown in Table 3.2.

Table 3.1: Labels of ESC-50 Dataset

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

Table 3.2: Labels of ESC-10 Dataset

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Crackling fire	Crying baby	Clock tick	Chainsaw
Rooster	Rain	Sneezing	-	Helicopter
-	Sea waves	-	-	-

Classes included in the labeled part of the dataset were arbitrarily selected with the goal of maintaining the balance between major types of sound events, all the while taking into consideration the limitations in the number and diversity of available source recordings and subjectively assessed usefulness and distinctiveness of each class (Karol J. Piczak, 2015). The Freesound database of recordings was queried for common terms related to the constructed classes. Search results were individually evaluated and verified by the author by annotating fragments containing events belonging to the given class. These annotations were then used to extract 5-second-long recordings of audio events (shorter events were padded with silence as needed).

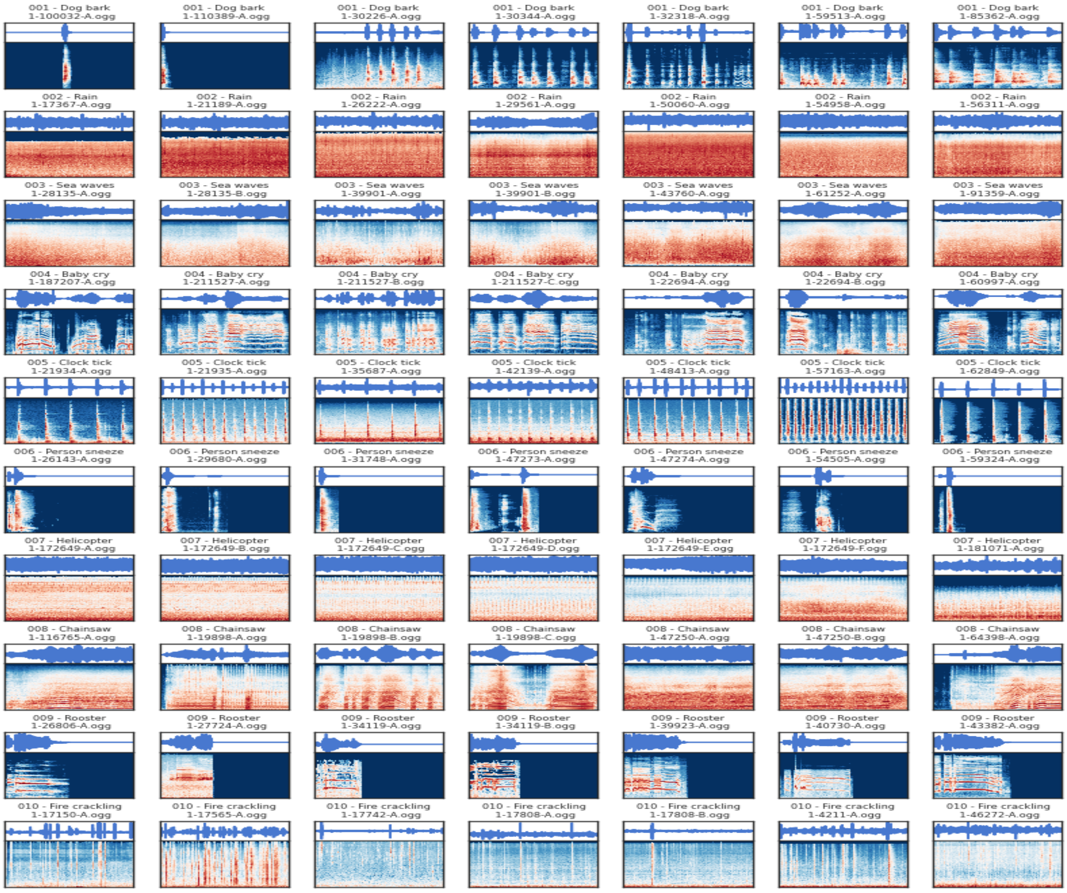


Figure 3.1: Spectogram and Signals of Audio in ESC Data

The extracted samples were reconverted to a unified format (44.1 kHz, single-channel, Ogg Vorbis compression at 192 kbit/s). The labeled datasets were consequently arranged into five uniformly sized cross-validation folds, ensuring that clips originating from the same initial source file are always contained in a single fold. Clips signals and Spectograms are shown in Figure 3.1.

