AN APPROACH BASED ON SOUND CLASSIFICATION TO PREDICT SOUNDSCAPE PERCEPTION THROUGH MACHINE LEARNING

A Ph.D. Dissertation

by

Volkan Acun

Department of
Interior Architecture and Environmental Design
İhsan Doğramacı Bilkent University
Ankara
June 2021

AN APPROACH BASED ON SOUND CLASSIFICATION TO PREDICT SOUNDSCAPE PERCEPTION THROUGH MACHINE LEARNING

The Graduate School of Economics and Social Sciences
of
İhsan Doğramacı Bilkent University

By

Volkan Acun

In Partial Fulfilment of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY IN INTERIOR ARCHITECTURE AND ENVIRONMENTAL DESIGN

DEPARTMENT OF
INTERIOR ARCHITECTURE AND ENVIRONMENTAL DESIGN
İHSAN DOĞRAMACI BİLKENT UNIVERSITY
ANKARA

June 2021

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy in Interior Architecture and Environmental Design.

Assist. Prof. Dr. Semiha Yılmazer
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy in Interior Architecture and Environmental Design.

Assist. Prof. Dr. Çağrı Imamoğlu
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy Philosophy in Interior Architecture and Environmental Design.

Prof. Dr. Arzu Gönenç Sorguç
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy Philosophy in Interior Architecture and Environmental Design.

Assoc. Prof. Dr. Yasemin Afacan
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy Philosophy in Interior Architecture and Environmental Design.

Assoc. Prof. Dr. Ozgül Yılmaz Karaman
Examining Committee Member

Approval of the Graduate School of Economics and Social Sciences

Prof. Dr. Refet Soykan Gürkaynak
Director

ABSTRACT

AN APPROACH BAED ON THE SOUND CLASSIFICATION TO PREDICT SOUNDSCAPE PERCEPTION THROUGH MACHINE LEARNING

Acun, Volkan

Ph.D., Department of Interior Architecture and Environmental Design

Supervisor: Assoc. Prof. Dr. Semiha Yılmazer

June 2021

A growing amount of literature and a series of ISO standards focus on concept, data collection, and data analysis methods of soundscapes. Yet, this field of research still lacks predictive models. We hypothesize that machine learning methods can be used to develop a predictive model by identifying the audio content of soundscapes and correlating it with individuals' perceived affective response to the soundscapes. Therefore, this research aims to identify machine learning-based sound classification methods for analyzing the audio content of soundscapes and using its output in a second model for evaluating the association between the audio content and perception of the soundscape. We focused on museum soundscapes to conduct our research. The methodology of this thesis is divided into two parts. For the first part, we used Convolutional Neural Networks for classifying the audio content of the soundscape. Due to their limitations, we used a different approach rather than the typical environmental sound classification methods. We used musical instruments for the training dataset and optimized the neural network for this type of task. The convolutional neural network classified the audio content of the soundscapes based on their similarities to the musical instruments of the dataset. We conducted an online soundscape perception survey to measure participants' affective responses to different museum soundscapes for the second part. To predict individuals' perception of soundscapes, we developed a feedforward neural network model. This model used the audio content output from the sound classification model and the

soundscape survey data to predict the perceived affective quality of soundscapes. We concluded the thesis by conducting statistical analyses to explore the association between the variable used in the predictive model.

Keywords: Soundscape, Auditory Perception, Machine Learning, Artificial Neural Networks, Sound Classification

ÖZET

MAKİNE ÖĞRENİMİ YOLUYLA İŞİTSEL PEYZAJ ALGISINI N TAHMİNİ İÇİN SES SINIFLANDIRMASI DAYALI BİR YAKLAŞIM

Acun, Volkan

Doktora., İç Mimarlık ve Çevre Tasarımı Bölümü

Tez Danışmanı: Doç. Dr. Semiha Yılmazer

Haziran 2021

Bir seri ISO standardı ve gün geçtikçe artmakta olan yayınlar, işitsel peyzajın kavram, veri toplama ve veri analizi yöntemleri üzerine odaklanmaktadır. Ancak literatürde tahmin yöntemlerine henüz rastlanılmamaktadır. İşitsel peyzajda kullanılmak üzere ses içerikleri belirlenerek ve bunların bireyin algısal tepkeleri ile korelasyonu yapılarak, bir tahmin yöntemi geliştirmek için makina öğrenme yöntemlerinin kullanılabileceğini öngörmekteyiz. Bu sebeple, bu tezin amacı işitsel peyzajların ses içeriklerinin analiz edecek makine öğrenimi tabanlı bir ses sınıflandırması metodu belirleyip, bunu kullanarak ses içeriği ve işitsel peyzajın algılanması arasındaki ilişkileri belirlenmesidir. Araştırmada, müzelerdeki işitsel peyzaja odaklandık. Bu çalışmanın metodolojisi iki bölümden oluşmaktadır. Birinci bölümde evrişimli sinir ağılarını (Convolutional Neural Networks-CNN) kullanarak işitsel peyzajların ses içeriklerini sınıflandırdık. Sınırlandırmaları sebebiyle, bu sınıflandırmayı yaparken, , normalde kullanılan çevresel ses sınıflandırması yöntemlerinden farklı bir yaklaşım kullandık. Sınıflandırmaları yaparken, eğitim verileri için müzik enstrümanlarından oluşan bir veri seti kullandık ve yapay sinir ağını (Artificial Neural Network-ANN) bu tür bir göreve uygun olacak şekilde optimize ettik. Evrişimli sinir ağı (CNN), ses içeriklerini her bir ses kaydının spektral özelliklerinin, eğitim verilerindeki müzik enstrümanlarıyla benzerliklerine göre

oransal olarak sınıflandırdı. İkinci bölümde ise, bireylerin farklı müzelerin ses ortamlarına verdikleri hissi tepkileri ölçmek için çevrimiçi bir işitsel peyzaj algısı araştırması yürüttük. Burada ayrıca bireylerin ses ortamlarını nasıl algılayacaklarını öngörecek bir İleri Beslemeli Sinir Ağı (Feedforward Neural Network-FFNN) modeli geliştirdik. Bu model, ses sınıflandırması modelinin çıktısı olan ses içerikleriyle ilgili verileri ve işitsel peyzaj algısı verilerini kullanarak ses ortamının algılanan hissi kalitesiyle ilgili bir öngörü hesapladı. Son olarak, tahmin modelinde kullandığımız farklı değişkenlerin arasındaki ilişkileri daha somut şekilde yansıtacak istatistik analizleri yaparak tezi tamamladık.

Anahtar Kelimeler: İşitsel Peyzaj, Ses Algısı, Makine Öğrenimi, Yapay Sinir Ağları, Ses Sınıflandırması

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Assoc. Prof. Dr. Semiha Yılmazer her encouragement, supervision, and guiding me throughout my graduate years. I want to thank her for the patience and encouragement she showed me during my Ph.D.

I would like to thank my dissertation monitoring commute members Assist. Prof. Dr. Çağrı İmamoğlu and Prof. Dr. Arzu Gönenç Sorguç for their patience, feedbacks, and encouragements. I would also like to thank the Examining Committee members Assoc. Prof. Dr. Yasemin Afacan and Assoc. Prof. Dr. Özgül Karaman Yılmaz for their feedbacks.

I want to thank Dr. Cengiz Yılmazer for his patience and guiding me when I struggled, especially with audio signal processing, and Dr. Patricia Davies for her insightful comments and suggestions. I also thank the Interior Design and Architecture Department graduate students for helping me reach out to participants for my survey.

I want to thank my father, Bülent Acun, for his endless support and for always being there for me. I would like to express my gratitude to Berkan Zıraman and Tuğcan Selimhocaoğlu for their suggestions during my research. Finally, I would like to thank the rest of my close friends for their support, encouragement, and helping me finding participants for my survey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier transform
DL	Deep Learning
FFNN	Feed Forward Neural Network
FFT	Fast Fourier Transform
FT	Fourier Transform
k-NN	k-Nearest Neighbors
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
RNN	Recurrent Neural Network
STFT	Short-Term Fourier Transform

CHAPTER I

INTRODUCTION

The soundscape approach is concerned with how individuals communicate with their environment through sound. According to sound quality research, only 30% of the noise annoyance is caused by the physical aspects of the sound, such as sound energy (Rainer, 1997). Therefore, it is hard to say that a decrease in sound levels will directly lead to increased quality of life. Barry Truax, one of the pioneers of soundscape research, states that a sound's meaning is partly related to its source, but it is mainly associated with the circumstances it is heard (Truax, 1984). The typically used energy transfer approach fails to fully reflect the subjective perception of acoustic environments since it cannot capture the listener context. Models built this way generalize poorly to different settings since they neglect the effect of perception of the sound environment on the listener (Ooi et al., 2020).

Soundscape started to be a research field of acoustics in the late sixties. In 1978, pioneer soundscape researchers defined it as an environment of sound perceived or understood by an individual or society(Schafer, 1977). In 2002 the European Commission introduced the Environmental Noise Directive 2002/49/EC (European Parliament and Council of the European Union, 2002), which aimed to establish a common approach among the member states to identify, protect, and plan quiet areas. This led the soundscape approach to receive significant attention. Due to its holistic basis, numerous

methods from different disciplines have proposed and applied research. Recently, research institutions are paying considerable attention to this subject. ISO working group TC43/SC1/WG54 so far published three standards regarding the definition(International Organization for Standardization., 2014), data collection(International Organization for Standardization, 2018), and data analysis (International Organization for Standardization, 2019) for soundscapes. European Research Council has also started a project to develop soundscape indices to supersede the energy-based measures for environmental sound quality (Kang et al., 2019; Ooi et al., 2020).

The models used for predicting soundscape perception are most commonly linear models, based on objective inputs of either raw audio files or psychoacoustic descriptors (Hong et al., 2020; Sun, Filipan, et al., 2019). However, correlating the subjective perception of soundscapes with objective measurements often end up requiring nonlinear computation. Few studies have incorporated Artificial Neural Networks (ANN), a subset of Artificial Intelligence (AI), to address this issue. However, they were mostly limited to specific sites and merely used AI as a data analysis tool (Puyana Romero et al., 2016).

AI is a discipline within computer science closely related to cognitive science, psychology, philosophy, linguistics, and mathematics (Kurfess, 2003). It aims to create systems that can exhibit some human-like intelligence. They are designed to simulate human reactions, such as understanding, analyzing, planning, defining, perceiving, linking data and variables. This can be achieved by either emulating the ways humans perform particular tasks or developing techniques more suitable for computers to execute. AI algorithms can perform these tasks at a decent speed with flexibility and have the ability to redesign themselves in long terms after a series of learning and testing cycles. More advanced deep learning algorithms are capable of experimentation and self-development without any human intervention.

Taking advantage of this rapidly growing new technology can be very useful in soundscape research. One of the uses of this technology is implementing Machine Learning (ML) algorithms, a broad category under AI, for evaluating peoples' subjective evaluation of soundscape. ML can provide the flexibility to observe subjective reactions to different sound elements. By learning to predict the responses to soundscapes, we can modify and design them to create the desired emotional responses.

Rather than using the psychoacoustic parameters or measuring the sound levels of acoustic environments, we decided to use the audio features that reflect the contents of the sound environment. Our primary research priority was to emphasize the context since both our pilot study(Volkan Acun & Yilmazer, 2019) and ISO 12913-1(International Organization for Standardization., 2014) conceptual framework placed context as the critical element of soundscape perception. With this regard, our research is concerned with developing an ML method for classifying the audio content of the soundscapes and using them to evaluate the soundscapes.

Identifying the sound source is the first step of the questionnaire suggested by the ISO 12913-2 standard (International Organization for Standardization, 2018). But this questionnaire only includes a minimal amount of sound sources, traffic, natural sources, human-generated sounds, and other sounds. Half of these sounds do not apply to most indoor environments. We will discuss this questionnaire in detail in the upcoming chapter, but we can say that using this questionnaire has its limitations for indoors.

The majority of the sounds we hear indoors originate from human activity. If we use this questionnaire to categorize the sounds individuals' hear indoors, almost everything will fall under the category of human-based sounds. We can always identify the sound sources beforehand and add them one by one to the questionnaire, but then we will be limited with the settings we identified those sound sources. In a situation where the research includes a large set of worldwide indoor environments, it will be next to impossible to

identify every sound event and add to the questionnaire. Because of this, we are proposing a new sound classification method to identify the audio content. In addition to the limitations we just listed, the existing sound classification methods need to train with countless different sound source types. The other alternative is using a sound event detection algorithm which requires manually identifying the onset and offset times of each sound event to train the network, which is very time-consuming. The model we are proposing uses the signal characteristic musical instruments as references. By doing so, we are cutting down on the parameters required to train the ML model, significantly reducing the amount of training data samples, and we do not need to identify the sound sources of that environment beforehand.

1.1. Aim and Scope of the Research

This research aims to develop a machine learning-based sound classification method for analyzing the audio content of soundscapes and using the output of this model to explore the association between the soundscape's audio content and perceived affective quality. Context is the most influential factor in the perception of soundscapes. We decided to use museums as our case study settings since they can host different contexts under a single public indoor environment. We conducted a pilot study in Rahmi Koç Museum, but we could not conduct the rest of the research in situ due to the pandemic and resort to online methods. Because of this, the scope of this research is limited to reproduced indoor environments. We used videos from various museums around the world and conducted an online questionnaire survey.

This study is divided into two parts (Figure 1). The first part focuses on developing a sound classification model to analyze the sound environment of each museum video and classify its audio content. Classification of environmental sounds is a challenging task. It requires hundreds of audio samples from different types of environmental sounds to train the classification algorithm. Since a sound environment can consist of countless different sound sources, we cannot classify every sound even with the most

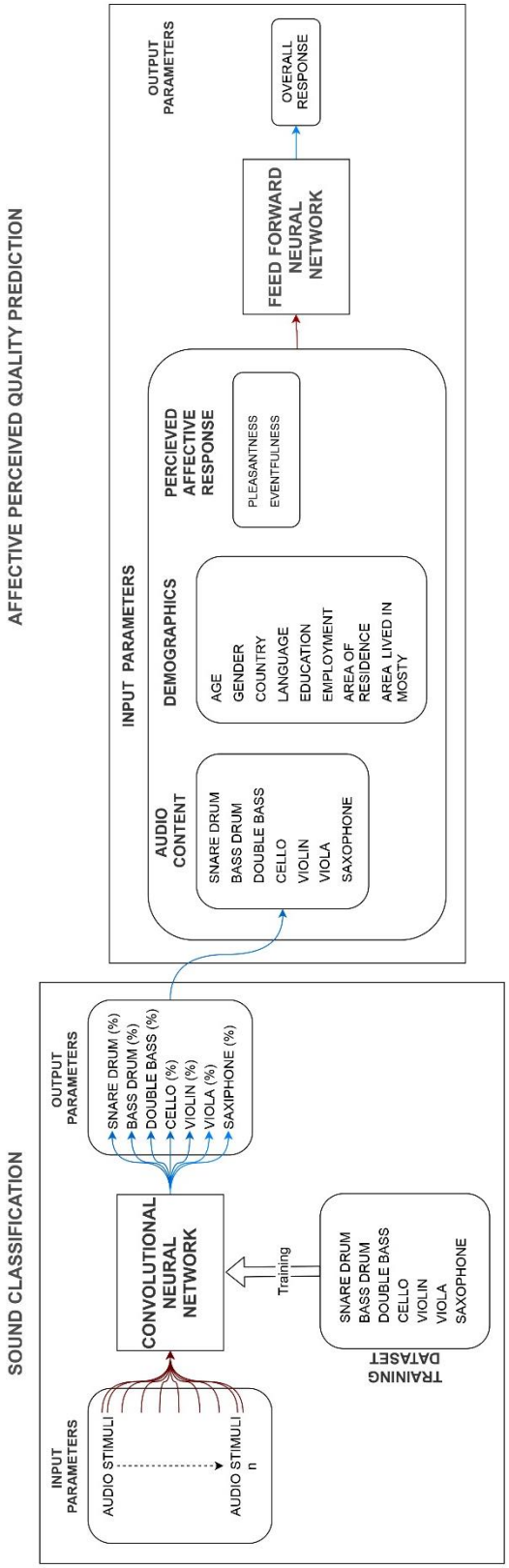


Figure 1: The framework of the research.

comprehensive training datasets. We developed a new approach to address this issue, which uses musical instruments as training samples rather than environmental sound source samples. This classification method is based on using the audio characteristics of the musical instruments like a filter. The Convolutional Neural Network (CNN) model compares the audio features of the recordings with the features of the musical instruments within the dataset. It then computes how much the recording resembles each musical instrument and produces an output in percentages.

The second part of the research is concerned with predicting individuals' perceived affective response to the soundscape. A soundscape perception survey is conducted during this phase of the study. Participants watched several museum videos and rated their sound environment based on their *Pleasantness* and *Eventfulness*. The output from the first phase had already classified the audio content of each video. The soundscape perception survey results are combined with this output and fed into a Feedforward Neural Network (FFNN) to develop a predictive model. Almost all data analysis of the research is conducted with machine learning algorithms. Due to this, the scope of this research is highly interdisciplinary. The main subject of soundscape is supplemented with machine learning and audio signal processing methods.

1.2. Structure of the Thesis

This thesis consists of six chapters. The introduction section provides a brief background and puts the research in context by stating its focus, scope, relevance, and importance. In this section, we explained the objectives of the thesis and how these objectives will provide value to the literature.

The second chapter focuses on conveying the theory behind the research tools we used. The leading theory behind this research is the soundscape approach. However, since soundscape is a highly interdisciplinary approach, we needed to use concepts from the

disciplines such as computer science and electrical engineering to supplement our research. With this regard, chapter two consists of the theoretical background of the soundscape approach, machine learning (ML), and audio signal processing.

Under this chapter, the section of *Soundscape Approach* explains the origins, definition, framework, and standard evaluation methods of soundscape research. We explained how our study relates to the standardization attempts and provided literature examples similar to our research. The second subsection of this chapter is about machine learning. Since the whole data analysis of this research is conducted with ML methods, it is crucial to explain the theory behind the ML application we used: Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FFNNs). The last subsection of chapter two is Audio Signal Processing, which analyzes the signal characteristics of the soundscape recordings. Sound classification methods are used hand in hand with CNNs to classify the audio content of soundscapes. To explain this in the best way possible, we started with the fundamentals of an audio signal and finished this chapter with the sound classification methods.

The third chapter explains the methods we used for this research. The majority of these methods are based on the theoretical background we provided in the previous chapter. In this chapter, we explained how we used these theories to develop our analysis methods. The chapter begins with the research design, which summarizes the study, stating the research questions and hypothesis. The research is divided into two parts the sound classification and soundscape perception survey. Each separate piece has its own data collection and data analysis parts. *Soundscape Perception Survey* uses an online questionnaire survey for data collection. The data collection is less evident for the *Sound Classification* part since it includes finding suitable environmental sound recordings and collecting audio samples for training datasets.

The fourth chapter presents the results of our research. We show the results of the sound classification part and discuss how musical instruments relate to the sound sources within the recordings. We also offer correlation results between the variables, which helps to understand how the FFNN model predicts the affective perceived quality of the soundscapes.

The fifth chapter discusses the findings of our research by comparing them with the literature. The limitations and how the network can be improved are also discussed in this chapter. We also present future applications of this model. Afterward, the sixth chapter concludes the thesis by summarizing the aim and findings of the research.

CHAPTER II

THEORETICAL BACKGROUND

This chapter will focus on conveying the necessary background theory about our research subject, tools, and relevant research examples. This thesis's main objective is to contribute to soundscape research by developing a new predictive model that can later be used to manipulate and design indoor soundscapes. While part of this chapter is about the theory behind the soundscape approach, a larger part focuses on providing the theory behind the data analysis tools we used. These tools are audio signal are machine learning, and audio signal processing, which are used hand in hand to analyze the indoor soundscape perception.

2.1. The Soundscape Approach

The soundscape approach emerged in 1969 when a research group from Simon Fraser University, headed by R. Murray Schafer, started on a new project called The World Soundscape Project(Westerkamp et al., 2006). The project grew out of an initial attempt to draw attention to the sonic environment. According to Schafer (1969), the quality of our sonic environment was in decline ever since the industrial revolution, and contemporary society was indifferent about it because of what he calls the “dominance of

eye culture". Schafer and his research group wanted to raise public awareness of this issue and find an ecological harmony between human society and the sonic environment surrounding them (Westerkamp et al., 2006). The outcomes of this project produced the seminal works (Schafer, 1977; Schafer et al., 1973; The World Soundscape Project, 1978), which now form the backbone of one of the most popular topics in acoustics.

Schafer defines the soundscape as "an environment of sound with emphasis on the way it is perceived and understood by the individual or a society"(Schafer, 1977). Until recently, most environmental policies focused on the noise control approach since sound was mainly considered in its epidemiological aspect of noise (Kang et al., 2016). However, many studies have proven that reducing sound levels not necessarily lead to an improved quality of life(V. Acun & Yilmazer, 2018; Aletta et al., 2017; Aletta & Astolfi, 2018; Brown et al., 2011). Unlike the traditional noise management approach, we cannot use the physical measurements of the sound energy on its own to make an inference about a sound environment. We need to combine the physical measurements with the individuals' interpretation of the sound environment. This made soundscape studies increasingly relevant as they focused on how people perceive and experience the sound environment.

Soundscape research considers the environmental sounds a resource to be managed rather than a waste (Kang et al., 2016). It considers the sonic environment and the listener as part of each other and not as isolated entities. Soundscape studies are concerned with the perceptual construct that is the information exchange between the listener and his/her environment, not with sound energy.

Soundscape research can be conducted in situ, simulated or reproduced, or recalled in memory (Aletta et al., 2016). Experiencing the soundscape in situ provides the most realistic experience. It also offers high ecological validity, yet; it suffers from low experimental control (Aletta et al., 2016). Results gathered from in situ investigations

will most likely only represent the case study setting and will not directly contribute to theory building or general knowledge.

On the other hand, recalling the soundscape from memory is the least realistic and indirect way of experiencing the environment (Aletta et al., 2016). Individuals' ability to recall the environment will significantly affect their responses during the survey. This method is relevant when the participants are familiar with the environment or residents since they can also provide information on how the environment changed over time or know about the typical conditions for the environment (Aletta et al., 2016).

Reproduced or simulated soundscapes are the midway between the other two. They can provide a high amount of experimental control while still providing a somewhat realistic way of experiencing the environment (Aletta et al., 2016). Thus, it is easier to investigate the causal relationships that can contribute to the theory building. One drawback of this model is that the results need to be validated in situ due to limited ecological validity (Aletta et al., 2016).

2.1.1. Standardization Attempts

As a result of its recent popularity, researchers have committed to a standardization process that led to establishing the ISO Working Group of WG54 ISO/TC 43/SC1 in 2008. This working group's purpose was to determine a fixed definition, develop standard data collection methods, analysis, and interpretation (International Organization for Standardization., 2014). According to ISO, the soundscape is an "acoustic environment as perceived or experienced and/or understood by a person or people, in context"(International Organization for Standardization., 2014). The ISO standard also proposed a framework for the perception of soundscape, which placed the context in the center (Figure 2) (International Organization for Standardization., 2014).

Regardless of the conceptual framework, there still wasn't a common consensus on the evaluation of soundscapes. To address this, the working group proceeded to the second

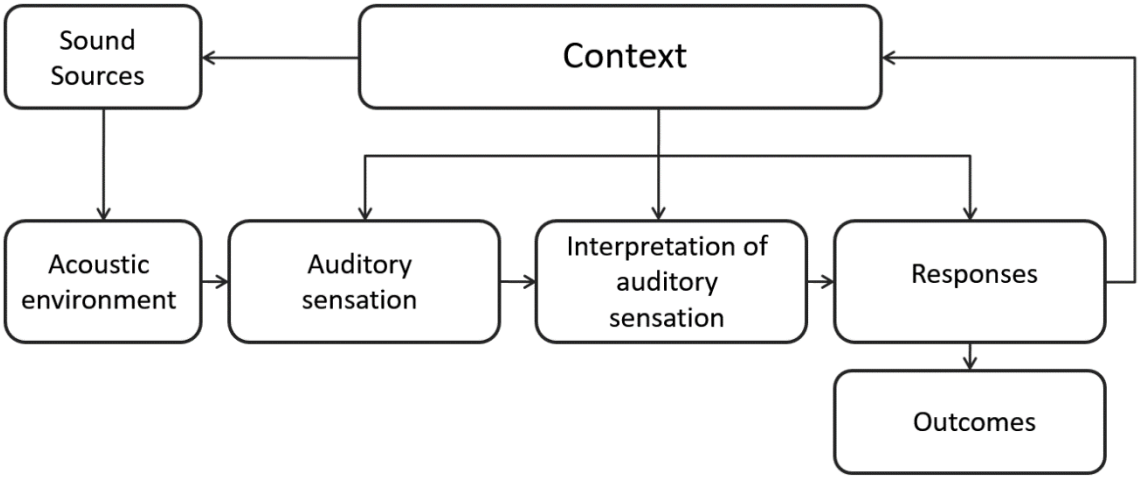


Figure 2: The conceptual framework of the soundscape perception as it is suggested by ISO 12913-1(International Organization for Standardization., 2014).

part of the ISO 12913 (International Organization for Standardization, 2018). This time, the working group's main objective was determining the minimum requirements and supporting information for data collection and reporting soundscape investigations (International Organization for Standardization, 2018). As a result, the working group suggested that data collection can be held through a soundwalk, a questionnaire survey, and a guided interview. They also proposed three protocols for the data collection methods mentioned above. These protocols are referred to as Methods A, B, and C.

Methods A and B are two alternative questionnaires. Method A questionnaire is more of a general-purpose questionnaire, while Method B is orientated towards soundwalks. On the other hand, Method C proposes a protocol for guided interviews, typically conducted off-site and focused on gathering qualitative data (Aletta et al., 2019). Method B is not within our scope since we did not use the soundwalk method as part of this thesis. For our pilot study, we conducted guided interviews as part of a grounded theory survey to explore the Rahmin Koç Museum's indoor soundscape. However, this was before the publication of the ISO 12913-2 (International Organization for Standardization, 2018), which proposed Method C; therefore, we did not have the chance to use this protocol.

The Method A of ISO 12913-2 is more suited to our requirements. This method's questionnaire consists of four parts. The first part is about sound source identification and proposes two slightly different sets of 5-point Likert scale questions, which should be selected based on site conditions. The second part is related to the perceived affective quality of the sonic environment. This part included eight response scales (5-point Likert scale) based on the principal components model of Axelsson, Nilsson, and Berglund (2010). The third and fourth both include only one question. The third part asks the participants, "Overall, how would you describe the present surrounding sound environment?". The question in the fourth part is about relating the context of the physical environment to the sound environment, which asks, "Overall, to what extent is the present surrounding sound environment appropriate to the present place?".

2.1.2. Soundscape Case Study Examples

Regardless of its recent popularity, soundscape studies are mostly limited to urban spaces at the moment. As it lacked a commonly accepted evaluation method, researchers proposed various methods to explore and evaluate soundscapes over the past decade. We can adopt many other techniques for both indoor and outdoor spaces. In this section, we will discuss the soundscape examples that are relevant to our objective.

As we mentioned previously, the perceived affective quality responses used in Method A are inspired mainly by Axelsson and his colleagues (2010). The researchers noticed the need for a model that can identify the dimensions of soundscape perception and conducted a listening experiment. In this experiment, they used 116 attribute scales to evaluate the excerpts of binaural recordings of urban soundscapes. They conducted a principal component analysis of the attribute scale values, which resulted in three components: *Pleasantness*, *Eventfulness*, and *Familiarity*. The first component, *Pleasantness*, is best explained by five attribute scales, Comfortable, Uncomfortable, Inviting, Appealing, and Disagreeable(Axelsson et al., 2010). The second component, *Eventfulness*, is explained by Eventful, Uneventful, Full of Life, Lively, and Mobile

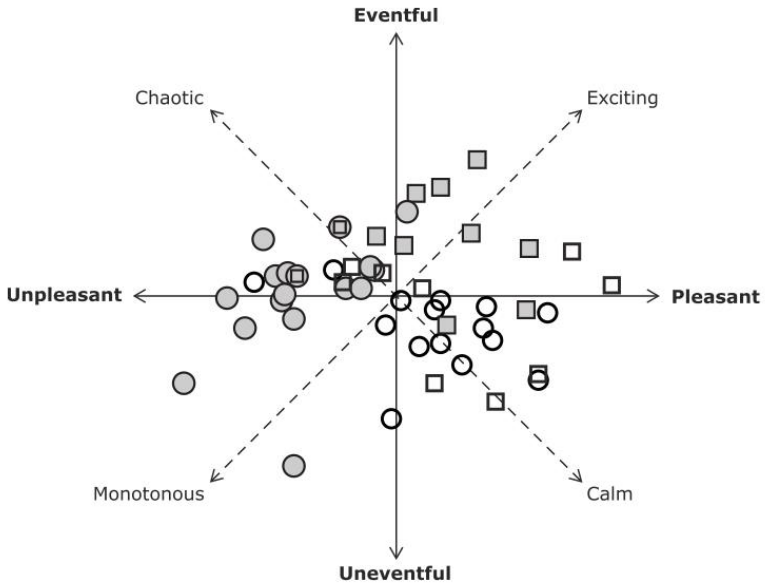


Figure 3: Circumplex figure used by Axelsson et al. (2010) to describe the principal components of soundscape perception:

(Axelsson et al., 2010). The last component, Familiarity, is best explained by Familiar, Rare, Commonplace, Common, and Real (Axelsson et al., 2010).

The authors prepared circumplex models based on the inter-correlation score of each component. The circumplex model suggested by the researchers represented two primary components of the soundscape as a mix of *Pleasantness* and *Eventfulness* (Figure 3). These primary components are supported by the secondary attributes of *Exiting* and *Calming*. This can be interpreted as an exciting and calm soundscape can be perceived as equally pleasant but differ in the amount of eventfulness (Axelsson et al., 2010). Similarly, a monotonous and chaotic soundscape can be equally unpleasant but have a different level of eventfulness. The authors compared this model with the peasantness-arousal model of Russel (1980). They argued that even though eventfulness is not synonymous with arousing, eventful soundscapes can be more arousing and activating than an uneventful soundscape (Axelsson et al., 2010). In their model, Exciting mixes high arousal with pleasure, while calmness mixes low arousal with pleasure.

Compared to Pleasantness and Eventfulness, the variance explained by Familiarity was rather low (8%). The authors suggest that the participants may have perceived the urban context similarly familiar since they come from the same cultural setting (Axelsson et al., 2010). This implication can be useful for us in two ways. Firstly, indoor soundscapes came in numerous settings compared to urban soundscapes, both in physical and acoustic environments. Consider the physical and the sound environment present in a park. While not completely each park can contain different elements, the mental image of a park consists of roughly similar features; thus, it will feel familiar to a certain degree. Each indoor environment, especially the public spaces, contains unique features that will not be familiar to the individuals unless they saw them before. The significance of Familiarity can potentially be much higher for indoor soundscapes compared to urban soundscapes due to this. Secondly, the authors' finding points towards the importance of the cultural background of the participants. A Japanese garden quite distinguishable from a regular park. It can be a familiar outdoor environment for individuals from a particular cultural background. However, it is an unfamiliar setting for someone from a different culture, and their affective response may differ.

