IMPROVING MASS SHOOTING SURVIVABILITY: A SYSTEMATIC MACHINE LEARNING APPROACH USING AUDIO CLASSIFICATION AND SOURCE LOCALIZATION

Jevon Mao[1], Marisabel Chang[2]

[1]Santa Margarita Catholic High School 22062 Antonio Pkwy, Rancho Santa Margarita, CA 92688
[2]Computer Science Department California State Polytechnic University, Pomona, CA 91768

ABSTRACT

Mass shootings have emerged as a significant threat to public safety, with devastating consequences for communities and individuals affected by such events [7]. However, a lack of widespread use of new technological infrastructure poses significant risk to victims [8]. This paper proposes a system to classify and localize gunshots in reverberant indoor urban conditions, using MFCC features and a Convolutional Neural Network binary classifier [9]. The location information is further relayed to users through a mobile client in real time. We installed a prototype of the system in a high school in Orange County, California and conducted a qualitative evaluation of the approach. Preliminary results show that such a mass shooting response system can effectively improve survivability.

KEYWORDS

Machine Learning, Public Safety, Acoustics, Directioning

1. INTRODUCTION

Mass shootings, particularly those that occur in public spaces, are a serious and pervasive problem in many parts of the world [10]. In the United States alone, there have been over 300 mass shootings since the start of 2021, resulting in hundreds of deaths and injuries (Gun Violence Archive, 2021). In response, researchers and developers have sought to utilize technological solutions to detect and mitigate mass shootings in real-time. One such solution is the use of mass shooting detection systems that employ audio classification and source localization.

Audio classification involves the analysis of sounds to discern their type or content, while source localization involves the determination of the location of a sound source. By integrating these technologies, mass shooting detection systems can identify the presence of gunshots and their location, enabling first respondent to respond more promptly to mass shooting events.

David C. Wyld et al. (Eds): ICDIPV, CBIoT, ICAIT, WIMO, NC, CRYPIS, ITCSE, NLCA, CAIML -2023
pp. 127-141, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131310

The usage and popularity of mass shooting detection systems that incorporate audio classification and source localization vary [11]. While some schools and other public venues have implemented such systems, others have not. The potential benefits of these systems include enhanced speed and effectiveness in detecting and responding to mass shootings, which may ultimately lead to the saving of lives and the reduction of injuries. However, the implementation and maintenance of such systems also carries costs, as well as potential privacy concerns.

Given the significant impact of mass shootings on public safety, the development and deployment of mass shooting detection systems that utilize audio classification and source localization is a topic of critical importance. These systems have the potential to prevent or mitigate the consequences of such events, thereby improving the safety of communities and individuals. As such, further research and development in this area is warranted to better understand the potential and limitations of these systems, as well as to address any remaining challenges.

Many existing classification and localization techniques, as well as public safety systems have been proposed to assist in the event of a mass shooting. One approach that has gained significant attention in recent years is the use of machine learning algorithms for audio classification and source localization. These algorithms are designed to identify the sounds of gunfire and locate the source of the shooting, in order to alert first responders and assist with the response. One example of this type of technology is the Deep Learning Gunshot Detector (DLGD), which was developed by researchers at the University of Maryland [12]. The DLGD uses a convolutional neural network to classify audio recordings as either gunfire or non-gunfire, and has been shown to achieve high accuracy in detecting gunshots in relatively controlled, theoretical environments.

Another example of an audio classification algorithm is the Gunshot Locator System (GLS), developed by ShotSpotter Inc. The GLS uses a combination of machine learning algorithms and cross-correlation techniques to identify the sounds of gunfire and locate the source. The system has been implemented in several cities across the United States, and has been credited with helping to reduce gun violence in those areas.

Another technology that has been developed to improve mass shooting survivability is the Moonlight app, which allows individuals to quickly and easily call for help in the event of an emergency. The app allows users to input their location and the type of emergency they are facing, and will automatically alert first respondents if the user is unable to do so themselves. The app uses location tracking and machine learning algorithms to detect potentially dangerous situations, and can automatically send an alert to authorities if the user does not respond to a prompt within a certain time frame. This can be particularly useful in situations such as during a mass shooting, as it allows individuals to quickly and discreetly alert authorities to the location and nature of the threat.