4. MODELS FOR AUDIO CLASSIFICATION

In this study, machine learning and pre-trained neural networks models are used for detecting audio.

4.1. Traditional Machine Learning Models

One of the crucial subsets of computer science is Machine Learning that was improved from a basis of quantitative learning and model identification. The machine learning theory is to create meaningful results that receive input and create outputs as accurate predictions using algorithms and statistical data. The output meaning in machine learning is new algorithms and methods to predict new inputs and systems. The models and algorithms use dynamic data, which means that they can feed themselves by data without remodeling or new codes. After this step, machine learning models are developed, and at the last point, it becomes artificial intelligence and deep learning at the advanced level. Machine Learning models are used in many different sectors, such as telco, bank, security, human behavior, anomaly detection, etc.

Data Cleaning- Preprocessing: In other words, data cleaning is the first step of machine learning processing. Noisy/dirty (unrelated to your result) data can be a reason for some anomalous and redundant results during the analysis and modeling part. Getting rid of complex or irrelevant data is essential to getting successful results from your model.

To obtain Machine Learning Model:

- Training, Validation, and Test Model
- Evaluation of model/algorithms performance
- Check the model inputs& outputs and optimize model.

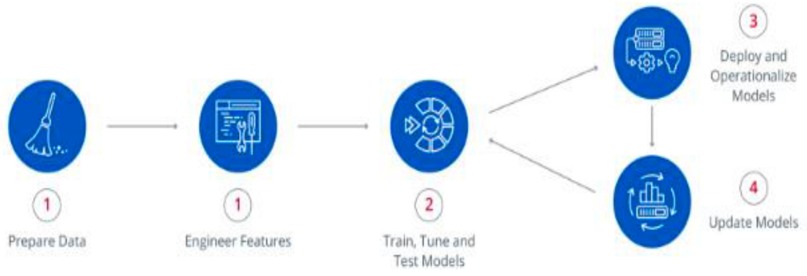


Figure 4.1: The Steps of Machine Learning

Machine Learning methods can be used for different aims depending on target output. Many methods having different purposes can be used together to get more accurate results. Machine learning algorithms can divide into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning is suitable for the cases where class/label is available for a certain dataset (training set) but is missing and needs to be predicted for other instances; regression and Classification methods belong to this category.

Another methodology in machine learning called Unsupervised learning is useful in cases where the challenge is to discover implicit relationships in each unlabeled dataset. There is no training set in this kind of learning algorithm. Clustering and association rules methods belong to this category (Lee J., 2016).

4.1.1. Support Vector Machine (SVM)

Support Vector Machine is one of the most popular methods nowadays and is often used in the field of machine learning algorithms. Basically, there is a plane called the hyperplane. A vector is created on this plane, separating the input variable fields. In SVMs, a hyper correlation is chosen to best distinguish class in points input variables.

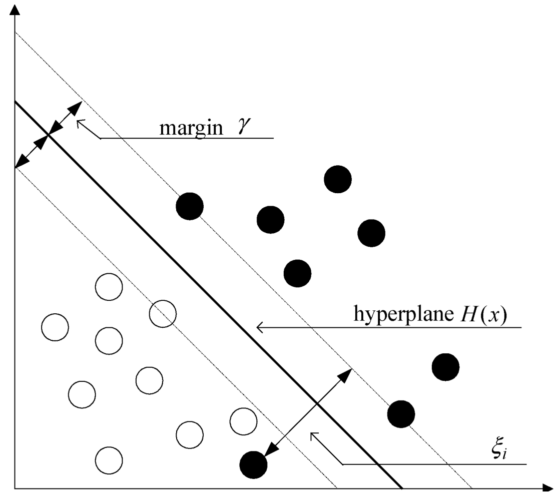


Figure 4.2: SVM Visual

The SVM learning algorithm finds the best coefficients that provide to classes to be separated by the hyperplane. The margin is the shortest distance between the hyperplane and the nearest data plane points. The best or most appropriate hyperplane to distinguish these classes is the line with the largest margin. These points only apply to the definition of hyperplasia and the construction of the regulator. This point is called support vectors. (Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong, & Yukon Chang,2005).