One of the more recent indoor soundscape investigations is conducted by Torresin, Albatici, and Aletta (2020). The authors aimed to define and analyze the dimensions underlying sound perception in indoor residential environments and to discuss the possible design implications for such a model. They performed principal components analysis on ninety-seven unidirectional attribute scales to reflect the perceived affective response to indoor soundscapes. The majority of these attributes scales are integrated from the research of Axelsson et al. (2010), while a small portion focus groups and other literature examples. The three main principal components of the study are Comfort, Content, and Familiarity. Comfort and Content together accounted for 83% of the total variance. Indoor soundscapes that are exposed to outdoor sounds, such as heavy traffic found to be annoying.

The authors' findings prove that the content of the soundscape directly affects the perception of the indoor soundscapes as well. Familiarity was also one of the main components of the model proposed by Axelsson et al. (2010), but the variance explained by *Familiarity* has increased compared to outdoor soundscapes. On the other hand, *Comfort* was one of the attribute scales that explained the principal component of *Pleasantness* in the model of Axelsson et al. (2010). By comparing the findings of these two studies, it is safe to assume that the dimensions that affect the perception of outdoor soundscapes are also relevant and can be used to evaluate indoor soundscapes. The one component that is left out for indoor soundscapes is *Eventfulness.* Since Torresin et al. (2020) focused on residential buildings' living rooms, it is logical that *Eventfulness* was not a significant component. However, we argue that eventfulness might become an essential component if the setting was a public space.

The research conducted by Yu and Kang (2009) is among the earliest attempts to use Artificial Neural Network(ANN) to evaluate soundscape quality. The main objective of the researchers was to find a suitable input variable for the ANN models. They conducted a sound level evaluation and acoustic comfort evaluation survey on seven case study sites. They used the data from these surveys to develop one model for the subjective assessment of sound level and another for the acoustic comfort evaluation. They concluded that a general model for sound level and acoustic comfort evaluation was not feasible for all case study sites due to the complex physical and social environments in the urban spaces and the different input variables required by the ANN models for different case study sites. However, they found that using specific models for certain types of locations or functions was more reliable but still did not yield very high model accuracy (Yu & Kang, 2009).

Boes and colleagues (2018) also explored the possibility of using machine listening techniques to achieve an attention-driven human-like auditory environment perception. The model proposed by the authors consists of a three-layer recurrent artificial neural network that identifies how the model notices the input sound. Their findings indicate

that the prediction of how different classes of sounds and soundscape quality indicators will be noticed is better or at least on par with classical sound level indicators (Boes et al., 2018). They concluded that applying new ANN approaches such as deep learning models could potentially advance soundscape research. ,

The most recent attempt to use Deep Learning for evaluating soundscapes was by Ooi et al. (2020). The authors aimed to develop a predictive deep learning model for perceptual attributes of outdoor soundscapes based on the acoustic recordings of the location. So far, the authors only used a single auditory stimulus, bird songs, and not the whole environmental recordings. They used a hundred thirty-second long recordings of bird songs from Macaulay Library and the Korean Broadcasting System. A hundred and twenty-seven participants evaluated the Pleasantness and eventfulness of these recordings on a 7-point Likert scale. Their model performed well as a regression model. Still, the classification performance of their model was relatively poor, which the authors attribute to the low amount of variance caused by the small sample size.

2.2. Machine Learning (ML)

Artificial Intelligence (AI) is a branch of computer science concerned with developing systems that can perform tasks similar to human intelligence. AI makes it possible for machines to perform human-like tasks, learn from experience and adjust to new inputs (Jefferson et al., 2016; Wittek, 2014). Since its early days, the main goal of AI was to construct accurate and testable theories about the mechanism of the human mind by mimicking human cognition. To achieve this, AI relies on Machine Learning (ML) and Deep Learning (DL) algorithms.

Most complex machines and software use algorithms to solve specific tasks (Parker, 2017). Some of these even have adaptive capabilities to adapt to varying conditions. ML, on the other hand, does not have a specific algorithm (Parker, 2017). ML processes the

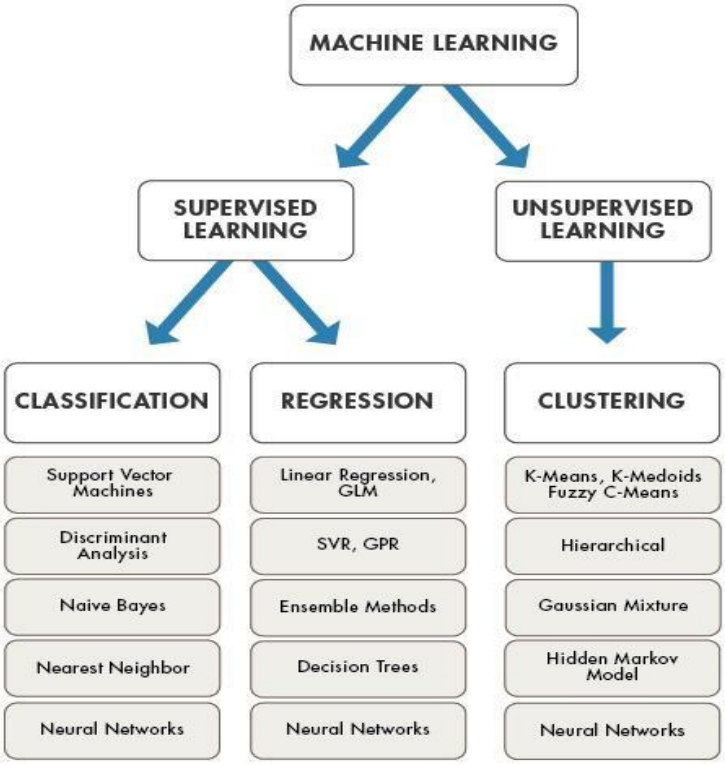


Figure 4: Categories of machine learning and examples of commonly used methods for each category (Boes et al., 2018)

data and produces an output based on a basic structure containing thousands of weights and coefficients (Parker, 2017). These weights are determined by a process called "*training*" which includes passing a large amount of data through the structure and updating the weights via a feedback process(Parker, 2017). The data are typically labeled with the "correct" answer during this process, also called the Ground Truth labels. The weights are adjusted through a feedback process, where the results are known in advance to obtain the desired output(Parker, 2017).

With the recent developments in technology, the performance of ML methods has improved considerably compared to conventional methods. The performance of an ML model depends on the amount of training data and the complexity of the algorithm's model. While it is unnecessary for all cases, the training process can require a significant amount of data depending on the desired outcome. However, a model whose

performance is maximized for a specific task will most likely perform worse on other tasks regardless of the amount of training data (Bianco et al., 2019).

Machine Learning problems can be examined under one of the two categories: *Supervised Learning* or *Unsupervised Learning* (Figure 04). Supervised Learning is the process of training the ML structure based on labeled data. Each observation of the predictor measurement comes with an associated response measurement (label)(Casella et al., 2013). We aim to fit a model that relates the response to the predictors (Casella et al., 2013). By doing so, we can construct a model that accurately predicts the responses of future observations or gain a deeper understanding of the association between the response and the predictors(Casella et al., 2013). A supervised learning model predicts the label of a new data sample, or the data within a test set, after training on a sample of labeled data instances(Wittek, 2014). Supervised Learning is the most commonly used ML category and used for classification (Figure 5) and regression problems (Bianco et al., 2019). Variables within the dataset are the determining factor to decide whether to use a classification or regression model. We can characterize these variables either as quantitative or qualitative (Casella et al., 2013). Quantitative variables consist of numeric values such as age, height, the value of a specific asset, and also referred to as numeric variables. Qualitative variables can take on more than two or more categories without any distinct ordering to the categories. Because of this, qualitative variables are also referred to as categorical variables. For example, eye color (blue, brown, green, etc.) is a

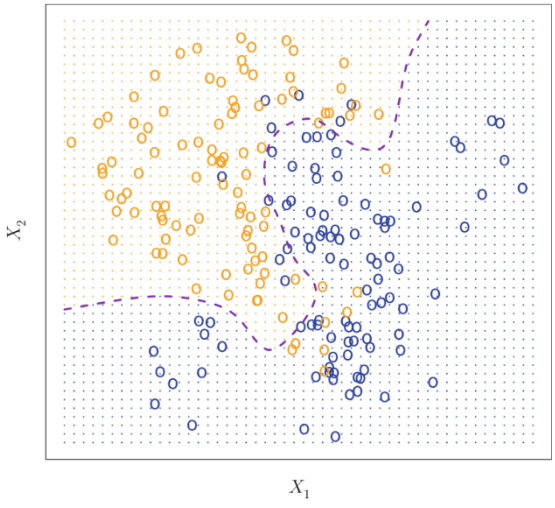


Figure 5: The goal of a classification model is identifying a decision surface to separate different data classes based on the given labelled training instances (Casella et al., 2013).

categorical variable with several categories but without a detailed ordering list (Casella et al., 2013).

We use regression models to address problems that include quantitative responses, while we use can classification models for those that consist of categorical variables (Casella et al., 2013). However, there may not always be a clear distinction between these two methods. Judging by the name, *Logistic Regression* should be a regression method that is supposed to be used with numeric data. Yet, we typically use it with categorical variables and classification purposes, but since its output estimates class probabilities, we can also think of it as a regression method (Casella et al., 2013). Some methods, such as K-Nearest Neighbors(KNNs) or Artificial Neural Networks(ANN), can be used with both numeric and categorical data as long as the data are coded through a preprocess before analyzing them(Casella et al., 2013).

In contrast to supervised Learning, observations are not associated with a label in Unsupervised Learning. Therefore, the training process needs to use unlabeled data to extract structure in the data independently (Casella et al., 2013; Wittek, 2014). Since there is no response label to predict, we cannot fit the model to a regression line. Unsupervised Learning needs to identify the relationship between variables or between observations. One method of achieving this is by arranging data samples into distinct groups based on their similarities, better known as clusters (Casella et al., 2013; Wittek, 2014). In its raw format, clusters can be nested in one another (Figure 06), and the density of the data can vary across the feature space, which makes clustering a challenging task (Wittek, 2014).

2.2.1 Deep Learning (DL)

The AI applications that are our primary concern are found within a subset of ML, called Deep Learning (DL). Artificial Neural Networks (ANN) are the core of DL. Even

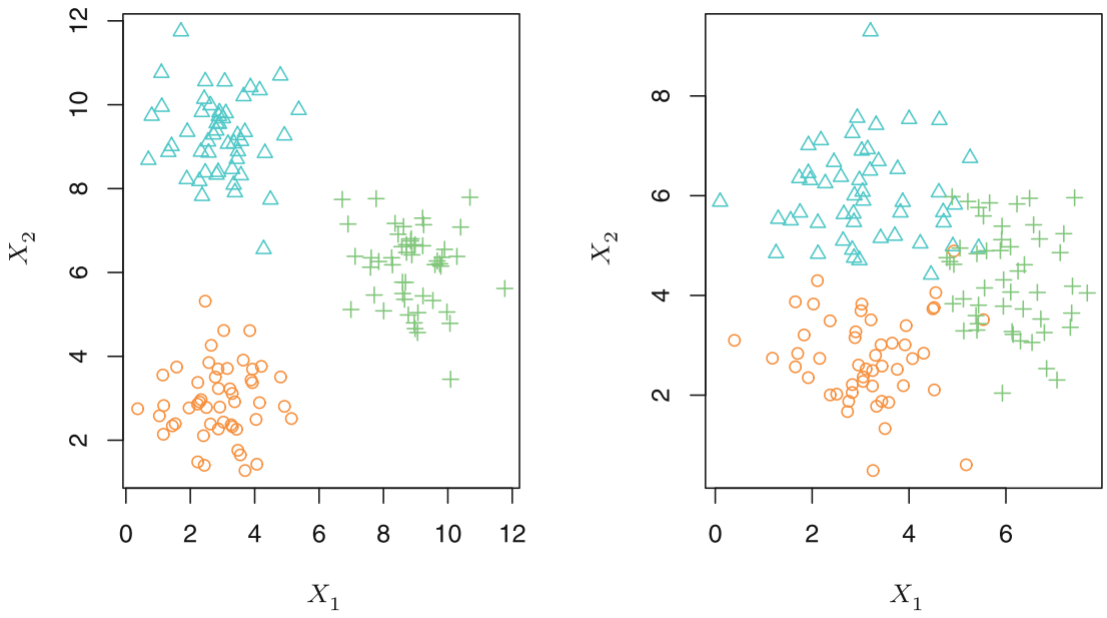


Figure 6: Example of a dataset that includes three distinct groups. While the clustering task is relatively easy when the groups are well-separated(left), it can become a very difficult to successfully identify the three groups when they start to overlap (Casella et al., 2013).

though there is no strict definition of what can be considered as a "deep" neural network, the general rule of thumb is that any network that has more than three hidden layers between the input and output layers can be considered as a "deep" neural network (Kotu & Deshpande, 2019a). ANNs originated from an effort to model the information processing ability of the human nervous system. The fundamental processing unit in the human nervous system is the neuron. Even though not as complicated as the biological neuron, the basic unit of an ANN is also referred to as a neuron (also called a node) since it serves a similar function. ANNs resemble the human nervous system in two respects:

1. Knowledge is acquired from the environment through a learning process (Puente, 2018).

2. The acquired knowledge is stored by interneuron connection strengths, also known as synaptic weights (Puente, 2018).

The functionality of a neuron on its own is simple, yet the connections between the neurons are organized into hierarchical layers and can be very complex(Puente, 2018). It is this interconnected nature of the neurons that defines the functionality of the network. ANN's task is to create a mapping between the behavior of a system and the environment it functions in (Kotu & Deshpande, 2019b). All ANN's simulate four basic functions of a human neuron (Profillidis, Botzoris, Profillidis, & Botzoris, 2019):

1. Receive information from the external environment (input).
2. Decide on what to do with the information (activate and take into account or ignore).
3. Process the information.
4. Present the output of the whole procedure.

ANNs are not a new invention. Using neural networks as computing machines is introduced by McCulloch and Pitts (1943) towards the middle of the Twentieth Century. Despite all their proven capabilities, ANNs were not widely adopted until the 2000s. Three reasons mainly caused this: the lack of readily available data, insufficient

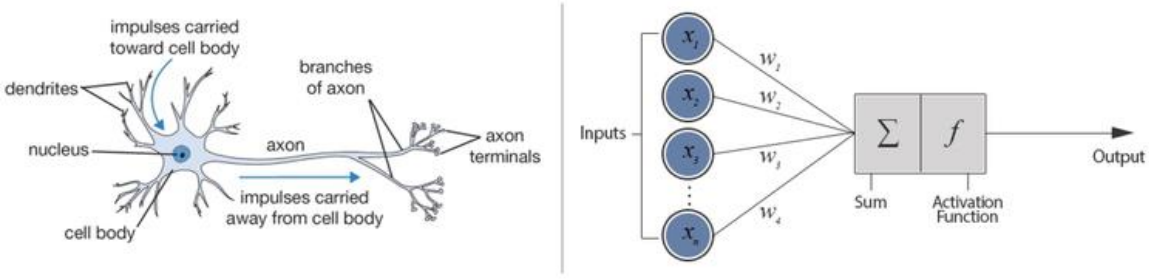


Figure 7: Comparison between a biological neuron and a perceptron.

hardware, and insufficient software (Kotu & Deshpande, 2019a). With the development of the internet in the 1990s, accessing data became more effortless than ever. Today, numerous websites, like Kaggle or GitHub, provide readily available and free-to-use databases. These websites also encourage the development of AI through competitions and free online courses for beginners. The rapid development of Graphics Processing

Units (GPU) of computers significantly decreased the computation time required to process the outcome of a complex ANN. Finally, due to open-source toolboxes and software, building and deploying DL algorithms have become very streamlined and simple. These advancements enabled the development of different types of ANN, which, in turn, increased the popularity and applications of DL. There are three main types of ANNs, Feedforward Neural Networks (FFNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). For the scope of this thesis, we used the FFNNs and CNNs.

2.2.1.1 Feedforward Neural Networks (FFNN)

Feedforward Neural Networks (FFNN) are the simplest type of ANNs. Its name comes from the fact that the information moves in only one direction in this type of network (Poznyak et al., 2019). It moves forward from the input layer to the hidden layer (if applicable) and then to the output layer. The *Perceptron* and *Multi-Layered Neural Networks* are some of the examples of FFNNs.

To understand how FFNNs work, we must examine their most fundamental element and the *Perceptron,* which is also the earliest ANN type. This simplest form of the neural network consisted of only two layers: an input and an output layer. They were developed for basic pattern recognition systems that weighted the data and tested if it exceeded a certain threshold in deciding (Kotu & Deshpande, 2019a).

A *Perceptron* primarily consists of weights, bias, and an activation function (Figure 07). Weights and activation functions are the conduits that carry over the relationship between the input and the output layers. The weights are equivalent to the *R* coefficient of a regression model. Each input parameter x_i has weights w_i associated with it. At its core, *Perceptron* computes the product of $\Sigma w_i x_i$, then passes it to the activation function. If the

activation function evaluates the weight of $\Sigma w_i x_i$ above a certain threshold, then the output is set to 1 (true) or otherwise to 0 (false).

The process of calculating the weights (w_i) is called "learning" or "training" the *Perceptron* (Kotu & Deshpande, 2019a). The data is presented to the Perceptron's inputs, and the output is computed according to the initial weights. The weights are modified based on the difference between the actual output and the desired output and xi inputs. The learning rule of a perceptron can be summarized as (Kotu & Deshpande, 2019a):

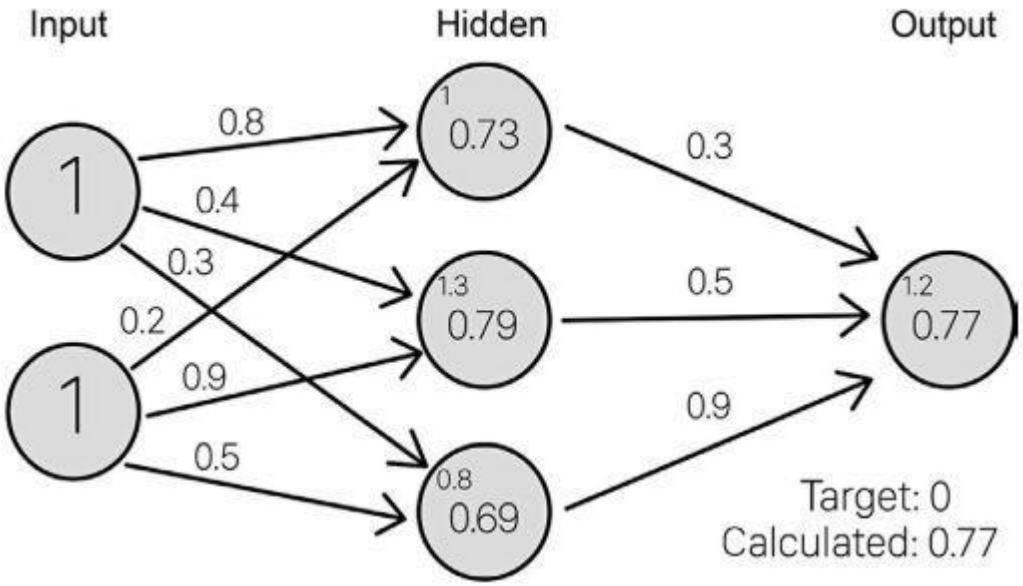
1. Initialize the weights to small random numbers.
2. Feed the inputs xi to the Perceptron and calculate the output.
3. Update the weights according to: $w_i(t + 1) = w_i(t) + \eta(d - y)x$

   Where:

   d is the desired output,

   t is the time step, and

   η is the learning rate, where $0.0 < \eta < 1.0$
4. Repeat steps 2 and 3 until the iteration error is less than a user-specified error threshold or until a predetermined number of iterations have been completed.



The significant limitation of this earliest ANN implementation was its poor performance with nonlinear data (Guarascio et al., 2018). AI scientists overcame this issue by adding a third type of layer between the input and output neurons. This new layer type is called the *Hidden Layer,* and neural networks with three different layers are called *Multi-Layer Perceptron (MLP).* Nonlinear activation functions are also implemented with this type of network.

First MLPs consisted of very few layers, an input layer, a maximum of 3 hidden layers, and an output layer. The neurons of each layer are interconnected hierarchically. Each neuron is only connected with the neurons of the next layer and never with the neurons of the same layer. This was good enough for basic pattern recognition and feature extraction tasks; therefore, they were used for image classification and classification. However, they were proven to be insufficient when a full-scale scene analysis was needed (Kotu & Deshpande, 2019a).

Figure 8: Architecture of a multilayer neural network (Guarascio et al., 2018).

Figure 08 shows the working principle of an MLP, which is similar to a *Perceptron*. In this figure, the input layer starts with a weight value of 1(weights are represented on the

arrows); the small numbers in the center neurons are the sums of the weight multiplied by the input values directed to them; the large numbers in the center circles represent the resulting output of the activation function controlling the “firing” of that node (Guarascio et al., 2018). The final value in the output node is the sum of the hidden layer neuron values multiplied by the weight associated with its link to the output node; processed through activation, they are transferred to the output layer (or to the second hidden layer if applicable).

The transfer between the layers is achieved through activation functions. Different activation functions (linear and nonlinear) can be selected for the transfers, depending on what the MLP needs to solve. During this transfer, activation functions decide whether the input will be taken into account and the neuron will be activated or not.

Activation functions are essentially a rule-based weighted averaging scheme (Kotu & Deshpande, 2019a). As we already described it at the *Perceptron*, the activation function evaluates the weighted average whether it passes a predetermined threshold or not. The position and scaling of the threshold are determined based on the desired functionality. The most commonly used activation functions are; Rectified Linear Unit (ReLu), Sigmoid, Hyperbolic Tangent(Tanh), and Softmax.

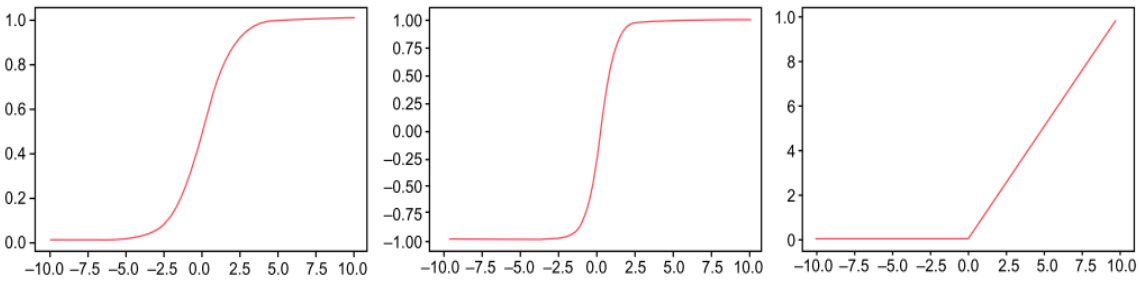


Figure 9: Some of the commonly used activation functions: Sigmoid(left), Tanh(Center), and ReLu (Right) (Kotu & Deshpande, 2019a).

Figure 9 shows examples of some of these activation functions. Even though Sigmoid and Tanh look similar, we can see that their scaling is different on a closer inspection. Tanh function scales between -1 to 1 while Sigmoid's output scales between 0 to 1. Due to this scaling difference, sigmoid is used when we want a probabilistic output. ReLu activation function, on the other hand, is very different when compared to the other two. The output linearly increases once the weighted average exceeds the threshold; otherwise, it is set to 0 (Kotu & Deshpande, 2019a). While not represented in Figure 9, Softmax is also a very crucial and commonly used activation function. The main purpose of this activation function is multi-class classification. These activation functions are nonlinear functions designed to provide the ANNs the ability to classify linearly nonseparable data.

Another essential concept in ANNs is *Backpropagation.* It was introduced in the 1980s to overcome some of the limitations of the perceptron training rules. Rather than reverting to the inputs every time the system encountered an error (learning rule no:2), the *Backpropagation* algorithm propagates errors from the output layer back to the input layer(Suk, 2017). This method enabled the network to work much faster and helped build more sophisticated networks capable of reading handwriting.

2.2.1.2. Convolutional Neural Networks (CNN)

Convolution is combining two functions to create a third function. *Convolutional Neural Networks* (CNNs) is a branch of ANNs that use a convolution operation in at least one layer (Aggarwal, 2018). The working principle of CNNs is similar to the feedforward neural networks. They are made out of layers that take the input, adjust weights, adds bias, and perform activations to produce an output. However, CNNs specialized in working with images as inputs. Because of this, their inputs are arranged in a grid structure and use two new types of layers (one of them being the convolution layer) to handle this input.

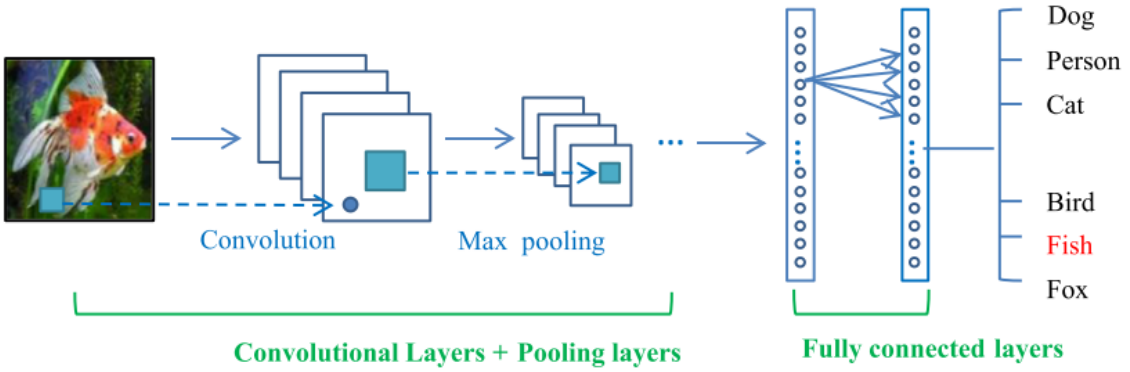


Figure 10: General architecture of a CNN (Guo et al., 2016).

Similar to the traditional feedforward neural network, CNNs are also inspired by the living organism. They are highly effective in solving computer vision problems. (Gu et al., 2018). A CNN consists of three layers: a convolution layer, a pooling layer, and fully connected layers (Gu et al., 2018; Guo et al., 2016). The typical architecture of a CNN can be seen in Figure 10. The convolution layer is responsible for learning the feature representations of the inputs. The data are fed to the convolution layer from the input image. As we mentioned previously, the input for this layer is typically organized into a 2D grid, where each grid point corresponds to a pixel of the input image (Aggarwal, 2018). However, we introduce a third dimension to this grid structure when we need to

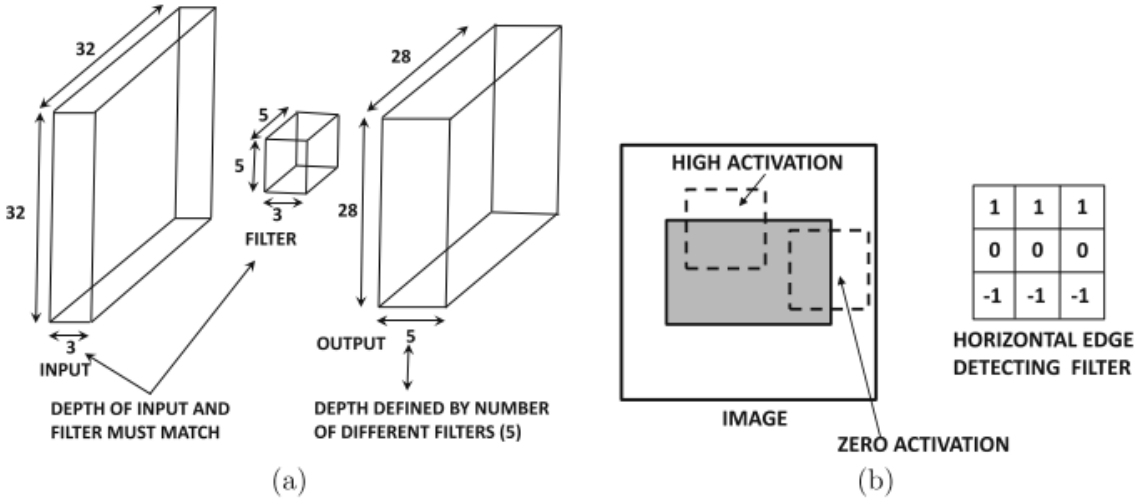


Figure 11: The convolution between an input layer of size 32×32×3 and a filter of size 5×5×3 produces an output layer with spatial dimensions 28 × 28 (a). Sliding a filter around the image tries to look for a particular feature in various windows of the image (b) (Aggarwal, 2018).

represent the image's color, thus turning it into a 3D grid structure (Figure 10). The three primary colors of the RGB color scheme are commonly used for color channels. If we have a 32x32 pixel image, we can capture the color by adding a depth dimension of 3; therefore, the spatial dimensions of the image now become 32x32x3(Figure 11) (Aggarwal, 2018).

2.2.1.2.1 Convolution Layer

Convolution operation uses two signals to create a third signal. In our case, the first signal is the input image, and the second one is the filter that we apply to it. The output of the convolution operation is the feature map. The convolution layer is composed of several filters, also commonly referred to as kernels. A filter, or kernel, is a set of parameters organized into groups of 3D units (or 2D if the input doesn't include color). These kernels aim to map the low-level features, such as edges, to a feature map. Therefore, convolving the input with a filter, a feature map is obtained as an output (Figure 12).

Since each filter captures a different input feature, several different filters are required to obtain a feature map. For example, a horizontal edge detection filter is demonstrated in

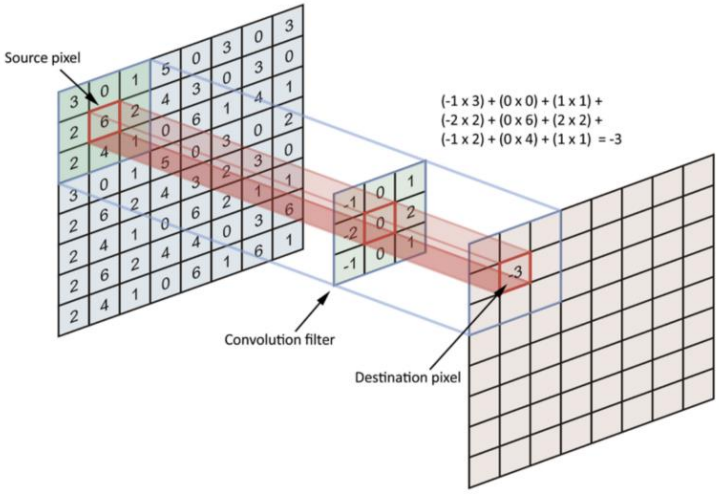


Figure 12: Operation of a Convolutional Layer (Stewart, 2019)

Figure 11. To create a full feature map of the input, we will also require a vertical edge detection filter. If we were trying to recognize a human face, these filters would be oriented towards finding low-level features like edges and curves of our facial features (eyes, nose, etc.). The main advantage of using the convolution layers are:

1. The weight-sharing mechanism reduces the number of parameters in a feature map.
2. The network learns the correlations between neighboring pixels with local connectivity.
3. Invariance to the location of the object.

2.2.1.2.2. Pooling Layer

A pooling layer is typically added after a convolution layer. Pooling layers are used to reduce the dimensions of feature maps and network parameters (Guo et al., 2016). Similar to the convolutional layers, computation of pooling layers also takes neighboring pixels into account. Filters of the pooling layer are designed to capture higher-level features. The curves and edges detected in the first convolutional layer are now beginning to form eyes, noses, ears, etc. Different pooling strategies are designed for different purposes and procedures, such as max pooling and average pooling (Guo et al., 2016).