Another option for individuals in the event of a mass shooting is the manual calling of 911. While this method has been in use for decades, recent developments in 911 technology, such as Enhanced 911 (E911), have made it easier for individuals to quickly and accurately communicate their location and situation to authorities. E911 technology uses GPS and other location services to automatically identify the location of the caller, which can be particularly useful in situations where the caller may not be able to communicate their location accurately.

Despite the potential of these technologies and products to improve mass shooting survivability, there are several issues that need to be considered. One of the main challenges is the accuracy of the algorithms and the quality of the data used to train them. If the algorithms are not properly trained or the data is not accurate, these technologies may not be effective in identifying or locating shooters. And most importantly, some of these technologies require specialized equipment or infrastructure, which may not be feasible or cost-effective in all locations.

In conclusion, different techniques involving machine learning algorithms and related technologies have made great attempts to tackle the issue of mass shootings. However, it is important to carefully consider the limitations and potential issues of these technologies, and to ensure that they are implemented in a way that maximizes their effectiveness. For commercial mass shooting detection technology, one common limitation in the scale of implementation is their lack of flexibility. For example, ShotSpotter can only be deployed to cover exterior public areas, using microphones sensors mounted on street lights. This greatly limits its implementation it cannot provide any indoor coverage, which is especially important for high risk indoor facilities like schools and public transportation hubs. Another limitation in the scale of implementation is its high costs. Oftentimes, the commercially available solutions require specialized equipment, installation, and maintenance as well as costly infrastructure for American cities. For instance, the well known Shot Spotter system costs "$65-90k per square mile per year, with a $10K per square mile one-time Service Initiation fee" (Citation). The lack of feasibility and financial resources to install those systems essentially renders any technological breakthroughs futile when they cannot be deployed on the frontline cities. Lastly, both commercial solutions and novel machine learning algorithms share a significant blindspot when it comes to involving end victims of the threat. These existing systems often lack a complete integrated pipeline that directly delivers updated information to victims, without handing the final control back to a traditional law enforcement situation. Most users would not be aware of a gunshot detection system installed to receive any psychological reassurance, and would significantly lack real-time information in the middle of a deadly attack.

Our method for improving mass shooting survivability is a machine learning approach using audio classification and source localization, utilizing a series of Raspberry Pi microphone sensors deployed around a building. The raspberry pi devices are equipped with on-device classification using a Recurrent Neural Network (RNN) to identify the sounds of gunfire. The audio recordings and timestamps are then synchronized across all the Raspberry Pi devices, and cross correlation is applied to calculate the possible source of the gunshot through Time Difference of Arrival (TDOA) [13].

One key feature of our method is the use of transfer learning to train the machine learning algorithms used for audio classification. Transfer learning is a technique in which a model that has been trained on a large dataset is fine-tuned for a specific task, allowing for more efficient and effective training. In our case, we used transfer learning to combine multiple audio training datasets, including the Urban8K audio training set, in order to achieve a high level of accuracy in predicting gunshots. Our approach resulted in a test set accuracy of 97.5% and a training set accuracy of 99.3%.

In addition to the use of transfer learning, we also developed an innovative time synchronization technique that syncs the time on each Raspberry Pi to within 0.001ms. This allows for accurate and precise calculation of the source of the gunshot using TDOA.

To complement our machine learning algorithms and sensor technology, we also developed an easy-to-use mobile app that allows individuals caught in a mass shooting situation to quickly and discreetly alert first responders to the location of the shooter. The app displays the threat on a map and enables live location sharing between nearby victims, as well as live audio and video feed with 911. The app is designed with a user-friendly interface and has been well received by testers.

In comparison to existing methods such as the Deep Learning Gunshot Detector (DLGD) and the Gunshot Locator System (GLS), our method offers several advantages. One key strength is the use of Raspberry Pi devices, which allows for a scalable and cost-effective solution that can be easily deployed in a variety of locations. Additionally, the use of RNN for on-device classification and TDOA for source localization provides a more accurate and reliable approach, as it is not reliant on a centralized system or specialized infrastructure.

Overall, our method for improving mass shooting survivability combines machine learning algorithms, sensor technology, and a user-friendly mobile app to provide a real-time, accurate, and reliable solution for identifying and locating shooters. The use of transfer learning, innovative time synchronization techniques, and a well-designed mobile app make our method a strong contender for improving mass shooting survivability.

