FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

A NEW SOUND FORENSICS APPROXIMATION: AN AUTOMATED LOCATION DETECTION METHOD İN MULTI-STOREY BUILDINGS USING ENVIRONMENTAL SOUND CLASSIFICATION

Mark Ndowobe OKABA

Master's Thesis

DEPARTMENT OF DIGITAL FORENSIC

ENGINEERING

JULY 2022

FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

Department of Digital Forensic Engineering

Master's Thesis

A NEW SOUND FORENSICS APPROXIMATION: AN AUTOMATED LOCATION DETECTION METHOD İN MULTI-STOREY BUILDINGS USING ENVIRONMENTAL SOUND CLASSIFICATION

Author
Mark Ndowobe OKABA

Supervisor
Assoc. Prof. Türker TUNCER

JULY 2022
ELAZIG

FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
TÜRKİYE

Department of Digital Forensic Engineering

Master's Thesis

Title:	A New Sound Forensics Approximation: An Automated Location Detection Method İn Multi-Storey Buildings Using Environmental Sound Classification
Author:	Mark Ndowobe OKABA
Submission Date:	06 June 2022
Defense Date:	04 July 2022

THESIS APPROVAL

This thesis, which was prepared according to the thesis writing rules of the Graduate School of Natural and Applied Sciences, Fırat University, was evaluated by the committee members who have signed the following signatures and was unanimously approved after the defense exam made open to the academic audience.

Signature

Supervisor:	Assoc. Prof. Türker TUNCER Firat University, Faculty of Technology	Approved
Chair:	Assoc. Prof. Dr. Şengül DOĞAN Firat University, Faculty of Technology	Approved
Member:	Asist. Prof. Dr. Mehmet BAYĞIN Ardahan University, Faculty of Engineering	Approved

This thesis was approved by the Administrative Board of the Graduate School on

..... / ..... / 20 .....

Signature

Prof. Dr. Kürşat Esat ALYAMAÇ
Director of the Graduate School

DECLARATION

I hereby declare that I wrote this Master's Thesis titled " A New Sound Forensics Approximation: An Automated Location Detection Method İn Multi-Storey Buildings Using Environmental Sound Classification " in consistent with the thesis writing guide of the Graduate School of Natural and Applied Sciences, Firat University. I also declare that all information in it is correct, that I acted according to scientific ethics in producing and presenting the findings, cited all the references I used, express all institutions or organizations or persons who supported the thesis financially. I have never used the data and information I provide here in order to get a degree in any way.

04 July 2022

Mark Ndowobe OKABA

PREFACE

Environmental sound classification is a fascinating field of study that has piqued the curiosity of many researchers. However, sound-based position recognition has made little progress since navigation only returns longitude and latitude. In this paper, we took a detailed look at solving this problem and its consequent application to the field of Digital Forensics. Traditional ambient sound classification algorithms rely on hand-crafted audio properties like MFCC, zero-crossing rate, and short-term energy, among others. These methods only consider one or more components of ambient sound and do not completely account for its diversity and complexity. Deep learning handcrafted model were implemented using a new center symmetric CS-Lblock-Pattern that utilizes S-boxes and hence helped in feature generation. The potential for this method is immense and although the study was not without limitation, we intend to collate more dataset to test model and hopefully deploy it for practical application in the field.

I want to specially thank my supervisor; **Associate Professor Türker TUNCER** for his guidance and effort in making the research a success

Mark Ndowobe OKABA
ELAZIG, 2022

TABLE OF CONTENTS

ABSTRACT

A New Sound Forensics Approximation: An Automated Location Detection Method İn Multi-Storey Buildings Using Environmental Sound Classification

Mark Ndowobe OKABA

Master's Thesis

FIRAT UNIVERSITY
Graduate School of Natural and Applied Sciences

Department of Digital Forensic Engineering

July 2022, Page: x + 35

Environmental sound classification has continued to attract research and applications designed specifically for digital forensic studies and cybercrime analysis because of its vast range of applications. [1]. Conventional ambient noise classification systems, for example, depend on hand-crafted extracted features such as MFCC, zero-crossing rate, and short-term energy. However, this might not help for position-based sound classification such as those in multi-storey buildings. This thesis is a research project aimed at resolving this issue using a centre-symmetric nonlinear pattern. We gathered ambient sounds from each floor of a multi-story structure, which we turned into a dataset and made publicly available. A substitution box (S-Box) is used to generate features in a new centre symmetric nonlinear pattern. After the feature generation technique we used, we dubbed the model CS-Lblock-Pattern. The dataset contains ten classes, one for each of the building's ten floors. Our investigations show that using SVM, we can get good classification accuracy for multi-storey hospital datasets, with an accuracy rate of 95.38%. The classification phase results clearly shows that ESC may produce effective results by applying a center-symmetric nonlinear pattern

Keywords: 3D spaces; environmental sound classification; iterative neighborhood component analysis; center symmetric Lblock pattern; statistical feature extraction

ÖZET

Yeni Bir Ses Adli Bilişim Yaklaşımı: Çevresel Ses Sınıflandırması Kullanan Çok Katlı Binalarda Otomatik Konum Algılama Yöntemi

Mark Ndowobe OKABA

Yüksek Lisans Tezi

FIRAT ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Adli Bilişim Mühendisliği Anabilim Dalı

Temmuz 2022, Sayfa: x + 35

Çevresel ses sınıflandırması, geniş uygulama yelpazesi nedeniyle dijital adli tıp çalışmaları ve siber suç analizi için özel olarak tasarlanmış araştırma ve uygulamaları çekmeye devam etmiştir. [1]. Örneğin, geleneksel ortam gürültüsü sınıflandırma sistemleri, MFCC, sıfır geçiş hızı ve kısa süreli enerji gibi el yapımı çıkarılmış özelliklere bağlıdır. Ancak bu, çok katlı binalardakiler gibi konuma dayalı ses sınıflandırması için yardımcı olmayabilir. Bu tez, bu konuyu merkez-simetrik doğrusal olmayan bir desen kullanarak çözmeyi amaçlayan bir araştırma projesidir. Çok katlı bir yapının her katından ortam seslerini topladık ve bunları bir veri kümesine dönüştürdük ve herkese açık hale getirdik. Yeni bir merkez simetrik doğrusal olmayan desende özellikler oluşturmak için bir ikame kutusu (S-Box) kullanılır. Kullandığımız özellik oluşturma tekniğinden sonra CS-Lblock-Pattern modeline isim verdik. Veri kümesi, binanın on katının her biri için bir tane olmak üzere on sınıf içerir. Araştırmalarımız, SVM'yi kullanarak, çok katlı hastane veri kümeleri için% 95,38 doğruluk oranıyla iyi bir sınıflandırma doğruluğu elde edebileceğimizi göstermektedir. Sınıflandırma fazı sonuçları, ESC'nin merkez-simetrik doğrusal olmayan bir desen uygulayarak etkili sonuçlar üretebileceğini açıkça göstermektedir

Anahtar Kelimeler: 3D konum tespiti; çevresel ses sınıflandırması; merkez simetrik Lblock deseni; istatistiksel özellik çıkarma; yinelemeli komşuluk bileşen analizi

LIST OF FIGURES

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVIATIONS

Abbreviations

dB	: Decibels
DNN	: Deep Neural Network
DWT	: Discrete Wave Transform
ESC	: Environmental Sound Classification
GMM	: Gaussian Mixture Model
GTSC	: Gammatone Spectral Coefficients
INCA	: Iterative Neighborhood Component
kNN	: k-Nearest Neighbour
L3	: Look, Listen and Learn
LD	: Linear Discriminant
MFCC	: Mel-frequency Cepstral Coefficient
MLP	: Multi-Layer Perceptron
RBM	: Restricted Boltzmann Machine
STFT	: Short Term Fourier Transform
SVM	: Support Vector Machine
TQWT	: Tunable Q-Wavelet Transform

1. INTRODUCTION

Machine learning models have been used, in recent years to classify environmental or ambient sound. This has led to significant progress in this field with relevant applications that includes music genre recommendations and alarm systems. Currently, there is a growing interest in applying deep learning approaches to classify environmental sounds, and advances in the image classification field are encouraging academics to use spectrogram images to classify sounds.

Music, unlike speech, has a strong structure and clearly defined sections, whereas speech and noises lack such frameworks. Because of this lack of organization, and because humans lack the processing capabilities, the human hearing system cannot adequately describe ambient noise. Advancement in this topic has been fueled by interest in developing deep learning models for autonomous classification of environmental sound. Because of increases in computing power using GPUs, these elegant solutions use neural networks that can handle a large quantity of data and complicated aspects. [2]. Converting audio sound to images and then using a neural network to analyze the spectrograms is the most frequent deep learning approach for environmental sound classification. Typically, they are supervised learning methods. Piczak's [3] research concludes that convolutional neural networks have the best spectrogram analysis accuracy rates. Features could be extracted from the audio stream using methods like Short-Term Fourier Transform (STFT), Discrete Wave Transform (DWT) and feature selection optimized for classification. Basic classifiers like KNNs, SVMs, and Random Forests have been shown to be less accurate than deep networks. The purpose of this study is to offer a novel model for environmental sound classification. Digital audio forensics, public safety, smart homes and cities, construction automation, location detection, and information security are just a few of the fascinating possibilities for this model.

A multilevel feature generating network using a combination of CS-Lblock-Pattern and Tunable Q-Wavelet Transform (TQWT) is demonstrated in this thesis, [4] as well as a statistical method for signal decomposition and feature extraction. Twenty decomposed signals were produced from each signal and this method is unique in that it uses CS-Lblock-Pattern to create a multilayer feature generating network. At each level, 300 features were created using primary textural feature extraction and statistical moments in the pattern. When this network is applied to the 20 deconstructed sounds and one raw environmental sound, it yields a total of 6300 characteristics. Iterative Neighborhood Component Analysis was used to pick features from the retrieved features (INCA) [5]. This is an advanced and highly optimized method of component analysis. We used LD and SVM [6] which are two shallow classifiers from the MATLAB Classification Learner Toolbox, in the classification step.

1.1. Problem Statement

Environmental sound classification (ESC) research, which focuses on detecting specific sound events such as dog barking, gunshots, and air conditioner sounds, has gotten a lot of attention in recent years. However, as earlier stated, classification of sound using nonlinear patterns has very few or no research. We intend to provide more research in this area to aid forensics studies such as location detection.

To attain our aims, a different number of research questions will be answered:

- How do we do ambient noise classification in digital forensic?
- Can audio clips from different floors in a multi-storey building be used to detect floors?
- How do varied degrees of noise in the input signals influence the trained neural network's prediction accuracy?
- How does the prediction accuracy of the trained neural network differ depending on the level of disparities in the input signals?

1.2. Purpose of study

The goal of this study is to use automatic classification of environmental noises to enable position detection using a center-symmetric nonlinear patter. Certain acoustic features associated with buildings can be used to distinguish residential multi-story structures. We were able to capture sound data from a multi-story hospital building in Elazig, Turkey, and have made it available to the public. In Turkey, hospital buildings have identical architectural structure, making it difficult to distinguish between residential and nonresidential areas. As a result, we provide an ESC model based on human factors. Due to the scarcity of datasets in this subject, the obtained data may be useful in furthering study in this area.

As previously said, the applications of ESC are highly diverse and intriguing. In forensic investigations, environmental sound classification is critical for criminal tracking, and this is a challenge for forensic scientist or investigators. We wish to demonstrate S-boxes' capacity to generate features by using microstructures, which are essential hand-crafted feature extractors. Linear patterns are commonly used in these microstructures. However, we want to test performance in a nonlinear pattern that employs the lightweight block ciphers' S-boxes.

