A System to Prevent Social Violence using Convolutional Neural Network

by

Sanzida Akter
19101584
Mostafa Nayeem Omar
18301026
Aanan Ehsan Siam
18101009
Fariha Rahman
19101038
Sadib Tanjib
19101332

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2023

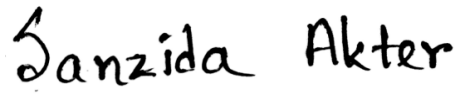
© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

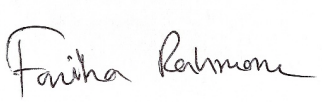
Student's Full Name & Signature:



Sanzida Akter
19101584

Mostafa Nayeem Omar
18301026

Aanan Ehsan Siam
18101009



Fariha Rahman
19101038

Sadib Tanjib
19101332

Approval

The thesis titled "A System to Prevent Social Violence using Convolutional Neural Network" submitted by

1. Sanzida Akter(19101584)
2. Mostafa Nayeem Omar(18301026)
3. Aanan Ehsan Siam(18101009)
4. Fariha Rahman(19101038)
5. Sadib Tanjib(19101332)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 26, 2023.

Examining Committee:

Supervisor:
(Member)

Md. Khalilur Rahman, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

Hereby, we consciously assure that for the manuscript 'A System to Prevent Social Violence using Convolutional Neural Network' the following is fulfilled:

- This material is the authors' own original work, which has not been previously published elsewhere.
- The paper is not currently being considered for publication elsewhere.
- The paper reflects the authors' own research and analysis in a truthful and complete manner.
- The paper properly credits the meaningful contributions of co-authors and co-researchers.
- The results are appropriately placed in the context of prior and existing research.
- All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper references.
- All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences. We all the authors agree with the above statements.

Date: 26-01-2023

Corresponding author's signature:

Sanzida Akter

Mostafa Sayeem Omar

Fariha Rahman

Sadib Tanjib

Abstract

Most women face violence in public and at home, including rape, physical and emotional abuse, mocking, and harassment. A social violence support system might allow people to seek aid from their friends, or relatives, or even request administrative assistance. The goal here is to detect clearly and reliably the screams of the individual in the position that is in any danger, that is, if the scream arose out of dread and horror, based on a particular collection of audios. Screams elicited by dread and panic usually have a shorter length, a higher frequency, and shrill pitches, whereas screams elicited by other emotions or intentionally have a longer duration, a fixed frequency, and pitch. In this sense, if we can use scream recognition to recognize dangerous and consequential circumstances in our society and inform the appropriate individuals at the appropriate moment, we will be able to avert these issues to a degree that will benefit both society and its citizens. To assist the wider populace, we have implemented a system using Convolutional Neural Network to identify screams automatically. This model will assist us in recognizing screams and sending SOS signals or messages to suitable contacts. As a result, people who are in danger will be able to call the people from their selected contacts or general authorities who are within their reach at any time. This system will not only assist victims in avoiding danger, but it will also provide them with a sense of security. On the other hand, the general authority will be able to use this software to limit the quantity of social and domestic violence.

Keywords: Social Violence; Spectrogram; Accuracy; Scream Detection; Support Vector Machine (SVM); Convolutional Neural Network (CNN)

Acknowledgement

First and foremost, in the name of Allah, the Most Merciful, the Most Compassionate, Alhamdulillah, all praise goes to the Almighty Allah Azzawajal, the Lord of the Worlds, and prayers and peace be upon Muhammad, His servant, and messenger. Allah's assistance and blessing in providing me with the opportunity and strength to carry out and accomplish the entire thesis work titled "A System to Prevent Social Violence using Convolutional Neural Network."

Second, we would like to thank our supervisor, Dr. Khalilur Rhaman, for his encouragement and assistance during our project. Without his encouragement, this study would have been insufficient.

Third, we would like to thank our parents for their unconditional support throughout our lives, particularly during the process of seeking a degree; it is only because of their undying love and prayers that we had the opportunity to complete this thesis.

Moreover, We would like to thank one of the Bracu Graduates Moh. Absar Rahman (Student ID - 22341105), and Sayantan Roy Sir for cooperating with us in our work.

Finally, we would like to express our gratitude to the researchers for providing us with feedback on their articles, which has helped us greatly in our subsequent work.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

API Application Programming Interface

CNN Convolutional Neural Network

DBM Deep Boltzmann Machines

DNN Deep Neural Network

GMM Gaussian Mixture Model

HMM Hidden Markov Model

LPC Linear Prediction Coefficients

$MFCC$ Mel-frequency cepstral coefficients

MVD Meta Vocal Dataset

NLU Natural Language Understanding

RMS Root Mean Square

$STFT$ Short-Time Fourier Transform

SVM Support Vector Machine

VAD Voice Activity Detection

ZCR Zero Crossing Rate

Chapter 1

Introduction

1.1 Classification of Violence

For both developed and developing countries, social violence is recognized as a major worldwide public health concern. Despite recent improvements in poverty and crime, Bangladesh is the most overpopulated country, with a population of over 169 million people. The country has seen a rise in social violence statistics over the past few years, and experts have attributed this rise to the growing economic disparity between the rich and poor. However, social and domestic violence against not only women and girls but also men and boys is a serious problem in Bangladesh.

Nonetheless, violence not only leads to physical but also sexual and psychological abuse, which imposes a significant toll on society in addition to the human cost of these occurrences. As an added cost, violence has a significant impact on healthcare, the criminal justice system, social services, and local economies. This classification of violent acts helps make sense of the many ways in which violence manifests itself in people's daily lives, from the home to the street. Although, Bangladesh has made significant strides in reducing crime over the past few years due in part to increased police presence and awareness campaigns aimed at reducing crimes such as domestic violence and assault.

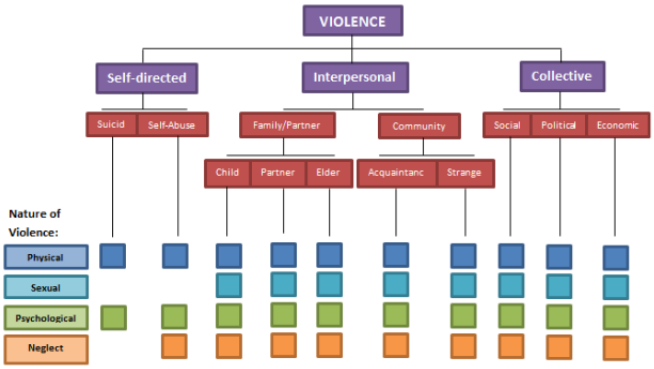


Figure 1.1: Classification of violence. [1]

[1] https://www.saferspaces.org.za/understand/entry/what-is-violence

1.2 Statistics of Violence in Bangladesh

Bangladesh has the highest rate of domestic violence in the world. Bangladesh has a domestic violence rate of 37.7 for every 1,000 women aged 15 to 49, which is higher than the worldwide average of 31.8 for every 1,000 women in this age group [3]. From the research, the violence rate has increased by 5% since 2009 and the country's government must gather information on how to prevent this from occurring. The domestic violence rate in Bangladesh has increased by 20% over the past decade. The number of reported incidents of domestic violence increased from 4,770 in 2012 to 5,941 in 2017. In addition, the number of women who reported experiencing violence increased from 13% to 16%. In addition to this increase in frequency, there was also a change in the type of abuse: while previous studies showed that physical abuse was the most common type reported, this one found that psychological abuse is now more prevalent than physical abuse. According to UNICEF [1], an estimated 1 in 3 women and girls experience violence, but only 2% of perpetrators are brought to court.

1.3 Problem Statement

Firstly, we need to know the term Social Violence, What does it mean? Why does it occur? Basically, social violence occurs when a person or group of people use their power to bully, intimidate, and threaten others. The goal is usually to get others to do what they want them to do, but this can lead to physical violence if it continues. Violence is a phenomenon that occurs on a global scale and is responsible for the lives of more than 1.6 million inhabitants annually. As a consequence, it is one of the significant causes of mortality around the globe. People who experience social violence often feel powerless to avoid it or protect themselves from its consequences. They may feel unsafe, stressed, and even depressed due to their inability to control their environment. Even though violence may occur everywhere, the great majority of its victims are killed or abused most brutally, even though no nation is immune to the problem of these violent crimes.

Sexual and gender-based violence are two types of social violence against women that are often used interchangeably. This kind of abuse is sometimes classified as a hate crime since it is directed against women and girls for being feminine. There is a long and troubled history of social violence against women, while the prevalence and rigidity of such acts have evolved throughout time and continue to differ across modern communities. In both interpersonal and social contexts, violence is being used to subjugate women. Aggression against women may be a symptom of the aggressor's underlying aggressive personality.

Another term for violence in Bangladesh is sexual harassment. Women in our country are being harassed in many other ways. The most significant areas where women are being harassed are public places, shopping centers, buses, the street, institutions, stations, and so on. Nowadays, it is difficult to find a place where harassment does not occur. "Eve-Teasing" is the terrifying form of sexual harassment that women face the most. Eve-Teasing has been a difficult crime to prove since attackers fre-

quently come up with inventive ways to attack women, which typically happen in public spaces, streets, and public transit. Eve-Teasing has evolved into an often brutal case of sexual harassment that can cause permanent physical and psychological damage to a girl's life.

Within the realm of scientific research, there is still a lot of ambiguity over where exactly violence against women originated. This is partly owing to the underrepresentation of several types of violence against women, including rape, sexual assault, and spousal violence, which is continually caused by cultural morality, taboos, smirch, and the sensitive nature of the subject. The reason for this underrepresentation is incompletely due to the fact that rape, sexual assault, murders, dowry abuse, female genital mutilation, forced marriage, and spousal violence are particularly prevalent types of violence against women. As is well recognized at the present time, drawing a clear picture of violence against women is difficult because there is a lack of safe and harmonious data.

Women in Bangladesh are often not treated with the same respect as males while they work. Women experience a great deal of mental and emotional strain as a result of the many obstacles they must overcome in order to successfully perform the many roles that are expected of them at the same time. The most prevalent kind of physical assault that may be seen in Bangladesh is that which occurs inside the home. The majority of individuals are aware of what constitutes domestic violence, which may include gang rape or violence connected to dowries, underage marriage, conjugal rape, abusive behavior, irritation, humiliation, and other forms of violence.

1.4 Scenery of Violence in Bangladesh

Physical abuse is one of the most horrific types of violence that has a significant deterrent effect in Bangladesh. Physical and mental health aren't the only things at risk when force is involved; a woman's sense of self and her future in society are also at risk. According to the research [19], the standard deviation for the annual number of victims of assault is 71, and the average annual number of victims is 750. There is a possibility mentioning the fact that child race is currently a major societal problem in Bangladesh. The percentage of children among the victims consistently exceeds 50%. The alarming rise in the murder rate among assault survivors over the last few decades is a major concern for modern civilization. Despite the fact that there have been fewer reported cases of gang rape in 2018 compared to the previous five years, the crime is nonetheless unacceptable. Before 2014 shown in figure 1.2, the number of rape victims was nearly consistent, and before 2018, the number of rape victims remained roughly stable for three years. Additionally, statistics show that the percentage of rape victims has ranged from 12% to 16% throughout the years, with the lowest recorded rate being in 2018 (12.1%).

Bangladesh's law enforcement has seen a shocking increase in rapes in recent years. As can be seen in figure 1.3, the Odhikar statistics unit (2001-2015) has recorded 77 cases of rape by law enforcement over that time period, with the number of such cases increasing by 14.29% from 2011-2015 and by 18.19% from 2001-2005. In

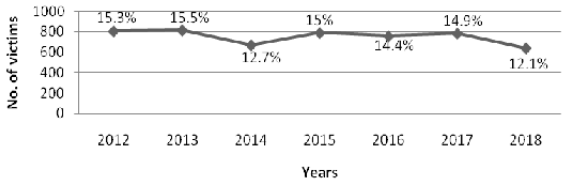


Figure 1.2: Trend of victims by Rape. [2]

second place on the list of most prevalent forms of violence against police is rape. Women are the most likely to be victims of law enforcement misconduct, despite the fact that they account for 66 percent of the total rate of misconduct perpetrated by law enforcement agencies. This is something that can be expected [7].