4.1.2. Random Forest Regression (RF)

Random Forests is one of the most efficient machine learning models for analysis, commonly used in statistics and analytics, making this model a work-friendly workload for machine learning. In this model, based on formal decision trees, many models are combined, and predictions are made. The model classes are created with the functions by f below. Therefore, it can be said that kinds of additional models can be created using nodes. On the other hand, it can be defined this model class as follows:

$$g(x) = f0(x) + f1(x) + f2(x) + \cdots \quad (1)$$

The sum of f is called g, which is the sum of simple basic, i.e., base models. It is a basic decision tree for every basic classifier or regression used here. This model is created to get better predictive performance. It is called a multi-variant wide-ranging technical model using multiple models. All the basic models created in random forests are implemented independently using a different sub-sample of the addressed data. When this model is implemented and the sample to be selected, namely the number of forests (turi.com, 2021).

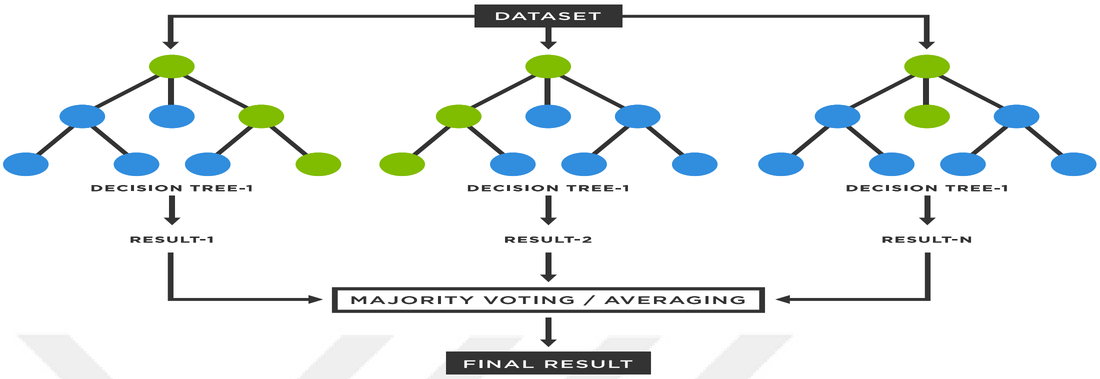


Table 4.3: Random Forest Classification (Random Forest Classification, tibco.com,2021)

4.1.3. K Nearest Neighbor (KNN)

KNN classifier is to classify unlabeled observations by assigning them to the class of the most similar labeled samples. Characteristics of observations are collected for both training and test dataset. It aims to group samples according to a relationship in terms of meaningful distance. To displaying them on a two-dimension plot, only two characteristics are employed. There can be any number of predictors, and the example can be extended to incorporate any number of characteristics. (Zhang, Zhongheng,2016)

There are two important concepts. One of them is the method to calculate the distance between sweet potato and other kinds of food. By default, the KNN() function employs Euclidean distance, which can be calculated with the following equation:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \tag{2}$$

where p and q are subjects to be compared with n characteristics. There are also other methods to calculate distance, such as Manhattan distance.

KNN concept is the parameter k which decides how many neighbors will be chosen for the KNN algorithm. The appropriate choice of k has a significant impact on the diagnostic performance of the KNN algorithm. A large k reduces the impact of variance caused by random error but runs the risk of ignoring a small but important pattern. The key to choosing an appropriate k value is to strike a balance between overfitting and underfitting (Zhang Z.,2014).

4.2. Transfer Learning Models (Resnet)

Deep convolutional neural networks are innovations for classification problems. Especially it is beneficial and provides successful results for image classification. Features and classifiers are integrated by deep networks naturally and end-to-end multilayer fashion, and the importance of features can be improved by the number of layers.

The learning process of machine learning models is supposed to behave like a human brain, which is the aim of transfer learning models. Transfer Learning stores the knowledge of its solution after facing a problem by thinking like us. After that, if the model encounters the same problem again, it will solve it with higher success and faster. In other words, Transfer Learning is to obtain modules that learn faster and more successfully by using less training data by hiding previous information. Therefore, transfer learning provides opportunities in training time, less data, better performance.

In this study, a pre-trained Resnet artificial neural network as a backbone is used for achieving better results from machine learning algorithms. A residual neural network (ResNet) is an artificial neural network (ANN) of a kind that builds on constructs known from pyramidal cells in the cerebral cortex. Residual neural networks do this by utilizing skip connections or shortcuts to jump over some layers. The reasons for skipping connections are avoiding the problem of vanishing gradients or mitigating the accuracy saturation problem, where adding more layers to a suitably deep model leads to higher training error (He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2016). Deep Residual Learning for Image Recognition). As you can see in Figure 4.4, the skipped connections on the networks are seen despite plain neural networks.