In order to prove the effectiveness of our method for improving mass shooting survivability, we used a combination of qualitative and quantitative analysis methods. For the qualitative analysis, we conducted a case study evaluating our entire system pipeline, including the machine learning algorithms, sensor technology, and mobile app. We performed a near 1:1 resemblance simulation of a real mass shooting scenario to demonstrate how our system improves mass shooting survivability. A power amplified speaker is used to replicate the loudness of a real gunshot in a building preinstalled with a Raspberry Pi sensor array that standbys 24/7. When the gunshot is played, the sensor array immediately registered the noise and successfully labeled the sound as gunshot through the deep learning classification model. The entire system was evaluated on different metrics, from time performance and accuracy performance to cost efficiency and ease of interaction. As part of the case study, we interviewed students and faculty about their experiences with the system and how they felt about its effectiveness. We asked questions such as whether they felt safer with the system in place, and whether they believed it would be useful in the event of a mass shooting. The responses show a significant increase in perception of safety among students and faculty.

To supplement the qualitative analysis, we also used quantitative methods to evaluate the precision of the classification model and the error margin of the localization system. For the classification model, we used a test set of audio recordings and compared the predictions of the model to the actual labels. This allowed us to calculate the accuracy of the model and determine whether it was effective in identifying gunshots. For the localization system, we used error margin analysis to assess the precision of the TDOA calculations and determine the potential error in the estimated location of the shooter.

Overall, our results showed that our method for improving mass shooting survivability was effective in both identifying gunshots and locating the source of the shooting. The combination of qualitative and quantitative analysis methods allowed us to evaluate the effectiveness of the system from multiple angles and provided a comprehensive understanding of its performance.

In addition to the analysis methods described above, we also conducted several experiments to further test and refine the system. For example, we conducted experiments to assess the impact of different microphone configurations and to evaluate the performance of the machine learning algorithms under various conditions. These experiments allowed us to identify areas for improvement and to optimize the system for maximum accuracy and reliability. Finally, we conducted experiments to evaluate the accuracy of our CNN model by using ESC-50 dataset, UrbanSound8K dataset and YAMNet pre-trained model.

In conclusion, our results demonstrate the effectiveness of our method for improving mass shooting survivability. The combination of machine learning algorithms, sensor technology, and a user-friendly mobile app allows for a real-time, accurate, and reliable solution for identifying and locating shooters. Further research and development is needed to continue to refine and improve the system, but the initial results are promising and suggest that our method has the potential to greatly improve public safety in the event of a mass shooting.
The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the comprehensive mass shooting response system; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the project, a few challenges have been identified as follows.

2.1. Obtaining dataset

The unique acoustic environment of our system's designed use case poses some unique challenges for audio classification. While many prior works explore into developing a good gunshot classification algorithm that maximizes true positives while reducing false positives, many of the traditional techniques and acoustic assumptions are not applicable here. It is well established that the spectral domain of a typical impulsive signal, like a gunshot, reflects more of its acoustic surroundings than of the signal itself. In other words, an impulsive signal that is easily absorbed, scattered, and reflected by objects and surfaces in its environment essentially acts somewhat as a mirror, reflecting the shape of its surroundings. For spectral domain based feature extraction methods like MFCC (Mel-frequency cepstral coefficient) or Mel Spectrogram, this can lead to reduced accuracy and differentiation of different impulsive signals. Many prior work's scope of research excludes the surrounding limitation of impulsive signals by exclusively addressing open field applications of gunshot detection, where there exists little to no reverberation. Other works acknowledge the spectral limitation and rather turn to the temporal domain of a gunshot sound by using template correlation, RMS threshold, or other similar, less sophisticated classification algorithms. However, prior success achieved with these methods are limited to extremely noiseless outdoor environments like forests and fields, where it can be assumed that no other high energy impulsive signals exist. Therefore, solely temporal domain based feature extraction is likely not applicable for our use case, where the system is intended to be deployed in schools, malls, and other urban locations. Numerous urban noises from jackhammers to car exhausts can easily resemble the high energy temporal characteristic of a gunshot. The complex and reverberant indoor environment can significantly distort audio signals, preventing algorithms from differentiating the spectral characteristic of a gunshot. In addition, the expected gunshot sound in our case has no predefined distance, weapon caliper, or direction. In other words, our algorithm must be able to account for the attenuation of audio signals in