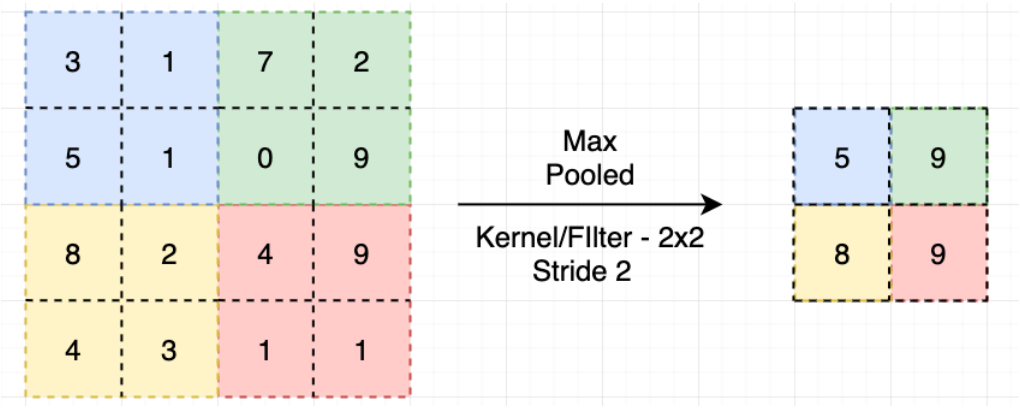


Figure 13: A 2x2 Pooling filter applied over a 4x4 feature map (Rana, 2020).

Figure 13 demonstrates how pooling is applied to the feature map. A max-pooling filter of 2x2 is applied with a stride of two. Stride is the amount of input units the filter shifts after each pooling operation. Since the example uses the max-pooling method, it takes a maximum of 4 numbers. The filter calculates the average of the four values within the given 2x2 depth slice for average pooling.

2.2.1.2.3. Fully Connected Layers

Fully connected layers identical to traditional neural networks (Guo et al., 2016). A CNN usually includes a series of convolution and pooling layer pairs. Fully connected layers are placed after the last pooling layer. They convert the 2D output of the final pooling layer into a 1D feature vector (Figure 14). Using more than one fully connected layer can increase the performance of the CNN. However, 90% of the network parameters are contained in these layers since the neurons are densely connected. Because of this, fully connected layers require a large amount of computational power (Aggarwal, 2018; Guo et al., 2016).

2.2.2. The Challenge: Overfitting and Underfitting

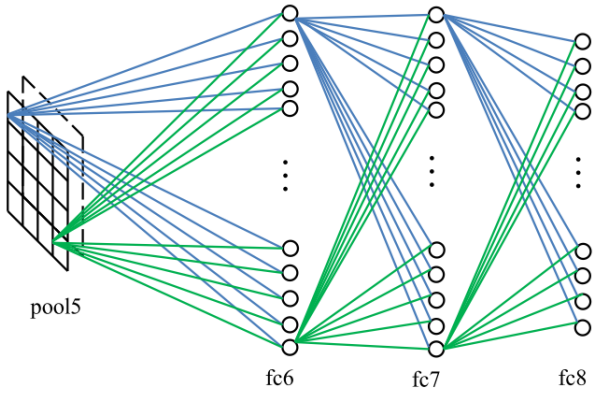


Figure 14: Fully connected layers of CNN (Guo et al., 2016).

A machine learning model's performance is measured by how well it can generalize on previously unseen data. As the model learns from the training data, it computes an error

measure called the training error. The goal of all ML models is to reduce this error as much as possible. However, this is not a simple optimization problem that can be solved by tuning some parameters. Machine learning algorithms also compute a generalization error called the test error, which should be as low as possible. Test error is the measure of an expected error on a new input data. Therefore, a good performing machine learning model should (Goodfellow et al., 2016):

1. Have a small training error (a model with low bias)
2. Have a small difference between test and training error (a model with low variance).

These two qualities of a good performing model correspond to two crucial machine learning problems, overfitting and underfitting. Underfitting occurs when the model cannot obtain a low error rate on the training data, thus have high training error. Overfitting occurs when there is a large gap between the training and test errors, in other words, when the model learns the training data too well (Figure 15).

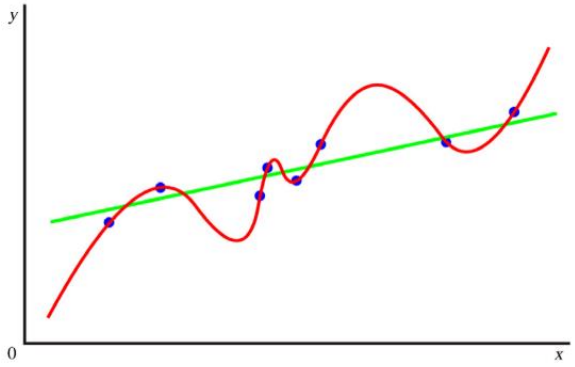


Figure 15: If the model is fitted around the red curve, it will perform too well since it is only appropriate for this particular data will not generalize well. Unless this is done for special reason, the green line will provide higher performance (Davies, 2017).

Overfitted models cannot generalize well and perform poorly on new inputs. The learning algorithm needs to cope with noise, distortions, and fuzziness in the data to satisfy this criterion. However, if the model learns the noise and the fuzziness in the data too well, it will negatively impact the model's performance on new datasets. The algorithm should learn to respond to the underlying population in which the training data have been drawn.

Still, it must not be too well adapted to that specific training set as this will cause it to respond poorly to other data samples drawn from the same population (Davies, 2017). Overall, there is a delicate balance between models' capability to discriminate and its capability to generalize. If the balance is not maintained, the model's performance will be less than desired.

One way to cope with overfitting and underfitting is using the suitable capacity for the complexity of the task. We can summarize the capacity as the number of trainable parameters of the model. Puente (2018) uses the example of a linear regression model to explain the capacity of a model. A linear regression algorithm uses linear functions as part of its hypothesis space. If we modify this algorithm to include polynomials (nonlinear) rather than just linear functions, its hypothesis space will increase hence its capacity. Keeping the model's capacity too marginal limits the model's ability to solve complex problems, leading to an underfitting model. But if the capacity of the model is higher than it is needed, the model will overfit. Therefore, an ML model must have the appropriate capacity for the complexity of the task (Puente, 2018).

Overfitting is more likely to happen with nonlinear and nonparametric models, as they are more flexible by nature. This type of model has more adjustable parameters than necessary for training the data, which is the typical cause of overfitting (Davies, 2017). Overfitting can be prevented by curtailing the training process. The model needs to be periodically tested during the training process to ensure that the point of over-adaptation is not reached (Davies, 2017). As for CNN models, a common method is using *Dropout* layers. A dropout temporarily removes random units from the network with all its incoming and outgoing connections (Puente, 2018).

2.3. Audio Signal Processing and Sound Classification

2.3.1. The Audio Signal

Sound energy occurs when a vibrating object comes in contact with air molecules (or with any gaseous medium), which forces air molecules to move with the vibrating object (Kadis, 2012). Sound propagation is originated from this energy transfer between the air molecules (Kadis, 2012). Our auditory system can use the sound energy to form high-level abstractions to provide rich information about our surroundings concerning locations and characteristics of sound sources (Dabrowski & Marciniak, 2017). Utilizing the functionality of our auditory system in a digital medium results in very useful applications such as the voice assistants of our mobile phones, automatic transcription of speech in videos, and music recognition. But we need to process the audio signal to utilize this functionality.

We can divide audio signals into tones and noise. A tone can be generally described as an audio signal capable of exciting an auditory sensation with a pitch (ANSI, 1995). Tones can be pure tones or complex tones. A pure tone consists of only a single frequency, while a complex tone combines two or more pure tones (*Tone,* 2009). Noise does not necessarily have a pitch. It is divided into different categories based on their temporal and spectral characteristics: stationary, non-stationary, broadband (e.g., white noise), and narrow-band noise (Mitrović et al., 2010). Noise-like sounds have a continuous spectrum, the tonal sounds, on the other hand, have line spectra most of the time (Mitrović et al., 2010).

The four basic dimensions of sound are *Duration, Loudness, Pitch,* and *Timbre. Duration* is the time between the start and the end of the audio signal. A sound signal has four phases: attack, decay, sustain, and release, but not all sounds necessarily have all four phases (Mitrović et al., 2010).

Loudness is one of the more commonly used psychoacoustic attributes of sound. The auditory sensation is mainly related to the sound pressure level changes induced by the audio signal (Mitrović et al., 2010; Wold et al., 1996). American Standards Association (ANSI) explains it more clearly by stating "*auditory sensation in terms of which sounds can be ordered on a scale extending from soft to loud*" (ANSI, 1995). Its unit is called "sone". Loudness sensation is related to sound pressure, frequency content, the waveform of the signal, and duration (Mitrović et al., 2010).

Pitch has several definitions in the literature, based on its use, such as spectral pitch and virtual pitch. For the scope of this thesis, we will use spectral pitch, which ANSI defines as *"that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high*" (ANSI, 1995). The unit of pitch i is called "Mel", which will be described detailly in the following sections. The perception of pitch is mainly related to the frequency content of sound, but sound pressure and waveform also affect its perception (Mitrović et al., 2010).

Timbre is one of the more complex attributes of sound. The ANSI defines it as *"'that attribute of auditory sensation which enables a listener to judge that two non-identical sounds, similarly presented and having the same loudness and pitch, are dissimilar"* (ANSI, 1995). To more clearly explain the timbre, Mitrović et al. (2010) describe the difference between the auditory sensations evoked by a piano and a violin while playing the same note.

Besides the more traditional sound features, which we mentioned briefly, newer features are used together with the features mentioned above. A taxonomy of features for audio classification has been suggested by Mitrović et al. (2010), which distinguishes features based on their domain. A domain is a representation a feature resides in after it is extracted. It is the most vital part of any signal processing application as it allows the interpretation of the feature data. The most typical domains of sound are the time-domain

and frequency domain of sound. Time-domain is the amplitude versus time representation of the audio signal in the waveform. On the other hand, the frequency domain is the amplitude versus frequency representation of the audio signal.

We convert the time-domain of the sound into the frequency domain using an operation called the Fourier Transform. The Fourier transform simplifies many operations required for signal processing, as it enables us to separate, group, and manipulate the spectral components of the audio signal. Short Term Fourier Transform (STFT) allows us to represent the audio signal in a time-frequency domain called the spectrogram, a vital representation of the audio signal's many features.

Based on the taxonomy suggested by Mitrović et al. (2010), the sound features are classified as; *Temporal Domain Features, Frequency Domain Features, Cepstral Features, Modulation Frequency Features, Eigen Domain Features,* and *Phase Space Features*. All categories contain various features of audio signals that can be used based on the desired type of classification. For this thesis, we will only cover three of these.

Temporal Domain represents the signals change over time, and its features are based on the features represented in waveforms such as amplitude, power, and zero-crossing rate (Mitrović et al., 2010; Sharan & Moir, 2016). It is the native domain of the audio signals. All temporal features can be directly extracted from the raw audio signal with low computational complexity. Zero-crossing rate is the rough estimation of the dominant frequency of the audio signal. The power of a sound is the energy transmitted per unit of time (Mitrović et al., 2010).

Frequency Domain Features are examined under two categories, physical and perceptual. Perceptual features consist of the previously mentioned *Loudness, Pitch,* and *Tonality*, in addition to *Brightness, Chroma*, and *Harmonicity*. The *Brightness* of sound measures the

higher frequency content of the signal (Wold et al., 1996). An example of Brightness is, putting your hand over your mouth as you speak. This makes your speech muffled because it reduces the brightness and the loudness of your speech. *Harmonicity* distinguishes periodic signals (Harmonic sounds) from non-periodic signals(noise-like, inharmonic sounds) (Mitrović et al., 2010). *Harmonicity* of the sound helps distinguish between musical instrument sounds (violin and drums) or wildlife sounds (bird song and dog bark).

A combination of *Temporal* and *Frequency Domain Features* of sound has been used for creating one of the earliest examples of content-based audio classification and retrieval systems by Wold et al. (1996). Their system is called the Muscle Fish. It was a commercially successful system that used the acoustical features we mentioned previously, such as loudness, pitch, timbre, Brightness, and bandwidth.

Cepstral Features are one of the more complex features of sound. They are frequency-smoothed representations of the log magnitude spectrum and capture timbre characteristics and pitch (Mitrović et al., 2010). A cepstrum is a domain that provides information about the changes in frequency for different spectrum bands (Sharan & Moir,

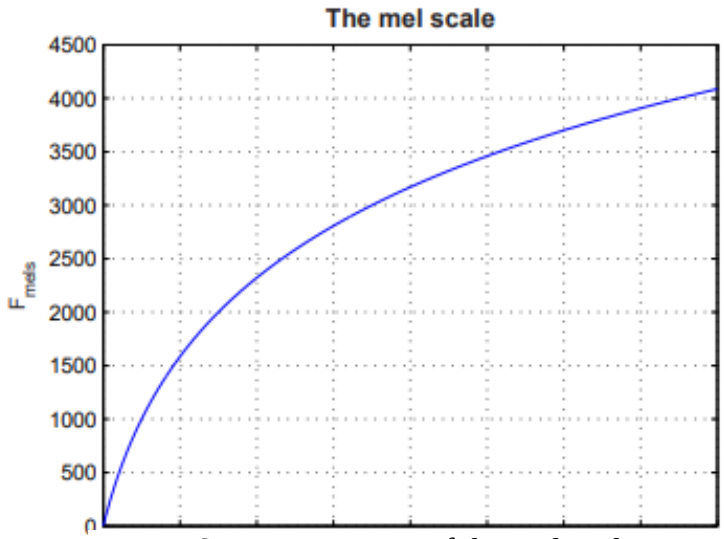


Figure 16: Representation of the Mel-scale.

2016). *Cepstral Features* are widely used for all fields of audio retrieval. We further discuss the applications of cepstral features in the following section.

2.3.2. The Mel-Scale and Mel Frequency Cepstral Coefficients (MFCCs)

The Mel-scale is the psychological scale of pitch (Stevens et al., 1937). Our hearing system is better at differentiating the changes in pitch at lower frequencies than at high frequencies (Sharan & Moir, 2016). When we listen to a pure tone, the difference between 200 Hz and 700 Hz is quite apparent, while the difference between 8000Hz and 8500Hz is barely noticeable. In 1937, Stevens, Volkmann, and Newman (1937) conducted series of laboratory experiments to address this issue. They found that our hearing system perceives pitch linearly between 0-1000 Hz and logarithmically above that threshold (Stevens et al., 1937). Mel-scale is the outcome of their experiment to approximate the human perception of pitch (Figure 16). In Mel-scale, sounds of equal distance also sound to us as they are in equal distance from one another (Umesh et al., 1999).

Mel-Frequency Cepstral Coefficients (MFCCs) are a prevalent audio feature and the primary audio feature for this thesis. AI scientists initially developed it to be used for speech recognition problems, but they found that it can also represent timbre reasonably well and started to be used for Music Information Retrieval (MIR) systems. Today MFCCs are used as the go-to audio feature for almost all types of audio classification problems.

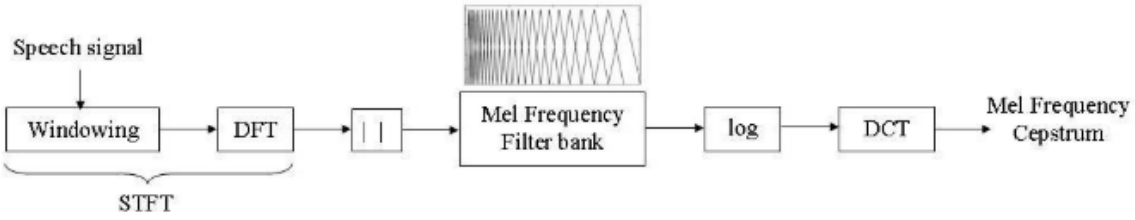


Figure 17: Computation of MFCCs (Kopparapu & Laxminarayana, 2010).

Figure 17 shows how we can compute the MFCCs. The basic procedure to calculate the MFCCs is (Nalini & Palanivel, 2016):

1. The audio signal wave is windowed, and Short-Term Fourier Transform is applied to each window to create the spectrogram.
2. The spectrogram is converted to Mel-Spectrogram by applying the Mel-Filterbank.
3. Taking the logarithmic magnitude of the filter bank
4. Applying Discrete Cosine Transform (DCT) to the filter bank outputs, which create the cepstral coefficients.

MFCCs are considered among the standard techniques of audio classification. However, they perform poorly in noisy conditions(Sharan & Moir, 2016). This issue can be overcome by training the algorithm for multiple conditions, but this also means acquiring large data sets to capture the variations in environmental conditions. MFCCs can work with different classifiers. Sharan & Moir (2017) compared the classification performance of Support Vector Machines (SVM) and DNN for robust sound classification in an audio surveillance application. They found that DNN produced better classification performance and also more noise-robust.

2.3.3. Sound Classification

All audio classification systems' primary goal is to identify perceptually similar audio content (Mitrović et al., 2010). Our brains can distinguish between different sounds and assign them to semantic categories and previously heard sounds. However, the audio signal lacks any semantic meaning for computers and is only represented by a numeric series of samples. This situation poses a semantic gap between the audio signals and the semantics of their contents. Semantic gap refers to the mismatch between the descriptions of a construct in different representation systems. Mitrović et al. (2010) use the example of Beethoven’s Symphony No. 9, which is a series of numeric values for a computer,

while for humans, it is a sequence of notes with high-level semantic concepts like motifs, themes, movements, and emotions.

Every environment contains many different sounds, which are generally divided into speech and non-speech. Sound classification systems cover the areas of content-based audio classification/retrieval (Mitrović et al., 2010), speech and non-speech recognition, audio surveillance (Sharan & Moir, 2015), sound event recognition (Jayalakshmi et al., 2018; Ozer et al., 2018; Stowell et al., 2015), environmental/general-purpose sound classification (Chu et al., 2009), and audio segmentation. Audio surveillance and sound event detection systems are used for security monitoring rooms and public transports, intruder detection in wildlife areas, and for monitoring the elderly (medical telemonitoring) (Sharan & Moir, 2016). Environmental sound recognition, on the other hand, is a more challenging subject. Sound environments include many different sound events which can be present in various combinations at any given time (Sharan & Moir, 2016). Sound recognition deals with the non-speech portion of this divide. It aims to recognize these sounds using signal processing and machine learning techniques (Sharan & Moir, 2016). In theory, they work almost identical to speech recognition systems that are widely used in smartphones, except the input signal is non-speech.

The output of a sound recognition algorithm depends heavily on the input data; therefore, a particular retrieval system is optimized for dealing with a specific dataset. In general, we can cluster sound classification systems we mentioned above under tree headings, speech recognition, music information retrieval, and environmental sound classification. The objectivity of the retrieval results depends upon the availability of standardized ground truths, which are not always readily available (Mitrović et al., 2010). This is especially the case for environmental sound classification. Ground truths are primarily available for speech and music retrieval systems. Still, the biggest challenge in these areas is related to legal and economic reasons (copyrights, high costs due to necessary transcription by humans). For the environmental sounds, on the other hand, the challenge comes from the fact that there is an almost infinite number of environmental sounds

(Mitrović et al., 2010). This also requires the very demanding job of creating a taxonomy of environmental sounds.

Sound Event Detection (SED) is a type of environmental sound classification method. It is concerned with recognizing the onset and offset times of a sound event in an acoustic scene and further labeling it (Adavanne et al., 2017). Sound event is a label that people would use to describe a recognizable event in the region of sound (Heittola et al., 2013). Since the SED systems are usually designed for specific tasks or environments, developing a system that can work in multiple environments or detect a large number of events is challenging. An event can be occurring at various loudness levels and various time durations, the noisiness of the sound recording, or have a limited number of examples to feed into an algorithm. Event types and variance within each category make the SED difficult even with well-represented labels, clean and undistorted signals. The overlapping sound events create an acoustic mixture within the signal that is more challenging.

Regardless of the different sound classification systems, they are fundamentally pattern recognition problems and are handled with similar techniques. A typical sound recognition system comprises three steps; signal preprocessing, feature extraction, and classification. Signal preprocessing prepares the sound for feature extraction. This process divides the sound signal into smaller frames, usually between 10-30 ms, and applies a window function to smoothen it (Sharan & Moir, 2016). Sound recognition systems typically use a sampling frequency of 16000 Hz, 22050 Hz, and 44100 Hz for the training data. Sampling frequency bands selection depends on what application is considered. Speech recognition systems commonly use 8000 Hz sampling frequency while ISO 12913-2 Soundscape Standard suggests 44100 Hz audio recording for environmental sounds (International Organization for Standardization, 2018). Based on the signal's sampling frequency, a frame size of 256, 512, or 1024 samples are used. These frames are chosen with a degree of overlap between adjacent frames to avoid loss of information around the borders of the samples (Sharan & Moir, 2016).

We use the preprocess to simplify the constantly changing nature of the audio signal. Despite this continuous change, on short time frames, the audio signals are statistically stationary. The size of the frame is essential because if it is too short, we will not have enough samples to make a reliable estimate. On the other hand, if it is longer, the audio signal will change too much (Wold et al., 1996).

After the audio samples are preprocessed, we proceed to the Feature Extraction step to get a higher-level understanding of the sound signal. As previously stated, all kinds of sound recognition systems are essentially pattern recognition problems. We obtain these patterns through the feature extraction process. Features represent specific properties of audio signals. Effective representation of an audio signal:

1. Should cover the most significant properties of sound for the selected task.
2. Should be robust under various circumstances
3. Should be general enough to describe various sound classes

Once we extract the audio features of the dataset, we load these to the input layer of the ANN. The network iterates over these features and learns the patterns of each different sound class. When the training process is over, we can use the network to make predictions about the type of previously unseen audio.

CHAPTER III

METHODS

3.1. Design of the Research

This research aims to identify machine learning-based sound classification methods for analyzing the audio content of indoor soundscapes and applying this model for evaluating the association between the audio content and perception of the soundscape. Thus, the research consists of two parts, the sound classification part and the soundscape perception part. We can summarize the research procedure in four steps:

1. Finding indoor soundscapes that include rich audio content,
2. Using an ML-based sound classification algorithm to classify the audio content of each indoor soundscape
3. Conducting a survey way to find out about individuals' perceived affective response to each soundscape.
4. Correlating the resulting audio content with individuals' perceived affective responses for developing a predictive model.

3.1.1. Research Questions

This study has one main research question, but it includes two sub-questions. These questions are:

Can we use machine learning to classify the sound sources that contribute to the audio content soundscapes?

a. Is there an association between the audio content of soundscape and the perceived affective quality of the acoustic environment?

b. Can we use the audio content of soundscapes to measure their perceived affective quality?

3.1.2. Hypothesis

Based on the aim of the research we prepared three research hypotheses. These hypotheses are:

1. Audio signal characteristics of musical instruments can be used to classify the audio content of soundscapes.
2. There is an association between the audio content of a soundscape and its perceived affective quality.
3. Machine learning methods can be used to predict the perceived affective quality of a soundscape with a reasonable accuracy.

3.2. The Preliminary Study

We conducted a pilot study in the Rahmi Koc Museum, located in Ankara (Volkan Acun & Yilmazer, 2019). The pilot study aims to familiarize with the physical and acoustic environment of the museum and identify potential issues that might present themselves during the research. We also wanted to test our survey tools, mainly our questionnaires. The pilot study used qualitative and quantitative research methods. We used Grounded

Theory (GT) to develop a theory and prepared a questionnaire survey to test his theory using the Structural Equation Modelling (SEM) method. The results and the methods we used in this pilot study had a major impact on the course of this thesis as they gave us the idea

The Rahmi Koç Museum consists of historic caravanserais of Çengelhan and Safranhan. Our pilot study was limited to the Çengelhan part of the museum. We selected this particular setting to understand individuals' perception of the soundscapes in a historical context.

We conducted semi-structured interviews with fifteen volunteering participants as part of the Grounded Theory. The majority of these participants were graduate students in the Interior Architecture and Environmental Design Department of Bilkent University, so they were considered experts due to their knowledge about the design principles and physical factors that affect perception. The rest of the participants consisted of friends and family who never been to this museum before. This composition of the participant group allowed us to obtain data samples that can enable the emergence of new categories for the theory and further develop the already existing ones effectively. The interior architects were more capable of providing design-influenced information and details about their observation of space which may go unnoticed by an ordinary visitor. Once the expert participants point out an aspect like this, we confirmed it with other participants and experts, ensuring the full development (grounding)of a category (Volkan Acun & Yilmazer, 2019).

We analyzed the interview transcriptions through the constant comparison method, consisting of open coding, axial coding, and selective coding. We generated seven categories as a result of the coding. We then arranged these categories in a graphical order based on their associations to prepare the conceptual framework of the theory (Figure 18).

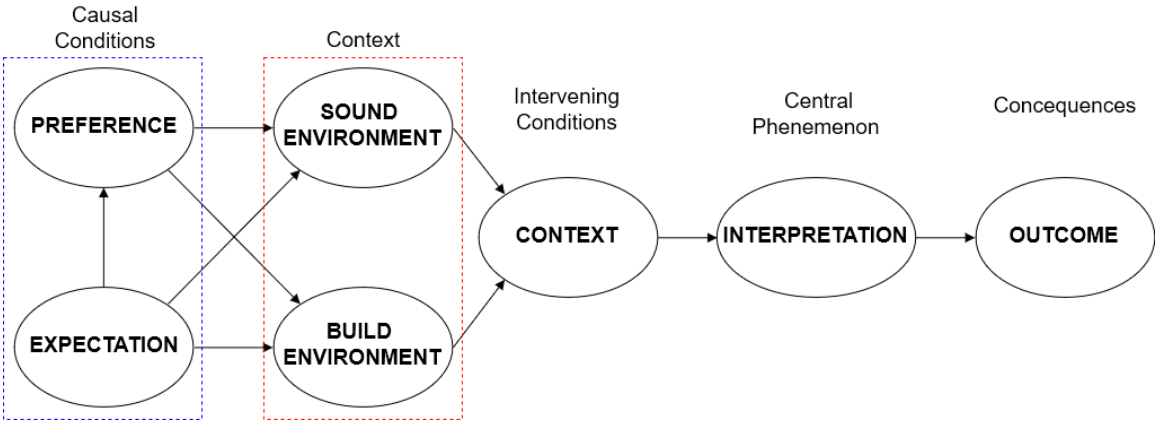


Figure 18: The conceptual framework based on the grounded theory method(Volkan Acun & Yilmazer, 2019)

We used the conceptual framework of part one as the basis for the second part of our pilot study. We prepared a questionnaire survey based on this conceptual framework. Most of the questions are influenced by the information we gathered from the semi-structured interviews, but we also used questions from the previous soundscape research (Davies et al., 2013; Mackrill, Cain, and Jennings, 2013). A total of one hundred and thirteen visitors participated in the questionnaire survey. All participants were randomly selected volunteers. We used IBM SPSS Statistics Software 21 and Smart PLS software for conducting the SEM analysis. Principle Component Analysis with Varimax rotation is used as part of the explanatory factor analysis to preprocess the data. Indicators with low factor loadings are removed, and the Kaiser-Mayer-Olkin score of the dataset was 0.733, which indicated the data was suitable for factor analysis. Confirmatory factor analysis is used for calculating the reliability and validity scores of Composite Reliability, Cronbach's Alpha, and Average Variance Extracted (AVE).

We used Partial Least Squares Structural Equation Modelling (PLS-SEM) technique. This model was more suitable for our data than regular SEM since our sample size was relatively small and the distribution of our data was non-normal. PLS-SEM is also most commonly preferred for explanatory research, which is not based on a well-established questionnaire. As the nature of our pilot study was similar, we decided to use this version of SEM.

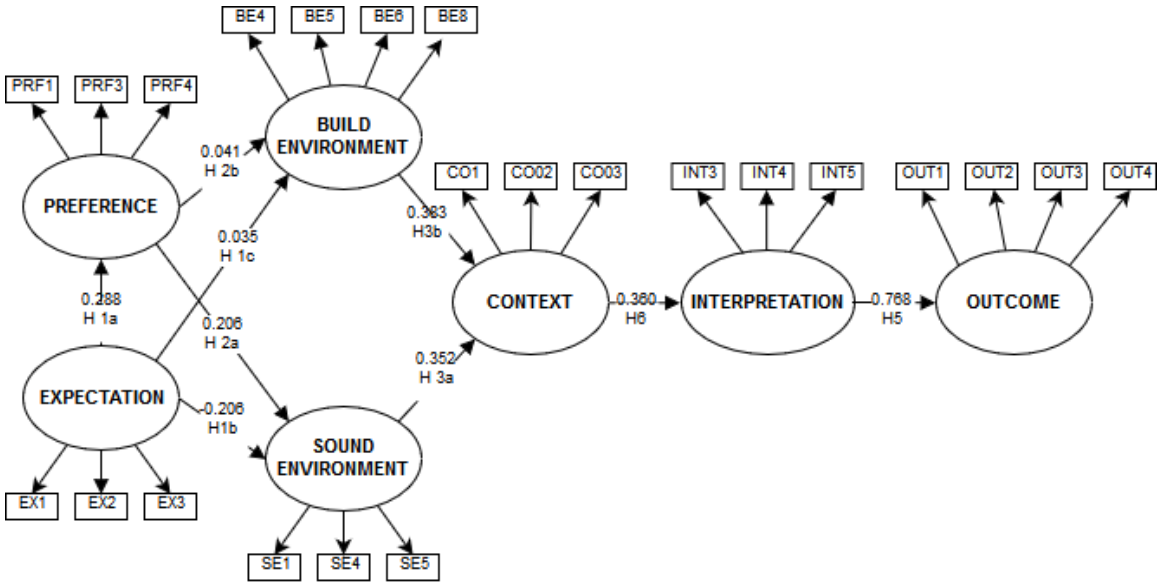


Figure 19: The SEM model with latent variables and path coefficients.

Figure 19 shows the latent variable path model we have created as a result of the SEM analysis. Five hypothesis paths out of nine showed statistical significance. These paths are Expectation – Preference, Build Environment – Context, Sound Environment – Context, Context – Interpretation, Interpretation – Outcomes. As a result of this analysis, we confirmed that context is the most crucial factor that affects the perception of indoor soundscapes, just as it was for outdoor soundscapes. Sound and Built Environments directly influence the Context, affecting our interpretation of the soundscape and leading to certain outcomes.

Seeing the influence of sound environment over the context made us think about methods to identify the composition of this sound environment. The ISO standard also identified the sound sources in the Method A questionnaire (International Organization for Standardization, 2018). Previously, we always used physical measurements of the sound levels to correlate with the perception of the sound environment. The findings of this pilot study, along with the ISO standard, inspired us to identify and use the audio content of the sound environment to see how it affects individuals' interpretation (or response) of the soundscape. This shaped the overall structure of the thesis as it leads us to focus on

using ML for both sound classification and prediction of the perceived affective quality of soundscapes.

3.3. Sound Classification with Machine Learning

As previously stated, our research consists of two parts. The first part aims to develop a sound classification model to analyze the audio content of soundscapes. The second part seeks to measure the perceived affective quality of soundscapes and correlate it with the classification results from the first part to develop a predictive model. We used a CNN model for the first part and an FFNN model.