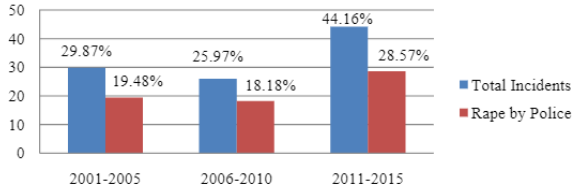


Figure 1.3: Rape by Law Enforcement Agencies. [3]

Acid violence is another horrifying kind of violence that became more common in Bangladesh between the years 2001 and 2005. This crime illustrates the use of acid as a method of retaliation by rejected suitors beginning in 1980. The number of people who are injured or killed by acid attacks in Bangladesh is now declining at an increasingly rapid rate. Acid Survivors, a well-known non-governmental organization in Bangladesh, has a documentation section that has recorded a total of 2,898 occurrences of acid attacks between the years 2001 and 2015, with a total of 3,254 victims. Figure 1.4 demonstrates that the proportion of acid sufferers from 2001–2005 and 2006–2010 was more than four and two times greater, respectively than the proportion of acid sufferers from 2011–2015 [7].

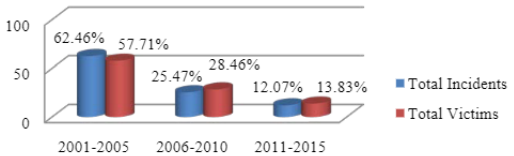


Figure 1.4: Acid violence against Women. [4]

According to Odhikar's research, between 2011 and 2018, a total of 2617 girls were targeted by stalkers, while 144 women were targeted for speaking out against sexual harassment. In addition, the rate of eve-teasing was at its maximum in 2011 and at its lowest in 2015, a comparison of the rates of eve-teasing over the last eight years. Based on figure 1.5, it would seem that the number of incidents of sexual

harassment is decreasing over time. The annual number of victims who are targeted by stalkers is an average of 278, with a standard deviation of 106. On a yearly basis, there are around 14 females who commit suicide as a direct result of being subjected to sexual harassment, and there are approximately four females who are murdered by stalkers. From 2012 to 2015, the percentage of people who were stalked and/or murdered by a Stalker decreased from 24.6% to 9.8%. After then, the number of victims rapidly increased to 13.9% in 2016, but after that year it started to gradually decrease again. However, after 2016, it started to steadily decrease once again [19].

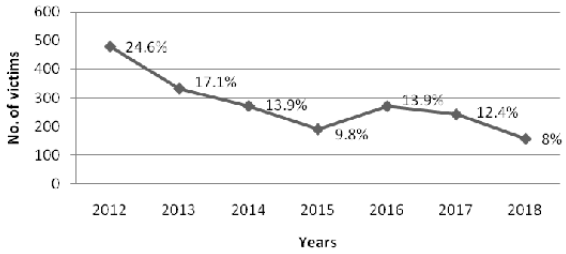


Figure 1.5: Trend of victims by stalkers. [5]

Additionally, our country has been cursed by the dowry system, but unfortunately, it is still widely practiced, particularly in rural regions. Between the years 2001 and 2018, there were a total of 5756 married women who fell prey to the dowry system. Among those women, 3273 were murdered, 2250 were subjected to physical violence, and 233 took their own life. When compared to the previous 17 years, the number of violent incidents that may be attributed to dowry was at its greatest in the year 2012. According to figure 6, this kind of violence had a dramatic reduction between the years 2012 and 2015, falling from 35.7% to 8.8%. Then, it went up somewhat in 2017, but then it went back down to 6.2% in the next year. There are an average of 329 victims of dowry-related violence per year, with a standard deviation of 236. [19]

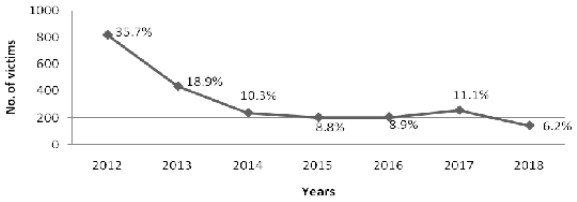


Figure 1.6: Trend of victims for dowry-related violence. [6]

1.5 Factors Contributing to Domestic Violence in Bangladesh

When women have greater opportunities to work outside the house, they are more likely to be victims of sexual assault and other types of harassment or intimidation. Regardless of their age, workers should expect to be harassed while traveling to their place of employment. Commuters, rickshaw drivers, shopkeepers, and bus passengers all contribute to the problem by making or listening to sexually provocative

remarks and whistling throughout the day. According to the collected data, the conductor and drivers of public transport acted inappropriately for around 70% of the female garment workers who traveled public buses. The other side of the coin is that bus passengers, particularly older men, have been known to push, shove, pinch, and make provocative remarks. Also, this violence includes non-verbal types of harassment that involve gazing, blinking, blowing, standing extremely near, and squeezing. There is an increased risk of violence against women working night shifts. In addition, seldom do incidences of workplace or commuter-related violence make it into the public eye, and few organizations keep a comprehensive record of such violence. Every single day, our nation is witness to an alarmingly high number of instances of this kind [7].

The United Nations Population Fund (UNFPA) ranks Bangladesh fourth among the world's countries in terms of violence against women. More than 60% of Bangladeshi men keep their wives from participating in social activities, while more than 65% of Bangladeshi men consider it acceptable to abuse physically their partners. In addition, the International Commission of Jurists (ICJ) notes that despite the fact that in 1997 the UN Convention on Elimination of all forms of Discrimination Against Women (CEDAW) Committee issued a statement expressing significant concern about the capacity of the government of Bangladesh to implement effective legislation to protect people from violence, as present situation has not changed and violence against women is still occurring undoubtedly. Mahila Parishad, a leading Bangladeshi women's group, released a study in which they estimated 3,625 women were assaulted throughout the country [4].

Moreover, domestic violence is a serious problem in both the country's rural and urban regions, affecting people of all socioeconomic backgrounds. It seems that, in the majority of these situations, the husband is the one who is guilty of the act of violence that was committed against the wife. It has even developed into a culture of acceptability among Bangladeshi populations, and this culture is being passed down from one generation to the next to the point that it is being formalized. Society as a whole learns to accept the enslavement of women, which has devastating consequences. It is believed that metropolitan women are less likely than rural women to experience domestic abuse because they are more aware of their rights and have easier access to human rights organizations. However, the likelihood of domestic violence victimization for a woman living in a rural region is the same as for a woman living in an urban one according to World Health Organization (WHO) in 1997. On the other hand, people tend to be more involved with and hierarchical within their societies in rural regions. In contrast, persons living in metropolitan regions often have little idea of what is happening in their neighborhoods. According to UNICEF, it happens because families that live in urban areas only have irregular interactions with one another. In situations like these, information regarding instances of domestic abuse may be difficult to get, and as a result, it is often disguised [4].

1.6 Survey on Violence Report

For better knowledge, we did also a survey on social and domestic violence to get to know the story by ourselves that how women and men are being treated in our society. From the survey, we got to know that 53% of total people have faced violence at least once in their life. Additionally, more than 27% of people got mugged or hi-jacked in their lifetime. We surveyed all ages, and we learned that most people especially women who are facing or have faced any kind of violent acts are in their 20s and 30s. They're also more likely to be single than married, which is interesting because it means that they may not have found partners yet who can protect them from predators. Also, the couples who are married or in a relationship, 17% of people get slapped by their partners. Home, pavements, bus or train stations, schools or colleges, and public transport are the most common place to get harassed in Bangladesh as we learned from our survey. Where 78% are unmarried and 22% are married.

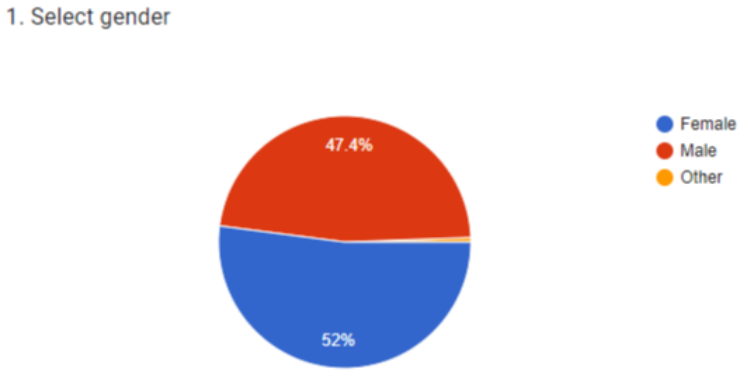


Figure 1.7: Percentage of Male and Female facing Violence

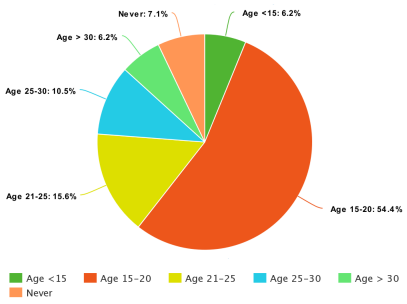


Figure 1.8: Percentage of Age (male and female) facing Violence

According to the results of the survey, women in Bangladesh face twice the risk as men. The humiliation takes place behind the locked doors, while the rage of the crowd can be heard outside. Women who have been the victims of domestic abuse are increasingly seeking help from parents or the legal system and the police. However, even if their requests for protection are granted, they are greeted with nothing

but scorn when they go back to their homes. Similarly, despite all the negative connotations associated with it, dowry remains a hallmark of marriage and one of the leading factors contributing to intimate partner abuse. Since dowry killings are less serious than murder, authorities have less incentive to investigate and punish them. This sort of violence is kept alive in our culture, mostly because of religious shame and traditional cultural traditions. As a result, the victims do not speak a single word because of society. This results in the silence being maintained. The vast majority of victims come from economically disadvantaged backgrounds, and as a result, they have very limited resources to devote to protracted legal fights. The only thing worse than being tortured or beaten by a spouse or in-law is having to go to court or ask for police protection in similar situations. Therefore, most women would rather bury their anger and die in silence than talk about their traumatic experiences. Despite this, society has always remained a dynamic organism incapable of change and development. It is impossible to make progress if half the population is kept in complete darkness. It's important to remember that the man and woman who brought us into this world are ultimately responsible for our existence. If women can fulfill the roles that are expected of them in society, then society will continue to advance. If not, then social development will be halted. This means that women, just like men, deserve respect and admiration in our communities. They may require some strength to find out how to escape the societal tyranny that they are experiencing. Therefore, their power must originate from the community as well as the administration.[4]