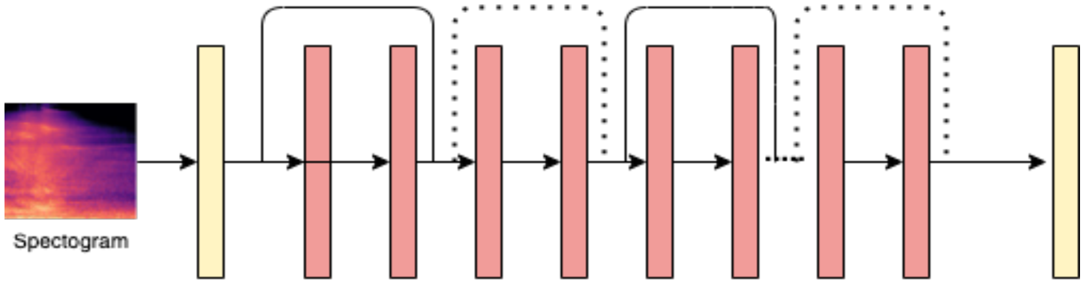


Figure 4.4: Transfer Learning (Resnet)

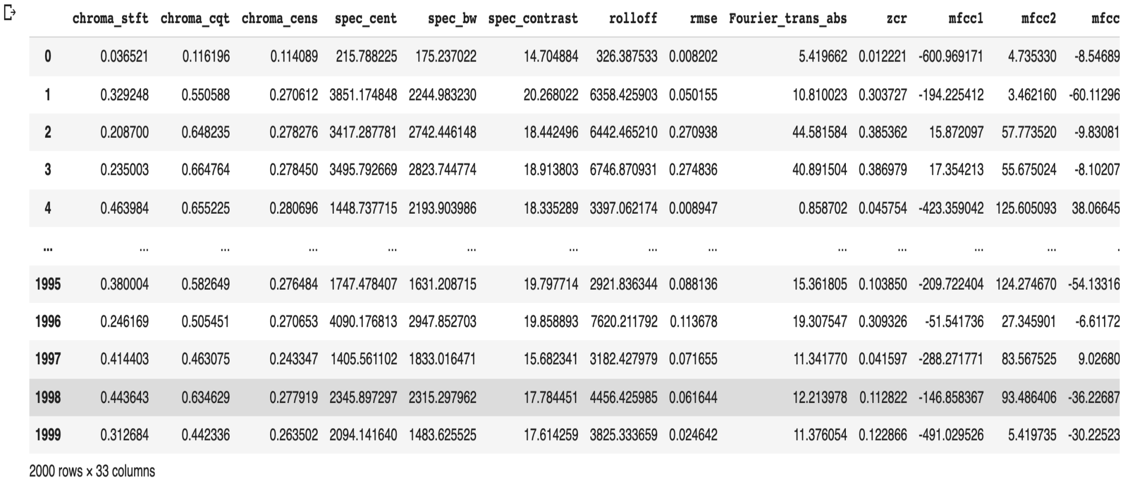
5. METHODOLOGY

This study uses machine learning and deep neural network algorithms to detect environmental audio properly.

The feature extraction methodologies used in this study are explained in detail in section 2. Most of them give vectors of different sizes as a feature. In this case, it is not possible to use different-sized feature set in algorithms. To avoid this situation, two data set were extracted from audios. The first dataset is created by calculating the mean for each feature extraction shown in Table 5.1; the other one is calculated with standard deviation. Both give the machine learning algorithms and feature selection methodologies.

In many of these values, contribution to the importance value of feature that is not normal and classification methods have been used to confirm these anomalies. With the applications of different classification techniques, the accuracy rate of the model was maximum with the SVM classification method. This rate does not show that the classification method is unmeaningful or has a low success rate; it can be developed with other machine learning techniques or deep neural networks.