1.3. Uniqueness of our work

The following are the uniqueness and contributions of our work:

- A new dataset was collected and made public to add to the pool of sound classification datasets available.

- To enable environmental sound classification, we provide a model based on the lightweight block cipher's S-boxes.
- We take advantage of the acoustic features of multi-story residential structures such as hospitals and courthouses.
- We used multilevel feature generation networks to extract features, and because deep learning networks are very powerful, this feature generating strategy improved classification.
- We provide an ESC model with good accuracy that could help forensic investigators with crime prevention

1.4. Hypothesis

This study aims to test the concept that sound analysis can yield relevant features that can be retrieved and put into machine learning models to enhance categorization, which can aid digital forensic scientists.

1.5. Thesis Structure

The following is the format in which the thesis is prepared and presented.

The first chapter serves as an overview of the research. It is a concise statement about the significance of the study topic, as well as a quick description of the method and the outcome.

The second chapter summarizes research that have attempted to address the same topic, either directly or indirectly, throughout the literature. We attempt to present the results of these additional investigations, as well as their techniques and materials, in this chapter.

Chapter three discusses in detail the scientific theories behind some of the popular concepts adopted in this research. It describes concepts such as sound and signal, representations of audio, and most importantly features in audio signal that can be used for deep learning classifications.

The materials we used for this experiment, particularly the dataset, are discussed in length in Chapter 4. This is the project's core, as it explains every facet of the developed model in great depth. It also describes the methods used, as well as the mathematical foundations for some of the theories and concepts included in the model.

The results and discoveries of our model are presented in Chapter 5. We utilize tables and charts to discuss the model's performance on the datasets that were used.

Our experiment concludes in Chapter 6, which highlights the contribution of our research to the body of knowledge. In addition, we made a recommendation.

2. LITERATURE REVIEW

The classification of environmental sound (ESC) is a significant and difficult task which has attracted a lot of research interest in the past and continue to do so. One of the interesting papers [7] achieved cutting-edge performance when utilizing discriminative models such filter bank features and wavelet-based features. For classification, machine learning methods such as KNN and SVM were used. However, the performance achieved, 67% using these approaches were not as satisfying as thought. One issue, as previously stated, is that classical classifiers lack feature extraction capability. In [8], They investigate using contrastive learning to learn audio representations from a variety of sources. They experimented with a variety of input formats, structures, and augmentations. It is discovered that by improving the concordance between two views of the same audio represented by the raw waveform and log-Mel filter banks, considerably superior representations can be learned. Furthermore, with a new cutting-edge accuracy of 90.5 percent, it extrapolates to the ESC-50 downstream classification problem.

More recently, we have seen an increasing use of neural networks and deep learning to make classification which has proven to increase the performance gain as well as reliability of models. Deep Neural Network can extract features from raw data. The authors of [9] created a classification model that feeds, using a long-short-term memory (LSTM), a deep convolutional neural network (CNN). The paper is distinguished by the presence of several feature channels composed of the Constant Q-Transform (CQT), Mel-Frequency Cepstral Coefficients (MFCC), Chromogram and Gammatone Frequency Cepstral Coefficients (GFCC). These multiples feature with deep convolutional neural network was never used for audio or sound processing previously. To improve performance, some data augmentation techniques were applied. The model was able to achieve cutting-edge performance on ESC-50 dataset. An 88.5 percent precision was achieved by the model which is higher than the human precision of 81.3.

Table 2.1 is the highlights table the literatures produced in this topic as well as several designs with model performances [10]

Table 2.1. Literature Review Table Showing Highlights of papers

Research Focus	**Accuracy (%)**	**Technology**
Efficient end-to-end audio classification with AclNet & CNN **[11]**	85.7	Data Augmentation with CNN mixup
(Between-class) Examples for Deep Sound Recognition **[12]**	84.9	Data Augmentation with EnvNet

Embeddings for Machine Listening Applications On Open-Set Classification with L3-Net [13]	85.0	Openll3 embeddings with x-vector network
Novel Phase Encoded Mel Filterbank Energies for ESC [14]	84.2	Phase encoded filter-banks energies (PEFBE) with CNN
Unsupervised Filterbank Learning Using Convolutional RBM for ESC [15]	83.0	GTSC and CNN with Convolutional RBM & Fusion
Knowledge Transfer from Poorly Labeled Audio using CNN for Sound Events [16]	83.5	Pretrained CNN on Audio
Unsupervised Audiovisual Learning using Deep Multimodal Clustering [17]	82.6	Unsupervised visual learning with CNN
Objects that Sound [18]	79.8	Stride 2 with L3 network
Look, Listen and Learn [19]	79.3	Pretrained on Audio-Visual Job on Convolutional Subnetwork (8-layered)
Multi-scale CNN for learning Environmental Sounds [20]	79.1	Feature Fusion with Multi-Scale Convolution
ESC using Novel TEO-based Gammatone Features [21]	82.0	Fusion of GTSC & TEO-GTSC using Convolutional Neural Networks
(Between-class Examples) DSR - Deep Sound Recognition [12]	73.3	CNN on raw wave files (18-layered) (Between-Class learning)

End-To-End CNN (EnvNet) for ESC [22]	71.0	Spectrogram and CNN
Image Recognition Network for ESC [23]	68.7	AlexNet on spectrograms, sampling rate of 16 kHz
Sound representations from unlabeled video [24]	74.2	Transfer learning from unlabeled videos to CNN raw wave files
Deep CNN for Raw Waveform [25]	68.5	CNN on raw wave files
Real-Time Audio Sound Classification Using Mixture-Based Model [26]	94.0	A database of classifying sound models
Audio Event Recognition Features and Kernels [27]	NA	SVM, GMM, MFCC
Audio Events Classification Using Deep Learning: Comparing Time and Frequency Domain for [28]	NA	On the ESC-10 and Freiburg-106, the discriminatory comparison of various signal representations was evaluated
Audio Event and Scene Classification: A Integra**ted** Approach using YouTube and strong labeling ESC-10 Data [29]	NA	Acoustic Activity Recognition using a combination of weakly labeled data
Masked Conditional Neural Networks for Automatic Sound Events Recognition [30]	NA	Masked ConditionaL Neural Network (MCLNN) and ConditionaL Neural Network (CLNN) for multi-dimensional temporal signal identification

DNN-based learning and transferring mid-level audio features for ESC [31]	N/A	Transfer learning from ESC-50 and other datasets
An Ensemble of Convolutional Neural Networks for Audio Classification [32]	88.7	Data Augmentation with CNN

3. THEORETICAL BACKGROUND

This section explains the theories that underpin the various concepts employed in this study.

3.1. Sound and Signal

A record of sound is called audio. It is often composed of a series of binary digits or, in the case of analog signals, a changing level of electrical voltage. The auditory frequencies capacity is limited around 20 to 20,000 Hz, which corresponds to human hearing's low and high thresholds. [33]. You will often find us referring to the ambient noise we collected as signal or sound in this thesis.

A microphone is a device that detects these differences and converts them into an electrical signal that may be used to represent sound. A speaker is a sound-producing device that receives an electrical signal. Because they transduce, or convert, signals from one form to another, microphones and speakers are referred to as transducers.

3.1.1. Parameters

Audio signals can be characterized based on their power level in decibels (dB), bandwidth, nominal level, and power level. The impedance of the signal line determines the connection between power and voltage. Single-ended or balanced signal pathways are possible. [33]

3.1.2. Periodic Signal

When a signal is repeated over a cycle of time or at a regular interval of time, it is called a periodic signal. This means that the pattern of a periodic signal is repeated across time

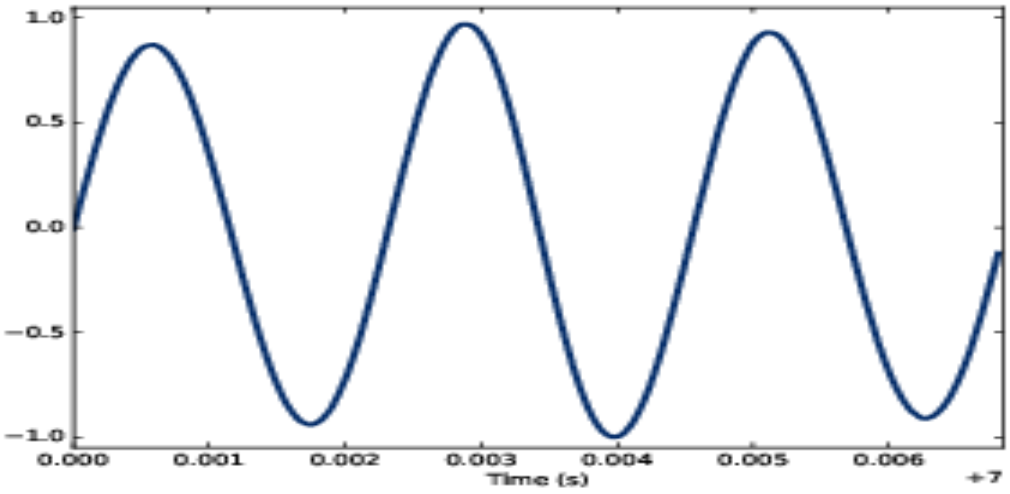


Figure 3.1. Periodic signal figure of a sound signal showing the regular crest and trough

Figure 3.1. Periodic signal figure of a sound signal showing the regular crest and trough

This signal has the same structure as the trigonometric sine function and resembles a sinusoid. This signal is periodic, as you can see. The period was chosen to depict three complete repetitions, commonly known as cycles.

3.1.3. Aperiodic Signal

Unlike periodic signals, non-periodic signals (also known as aperiodic signals) do not have a single frequency. Instead, they are dispersed throughout a wide frequency range. A spoken signal, for example, has a frequency range of roughly 100 Hz to a few thousand Hz (for telephone-quality speech, a range of 300 Hz to 3400 Hz is often assumed). We can see how irregular non-periodic signal is in **Figure 3.2**.