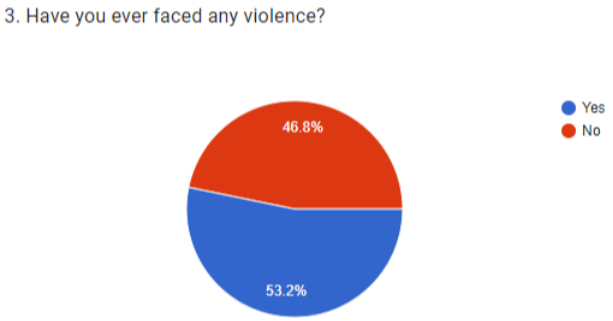


Figure 1.9: Percentage of People faced Violence Once in Their Life

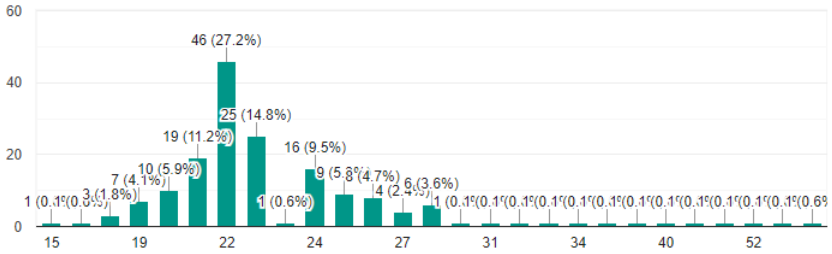


Figure 1.10: Present Age of People who responded in the survey

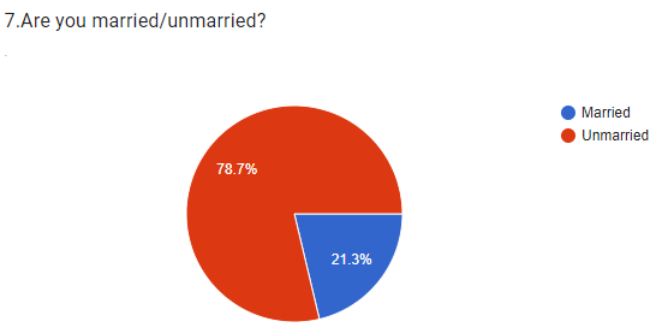


Figure 1.11: Percentage of People who are married and unmarried

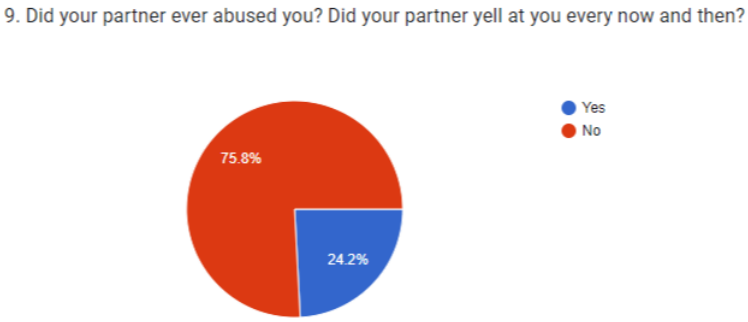


Figure 1.12: Percentage of Married People abused by Their Partner

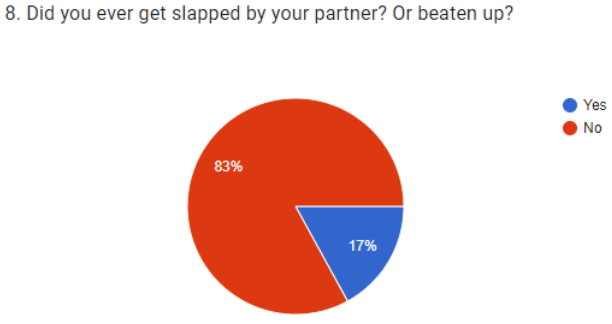


Figure 1.13: Percentage of People facing Domestic Violence

1.7 Aims and Objectives

In this research, we aim to create a system that will be used by people of all genders around the country. We want to build a system that will help victims of domestic and social violence by providing them with a platform where they can not only call for help but also record the audio as evidence on their devices, Also, the system will be able to call or send messages for help when they're in danger, even if they can't touch their phones since this system will be an automated system based on scream detection. It will detect any screams and automatically start recording in the background, so the user does not have to touch the device. Additionally, in case of any emergency, people can call their parents or the police when the saved code word will be called more than 2 times. Moreover, a victim can also use some special keywords to activate the system and get emergency help through it.

We believe in the power of technology to make the world a better place. Our goal is to create a platform where victims of domestic abuse or sexual assault can report incidents to law enforcement, but also where their abusers can be reported as well. We think that this is the first step toward ending both forms of violence by putting everyone in the loop. In addition to reporting incidents, we hope to provide a way for victims and survivors alike to have access to resources that could help them heal from the trauma they have experienced.

Chapter 2

Background and Related Work

2.1 Violence Safety System

According to research, an Android app called Abhaya assures women's safety [5], it helps people who are in danger by recognizing the location of an individual. Initially, the application user has to undertake the contact numbers of police, family members, and friends into the app and store it. The user must launch the application while on the go and click "start" whenever the situation calls for it. As soon as the application launches, it calls the first registered contact number saved in the system and sends a message to every contact number with the victim's location. This app's distinctive feature is that until the stop button is selected, a message with a location is continuously sent to the registered contact numbers every five minutes. Therefore, this program allows for continuous victim location tracking.

Women in society are provided with a safe and secure atmosphere with another app named WE'RSAFE App [9]. For anyone who is considering committing a crime against a woman, this app will prevent it, and the rate of crime against women will be minimized eventually. This app will operate as a weapon for women, ensuring their protection and security, and it will work on any smartphone using the Android operating system. The first page is Splash, which is a screen that displays the app's logo for a few seconds before loading the components. On the phone's screen, the same logo appears as an icon. The Login and Signup buttons are located on the Sign-in Page. The Login button allows already registered users to access our app, while the Signup button allows new users to create an account and access our app for the first time. The Signup page is a screen that asks new users to fill out information such as their username, password, phone number, and email address. If he or she is logged out of the app after becoming a registered user, only the login page will appear. The Key Page is a screen that displays the app's main functionality, such as calls, recordings, and alarms. When the user presses the Emergency Call button, an emergency contact number that the user has saved is automatically dialed. When the user selects the "Available Hospitals and Police Stations" button, an offline list of various places appears. When you select a location, a list of surrounding hospitals and police stations is provided individually. Update profile, add or modify contacts, tips feed, and other choices are available in the menu. The "Add or Modify Contacts" option is a screen that asks for the user's name and phone number. The user can view, add, amend, and delete information. With more research and

development, the concept could be put on a small wearable device such as a watch, pendant, or bracelet that uses GSM and GPS modules. When this app is triggered, the location is collected by the GPS module and decoded into a legitimate link in Google maps, which will be sent through text message to the enrolled relatives and friends.

Moreover, another research proposed a hardware-software-based system for preventing Social Violence [12]. In this method, a user has to use a safety band along with the safety app. Users can wear the safety band on their arms. The software will work when the trigger button is pressed. There is a button named "Danger"; When a user pushes the button, the local area network is activated and the script file is opened. The server will check if the script file is on or off. When the script file is turned on, the GPS and GSM module, generated by a microcontroller, will send a help message along with the location to the previously saved emergency contacts as well as to the closest police station. After receiving the message, the police will press a button that says that the message is received and then the LED light of the band will receive the notification, eventually it will be turned on. If the victim has been rescued, the police can press the "Final Acknowledgement button" which will turn off the light and send a rescue message to the victim's emergency contacts.

Furthermore, there is an application of Machine Learning along with scream detection, speech recognition, and various deep learning techniques. This study aspires to develop a sound automated system that uses long short-term memory and support vector machines (SVM) to identify any casualties in fire crises [17]. This research is for creating a custom dataset of raw audio files that pass through a sound preprocessing block, followed by an audio feature extraction block to extract necessary features. Furthermore, the t-SNE algorithm enhances a similarity measure between occurrences in high-low dimensional spaces. As for the feature extraction process, the model used a sample rate of 16 kHz and 16 features were extracted per 32 into ms frame duration. Moreover, 60% of the data from the dataset were used for training, more than 10,000 samples for validation and the remaining audio was tested. The implementation of the LSTM network is well suited because of its outstanding iterative time series recurrent classification, and the fact that SVM is a lot simpler model, making it acceptable for use in AESV as this application works with audio recordings. In noisy situations, LSTM networks outperformed DNN and Convolutional Neural Networks (CNN).The trials involved calculating the average of the greatest 10% ZCR across each audio, which produced an input shape for SVM classification and helped to limit the number of features to prevent overfitting. The LSTM cell also has an output gate, a forged gate, and an input gate. The input gate participates in updating the state of the cell, and the output gate updates the value of the hidden unit. To design a noise-resistant model for identifying screams, a dataset with nine different levels of SNR ranging from -10 dB to 30 dB was also created, and the analysis sections for SVM and LSTM displayed the outcome of a combined dataset that includes all the SNR levels. Due to its greater scream prediction accuracy, performance stability in low SNR, and computational economy needed while running on autonomous embedded systems, the SVM is taken into consideration for the final experiment.

A system called "ME TOO APP" also works as a social violence help app [13]. In this work, there are many features of the suggested system. For example, A user can make an online complaint through the system. The main objective of the online complaint management system is to make complaints easier for monitoring, tracking, and resolving, as well as to provide the company with effective tools for recognizing problems. The system can track a user's location since tracking is necessary to notify relatives, and friends, via GPS. Scream, however, it is highly unlikely to make phone calls in a crisis, there are some situations where users can simply click on a button to get help. When it is initiated, a screaming alarm makes a loud noise. In most cases, this button works as a panic button. The alarm is a loud noise that alerts others that someone requires assistance. Moreover, it will include contact information as well as a voice keyword. The user makes a contact list of people to whom he or she wants to ask help, and a keyword or voice is stored for recognition purposes. The database is stored in mobile memory.SQLite is the database used. At least 2 contact numbers should be included in the database. Voice recognition: the voice recognition module identifies the keyword spoken by the user. Then the victim's spoken keyword will be compared to the registered keyword. This keyword will be found in the converted text. If the keyword is found, the message will be sent, A location is searched using position information, and an actual address is provided via message. The Internet is needed for the process. When GPS is turned off, the system will not be able to pinpoint the user's exact location, it will only send the location's position information. There will be a message-sending module for the user. The GPS will obtain longitude and latitude coordinates. A previously saved emergency message will be sent to registered contact numbers, along with the user's longitude and latitude, as well as his or her exact address. If the user's mobile network is not available, the message is queued and sent when networks become available. when a message is sent, it is followed by a notification.

2.2 Scream Detection and Audio Classification using Different Classifiers

For scream detection technology, this research prefers a system to see if a trained machine learning can recognize screaming noises in audio streams captured in the house [16]. Children with autism and other behavioral disorders frequently give erroneous accounts of their behavior. A scientific approach to assessing behavioral issues would eliminate the need for these erroneous reports. The goal of the research is to recognize human screams in a steady audio recording from within a house. The model trained the classifier on an audio database, then used some of the TV shows to approximate clinical data for validation. Their training data is made up of a few seconds of sound snippets taken from YouTube and portions of the AudioSet data collection. Audio recordings of families are the model's target use case. Episodes of the TV shows were utilized as stand-ins in the absence of such recordings. Scream or non-scream was assigned to every second of audio data. If they are generalizable, classifiers for clinical recordings could learn from open databases. Post-processing helps to reduce false positives and isolate the true episodes of screaming. A public database is used to train a scream-detecting classifier for clinical audio recordings. If shouting information from audio is utilized to segment a home video feed, the

resulting segments can be helpful for diagnosing behavioral problems or managing illnesses. A portion of the publicly accessible Audio Set data set and a set of audio data taken from the TV show, which was chosen for its closeness to clinical data, were the two sets of audio samples that were created to test the model. Scream events were manually annotated for the TV show audio data, and already-existing annotations for the AudioSet data were improved. A convolutional neural network that has been previously trained on AudioSet was used to extract audio features.

Another study aims to follow various feature and classifier combinations [8] for audio detection where Deep Neural Networks propagate the mechanism, which consists of a network of individual cells known as units, and the units reflect unknown causes or factors in the input data. In the proposed study, the data were manually separated into four main sound categories: Scream, Shout, Conversation, and Noise. Although, there can be a lot of noise in the Conversation, Scream, and Shout classes. These classes are distinguished from the noise class by the absence of voice or vocal signals. As for the dataset, 33 Screams, 292 Shouts, 997 Conversations, and 1429 Noises were used for training and testing. The test set consisted of 17 Screams, 92 Shouts, 207 Conversations, and 396 Noises. The authors employed MFCC coefficients of the 12th order, which are also energy terms, and concentrated on MFCC along with energy vectors every 10 consecutive frames. Furthermore, the authors have used the library of Python's 'pdnn' to implement Deep Belief Networks and Deep Neural Networks based on the 'theano' framework where the Deep Neural Network is trained using the same data as the Deep Belief Network, and experiments are conducted with both networks. The authors accomplished a 6.2 percent of error rate for Scream versus Shout and a 6.5 percent error rate for Shout versus Noise in the classification process. However, in the three-class problems such as Shout-Conversation-Noise, the authors achieved a 28.7 percent error rate. Point to be noticed that, 21.6 percent of the Conversations are correlated with Noise, also, 34.8 percent of the Shouts are correlated with Conversations.