Environmental sound classification is a very challenging task. Supervised classification algorithms require numerous labeled instances of each classification class for training. We need to have hundreds of labeled audio samples for each sound source that can be present in a sound environment to classify everything. This means that we either need to find sound samples for even the most insignificant sound source or narrow down the scope of classification only to include the significant sources for us. However, this too presents a challenge. Since the perception of soundscapes depends on the context, a sound source that can be insignificant for one environment can be significant for another one. For example, people will most likely not pay much attention or notice the sound of a cashier typing on his/her keyboard in a shopping mall. But typing sounds of a coworker in your office will get noticed in an office setting (Volkan Acun & Yilmazer, 2018). A sound environment can also be predominantly composed of sound sources that were not present in our training dataset and will not correctly detect the audio content of that sound environment. Because of this, it is a very challenging task to limit the scope of the classification. You will either have a very limited set of environmental sound classes or will need to train the network with thousands of audio samples from hundreds of sound classes. Even if you decide to use the second approach, it is still tough to find enough audio samples for many audio events, such as dropping an object to the ground.

To cope with this issue, we came up with a different sound classification approach. We trained the network with musical instruments rather than training the network with audio samples like footsteps, speech, laughter, etc. Each type of musical instrument has different audio characteristics. Suppose we can tune the neural network to provide an output in probabilities instead of binary (yes or no). In that case, we can see portions of the audio recording that are similar to the characteristics of the musical instruments. For example, if the model classifies 60% of the recording as cello and 40% as violin, it does not mean a cello plays 60% of the recording. It means that 60% of the recording has similar spectral characteristics that are found in a cello. We use the musical instruments less like a sound source and more like a filter in this approach. Spectral characteristics of the musical instruments are acting like a filter that detects audio events most similar to them. Using this new approach will significantly reduce the number of parameters (number of classes, training samples, time required to train the network, etc.) needed for the neural network. Since we are using musical instruments like a filter, we do not need to decide which sound sources to classify; hence we are not narrowing down the scope of the classification. However, deciding on the type of musical instrument is very important, which we will explain further in the thesis.

A typical sound classification procedure consists of three phases; data preprocessing, feature extraction, and training. Since we are using a machine learning model based on supervised learning, we need to have a large dataset of training samples. As we stated previously, our dataset consists of musical instruments. The samples in this set can have different sampling frequencies, different durations, or they can potentially have empty portions in their recordings that contain no valuable audio information. Preprocess phase is used to deal with these as some networks require a particular sampling frequency or duration. Also, “cleaning” the samples by removing the empty portions speeds up the computation. However, the essential function of the preprocess phase is dividing each audio sample into smaller time frames and applying window functions to capture the information about the audio signal used for feature extraction.

We used three audio features for the feature extraction phase. MFCCs are the main audio feature we used for sound classification, which we already mentioned in detail under the Theoretical Background section. We also used Tonnetz, and Spectral Contrast features as these audio features are frequently used for Musical Information Retrieval (MIR), and we are, in fact, dealing with musical instruments in our dataset.

Once the features of all audio samples are extracted, we trained the CNN model with them. The algorithm prepared a predictive model based on what it learned from the features, and we used this to predict new inputs. These inputs are the recordings of the museums we used in the survey. Normally, these recordings have an average duration of thirty seconds. However, we divided these recordings into two-second-long parts since making a prediction for lower time frames provides more reliable outcomes. The model analyzed every two second-log parts and produced an output that shows how the contents of that part match the training set's musical instruments. We reconstructed the whole recording by combining these outcomes. We will explain this process in detail in the following sections.

3.3.1. Software and Libraries

We used Python version 3.7.7. as the programming language for all coding purposes. Besides the standard build-in libraries of Python, we used Pandas, NumPy, SciPy, and Matplotlib for reading, plotting, and writing the data. We used Librosa for audio resampling and feature extraction. For building the CNN, we used MIT's Keras with TensorFlow 2.3 as a backend. Scikit-Learn library is used for computing class weights and splitting the data into train and test sets.

3.3.2. The Dataset

To determine which musical instruments to include in the dataset, we prepared an experimental procedure. We compared the classification results of different sets of

musical instruments in different soundscape recordings for this experiment. We are yet to mention the sound classification model we used, but at this point, we already had a working classification model that we used recursively to train and test the accuracy of each dataset. We will give the details of this model further into the thesis.

Literature survey showed that previous researchers studied associating emotions with musical instrument types such as string, wind, and percussion instruments (Rajesh & Nalini, 2020). Based on this, we decided to create three sets of musical instruments. Our main principle was selecting musical instruments that cover successive frequency spectrums. When combined, the frequency range of these different types of musical instruments will cover a broad range. So, the lower frequencies will be covered by an instrument with a low-frequency range, while the higher frequencies will be covered by instruments with a high-frequency range. It is unavoidable to have a certain degree of overlap between the frequency ranges of different instrument classes, such as violin and viola. Since we are using different audio features like MFCCs, the network's classification will depend on how similar the sound is to most of these features and not only to the frequency.

The first group consisted of four string instruments (Figure 20): violin, viola, cello, and double bass, along with two percussion instruments of snare drum and bass drum. Percussion instruments are meant for classifying the impact sounds within the recordings, such as footsteps or door slam. We used one hundred and eighty audio samples for this first group for each string instrument, a hundred snare drum, and ninety-eight bass drum samples, adding up to nine hundred and eighteen samples.

For the second group, we kept the two percussion instruments but replaced the strings with wind instruments. These wind instruments are tuba, saxophone, trumpet, and flute. We again used one hundred and eighty samples for each wind instrument, a hundred

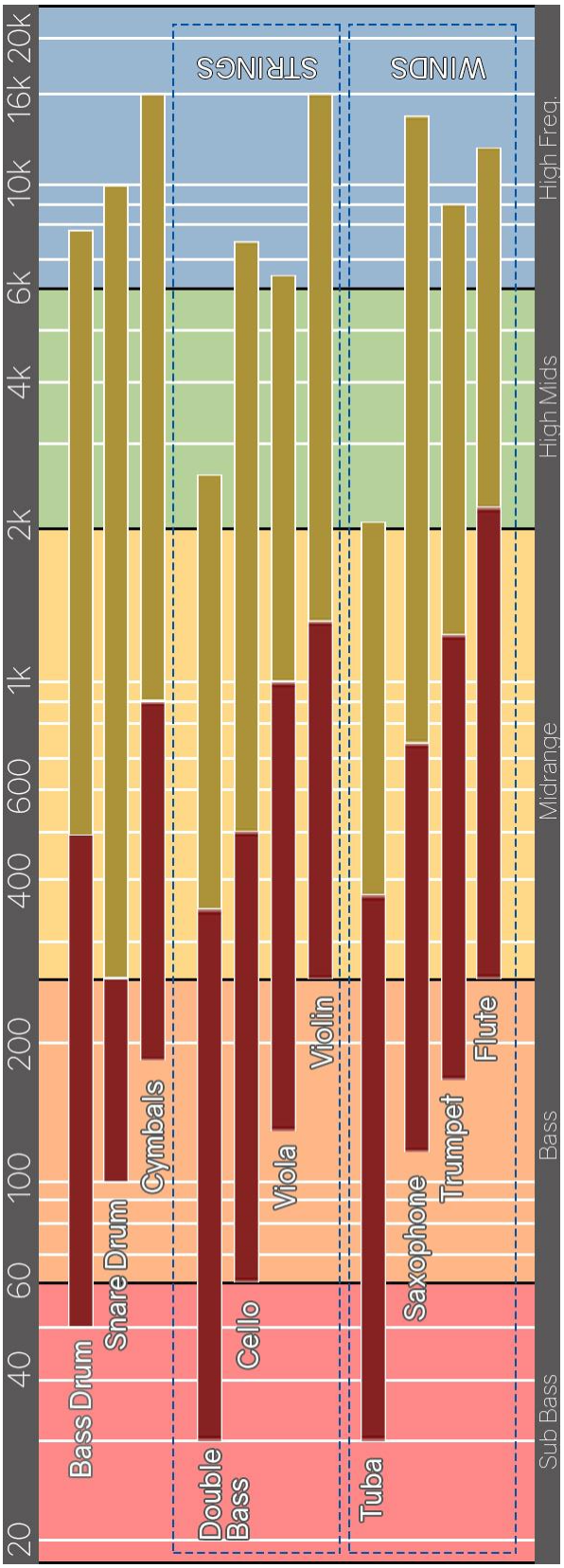


Figure 20: The frequency ranges of the musical instruments considered for the training

snare drum, and ninety-eight bass drum samples, with hundred and eighteen audio samples in total.

The third group consisted of a combination of both the string and wind instruments. The purpose of the third set is to compare the results of this set with the results of the previous two and see which instruments pairs work best with each other. We used identical audio samples for each instrument class as we used in previous groups, with only a hundred cymbals. This group consisted of one thousand seven hundred and thirty-eight audio samples in total.

Table 1: Classification results for each musical instrument group for four different sound environments.

	Test Environment 01			Test Environment 02			Test Environment 03			Test Environment 04		
	Group 1	**Group 2**	**Group 3**	**Group 1**	**Group 2**	**Group 3**	**Group 1**	**Group 2**	**Group 3**	**Group 1**	**Group 2**	**Group 3**
Bass Drum (%)	2.9	34	12.3	0.3	0	0	4.9	2.8	3.8	1.3	0	0
Snare Drum (%)	2.2	9.2	2.5	50.1	49.2	50.1	1.6	10.6	0.7	32.9	56.6	48.7
Cymbals (%)	-	-	0.5	-	-	1.6	-	-	0	-	-	0.3
Double Bass (%)	1.5	-	2.5	16.3	-	12.5	34.5	-	51.8	18.6	-	11
Cello (%)	1.2	-	4.5	5.2	-	7.8	0.7	-	0.6	0.7	-	0.3
Viola (%)	47.1	-	37.8	9.9	-	2.8	32.2	-	24.2	16.6	-	0.8
Violin (%)	45.6	-	39.9	17.8	-	13.7	25.6	-	17.1	29.9	-	6.3
Tuba (%)	-	8	0	-	1.8	0.2	-	2.4	0.1	-	1.4	1.4
Saxophone (%)	-	2.2	0	-	46	9.2	-	36.6	1.4	-	41.1	28.6
Trumpet (%)	-	0.4	0	-	2.6	0	-	29.4	0	-	0.6	0
Flute (%)	-	46.1	0.4	-	0	0	-	18.2	0.2	-	0	0

We collected audio samples from the Philharmonia Orchestra (*Sound Samples | Philharmonia*, n.d.) data set for each of these musical instrument types. We trained our CNN model with each instrument group and performed sound classification on recordings of four different soundscapes each time. These soundscapes all consisted of different sound sources, different loudness levels and contained a different amount of noise in the recordings. Table 1 shows the results of this process. Results indicate that it is not efficient to use the string and wind instruments together.

Group 3 consists of the string instruments of group 1 and wind instruments of group 2. According to the classification results of group 3, it is clear that the string instrument masked the wind instruments of group 2 in almost all conditions. The saxophone is the only instrument that is not completely masked by string instruments. We also introduced cymbals to group 3 to see if they can classify high-frequency impact sounds, but their performance was poor for every condition. We decided to use seven musical instrument types for our training dataset: bass drum, snare drum, double bass, cello, violin, viola, and saxophone. All audio samples are converted into WAV format and sampled to uncompressed 16-bit resolution, 44100 Hz mono format. This final group consisted of one thousand and ninety-eight samples in total. Like previous groups, we used one hundred samples for snare drum, ninety-eight samples for bass drum, and one hundred and eighty samples for the rest of the musical instruments.

3.3.3. Audio Signal Processing

The first step of audio signal processing is preprocessing the audio files. Preprocessing starts with dividing each audio file (from the training set) into smaller frames by applying a window (framing) in the time domain. We then apply a signal envelope to each window to clean the signal and avoid spectral leakage. Feature Extraction procedures follow preprocess. Short-term Fourier Transform (STFT) is applied to each window to convert the time domain into the frequency domain, which is then used for obtaining Mel-spectrograms. These spectrograms are used for calculating Mel-Filter Bank energies. The

feature extraction process is completed by applying a Discrete Cosine Transform (DCT) to filter bank energies, which provides the Mel-Frequency Cepstral Coefficients (MFCCs).

3.3.3.1. Preprocessing

Preprocessing prepares the data for feature extraction. If the audio files are not sampled at the same frequency or bit depth, they are sampled to the same sampling characteristics in this step. In some cases, all audio inputs may also require to have the same duration, as some networks do not accept different input sizes (Puente, 2018), but this is not the case for the model we are using.

All audio samples used for this model have a sampling rate of 44100 Hz, which means that they have a Nyquist frequency of 22050 Hz. Since most changes occur at lower frequency content, downsampling the data allows us to increase the computing speed without losing valuable information. Therefore, we have downsampled the audio files down to 16000, so all of the audio files ended up having a Nyquist frequency of 8000 Hz.

Downsampling is followed up by framing the audio signals. In basic terms, framing is splitting the audio signal into smaller frames. Since the audio signal is constantly changing, we cannot make a generalizable assumption using the whole duration of the signal. We can, however, assume that the signal will be statistically stationary if we take a sample from a very short time frame. By analyzing this short time frame, we can approximate the frequency contours by concatenating adjacent frames. The most commonly used frame sizes vary between 20-40ms, depending on the desired task (Sharan & Moir, 2016). If the frame size is too short, there will not be enough samples to make a reliable assumption about the signal. If it is too long, the signal will change too much over time and defeat the purpose of framing (Wold et al., 1996). This is the reason frame size is important. We used a 25 ms frame size with a 10ms stride (10% overlap). Afterward, we apply a window function over each frame to smoothen the signal and avoid any spectral leakage (Sharan & Moir, 2016). The Hanning window is one of the

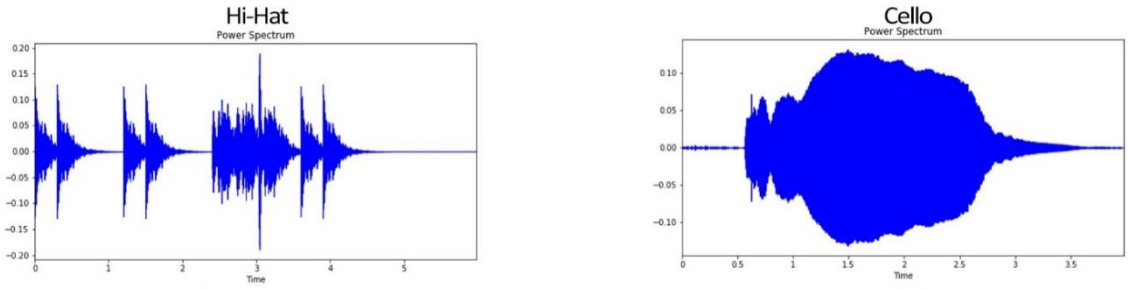


Figure 21: Power spectrum of two audio signals in the time domain.

most commonly used windows. Once the preprocessing is over, the signal's power spectrum is calculated (Figure 21) and plotted over the time domain graph.

Using the Pandas library, we applied a rolling window to each window to build the signal envelope and clean the signal. This is required when the signals contain dead spaces which do not have any audio information. The Signal envelope uses amplitude modulations to remove this dead space. The envelope outlines the upper and lower boundaries of the waveform by smoothly connecting the peaks without overfitting (Jarne, 2018).

3.3.3.2. Feature Extraction

Feature extraction provides the ML algorithm with the best possible data to perform in the best possible way. In this step, features are generated by taking raw, unstructured data and defining features for potential use in statistical analysis (Puente, 2018). The features used for this model are MFFCs, Tonnetz, and Spectral Contrast.

Tonnetz and *Spectral* contrast are typically used for music genre classification or musical information retrieval systems. We choose them for this model since the dataset consists of musical instruments. *Tonnetz* allows spatial repetitions of tonal distances or relationships by computing tonal centroids (Harte et al., 2006). *Spectral Contrast* is used for representing the relative spectral characteristics of music (Jiang et al., 2002).

While these features generate information about the musical instruments, their contribution is minimal. The contribution of the MFFCs, on the other hand, is preeminent, and it is crucial for this model. We have provided the theoretical background of the MFCCs previously, but we have not explained how they are calculated. We will explain how they are calculated in detail since it is our main audio feature.

3.3.3.2.1. Fourier Transform

Fourier Transform (FT) is the first step of almost every audio feature. The audio signal's more distinguishable features become more apparent for the first time with this operation. FT converts the audio signal from the time domain to the frequency domain (amplitude/time to amplitude/frequency) (Puente, 2018). The sampling rate of the signal has no direct influence on the frequency resolution of the FT, but it does influence the frequency coverage of the FT. There are different classes of FT, such as Short-Term Fourier Transform (STFT), Fast Fourier Transform (FFT), and Discrete Fourier Transform (DFT).

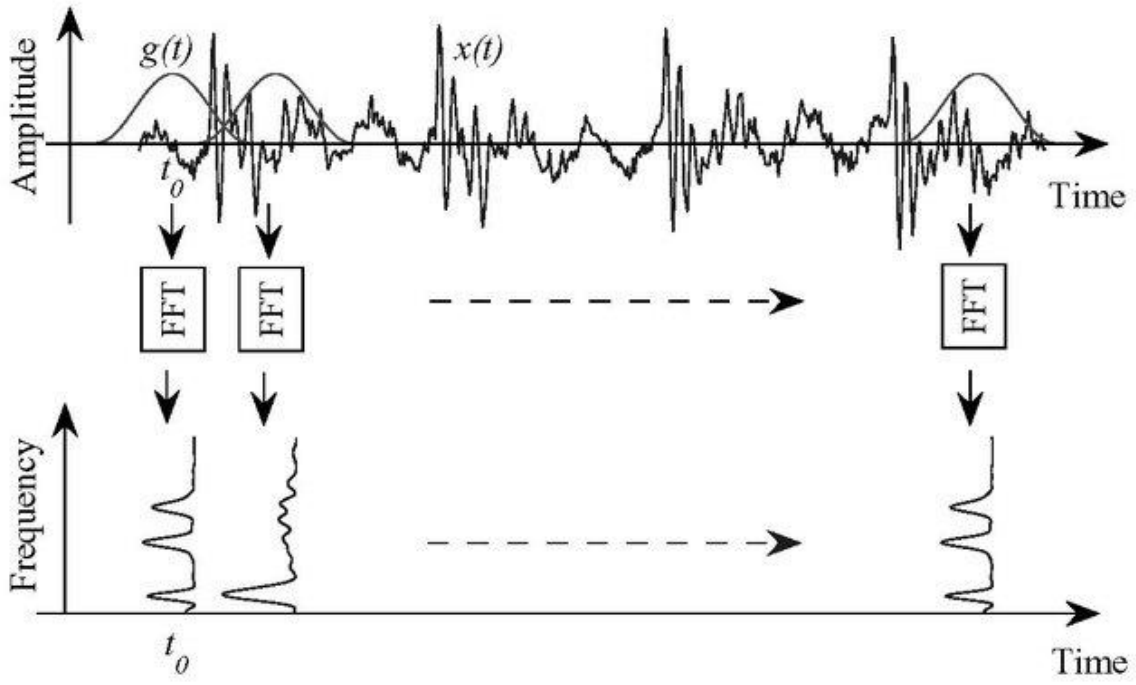


Figure 22: Short-Term Fourier Transform on an audio signal. (Gao & Yan, 2006).

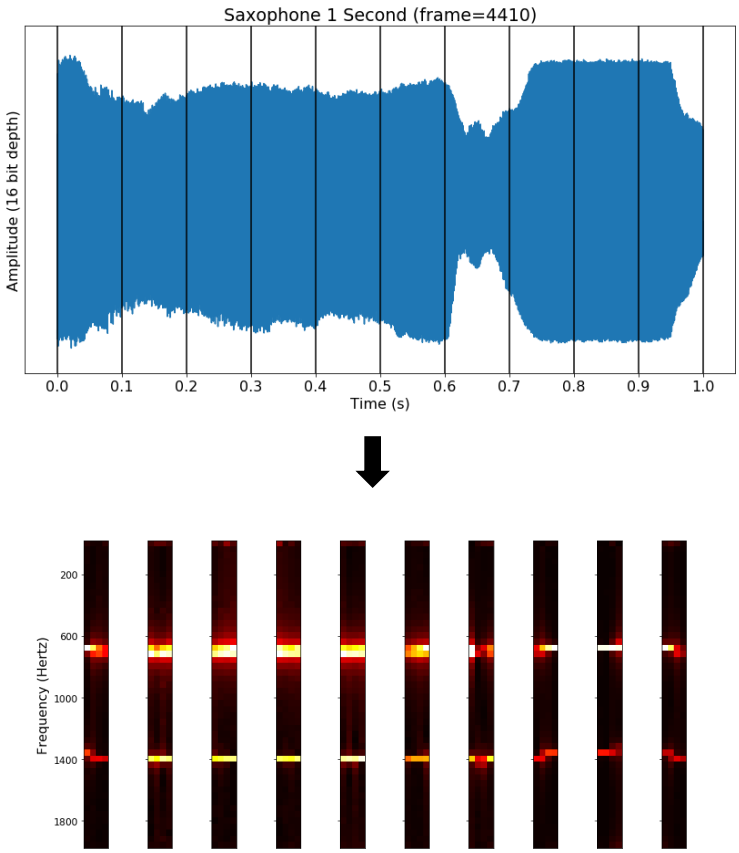


Figure 23: Construction of a spectrogram by applying STFT to the power spectrum of an audio signal.

Because the signal is statistically stationary in short time frames, STFT is a commonly used FT method. STFT (Figure 22) is a sequence of Fourier Transforms of a windowed (time frame) signal (Kehtarnavaz, 2008). It is used in situations where the frequency components of a signal vary over time to provide the time-localized frequency information. While the FFT provides information about the frequency content averaged over the entire signal time interval, STFT provides frequency information of a short period of the audio signal (Kehtarnavaz, 2008).

The output of the STFT is periodograms for each window. By stacking these periodograms together, we obtain the spectrogram of the audio signal (Figure 23). A

spectrogram is the time-frequency representation of the audio signal. It is an essential tool for audio analysis as they describe the evolution of the frequency components over time.

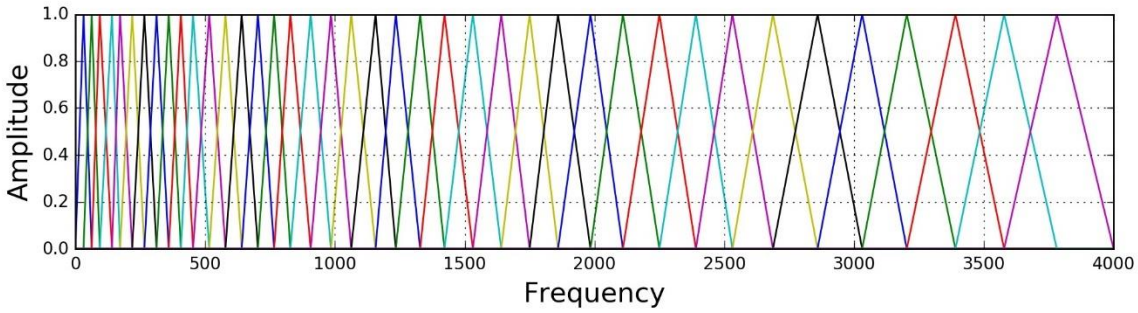


Figure 24: The Mel-filter bank. Each different-colored triangle represents a filter.

3.3.3.2.2. The MFCCs

We mentioned previously how the human ear perceives sound nonlinearly and provided theoretical background about this phenomenon. In this section, we will explain how we prepared the Mel-spectrograms and the MFCCs.

In the previous step, we calculated the magnitude spectrum of the audio signal (spectrogram). We apply the Mel- Filterbank to the magnitude spectrum of the signal. Mel-Filterbank is a set of 20-40 overlapping triangular filters defined by their center frequencies. To apply this filter bank, we first need to choose a lower and upper frequency, and we also need to specify the number of Mel filters we will use. Each triangular filter has a response of 1 at the center frequency and decreases towards zero until it reaches the center frequency of its adjacent filters (Figure 24). For our feature extraction method, we used 26 Mel-filter banks.

The resulting filter bank coefficients can be highly correlated with each other, which can cause problems for machine learning algorithms (Kopparapu & Laxminarayana, 2010). After taking the logarithm of all filter bank energies, Discrete Cosine Transform (DCT) is applied to the filter bank coefficients (Figure 24). DCT is used to decorrelate the output

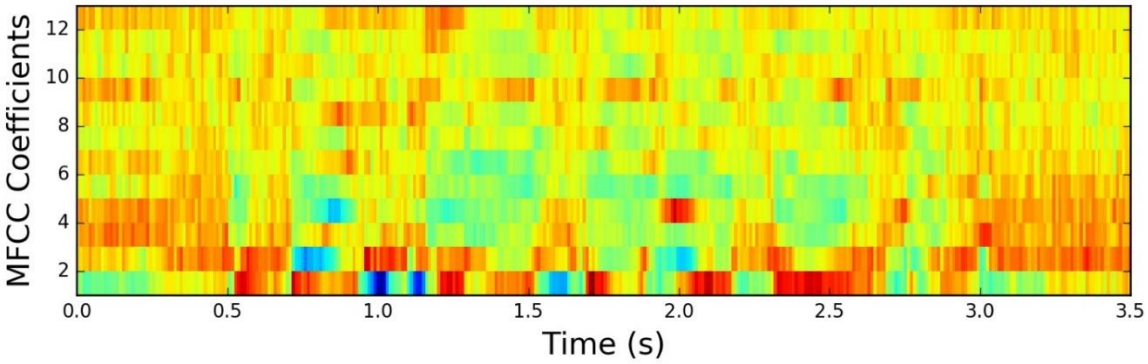


Figure 25: MFCC of the audio signal after taking DCT of the filter bank energies.

data (Kopparapu & Laxminarayana, 2010; Mitrović et al., 2010). It is almost like a low pass filter for the audio signal. Our eyes and ears are analog objects. The amplitudes are often too similar to their neighbors in many audio signals and graphic images for us to distinguish the distortions around the edges. If we remove some higher frequency content, our brain hardly perceives the difference, but the number of parameters required for the analysis is significantly reduced, lowering the computation time. Therefore, with the DCT, filter bank coefficients are decorrelated by removing high-frequency content and compressing down the data to lower frequencies.

Typically, the MFCCs between 2-13 are retained and used for representing the shape of the spectrum (Figure 25). We kept these MFFCs and removed the rest as they represent fast changes in the filter bank coefficients. The details provided by these do not significantly contribute to the sound classification. Different applications of sound classification have different requirements, and some applications choose to retain more coefficients. A limitation of DCT is that it is a linear transformation, and because of this, it discards some of the highly non-linear signals. After obtaining the MFFCs, we apply a mean normalization to the output. Normalization is used to balance the spectrum and improve the Signal-to-Noise ratio (Fayek, 2019). This is done by subtracting the mean of each coefficient from all frames (Figure 26).

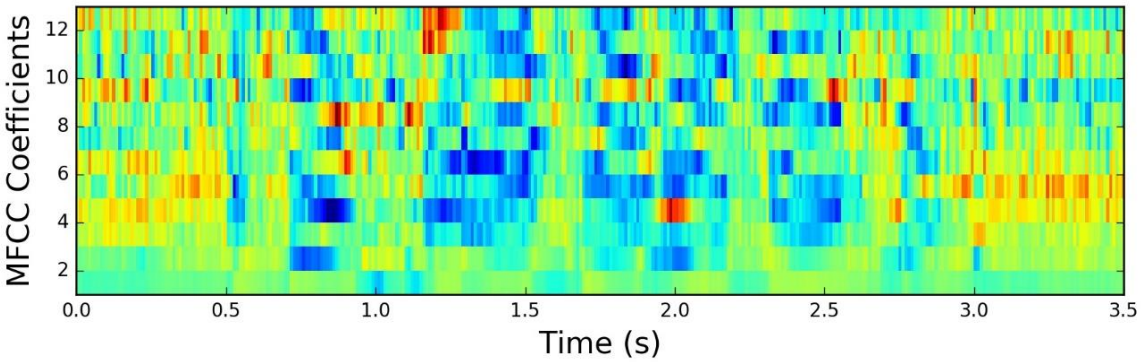


Figure 26: MFCCs after applying mean normalization.

This feature extraction procedure is performed for every audio simple within the dataset. Each class of musical instruments is stored in a different folder. We used Pandas to use the folder name as the training label for each audio sample. Our preprocess and feature extraction algorithms processed the audio samples and associated them with their respective labels. The extracted features are saved as NumPy files in the project directory. Unless we make changes to the dataset, we only need to process the audio samples once. We can use the NumPy file to load the features to the neural network whenever it is required. This is especially useful when we need to tune the hyperparameters of the neural network.

3.3.4. The CNN Model

We used a CNN model for sound classification. Since this network type is one of the more complex applications of ML, we decided to build out a network upon an existing one. There are specific criteria we require from the CNN algorithm to satisfy. It should use the training dataset to predict an entirely new set of audio files, which we never fed to the network before (recordings of the museums) with reasonable accuracy. After a thorough search in GitHub, we decided to develop our CNN model based on a GitHub repository named pyAudioClassification (https://github.com/micah5/pyAudioClassification. You can see the code for our CNN algorithm in Appendix E1.

Figure 27 is the graphical representation of the CNN model we used for this study. This model uses four 1D convolution layers. The dimension of the convolution layer is related to the size of the kernel. A 1D convolution layer uses a one-dimensional matrix of 3 for our case. If this was a 2D convolution layer, the kernel size would be a 3x3 matrix

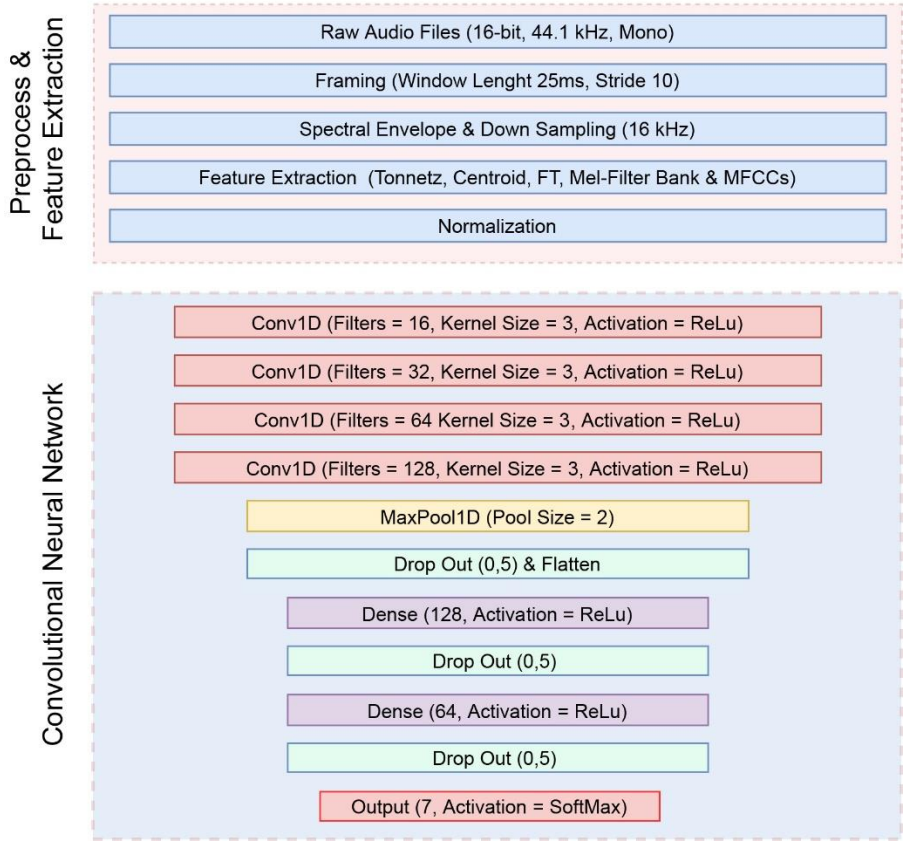


Figure 27: Representation of the CNN architecture used for this classification task.