Table 5.1: Mean of Feature Extraction For ESC-50



	chroma_stft	chroma_cqt	chroma_cens	spec_cent	spec_bw	spec_contrast	rolloff	rmse	Fourier_trans_abs	zcr	mfcc1	mfcc2	mfcc
0	0.036521	0.116196	0.114089	215.788225	175.237022	14.704884	326.387533	0.008202	5.419662	0.012221	-600.969171	4.735330	-8.54689
1	0.329248	0.550588	0.270612	3851.174848	2244.983230	20.268022	6358.425903	0.050155	10.810023	0.303727	-194.225412	3.462160	-60.11296
2	0.208700	0.648235	0.278276	3417.287781	2742.446148	18.442496	6442.465210	0.270938	44.581584	0.385362	15.872097	57.773520	-9.83081
3	0.235003	0.664764	0.278450	3495.792669	2823.744774	18.913803	6746.870931	0.274836	40.891504	0.386979	17.354213	55.675024	-8.10207
4	0.463984	0.655225	0.280696	1448.737715	2193.903986	18.335289	3397.062174	0.008947	0.858702	0.045754	-423.359042	125.605093	38.06645
...	...	...	...	...	...	...	...	...	...	...	...	...	.
1995	0.380004	0.582649	0.276484	1747.478407	1631.208715	19.797714	2921.836344	0.088136	15.361805	0.103850	-209.722404	124.274670	-54.13316
1996	0.246169	0.505451	0.270653	4090.176813	2947.852703	19.858893	7620.211792	0.113678	19.307547	0.309326	-51.541736	27.345901	-6.61172
1997	0.414403	0.463075	0.243347	1405.561102	1833.016471	15.682341	3182.427979	0.071655	11.341770	0.041597	-288.271771	83.567525	9.02680
1998	0.443643	0.634629	0.277919	2345.897297	2315.297962	17.784451	4456.425985	0.061644	12.213978	0.112822	-146.858367	93.486406	-36.22687
1999	0.312684	0.442336	0.263502	2094.141640	1483.625525	17.614259	3825.333659	0.024642	11.376054	0.122866	-491.029526	5.419735	-30.22523

2000 rows × 33 columns

That success rate is not too bad and unmeaningful in baseline methods, but the purpose of the thesis is that it is possible to provide a more efficient and powerful method for sound detection in baseline machine learning algorithms. To find the best fitting solution to increase the model accuracy with a contribution of the best-selected attributes or using different machine learning algorithms or neural networks is the main aim of this study. For that purpose, the correlation matrix of the data was created to understand the relation and effect between features in Figure 5.1

To build a predictive model with best fitted, the correlation and regression of the features of audio were determined. Full features were taken as input to understand the existing model is successful or not in the python notebook. After that, it was found what are the most important features by using the feature selection method, and a classification algorithm was applied.

The correlation matrix shows that bigger and darker circles are notated as the most relevant feature with the other. As you can see from the heatmap of the correlation matrix, MFCC features are related between them. And spectral features have the same situation with MFCC. ZCR and Spectral Centroid features have a high score on the matrix.