A work proposed a system that uses a self-recorded scream-based database by using MFCC with 100% accuracy [6]. In this work, the use of MFCC as features and deep learning for classification is proposed for classifying normal noises from scream noises. The study shows MFCC feature extraction techniques and procedures to alter the features into an input vector followed by Deep Boltzmann Machines (DBM). In the dataset, there were two kinds of sounds: Scream 'ah ' and simple 'ah' which are self-collected datasets that are captured at a sample rate of 16 kHz with 16-bit accuracy. The audios of the screaming dataset hold several screams of 60 different people of every age which are recorded in various environments. Altogether, the dataset contains 240 audio of which 130 of the audios are scream sounds and the rest 110 audios are normal sounds. For training, 92 screams and 78 non-scream sounds are used, whereas, for testing, 38 screams and 32 regular audios are employed. Furthermore, the authors applied MFCC technique to extract the features from the dataset as the usage of logarithms and Mel scale for transformation in MFCC provides a distinct advantage. The Hamming window was placed at 25 ms and the sampling rate was placed at 10 ms. A set of N vectors represents the output feature matrix extracted from the audio input stream. Furthermore, the features gathered are extremely diverse and can aid in the classification of objects with great precision.

After extraction of the features, the sounds were normalized to the neural network which can give accurate results. The suggested system comprises two visible levels and two secret layers with 500 and 1000 hidden units each. During the evaluation procedure, the dataset is permuted at random from the original dataset, creating five separate datasets. Then the system is assessed across all datasets. The tests were carried out to produce the MFCC's smallest yet most diversified feature matrix as the time utilization increases dramatically as the input vector size grows, it is critical to reduce the input vector size. As a result, it is found that as the batch size is reduced, the accuracy decreases. A normal classification approach using the same dataset was also used to further assess the system. The GMM algorithm is used for extracting the features, and SVM is used to classify in their study. After comparing performance to the suggested model, the best-performing parameters were chosen, and five randomly chosen datasets were examined on the systems. Furthermore, the proposed model was found to be 100 percent accurate.

Moreover, in another study of classifying heavy metal music, the topic of detecting and classifying extreme vocal approaches in heavy metal music, specifically the identification of diverse scream styles [18]. The key contributions of this work include a carefully annotated dataset of over 280 minutes of heavy metal songs from diverse genres, as well as a systematic exploration of multiple input feature representations for the classification of heavy metal vocalists. In this study, scream detection and localization in urban sound, scream detection in subways, scream and shout recognition in noise, and scream detection for home applications are all examples of previous work in this area. However, the authors have used the Metal Vocal Dataset (MVD) which consists of 57 old songs from different genres such as groove metal, death metal, progressive metal, black metal, and metal core. About 70 percent of the data from the datasets are used for training, 15 percent for testing, and the rest for calculating accuracy. Both the training and test or validation sets of songs from the same bands were eliminated. Based on the perceived sound, the authors divided fry screams into three categories: high, mid, and low. Sing, High Fry scream, Mid-Fry scream, Low Fry scream, and Layered scream were the class designations assigned to the vocal events. The authors have divided the sample audios into 5 feature sets and each audio block was classified using two multi-class classifiers based on the feature vector. The authors achieved the greatest performance and perform similarly with recall above 82 percent after applying all of these characteristics with SVM and using CNN the results show that the spectrogram input can detect the existence of screaming with an accuracy of 87.6 percent, which is a significantly higher than any SVM-based technique. It appears that CNN can effectively recognize spectral patterns by utilizing the information in the spectrogram.

In addition, different audio classifications are possible using a different classifier, for example, Support Vector Machine (SVM). Many steps are taken in the experiment of a frame-based and multi-class SVM for audio classification [2]. The audio feature of this model is integrated with perceptual features such as LPC, LPCC, and MFCCs to produce an audio feature set. The accuracy rate was 91.7 % for the proposed system. Firstly, they extracted the audio signals according to their necessary features. Then they used some algorithms for the classification of the audio snippets. Before this work, there are many other features and methods proposed

by different authors. For example, Foote brings the use of MFCC plus energy as an audio feature that follows the NN rule. Li, on the other hand, describes how the perceptual and cepstral feature sets are used for audio classification. For audio classification, the same author introduces a novel classifier called the nearest feature line (NFL), which is more efficient and produces better results than NN-based classifiers. Using the same feature, the author eventually employed SVM with a binary tree structure to recover audio categorization difficulties. To increase audio classification accuracy, the recommended frame-based multiclass SVM employs the basis of means and standard deviations of each individual feature in an audio file. An innovative audio feature based on ICA, as well as overall spectrum power, brightness, bandwidth, pitch, and MFCCs, are utilized to choose features. For the computation of MFCCs, the log powers of the critical-band filters are determined first. The authors also applied Discrete cosine transform (DCT) to the log powers to work with the correlations. To run a new audio feature, ICA transform is used instead of DCT to the log powers. A mel-frequency ICA-based feature is used for the audio classification among all the critical-band log powers. For the result, four different sections are used - news, advertisement, sports, and music.LPC, LPCC, and MFCC features are extracted for the signals and they are used as input in SVM. SVM is a modeling technique used in this work for audio classification. The method shows 94% accuracy in this experiment.

On the other hand, the use of Convolutional Neural Networks (CNNs) has shown great victory in image classification and has potential in the audio domain. In this study, the soundtracks of a dataset have been categorized consisting of 7 crores in training films (5.24 million hours) using a variety of CNN architectures. The dataset contains 30,871 video-level labels [10]. They take a look at four different kinds of fully linked Deep Neural Networks (DNNs)—AlexNet, VGG, Inception, and ResNet.

Convolutional Neural Network (CNN) architectures like AlexNet, VGG, Inception, and ResNet have led to significant improvements in image classification performance on large datasets like ImageNet.In this study, it has been figured out whether CNNs trained on similarly large datasets can achieve satisfactory results when applied to audio classification issues. This study uses the YouTube-100M dataset to inquire into the following questions: how do well-known Deep Neural Network (DNN) architectures perform on video soundtrack categorization; how do the results change when the train set and labeling vocab sizes vary; and to what extent do the resulting models may be used to automated external defibrillation (AED). There are a total of 100 million videos in the YouTube-100M data set, including 70 million used for training, 10 million used for evaluating, and 20 million used for validation. There are 5.4 million training hours worth of videos with an average length of 4.6 minutes apiece. There are 30,871 distinct labels used to categorize these films into their respective subject areas. Frames of 960 milliseconds (ms) duration are used to separate the sounds. About 20 billion instances were generated from the 70 million videos. The metadata of the video's parent is copied into each individual frame. A short-time Fourier transform using 25-ms windows every 10 ms is used to deconstruct the 960-ms frames. The final spectrogram is integrated into 65 frequency bins that are separated by miles, and the magnitude of each bin is log-transformed after a tiny offset is added. This is done to eliminate any numerical difficulties that may

arise. TensorFlow was utilized throughout all trials, with the Adam optimizer being used to conduct the training in an asynchronous fashion across many GPUs.One million videos were utilized for the 30K labels, 100,000 films were used for the 3087 most frequent labels and 12,000 videos were used for the 400 highest frequent labels; all three evaluation sets included around 33 samples per class. The authors used the classifier on every frame across the duration of the assessment videos (960 ms). The scores generated by the classifier were then averaged over all video segments. The research team used a completely integrated DNN as their standard of comparison before testing out different networks that were closely based on existing, effective picture classification systems. Researchers trained all network topologies using 3000 labels and 70000 movies, then compared results after 5,000,000 mini-batches of 128 inputs. Some networks train more quickly than others, therefore comparing after a set wall-clock period will provide somewhat different results without altering the relative ranking of the designs' performance. Data for ResNet after 405 hours of training (or 17 million mini-batches) is included to demonstrate the model's steady performance gains. After 13 million mini-batches, they slowed the learning rate down by a factor of 10. The availability of such a massive training set enables them to study the effects of training set size on accuracy. Experts have around 20 billion 960 milliseconds of training examples thanks to the 70 million films and the average of 4.6 minutes for each movie. With 20 GPUs and ResNet-50's training speed of 11 mini-batches per second, the network would need 23 weeks to see each pattern once (one epoch).

In another work, Convolutional neural networks (CNNs) excel at classification problems when high-dimensional sensory data (like images or audio) is provided as input. End-to-end learning, whereby the raw pixels of pictures are fed directly into the CNNs, has become a common method in image categorization. In contrast, most CNN-based models for audio classification still rely on spectrogram-based inputs like Mel-spectrograms, which need substantial handmade architecture in the time-frequency transform [15]. For use in music categorization tasks, for instance, Mel-spectrograms with 128 bins are often used. More often than not, 80 bins or fewer are employed in voice recognition and ambient sound categorization tasks. Waveforms were used to train CNNs classifier for music tagging, and the results were compared to mel-spectrogram input. In order to recognize speech from waveforms, this study used Convolutional Long short-term memory Deep Neural Networks (CLDNNs). New convolutional neural networks (CNNs) with tiny filter sizes were suggested by the authors. They named the CNN architecture they developed, which improves performance in music auto-tagging, SampleCNN since the initial convolutional layer accepts a tiny grain of samples instead of a standard frame-level size. With the property that untrained model performances are strongly linked to trained model performances, the authors also conducted a thorough assessment of CNN models for (music) audio classification. They demonstrated that SampleCNNs outclass other waveform-based CNNs that used large filters. In this study, researchers undertake extensive experiments on three distinct audio classification tasks to explore SampleCNN and its expanded architectures. The tests compare the performance of several expanded versions of the SampleCNN design and concentrate on the architecture itself for analysis. The findings confirm SampleCNN's efficacy and demonstrate that the SE block is, on average, preferable for the audio

categorization tasks.

Besides, SampleCNN's structure in light of spectrogram-based CNNs and WaveNet has been examined. The models that came out on top in the ImageNet competition marked many significant turning points in the advancement of image categorization. While VGGNets use 3*3 filters for each layer, SampleCNN uses 1D filters, convolution, and pooling. Since ResNets and SENets were built upon VGGNets, it is only logical to include their individual modules in SampleCNN to achieve even greater audio categorization gains.

In this research, in order to train a highly deep CNN with skip connections, Residual Networks (ResNets) are used. Backpropagation of gradients in neural networks is made smoother by the use of shortcuts, which even make it possible to train a ResNet with 1001 layers. Their standard building block is called a Res-n Block, and it has one skip connection. When n is 1, researchers just deal with the first convolutional layer, and when n is 2, they only deal with the second. The SE block is a supplementary architectural element that can be incorporated into existing plans. The squeeze and excitation processes are the two basic components. During the squeezing process, a statistic for each channel is extracted; this statistic is known as the channel-wise statistic. It is implemented using a pooling layer that takes into account the global average over time. The weight is determined for each channel using the channel-wise data as inputs in an operation known as excitation. Through the two fully-connected layers, the weights are learned. As one dives further, one often experiences excitations that are more specialized to one class than previous ones. To maximize the benefits of both blocks, the ReSE-n block is constructed by fusing the Res-n block with the SE block. The SE block may connect to the Res-n block in a number of different places.This research aims to validate SampleCNN's generalizability and efficacy by testing it across three distinct audio domains: music, voice, acoustic scene sound. This study aims to validate SampleCNN's generalizability and efficacy by testing it across three distinct audio domains: music, voice, acoustic scene sound.

In another study,a gated CNN for audio classification [14] has been presented. The data that has been used is taken from YouTube videos which are dynamically labeled with one or several audios without having the timestamps of the audio events, which is why it has been referred to as weakly labeled data. This challenge clearly defines two subtasks using this weakly labeled data: audio tagging and sound event detection. They have proposed a convolutional recurrent neural network (CRNN) with learnable gated linear units (GLUs) non-linearity applied on the log Mel spectrogram. Furthermore, the authors suggest a strategy to anticipate the locations from the data. In audio classification tasks, CRNNs have been used. And, for audio tagging, a CRNN-based method has been used for predicting the tags. The audio waveforms are first changed to T-F representations such as log Mel-spectrogram. Convolutional layers are then added to the T-F representations to extract high-level features. A bidirectional recurrent neural network (Bi-RNN) is then utilized to record temporal context information, followed by a feed-forward neural network to predict the posteriors of each audio class at each frame. Finally, the predicted probability of each audio tag is derived by averaging the posteriors of all frames,

and a binary cross-entropy loss was imposed between the projected probability and the ground truth of an audio recording during the training phase. They have also recommended for the use of gated linear units (GLUs) as activation functions in the CRNN to replace the usual ReLU activation functions. In a lightly supervised mode, a temporal attention-based localization strategy has been proposed to localize the occurring events as well as the chunk. On the audio tagging, the final system received a 57.7% score.