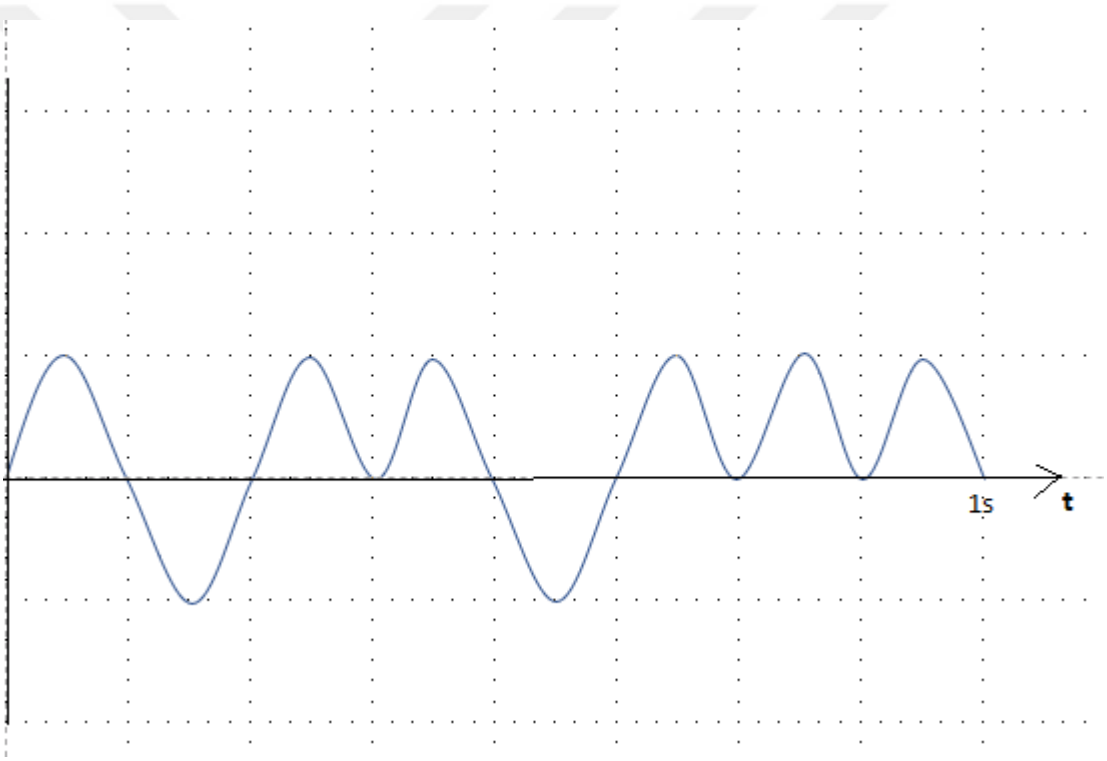


Figure 3.2. Picture of Non-periodic signal of a sound signal

The inverse of the period is the number of cycles per second, which is properly referred to as the frequency of a signal and is measured in Hertz, abbreviated "Hz." (Because the number of cycles is a dimensionless quantity, a Hertz is genuinely a "per second"). A sine wave or cosine wave is a simple periodic analog signal that cannot be broken into smaller signals. A wave is a representation of a signal that has been analyzed at multiple points in time. Each moment in time is referred to as a frame (a term borrowed from movies and video). A sample is the measurement itself, while the terms "frame" and "sample" are sometimes used interchangeably.

Sound signals are always treated in relation to their ambient or keynote context in soundscape investigations because they supplement that environment in the same way that figure and ground complement visual perception. As a result, analyzing signals exposes crucial information about the total sound environment. For example, during this century, the growth in the ambient noise level of cities has been closely related to the increase in the level of emergency warning signals.

Sound signals can be investigated in a variety of ways:

- Subjectively, according to their perceived meanings.
- Historically, based on their development within a certain social setting.
- In comparison to other civilizations or periods, according to kind and purpose.
- In terms of their connotative and associative meanings, symbolically. [34]

However, in the context of our work we are working with the generic features of sound to enable classification.

3.2. Audio Signal Representation

The acoustic properties of a sound recording can be viewed in a time-frequency "image" of the audio wave, to the point that the contributory sources can frequently be separated in the visual domain employing gestalt grouping concepts. [35]

Analyzing sound in the cochlea via frequency is the first step in human auditory perception. As a result, the time-frequency representation of sound is an excellent first step for machine-based segmentation and labeling. The spectrogram and an auditory representation are two essential audio signal representation that aid in the visualization of spectro-temporal sound characteristics. While the former use the Fourier transform to analyze time-changing data, the latter employs auditory perception information to emphasize perceptually significant portions of the signal. [36]

3.2.1. Spectrogram

The Fourier transform of an acoustic signal yields two real-valued frequency functions, the phase spectrum, and the amplitude, which are employed in spectral analysis. At short intervals, Fourier transform spectra of overlapping window function segments have been generated to detect the signal's moment features. Real-world signal waveforms in the time domain that sound comparable. The varying phase relationships between frequency components cause a lot of variability in audio signal processing. The matching magnitude is more perceptually important than the short-time phase spectrum; hence it's left out of the signal representation [37]. The continuous magnitude spectra are used to generate a spectrogram, which is a graphic representation of the signal's time-frequency content. From **Figure 3.3** the spectrogram representation shows the frequencies of the voiceprints in a power-intensity form.

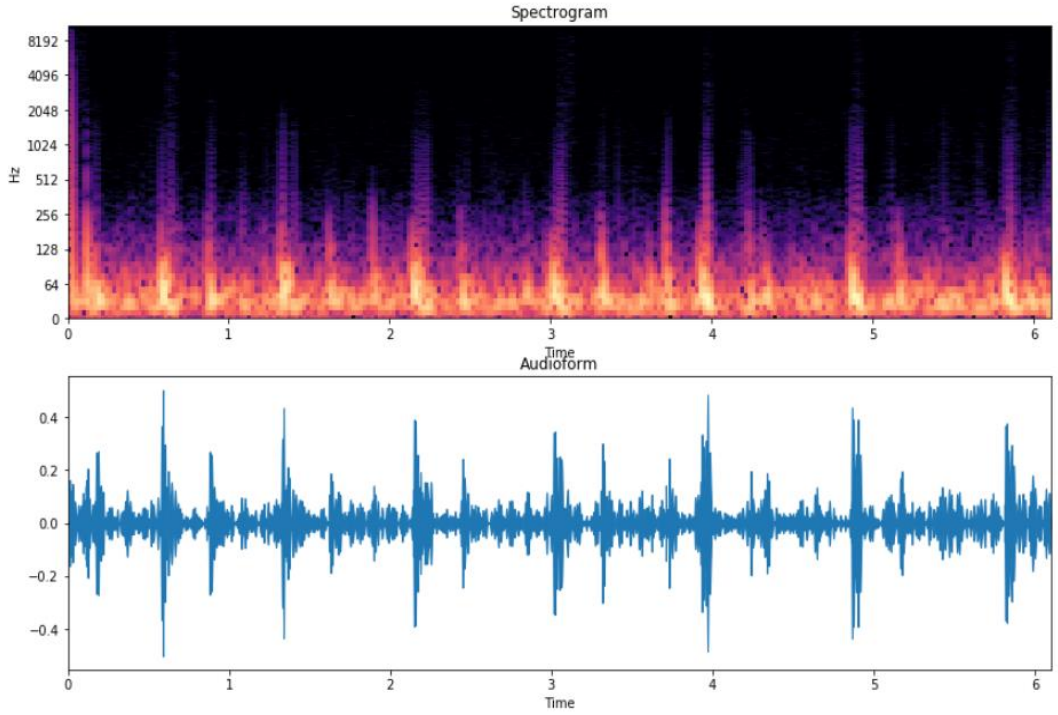


Figure 3.3. Sound Signal Representations in Spectrogram and Audio Form

3.2.2. Auditory Representation

A computational auditory model's output is often an audio representation that expresses a physical quantity at points along the auditory system. The model describes the functions of the auditory nerve's outer, middle, and inner ear in converting sound energy into neuronal code. The computer models are built on heuristics of auditory phases of processing, with model parameters fitting to experimental observations from psychoacoustical research findings. Cochlear models, for example, replicate the band-pass filtering effect of the basilar membrane and the resulting activation rates of hair-cell neurons because of cochlear location. [38] [39]. **Figure 3.4** shows the representations of the cochlear model.

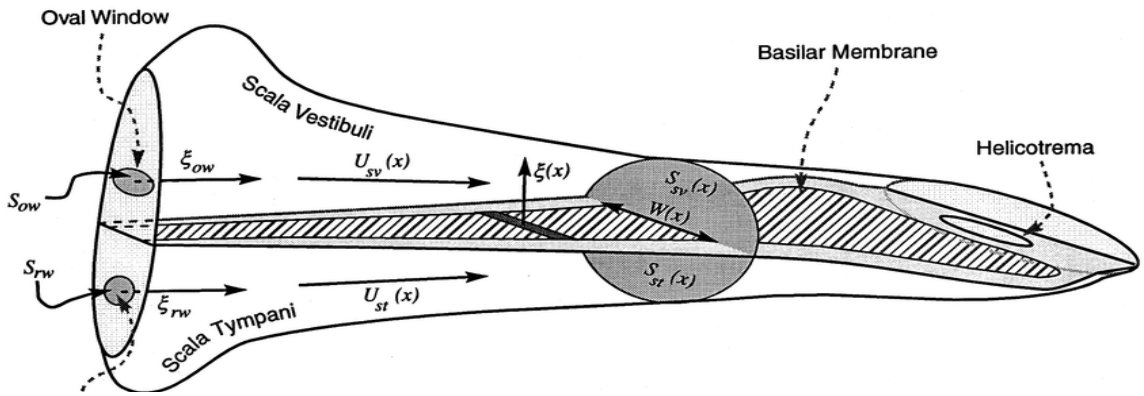


Figure 3.4. How the Cochlear Model of Audio Representation looks like
Image Credit: Glenis Raewyn Long

3.3. Acoustic Signal Processing

Machine learning is the application of algorithms and statistical models to efficiently complete a task without having to explicitly program the instructions for doing so. Instead, the algorithms use the data to learn how to accomplish the intended function. A training dataset is used in supervised learning, and each sample is labeled with the right output. manual annotation by people inspecting the data is usually used to generate these labels, which is a time-consuming and costly operation. models are trained without access to labeled data in unsupervised learning. this is frequently employed in cluster analysis (automatic discovery of sample groups).

Regression and classification can both benefit from supervised learning approaches. The purpose of regression is to forecast a continuous real-valued variable, and classification is to predict a discrete variable.

This research thesis examines a new model and introduces a subfield of ambient sound classification. This is the first study to use environmental sound classification to solve the challenge of multi-story building location recognition. Other noteworthy studies in digital sound classification are listed in the Literature Review. It's understandable that ESC is a difficult problem for machine learning to solve, hence numerous ways have been created over time to allow humans to accurately classify ESC levels in machines. Deep learning and convolutional neural networks are two of these methods. Converting digital sounds to pictures and creating Mel-frequency Cepstral Coefficients is a prominent method for extracting features. CNN models, on the other hand, are developed for computer vision and have a significant time load. As a result, we set out to create an effective learning model to address this problem and contribute to digital forensic research.

3.4. Audio Features for Classification

Audio features are descriptions of sounds or audio signals that can be put into statistical or machine learning algorithms to create intelligent systems. Machine Learning has the potential to revolutionize how digital researchers investigate pattern evidence as well as other types of evidence. Although no automated system can fully replace people in the interpretation and analysis of evidence, algorithms can serve to reduce some of the subjectivity that imbues most forensic subject areas, as well as approximate the level of uncertainty in forensic inferences, with the right background data sources and comprehensive verification and debugging [40].

Various characteristics capture various components of sound. Generally, audio features are classified according to the following criteria:

- High-level, mid-level, and low-level aspects of musical signals are classified as levels of abstraction.

- Temporal Scope: Instantaneous, segment-level, and global time-domain properties.
- Musical Aspect: Beat, rhythm, timbre (sound color), pitch, harmony, melody, and other acoustic qualities.
- Signal Domain: Features in the time, frequency, or both domains.
- Hand-picked features for classical ML modeling or automatic feature extraction for deep learning modeling are two approaches to machine learning. [41]

Our work focuses on using hand-picked features for the model since the potential for deep learning models is enormous. Unstructured audio representations, such as the spectrogram or MFCCs, are considered in the Deep Learning technique. It is capable of extracting patterns on its own. Because feature extraction is automated, this became the favored method by the late 2010s. It's also aided by the vast amount of data and computing power available. . [41]

Spectrograms, Mel-spectrograms, and Mel-Frequency Cepstral Coefficients are examples of commonly used features or representations that are directly input into neural network designs (MFCCs). Discussion on our feature generation and extraction will be discussed extensively in the next chapter.

4. MATERIALS AND METHOD

This section describes the materials and methodology that have been employed to create a reliable model for our audio classification. It includes the dataset used, model description as well as feature generation, extraction, and model implementation.