accordance with the inverse square law and be agnostic to the weapon type as well as firing location. To address the various classification challenges, a sophisticated deep learning solution trained on a large and diverse dataset of gunshot sounds is likely required. However, obtaining such a dataset can also be difficult, as it may require access to a wide range of firearms and the ability to safely and legally fire them in controlled conditions. In addition, the dataset may need to include a variety of different background noise and environmental conditions in order to adequately prepare the model for real-world situations.

2.2. The accuracy and reliability of the method

Source localization, or the determination of the location of a sound source, is another important aspect of improving mass shooting survivability. One common method for source localization is multilateration based on time difference of arrival (TDOA), which involves the use of multiple microphones to determine the location of a sound source based on the difference in arrival time of the sound at each microphone. While TDOA can be an effective method for source localization, it is not without its challenges.

One major difficulty of using TDOA is the accuracy and reliability of the method. The accuracy of TDOA can be affected by factors such as the distance between the microphones, the orientation of the microphones, and the presence of obstacles or reflections in the environment. If the microphones are not properly spaced or oriented, the TDOA method may produce inaccurate results. Additionally, TDOA localization can be sensitive to noise and other distortions in the audio signal, which can reduce its accuracy and reliability.

Another challenge of using TDOA is the computational complexity of the method. The TDOA method requires the use of complex algorithms to calculate the difference in arrival time of the sound at each microphone, which can be computationally intensive. A typical cross correlation algorithm like the GCC-PHAT (Generalized Cross Correlation with Phase Transform) performs a full fourier transform on both signals. The computational intensity further increases as the audio sample rate from the microphones is increased or when more microphone sensors are used to compute the localization. Lowering the sample rate risks lowering the Nyquist Frequency and losing valuable high frequency information, while decreasing the number of microphones inherently decreases the accuracy of multilateration. This poses additional challenges for implementing TDOA in real-time systems or in resource-constrained environments, including the Raspberry Pi single board computer we used for prototyping.

2.3. Developing a system for improving mass shooting survivability in real-world environments

Developing a system for improving mass shooting survivability in real-world environments is a challenging task, due to the unpredictable nature of real-world environments. Sound can be attenuated or reflected by walls, objects, and other environmental factors, which can make it difficult to accurately classify and locate sounds. Additionally, other sounds in the environment, such as screaming, bag popping, and other noises, can resemble gunshots and create false positives for the system.

The frequency characteristics of gunshots can also be affected by the environment, such as by the presence of reverberation or other acoustic effects. These factors can make it difficult to develop a reliable and accurate system for improving mass shooting survivability in real-world environments. It is therefore important to carefully consider the potential impact of these factors when designing and testing the system, and to use appropriate techniques to mitigate their effects.

3. SOLUTION

Our proposed system for improving mass shooting survivability consists of three main components: a microphone array, a real-time gunshot classification module, and a source localization module. These components work together to provide real-time, accurate information about the location of a shooter in a mass shooting situation, enabling individuals to take appropriate action to protect themselves and others.

The microphone array is composed of multiple Raspberry Pi devices, each equipped with an omnidirectional microphone sensor with a wide frequency response and high sensitivity. These Raspberry Pis are mounted in different locations in different rooms to form the microphone array, allowing for wide coverage and the ability to localize sounds from multiple directions. The Raspberry Pis in the microphone array are connected to a central server through a secured tunnel, enabling the transfer of audio data and localization information in real-time.
The real-time gunshot classification module is implemented using a recurrent neural network (RNN) trained on a large dataset of gunshot sounds. The RNN is designed to recognize the unique characteristics of gunshots and distinguish them from other types of sounds. Each Raspberry Pi in the microphone array is capable of running this RNN in real-time to classify incoming audio as either a gunshot or some other type of sound. When combined, the Raspberry Pis in the microphone array can use time difference of arrival (TDOA) to localize the source of the gunshot.