In another research, the influence of audio preparation on music tagging was examined using deep neural networks [11]. The study used multiple temporal frequency visualizations, logarithmic magnitude compression, frequency weighting, and scaling to conduct thorough audio processing studies. They demonstrated that, except for magnitude compression, many commonly used input preprocessing techniques are unnecessary. They asserted that a neural network may represent any function, but this does not mean that it can efficiently learn any function. Mel-spectrograms, for example, have recently been recommended. Despite their reduced size, Fourier transformations were deemed to convey adequate information. The following experiment has been implemented in Python deep learning frameworks. The area under the curve - of receiver operating characteristic (AUC) was utilized as a measure in all of these experiments. Although it can be less than 0.5 in theory, AUC is usually between 0.5 and 1.0 because random and perfect predictions have AUCs of 0.5 and 1.0, respectively. The authors noted that AUC scores were measured on the test set, with the results displayed on the left as AUC values of 15 repeated experiments and their standard deviation on the right. A low standard deviation indicates that they can obtain a dependable, accurate score by repeating the same trials enough times. The two largest variations between the average AUC scores and those of experiments 4 and 8 (AUC differences of 0.0028 and 0.0026, respectively) imply that there could be up to a 0:005 AUC difference between experiment instances. Based on the experiment, they assumed that an AUC difference of 0:005 is non-significant in this study and demonstrated that different input preprocessing procedures can affect performance.

Chapter 3

Working with Dataset

3.1 Dataset Description

Since we are working on audio files, we collected two types of data - scream and non-scream audio. There are about 1600 scream and 1600 non-scream audio files in the dataset. The scream files contains different kinds of scream , e.g. screams of joy, danger , fear , pain , wounded etc.The scream audio files are collected from different ages of people (children to adult). The non - scream file contains the sounds of anything except screams . For example, dog barking, birds chirping , baby crying etc.

3.2 Data Collection and Challenges

We have collected the audio files of both scream and non-scream from each individual person after asking their permission. We have collected data from our family members, friends , known people , children , students from Brac University as well as other university or school or college students. We told each person to give 5 or 6 kinds of scream or non-scream files. We recorded them in our device and saved them in local and cloud storage.

While collecting the audio files, we had to face some challenges.We had a target of collecting more than a thousand audio files for both scream and non-scream. It was difficult to ask everyone for 6 to 7 kinds of different screams at a time and some people were feeling uncomfortable to give us data .However, after discussing the importance of our work, they cooperated us in collecting data properly. Our data collection process took long time for this reason . After finishing the collection phase, we checked the files one by one if they are executable or not.

3.3 Data Pre-processing

3.3.1 Data Labeling and Naming Format

We collected data from people of different ages(both male and female). After collecting all of them, we labeled them into two categories - scream and non-scream. The scream audio files were collected in one folder and non-scream files were collected in one particular folder. After shifting the scream and non-scream records into two different folders, we renamed all the files as audio1, audio2, audio3, etc. in this format.

3.3.2 Resize

After collecting all the scream and non-scream audio files, we resized all the data into smaller portions. Some of the audio files were too long. We resized them, cut the audio files into different parts, and made them into 2 to 5 seconds audio files. We also cut those files which had bad sound quality and can not be executed properly.We also converted the audio files from mp3 to .wav format one by one.

3.3.3 Analysis of Audio Files

The audio files were read by using some python libraries and it shows the frequency of a particular audio file, spectrogram and Mel-spectrogram. For playing the scream audio file, we gave the serial number of the array and ran the code simply, thus we plotted the graphs easily. At first, we imported the necessary libraries (e.g. librosa, matplotlib, pandas) for running a scream audio file. All the audio files are then resampled into a sample rate of 22kHz, normalized, and downmixed to mono. Secondly, we played a scream audio file and read the raw data and the sample rate of that particular audio. The sample rate is 22050 for the file we used. We also found the shape of the audio file which is 105318. After that, we have the raw audio data in a graph. In the next step, we trimmed the audio to have a clear version of the previous one but since this is too much detailed and we are not interested in the detailing of the audio, we further zoomed the audio file by slicing it and getting the multiple different frequencies at different times overlapping.

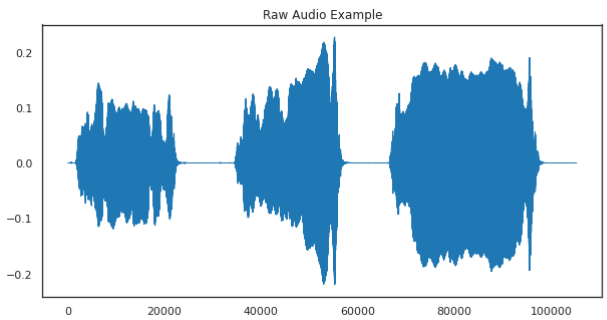


Figure 3.1: Raw Audio Example

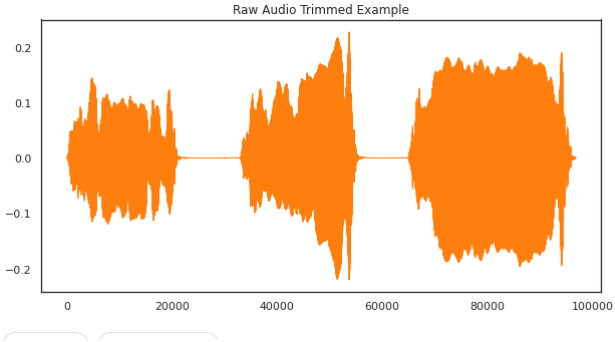


Figure 3.2: Raw Audio Trimmed Example

Now, we are going to apply a Fourier transform to the audio data to extract which frequencies are sounding at different parts of the audio file. We first applied Short Time Fourier Transformation (STFT) and after that, we used the output to run another transformation - amplitude to decibel transformation. Next, we make a spectrogram that basically shows frequency over time.

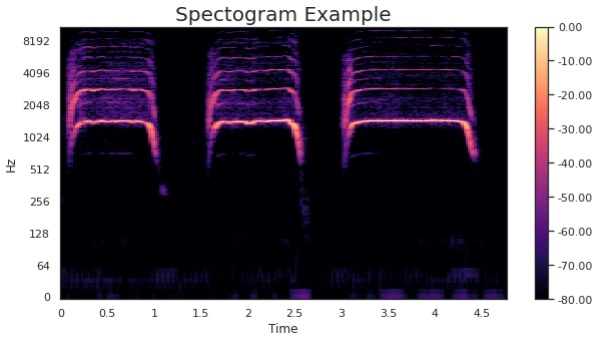


Figure 3.3: Spectrogram of the raw audio

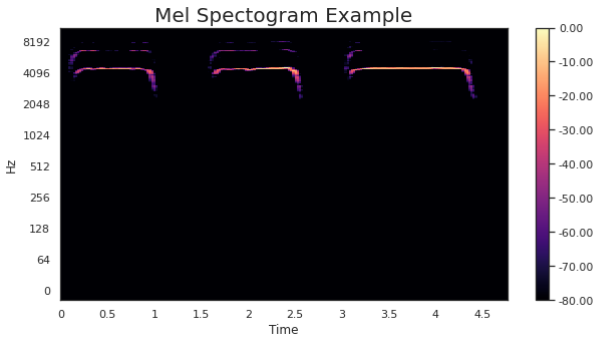


Figure 3.4: Mel Spectrogram

Finally, we created a mel-spectrogram which is the same as the previous one but 'mel' actually stands for 'melodic' because we are going to use this transform to express the frequencies that we can hear in audio usually.

$$zrc = \frac{T}{T-1} \sum_{t=-1}^{T-1} 1_{R<0}(s_t s_{t-1})$$

Figure 3.5: Here, s is a signal of length T and 1R<0 is an indicator function.

Chapter 4

Working with CNN

4.1 Sample Rate Conversion

The amount of data samples per second that are extracted from a signal to generate a discontinuous signal is known as the conversion of the rate of the data sample.

We have almost 1600 amount of scream and non-scream files. The quality of each audio file is not the same. The frequency of each audio file was also different before conversion. The audio sample rate is measured in kilohertz(kHz) and the ideal rate for same rate audio is 44.1 kHz. By converting the sample rate of audio, we will have high-quality audio data files as well as it will give us the accurate result for audio classification.

For our work, we converted the sample rate of each audio file to 16 kHz. At first, we tried with 48kHz but the result of classification was not good so we shifted to 16 kHz as per the size and duration of each audio file. After the conversion of the audio sample rate, we got better-quality audio files for training the dataset.

4.2 Audio Classification using CNN

The next step is a classification of the converted data using a Convolutional Neural Network (CNN). Though we first trained our dataset with SVM but SVM model did not show up better result while detecting screams.

However, CNN plays a better role in audio classification rather than SVM because CNN makes images of the audio and then analyzed the images to have a better version of the audio. In many types of research, CNN outperformed rather than other supervised models for audio classification. CNN can detect necessary features by scanning the image version of the audio files without any supervision. So the prediction or accuracy of the model is better than any other model.

4.3 Results from Audio Classification using CNN

For our dataset, we trained our dataset using CNN. At first, we figured out the waveform of audio files in different plots. We have two classes for our model: scream and non-scream. In the next step, we figured out the spectrogram of the audio after converting the sample rate to 16 kHz. A spectrogram is the visual representation of the audio signal frequency.

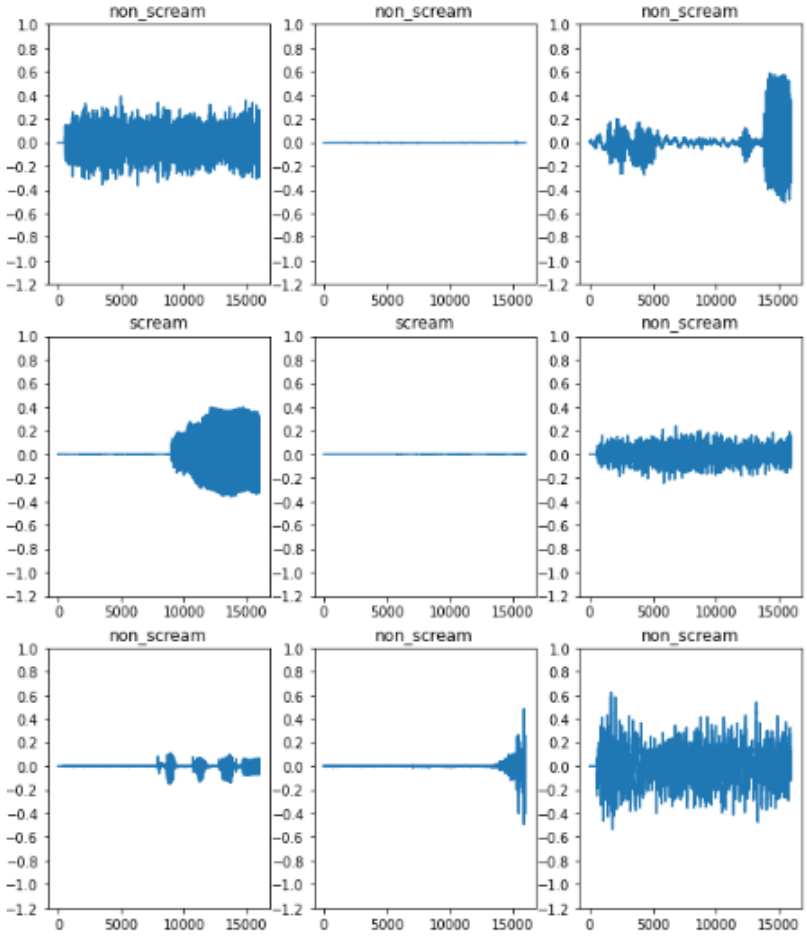


Figure 4.1: Waveform of Scream and Non-Scream Audio

We then checked if each audio files are executable after converting the sample rate, having their waveform and spectrogram together as output.[Figure 4.2 is showing the waveform and spectrogram output of a converted (16kHz) scream audio file].