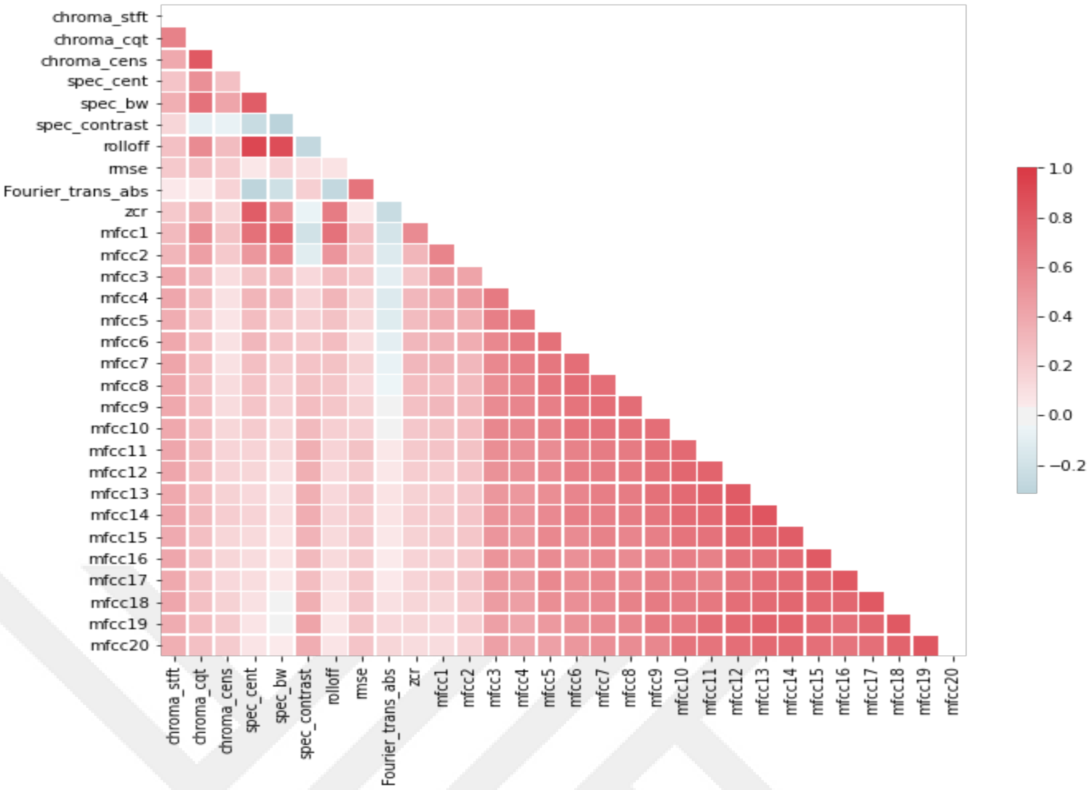


Figure 5.1: Correlation Matrix

Classification machine learning algorithms are applied with two methodologies. The first feature extraction data set is dividing %60-%40 as train and test group. Another one is k-fold validation, k equals 5. After that, a mean of 5-fold is calculated, as is shown in the table below.

Although feature extraction methods mentioned in section 2.1.1, are used in machine learning models, Mel-spectrograms are used as an image mentioned in section 2.1.2 for implementing resnet models. Dataset of resnet models are created according to two ways; k fold (k=5) and %80 -%20 train-test split in shape of class hierarchical. Each model is trained with the parameter of max_epoch as 10.

In this experiment, Lightning Flash, an extension of the Pytorch Lightning framework, is based on transfer learning modeling. It provides to high-level deep learning framework for POC, prototyping, baselining, and solving deep learning problems. It features a set of tasks for you to use for inference and finetuning out of the box and an easy-to-implement API to customize every step of the process for full flexibility

(Lightning Flash, Quick Start, 2021). In this experiment, ESC–50 and ESC-10 dataset is used. Thanks to Lightning Flash, pre-trained models are used to solve many problems in literature.

6. RESULT & INFERENCES

In this study, we have proposed a supervised sound detection algorithm based on SVM, RFC, K-NN with a combination of different feature extraction methods and transfer learning algorithms. Machine learning algorithms outperform classical sound detection methods. Experimental results on SVM present a higher accuracy value than literature baseline algorithms thanks to different feature extraction. Pre-trained resnet algorithms give a result at the state-of-art level.

To discover the best classification performance, 30 features are extracted and analyzed the correlation between them. As a result of this study, the SVM accuracy rate is higher than most of the studies in the literature (github.com,2020). The studies related to this data set; machine learning algorithms' results are not enough to set up a real system. However, the results of neural networks algorithms will perform more accuracy rate in the literature.

In addition to this, we believe that systems become relatively complex, and the accuracy of the classifier decreases if the number of classes increases. The solution to this situation is to extend the data set for similar sound classes. As it can be seen from these kinds of studies, the ESC-50 dataset is not only used itself in most of them; private datasets or data augmentation are used. In this study, we use only ESC-50 and ESC-10 datasets, and a comparison is made with only ESC datasets.

As it is shown in the literature and in this study, machine learning algorithms are indeed a viable solution for environmental sound detection. Conducted experiments show that neural networks outperform more effectively. Although taking into consideration much longer training times, the result is not groundbreaking; it shows that neural networks are effectively used in environmental sound classification tasks even with limited datasets and simple data augmentation.

The machine learning models provide low-level accuracy despite the neural networks. The models in this study, train-test, and validation have done for ESC-50 and ESC-10 datasets. The accuracy of baseline models mentioned in the article of Karol J. Piczak, which models use ZCR and MFCC feature extraction methods and cross-validation (k=5), as shown in Table 6.1 and Table 6.2. The experiments in this study with the same models and feature extraction methods mentioned above show that SVM model accuracy is higher than literature baseline SVM model.