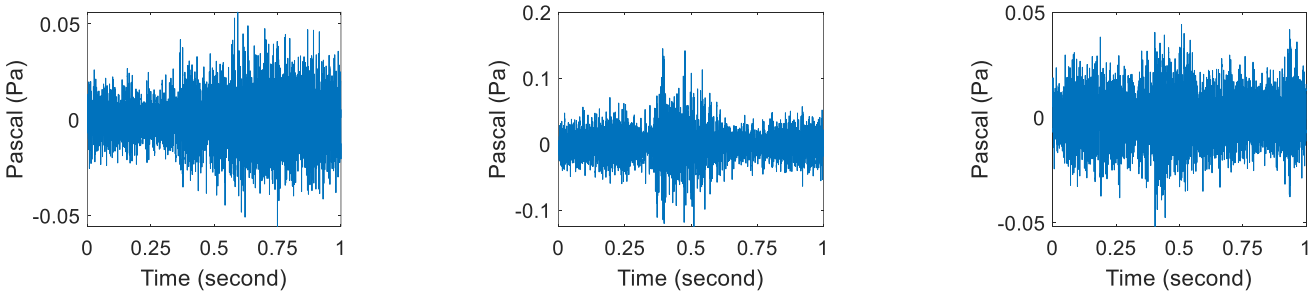
4.1. Dataset

We gathered a new ESC dataset for this study to determine sound classification in a multi-story building. On weekdays, we recorded environmental sound from each of the Firat University Hospital's ten floors. Because it is a residential building, the hospital was chosen. Because the sample rate of the gathered audio is 48kHz, the audio was collected and separated into segments with a frame length of 48000 using a Huawei Mate 20 Lite. Each of the ten classes in the dataset represents a different floor of the building. The features of this dataset are listed below, and it was published on web.firat.edu.tr/turkertuncer/hospital.rar. **Table 4.1** shows the number of segments obtained from each of the ten classes. Illustratively, ten random samples from each of the classes was picked and the time-frequency graphs shown in **Figure 4.1**.

Table 4.1. Audio classes table and number of segments obtained from each

Class	***Number of segments***	***Class***	***Number of segments***
1	326	6	355
2	330	7	346
3	317	8	397
4	315	9	317
5	302	10	305

Figure 4.1 displays graphical illustrations of the categories.



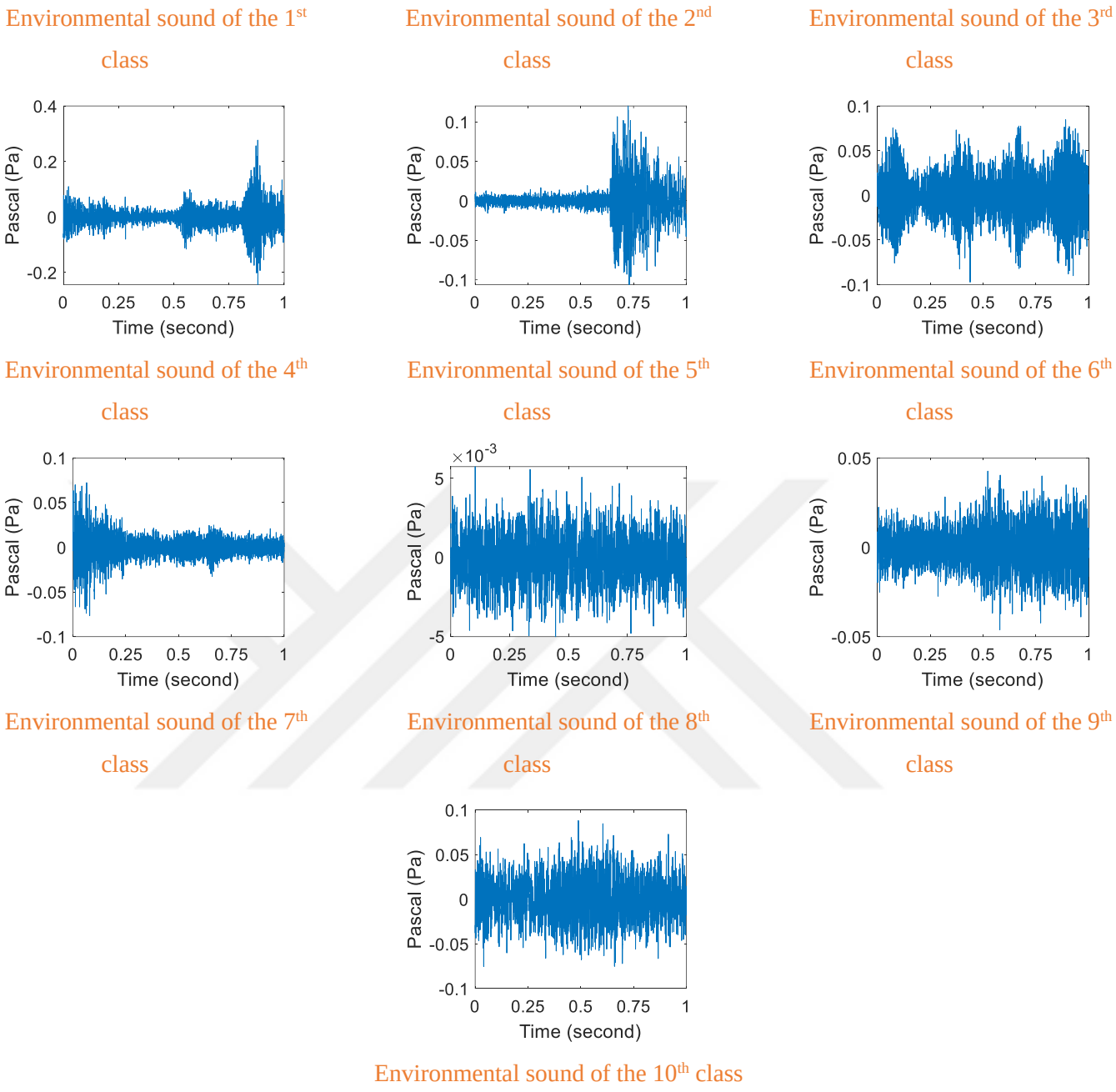


Figure 4.1. Graphical representations of the time-frequency graphs of random classes of the audio classes

4.2. Center Symmetric LBlock Pattern

A novel nonlinear handcrafted feature generating function has been presented in the research work. The work utilizes S-Box to postulate nonlinear pattern. As shown in the literature, cryptography-based feature generators can tackle one-dimensional signal classification issues with excellent accuracy [42] [43]. S-Boxes are nonlinear structures as seen in block ciphers and this was primarily the reason why we picked it having proven its nonlinearity in cryptology.

The 4-bit S-Boxes are commonly used in lightweight block ciphers. A nonlinear new generation microstructure can be presented by using the S-Boxes. The selected S-Box is shown below in Table 4.2. [44] [45]. **Table 4.2** and **Figure 4.2** shows how the mapping of values is achieved through S-Box and Signum Function, respectively.

Table 4.2. S-Box table of the LBlock Cipher with mappings of values

x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
S(x)	15	10	16	1	14	5	11	12	2	3	9	4	8	7	13	6

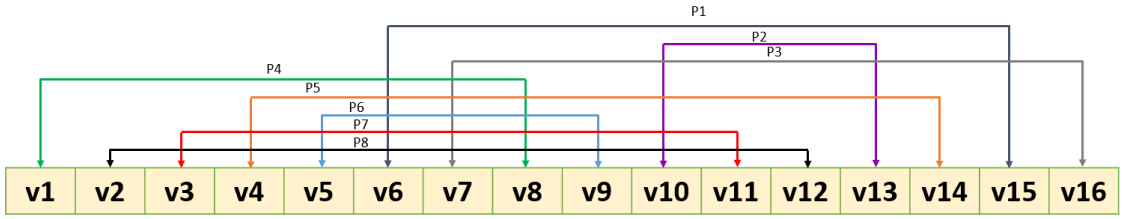


Figure 4.2. Representing the proposed CS-Lblock Pattern for graphical understanding

Binary features have been obtained and transformed to digits using Figure 4.2 and the CS-Lblock-Pattern to create a map or feature signal. This feature for our model will be the histogram.
The following are the steps of the feature generation function used in the CS-Lblock-Pattern.

0. Load audio data
1. Split the signal into sixteen (16) overlapping pieces.

   $$block^i = \ i = \{1,2,\dots,lngt-15\}, \qquad j = \{1,2,\dots,16\}, \qquad sound(j+i-1) \quad (4.1)$$

   here block, sound, and lngt are 16-dimensional overlapped blocks, inputting 1-dimensional signal, and signal lengths, respectively. Equation 4.1 defines overlapped block splits using 16 sized blocks theoretically.

2. Create binary features using the pattern shown in Figure 4.2 and the signum function.

   $$bt^i(k) = signum\left(block^i\big(S(k)\big), block^i\big(S(17-k)\big)\right), k = \{1,2,\dots,8\} \quad (4.2)$$

   $$signum(x,y) = \begin{cases} 0, & x-y<0 \\ 1, & x-y \geq 0 \end{cases} \quad (4.3)$$

   where bt^i are the bits extracted from i[th] block, $signum(.,.)$ represents signum function and x, y are input parameters of the signum function, $S(.)$ represents the used LBlock S-box which is defined in Table 4.2. Equations. 4.2-4.3 define the bit generation of the presented center symmetric and nonlinear feature generator. By deploying these equations, eight bits are generated from each overlapping block with a length of 16.

3. Create a map signal by converting the generated bits (bt) into a decimal value. Equation 4.4 describes this procedure.

$$map(i) = \sum_{k=1}^{8} bt^{i}(k) * 2^{k-1} \tag{4.4}$$

Equation 4.4 is used to transform the resulting 8 bits to a decimal value, which is then used to generate the map signals.

4. Obtain the histogram of the map signal. An 8-bit code is used to annotate the map signal and as a result, the length of the retrieved histogram is estimated as 2^8=256.

$$histo(t) = 0; t = \{1,2,\dots,256\} \tag{4.5}$$

$$histo\big(map(i)\big) = histo\big(map(i)\big) + 1 \tag{4.6}$$

The initial value assignment of the histogram (histo) is shown in Equation 4.5. Equation 4.6 defines histogram extraction mathematically.

The retrieved histogram is used as a 256-dimensional feature vector. The feature generating technique of the CS-Lblock-Pattern is defined by the steps listed above (see steps 0-4). Figure 4.3 depicts a numerical model of the provided CS-Lblock-Pattern for clearer understanding.