In the next step, we evaluated the audio classifier with the training and validation loss to assess the fitness and performance. From the output , validation loss and training loss both decreases and stabilizes at one point and again the validation loss is decreasing. So , the model is a good fit with our dataset.[Figure 4.4 is showing the training and validation loss. Here, the yellow line showing validation loss and the blue line is showing training loss]

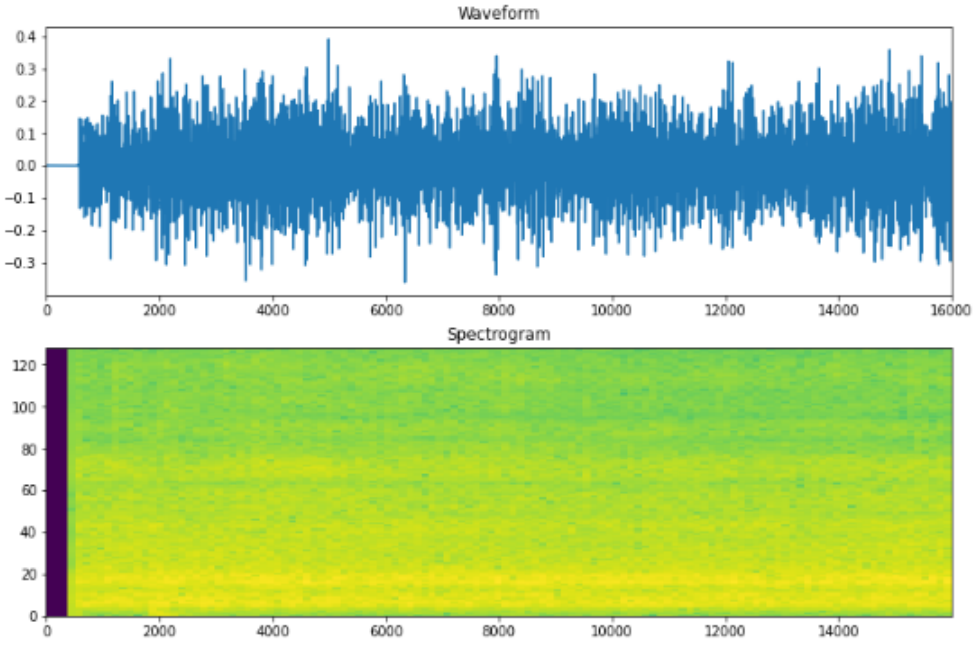


Figure 4.2: Waveform and Spectrogram of an Individual Audio File

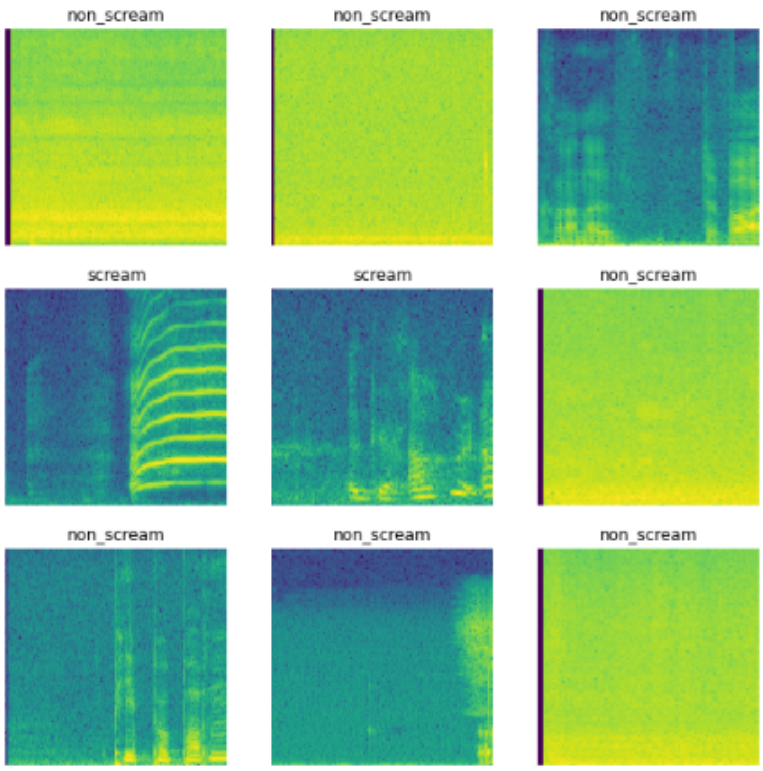


Figure 4.3: Spectrogram of Scream and Non-Scream Audio

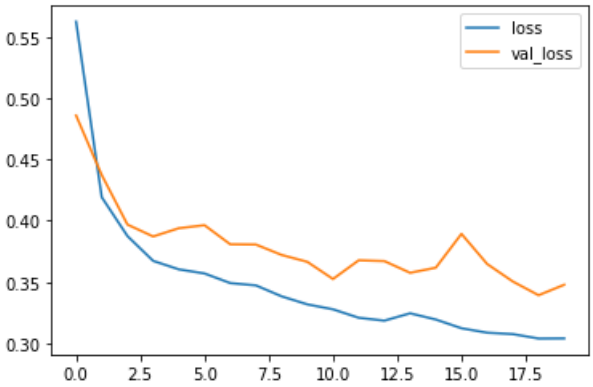


Figure 4.4: Epoch

We did performance measurements of the CNN classifier using the Confusion Matrix. The Confusion Matrix is a machine learning performance measurement that is useful for having precision, accuracy, or recall. For this model, we had two labels, scream and non-scream. We found a prediction of the model using the confusion matrix theory. The figure(below) shows how Confusion Matrix shows the prediction of a model .

[Here , TP = True Positive, FP = False Positive , FN = False Negative, TN = True Negative]

For a two-class classification problem, the prediction of the CNN model is measured based on the diagram below (Figure 4.5).

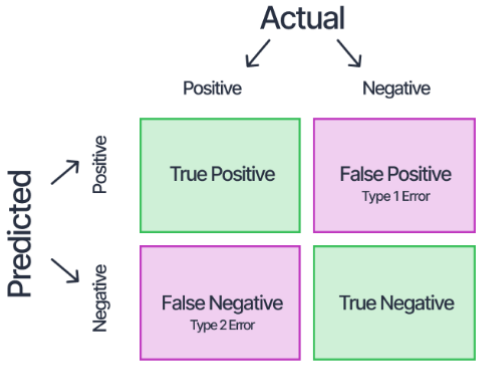


Figure 4.5: Confusion Matrix

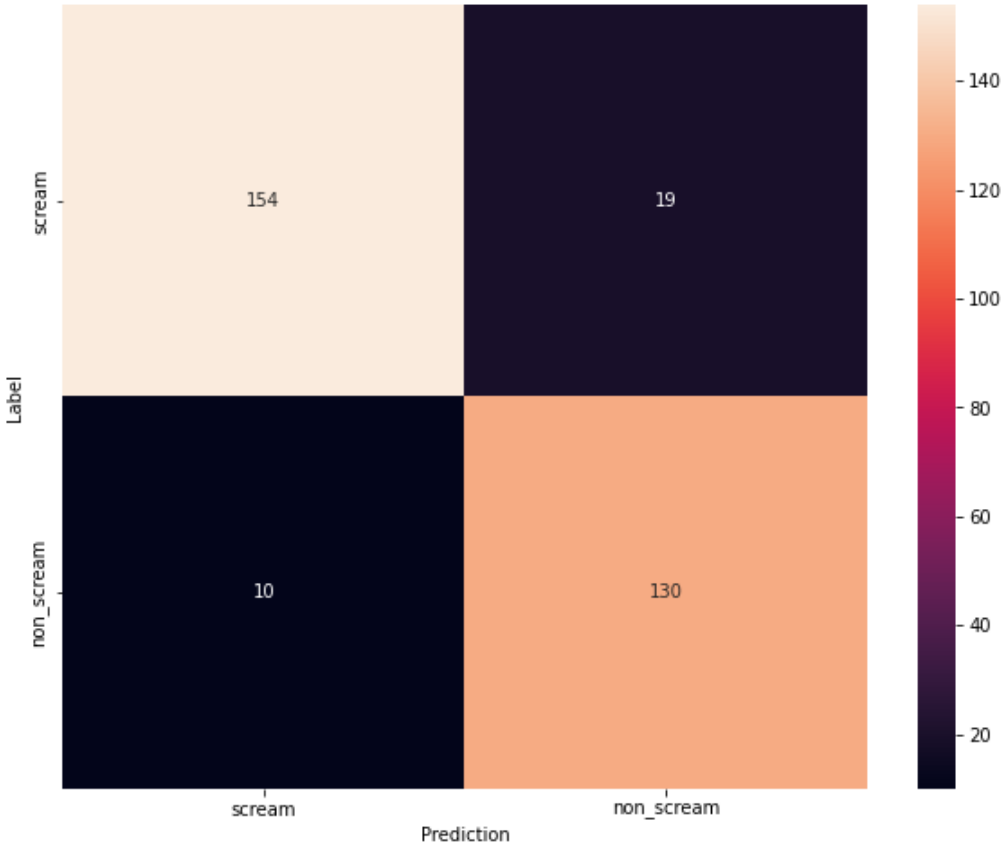


Figure 4.6: Prediction using Confusion Matrix

From the confusion matrix result, we can see that among 173 files of scream data 154 files (TP) are detected correctly as scream and the rest 19 files are detected wrongly as non-scream. Similarly , among 140 files of non-scream data, 130 files (TN) are identified correctly as non-scream but the rest 10 files are identified wrongly as non-scream.

Based on the confusion matrix result or prediction , the accuracy for the training data was 91% for CNN using the primary dataset.

Chapter 5

Work Plan and Proposed Methodology

5.1 Planning and Work-flow

5.1.1 Launch Phase

The system will first ask permission for storage, microphone, and location from the user, and all of their needs to be granted in order for the system to utilize all the functions. Then, the user has to set a set of the mobile number of his/her close ones. After that, he/she needs to provide the mobile number of the local police. After setting up all the numbers mentioned above, the app will run in the background.

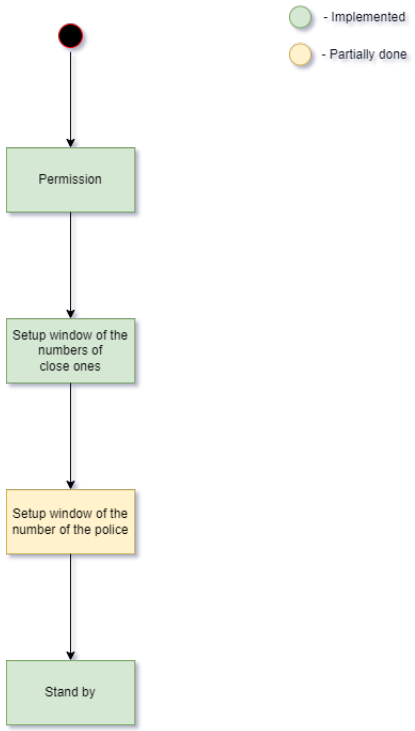


Figure 5.1: Launch Phase

5.1.2 Running on Background

The system will remain on standby mode until it gets activated through a trigger button press or special keyword. Any of the actions will lead to running up two processes which will run in parallel. One is detecting whether the sound is a scream or not and the other is recording. If the sound is detected as non-scream, then it will not save the record and will go back to standby mode. But if it is detected as a scream then it will first save the record to the local storage and simultaneously, it will fetch the current location of the person and will send a customized message to the numbers which were given earlier in the app using Twilio along with the location it fetched so that people can know the person is in danger and will be able to locate. After all the process is being done, it will again go back to standby mode.