Table 6.1: Accuracy of baseline Models for ESC-50

	TEST		**K_FOLD VALIDATION**		**Baseline Models Result**
	Mean of Feature Set	Standard Dev of Feature Set	Mean of Feature Set	Standard Dev of Feature Set	MFFC & ZCR Feature K_FOLD
K-NN	37,13%	35,88%	30,15%	31,00%	33,20%
RANDOM FOREST	49,63%	46,00%	41,35%	43,00%	44,30%
SVM	48,63%	43,75%	42,15%	41,25%	39,50%

Table 6.2: Accuracy of baseline Models for ESC-10

	TEST		**K_FOLD VALIDATION**		**Baseline Models Result**
	Mean of Feature Set	Standard Dev of Feature Set	Mean of Feature Set	Standard Dev of Feature Set	MFFC & ZCR Feature K_FOLD
K-NN	69,38%	75,00%	65,75%	73,50%	66,70%
RANDOM FOREST	83,13%	75,00%	83,00%	79,25%	72,70%
SVM	79,38%	77,50%	78,50%	77,70%	67,50%

ESC-50 and ESC-10 datasets contain clear, minimum noise audio. The difference between them is the number of classes mentioned in section 3. These datasets have not only advantages but also disadvantages. Clearance, minimum noise, and using a high-quality record system & parameters for recording are advantages. Otherwise, the number of sound clips is not enough for properly modeling to obtain a high accuracy rate. Additionally, some classes seem to be to other ones such as airplane/chain-saw, baby-crying/clock-alarm, etc. This situation can be understood from the human accuracy rate (%81,3), which is listened by people in literature. (Karol J. Piczak,2015).

The accuracy of models is related number of classes directly. If the number of classes increases, complexity can increase. That's why learning rate and accuracy are not satisfied, and accuracy is low such as in this study. ESC- 50 has 50 classes despite ESC-10 having ten classes. As you can see in Figure 6.2, the accuracy rate according to validation and test scenario proves this situation. In these experiments, pretrained resent models are used as transfer learning models. The first thing that catches your eyes is the high accuracy rate despite standard machine learning models. Most of the results are a level of state of the art. Models for ESC-10 dataset, the best model in our study having high accuracy % 87,25 is better than one related work of them which is given on Figure 6.5. The same situation is valid for models which are trained with the ESC-50 dataset approximately, as it is shown in Figure 6.5. In this comparison, models using only ESC 50 and ESC-10 datasets are considered. Because most of the related works have used the ESC-50 dataset and private/public dataset together. Taking into consideration, the most successful model in terms of accuracy using the ESC-50 dataset is our resnet101_32x8d model in the related works, while resnet101 is best one for ESC-10.

The aim of this study was to evaluate whether transfer learning and machine learning model can be successfully applied to audio classification tasks, especially considering the limited nature of datasets available in this field. Conducted experiments show that transfer learning and neural network are more suitable solutions for this problem.

Table 6.3: Accuracy of Resnet Models for ESC-10

ESC-10 DATASET		
Model	**Accuracy (K_FOLD)**	**Accuracy (TEST)**
resnet101	87,25%	95,00%
resnet50	85,25%	88,75%
resnext101_32x8d	83,75%	90,00%
resnet18	83,50%	85,00%

Table 6.4: Accuracy of Resnet Models for ESC-50

ESC-50 DATASET		
Model	**Accuracy (K_FOLD)**	**Accuracy (TEST)**
resnet101	64,25%	76,25%
resnet50	63,30%	68,50%
resnext101_32x8d	68,95%	72,25%
resnet18	64,85%	69,25%

Table 6.5: Accuracy for the models using ESC-50 in Literature

Models in Literature (ESC-50)	Accuracy
Piczak ConvNet (Piczak, 2015b)	64,50%
auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks:Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, Björn Schuller(2017).	64,30%
Classifying environmental sounds using image recognition networks: Venkatesh Boddapatia , Andrej Petefb , Jim Rasmussonb , Lars Lundberga(2017)	63,20%
Soundnet: Learning sound representations from unlabeled video:Yusuf Aytar, Carl Vondrick, Antonio Torralba (2016); 5-layer CNN trained on raw audio of ESC-50 only	65,00%
Resnext101_32x8d (Ours)	68,95%

Table 6.6: Accuracy for the models using ESC-10 in Literature

Models in Literature (ESC-10)	Accuracy
Piczak ConvNet (Piczak, 2015b)	80,50%
Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from betweenclass examples for deep sound recognition," arXiv preprint arXiv:1711.10282, 2017	91,30%
J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283, 2017.	91,70%
Resnet 101 (Ours)	87,25%

One of the possible questions open for the future of this study is whether neural networks based on transfer learning and machine learning methods could be more effective in this topic. Not only in this study but also related works indicate that using transfer learning models can go up with improvements in technology and increasing datasets in this area. Because transfer learning models are better than classic models in terms of training time, less data, better performance. Finally, sound detection or related studies will increase day by day, and the products of these research areas will be part of our life.