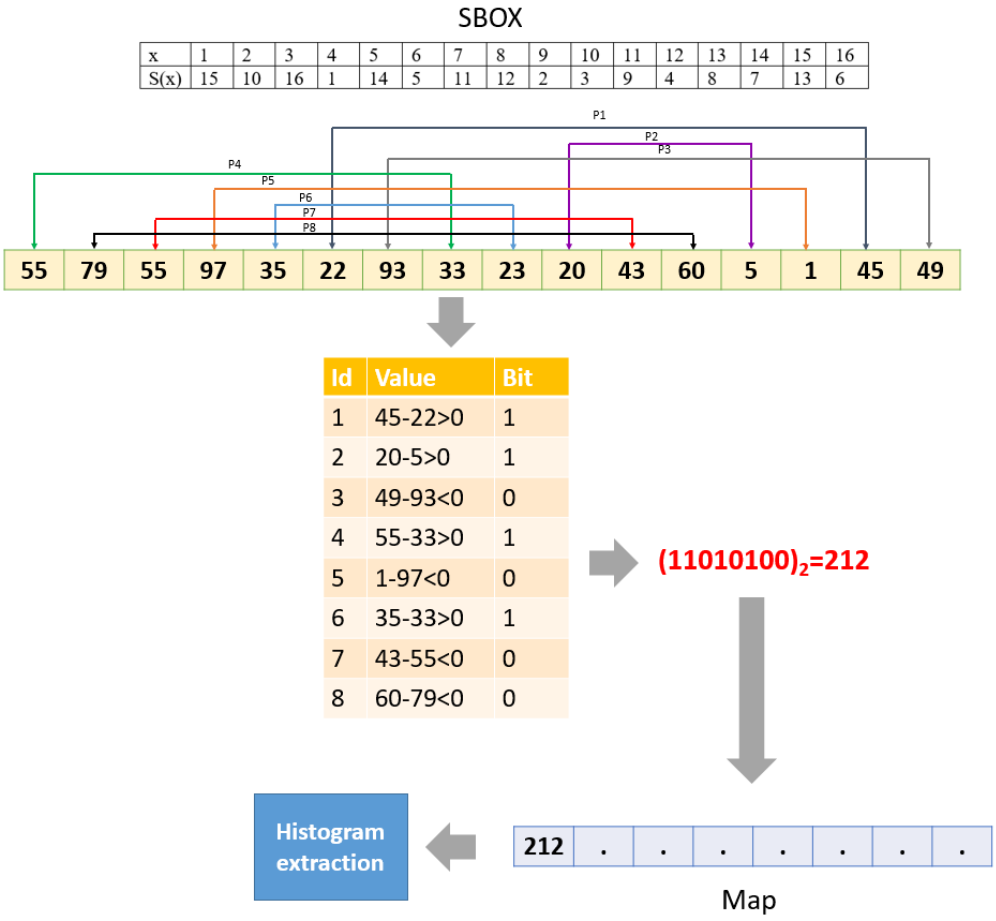


Figure 4.3. Visual summary of the CS-Lblock-Pattern Model for each step of the feature extraction process

We can see from **Figure 4.3** the summary of the feature extraction and engineering process.

4.3. Environment Sound Classification using CS-Lblock-Pattern Model

A model of ESC which can be utilized in the recognition of floors in multi-story structures in introduced in this thesis. This new model utilizes the CS-Lblock-Pattern with the INCA-based model. Past exploration works have recognized ESC as one of most muddled research viewpoints in the space of AI and security. This original model will be profoundly advantageous in the identification of ambient noise. This model comprises of three phases which incorporate the feature generation or extraction phase, feature selection phase and the classification phase. CS-Lblock-Pattern and statistical moments is utilized in the feature generation phase. The feature selection phase is a discriminative stage which utilizes the INCA selector, while two simplistic classifiers are used in the classification phase.

To create a feature in multi-level, a decomposition model is needed. A TQWT will be used as a decomposition model during this phase. Applying the 22 statistical moments and the CS-Lblock-Pattern, each signal yielded 300 features. A total of 6300 features were created from a sound using 20 decomposed signals and a raw sound source. The INCA was then used as a feature selector to tackle the challenge of selecting the optimal number of features. The categorization results are then generated using shallow classifiers. Below is a breakdown of the means in the system.

Step 0: Load audio data.

Step 1: On the acoustic signal, apply TQWT with the Q-factor (Q), number of levels (j) and redundancy (r) parameters set to 1, 19, and 3, respectively. Using the settings supplied, generate 20 fragmented noises.

Step 2: The 22 statistical moments with postulated CS-Lblock-Pattern can be used to generate features. For every level, 300 features were generated. The hypothesized CS-Lblock-Pattern was used to construct 256 of them, with approximately 22 formed using statistical moments of nonlinear textural qualities and 22 formed using statistical moments of decomposed sound.

Step 3: To produce the final feature vector, concatenate the retrieved features having dimension 21 x 300=6300.

Step 4: Choose the most explanatory characteristics for the final feature vector using the ICA selector. However, 523 features were chosen as the most useful for this situation.

Step 5: To offer an example of this model, the INCA sends the selected feature vector to traditional classifiers. **Figure 4.4** illustrates the block diagram and depicts the entire framework.

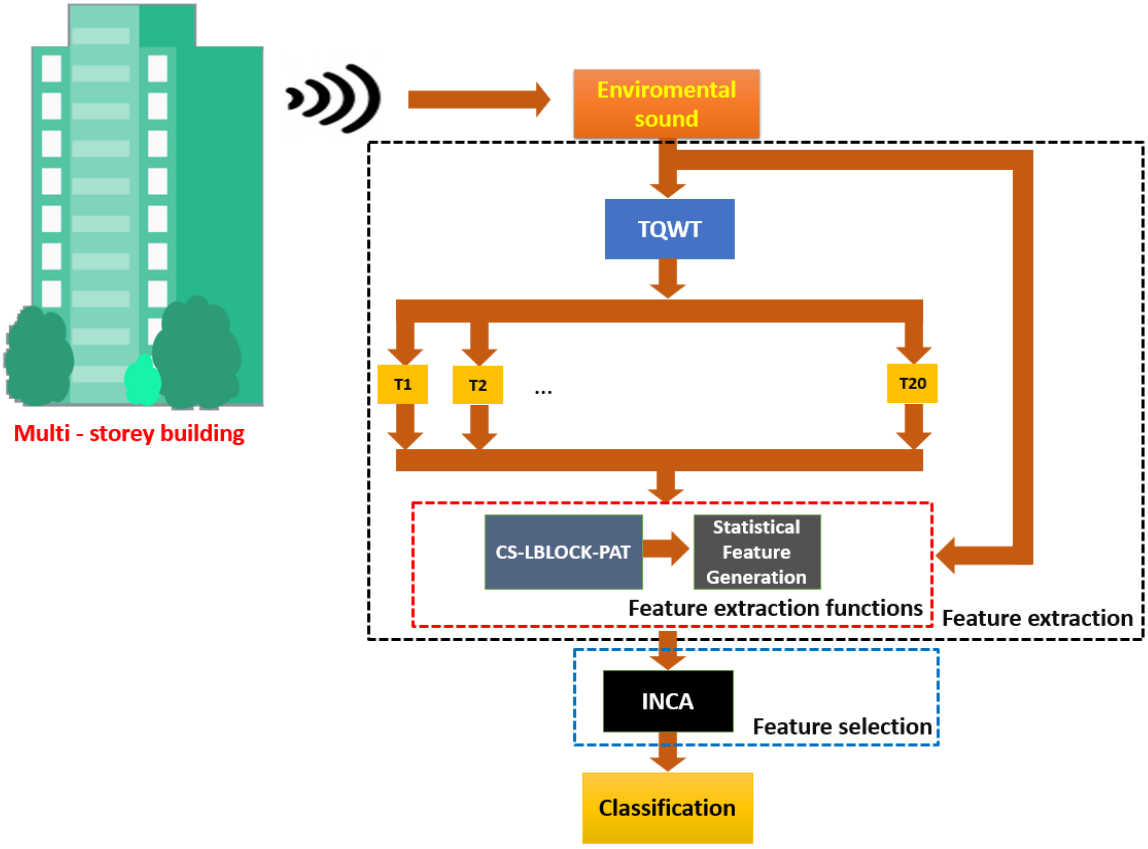


Figure 4.4. Summary Figure of the ESC Model illustrating the entire process of engineering

4.4. Feature Engineering

This segment delves into the details of multi - layer and hand-crafted features engineering. This model employs two feature generating strategies, as shown in **Figure 4.4**. This is the generation of statistical and textural features. Statistical and textural feature generation are the two models for feature engineering strategy. As a result, there are three sections in this study that discuss the proposed features network. These segments include:

- TFG or Textural Feature Generation
- SFG or Statistical Feature Generation
- The proposed multilevel feature generating network's general steps

4.4.1. Textural Feature Generation

As shown in the model in section 3, as a feature generating function to produce textural characteristics the postulated CS-Lblock-Pattern will be used. In the general phases, CS-Lblock-Pattern(.) work is used to express this method.

4.4.2. Statistical Feature Generation

The statistical moment is one among the foremost utilized hand-crafted feature extraction functions, and its primary objective is to assist produce features by utilizing a mixture of statistical and textural feature generation function. During this study, 22 statistical moments are used, and therefore the resulting equation is provided below.

$$bf(1) = \frac{\sum_{i=1}^{L} sound_i}{L} \tag{4.7}$$

$$bf(2) = \sqrt{\frac{\sum_{i=1}^{L}(sound_i - bf(1))^2}{L}} \tag{4.8}$$

$$bf(3) = \sqrt{\frac{\sum_{i=1}^{L}(sound_i)^2}{L}} \tag{4.9}$$

$$bf(4) = -\sum_{i=1}^{L} prob(sound_i) * \log\left(prob(sound_i)\right) \tag{4.10}$$

$$bf(5) = \sum_{i=1}^{L} \frac{i * sound_i - b(1)}{b(1)} \tag{4.11}$$

$$bf(6) = \frac{\sum_{i=1}^{L} |signal_{i+1} - signal_i|}{L} \tag{4.12}$$

$$bf(7) = \frac{\sqrt{L(L-1)}}{L-2}\left(\frac{\frac{1}{L}\sum_{i=1}^{L}(sound_i - b(1))^3}{\frac{1}{L}\sum_{i=1}^{L}(sound_i - b(1))^2}\right) \tag{4.13}$$

$$bf(8) = \frac{L-1}{(L-2)(L-3)}\left[(L+1)\left(\left(\frac{\frac{1}{L}\sum_{i=1}^{L}(sound_i - b(1))^4}{\frac{1}{L}\sum_{i=1}^{L}(sound_i - b(1))^2}\right) - 3\right) + 6\right] \tag{4.14}$$

$$bf(9) = sound(\frac{L}{2}) \tag{4.15}$$

$$bf(10) = min\{sound\} \tag{4.16}$$

$$bf(11) = \max\{sound\} \tag{4.17}$$

$$bf(12) = \frac{bf(1)}{bf(2)} \tag{4.18}$$

$$bf(13) = \max\{sound\} - median\{sound\} \tag{4.19}$$

$$bf(14) = \max\{sound\} / median\{sound\} \tag{4.20}$$

$$bf(15) = \max\{sound\}/b(1) \quad (4.21)$$

$$bf(16) = bf(1)/min\{sound\} \quad (4.22)$$

$$bf(17) = bf(1)/median\{sound\} \quad (4.23)$$

$$bf(18) = bf(1) - min\{sound\} \quad (4.24)$$

$$bf(19) = bf(1) - median\{sound\} \quad (4.25)$$

$$bf(20) = bf(2) - median\{sound\} \quad (4.26)$$

$$bf(21) = bf(2)/median\{sound\} \quad (4.27)$$

$$bf(21) = bf(2)/median\{sound\} \quad (4.28)$$

where bf denotes the extracted statistical features and prob denotes the sound's probability. The statistical properties are defined by the provided equations (Equations. 4.7-4.28). Statistical moments are commonly employed to construct statistical features.

4.4.3. Steps in the Feature Engineering Process

0: Load audio data

1: 20 sub-bands (SB) is generated by applying TQWT

$$SB = TQWT(sound, 1, 3, 19), \ , r = 3, \ J = 19, \ Q = 1 \quad (4.29)$$

TWQT is the key decomposer, and it uses the Q-factor, Level number (J), and redundancy (r) parameters to create sub-bands (TWQT). The oscillatory value is determined by the Q value. When the specified parameter in equation 29 is employed, twenty sub-bands are created. These sub-bands and the feature creation routines used are the medium for generating high-level and mid-level features.