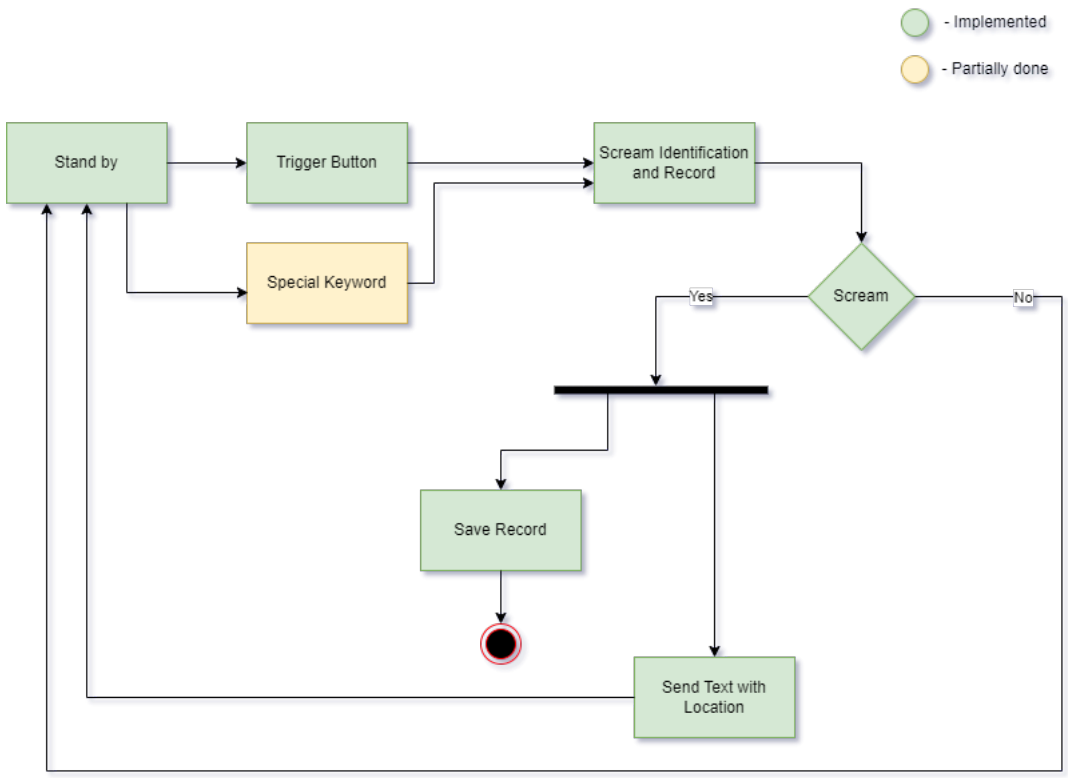


Figure 5.2: Running in Background Phase

5.2 Methodology

5.2.1 Developing the System on Flutter using Different APIs

Developing a system that detects screaming and delivers alerts to chosen contacts is a difficult but doable undertaking. Creating such a system entails constructing an algorithm that can recognize screaming properly, as well as providing an interface that allows users to pick which contacts they wish to get notifications from. Once the algorithm and interface are complete, the backend and front end of the system must be implemented. This entails building code that allows the system to accept user input, process it, and then send alerts to the specified contacts. Finally, the system must be tested and adjusted to guarantee that it functions properly. It is feasible to create a system that automatically detects shouts and delivers alerts to designated recipients using the correct design and development tactics. Developers may design efficient and effective software for this purpose by utilizing the diverse programming language Flutter.

Firstly, after creating a new Flutter project and installing the sound processing packages, we need the necessary packages to import. Permission handler is one such package, and as the name implies, it manages most of the permissions necessary for the system we have built. Permissions are not automatically provided to programs on most operating systems. Developers must ask the user's permission while operating. This plugin provides a cross platform (iOS, Android) API for requesting and checking permissions. You may also access the device's system settings to allow users to provide authorization. On Android, you may provide a reason for asking permission.

Using the Flutter audio recorder, we can record a scream and save it as an audio file with the Audio classification Tflite package for flutter. The tflite audio package will identify audio from stored audio files. Although at this moment, only mono wav files are available, it is also capable of detecting audio recordings. It also allows you to fine-tune recording/inference parameters. Finally, it automatically reshapes and transposes audio sources. This plugin can handle a variety of model types, including raw audio, decoded wav, spectrogram, Melspectrogram, and MFCC inputs.

For our case, we created the mentioned model to suit our system. This model helped us to process the dataset we collected and detect screams. Then this model was converted to tflite model so that we can deal better with our system. The process of running a TensorFlow Lite model on-device to make predictions based on input data is referred to as inference. An interpreter is required to execute an inference in the TensorFlow Lite model. The TensorFlow Lite interpreter is intended to be lightweight and quick. To achieve low load, startup, and execution time, the interpreter employs static graph ordering and a bespoke (less dynamic) memory allocator.

Twillo_Flutter is a package for both Android and iOS that assists developers with Twilio API services. We were able to alert the appropriate contacts since it can send SMS programmatically, obtain all SMS associated with a Twilio account, receive further information on each SMS sent from a Twilio account, and send WhatsApp messages programmatically. The only drawback was that several of the described

services were paid, and we could just test the auto-produced SMS section.

For obtaining the current and known locations, a Flutter geolocation plugin that enables simple access to platform-specific location APIs was employed. To find out where the device is right now, we used the getCurrentPosition function. The following parameters can be used to fine-tune the results: **desiredAccuracy:** the precision with which the system wants to get location data, and timeLimit is the maximum time allowed to obtain the current location. When the time limit is reached, a TimeOutException is raised, and the call is terminated. There is no limit set by default. In terms of the last known position, we utilized the getLastKnownPosition function to obtain the last known location saved on the device (note that this can result in a null value when no location details are available).

Chapter 6

Experiment with Interface

6.1 Results from The System

We were successful in developing a system that can run our scream detection algorithm. According to the presented dataset, we were able to show sufficient proof that using CNN, it is actually possible to identify screaming and inform nearby individuals or established safety contacts to notify. The figures (below) are showing how the proposed model is asking permission from the user while using the app, detecting screams, and sending messages to the selected contact list [Figure 6.4, Figure 6.5, Figure 6.6] .

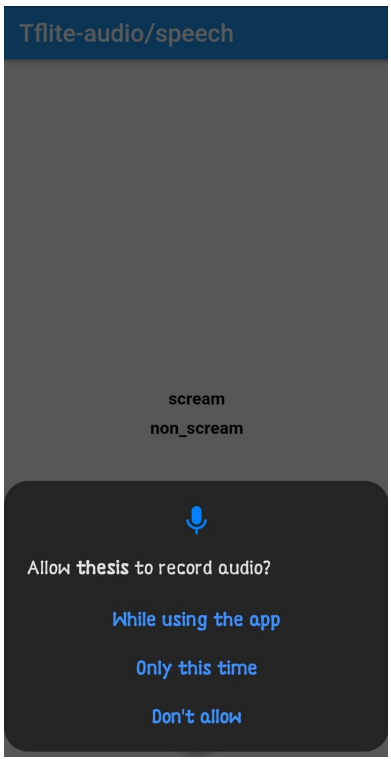


Figure 6.1: Asking Permission

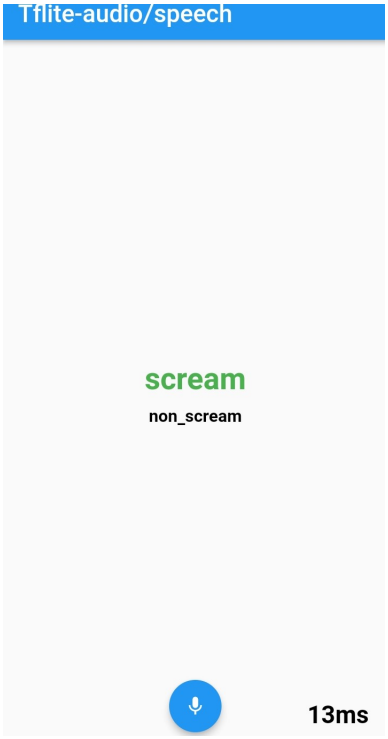


Figure 6.2: Detecting Scream

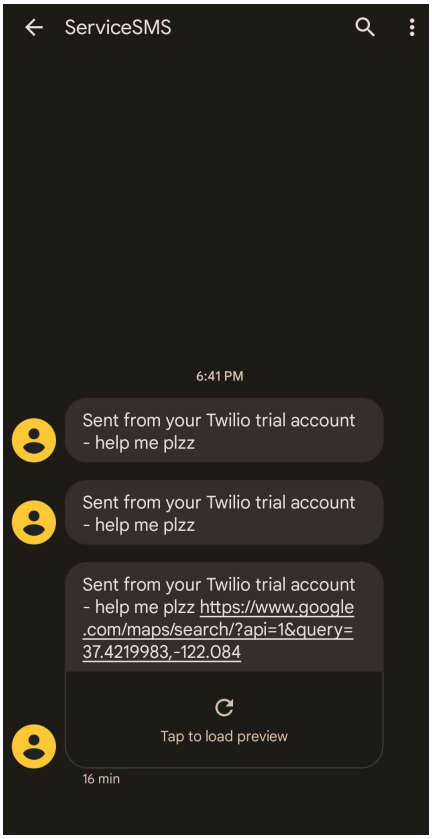


Figure 6.3: Sending Message and Location to The Selected Contact

Chapter 7

Research Contribution and Challenges

7.0.1 Contribution

By seeing suspicious behaviors, AI can help investigators identify criminals more quickly and avoid crimes. This increases public safety while also increasing community faith in the criminal justice system. Our method can significantly lower the number of such incidents, particularly in our nation, where societal violence is a widespread problem. By alerting law authorities in advance as soon as our technology detects the victim's screams and communicates their position, we can further assist our society.

Given the scarcity of scream-related data, we gave a sufficient number for future researchers to use. To develop a social violence support model for the study, we have acquired audio data that contains both screams and non-screams. Speaking voices, speeches, and a range of sounds, including a baby crying, a dog barking, birds chirping, etc., have all been captured on tape. The scream files contain screams from both men and women of different ages that convey a range of emotions, including fear, agony, pleasure, hurt, and happiness. All these details enabled us to develop a system that makes it simple to alert the appropriate parties. Our dataset not only covered different audio events but also contains speeches and dense noise so that it can better detect screams. Our primary dataset contains versatile screams, noises, and audio events that cover a wide variety of sectors so that we could enrich the model further.

Our proposed model will help society to prevent social violence that is more private in nature and tackles the core safety of the user as the usage of mobile phones increases, then it becomes a powerful tool to use such models and systems in our daily lives. We really wanted to establish such a system so that the overall people of the country, be it women, men, children or the elderly can have a safer environment in society.

We worked with tflite model, a suite of tools that makes it possible for developers to execute their models on mobile, embedded, and edge devices, enabling on-device machine learning. This enabled us to use the CNN model in mobile phone apps.

We are also developing speech-to-intent for our system so that we can utilize trigger phrases in public to launch the system automatically and send alerts to the user's selected contacts as well as the local authority administration. This will assist in lowering the incidence of social violence in society. As we can currently use it on our computers, this is still a work in progress, but we are attempting to leverage speech-to-intent on phones to make it more effective and simpler to use. We haven't made much progress in this area because of the restrictions put in place by phone makers.

7.0.2 Challenges

One of the most difficult aspects of our research was developing a model that might identify a scream because there was no data on what attributes a scream should have. It is a more recent study in which researchers are attempting to distinguish screams from other types of auditory occurrences. On that topic, another major obstacle was the dataset for our model, since we were unable to obtain sufficient data because scream is not widely available, making our data set collecting more difficult. Following the dataset and the model, we moved on to the actual process of creating the application, where we encountered compatibility issues with the phones we were currently using, which were affordable phones, as well as emulators, as we were unable to run two models, one for scream detection and another for speech to intend model, properly on any of the affordable phones. On the other hand, an application running in the background requires permission from the administrator settings, which means that if we didn't explicitly grant permission to run it, it wouldn't because it is a privacy issue, and most phone manufacturers do not allow us to have administrative access right from the back when we buy a phone, which is especially noticeable in iOS. Finally, we encountered freemium difficulties with several APIs, such as a few capabilities that we desired being pay-locked and only trial versions supporting the few functions that were insignificant for our study to focus on. One of these concerns was with the Twilio API, where we could send text messages to the chosen phone numbers but not to the contacts in bulk. Another concern was cloud storage, which was expensive for a secure cloud storage service since the data stored would demand more space than estimated. Many of the problems may have been avoided by simply using a cloud server to save or process data, but this would have resulted in less security, privacy, and user access to the key files. For example, consider what most voice recognition platforms, such as Google, Apple, and Amazon, are doing. Because they passively gather data from users, and consumers do not have full access to that data at any time.