The first *Convolution Layer* starts with a filter size of 16. Filter size progressively increases as the model iterates through the *Convolution Layers*. Typically, the number of filters is increased by powers of two, so the number of filters starts at 16, and it reaches 128 by the fourth layer. Starting with a smaller number of filters helps to collect as much local information as possible. Increasing the number of filters at each *Convolution Layer* helps to gather more global and higher-level features. Each convolution layer uses a kernel size of 3 with a stride of 1 to create feature maps. For all four layers, the Rectified

Linear Unit (ReLu) activation function is used. This nonlinear activation function changes negative values into zero while leaving the positive values unchanged (Chen et al., 2019; Parker, 2017).

A 1D *Pooling Layer* of size 2 is added after the convolution operations *to* reduce the dimensions of the feature map. *Pooling Layers* reduce the parameters and dimensions of the convolution layer while increasing the efficiency of the model. The standard convention is adding a *Pooling Layer* after each *Convolution Layer*. In this model, a single max-pooling layer is placed after the last convolution layer. It makes sense to put a *Pooling Layer* after each *Convolution Layer* in models with large input spaces as it also increases the efficiency of the model. Since the input space is not particularly large in this model, it was unnecessary to place one after each convolution operation. Arranging the layers like this allows the model to build upon the features progressively it has acquired from the previous Convolution Layer and learn different features of the MFFCs.

A *Dropout* layer usually follows the Pooling Layer. *Dropout* prevents the model from overfitting and increases the generalizability of the model (Goleman, Daniel; Boyatzis, Richard; Mckee, 2019; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). If we did add a *Pooling Layer* after each *Convolution Layer*, we would also be required to add a *Dropout* after each *Pooling Layer*. Dropout prevents overfitting by dropping units, along with their connections, from the network during training (Srivastava et al., 2014). This means destroying the input space after each pooling layer. Since the size of our input space allows us to continue without pooling after each convolution, we build up upon our feature map each time.

Flatten layer is added after the *Dropout* to convert the input space into a 1D array. Flatten is used to prepare the data before feeding it into the *Fully Connected Layers*. These layers are also called *Dense* layers because the neurons of these layers are very densely connected and constitute the bulk of the network's computation power. This time

the number of dense layers gradually decreases, starting from 128 and down to 7, which is the number of classes. A *Dropout* is added between each *Dense Layer* to avoid overfitting. The first two *Dense Layers* use the ReLu activation function, but the third *Dense Layer* uses a different activation function since it is the output layer.

The purpose of the last *Dense Layer* is to make the values produced by the network to be interpretable by humans. The most commonly used activation functions for the output layer are Softmax and Sigmoid functions. The difference between them is, Softmax turns the network output into a probability between 0 to 1, while Sigmoid provides a binary output (Aggarwal, 2018; Passricha & Aggarwal, 2019). Both the SED and sound classification models make use of both. If the goal is to see which sounds are present in the given sound environment, a Sigmoid should be used. Since this research aims to know the probability, we used the Softmax activation function. The model is trained through 500 iterations (batch size). In the end, the model obtained a training accuracy of 86%.

3.4. The Soundscape Perception Survey

The Soundscape Perception Survey aims to identify participants' perceived affective responses to the different soundscapes. The main feature of the survey is the *Listening Tests*, in which the participants rate each given sound environment based on given scales. Under normal circumstances, this test is conducted in an anechoic chamber or a quiet room. Due to the extraordinary circumstances we are going through, it was not possible to find enough individuals who volunteer to come to the university for *Listening Tests*. We had to resort to an online survey, which also had its challenges and limitations.

The major challenge of online surveys is the lack of control. We had to find methods to cope with various limitations. Typically, participants had to go through an audiometry test to determine if they are eligible for the listening test(Burkard, 2016). We needed to

find an online alternative. We also had to ensure that participants use earpieces (headphones or earphones) during the survey and not speakers since it can significantly affect the test results (audiometry and listening tests). Ideally, the participant would use a headphone calibrated to 55 dB(A) to hear the full range of the audio present in the video. They should never change the volume between videos since it would alter their perception of different soundscapes completely. Addressing these limitations was the guiding principle while designing the questionnaire and selecting an online survey tool.

3.4.1. Sound Environments

One of the main challenges we faced due to the pandemic is finding a suitable case study setting. Rahmi Koç Museum and Erim Tan Museum were initially planned to be the case study settings. However, after visiting both locations multiple times, we saw that the number of visitors had dramatically decreased and both museums lacked visitors. Since we used Rahmi Koç Museum for our pilot study (Volkan Acun & Yilmazer, 2019), we knew how the museum environment usually is. Human-generated sounds account for the majority of the sounds that we hear in indoor spaces. Sound sources (wind, water, animal, traffic, etc.) that contribute to the outdoor soundscapes are not present in most indoor soundscapes. Therefore, the soundscape characteristic of a public interior space like a museum is highly dependent on the presence of human activity. Because of this, both museums lacked the critical element that contributed to their soundscapes were very dull.

To start the data collection within the given time margin, we decided to use videos of museums as audiovisual stimuli. The notion behind using the visual material in addition to the audio material is to provide context for the sound environment. The focus of the study is still the audio stimuli found within the videos. We also reminded participants to pay attention and evaluate the videos based on the audio stimuli during the survey.

We searched online for online museum tours/walking museum tours to find suitable candidates for the survey. The videos must meet specific requirements for us to consider. The videos must provide an experience as close as possible to a typical tour in an interior space. This means that videos must not be any music or text added to them through video edit. They shouldn't have any narration as well. They all must be just plain walks in the space, the only sound originating from the room. The videos must have HD quality and provided a source for their recording/video equipment. Videos must have image stabilization and should not have any swift camera movements, creating discomfort to the viewer. Finally, they should have rich sound environments with various sound activities. Based on these requirements, we selected videos of ten indoor environments. These environments are Kröller-Müller Museum, Louvre Museum, London Natural History Museum, National Museum of Scotland, The British Museum, Weserburg Museum of Modern Art, Victoria and Albert Museum, Smithsonian National Museum of Natural History, and Marine Parade Public Library.

As the names of the museums suggest, each case study environment has different spatial characteristics and includes various types of contents. While some of these environments are located in historic buildings, some are in very modern ones. There are both minimalistic designed modern art museums and natural history museums with lots of objects and colorful exhibitions. This is a conscious decision made to increase the variance of the data later on.

We used these videos for listening tests as audiovisual stimuli to measure the perceived affective quality of the soundscape. The duration of the original videos varied between 27 minutes and 80 minutes; therefore, it was impossible to use them as they were. Duration should be short enough to be suitable for a person's attention span. Based on the literature survey, we saw that the most commonly used audiovisual stimuli used for listening tests are primarily between 30 seconds to 1 minute(Boes et al., 2018; Eronen et al., 2006; Ooi et al., 2020; Sun, De Coensel, et al., 2019). Since we conducted the survey online and did not control over the participant, we decided to use a 30-second duration

for the videos to reduce the chance of participants dropping out of the survey. Longer duration will also increase the number of sound sources within the video which would cause issues during evaluation. Appendix D1 includes a QR code for one of these videos.

Each video is downloaded and resampled to have 44100hz, 16-bit resolution audio. We have gone over each museum video multiple times to determine the onset and offset times of each potentially valuable audio event. These periods are cut into individual samples of 30-second-long videos. In the end, we have prepared 72 videos this way. These videos were later added to the online questionnaire, while the content of their audio files was analyzed and classified by the CNN algorithm.

3.4.2. Data Collection

We held the data collection through an online questionnaire survey. Before conducting the questionnaire survey, we applied the Ethics Committee of I.D. Bilkent University for approval (Ethics Committee Approval no: 2020_10_23_01, Appendix A). The questionnaire included a *Listening Test* part consisting of many videos from different sound environments. Therefore, we needed the online survey service to be capable of handling large numbers of videos. We considered three online survey services for the survey. To find the most suitable one, we prepared the same questionnaire for all three survey services and distributed them to a small number of participants. We also asked the participants to note each part they find confusing and the time it took them to complete the survey. Based on the feedback from the participants, we identified problems and improved the content of the questionnaire. Unfortunately, every participant experienced an issue with the order of videos while using Jot Forms. We decided to use Kwik Survey as our primary survey tool and kept Google Forms as a backup. While Google Forms was free and did not present any issues, Kwik Surveys provided more options and had more detailed question logic jumps. It also offered a superior user experience, making the participants slightly more eager to carry on with the survey.

A potential limitation of conducting an online survey was the streaming issues caused due to the internet connection speed. Streaming the videos with limited internet speed can cause the videos to freeze while watching. Based on the feedback we got from the participants we know that at least one of them experienced this issue. Unfortunately, there is no easy solution to address this issue other than advising participants to only start the survey once they have a stable internet connection.

Perhaps the greatest challenge of the data collection was finding participants who would complete the survey. During our pilot questionnaire test, some participants expressed that it took nearly thirty minutes to complete the questionnaire since it contained many videos. Some said that parts of the questionnaire were too complicated, which almost caused them to leave the survey. This was a very problematic issue since we needed to use many videos in the Listening Tests part to gather information on the audio content of different sound environments. If the survey contained a small number of videos, participants would have been more likely to complete the survey. However, since they would respond to a small number of videos, the data variance extracted from the sound environments would also be minimal. For example, if 100 participants watch five videos and answer the questions based on that five videos, the model we create from that data would not be a generalizable one since it will overfit almost immediately. So this issue presented us with one of our significant challenges, the survey should not be too long, but it must also contain enough audiovisual stimuli to create a model that generalizes well. We tackled this problem by creating three different video groups. Each video group included the same questions but used a different set of videos as the audiovisual stimulus. The questionnaire of one of the video groups can be accessed with the following link: https://freeonlinesurveys.com/s/L196dMLQ#/0. Appendix D2 also includes a QR code link for this online questionnaire.

3.4.2.1. Structure of the Questionnaire

The questionnaire consists of three main parts. It is also available in English and Turkish since it was crucial to gather information from individuals with different cultural backgrounds. The first part of the questionnaire is concerned with demographic data. The questions regarding this part of the survey can be seen in Appendix C.

The second part of the questionnaire solely focuses on managing the limitations of online listening tests. We searched the literature and found that online audiometry tests can be conducted via mobile devices if they are biologically calibrated(Masalski et al., 2016). We added a link to a simple and straightforward hearing test to our questionnaire. A typical pure-tone audiometry test provides an audiogram as output after completing the test, which can be hard to interpret by the participants. The one we added tells the participant if he/she has hearing loss or not, and if so, its amount. Ease of use was one of the primary reasons that made us decide on this particular test. We used a different hearing test during the pilot of the questionnaire. However, the feedbacks indicated that its instructions were hard to follow, its output was even harder to understand, and its results were not very reliable. It also took some participants as long as 10 minutes to complete the hearing test alone. The audiometry test we ended up using is much simpler and straightforward, which we also tested again with the same participants before adding to the final version of the survey. After participants complete this hearing test, they are directed back to the survey page and respond to the question, “Did the test identified any hearing loss?”. This test also required the participants to put on headphones. We have also added reminders in multiple parts of the survey that they must use earpieces at all times. The second part of the questionnaire ends with a white noise sample. We asked the participants to use this as a reference for adjusting the volume of their device to a comfortable level and never change it until the end of the survey.

The main part of the survey is the third part, which is the listening tests. This part measures the perceived affective quality of the audio stimuli. Twenty videos from

different indoor environments are presented to the participants as audiovisual stimuli. After each stimulus, participants are asked to evaluate the pleasantness, eventfulness, and how they feel about the soundscape overall. We reminded the participants to make their evaluation based on the sound environment (audio stimuli). The three questions used here are part of data collection Method A of ISO 12913-2:2018(International Organization for Standardization, 2018). Pleasantness and eventfulness are also two of the three components(the third one being familiarity) of soundscape perception according to the highly acclaimed principal component model of Axelsson et al. (2010), which we discussed in detail previously. Turkish translation of pleasantness and eventfulness is determined based on the research carried out by Ozcevik et al. (2014). Questions use a 5-point Likert scale, with 1 corresponding to "Completely Disagree" and 5 corresponding to "Completely Agree."

3.4.3.2. Video Groups

As previously stated, collecting participants' perceived affective response data based on a small number of audiovisual stimuli will greatly limit the model's generalizability. It will most likely cause the training set to overfit during data analysis. On the other hand, conducting the survey with 60 different audiovisuals would make it very difficult to find participants that are willing to complete it. The duration of each stimulus is 30 seconds, so using 60 of them would make the *Listening Tests* part of the survey last at least half an hour. To deal with this important issue, we distributed videos into three video groups. Each group consists of the same set of questions but has different videos as audiovisual stimuli in the *Listening Test* part. The participants were allowed to watch the videos as many times as they want before evaluating them. The combination of participants' demographic information, the audio content of the video, and evaluation of pleasantness, eventfulness, and overall impressions, formed a sample for the dataset, which is later used for the soundscape model. The *Video Group 1* has 56 x 20 = 1120, *Video Group 2* has 63 x 20 = 1260, and *Video Group 3* has 57 x 20 = 1140 samples (Figure 28), adding up to 3520 total samples.

Out of the 72 videos we prepared based on the museum videos, 60 are selected and distributed between the three groups. This process was critical since the audio characteristic of the audiovisual stimuli should be similar between each survey group. If one survey group mainly consisted of loud and more eventful audio stimuli while the others consisted of quieter and uneventful stimuli, it would decrease the variance caused by the relation between demographic information and perceived affective quality.

We needed to ensure that different audio contents of indoor soundscapes are represented equally in each survey group. To achieve this, we implemented a procedure similar to qualitative coding. We assigned multiple labels to each video based on their audio content. For example, the entrance of the Louvre Museum has a substantial volume which causes lots of echo and background noise. In the video, people clustered around a tour guide, listening to him talking about the building, children running around, and sometimes yelling at each other. We labeled this video as laud, echo, human speech, children, and large space. By applying this labeling strategy, we formed a labeling framework that consists of labels for every video we have. Appendix B includes the frameworks we prepared. We used these frameworks while choosing the content of each video group. These ensued video groups consist of a broad range of sound environments. This method maximizes the difference between the audio stimuli and participants' perceived affective response, allowing the data samples' saturation, thus improving the model's generalization capabilities.

Figure 28: The number of recordings and participants of each survey group.

3.4.3.3. Participants

We reached out to the participants by snowball sampling. A total of 175 individuals, 117women and 58 men, have participated in the study. The majority of the participants varies between 25 to 34 (56.5%), while the second-highest age cohort is the range between 18 to 24(28%) (Figure 29). These cohorts are followed by the 35 to 44 range (6.8%), 45 to 54 range (5.7%), and 55 to 64 range (2.8 %). The test was available both in Turkish and English to reach out to individuals from different cultural backgrounds. This would ensure maximum variance within the dataset. Regardless of our efforts, the vast majority of the participants were from Turkey (88%) (Figure 30).

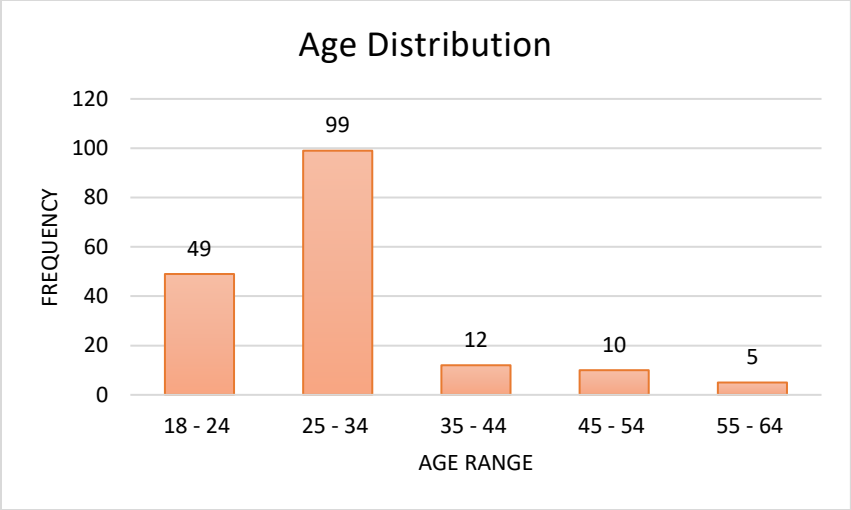


Figure 29: Age distribution of the participants.

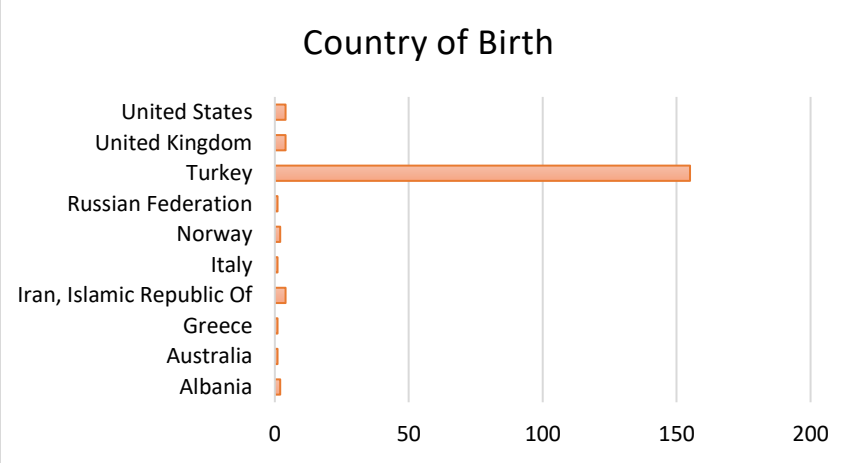


Figure 30: Participant's country of birth.

3.4.3. Data Analysis

The main purpose of data analysis is to develop a predictive model. While we mainly focused on using an FFNN model for this purpose, we also looked into other machine learning approaches to compare and find out if there is a more suitable one. After searching, we also decided to use K-Nearest Neighbors(K-NN) classification algorithm. This classification algorithm contrasts the FFNN algorithm as it does not include a training phase. Because of this, it is considered as a "Lazy Learning Algorithm." As we briefly stated under the theoretical background, K-NN uses feature similarity with the nearest data samples to predict the class of a new sample (Jefferson et al., 2016).

The audio content of the soundscapes, participants' *Pleasantness* and *Eventfulness* responses, and their demographic information were used as input variables (independent variables), while the *Overall Response* to the soundscape is used as the output variable (dependent variable) for both FFNN and K-NN models. Therefore, we can say that we developed both an FFNN and a K-NN model as part of data analysis. We compared their classification results to see how our data performs with different applications of machine learning. You can see the code we used for the FFNN model in Appendix E2 and K-NN model in Appendix E3.

3.4.4.1. Preprocess for FFNN and K-NN

Similar to the previously described CNN model, this data analysis starts with preprocessing the data. Since the logic behind the preprocessing phase is identical for both FFNN and K-NN algorithms. Both methods have undergone the same preprocess with only one exception; the dataset for K-NN cannot contain any zeros. Other than this exception, the process is identical for both of them. Because of this, the preprocess phase covered under this subsection applies to both models and will not be explained separately for each model. The preprocess prepares the data before running the classification algorithm. The main subjects preprocess phase addressed here are:

1. Ensuring the data does not contain any missing values or incompatible data types (or any zeros for K-NN).

2. Encoding the categorical data to converting them into numeric ones.
3. Splitting the dataset into train and test (and sometimes validation) sets.
4. Future scaling, which ensures different variables have similar magnitudes.

The first step is ensuring that the data types are consistent and handle the missing values, if any are present. The data type for one of the variables was different from the rest and contained missing or infinite values. The mean replacement method is typically used for handling the missing data. However, since our audio content variables range between 0 and 100, replacing the missing values through mean replacement would have violated the data integrity. Because of this, we replaced the missing values with zeroes for FFNN and with ones for K-NN since zeros cannot be used for K-NN. We also changed the rest of the zeros within the dataset with ones for K-NN.

After handling the missing/infinite values and zeros for K-NN, we encoded the categorical variables to numeric values. Within our dataset, two variables included categorical data. This type of data is the variables that have label values rather than a numeric value, and they must be converted to numeric values for the ML algorithm to operate. In our case, participants' country of birth and their preferred language were categorical variables. One-Hot Encoder method from Scikit-Learn library is used for converting these categorical values into numeric ones. This method splits the columns with categorical data into multiple columns and assigns a binary value for each unique category value (Figure 31). The binary value is essentially a notion of true or false, indicating if the corresponding label is true or false for the data sample.

After encoding the categorical data, we split the dataset into train and test sets. The notion behind splitting into train and test groups to validate the training results and obtain accuracy metrics. We split 80% of the total data into the training set and the remaining 20% into the test set, which is the typical practice.

BRIDGE-TYPE (TEXT)	BRIDGE-TYPE (Arch)	BRIDGE-TYPE (Beam)	BRIDGE-TYPE (Truss)	BRIDGE-TYPE (Cantilever)	BRIDGE-TYPE (Tied Arch)	BRIDGE-TYPE (Suspension)	BRIDGE-TYPE (Cable)
Arch	1	0	0	0	0	0	0
Beam	0	1	0	0	0	0	0
Truss	0	0	1	0	0	0	0
Cantilever	0	0	0	1	0	0	0
Tied Arch	0	0	0	0	1	0	0
Suspension	0	0	0	0	0	1	0
Cable	0	0	0	0	0	0	1

Figure 31: An example showing how One-Hot Encoder handles the categorical data of "Bridge Type". OneHotEncoder creates a new column for each unique value and assigns binary value to the column (Dinesh Yadav, 2019).

All variables of the training set should have a similar scale before we can be feed into the network. Highly varying values will cause the features with high magnitudes to have weight a lot more than the features with low magnitudes, which will impact the results. The data frame needed needs to be normalized to suppress this effect. We applied feature scaling to the training set to arrange the values in the same range, so no feature is dominated by the other. StandardScaler from Scikit-Learn's Preprocess library is used for feature scaling. Scaler standardizes each input variable by subtracting the mean and dividing by the standard deviation(Scikit-Learn, n.d.). This process shifts the data distribution to have a mean of 0 and a standard deviation of one. After this last step, the data is ready to be used for the FFNN and K-NN algorithms.

3.4.4.2. The FFNN model

We used the Keras library to build the Feedforward Neural Network. The input layer of the network consists of seventeen neurons, a neuron for each input variable. Seven of these input variables are the musical instrument percentages we use as the indicator of the audio content. Eight of them are the demographic variables which use to measure the participants' background. The last two variables are the perceived affective response variables of pleasantness and eventfulness scores. Therefore, the input layer of our model has an input dimension of seventeen and uses *Rectified Linear Unit* (*ReLu*) activation function.

The output layers consist of just one variable: the participants' overall response to the soundscape. This layer has only one variable but includes six neurons. Since we used a 5-point Likert scale for the questionnaire, the output variable can have five different values. We added an extra neuron because Keras required us to add one more than the total values. Similar to the CNN network, the *Softmax* activation function is used for the output layer. Because of this, the output from the network will show the probability of each overall perception value.

The number of hidden layers and the number of neurons in each hidden layer are crucial parameters. There is no one common rule for deciding on the number of neurons for the hidden layers. However, there are several methods such as using an amount between the size of the input layer and the size of the output layer, using the 2/3 the size of the input layer plus the size of the output layer, using a number smaller than twice the size of the input layer. There is also the method of using $\sqrt{n*m}$, where n is the number of input layers and m is the number of output layers (Masters, 1993). We used

We experimented with combinations of a different number of hidden layers and neurons in these hidden layers. Accuracy and loss plots are used for comparing the performance of different combinations (Figure 32). An Accuracy plot shows the difference between the training and test accuracy. The main goal here is the find the combination that

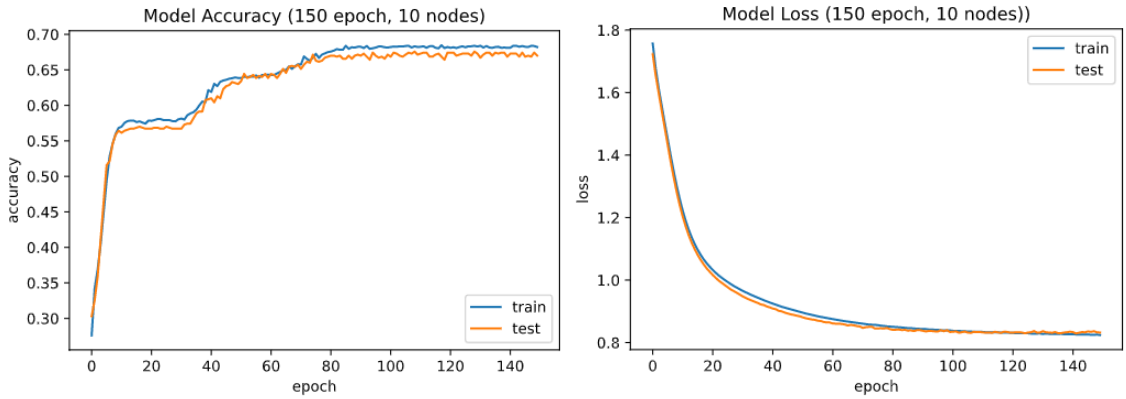


Figure 32: Accuracy plot(left) and Loss plot(plot) showing the difference between train/test accuracy and loss over given number of epochs.

provides the highest accuracy without causing the model to overfit dramatically. We spot the overfitting when the train and test accuracies start to diverge noticeably from each other.

The loss plot is similar to the accuracy plot. It shows the amount of train and test loss for each epoch. Spotting the point of overfitting and limiting the number of epochs to that point is another way to deal with overfitting. Plotting the accuracy and loss also allowed us to find the optimum number of epochs to train the network. We also tested *Adaptive Momentum Estimation* (ADAM*)* and *Stochastic Gradient Descent* (SGD) optimizers to see which one performs better. Based on these, the best performing model presented 70% training accuracy, consisted of three hidden layers with ten neurons in each, and trained with SGD optimizer for 150 epochs.

3.4.4.3. The K-NN model

We used Scikit-Learn to create the K-NN model. Model is initialized using thirteen neighbors(k) and using Euclidean as the distance metric to measure the distance between the new data point and the surrounding thirteen data points. The two accuracy metrics we used for the model are test accuracy and f1 score. As simple as it was to prepare it, the results obtained from the model were not satisfactory. Both metrics were quite low, with the f1 score being 45% and test accuracy being only 42%.

To test the accuracy metrics further, we used the k-Fold Cross Validation method. This method essentially splits the dataset into k number of pieces, which is ten folds in our case. It uses nine folds for training and the remaining one-fold for testing. The cross-validation procedure runs for ten iterations while changing the test fold each time and calculating the accuracy. The mean accuracy of ten folds is provided in the end. The cross-validation result for the model was 40%, which is even lower than the previous two metrics.

Hyperparameters are one of the important factors in the performance of models. To see which hyperparameters can provide the highest performance, we used the *GridSearchCV* function from the *Scikit-Learn* library. This function loops through predefined hyperparameters and fits the model to the training set. In the end, it provides the highest performing parameters. However, even with the highest performing set of parameters model only achieved 50% accuracy.

CHAPTER IV

RESULTS

The results of the study will be presented in this chapter. The results are divided into two the sound classification results and soundscape perception results.

4.1. Sound Classification Results

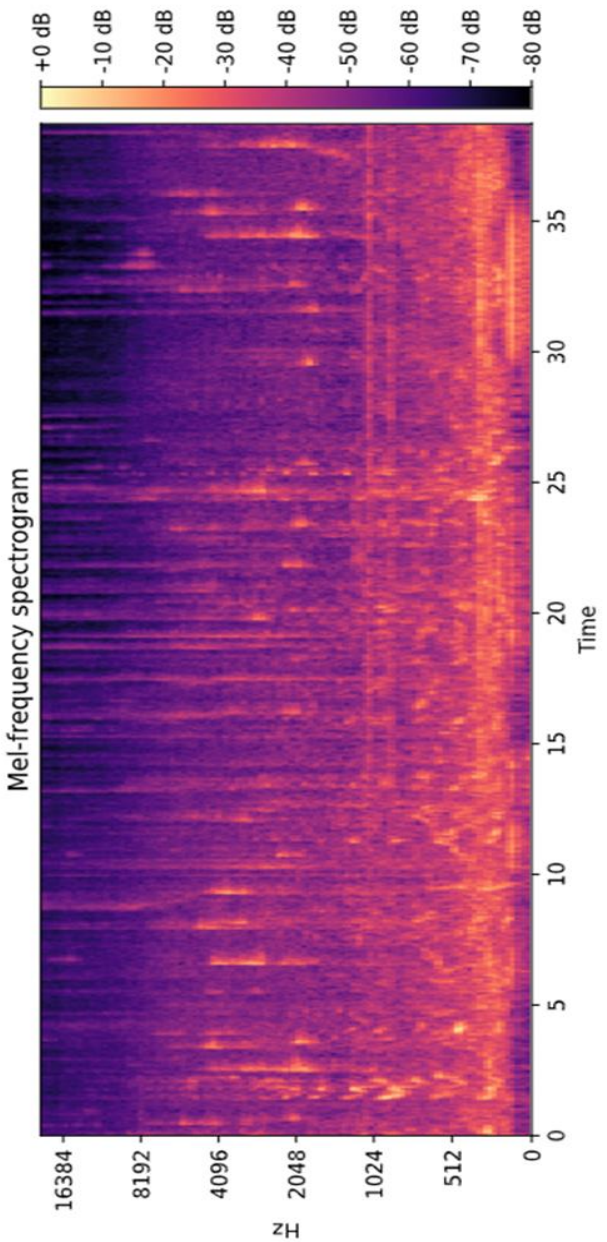
The final dataset used for the CNN model consisted of seven musical instrument classes, bass drum, snare drum, double bass, cello, violin, viola, and saxophone. Table 2 presents the dominant frequency ranges of each class. The dominant frequencies are calculated using the peak frequencies of each audio sample within the musical instrument classes. Since we cleaned the audio samples during preprocessing, samples mostly do not contain noise. However, there can still be a small amount of noise that was not removed with the filter. Like the snare drum, some instruments tend to resonate excessively due to their material of choice, which creates overtones and harmonics. These are not removed with the filter; thus, the signal's frequency domain can consist of broadband frequencies. To cope with this, we only calculated the peak frequencies. The frequency range of the instruments extends beyond what is shown in Table 2. This table presents the frequency ranges each instrument class is dominant.

Table 2: The dominant frequencies of the musical instrument classes used in the dataset.