2: Extraction of fused properties from a audio signal with its sub-bands.

$$feat^1 = concat\big(CS - Lblock - Pat(sound), statis(sound), statis(CS - Lblock - Pat(sound))\big) \quad (4.30)$$

$$feat^r = conc\big(CS - Lblock - Pat(SB^{r-1}), statis(SB^{r-1}), statis(CS - Lblock - Pat(SB^{r-1}))\big) \quad (4.31)$$

$$r = \{2,3, \dots ,21\} \quad (4.32)$$

CS-Lblock-Pattern(.) and Statis(.) are CS-Lblock-Pattern nonlinear text and statistical feature generating functions correspondingly, in Equations 4.30-4.32, while conc(.) symbolizes the concatenation function.

3: To acquire the final feature vector, convolve the feature vectors with sizes of 300.

$$X = conc(feat^1, feat^2 \ldots, feat^r) \quad (4.33)$$

where X has a magnitude of 6300 and represents the final feature vector.

4.4.4. Feature Selection

In feature selection, the INCA, which is a more advanced and sophisticated version of the NCA, is used as the feature selector. The INCA is used to identify and select the 6300 most informative characteristics extracted from the 21 signals. Tuncer et al. [46] proposed the INCA, which was designed to automatically determine the optimal number of characteristics. Figure 4.5 depicts the INCA selector's procedure.

Procedure: $INCA(X, target, C)$

Input: Extracted features (X) with size of d x K, target with size of d and classifier (C).

Output: Final feature vector ($last$) with size of j.

00: Load X

01: for i=1 to K do

02: $X(:,i) = \frac{X(:,i) - min(X(:,i))}{max(X(:,i)) - min(X(:,i))}$; //Apply min-max normalization to each column of the feature vector. NCA is a distance based feature selector. In order to use NCA effectively, normalization must be performed.

03: end for i

04: $endex = NCA(X, target)$; // Calculate indices of the features by using NCA. These indices are ordered by descending.

05: for k=1 to 900 do // Apply iterative NCA. Select number of features from 100 to 1000. The main aim of this step is to reduce computational complexity of this method.

06: for l=1 to 99+k do

07: $f^{NCA}(:,l) = X\big(:, endex(l)\big)$; // Iterative NCA feature selection process.

08: end for l

09: $error(k) = C(f^{NCA})$; // Calculate error rates of each selected feature vector. In this work, kNN is used with 10-fold cross validation.

10: end for k

11: $[minimum, index] = \min(error)$; // Find minimum error and index of it.

12: for j=1 to $index + 99$ do // Select optimal features.

13: $last(:,j) = X\big(:, endex(j)\big)$;

14: end for j

Figure 4.5. Pseudo code of the INCA Feature Selection Process

Figure 4.5 represents the pseudo code and it depicts how the INCA employs three variables: the feature matrix, the target, and the classification algorithm, while the NCA serves as a distance-based controller. Because of these pairings, the features are subjected to a min-max normalization, as shown in lines 01-03. The normalized feature vector is subjected to NCA, and sorted indices are generated. Because the INCA is complicated, various characteristics in the [100,1000] range have been introduced to make it easier to understand. A classifier is used to compute the error/loss values for each feature vector. The LD was the most popular classifier because it uses 10-fold cross validation to quantify the value of error/loss. Lines 4.11-4.14 describe the process of optimal feature selection, which resulted in the 523 most informative features being chosen.

4.4.5. The Classification System

The classification phase of our suggested ESC classification approach is the final step. In this step, measurements are made using two shallow classifiers, the LD and SVM classifiers.

Linear Discriminants (LD) are a statistical dimensionality reduction strategy that provides the best discrimination between different classes. They're used in machine learning to find the optimal linear combination of attributes for separating two or more object classes. Pattern recognition, picture retrieval, and speech recognition are just a few of the applications where it's been used. The method works by estimating discriminant functions from a training data set. Regarding the characteristic vector, these discriminant functions are normally linear and take the form [47]

$$f(t) = w^t x + b_0 \tag{4.33}$$

with w representing the weight vector, b_o denotes the threshold and x represents the characteristic vector. Because the LD classifier does not employ parameters in the classification learner toolbox of MATLAB, no parameter was set for the LD classifier in this study [48].

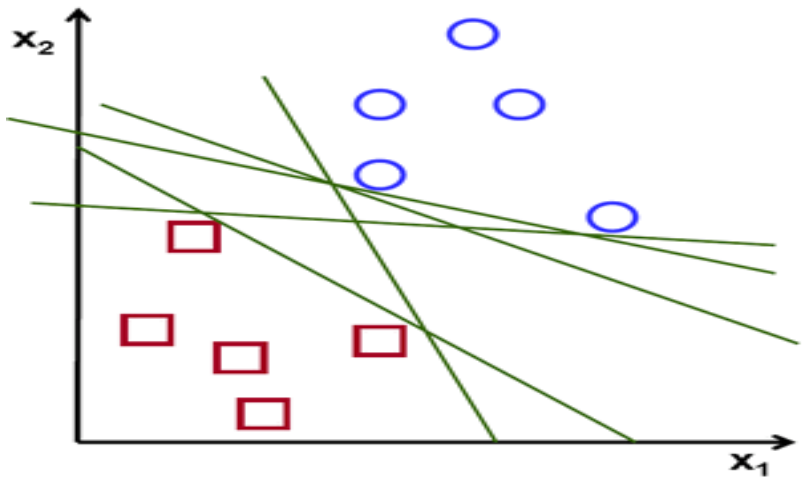


Figure 4.6. Different hyperplanes figures showing how LD can be employed

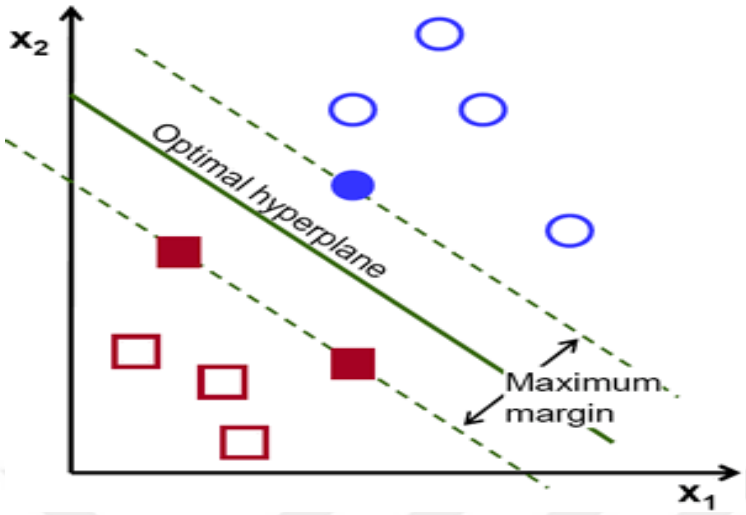


Figure 4.7. Possible Hyperplanes that illustrate how optimal planes are obtained in LD

The two figures; **Figure 4.6** and **Figure 4.7** represents Different hyperplanes and Optimal Hyperplanes, respectively.

SVM, or Support Vector Machine, is a non-parametric supervised learning model. They use the kernel approach to translate data to high-dimensional feature spaces for non-linear regression and classification problems [49]. In a high-dimensional or infinite-dimensional space, SVMs build a hyperplane or set of hyperplanes that may be used for categorization, regression, and other tasks. The SVM classifier, like the LD, is a basic classifier that employs variable kernels.

The 2nd-degree polynomial or quadratic kernel SVM was used in this study. The value of C (box constraint level) is 1, and the classification approach is one-vs-all [50]

SVM has the following drawbacks:

- If the set of features is higher than the sum of samples, it is critical to prevent over-fitting while picking regularization terms and Kernel functions.
- Probability estimates are obtained through a time-consuming five-fold cross-validation technique rather than directly by SVMs.

In these two basic classifiers, the MATLAB classification learner toolbox is utilized, using a k - fold cross validation (10) for feature verification and validation. Both classifiers share properties.

4.4.6. Experimental Setup

We programmed the INCA and CS-Lblock-Pattern model on a 2020a MATrix LABoratory (MATLAB) software on a home computer with a minimal configuration. The ambient noises

collected from each of the multi-story structure's 10 floors were saved in m4a format to create a dataset. To create a rapid reaction model, the dataset was tagged and split into one-second duration segments. The model was then programmed in statistical feature generation, CS-Lblock-Pattern, and INCA methods, and MATLAB was utilized to categorize the specified feature.

5. RESULTS AND DISCUSSION

The segment contains the findings of the experiments. The following are the outcomes of our trials utilizing the postulated CS-Lblock-Pattern model. The section also goes into detail about the result discussion.

5.1. Results

According to the literature, several variable metrics have been used in classification performance testing. This study uses several true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) to estimate F1-score (F1), precision (pre), recall (rec) and performance metrics accuracy (cac) [51].

Table 5.1 Is the mathematical equality table and it shows how different important parameters can be represented mathematically.

Table 5.1. Performance metrics' table showing the mathematical equality of the parameters

Metric	Equation	No
$F1$	$\frac{2tp}{2tp + fn + fp}$	(34)
pre	$\frac{tp}{tp + fp}$	(35)
rec	$\frac{tp}{tp + fn}$	(36)
cac	$\frac{tp + tn}{tp + tn + fp + fn}$	(37)

Table 5.2. The computed results table for every classifier used is shown for the LD and SVM

Classifier	Metric	Performance (%)
SVM	$F1$	95.27
	Classification accuracy	95.32
	Average recall	95.27
	Average precision	95.26
LD	$F1$	93.54
	Classification accuracy	93.53
	Average precision	93.58
	Average recall	93.50

Table 5.2 lists the performance parameters in this section demonstrating how accurate our model was in the prediction process.

The LD & SVM Classifiers Confusion Matrices are shown in Tables 5.3 and 5.4 to demonstrate the results in detail.

Table 5.3 and **Table 5.4** are the confusion matrices tables and they show the classification prediction numbers for the LD and SVM classifiers respectively.