The source localization module is implemented using a cross-correlation technique and a distance equation that accounts for the speed of sound. The cross-correlation technique is used to determine the time difference of arrival (TDOA) of the gunshot at each microphone in the array, while the distance equation is used to estimate the location of the gunshot source based on the TDOA values. By minimizing the distance equation between each microphone in the array and the sound source, the system is able to accurately estimate the location of the gunshot source.

The information from the microphone array and the source localization module is relayed to a backend server, which then pushes the information onto mobile devices through a secured tunnel. On these mobile devices, the user can see the best estimated location of the shooter in real-time, allowing them to formulate a better escape plan in the event of a mass shooting.

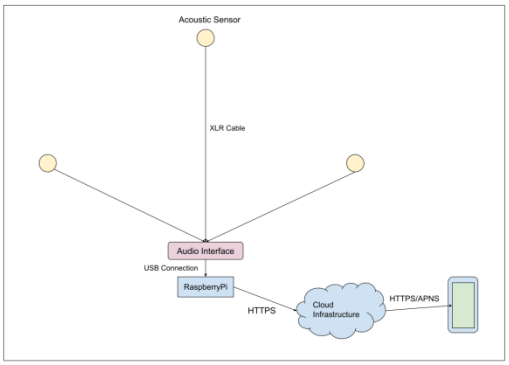

Figure 1. Overview of the solution

3.1. Hardware

The prototype system hardware architecture consists of a Raspberry Pi 4 SBC (Single Board Computer), a Behringer Audio Interface, and 3 Audio Technica AT803 Omnidirectional Condenser Lavalier microphones. The Raspberry Pi 4 configured with 8GB of memory is chosen as the main processing unit for this paper due to its practical, small size for deploying in real world scenarios, relatively low power consumption, and powerful CPU processing power necessary to run deep learning models. An XLR audio interface is a type of digital interface that is used to connect professional audio equipment, such as microphones and mixing consoles, to a computer. XLR connections are balanced, which means that they are less prone to noise and interference than unbalanced connections. This industry standard hardware is often used for recording and producing audio in professional settings, also making it an ideal candidate for our TDOA use case. The Behringer Audio Interface supports 4 XLR input channels, aggregating the analog signals from each microphone and converting them to digital signal for further processing. USB microphones are prone to severe clock drifts, as each microphone has its own ADC (Analog-to-digital converter) that is not synchronized with other microphones. When performing TDOA multilateration, this can lead to severe complications when sound travels at roughly 343 meters per second at 20 Celsius, and even milliseconds of time drift from USB microphones will lead to huge inaccuracies. The AT 803 lavalier microphone is chosen for its flat frequency response across spectrum, wide 30-20,000 Hz frequency response, high maximum sound pressure level at 135 dB and low sensitivity. An ideal acoustic sensor for gunshot classification and localization should have a flat frequency response to accurately reflect the sound wave, wide frequency response to capture all the spectral characteristics, high maximum sound pressure level to fully capture high energy impulsive signals, and low sensitivity to avoid clipping when recording loud gunshots in a highly reverberant indoor environment.
AI Algorithm

3.1.1. Gunshot detection

After assessing well established binary classifier techniques for gunshot detection, we chose to implement a MFCC audio feature extraction and feed the coefficients into a CNN (Convolutional Neural Network). When the microphone records a sound, the raw acoustic information is stored as a time series in an array with each sample representing signal strength at the instant. To get an acceptably accurate classifier, we must employ feature extraction for dimensionality reduction rather than feeding in the entire raw time series. MFCC (Mel-Frequency Cepstral Coefficients) is a widely used feature extraction technique in the field of speech and audio signal processing. It is based on the idea of representing the spectral characteristics of a signal in the frequency domain using a set of coefficients that capture the shape of the spectrum in a compact form. MFCC has several advantages for gunshot detection applications. First, it is robust to noise and can effectively filter out background noise, making it well-suited for use in noisy environments. Second, it is highly discriminative and can accurately differentiate between different types of gunshots, even when they are recorded under different conditions. Finally, MFCC is computationally efficient, which makes it well-suited for use in real-time gunshot detection systems and competitive against other extraction algorithms like LPC and impulsivity parameter of stable distribution. Compared to LPC, MFCC is more robust to noise. LPC relies on the assumption that the signal can be modeled as an all-pole filter, which is not always the case in real-world signals. In contrast, MFCC is based on a more general model of the spectrum that is less sensitive to modeling errors. The impulsivity parameter of a stable distribution is a measure of how "peaky" or "impulsive" the distribution is, but it does not provide any information about the valuable spectral characteristics of a signal, like MFCC does. We used the Slaney warping formula, a popular implementation for MFCC. The warping formula is used to warp the frequency scale of the power spectrum so that the distance between frequency bins is