	Bass Drum	Snare Drum	Double Bass	Cello	Viola	Sax.	Violin
Lower Range (Hz)	10	171	53	91	218	177	218
Upper Range (Hz)	86	1816	417	898	1743	2719	5311
Average (Hz)	54	505	123	365	424	672	497
Standard Dev. (Hz)	18	975	94	210	517	487	574

As stated previously, we conducted a small test with three different training datasets to determine which instrument types to include. We trained the network with each dataset. Training accuracy for Group 1(string Instruments) was 75%, for Group 2(wind instruments) 85%, and Group 3(combination of string and wind) 78%. The instrument types of Bass drum and Snare drum were present in all three datasets since they are used for classifying the impact sounds within the recordings. Like the rest of the instruments, the classification output of the snare drum was slightly different in each dataset. Since it was one of the only two constant instruments in this testing procedure, we conducted a statistical test to see if there is a statistically significant difference between the classification result of snare drums in each setting.

We conducted a one-way ANOVA F-test to compare the output percentage of the snare drum on three datasets (H0: $\mu 1 = \mu 2 = \mu 3$, H1: at least one of the means is different). The test is conducted using IBM SPSS Statistics 25. There was no statistically significant difference between the snare drum output at the $p>0.5$ alpha level for the three datasets. ($(F=2,207) =1.684$, $p=0.18$). Therefore, even though the classification result was slightly different with each dataset, we can say that the algorithm reliably classified the audio content whose spectral characteristic was similar to the snare drum in each dataset.



	Time Frame (2 second intervals)																									
	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	17-18	19-20	21-22	23-24	25-26	27-28	29-30	31-32	33-34	35-36	37-38	39-40	Average
Viola (%)	15	19	4	12	10	10	19	10	13	4	6	12	7	4	14	0	0	24	5	15	10.15
Violin (%)	23	32	10	23	20	19	29	15	21	10	13	19	14	9	21	1	1	44	10	22	17.8
Cello (%)	7	0	1	3	3	11	4	7	5	9	10	16	2	5	14	1	4	7	1	1	5.55
Snare Drum (%)	54	48	84	59	65	39	38	58	56	57	61	18	75	79	35	0	2	9	83	61	49.05
Double Bass (%)	1	0	2	3	1	21	9	9	5	19	9	34	1	3	12	97	92	16	1	1	16.8
Bass Drum (%)	0	0	0	0	0	1	1	1	0	1	0	1	0	0	1	0	0	0	0	0	0.3

Figure 33: Comparison between the Mel-spectrogram of a video and the classification output

To classify the audio content, we extracted the audio of each museum video. Similar to how short-term Fourier transform is computed on a smaller frame of the recording, we performed the classification on smaller time frames. Performing the classification on the whole duration of the video can lead to inaccurate results as the audio signal fluctuates too much on longer time frames. Using smaller time frames for classifying the sound sources can yield higher accuracy. Because of this, we split the museum recordings into smaller frames of two, five, and ten-second long periods and performed the sound classification on these pieces. We later reconstructed the audio recording by combining the classification result of each time frame.

Figure 33 an example of this procedure. In this example, the recording of the museum is deconstructed by split into two-second-long time frames before loading into the CNN model. Sound classification is performed on each of these frames. The resulting classification probabilities are combined based on their respective time frames. The output of this procedure is the reconstruction of that museum's soundscape. The Mel-spectrogram at the top portion of this figure uses frequency, amplitude, and time to represent the soundscape. Our reconstructed output does this by using the musical instrument classification percentages and time periods.

To explain the logic behind how our network classified this recording, we need to know more about this sound environment and examine the Mel-spectrogram. The audio recording shown in this example is taken from an exhibition about minerals, rock formations, and caves in London's Natural History Museum. This environment was a relatively quiet one, containing only the sounds of human speech, water droplets, air conditioning, and sometimes an unrecognizable thematic background sound that is probably part of the exhibition. Upon inspecting the Mel-spectrogram in Figure 33, three continuous sound bands are visible. The wider one is a broadband sound that continues throughout the recording at lower frequencies. The two narrower ones are located around 1024 Hz, and they start to become apparent towards the middle periods of the recording. In addition, there are also specific sound events regularly happening in the higher

frequencies. Lastly, there are two occasions where particular sound events appear in a very broad frequency range within the same time frames of two to three seconds and around the twenty-fifth second.

After listening multiple times and comparing what we hear with what we see in the Mel-spectrogram, we can understand some of these patterns. The broadband low-frequency sound belongs to the air conditioning unit. The other two continuous sounds originate from the thematic sounds. The discrete sound events on higher frequencies are the water droplets and human speech. The particular event right at the beginning and around the twenty-fifth second is also speech, but they are amplified because the recording device was closer to the speaker at these times, whereas the rest of the speech was happening in the background, far away from the recording device.

If we compare the classification output with the Mel-spectrogram, we can see that each sound event is also visible in the classification table. This table shows the classification probabilities of musical instruments for every two-second time frame of the recording. The broadband sound generated by the air conditioner is visible in the table and marked with the purple highlight. It is distinguished by the high percentage of snare drum presence throughout the recording. There is also a very distinct high amplitude low-frequency sound event between the thirteenth and thirty-fifth seconds. It is even tough to understand what this sound belongs to by listening to it as a human, but we think it is part of the thematic sound or the recording device comes very close to an air condition outlet. If we check these time frames in the classification table, we can see that this event is clearly present (highlighted with red). This dominant sound event masked the rest of the frequencies because of its high intensity and proximity to the recording device. This is one of the limitations we will discuss in the following chapter. The two continuous sounds originating from the thematic background sound are highlighted with blue. We can also see the speech, highlighted with orange, at the beginning and twenty-fifth second, as well as some of the other brief broadband sounds.

By classifying the recordings, we prepared a classification table for all sixty museum soundscapes. By comparing the Mel-spectrograms with the classification tables, we learned which sound events (or sources) correspond to particular musical instruments. For example, the bass drum is almost non-existent in Figure 33 because it mostly corresponds to impact sounds such as footsteps. This specific recording did not significantly impact sound; hence, the algorithm did not classify anything as a bass drum. We initially thought the snare drum would correspond to mid to higher-frequency impact sounds, but to our surprise, it corresponded to background noise and other types of broadband sounds, such as the air conditioner. In the video of the cafeteria of the Kröller-Müller Museum, we observed that the kitchenware's sound, which consists of mid and high-frequency content, is classified as viola since it is the instrument that contains the highest frequencies. Human-generated sounds, such as speech, are usually classified as cello or violin, but this largely depends on the gender and age of the speaker. As women and children have a higher-pitched voice, they are even classified as viola.

We tried to use minimal frequency overlap, but Figure 20 and Table 2 show that it is inevitable to have at least some overlap between the instruments. However, our classification algorithm is not solely based on frequencies. MFCCs is the main audio feature for classification, including timbre characteristic to calculate the cepstral coefficients. The MFCCs are supplemented by Tonnetz and Spectral Contrast feature. The classification is probabilities are computed based on all three of these features. We are using the frequencies because the frequency domain is the foundation for all three audio features that we are using, and it is also the most tangible way to describe the sounds.

4.2. Soundscape Perception Survey Results

The sound classification probabilities of the CNN model represent each video's audio content for the soundscape perception survey. The results we gathered from the survey are divided into two categories, demographic information and perceived affective

response. The combination of audio content, demographic information, and perceived affective response forms the input parameters of our FFNN and K-NN models. Upon comparing the models' test accuracy, we saw a vast difference between the models. FFNN model's performance is superior to K-NN, with a test accuracy score of 69%. We used *GridSearchCV* to estimate the K-NN model's optimum parameters, but the model achieved only 50% accuracy even with the optimum parameters. Because of this, we confirmed that the FFNN model is the more suitable machine learning method for our data and proceeded with that model.

For the FFNN model, we used individuals' overall response to the soundscape as the output parameter while training the data, which vary between 1 for very negative and 5 for very positive. Therefore, the outcome of this research is calculating how individuals' will respond to a soundscape based on its audio content and their demographic background. We can manually enter input parameters to the model, which will then use the prediction function and show us how likely someone from a particular background can appreciate the sound environment with that audio content.

We concluded our research once the FFNN model was finalized. However, due to the nature of the machine learning methods, we lack information regarding the associations between our research variables. By obtaining this information, we will predict the soundscape perception with a machine learning method and further understand how well our initial hypothesis of using the musical instruments to measure the audio content of the soundscape works. To achieve this, we conducted a series of linear regression and correlation analyses.

Before going into the statistical analyses, we started with the Seaborn library of Python to plot the associations between perceived affective response variables to explore the data. Figure 34 shows associations between perceived affective response variables. At the same time, these are not strict regression plots. They are good in terms of providing an

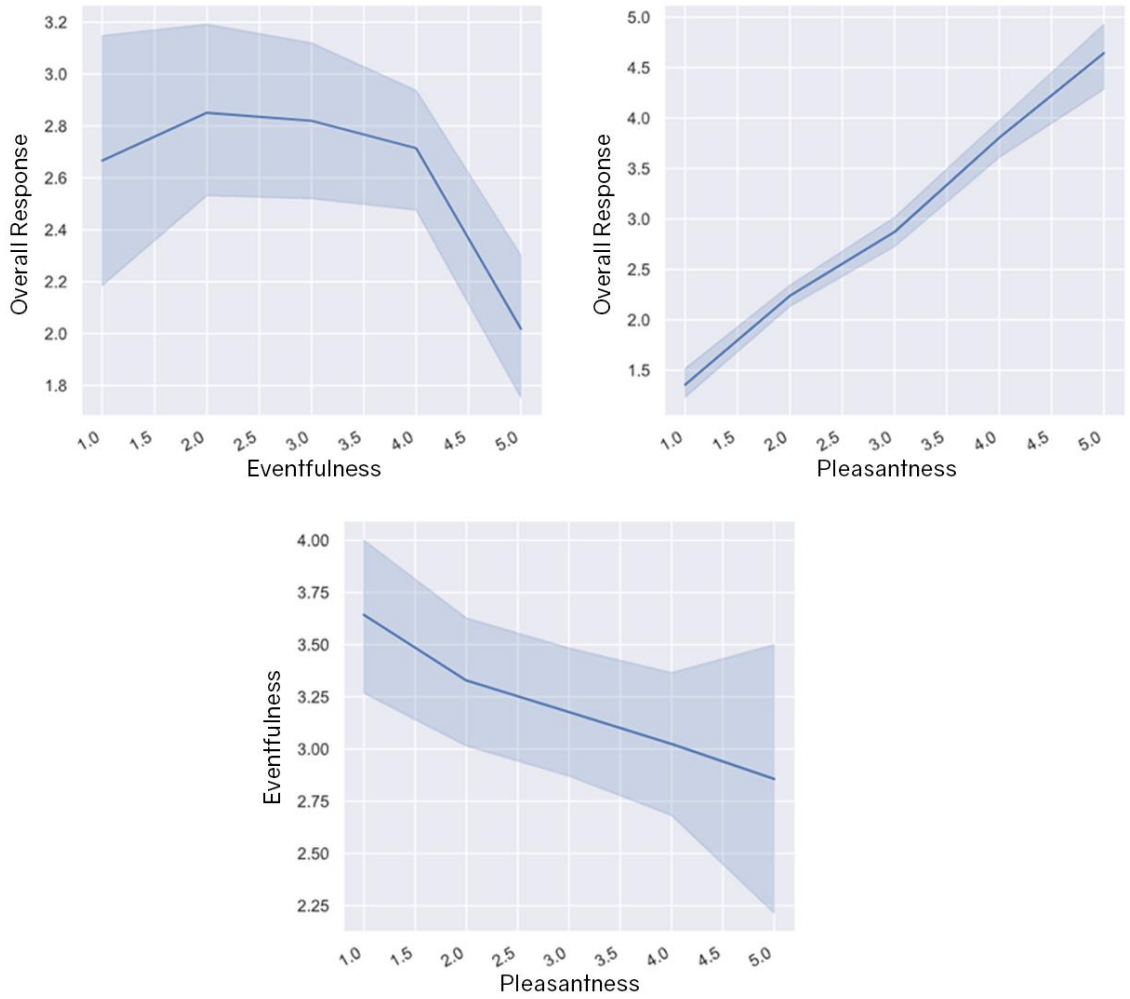


Figure 34: The associations between the three perceived affective response variables. The light blue area around the regression line represents the standard deviation.

initial idea about the association between different variables. These plots are very straightforward and do not require choosing a type of regression. Because of this, they can easily show the nonlinear relationships, whereas if we were doing regression from the start, we would have to try at least a couple of different scatter plots and regression methods to understand the form of the relation.

Figure 34 shows an evident positive relationship between the *Pleasantness* and *Overall Response* to the soundscape (top right plot). This plot is almost demonstrating a perfect positive relationship line. It also shows a low amount of standard deviation. The plot at the bottom shows a potential negative association between *Eventfulness* and

Pleasantness. The plot on the top right is an interesting one as it shows a possible nonlinear association between *Overall Response* and *Eventfulness.* This plot suggests that an uneventful soundscape will not directly lead to a positive response contrary to common beliefs. We will discuss this further in the next chapter.

Table 3: Chi-Square test for association results between the demographic and perceived affective response variables.

	Pearson Chi-Square Value	**Asymptotic Significance (2-sided)**
Age - Pleasantness	61.273	0.000
Age - Eventfulness	66.586	0.000
Age - Overall Response	48.115	0.000
Gender - Pleasantness	2.869	0.580
Gender - Eventfulness	27.149	0.000
Gender - Overall Response	7.099	0.131
Country - Pleasantness *	111.887	0.000
Country - Eventfulness *	180.953	0.000
Country - Overall Response*	122.883	0.000
Language - Pleasantness*	80.407	0.000
Language - Eventfulness*	122.836	0.000
Language - Overall Response*	89.565	0.000
Education - Pleasantness	26.016	0.011
Education - Eventfulness	39.445	0.000
Education - Overall Response	10.444	0.577
Employment - Pleasantness	85.365	0.000
Employment - Eventfulness	88.246	0.000
Employment - Overall Response	67.656	0.000
Area01 - Pleasantness*	16.496	0.036
Area01 - Eventfulness*	15.064	0.058
Area01 - Overall Response*	20.095	0.010
Area02 - Pleasantness*	27.485	0.001
Area02 - Eventfulness*	19.028	0.015
Area02 - Overall Response*	25.520	0.001

* Data count less than expected in some cells

We proceeded to statistical analyses after using plots to explore the relationship between the perceived affective response variables. The first statistical test we conducted is the Chi-Square test for association. Since most of our demographic variables are categorical, we could not use Spearman's rho or Pearson's r correlation tests. Chi-square test for

association requires at least one of the variables to be categorical(nominal), and each variable should include at least two categories.

Table 3 shows the results of the statistical analysis we conducted to examine the association between demographic and perceived affective response variables. The null hypothesis(H_0) for the Chi-square test is that the two variables are not associated, while the alternate hypothesis is the variables are associated (H_A). These results suggest there is an association between most of the variable pairs. However, the data distribution is less than ideal for half of the demographic variables: country, language, the type of area currently living in, and the type of area participants spent most of their lives. For example, 86% of the participants are from Turkey, and only one participant is from

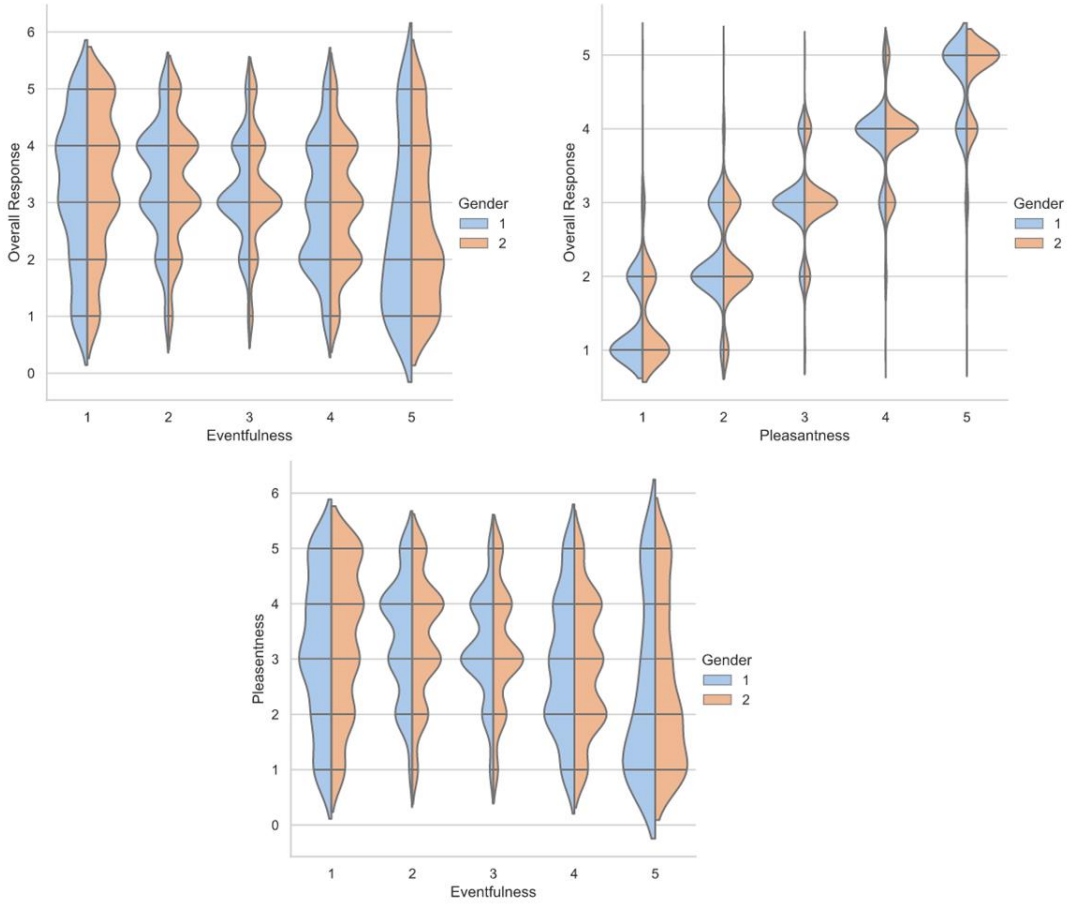


Figure 35: The distribution of perceived affective response variables based on gender.

Australia and Norway. Due to this uneven distribution, the data significance results obtained for these variables are unreliable and cannot be used.

As for the variable that are distributed normally, a statistically significant association is observed between Age and all three perceived affective response variables (Pleasantness= $\chi 2(16, N=3500) =61$, $p=0.000$, Eventfulness= $\chi 2(16, N=3500) =66$, $p=0.000$, Overall Response= $\chi 2(16, N=3500) =48$, $p=0.000$). There is a statistically significant relationship between Gender and Eventfulness ($\chi 2(16, N=3500) =27$, $p=0.000$). A statistically significant relationship also exists for Education-Pleasantness ($\chi 2(16, N=3500) =26$, $p=0.011$) and Education-Eventfulness ($\chi 2(16, N=3500) =39$, $p=0.000$). Figure 35 shows the associations between the perceived affective response variables based on gender. Even though these plots do not include a regression line, the Overall Response and Pleasantness association is quite evident (top right). However, there is a visible difference between the genders in terms of their perceived affective response.

We conducted a correlation between all numeric variables to see associations between them. We mainly focused on the intercorrelations between the three perceived affective response variables of *Pleasantness*, *Eventfulness*, and *Overall Response* and the demographic and audio content variables (musical instruments). We used Spearman's rho correlation coefficients since these data samples are ordinal (Likert scale data) and interval (audio content probabilities). We only included two demographic variables since the rest were categorical variables (gender, preferred language, country of birth, etc.). Because of this, the only demographics variables we used are age and education level. Spearman’s rho correlation coefficients for the audio content, numeric demographic variables, and perceived affective response variables.

Table 4: The Spearman's rho correlation coefficients for the audio content, numeric demographic variables, and perceived affective response variables.

	Pleasantness		**Eventfulness**		**Overall Response**	
	Correlation Coefficient	Sig. (2-tailed)	**Correlation Coefficient**	Sig. (2-tailed)	**Correlation Coefficient**	Sig. (2-tailed)
Violin	.080**	0.000	-.254**	0.000	.083**	0.000
Viola	0.017	0.311	-.233**	0.000	0.029	0.086
Saxophone	-.232**	0.000	.378**	0.000	-.239**	0.000
Cello	.087**	0.000	-.102**	0.000	.067**	0.000
Double Bass	-.051**	0.003	.212**	0.000	-.065**	0.000
Snare Drum	-.133**	0.000	.227**	0.000	-.135**	0.000
Bass Drum	0.020	0.227	0.030	0.073	0.032	0.055
Age	-0.020	0.237	.051**	0.003	-.042*	0.012
Education	-0.020	0.237	0.009	0.600	-0.026	0.117
Pleasantness	1.000	-	-.186**	0.000	.857**	0.000
Eventfulness	-.186**	0.000	1.000	-	-.191**	0.000

* p<0.5 **, p<0.1

Table 4 shows the results of Spearman's rho Correlation Coefficients for our variables. The intercorrelations between audio content variables are not presented in the table since they will not be very meaningful. We cannot interpret the association between double bass and cello. Due to this, we removed the columns with those variables to make the table more readable. The findings indicate that there is a very high positive association between the *Overall Response* of the soundscape and its *Pleasantness* ($r(3498)=.857$, $p<.01$). There is a weak negative association between *Eventfulness* and *Overall Response* to the soundscape ($r(3498)=-.191$, $p<.01$). This indicates that pleasant soundscapes will be perceived positively while eventful ones will be perceived as negative. The negative association between *Pleasantness* and *Eventfulness* ($r(3498)=-.186$, $p<.01$) further supports this as eventfulness decreases the pleasantness of the soundscape. These correlation results are consistent with the plots we prepared to explore the data.

Among the two demographic variables, Age has a statistically significant negative association with the Overall Response ($r(3498) = -.042$, $p<.05$) and a positive one with Eventfulness ($r(3498) = .051$, $p<.01$). However, the correlation coefficients of both associations are very low, and the level is association is very weak. Education level does not have a statistically significant association with any of the variables.

The correlation coefficients between the audio content and perceived affective response variables show a general trend consistent with Pleasantness, Eventfulness, and Overall Response correlations. *Bass Drum* is the only audio content variable with no statistically significant association with any perceived affective response variables. *Saxophone* has a statistically significant moderate positive association with *Eventfulness* ($r(3498) = .378$, $p<.01$) and low negative association with *Pleasantness* ($r(3498) = -.232$, $p<.01$). It is also negatively correlated with the *Overall Response* ($r(3498) = -.239$, $p<.01$). We can interpret this as the soundscapes that contain sound sources with similar spectral characteristics to a saxophone are perceived as eventful. Since *Eventfulness* and *Pleasantness* were negatively correlated, it makes sense that Saxophone is also negatively correlated with *Pleasantness*. Similarly, *Pleasantness* was very positively correlated with *Overall Response*. Thus, *Saxophone* is negatively correlated with *Overall Response*.

Snare drum is another audio content variable that is positively correlated with *Eventfulness* ($r(3498) = .227$, $p<.01$)., and negatively correlated with *Pleasantness* ($r(3498) = -.133$, $p<.01$) and *Overall Response* ($r(3498) = -.135$, $p<.01$). We already know that this musical instrument corresponds to broadband sounds such as background noise and ventilation. Since eventful environments will contain more activity, they will also include more background noise, which is universally unpleasant, and individuals tend to respond negatively. Exposure to background noise for extended periods can lead to a negative physiological state such as fatigue.

Violin (r (3498) = -.254, p<.01), Viola (r (3498) = -.233, p<.01), and Cello (r (3498) = -.102, p<.01) are the only audio content variables that have negative statistically significant association with the eventfulness. However, the viola is not significantly associated with the rest of the perceived affective response variables. *Violin* and *Cello,* on the other hand, are significantly associated with *Pleasantness* (*Violin*=(r(3498)= .080, p<.01), *Cello* = (r(3498)= .087, p<.01)) and *Overall Response*(*Violin*=(r(3498)= .083, p<.01), *Cello*=(r(3498)= .067, p<.01)). The correlation coefficients of both variable pairs are very low, suggesting that their effects on *Pleasantness* and *Over Response* are minimal—these three musical instruments mostly corresponded to speech and sounds generated by children.

Table 5: Linear regression models between perceived affective response variables

F	***p-value***	**Std. Error**	**R**	**R^2**	**R^2 Adjusted**	***x*1**	***x*2**	**VIF**
9638.347	0.000	0.592	0.857	0.734	0.734	X		1.000
138.695	0.000	1.124	0.195	0.038	0.038		X	1.000
4839.836	0.000	0.591	0.857	0.735	0.734	X	X	1.039

*X*1: Pleasantness, *X*2: Eventfulness, Dependent variable: Overall Response

Table 5 presents the linear regressions between the perceived affective response variables. Since Overall Response is the output layer of our FFNN model, the linear regression models are adjusted to demonstrate the effect of Pleasantness and Eventfulness over the Overall Response. We conducted the linear regression one by one for each independent variable and then conducted a multiple linear regression using both variables. All regression models reached statistical regression at alpha level p<0.01. Similar to the correlations, the highest performing variable was Pleasantness F (2, 3497) = 9638.347, p < .000, R2 = .734. Even though Eventfulness reached statistical significance, its model parameters are not very suitable since it has a high standard error and very low R2 coefficient compared to others F (2, 3497) = 138.695, p < .000, R2 =.0.038 The multiple regression model performed almost identical to Pleasantness. F (2, 3497) = 4839.836, p < .000, R2 = .735. These results further point out the low effect of *Eventfulness* on the dependent variable. VIF score does not indicate any collinearity between the independent variables; therefore, the multiple regression model is suitable for our purpose.

Table 6:Linear regressions between the musical instruments and perceived affective response variables.

F	*p-value*	Std. Error	R	R^2	R^2 Adjusted	*DV1*	*DV2*	*DV3*
58.120	.000[b]	1.159	.323[a]	0.104	0.103	X		
141.913	.000[b]	1.152	.471[a]	0.221	0.220		X	
53.268	.000[b]	1.091	.311[a]	0.096	0.095			X

DV1: Pleasantness, DV2: Eventfulness, DV3: Overall Response, Independent variables: Musical Instruments

In Table 6, the linear regressions between musical instruments (audio content) and perceived affective response variables can be seen. In these regressions, we kept the independent variables constant but changed the dependent variable to see the effect of audio content on perceived affective response. The alpha levels indicate that all three independent-dependent variable combinations reached statistical significance at $p<0.000$. However, the R2 coefficients of all pairs are very low.

Upon further inspecting the regression coefficients of each variable pairs in Tables 7,8 and 9, we can see that the collinearity statistics are very high. VIF scores are over the cut-off value of 10, and the Tolerance scores are below 0.1 for most of the variables. This indicates that musical instruments are highly correlated with each other, which was also evident in the correlation coefficients previously. This suggests that the independent variables are not independent. Multicollinearity reduces the precision of the regression model by effecting the regression coefficients and p-values.

In terms of statistical significance, most independent variables reached statistical significance at least at alpha level $p<0.05$. Some variables, such as Double Bass and Snare drum in Table 6, may lack the variance to research statistical significance. A larger dataset can provide a larger variance for the variables, making the associations more evident.

Table 7: Multiple linear regression coefficients of musical instruments and Pleasantness.

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	2.570	0.633		4.062	0.000		
Viola	0.034	0.007	0.536	5.134	0.000	0.024	42.455
Violin	-0.027	0.007	-0.373	-3.939	0.000	0.029	34.886
Saxophone	-0.072	0.009	-0.249	-8.381	0.000	0.291	3.432
Cello	0.056	0.020	0.046	2.756	0.006	0.905	1.105
Double bass	0.006	0.007	0.111	0.917	0.359	0.018	56.700
Snare drum	0.004	0.007	0.064	0.668	0.504	0.028	35.270
Bass drum	0.015	0.006	0.196	2.330	0.020	0.036	27.580

a. Dependent Variable: Pleasantness

Table 8: Multiple linear regression coefficients of musical instruments and Eventfulness.

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	4.749	0.629		7.549	0.000		
Viola	-0.055	0.007	-0.801	-8.229	0.000	0.024	42.455
Violin	0.011	0.007	0.139	1.571	0.116	0.029	34.886
Saxophone	0.107	0.009	0.347	12.543	0.000	0.291	3.432
Cello	0.004	0.020	0.003	0.177	0.860	0.905	1.105
Double bass	-0.017	0.007	-0.288	-2.564	0.010	0.018	56.700
Snare drum	-0.019	0.007	-0.256	-2.890	0.004	0.028	35.270
Bass drum	-0.030	0.006	-0.366	-4.667	0.000	0.036	27.580

a. Dependent Variable: Eventfulness

Table 9: Multiple linear regression coefficients of musical instruments and Overall Response.

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	2.459	0.595		4.129	0.000		
Viola	0.031	0.006	0.517	4.930	0.000	0.024	42.455
Violin	-0.020	0.006	-0.294	-3.095	0.002	0.029	34.886
Saxophone	-0.063	0.008	-0.232	-7.798	0.000	0.291	3.432
Cello	0.045	0.019	0.040	2.352	0.019	0.905	1.105
Double bass	0.006	0.006	0.126	1.043	0.297	0.018	56.700
Snare drum	0.006	0.006	0.096	1.006	0.315	0.028	35.270
Bass drum	0.016	0.006	0.228	2.698	0.007	0.036	27.580

a. Dependent Variable: Overall Response

CHAPTER V

DISCUSSION

The hypothesis of this research was machine learning can be used to classify the sound sources that contribute to the soundscape, and perception of the soundscapes can be predicted based on these sound sources. We first developed an environmental sound classification model with musical instruments and identified the audio content of soundscapes. We conducted a soundscape perception survey and combined its findings with the audio content data to develop an FFNN model which predicts individuals' overall response to the soundscapes.

5.1. Demographic Variance

Conducting an online survey enabled us to reach a broader participant base, but it also limited us considerably due to lack of control. One of the limitations caused by this was regarding the demographics. Based on our findings from the pilot study, we initially thought that demographics would significantly affect the perception of soundscapes. Yet, we did not observe this influence as much as we thought we would, and when there is indeed a statistically significant association with demographics, the data samples are not enough. This is caused due to the lack of variance within the demographic data. Since we did not have complete control over the participant selection, the demographic information

was not evenly distributed, as we reported under section 3.4.3.2. Participants. More than half of the participants were aged between 25 and 34(56.7%), more than a quarter aged between 18 and 24(28%). Only one-quarter of the participants were older than thirty-five years old, which meant that the demographic data barely represented them. The correlation scores shown in Table 4 did present statistically significant difference between participants' Age and their perceived Pleasantness and Overall Response to the soundscape, but the amount of the association is low. A stronger association might be present, but we do not have the dataset to observe it. Similarly, Figure 35 did not demonstrate a noticeable difference between participants' perceived affective responses regarding gender, but 67% of the participants were female.