Table 5.3. The Confusion Matrix for the LD Classifier showing the classification prediction

True Class	**Classification Prediction**									
	1	2	3	4	5	6	7	8	9	10
.1	297	19	5	4	0	0	1	0	0	0
.2	28	286	4	12	0	0	0	0	0	0
.3	7	9	283	12	0	3	3	0	0	0
.4	15	10	3	284	1	0	0	2	0	0
.5	1	0	2	1	280	3	13	1	1	0
.6	2	1	3	3	3	334	8	1	0	0
.7	1	1	5	2	2	2	333	0	0	0
.8	1	0	0	6	6	4	0	380	0	0
.9	0	0	1	0	0	1	0	0	315	0
.10	0	0	0	0	0	0	1	0	0	304

Table 5.4. The Confusion Matrix for SVM Classifier showing the classification prediction

True Class	**Classification Prediction**									
	1	2	3	4	5	6	7	8	9	10
.1	307	11	4	2	0	0	2	0	0	0
.2	15	296	7	12	0	0	0	0	0	0
.3	3	7	291	10	1	1	4	0	0	0
.4	9	11	4	287	2	0	0	1	1	0
.5	0	0	3	1	289	2	7	0	0	0
.6	0	2	2	1	1	347	2	0	0	0
.7	1	2	6	1	3	2	329	2	0	0
.8	2	0	0	4	3	0	0	388	0	0
.9	0	0	0	1	0	0	0	0	316	0
.10	0	0	0	0	0	0	0	0	0	305

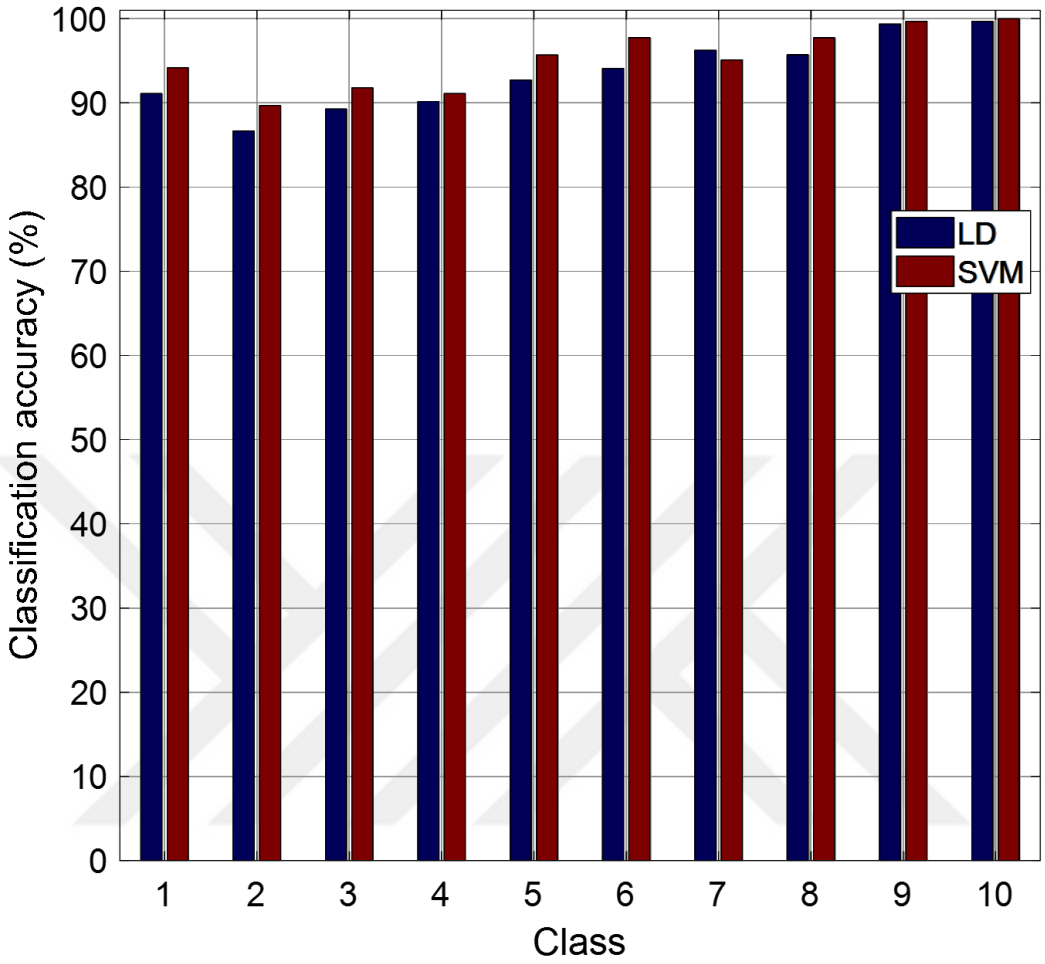


Figure 5.1. The accuracy rates estimated using the LD and SVM classifiers for each class (floor)

Figure 5.1 is the accuracy rate values for each classifier compared side by side. It is evident that the highest performances were in the 10^{th} class.

5.2. Discussions

The goal of this study was to solve ESC using a center-symmetric nonlinear patter. To that purpose, we gathered a reliable dataset from Turkey's Firat University Teaching Hospital. The acquired data was then divided into one-second frames to create a quick reaction system. Furthermore, we were able to produce a novel non-linear pattern in our study by using one of a lightweight block cipher's S-Boxes (LBlock cipher). We dubbed the model CS-Lblock-Pattern because the pattern was symmetrically centered. The main purpose of this new non-linear pattern is to record and collect nonlinear information from a sound stream. This new pattern will be able to generate about 256 patterns, which will aid in the detection of nonlinearities. We have included

statistical elements to aid in the implementation of the planned network. We also made use of the INCA as a feature selector. Figure 5.2 depicts the INCA feature selection procedure.

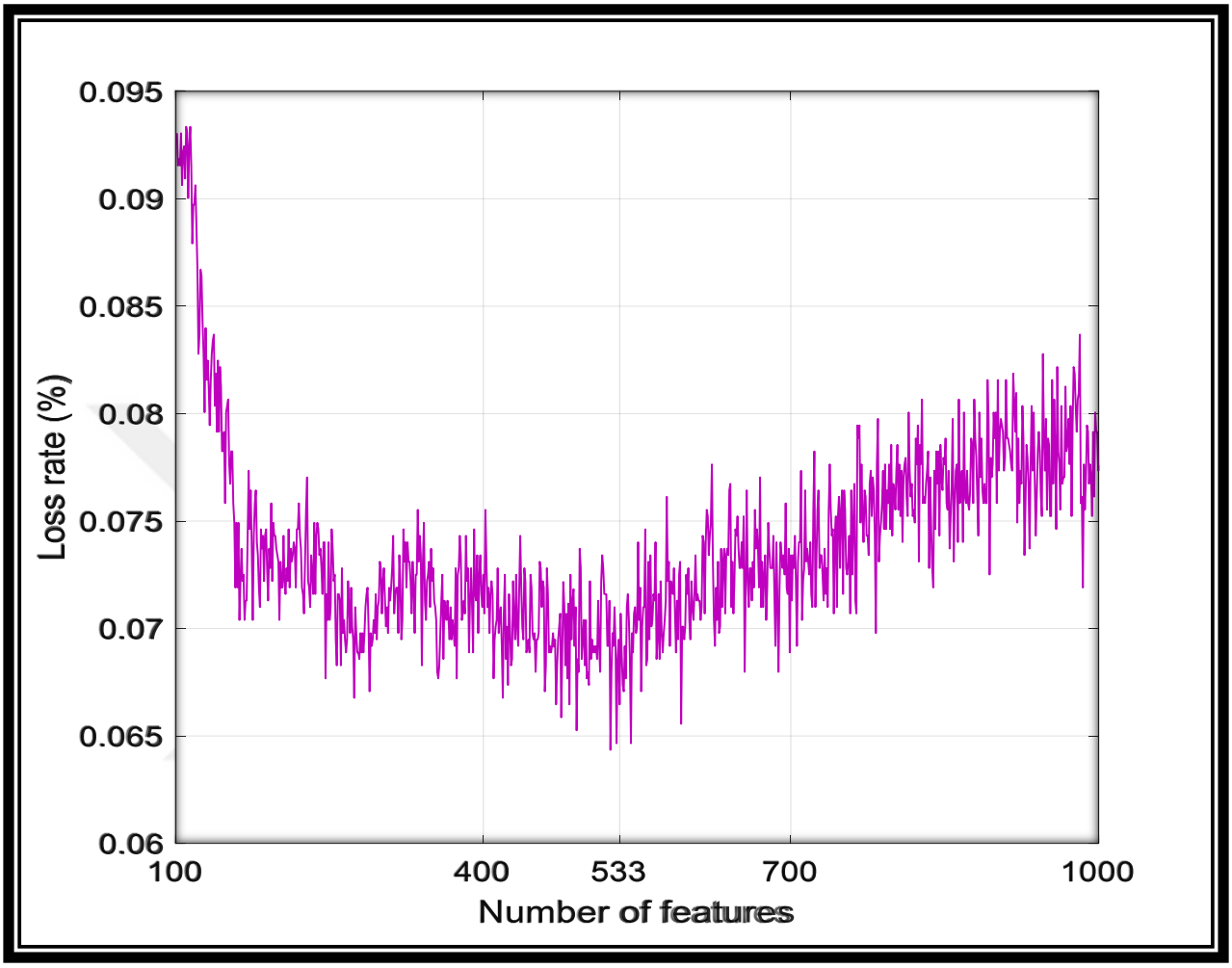


Figure 5.2. INCA's feature selection showing loss function against number of features

Figure 5.2 is the plot of the loss rate against number of features and this allows the loss function to know what features are in the target. We were able to demonstrate the efficacy of the novel feature generation and selection methodologies proposed in this thesis using both classifiers., as the results show that both classifiers achieved an accuracy of 95.38%.

The following are some of the ways in which the proposed model is thought to be advantageous.

- Because standard navigation programs only provide information in two dimensions of longitude and latitude, our newly developed model will be able to classify ambient noise in multi-storey buildings.

- The study also established a new model (CS-Lblock-Pattern) that can be used to analyze the non-linear aspects of environmental sounds and was able to produce a novel ESC data set that can be easily accessible by anyone.
- A multilevel feature generating network was described in the paper, as well as a highly accurate ESC model.

5.3. Importance of the Study

• In the fields of security applications and digital forensics, the proposed approach will be quite useful.

• Using the findings of this study, a new sound forensics toolkit may be created.

• Our proposed model can be employed in environmental sound classification when combined with deep learning and online services.

• Our research will pave the way for the creation of a slew of new non-linear patterns.

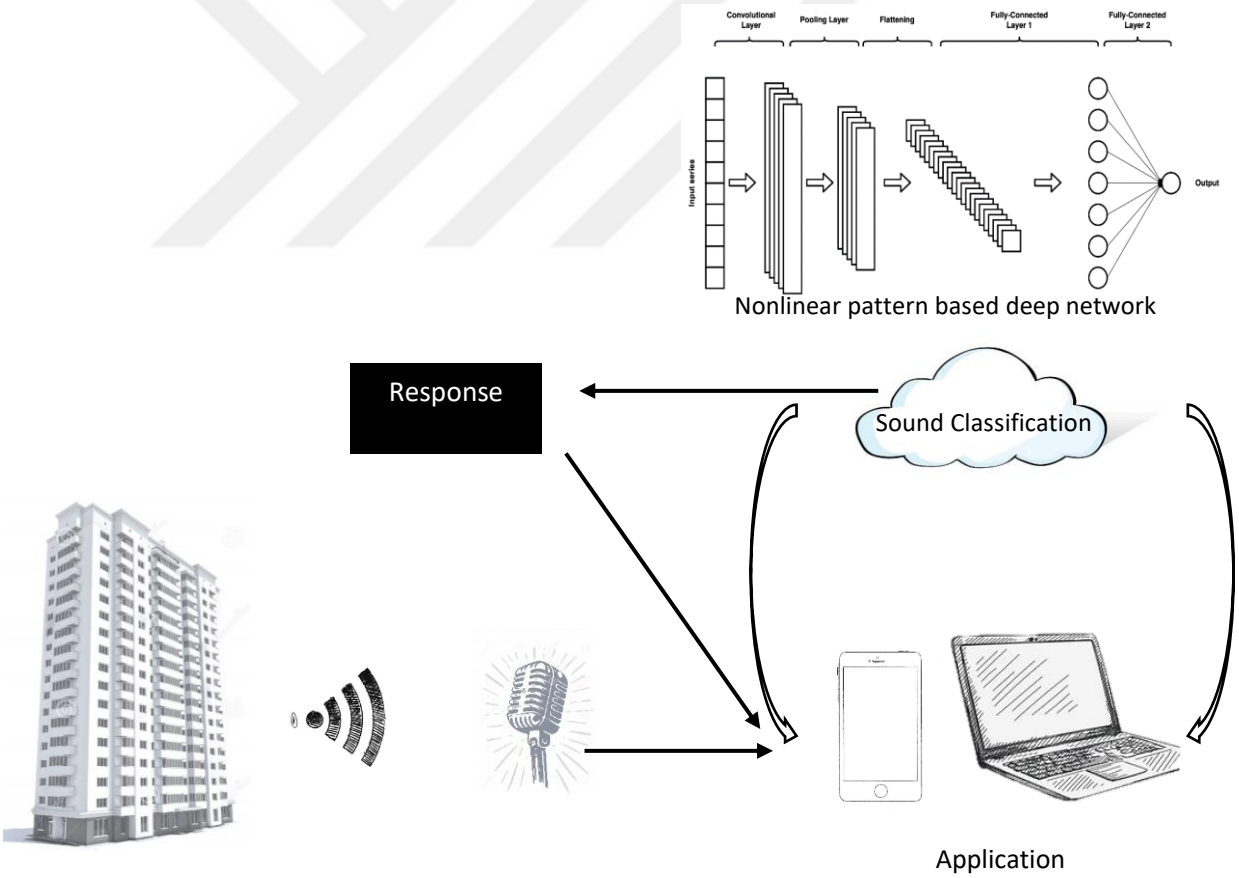


Figure 5.3. Snapshot figure showing route for future or planned work

Figure 5.3 is the route for our planned future work, demonstrating the general environmental sound classification process.