Chapter 8

Experiment with SVM

For audio classification, we separated our entire work into two segments: one is concerned with detecting screams using CNN and the other is concerned with extracting the features of the audio data for training into SVM. The performance of working with SVM model and tflite was not successful so we could not integrate the SVM model with tflite. Thus, we worked with the CNN model which worked better than SVM in the recording process as well as integration with tflite.

To implement the SVM model, we worked on two main aspects: data pre-processing and model training, therefore we first need to process the raw audio data. We imported over 3000 audio files for data pre-processing for this research. For training the SVM-based model, we accept only audio data both scream and non-scream collected from various people. First, we used FFmpeg to sample size the audio to 22 kHz and extract the audio feature from the input dataset as this is the default sample rate in Librosa Library.

Secondly, we played a scream audio clip and read the raw data as well as the sample rate. After that, we have the raw audio data in graph form. In the next step, we trimmed the audio to have a clear version of the previous one but since this is too much detailed and we are not interested in the detailing of the audio, we further zoomed the audio file by slicing it and getting the multiple different frequencies at different times overlapping.

The audio files were read using Python libraries, and each audio file's frequency, spectrogram, and Mel-spectrogram were represented. Similarly, many extraction features are used from Librosa Library to extract the features from every WAV file of our dataset and store a total of 12 features in our CSV file. We used librosa to extract features from audio data as Librosa is a Python audio analysis and automatic speech recognition tool. It provides the building blocks for developing audio information retrieval systems. There are many Spectral Features in Librosa. The main features that we used for this research are:

1. **Chroma STFT:** To construct a chromatogram for an audio dataset, use a short-time Fourier transform on a waveform or power spectrogram.

2. **Spectral Centroid:** Spectral centroid is a calculation of digital signal to distinguish a measure the mass in the center of gravity.

3. **Spectral Bandwidth:** The spectral bandwidth is described as the width of the light band at one-half of its maximum. A spectrophotometer's spectral bandwidth is proportional to the physical slit width and optical dispersion of the monochromator system.

4. **Spectral Rolloff:** This is the frequency that gives the percentage of total spectral energy and calculates the roll off frequency for each frame of the audio signal. The percentage is 85% by default.

5. **MFCC:** It uses the MFCC algorithm to extract elements from the screaming sound which works very well with very strong noise, and the characteristics are very diverse, helping to achieve a very accurate classification.

6. **Melspectrogram:** For audio frequencies exceeding a particular threshold are rendered logarithmically.

7. **Fourier Tempogram:** It is the magnitude spectrogram of the novelty function. The visualization of a Fourier tempogram reveals the dominant tempo over time.

8. **RMS:** RMS stands for Root-mean-square. It is used for calculating a root-mean-square value for each frame of a signal that comes from the audio samples or a spectrogram. It is faster because it does not need STFT calculation and it gives an accurate representation of energy over time from the spectrogram.

9. **Poly Features:** Polynomial features are calculated by raising existing features. For instance, if the dataset has only one input feature X, the polynomial feature adds a new feature (column) to the value calculated by squaring the value of X. X2.

10. **Tempogram:** An intermediate-level representation of tempo information is used in the dataset to define tempo variation and local pulses for the audio stream.

11. **Inverse.MFCC to Mel:** It is used to invert the Mel frequency cepstrum coefficient to approximate the mel power spectrogram. First, INVERSED CT is applied to MFCC. Then use librosa.db to power to map the decibel scale results to the spectrogram.

12. **Zero Crossing Rate (ZCR):** ZCR is the speed at which the signal changes from negative to zero, positive, or positive to zero, negative, but in some situations, only "positive" or "negative" intersections are counted, not all intersections. .. Its importance is widely recognized in speech recognition and is an important factor in the classification of percussion instruments.

However, to process our dataset, we used 11 features that are in Real Numbers and discarded all the complex numbers. After preprocessing, we used Librosa Library features to train with Support Vector Machine (SVM) model that can identify how much a person is screaming or not screaming i.e., if they are angry or happy. For training, we used the MFCC, Zero crossing rate, etc. feature vectors to predict the phoneme to make the model more resilient so that the model can detect fear scream

whenever a person shouts in times of danger.

Although the SVM model reached an accuracy level of 89%, the SVM-based model showed low accuracy than the CNN model and SVM could not detect scream or non-scream sound properly.Things have recently gotten a little clearer by employing Convolutional Neural Networks (CNN), where the model shows better accuracy than SVM ,has the capacity to learn patterns with spatial hierarchies and translation invariant, and without any human oversight, it automatically finds the key characteristics of any audio data.Moreover, integration with tflite was successful with CNN.

Chapter 9

Comparison with Existing System

	Scream Detection	Registration	Audio Recording Evidence	Offline Speech to Intent	SMS & Victim Location
Our System	yes	no	yes	yes (Partially)	yes
Abhaya	no	yes	no	no	yes
WE'RSAFE	no	no	no	no	yes
SafeBand	no	no	no	no	yes
#MeToo	no	no	no	yes	yes

Figure 9.1: Comparison with Existing System

In the table, comparison has been made between the existing systems such as Abhaya, WE'RSAFE, SafeBand and Metoo with our system where it is seen that it terms of SMS and sending victim's location all the system has the capability to perform the task. In terms of Scream detection only our system has the ability to perform the task where no other system mentioned here can do it. Except Abhaya, our system along with the others does not require registration which makes it convenient. Only our system has the ability to record the audio event when a violence is occurring which will be saved in the local storage as an evidence for future reference purpose. And lastly, in terms of Offline speech to intent only MeToo and our system (partially) has the ability to use it.

Chapter 10

Future Work

10.1 Automatic Detection

Whenever a person is in danger, it is difficult to use a device properly most of the time. Besides, the attacker might be well known about the fact that the victim is carrying something defensively. For this reason, we will implement automatic detection by using speech-to-intent. Speech-to-intent is a Natural Language Understanding (NLU) engine that basically converts a spoken utterance into text. If a person uses any special keywords like " help me" or " I am in danger", the system can convert the utterance into speech. In this way, the victim does not need to use the device or app directly. So we will integrate speech-to-intent with our system for the users in the future.

Moreover, we will work on the scream and non scream audio with their frequency so that the system can distinguish properly between fear, pain, joy, happy scream etc. and give users accurate result while detecting scream.

10.2 Saving Records in Cloud Storage

For the proposed system, we are now saving the audio records in the local storage of an android device. By using those records, a victim can report to the police or administration to get proper help in the future. Every audio file, time duration, and time everything can be used as evidence. By saving in cloud storage, the evidence will stay secure, it can be accessed easily, and will be convenient while sharing. Data is stored in the cloud throughout multiple unnecessary servers, so even if one of the cloud services crashes, data will still be maintained by the other servers, keeping it safe and under control. Moreover, if the user's device is lost, so the backup data from cloud storage can be used as evidence. We will implement the feature for saving audio records in cloud storage in our future work.

10.3 Better User Interface

Although we were able to establish a system, its user interface and design still need improvement in order to improve the fluidity and quality of life. An improved user

interface will not only assist the user manage the emergency contacts more effectively, but it will also help the user keep track of changes and stored data.

10.4 Auto Call

We will have an automatic call function in the system to the local authorities would be another enhancement over our work, but it would be a challenging future development given the problems with permits and S.O.S. features that we have already discussed.

10.5 Group Message or Call

Another section will be added named the group messaging or call feature, which allows users to send messages or call specific groups of individuals like their neighbors or the local police station, which can aid in this effort and result in greater aid for the victims. There are always other individuals who can help because not everyone is constantly accessible. By having this feature, if a person is in danger and there are other people staying in that area who is/are using the same system on their device, will get a notification of nearby danger along with the location. Thus, those people who will have notification can help the victim instantly.

10.6 Special Trigger or Unique Trigger

We will include a feature named special trigger or unique trigger for emergency help. This feature will be used for sending emergency help to the police. It is sometimes difficult to call by typing a number instantly in any dangerous situation. By using this feature, if a person presses the phone button 3 times simultaneously, there will be a message with the location which will be sent to the police or the user's selected contact list.

Chapter 11

Conclusion

When it comes to social violence, occurs all over the world and technology is upgrading day by day. Thus, our aim is to take help from technology to prevent social violence through our model which will be able to detect clearly and reliably the screams of the individual in the position who is in any danger, that is, if the scream arose out of dread and horror, based on a particular collection of audios. By using the CNN model in the system, we are able to detect screams properly and notify the people from the selected contacts via message. In our future work, we will attempt to classify the screams to a greater extent, e.g. measure the violent level of screams. Besides, we will add many more features to help people in danger using many options from this system. We hope that one day, the developed system will be able to extract the scream from a very noisy environment and be able to measure the violence level. Moreover, it will help people to convert their utterances into texts and send automatic calls or messages to the police.

Bibliography

[1] S. Kapoor *et al.*, "Domestic violence against women and girls," Tech. Rep., 2000.

[2] J.-C. Wang, J.-F. Wang, C.-B. Lin, K.-T. Jian, and W. Kuok, "Content-based audio classification using support vector machines and independent component analysis," in *18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, vol. 4, 2006, pp. 157–160.

[3] W. H. Organization *et al.*, "Adolescent pregnancy [electronic resource]: Unmet needs and undone deeds: A review of the literature and programmes," 2007.

[4] N. Sarker and S. Yesmin, "Domestic violence in bangladesh: Analyzing from contemporary peace and conflict perspectives," *Peace and Security Review*, vol. 5, no. 10, pp. 74–90, 2013.

[5] R. S. Yarrabothu and B. Thota, "Abhaya: An android app for the safety of women," in *2015 Annual IEEE India Conference (INDICON)*, 2015, pp. 1–4.

[6] M. Z. Zaheer, J. Y. Kim, H.-G. Kim, and S. Y. Na, "A preliminary study on deep-learning based screaming sound detection," in *2015 5th International Conference on IT Convergence and Security (ICITCS)*, IEEE, 2015, pp. 1–4.

[7] K. Abusaleh and A. Mitra, "Trends and patterns of violence against women in bangladesh," *Glob J Hum Soc Sci*, vol. 16, no. 6, pp. 28–34, 2016.

[8] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 6460–6464.

[9] T. Dey, U. Bhattacharjee, S. Mukherjee, T. Paul, and R. Ghoshhajra, "Advanced women security app: We'rsafe," *Int. Inform. Eng. Technol. Assoc*, vol. 4, no. 2, pp. 47–51, 2017.

[10] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2017, pp. 131–135.

[11] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "A comparison of audio signal preprocessing methods for deep neural networks on music tagging," in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, 2018, pp. 1870–1874.

[12] M. N. Islam, N. T. Promi, J. M. Shaila, M. A. Toma, M. A. Pushpo, F. B. Alam, S. N. Khaledur, T. T. Anannya, and M. F. Rabbi, "Safeband: A wearable device for the safety of women in bangladesh," in *Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia*, 2018, pp. 76–83.

[13] J. A. Sheikh and Z. Fayyaz, "# metoo: An app to enhancing women safety," in *International Conference on Applied Human Factors and Ergonomics*, Springer, 2018, pp. 546–553.

[14] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 121–125.

[15] T. Kim, J. Lee, and J. Nam, "Comparison and analysis of samplecnn architectures for audio classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 285–297, 2019.

[16] R. O'Donovan, E. Sezgin, S. Bambach, E. Butter, S. Lin, *et al.*, "Detecting screams from home audio recordings to identify tantrums: Exploratory study using transfer machine learning," *JMIR Formative Research*, vol. 4, no. 6, e18279, 2020.

[17] F. S. Saeed, A. A. Bashit, V. Viswanathan, and D. Valles, "An initial machine learning-based victim's scream detection analysis for burning sites," *Applied Sciences*, vol. 11, no. 18, p. 8425, 2021.

[18] V. Kalbag and A. Lerch, "Scream detection in heavy metal music," *arXiv preprint arXiv:2205.05580*, 2022.

[19] H. Akhter, "A trend analysis on domestic violence against women in bangladesh,"