Previous studies pointed out that the acoustic environment individuals' got exposed in their living environment significantly affected their evaluation of sound levels in urban environments (Yu & Kang, 2008). To see if this effect is also true for indoor sound environments, we asked the participants what kind of residential area they are currently living in and what type of area they spent most of their lives. Individuals who spend most of their life in a rural area can consider the major cities very loud, while it is just normal for people born and who spent their whole lives in the city. However, our results

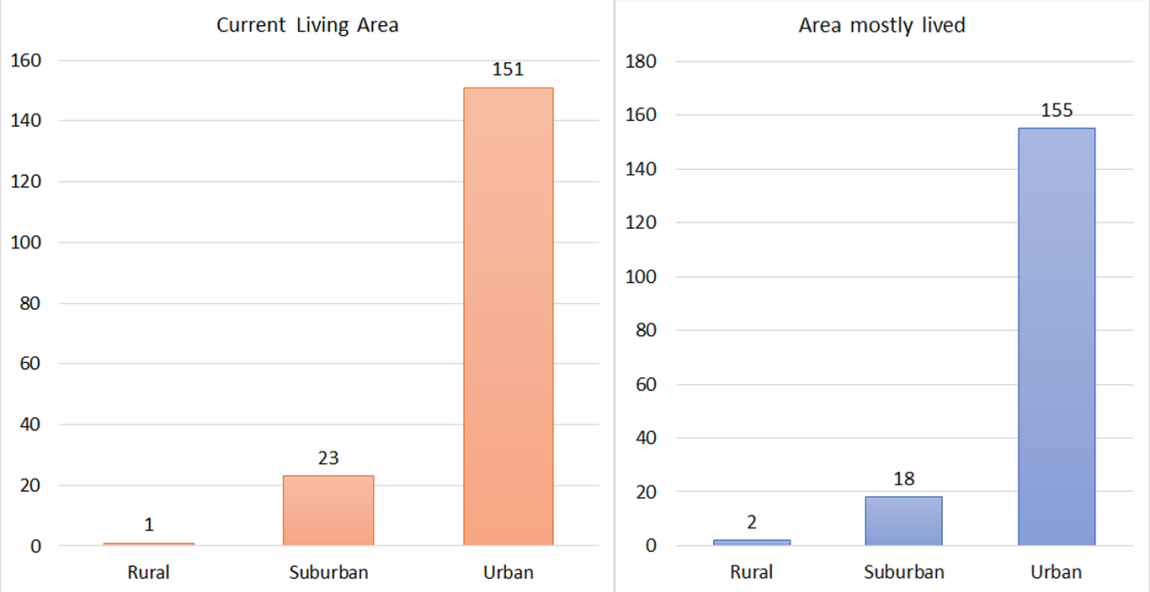


Figure 36: Participants' current type of living area and the type of area they spent most of their lives in.

indicated no significant difference. This is also caused due to the lack of variance within the data. In Figure 36, we can see that the vast majority of the participants lived and currently living in an urban environment. From our sample size of one hundred and seventy-five participants, 85% live in an urban setting, and 88% are from Turkey. The research of Yu and Kang (2008) included participants in nine cities from six countries and with a sample size of more than ten thousand. Therefore, we needed a far larger sample size and variance to observe the effect of demographic and cultural background on evaluating soundscapes.

If the demographical data had enough variance for each group, we would have seen stark contrasts between the perceived affective response of each group. For example, based on what we learned from our pilot study, we were expecting that those who spent their life in a rural area would respond to eventful soundscapes more negatively when compared to someone who lived in an urban center. The ones that lived in the city are more used to sounds emanating from the chaotic and loud events like the traffic, and therefore will be less discerning towards eventful soundscapes. The study conducted by Yu and Kang (2008) also indicated that individuals' evaluation of the soundscape is affected by the acoustic environment of their homes.

5.2. Multi-event Classification

One of the limitations of sound classification occurs when multiple sound events happen simultaneously (polyphonic). This is a significant drawback for Music Information Retrieval (MIR) systems (Chithra et al., 2015). It is especially hard to retrieve the audio information when multiple instruments or vocal performers are performing simultaneously. Dealing with this issue uses methods similar to Sound Event Detection (SED) models. In the SED models, environmental sound recordings are annotated (fully or partially labeled) based on each sound event's onset and offset time. These annotations are used to train the neural network, along with the environmental sound recordings. MIR systems that are designed to detect multi-instruments in polyphonic recordings are

designed similarly. Each recording in the training set is annotated based on the instrument classes (Anhari, 2020).

In our examples, we also observed instances where a particular musical instrument dominates the classification of a time frame while more than one sound event was happening simultaneously. As much as this is a limitation of our model, its consequences are not as vital as MIR models. The aim of the classification model was never about identifying each sound source within the soundscape. We are using the spectral characteristics of musical instruments as a reference to analyze the spectral character of the soundscape and make inferences about its audio content based on their similarities and differences. Due to this, not detecting the simultaneous sound events does not significantly impact our results.

When Schafer and his team introduced the soundscape approach, they defined three features of soundscapes, *Keynote Sounds*, *Signal Sounds*, and *Soundmarks* (Schafer, 1977). *Keynote Sounds* are the fundamental tone that all other acoustic materials modulate on (Schafer, 1977).In other words, they are the background sounds that we do not perceive consciously, but we feel their absents when they stop. On the other hand, *Signal Sounds* are the foreground sounds that we listen to consciously. In some cases, signal sounds are specifically meant to be heard, such as alarms and car horns. In a figure-ground relationship, *Keynote Sounds* are the ground, while *Signal Sounds* are the figure.

Being a composer, Schafer was inspired by the music theory when he introduced the soundscape features. There is always a keynote sound in polyphonic/multi-instrument classification problems, such as the bass guitar or double bass playing in the background and multiple instruments playing in the foreground as the signal sound. Almost all of our museum recordings also include keynote sounds in terms of background sounds. These mainly occur due to the building's mechanical systems or human activity happening

further away from the listener. In most cases, signal sounds originate from sound events near the listener (unless it is an actual signal like an alarm). Especially in large and crowded indoor spaces, everything can become a part of the background noise, turning into a keynote sound, while few sound events remain the signal sounds. These are the sounds that we pay attention to while the rest blends into the background. The classification model we developed tends to provide a higher classification probability on dominant sound events. From a machine learning perspective, this is a limitation, but this limitation is not evident in the final result from a soundscape perception. Sounds are needed to be paid attention to by individuals to contribute to the overall perception of the soundscape (Boes et al., 2018; De Coensel et al., 2009). In multi-classification, the model we use tends to favor the dominant signal sounds when computing the probabilities rather than the background noise that forms the keynote sounds. These are the sounds individuals pay attention to and the ones that contribute to the soundscape perception. Because of this, what seems to be the limitation of the model is actually making the model behave like the notice-event model of De Coensel et al. (2009). This model suggests that the perception of environmental sounds is based on consciously noticed sounds (De Coensel et al., 2009). Noticing a sound event is based on its sound characteristics, particularly on the signal-to-noise ratio and the attention paid to it by the listener.

5.3. Data Augmentation

One possible improvement that can be included in the future study is introducing more musical instruments to the classification data set to get more variance in classification probabilities. We are currently using a dataset that is more focused on the frequency range between sub-bass and midrange or between 20Hz to 2000Hz. This does not mean that they do not cover above 2000Hz since the dominant frequencies of viola and violin are above that range. However, the precision at higher frequencies is not as good as sub-bass to midrange area due to fewer training samples. A more comprehensive dataset can be created and used for a more detailed classification output by conducting extensive testing with numerous instruments.

Data augmentation is typically used for increasing the size of the dataset by modifying the existing data. Larger datasets become more accessible every year, but state-of-the-art ML models require data parameters in millions. The idea behind data augmentation is deforming the labeled data without changing its semantic meaning (Salamon & Bello, 2017). The network's generalization on unseen data is improved by training it on these additional data samples. In environmental sound classification, the most common application of data augmentation is stretching the sample time, shifting its pitch, and shifting its time (Salamon & Bello, 2017).

There are many examples of increasing the size of the training data set with data augmentation(Aytar et al., 2016; Mushtaq & Su, 2020; Salamon & Bello, 2017; Sze et al., 2017), but its applications do not have to be limited with the training dataset. We can also use data augmentation for environmental sound recordings for future studies. Finding museum videos with different sound content is a time-consuming task. Each video should be different from the rest but should also contain just the right amount of different sound events to be meaningful to evaluate. Data augmentation can control creating synthetic soundscapes, which we can form for our research purposes. For example, we can manipulate the existing museum recordings and simulate different audio conditions of the original soundscape and compare the participants' responses to the soundscape. The fact that indoor soundscapes contain large amounts of speech and typical data augmentation methods like time-stretching will not be suitable. However, data augmentation is used for speech recognition(Rebai et al., 2017; Song et al., 2020), so we can find a method for applying it to indoor soundscapes without making the speech in the recordings meaningless.

5.4. The Effect of Eventfulness on Perception

We showed the differences between the variables using Spearman's rho Correlation Coefficients, linear regression models, and plotting the associations. Among these, one of the interesting findings is the association between *Eventfulness* and *Overall Response*. The relation we observed between these two variables in Figure 34 supports the idea that quiet settings do not necessarily lead to an overall positive response. Typically, we would expect uneventful settings to lead to a positive response, but this figure suggests that it leads to a neutral and maybe even negative response to the soundscape.

One explanation for this is based on the affect model of Russell (1980) and principal soundscape components of Axelsson et al. (2010). As we previously discussed in Chapter 2: Theoretical Background, our *Eventfulness* and *Pleasantness* variables come from Axelsson and his colleges (2010). In their model, authors compared their eventfulness components with Rusell's (1980) arousal component. According to this, exciting and calm soundscapes are both pleasant. The difference between them is the degree of eventfulness. An exciting soundscape is an eventful one, which mixes high arousal with pleasure, while a calm soundscape combines low arousal with pleasure (Axelsson et al., 2010). When low arousal is mixed with a low amount of pleasure, the soundscape becomes monotonous. This can be one of the reasons why uneventfulness is not as desired as we thought it would. While the results do not suggest that uneventfulness leads to a positive response to the soundscape, it certainly indicates that some activity is desired.

Figure 37 shows the Kernel Density Estimation(KDE) of *Eventfulness* and *Overall Response* we prepared with Seaborn. A KDE plot estimates the probability density function from the data samples and plots the distribution of variables for comparison. This plot shows that when the eventfulness is high, the overall response is negative, and when it is low, the overall response tends to be positive. However, the densest part of the plot is around the middle part of the figure. Participants ' response to the soundscape was

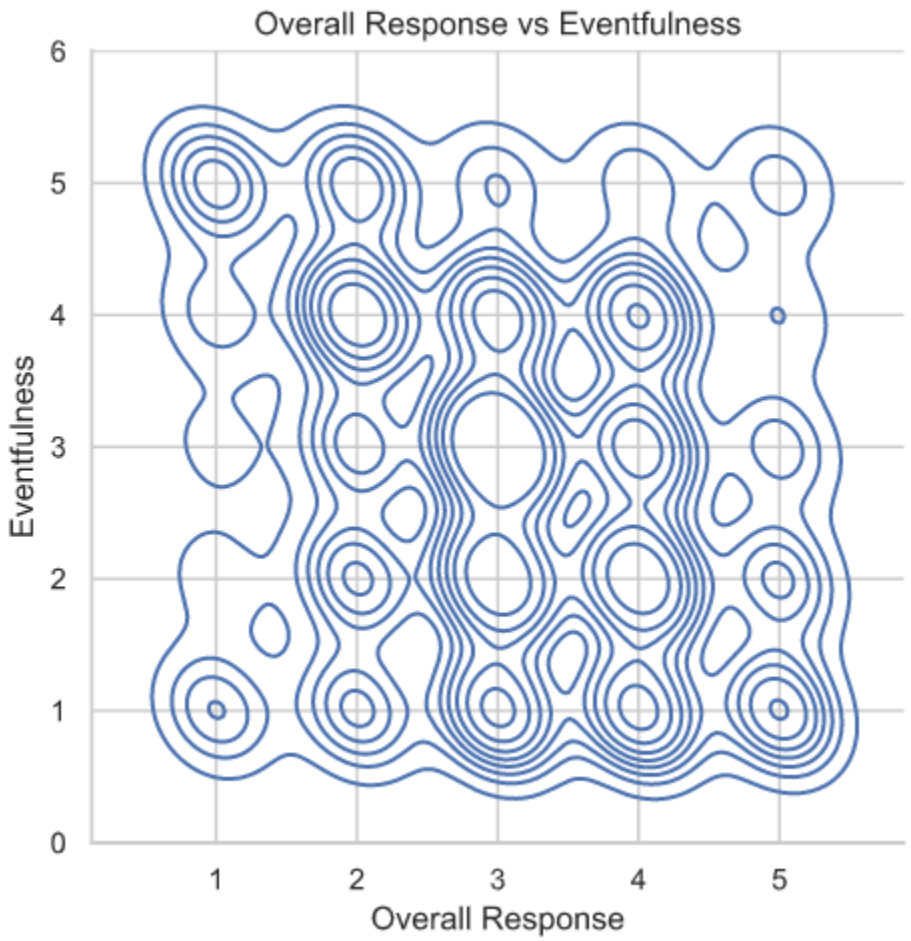


Figure 37: Kernel Density Estimation graph that shows the distribution of Eventfulness and Overall Response data samples

mostly indifferent or positive when there are a moderate amount of eventfulness(between 2 to 4) participants. This does suggest that some degree of *Eventfulness*, thus sound, is desired.

The other reason some degree of *Eventfulness* is desired is due to expectation. According to Truax (1984), soundscape competence is tacit knowledge. We learn the acoustic elements that form the soundscapes and associate meaning to them through experience. Because of this, individuals expect certain sounds to be present in particular settings. Imagine going to a museum you have been to before. You expect to see and hear certain elements because of your previous experience. When you know that the museum is

usually full of people but currently almost empty, it will affect how you perceive the environment.

One advantage of having a certain amount of eventfulness is the masking effect caused by the activity. We previously observed this in open-plan offices(Volkan Acun & Yilmazer, 2018). In the office environment, individuals' expressed that complete silence is just as bad as noise. When there are too few sounds within the environment, every sound event becomes more annoying and disturbing. Having a certain amount of background noise makes the regular sound events less disturbing as they can blend into the background noise. In terms of speech, background noise reduces speech intelligibility, thus reducing the disturbance caused by it.

5.5. The Potential Effect of Expectation and Preference

While not directly part of the soundscape survey we conducted, we need to discuss the effect of expectation and preference on the perception of soundscapes. We hypothesized from the beginning that expectation and preference have an indirect impact on the interpretation of the soundscape. These two categories emerged right at the beginning, during the grounded theory phase of our pilot study. The expectation was the only exogenous variable of our conceptual model, while preference was an endogenous variable (only depending on expectation). We tested their associations with other categories of the conceptual model with the SEM approach, but both categories performed poorly, and the hypotheses paths associated with them did not show any statistical significance.

Based on our previous qualitative soundscape research experience(Volkan Acun et al., 2016, 2018; Volkan Acun & Yilmazer, 2018, 2019; Yilmazer & Acun, 2018a, 2018b) and literature (Bruce & Davies, 2014), expectation has an important effect on the interpretation of soundscapes. Our implementation of the SEM approach failed to measure this effect. This can potentially be caused due to an unobserved variable that we

were unable to include in the model. Their relation in the model can also be different than what we initially thought as well. Since the SEM part of the pilot study was concerned with only measuring the strength of the relationships between the categories of the conceptual model, we did not change the layout of the model. Lastly, the associations of these two variables might not be measured with linear models.

Artificial Neural Networks are especially powerful in terms of measuring non-linear relationships. We were initially planning on using this advantage for further addressing the issue with expectation and preference. However, since the pandemic forced us to conduct the survey online, we could not address this issue. Soundscape expectation is especially hard to measure in an online format as it depends on prior knowledge and familiarity. The museum videos we used for the survey are unfamiliar environments for almost every single participant. Because of these, we could not further investigate the effect of expectation. We are determined to address this issue for future research by identifying the unobserved variables and testing out nonlinear models.

CHAPTER VI

CONCLUSION

This research aimed to develop a machine learning-based sound classification model for analyzing the audio content of soundscapes and using this model to evaluate the association between the audio content and perception of the soundscape. To achieve this goal, we used machine learning and statistical analysis methods and a soundscape perception survey, which we had to conduct online. We first developed a convolutional neural network (CNN) model to classify the sound sources that contribute to the audio content of the soundscape. This model computed the audio content of the soundscape in terms of classification percentages. We then developed a feedforward neural network (FFNN). This network combined the audio content output of the CNN model with the perception data of the questionnaire survey to measure and make a prediction about the overall response to the soundscape. Finally, we conducted a series of statistical analyses to further visualize the associations between the variables we used in the FFNN model.

Based on the outputs of the machine learning models and the statistical analyses, we can conclude that audio content can be used for measuring the perceived affective quality of soundscapes. The effect of sound sources on the perception of soundscapes is already established, yet it is typically used for describing the soundscape rather than measuring how it is perceived. This can be seen from the fact that even the ISO 12913-2 standard

includes minimal sound sources as part of the data collection process, and even those are oriented towards outdoor soundscapes. The model we are proposing is not just addressing this issue, but because we do not have a pre-determined list of sound sources, it can be used for any soundscape. We used museums as case study settings in this research only to confirm our hypotheses.

Our main research question was if we can use machine learning to classify the sounds that contribute to the audio content of soundscapes. Using machine learning to analyze the sound recordings is initially an environmental sound classification problem. There are certain methods to classify environmental sounds, but they suffer from significant drawbacks. Their classification scope is limited to the number of classes in their datasets or requires considerable time and computation power. We addressed these drawbacks by using musical instruments for the training dataset rather than environmental sound sources. This allowed us to use musical instruments as indicators to capture the audio content of the soundscape, by classifying the audio content based on how similar their spectral characteristics are to a particular musical instrument. The CNN model used to classify the sound sources provided a reasonable accuracy score of 75% and confirmed this method is suitable for the task.

One of the research questions of this thesis was learning if there is an association between the audio content of the soundscape and its perceived affective quality. Based on the correlation results, we can say that there is a statistically significant association between these two. Our correlation results indicated that there is at least a low to moderate association between the musical instruments, which we used as indicators of audio content, and the perceived affective response variables of *Pleasantness*, *Eventfulness*, and *Overall Response*. There are very few cases where there is no statistically significant association between a particular audio content indicator and a perceived affective response variable, but this primarily due to the amount of variance. Based on the results, we are confident the strength of these associations will improve with a higher amount of data.

The statistical analyses showed us how the presence and amount of a particular audio content indicator affect the soundscape perception. For example, background noise and ventilation sounds are classified under snare drum by the CNN model. Statistical analysis showed us that the indicator of the snare drum is positively correlated with *Eventfulness*, which is negatively correlated with *Pleasantness* and *Overall Response*. This one example illustrates how we can make an inference about the perception of soundscapes by using the correlation between the audio content and perceived affective response. These correlation and regression results, however, are prone to nonlinearity. These statistical methods may not fully explain the variables if the association between them is nonlinear. This was the case for the association between our *Eventfulness* and *Overall Response* variables. The FFNN model we build computes similar relations, but instead of using correlations and linear regression, it adjusts the weights of input parameters and uses the activation functions. It provides a more reliable prediction since neural networks are developed for solving nonlinear classification and regression problems.

Conducting an online questionnaire survey had its advantages and challenges. Its major advantage was the ease of reaching participants since they do not need to come to the laboratory environment. This enabled us to collect the data in a relatively short amount of time, but it came at the expense of having minimal control over the experimental procedure. By adding a hearing test and adjusting the sound levels before starting the listening tests, we tried to recreate the methods used in typical face-to-face listening tests as much as possible. Regardless of how well we implement these procedures, we had to trust that the participants followed the instructions. Despite all the limitations, conducting an online survey enabled us to compare the soundscapes of various museum environments, which would not be possible in the given time frame otherwise

Another challenge we faced due to the pandemic was regarding the recordings of soundscapes. We were initially planning on recording the soundscapes of Rahmi Koç Museum and Erim Tan Museum from Ankara. We conducted a pilot study in one of these museums; therefore, we were familiar with the environment. This might not be a

limitation since we ended up using videos from several different museum environments, but we again lacked control over these environments. If we had control over the recordings, we would be able to choose the time and setting. Using videos, on the other hand, allowed us to measure the responses for different museum contexts. A potential future study can use in-situ recordings of several pre-determined environments. Data augmentation can be used for these recordings to simulate different environmental conditions, such as crowd density and examine how it affects individuals' responses.

Another topic that can be addressed in a future study is expectation and preference. Our pilot study failed to identify any association with any factors for these two variables. We could not address this topic in the thesis since we collected the data online, but we are confident that expectation and preference affect the perceived affective quality of soundscapes. In a future study, these factors can be addressed by adding new variables that can potentially visualize their effects and through nonlinear models.

This research is a proof-of-concept attempt to show that machine learning can analyze and predict soundscape perception. Once we predict how individuals' will respond to soundscapes, we can use this knowledge to manipulate the soundscapes, essentially designing them. We were initially planning a third part of the study in which the output from the FFNN model was going to be used for composing positive soundscapes (or negative). We had to abandon that part due to time constraints, but this can still be accomplished in a future study. In a study like this, we can generate synthetic soundscapes based on the predictions of this model. We can then conduct a listening test with synthetic soundscapes and compare its results with perceived affective response predictions.

REFERENCES

Acun, V., & Yilmazer, S. (2018). Understanding the indoor soundscape of study areas in terms of users' satisfaction, coping methods and perceptual dimensions. *Noise Control Engineering Journal*, *66*(1).

Acun, Volkan, & Yilmazer, S. (2018). A grounded theory approach to investigate the perceived soundscape of open-plan offices. *Applied Acoustics*, *131*, 28–37. https://doi.org/10.1016/j.apacoust.2017.09.018

Acun, Volkan, & Yilmazer, S. (2019). Combining Grounded Theory (GT) and Structural Equation Modelling (SEM) to Analyze Indoor Soundscape in Historical Spaces. *Applied Acoustics*, *155*, 515–524. https://doi.org/doi.org/10.1016/j.apacoust.2019.06.017

Acun, Volkan, Yilmazer, S., & Orhan, C. (2018). Indoor Soundscape of Historical Spaces: The Case of Çengelhan Caravanserai. In E. C. and E. on N. C. Engineering (Ed.), *Euronoise 2018*. http://www.euronoise2018.eu/docs/papers/415_Euronoise2018.pdf

Acun, Volkan, Yilmazer, S., & Taherzadeh, P. (2016). Perceived Auditory Environment in Historic Spaces of Anatolian Culture : a Case Study on Haci Bayram Mosque. *23rd International Congress on Sound & Vibration (ICSV23)*. https://doi.org/DOI: 10.1080/01426399208706361

Adavanne, S., Pertila, P., & Virtanen, T. (2017). Sound event detection using spatial features and convolutional recurrent neural network. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 771–775. https://doi.org/10.1109/ICASSP.2017.7952260

Aggarwal, C. C. (2018). Neural networks and deep learning : a textbook. In *Machine Learning*. Springer. https://doi.org/10.1111/j.1464-5491.2005.01480.x

Aletta, F., & Astolfi, A. (2018). Soundscapes of buildings and built environments. *Building Acoustics*, *25*(3), 195–197. https://doi.org/10.1177/1351010x18793279

Aletta, F., Botteldooren, D., Thomas, P., Devos, P., Van de Velde, D., De Vriendt, P., &

Vander Mynsbrugge, T. (2017). Monitoring Sound Levels and Soundscape Quality in the Living Rooms of Nursing Homes: A Case Study in Flanders (Belgium). *Applied Sciences*, *7*(9), 874. https://doi.org/10.3390/app7090874

Aletta, F., Guattari, C., Evangelisti, L., Asdrubali, F., Oberman, T., & Kang, J. (2019). Exploring the compatibility of "Method A" and "Method B" data collection protocols reported in the ISO/TS 12913-2:2018 for urban soundscape via a soundwalk. *Applied Acoustics*, *155*, 190–203. https://doi.org/10.1016/j.apacoust.2019.05.024

Aletta, F., Kang, J., & Axelsson, Ö. (2016). Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landscape and Urban Planning*, *149*, 65–74. https://doi.org/10.1016/j.landurbplan.2016.02.001

Anhari, A. K. (2020). Learning multi-instrument classification with partial labels. *ArXiv*, 1–4.

ANSI. (1995). *ANSI S3.20-1995, Bioacoustical Terminology*. ANSI.

Axelsson, Ö., Nilsson, M. E., & Berglund, B. (2010). A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*, *128*(5), 2836–2846. https://doi.org/10.1121/1.3493436

Aytar, Y., Vondrick, C., & Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems(NIPS 2016)*, 892–900.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., Deledalle, C. A., & Li, W. (2019). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America 146, 3590*(2019). https://doi.org/10.1121/1.5133944

Boes, M., Filipan, K., De Coensel, B., & Botteldooren, D. (2018). Machine Listening for Park Soundscape Quality Assessment. *Acta Acustica United with Acustica*, *104*(1), 121–130. https://doi.org/10.3813/AAA.919152

Brown, A. L., Kang, J., & Gjestland, T. (2011). Towards standardization in soundscape

preference assessment. *Applied Acoustics*, *72*(6), 387–392. https://doi.org/10.1016/j.apacoust.2011.01.001

Bruce, N. S., & Davies, W. J. (2014). The effects of expectation on the perception of soundscapes. *Applied Acoustics*, *85*, 1–11. https://doi.org/10.1016/j.apacoust.2014.03.016

Burkard, R. (2016). Hearing Disorders. In *International Encyclopedia of Public Health* (pp. 512–519). Elsevier Inc. https://doi.org/10.1016/B978-0-12-803678-5.00198-3

Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. https://doi.org/10.1016/j.peva.2007.06.006

Chen, Y., Guo, Q., Liang, X., Wang, J., & Qian, Y. (2019). Environmental sound classification with dilated convolutions. *Applied Acoustics*, *148*, 123–132. https://doi.org/10.1016/j.apacoust.2018.12.019

Chithra, S., Sinith, M. S., & Gayathri, A. (2015). Music information retrieval for polyphonic signals using hidden Markov model. *Procedia Computer Science*, *46*, 381–387. https://doi.org/10.1016/j.procs.2015.02.034

Chu, S., Narayanan, S., & Kuo, C.-C. J. (2009). Environmental Sound Recognition With Time–Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1142–1158. https://doi.org/10.1109/tasl.2009.2017438

Dabrowski, A., & Marciniak, T. (2017). Audio signal processing. In *Digital Systems and Applications* (Vol. 185, Issue 2008, pp. 169–170). Springer. https://doi.org/10.1201/9780849386206

Davies, R. E. (2017). Basic classification concepts. In *Computer Vision* (pp. 365–398). Academic Press. https://doi.org/10.1016/b978-0-12-809284-2.00013-7

Davies, W. J., Adams, M. D., Bruce, N. S., Cain, R., Carlyle, A., Cusack, P., Hall, D. A., Hume, K. I., Irwin, A., Jennings, P., Marselle, M., Plack, C. J., & Poxon, J. (2013). Perception of soundscapes: An interdisciplinary approach. *Applied Acoustics*, *74*(2), 224–231. https://doi.org/10.1016/j.apacoust.2012.05.010

De Coensel, B., Botteldooren, D., De Muer, T., Berglund, B., Nilsson, M. E., & Lercher, P. (2009). A model for the perception of environmental sound based on notice-events. *The Journal of the Acoustical Society of America*, *126*(2), 656–665. https://doi.org/10.1121/1.3158601

Dinesh Yadav. (2019, December 6). *NLP: Building Text Cleanup and PreProcessing Pipeline | by Dinesh Yadav | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G., & Huopaniemi, J. (2006). Audio-based context recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(1), 321–329. https://doi.org/10.1109/TSA.2005.854103

European Parliament and Council of the European Union. (2002). Assessment and management of environmental noise (EU Directive). In *Official Journal of the European Communities* (Issue L189, pp. 12–25). https://doi.org/10.1016/j.jclepro.2010.02.014

Fayek, H. (2019). *Speech Processing for Machine Learning : Filter banks , Mel-Frequency Cepstral Coefficients (MFCCs) and What ' s In-Between Pre-Emphasis*. https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

Gao, R. X., & Yan, R. (2006). Non-stationary signal processing for bearing health monitoring. *International Journal of Manufacturing Research*, *1*(1), 18–40.

goleman, daniel; boyatzis, Richard; Mckee, A. (2019). Dive into Deep Learning. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks.

Pattern Recognition. https://doi.org/10.1016/j.patcog.2017.10.013

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. *Proceedings of the ACM International Multimedia Conference and Exhibition*, 21–26. https://doi.org/10.1145/1178723.1178727

Heittola, T., Mesaros, A., Eronen, A., & Virtanen, T. (2013). Context-dependent sound event detection. *Eurasip Journal on Audio, Speech, and Music Processing*, *2013*(1), 1–13. https://doi.org/10.1186/1687-4722-2013-1

Hong, Y. J., Ong, Z.-T., Lam, B., Ooi, K., Gan, W.-S., Kang, J., & Tan, S.-T. (2020). Effects of adding natural sounds to urban noises on the perceived loudness of noise and soundscape quality. *Science of the Total Environment*, *711*(1). https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.134571

International Organization for Standardization. (2014). *ISO 12913-1 Acoustics — Soundscape — Part 1: Definition and conceptual framework*. ISO.

International Organization for Standardization. (2018). *ISO/TS 12913-2:2018 Acoustics — Soundscape — Part 2: Data collection and reporting requirements*. ISO.

International Organization for Standardization. (2019). *ISO/TS 12913-3:2019 - Acoustics — Soundscape — Part 3: Data analysis*. https://www.iso.org/standard/69864.html

Jarne, C. (2018). A heuristic approach to obtain signal envelope with a simple software implementantion. *Anales de La Asociacion Fisica Argentina*, *29*(2), 51–57. https://doi.org/10.31527/analesafa.2018.29.2.51

Jayalakshmi, S. L., Chandrakala, S., & Nedunchelian, R. (2018). Global statistical features-based approach for Acoustic Event Detection. *Applied Acoustics*, *139*(February), 113–118. https://doi.org/10.1016/j.apacoust.2018.04.026

Jefferson, J., Reinders, J., & Sodani, A. (2016). Machine Learning. In *Intel Xeon Phi*

Processor High Performance Programming (pp. 527–547). https://doi.org/10.1016/B978-0-12-809194-4.00024-7

Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., & Cai, L. H. (2002). Music type classification by spectral contrast feature. *Proceedings - 2002 IEEE International Conference on Multimedia and Expo, ICME 2002*, *1*, 113–116. https://doi.org/10.1109/ICME.2002.1035731

Kadis, J. (2012). Sound. In *The Science of Sound Recording*. Focal Press. https://doi.org/10.1016/B978-0-240-82154-2.00003-X

Kang, J., Aletta, F., Gjestland, T. T., Brown, L. A., Botteldooren, D., Schulte-fortkamp, B., Lercher, P., Kamp, I. Van, Genuit, K., Luis, J., Coelho, B., Maffei, L., & Lavia, L. (2016). Ten questions on the soundscapes of the built environment. *Building and Environment*, *108*, 284–294. https://doi.org/10.1016/j.buildenv.2016.08.011

Kang, J., Aletta, F., Oberman, T., Erfanian, M., Kachlicka, M., Lionello, M., & Mitchell, A. (2019). Towards soundscape indices. *Proceedings of the 23rd International Congress on Acoustics*, *September*, 2488–2495. https://www.researchgate.net/publication/335661596_Towards_soundscape_indices

Kehtarnavaz, N. (2008). Frequency Domain Processing. In *Digital Signal Processing System Design* (pp. 175–196). Academic Press. https://doi.org/10.1016/b978-0-12-374490-6.00007-6

Kopparapu, S. K., & Laxminarayana, M. (2010). Choice of Mel filter bank in computing MFCC of a resampled speech. *10th International Conference on Information Sciences, Signal Processing and Their Applications, ISSPA 2010*, *May*, 121–124. https://doi.org/10.1109/ISSPA.2010.5605491

Kotu, V., & Deshpande, B. (2019a). Chapter 10 - Deep Learning. In V. Kotu & B. Deshpande (Eds.), *Data Science (Second Edition)* (Second Edi, pp. 307–342). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00010-1

Kotu, V., & Deshpande, B. (2019b). Chapter 10 - Deep Learning. In V. Kotu & B.