REFERENCES

Çakır, E. (2019). Deep neural networks for sound event detection. (Doctoral Dissertation, Tampere University, Finland).
Retrieved from https://tutcris. tut.fi/portal/files/17626487/cakir_12.pdf

Karasulu, B. (2019). Çoklu Ortam Sistemleri İçin Siber Güvenlik Kapsamında Derin Öğrenme Kullanarak Ses Sahne ve Olaylarının Tespiti . Acta Infologica , 3 (2) , 60-82 . Retrieved from https://dergipark.org.tr/en/pub/acin/issue/51284/590690

Chengjin Xu, Junjun Guan, Ming Bao, Jiangang Lu, Wei Ye, "Pattern recognition based on time-frequency analysis and convolutional neural networks for vibrational events in φ-OTDR," Opt. Eng. 57(1), 016103 (2018), doi: 10.1117/1.OE.57.1.016103.

Sejdić, E., Djurović, I., & Jiang, J. (2009). Time–frequency feature representation using energy concentration: An overview of recent advances. Digital Signal Processing, 19(1), 153–183.doi:10.1016/j.dsp.2007.12.004

Lena Sophie Brüder .(2013). Music classification using Constant-Q based Features, Master Thesis Ruhr-University Bochum

Grey, J. M., Gordon, J. W., 1978. Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America 63 (5), 1493–1500, doi:10.1121/1.381843

Nicole Kreuziger Keppy, Michael Allen, Ph.D., Thermo Fisher Scientific, Madison, WI, USA(2008).Understanding Spectral Bandwidth and Resolution in the Regulated Laboratory

Korkmaz, Y. ve Boyacı, A. (2018). Adli bilişim açısından ses incelemeleri. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 30, 329–343.

PICZAK, Karol J. ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. 2015. p. 1015-1018

"ESC-50 Data Set & Studies" Available at: https://github.com/karolpiczak/ESC-50/blob/master/README.md (Accessed: 10 May 2020).

Lee J. (2016) The 10 Algorithms Machine Learning Engineers Need to Know Available at: https://www.kdnuggets.com/2016/08/10-algorithms-machine-learningengineers. html/2 (Accessed: 26 Mar. 2018)

Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong, & Yukon Chang. (2005). Audio classification and categorization based on wavelets and support vector Machine. IEEE Transactions on Speech and Audio Processing, 13(5), 644–651.doi:10.1109/tsa.2005.851880

"Random Forest Regression." Available at:
https://turi.com/learn/userguide/supervisedlearning/random_forest_regression.html
(Accessed: 01May 2021).

"Random Forest, What is a Random Forest?" Avaliable at:
https://www.tibco.com/reference-center/what-is-a-random-forest
(Accessed: 10 Dec 2021).

Zhang, Zhongheng. "Introduction to machine learning: k-nearest neighbors" *Annals of Translational Medicine* [Online], Volume 4 Number 11 (20 April 2016)

Zhang Z.(2014) Too much covariates in a multivariable model may cause the problem of overfitting. J Thorac Dis 2014;6:E196-7

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2016). Deep Residual Learning for Image Recognition 016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE. pp. 770–778

"Lightning Flash, Quick Start " Avaliable at https://lightning-flash.readthedocs.io/en/latest/quickstart.html (Accessed: 10 Sep 2021)
Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, Björn Schuller(2017).auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks

Venkatesh Boddapatia , Andrej Petefb , Jim Rasmussonb , Lars Lundberga (2017)
Classifying environmental sounds using image recognition networks

Yusuf Aytar, Carl Vondrick, Antonio Torralba (2016)
SoundNet: Learning Sound Representations from Unlabeled Video

Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from betweenclass examples for deep sound recognition," arXiv preprint arXiv:1711.10282, 2017

J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283, 2017

CURRICULUM VITAE

Personal Information
Name Surname: Mesut Acar

Academic Background
Undergraduate Education: Information Technologies, Kadir Has University
Industrıal Engineering (Double Major), Kadir Has University
Graduate Education: Management Information Systems, Kadir Has University
Foreign Languages: English

Work Experience
Name of Employer and Dates of Employment:

Turkcell İletişim Hizmetleri A.Ş: 09/2016 - 11/2019
Turkish Radio Television: 12/2019 -