6. CONCLUSIONS

To create a high-accuracy ESC model, this study combined a multistage handmade (statistical features and nonlinear) feature generating model with the INCA-based feature selector. The primary purpose of the model includes utilizing the given model to classify ambient noise in a multi-story structure. The ESC model is divided into three stages: feature generation, feature extraction, and feature generation utilizing TQWT and the postulated CS-Lblock-Pattern. The most relevant characteristics were picked using the INCA selector. The presented feature generating network, which is built on the CS-Lblock-Pattern and TQWT algorithms, extracts 6300 characteristics from each audio, with INCA selecting the most relevant 523. Classification of features was done using SVM and LD classifiers, with success rates of 95.32 percent and 93.53 percent, respectively. The collected findings indicated the effectiveness of the INCA-based and CS-Lblock-Pattern system. The data obtained had also been utilized to improve the effectiveness of the ESC-based system.

Audio forensics is the branch of forensic science concerned with the acquisition, analysis, and evaluation of audio recordings. These recordings are frequently used as evidence in legal proceedings. This model could solve environmental sound classification for crime prevention with more research.

RECOMMENDATIONS

The task of environmental or ambient noise identification is crucial in audio classification. The suggested model's success, like that of all machine learning and deep learning tasks, is contingent on a good dataset. Therefore, we gathered a new dataset that is specifically designed for its use. We urge that the volume of data be increased to develop this research further. We've made our dataset public, and we're hoping to gather more data to evaluate the resilience of our method and model before deploying a dependable machine learning application.

REFERENCES

[1] M. Obaid, A. Qahtani, S. Abdulaziz and A. Mazyad, "Environment Sound Recognition for Digital Audio Forensics Using Linear Predictive Coding Features," in *International Conference on Digital Information Processing and Communications*.

[2] V. Boddapatia, A. Petefb, J. Rasmussonb and L. Lundberga, "Classifying environmental sounds using image recognition networks," *Science Direct*, p. 9, 2011.

[3] K. J. Piczak, "Environmental Sound Classification," in *2015 IEEE International Workshop on Machine Learning for Signal Processing*, Boston.

[4] I. Selesnick, "EE Web Engineering," August 2011. [Online]. Available: https://eeweb.engineering.nyu.edu/iselesni/TQWT/. [Accessed 2021].

[5] T. Tuncer and S. Dogan, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Research Gate*, November 2020.

[6] R. Sunil, "Analytics Vidhya," September 2011. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. [Accessed 02 April 2021].

[7] V. Boddapatia, A. Petef, J. Rasmusson and L. Lundberg, "Classifying environmental sounds using image recognition networks," in *International Conference on Knowledge Based and Intelligent Information and Engineering*, 2017.

[8] L. Wang and A. v. d. Oord, "Multi-Format Contrastive Learning of Audio," *Cornell University arXivLabs*, 2021.

[9] J. Sharma, O.-C. Granmo and M. Goodwin, "Environment Sound Classification using Multiple Feature Channels and Deep Convolutional Neural Networks," *Journal of Latex Class File*, vol. 14, 2015.

[10] K. J. Piczak, "Github," Microsoft, 2016. [Online]. Available: https://github.com/karolpiczak/ESC-50. [Accessed 25 March 2021].

[11] J. J. Huang, J. Jose and A. Leanos, "Aclnet: Efficient End-To-End Audio Classification Cnn," *arXiv*, vol. 1, no. 1811.06669, 2018.

[12] Y. Tokozume, Y. Ushiku and T. Harada, "Learning from Between-class Examples for Deep Sound Recognition," in *ICLR*, 2018.

[13] K. Wilkinghoff, "On Open-Set Classification with L3-Net Embeddings for Machine Learning Applications," in *EUSIPCO*, 2020.

[14] D. Agrawal, H. Patil and R. N. Tak, "Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification," in *International Conference on Pattern Recognition and Machine Intelligence*, 2017.

[15] H. B. Sailor, D. M., H. Agrawal and A. Patil, "Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification," *INTERSPEECH*, p. 10, 2017.

[16] A. Kumar, M. Khadkevich and C. Fugen, "Knowledge Transfer From Weakly Labeled Audio Using Convolutional," arXiv, 2018.

[17] D. Hu, F. Nie and X. Li, "Deep Multimodal Clustering for Unsupervised Audiovisual Learning," CVPR, 2017.

[18] R. Arandjelovi and A. Zisserman, "Objects that Sound," arXiv, 2018.

[19] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," arXiv, 2017.

[20] B. Zhu, W. Changjian, F. Lei, Z. Lu and Y. Peng, "Learning Environmental Sounds with Multi-scale Convolutional Neural Network," arXiv, 2018.

[21] D. M. Agrawal, H. B. Sailor, M. H. Soni and H. A. Patil, "Novel TEO-based Gammatone Features for," in *European Signal Processing Conference, EUSIPCO*, 2017.

[22] Y. Tokozume and T. Harada, "Learning Environmental Sounds With end-to end convolutional neural network," in *ICASSP*, New Orleans, 2017.

[23] V. Boddapati, A. Petef, J. Rasmusson and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Elsevier Science Direct*, vol. 112, pp. 2048-2056, 2017.

[24] Y. Aystar, C. Vondrick and A. Torralba, "SoundNet: Learning Sound," in *Neural Information Processing, NIPS, 30th*, Barcelona, 2016.

[25] W. Dai, C. Dai, S. Qu, J. Li and S. Das, "Very Deep Convolutional Neural Networks For Raw Waveforms," arXiv, 2016.

[26] M. Baelde, C. Biernacki and R. Gref, "A mixture model-based real-time audio sources classification method," in *The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, 2017.

[27] A. Kumar and B. Raj, "Features and Kernels for Audio Event Recognition," arXiv, PA, 2016.

[28] L. Hertel, H. Phan and A. Mertins, "Comparing Time and Frequency Domain for Audio Event Recognition Using Deep Learning," arXiv, 2016.

[29] A. Kumar and B. Raj, "Audio Event and Scene Recognition: A Unified Approach using Strongly and Weakly Labeled Data," arXiv, 2017.

[30] F. Medhat, D. Chesmore and J. Robinson, "Masked Conditional Neural Networks for Automatic Sound Events Recognition," *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 389-394, 2017.

[31] S. Mun, S. Shon, W. Kim, D. Han and H. Ko, "Deep Neural Network based learning and transferring mid-level audio features for acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[32] L. Nannia, G. Maguoloa and S. Brahnamb, "An Ensemble of Convolutional Neural Networks for Audio," arXiv.

[33] Wikipedia, "Audio SIgnal," Wikipedia, 27 June 2021. [Online]. Available: https://en.wikipedia.org/wiki/Audio_signal. [Accessed 29 April 2022].

[34] B. Truax, "Handbook for Acoustic Ecology," 2022. [Online]. Available: https://www.sfu.ca/sonic-studio-webdav/handbook/Sound_Signal.html. [Accessed 30 April 2022].

[35] C. S. Department, University of Sheffield, [Online]. Available: http://spandh.dcs.shef.ac.uk/research/asa.html. [Accessed 29 04 2022].

[36] P. Rao, "Audio Signal Processing," Department of Electrical Engineering, Indian Institute of Technology, Bombay, 2007.

[37] S. Oppenheim and L. J, "The Importance of Phase in Signals," *IEEE*, no. 529-550, p. 69.

[38] R. D. Patterson, "Auditory Images: How Complex Sounds Are Represented in the Auditory System," *J. Acoustic Soc. Japan*, vol. 21, p. 14, 2000.

[39] L. R. and F. D. L., "Experiments with a Computational Model of the Cochlea," in *International Conference on Acoustics, Speech and Signal Processing*, 1986.

[40] A. Carriquiry, H. Hofmann, X. H. Tai and S. V. Plas, "Machine learning in forensic applications," *Royal Statistica Societt,* 2019.

[41] D. M. Content, "Audio Feature Extraction," Devopedia, 8 May 2021. [Online]. Available: https://devopedia.org/audio-feature-extraction. [Accessed 25 04 2022].

[42] G. Algan and I. Ulusov, "Image Classification with Deep Learning in the Presence of noisy labels: A Survey, Knowledge-Based Systems," vol. 106771, 2021.

[43] J. Salamon and J. Bello, "Deep Convolution Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, Vols. 279-283, p. 24, 2017.

[44] X. Fan, T. Sun, W. Chen and Q. Fan, "Deep Neural Network Based Environment Sound Classification and its Implementation on Hearing Aid App," *Measurement*, vol. 107790, 2020.

[45] S. Chandrakala and S. Jayalakshmi, "Generative Model Driven Representation in Learning in a Hybrid Framework for Environmental Audio Scene and Sound Event Recognition," *IEEE Transactions on Multimedia*, vol. 22, pp. 3-14, 2019.

[46] T. Tuncer, S. Dogan, F. Özyurt, S. Belhaouari and H. Bensmail, "T. Tuncer, S. DoNovel Multi Center and Threshold Ternary Pattern Based Method for Disease Detection Method Using Voice," *IEEE Access*, Vols. 84532-84540., p. 8, 2020.

[47] P. Rocha, F. Rodrigues, J. Paulo and d. V. Madeiro, "The Issue of Automatic Classification of Heartbeats," *Science Direct*, pp. 169-193, 2019.

[48] M. Fraz, P. Remagnino, A. Hoppe and S. Barman, "Retinal image analysis aimed at extraction of vascular structure using linear discriminant classifier," in *M. Fraz, P. Remagnino, A. Hoppe, S. Barman, Retinal image analysis aimed at extraction of vascular structure using liInternational Conference on Computer Medical Applications (ICCMA), IEEE*, 2013.

[49] M. Hearst, "Support Vector Machines," *IEEE Intelligent Systems.*

[50] S. Keerthi, S. Shevade, C. Bhattacharyya and K. Murthy, "A fast iterative nearest point algorithm for support vector machine classifier design," *IEEE Transactions on Neural Networks*, Vols. 124-136, p. 11, 2000.

[51] T. Tuncer and S. Dogan, "Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals," *Knowledge-Bases Systems*, vol. 104923, p. 186, 2019.