Deshpande (Eds.), *Data Science (Second Edition)* (Second Edi, pp. 307–342). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00010-1

Kurfess, F. J. (2003). Artificial Intelligence. In *A Brief History of Computing* (3rd ed., pp. 609-629,). Academic Press. https://doi.org/10.1016/B0-12-227410-5/00027-2.

Mackrill, J., Cain, R., & Jennings, P. (2013). Experiencing the hospital ward soundscape: Towards a model. *Journal of Environmental Psychology*, *36*, 1–8. https://doi.org/10.1016/j.jenvp.2013.06.004

Masalski, M., Kipiński, L., Grysiński, T., & Krêcicki, T. (2016). Hearing tests on mobile devices: Evaluation of the reference sound level by means of biological calibration. *Journal of Medical Internet Research*, *18*(5). https://doi.org/10.2196/jmir.4987

Masters, T. (1993). Practical Neural Networks Recipes in C++. In *Book*. Morgan Kaufmann. https://books.google.de/books/about/Practical_Neural_Network_Recipes_in_C++.html?id=7Ez_Pq0sp2EC&redir_esc=y

Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Features for Content-Based Audio Retrieval. In *Advances in Computers* (Vol. 78, pp. 71–150). Elsevier. https://doi.org/10.1016/s0065-2458(10)78003-7

Mushtaq, Z., & Su, S. F. (2020). Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images. *Symmetry*, *12*(11), 1–34. https://doi.org/10.3390/sym12111822

Nalini, N. J., & Palanivel, S. (2016). Music emotion recognition: The combined evidence of MFCC and residual phase. *Egyptian Informatics Journal*, *17*(1), 1–10. https://doi.org/10.1016/J.EIJ.2015.05.004

Ooi, K., Hong, J., Bhan, L., Ong, Z., & Woo-Seng, G. (2020). A deep learning approach for modelling perceptual attributes of soundscapes. *Inter Noise 2020*.

Ozcevik, A., Yuksel Can, Z., Gurbuz, H., & Poyraz Acar, I. (2014). An Applied Approach to the Examination of Urban Acoustic Comfort: The Soundscape Concept

- Statistical Analysis. *MEGARON / Yıldız Technical University, Faculty of Architecture E-Journal*, *9*(1), 45–54. https://doi.org/10.5505/megaron.2014.46338

Ozer, I., Ozer, Z., & Findik, O. (2018). Noise robust sound event classification with convolutional neural network. *Neurocomputing*, *272*, 505–512. https://doi.org/10.1016/j.neucom.2017.07.021

Parker, M. (2017). Introduction to Machine Learning. In *Digital Signal Processing 101* (pp. 347–359). Newnes. https://doi.org/10.1016/b978-0-12-811453-7.00026-3

Passricha, V., & Aggarwal, R. K. (2019). End-to-End Acoustic Modeling Using Convolutional Neural Networks. In *Intelligent Speech Signal Processing* (pp. 5–37). Academic Press. https://doi.org/10.1016/b978-0-12-818130-0.00002-7

Poznyak, T. I., Chairez Oria, I., & Poznyak, A. S. (2019). *Chapter3 - Background on dynamic neural networks* (T. I. Poznyak, I. Chairez Oria, & A. S. B. T.-O. and B. in E. E. Poznyak (eds.); pp. 57–74). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-812847-3.00012-3

Puente, S. A. (2018). *Single and Multi-Label Environmental Sound Classification Using Convolutional Neural Networks*. Chalmers University of Technology.

Puyana Romero, V., Maffei, L., Brambilla, G., & Ciaburro, G. (2016). Modelling the soundscape quality of urban waterfronts by artificial neural networks. *Applied Acoustics*, *111*, 121–128. https://doi.org/10.1016/j.apacoust.2016.04.019

Rainer, G. (1997). Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information. *Acta Acustica United with Acustica*, *83*(5), 765–774.

Rajesh, S., & Nalini, N. J. (2020). Musical instrument emotion recognition using deep recurrent neural network. *Procedia Computer Science*, *167*, 16–25. https://doi.org/10.1016/j.procs.2020.03.178

Rana, K. (2020, April 3). *Pooling Layer — Short and Simple*. Artificial Intelligence in Plain English. https://ai.plainenglish.io/pooling-layer-beginner-to-intermediate-fa0dbdce80eb

Rebai, I., Benayed, Y., Mahdi, W., & Lorré, J. P. (2017). Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, *112*, 316–322. https://doi.org/10.1016/j.procs.2017.08.003

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Salamon, J., & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381

Schafer, R. M. (1969). *The new soundscape : a handbook for the modern music teacher* (p. 65 p.). https://doi.org/papers3://publication/uuid/28275828-EB29-4C2E-999B-92566728276B

Schafer, R. M. (1977). *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books.

Schafer, R. M., Truax, B., Colin, M., Davis, B., Huse, P., & Broomfield, H. (1973). *The Vancouver Soundscape*. Simon Fraser University.

Scikit-Learn. (n.d.). *sklearn.preprocessing.StandardScaler — scikit-learn 0.23.2 documentation*. Retrieved December 19, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Sharan, R. V., & Moir, T. J. (2015). Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM. *Neurocomputing*, *158*, 90–99. https://doi.org/10.1016/j.neucom.2015.02.001

Sharan, R. V., & Moir, T. J. (2016). An overview of applications and advancements in automatic sound recognition. *Neurocomputing*, *200*, 22–34. https://doi.org/10.1016/j.neucom.2016.03.020

Sharan, R. V, & Moir, T. J. (2017). Robust acoustic event classification using deep neural networks. *Information Sciences*, *396*, 24–32. https://doi.org/10.1016/j.ins.2017.02.013

Song, X., Wu, Z., Huang, Y., Su, D., & Meng, H. (2020). SpecSwap: A simple data augmentation method for end-to-end speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-Octob*, 581–585. https://doi.org/10.21437/Interspeech.2020-2275

Sound samples | Philharmonia. (n.d.). Retrieved December 19, 2020, from https://philharmonia.co.uk/resources/sound-samples/

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 345–350. https://doi.org/10.1016/0370-2693(93)90272-J

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190. https://doi.org/10.1121/1.1915893

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, *17*(10), 1733–1746. https://doi.org/10.1109/TMM.2015.2428998

Suk, H.-I. (2017). An Introduction to Neural Networks and Deep Learning. *Deep Learning for Medical Image Analysis*, 3–24. https://doi.org/10.1016/B978-0-12-810408-8.00002-X

Sun, K., De Coensel, B., Filipan, K., Aletta, F., Van Renterghem, T., De Pessemier, T., Joseph, W., & Botteldooren, D. (2019). Classification of soundscapes of urban public open spaces. *Landscape and Urban Planning*, *189*, 139–155. https://doi.org/10.1016/j.landurbplan.2019.04.016

Sun, K., Filipan, K., Aletta, F., & Renterghem, T. Van. (2019). Classifying urban public spaces according to their soundscape. *23rd International Congress on Acoustics, September*, 6100–6105.

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. (2017). Efficient Processing of Deep

Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, *105*(12), 2295–2329. http://arxiv.org/abs/1703.09039

The World Soundscape Project. (1978). *The Handbook for Acoustic ecology* (B. Truax (Ed.)). https://www.sfu.ca/~truax/handbook2.html

Tone. (2009). Encyclopedia Britanica. https://www.britannica.com/science/tone-sound

Torresin, S., Albatici, R., Aletta, F., Babich, F., Oberman, T., Siboni, S., & Kang, J. (2020). Indoor soundscape assessment: A principal components model of acoustic perception in residential buildings. *Building and Environment*, *182*, 107152. https://doi.org/10.1016/j.buildenv.2020.107152

Truax, B. (1984). Acoustic communication. In *Communication and information science*.

Umesh, S., Cohen, L., & Nelson, D. (1999). Fitting the Mel scale. *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *1*, 217–220. https://doi.org/10.1109/icassp.1999.758101

Westerkamp, H., Woog, P., A., Kallmann, H., & Truax, B. (2006). *World Soundscape Project*. http://www.thecanadianencyclopedia.com/en/article/world-soundscape-project/,

Wittek, P. (2014). Machine Learning. In *Quantum Machine Learning: What Quantum Computing Means to Data Mining* (pp. 11–24). Academic Press. https://doi.org/10.1016/B978-0-12-800953-6.00002-5

Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, *3*(3), 27–36. https://doi.org/10.1109/93.556537

Yilmazer, S., & Acun, V. (2018a). A Structural Equation Modelling Approach for Indoor Soundscape: Adaptive re-use in Çengelhan Caravanserai. *AESOP 18*.

Yilmazer, S., & Acun, V. (2018b). A grounded theory approach to assess indoor soundscape in historic religious spaces of Anatolian culture: A case study on Hacı Bayram Mosque. *Building Acoustics*. https://doi.org/10.1177/1351010X18763915

Yu, L., & Kang, J. (2008). Effects of social, demographical and behavioral factors on the sound level evaluation in urban open spaces. *The Journal of the Acoustical Society of America*, *123*(2), 772–783. https://doi.org/10.1121/1.2821955

Yu, L., & Kang, J. (2009). Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach. *The Journal of the Acoustical Society of America*, *126*(3), 1163–1174. https://doi.org/10.1121/1.3183377

APPENDICES

APPENDIX A

(Ethics Committee Approval)

Bilkent University
Office of the Provost

November 17, 2020

To Whom It May Concern:

By the attached letter, which is in Turkish, the following research has received the approval of the Ethics Committee for Research Projects Involving Human Participants (Internal Review Board of Bilkent University). The investigators may collect data starting the date of the attach letter (2020_10_23_01).

Research Project Title: *Soundscape Evaluation Through Artificial Intelligence*

Names of the Principal Investigators: Semiha Yılmazer, Volkan Acun

Sincerely,

Professor H. Altay Güvenir
Chair
Ethics Committee for Research Projects Involving Human Participants

Bilkent University, TR-06800 Bilkent, Ankara, Turkey • Phone: +90-312-290-1213 • +90-312-290-1214 • Fax: +90-312-266-4152

APPENDIX B

Appendix A1: The framework showing the content of the videos used in Video Group 1.

Video Group 01		
#	**Recording Name**	**Contents**
1	R1_Louvre_00_30_01_01.mp4	Large, Loud, Human, Echo, Ambient
2	R24_Scotland_02_2_02_50.mp4	Moderate, Large, Semi-loud, Human, Children, Speech
3	R11_London51_10_51_40.mp4	Loud, Children, Rhythm, Human, Speech
4	R5_Louve_06_20_06_52.mp4.mp4	Large, Semi-loud, Echo, Human
5	R22_BRIT_07_30_07_30.mp4	Moderate Footsteps, Human
6	R25_Kröller_00_07_00_43.mp4	Quiet, Footsteps, Speech
7	R33_Kröller_35_07_35_50.mp4	Quiet, Human, Speech, Footsteps, Electronic
8	R40_NaturalHist_08_20_08_52.mp4	Loud, Thematic, Speech, Mix
9	R8_London_06_20_06_50.mp4	Loud, Echo, Speech, Children
10	R28_Kröller_04_55_05_30.mp4	Quiet, Dining area, Human
11	R38_Wesserburg_18_15_18_46.mp4	Quiet, Thematic, Footsteps, Echo
12	R29_Kröller_14_40_15_10.mp4	Quiet, Footsteps, Speech, Thematic
13	R42_NaturalHist_14_40_16_12.mp4	Semi-loud, Speech, Water, Thematic, Human, Electronic
14	R26_Kröller_01_05_00_57.mp4	Quiet, Footsteps, Speech,
15	R4_Louve_05_32_06_02.mp4.mp4	Moderate, Footsteps, Human, Speech
16	R15_Maasricht_11_35_12_05.mp4	Moderate, Music, Footsteps, Speech
17	R32_Kröller_30_37_31_10.mp4	Quiet, Thematic, Footsteps, Music
18	R30_Kröller_28_37_29_10.mp4	Moderate, Thematic, Footsteps, Music
19	R10_London50_00_51_10.mp4	Loud, Thematic, Bass, Children, Human
20	R6_Louve_07_20_07_50.mp4	Large, Semi-Loud, Human, Echo, Speech

Appendix A2: The framework showing the content of the videos used in Video Group 2.

Video Group 02		
#	**Recording Name**	**Contents**
1	R2_Louve_01_20_01_50.mp4	Loud, Large, Human, Echo
2	R22_BRIT_07_30_07_30.mp4	Loud, Human, Footsteps
3	R36_MarineLibrary_08_44_09_18.mp4	Moderate, Footsteps, impact, echo, speech
4	R34_Kröller_40_07_30_50.mp4	Quiet, Speech, Kitchenware, footsteps,
5	R18_Louvre_52_00_52_30.mp4	Moderate, Speech, Background noise,
6	R16_Maasricht_12_55_13_25.mp4	Quiet, Speech, Echo, Music
7	R26_Kröller_01_05_00_57.mp4	Quiet, Footsteps, Speech,
8	R31_Kröller_29_10_29_42.mp4	Quiet, Thematic Sounds, Footsteps, Electronic Sounds
9	R3_Louve_02_00_02_30.mp4.mp4	Loud, Large Space, Echo, Background Noise
10	R39_Wesserburg_20_04_20_36.mp4	Moderate, Footsteps, Electronic Sound, Thematic Sounds
11	R43_Scotland_26_04_26_35.mp4	Quiet, Speech, Electronic, Announcements, Signal Sounds
12	R41_NaturalHist_10_20_10_52.mp4	Loud, Music, Children, Footsteps, Running
13	R47_Kröller_26_58_27_35.mp4	Quiet, Footsteps, Music
14	R7_London_05_30_06_00.mp4	Loud, Speech, Children, Background Noise
15	R9_London_08_00_08_30.mp4	Loud, Background Noise, Children, Speech, Yelling
16	R13_London_53_00_53_30.mp4	Moderate, Thematic Sounds, Speech, Background Noise
17	R23_Scotland_01_21_01_51.mp4	Moderate, Large Space, Children, Echo
18	R46_Kröller_10_20_10_55.mp4	Quiet, Speech, Footsteps, Cash Register, Environmental Sounds
19	R44_Scotland_27_19_27_57.mp4	Quiet, Speech, Environmental Sounds, Thematic Sounds
20	R45_Scotland_32_04_32_41.mp4	Quiet, Children, Speech, Environmental Sounds

Appendix A3: The framework showing the content of the videos used in Video Group 3.

Video Group 03		
#	**Recording Name**	**Contents**
1	R6_Louve_07_20_07_50.mp4	Loud, Large, Echo, Speech, Background Noise,
2	R12_London_52_12_52_52.mp4	Loud, Background Noise, Thematic Sounds, Speech
3	R53_Bookshop_00_57_01_27.mp4	Quiet, Music, Footsteps, Environmental Sounds
4	R60_NATUR_03_24_03_54.mp4	Moderate, Thematic Sounds, Music, Electronic
5	R59_Scotland_23_11_23_41.mp4	Quiet, Speech, Bells, Environmental Sounds
6	R27_Kröller_04_20_04_53.mp4	Quiet, Footsteps, Speech, Kitchenware
7	R49_V&A_40_15_40_48.mp4	Moderate, Speech, Echo, Footsteps, Environmental Sounds
8	R37_Wesserburg_12_24_12_56.mp4	Moderate, Electronic Sounds, Footsteps, Background Noise
9	R50_V&A_52_01_52_31.mp4	Moderate, Footsteps, Signal Sound, Speech, Alarm
10	R54_Bookshop_01_08_40_01_09_10.mp4	Quiet, Music, Speech, Footsteps,
11	R51_V&A_53_22_53_54.mp4	Loud, Footsteps, Speech, Announcement
12	R55_Kröller_11_10_11_40.mp4	Quiet, Speech, Footsteps, Environmental Sounds
13	R52_V&A_1_14_52_1_15_23.mp4	Loud, Children, Background Noise, Speech
14	R56_Scotland_07_18_07_48.mp4	Moderate, Music, Thematic Sounds, Background Noise
15	R61_NATUR_10_01_10_31.mp4	Moderate, Thematic Sounds, Children, Running, Music
16	R62_NATUR_13_17_13_47.mp4	Loud, Speech, Ventilation, Background Noise
17	R48_V&A_33_35_34_05.mp4	Quiet, Speech, Footsteps, Environmental Sounds
18	R58_Scotland_22_03_22_33.mp4	Moderate, Thematic Sounds, Electronic Sounds, Background Noise
19	R57_Scotland_11_35_12_05.mp4	Quiet, Music, Speech, Background Noise
20	R63_NATUR_05_40_06_10.mp4	Loud, Thematic Sounds, Background Noise, Children, Speech

APPENDIX C

(Demographic Information Questionnaire)

DEMOGRAPHIC INFORMATION:

1. Your age?

[ ] Male [ ] Female

2. Your gender?

[ ] Between 18 - 24 [ ] Between 24 – 34 [ ] Between 35 - 44

[ ] Between 44 – 54 [ ] Between 55 – 64

3. In which country you were born?

_____

4. Which language do you feel most comfortable speak?

_____

5. What is the highest degree or level of school you have completed?

[ ] Less than a high school diploma [ ] High school diploma or equivalent
[ ] Bachelor's degree [ ] Master's degree
[ ] Doctorate

6. Your employment status?

[ ] Employed full-time [ ] Employed part-time
[ ] Unemployed [ ] Retired
[ ] Student

7. What kind of settlement area are you currently living in?

[ ] Urban [ ] Suburban [ ]Rural

8. What kind of settlement area you spent the majority of your life in?

[ ] Urban [ ] Suburban [ ]Rural

9. Do you have any hearing problems that you are aware of?

[ ] Yes [ ] No

APPENDIX D

(QR Codes)

Appendix D1: QR code for one of the videos that is included in the soundscape perception survey.

Appendix D2: QR code for the soundscape perception survey questionnaire.

APPENDIX E

(Python Codes for the CNN and FFNN models)

Appendix E1: The code used for the feature extraction and development of the Convolutional Neural Network based classification model based on https://github.com/micah5/pyAudioClassification.

```
from __future__ import absolute_import
import os
from pyaudioclassification.feat_extract import parse_audio_files,
parse_audio_file
import numpy as np
import pyaudioclassification.models
from keras.utils import to_categorical
from keras.optimizers import SGD

def feature_extraction(data_path):
    r = os.listdir(data_path)
    r.sort()
    features, labels = parse_audio_files(data_path, r)
    return features, labels

def train(features, labels, type='cnn', num_classes=None,
print_summary=False,
    save_model=False, lr=0.01, loss_type=None, epochs=50,
optimizer='SGD', verbose=True):
    labels = labels.ravel()
    if num_classes == None: num_classes = np.max(labels, axis=0)

    model = getattr(models, type)(num_classes)
    if print_summary == True: model.summary()

    if loss_type == None:
        loss_type = 'binary' if num_classes <= 2 else 'categorical'
    model.compile(optimizer=SGD(lr=lr),
                  loss='%s_crossentropy' % loss_type,
                  metrics=['accuracy'])

    if loss_type == 'categorical':
        y = to_categorical(labels - 1, num_classes=num_classes)
    else:
        y = labels - 1

    x = np.expand_dims(features, axis=2)

    model.fit(x, y, batch_size=64, epochs=epochs, verbose=verbose)

    return model

def predict(model, data_path):
    x_data = parse_audio_file(data_path)
    X_train = np.expand_dims(x_data, axis=2)
    pred = model.predict(X_train)
    return pred
```

```
def print_leaderboard(pred, data_path):
    r = os.listdir(data_path)
    r.sort()
    sorted = np.argsort(pred, axis=1)
    count = 0
    for index in (-pred).argsort(axis=1)[0]:
        print('%d.' % (count + 1), r[index + 1],
str(round(pred[0][index]*100)) + '%', '(index %s)' % index)
        count += 1



import glob
import os
import librosa
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.pyplot import specgram
import soundfile as sf
from tqdm import tqdm

def extract_feature(file_name):
    """Generates feature input (mfccs, chroma, mel, contrast, tonnetz).
    -*- author: mtobeiyf https://github.com/mtobeiyf/audio-
classification -*-
    """
    X, sample_rate = sf.read(file_name, dtype='float32')
    if X.ndim > 1:
        X = X[:,0]
    X = X.T
    X = np.asfortranarray(X)
    stft = np.abs(librosa.stft(X))
    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate,
n_mfcc=40).T,axis=0)
    chroma = np.mean(librosa.feature.chroma_stft(S=stft,
sr=sample_rate).T,axis=0)
    mel = np.mean(librosa.feature.melspectrogram(X,
sr=sample_rate).T,axis=0)
    contrast = np.mean(librosa.feature.spectral_contrast(S=stft,
sr=sample_rate).T,axis=0)
    tonnetz =
np.mean(librosa.feature.tonnetz(y=librosa.effects.harmonic(X),
sr=sample_rate).T,axis=0)
    return mfccs, chroma, mel, contrast, tonnetz

def parse_audio_files(parent_dir, sub_dirs, file_ext=None,
verbose=True):
    # by default test for only these types
    if file_ext == None:
        file_types = ['*.ogg', '*.wav']
    else:
        file_types = []
        file_types.push(file_ext)
    features, labels = np.empty((0,193)), np.empty(0)
    for label, sub_dir in enumerate(sub_dirs):
        for file_ext in file_types:
            # file names
```

```
            iter = glob.glob(os.path.join(parent_dir, sub_dir,
file_ext))
            if len(iter) > 0:
                if verbose: print('Reading', os.path.join(parent_dir,
sub_dir, file_ext), '...')
                for fn in tqdm(iter):
                    ext_features = get_ext_features(fn)
                    if type(ext_features) is np.ndarray:
                        features = np.vstack([features, ext_features])
                        labels = np.append(labels, label)
    return np.array(features), np.array(labels, dtype = np.int)

def get_ext_features(fn):
    try:
        mfccs, chroma, mel, contrast, tonnetz = extract_feature(fn)
        ext_features = np.hstack([mfccs, chroma, mel, contrast,
tonnetz])
        return ext_features
    except Exception as e:
        print("[Error] extract feature error. %s" % (e))
        return None

def parse_audio_file(fn):
    features = np.empty((0,193))
    ext_features = get_ext_features(fn)
    features = np.vstack([features,ext_features])
    return np.array(features)
```

```
import numpy as np
from keras.models import Sequential
from keras.layers import Dense, Dropout, Activation
from keras.layers import Conv2D, BatchNormalization, MaxPooling2D

def cnn(num_classes):
    from keras.layers import Embedding
    from keras.layers import Conv1D, GlobalAveragePooling1D,
MaxPooling1D, Flatten

    activation = 'softmax' if num_classes > 2 else 'sigmoid'
    model = Sequential()
    model.add(Conv1D(64, 3, input_shape=(193, 1)))
    model.add(Activation('relu'))
    model.add(Conv1D(64, 3))
    model.add(Activation('relu'))
    model.add(MaxPooling1D(3))
    model.add(Conv1D(128, 3))
    model.add(Activation('relu'))
    model.add(Conv1D(128, 3))
    model.add(Activation('relu'))
    model.add(GlobalAveragePooling1D())
    model.add(Dropout(0.5))
    model.add(Dense(7, activation='softmax'))
    model.summary()
```

```
    model.compile(loss='categorical_crossentropy',
                  optimizer='SGD',
                  metrics=['acc'])

    return model




def cnn2d(num_classes):
    from keras.layers import Conv2D, BatchNormalization, MaxPooling2D
    model = Sequential()
    # Conv Layer #1
    model.add(Conv2D(8, (3, 3), padding='same', input_shape=(193, 193,
1)))
    model.add(Activation('relu'))
    model.add(MaxPooling2D(2,2))
    model.add(BatchNormalization(axis=1))
    # Conv Layer #2
    model.add(Conv2D(16, (3, 3), padding='same'))
    model.add(Activation('relu'))
    model.add(MaxPooling2D(3,3))
    # Conv Layer #3
    model.add(Conv2D(32, (3, 3), padding='same'))
    model.add(Activation('relu'))
    model.add(MaxPooling2D(3,3))
    # Conv Layer #4
    model.add(Conv2D(32, (3, 3), padding='same'))
    model.add(Activation('relu'))
    model.add(MaxPooling2D(3,3))
    # Flatten
    model.add(Flatten())
    model.add(Dense(64))
    model.add(Activation('relu'))
    model.add(Dropout(0.5))
    model.add(Dense(12))
    model.add(Activation('sigmoid'))
    model.compile(loss='categorical_crossentropy',
                  optimizer='adam',
                  metrics=['acc'])

    model.summary()

    return model
###  make a prediction

import numpy as np
from pyaudioclassification import feature_extraction, train, predict,
print_leaderboard

parent_dir = 'e:\Machine Learning\Pro_03_env'
#%%
# step 1: preprocessing
if np.DataSource().exists("./feat_G04_FINAL.npy") and
np.DataSource().exists("./label_G04_FINAL.npy"):
```

```
    features, labels = np.load('./feat_G04_FINAL.npy'),
np.load('./label_G04_FINAL.npy')
else:
    features, labels = feature_extraction('./data/')
    np.save('./feat_G04_FINAL.npy', features)
    np.save('./label_G04_FINAL.npy', labels)

# step 2: training
if np.DataSource().exists("./model_G04_FINAL.h5"):
    from tensorflow.keras.models import load_model
    model = load_model('./model_G04_FINAL.h5')
else:
    model = train(features, labels, type='cnn', print_summary=True,
lr=0.03, loss_type='categorical',
              epochs=200, optimizer='SGD')
    model.save('./model_G04_FINAL.h5', include_optimizer=False)

# step 3: prediction
pred = predict(model,
'./Prediction/Survey_G03_Pred/R56_Scotland_07_18_07_48.wav')
print_leaderboard(pred, './data/')
```

Appendix E2: The code used for the Feedforward Neural Network.

```
####### THE FFNN FINAL WORKING MODEL #########
# plots accuracy and loss graphs
import tensorflow as tf
from tensorflow import keras
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score

# Load the data
dataset = pd.read_csv('THESIS_DATA_MASTER.csv')
print(len(dataset))
print(dataset.head())

####PREPROCESS#####

dataset.dtypes
dataset['Overall'] = dataset.Flute.astype(int)
dataset['Overall'] = dataset['Overall'].fillna(0).astype(np.int64)
dataset.dtypes
#####

# Define the features and the output
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values


#Label Encoder or OneHotEncoder (Used for country and languages here)
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(),
[9,10])], remainder='passthrough')
X = np.array(ct.fit_transform(X))


# Train Test split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=0)

# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)

# Check the shape
y_train.head()
print(x_train.shape)
print(y_train.shape)
```

```
##### MODEL #####

model = tf.keras.models.Sequential()
# input layer
model.add(tf.keras.layers.Dense(units=17, kernel_initializer='uniform',
activation='tanh'))
# 1st hidden layer
model.add(tf.keras.layers.Dense(units=10, activation='ReLu'))
 # 1st hidden layer
model.add(tf.keras.layers.Dense(units=10, activation='ReLu'))
# 3rd hidden layer
model.add(tf.keras.layers.Dense(units=10, activation='ReLu'))

# Output layer
model.add(tf.keras.layers.Dense(6, activation='softmax'))

#adam = tf.keras.optimizers.Adam(lr=0.01)
sgd = tf.keras.optimizers.SGD(lr=0.01, momentum=0.9)
model.compile(loss='sparse_categorical_crossentropy',optimizer='sgd',me
trics=['accuracy'])
history = model.fit(X_train,y_train,batch_size=80, epochs=150,
validation_data=(X_test, y_test))
model.summary()


####### METRICS PLOTS #######
print(history.history.keys())
# Get training and test loss histories
training_loss = history.history['loss']
test_loss = history.history['val_loss']
# Create count of the number of epochs
epoch_count = range(1, len(training_loss) + 1)
# Visualize loss history
plt.plot(epoch_count, training_loss, 'r--')
plt.plot(epoch_count, test_loss, 'b-')
plt.legend(['Training Loss', 'Test Loss'])
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.show();


# Plot Accuracy
plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('Model Accuracy (150 epoch, 10 nodes)')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='lower right')
plt.show()

# Visualize Loss 2.0
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('Model Loss (150 epoch, 10 nodes))')
plt.ylabel('loss')
```

```
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper right')
plt.show()
```

Appendix E3: The code used for the k-Nearest Neighbors algorithm

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
from sklearn.metrics import roc_auc_score

dataset = pd.read_csv('THESIS_DATA_MASTER.csv')
print(len(dataset))
print(dataset.head())
#dataset  = dataset.fillna(1)



# Converts data types
dataset.dtypes # check the data types
dataset['Overal'] = dataset['Overal'].fillna(0).astype(np.int64)
dataset.dtypes

#split dataset
# Define the features and the output
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values


#Label Encoder or OneHotEncoder (Used for country and languages here)
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(),
[12,13])], remainder='passthrough')
X = np.array(ct.fit_transform(X))

# Train test split
X_train, X_test, y_train, y_test = train_test_split(X, y,
random_state=5, test_size=0.2)


# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)

#to find the number of K
import math
math.sqrt(len(y_test))

# Neighbourhood Compotent Analysis
from sklearn.neighbors import NeighborhoodComponentsAnalysis
nca = NeighborhoodComponentsAnalysis(random_state=42)
```

```
nca.fit(X_train, y_train)

# Define the model: Init K-NN
classifier = KNeighborsClassifier(algorithm='auto', leaf_size=30,
metric='minkowski',
            metric_params=None, n_jobs=1, n_neighbors=10, p=2,
weights='distance')

# Fit the Model
classifier.fit(X_train, y_train)
print(classifier.score(X_test, y_test))

classifier.fit(nca.transform(X_train), y_train)
print(classifier.score(nca.transform(X_test), y_test))
# Predict the test set results
y_pred = classifier.predict(X_test)
y_pred

#Evaluate Model
cm = confusion_matrix(y_test, y_pred)
print (cm)
print(f1_score(y_test, y_pred, average='micro'))
print(accuracy_score(y_test,y_pred))
print(roc_auc_score(y_test, y_pred))

#Evalute in cross validation
scores = cross_val_score(classifier, X, y, cv=10, scoring='accuracy', )
print(scores)
print(scores.mean())



#### Estimate optiumum parameters ####
dataset_model = dataset.copy()
#List Hyperparameters to tune
leaf_size = list(range(1,50))
n_neighbors = list(range(1,30))
p=[1,2]
#convert to dictionary
hyperparameters = dict(leaf_size=leaf_size, n_neighbors=n_neighbors,
p=p)
#Making the model
clf = GridSearchCV(classifier, hyperparameters, cv=10)
best_model = clf.fit(X_train,y_train)
#Best Hyperparameters Value
print('Best leaf_size:',
best_model.best_estimator_.get_params()['leaf_size'])
print('Best p:', best_model.best_estimator_.get_params()['p'])
print('Best n_neighbors:',
best_model.best_estimator_.get_params()['n_neighbors'])
#Predict testing set
y_pred = best_model.predict(X_test)
#Check performance based on accuracy
print(accuracy_score(y_test, y_pred))
#Check performance based on F1 score
(f1_score(y_test, y_pred, average='micro'))
```