proportional to the perceived change in frequency by the human ear. We chose a coefficient size of 40, essentially condensing the dimensionality of a 3 second audio clip recorded at a sample rate of 44.1 kHz into 40 float numbers before feeding it into the CNN.

Our gunshot classifier architecture is a convolutional neural network (CNN). The CNN consists of four convolutional layers followed by a global average pooling layer and a fully-connected sigmoid output layer.

The convolutional layers apply a set of filters to the input audio data and extract features from the data using a sliding window approach. The kernel size of the filters is 2, which means that each filter looks at a small window of 2 samples at a time. The padding parameter is set to "same," which means that the spatial dimensions of the output of the convolutional layers will be the same as the input, allowing the CNN to preserve spatial information in the data. The activation function used in the convolutional layers is the rectified linear unit (ReLU) function, which helps to introduce nonlinearity into the model.

After each convolutional layer, there is a max pooling layer that downsamples the output of the convolutional layer by taking the maximum value within a small window of samples. This helps to reduce the spatial dimensions of the data and reduce the computational complexity of the model. The dropout layers after each pooling layer help to prevent overfitting by randomly setting a fraction of the input units to zero during training.

The global average pooling layer aggregates the output of the convolutional layers across the spatial dimensions, resulting in a single vector per channel. This reduces the number of parameters in the model and helps to prevent overfitting. The sigmoid output layer produces a probability score for each class (in this case, two classes representing gunshot and non-gunshot sounds) using the sigmoid activation function.

This architecture learns the hierarchical representations of the audio data, capturing the complex spectral patterns characteristic of gunshots. The use of max pooling and dropout layers helps to regularize the model and reduce the risk of overfitting. The final sigmoid output layer allows the model to produce probabilistic predictions, which can be useful for tasks such as anomaly detection.

3.1.2. Positioning

To localize the acoustic source, we chose to work with the well established multilateration technique based on TDOA (Time Difference of Arrival). Multilateration is a technique that is used to determine the position of a source of a signal in space based on the time-of-arrival (TOA) of the sound at different locations. The TOA of a signal can be measured using various techniques, such as cross-correlation which compares the waveform of the signal at two different locations and computes a measure of the similarity between them. Cross-correlation is a statistical measure of the similarity between two signals as a function of the time lag between them. It is often used to estimate the time delay or time-of-arrival (TOA) of a signal by determining the time lag that maximizes the similarity between the two signals. Cross-correlation is widely used in a variety of applications, including audio and speech processing, image processing, and radar signal processing. To compute the cross-correlation between two signals, x(t) and y(t), the signals are first aligned in time by shifting one of the signals by a certain time lag. The cross-correlation function, rxy(tau), is then defined as the integral of the product of the aligned signals over all time:

$$rxy(tau) = \text{integral from } -\infty \text{ to } \infty \text{ of } x(t)*y(t-tau)\, dt$$

The cross-correlation function is a measure of the similarity between the signals as a function of the time lag tau. The time lag that maximizes the cross-correlation function is taken to be the time delay or TOA of the signal y(t).This measure can be used to estimate the TDOA of the signal by taking the argmax of the cross correlation function. We chose the GCC-PHAT cross correlation implementation because it is robust to noise and can accurately estimate the TOA of a signal even in the presence of other signals or interference.. GCC-PHAT (Generalized Cross-Correlation with Phase Transform) is a method for estimating the time delay or time-of-arrival (TOA) of a signal using cross-correlation. It is a variant of the cross-correlation method that is specifically designed for use with microphone arrays, where the signals from the microphones are highly correlated due to the close proximity of the microphones. Taking the time difference measurements and the known positions of the measurement locations, we obtain a system of hyperbolic equations. The equations are fed into a scalar function optimizer to solve for the location of gunfire.

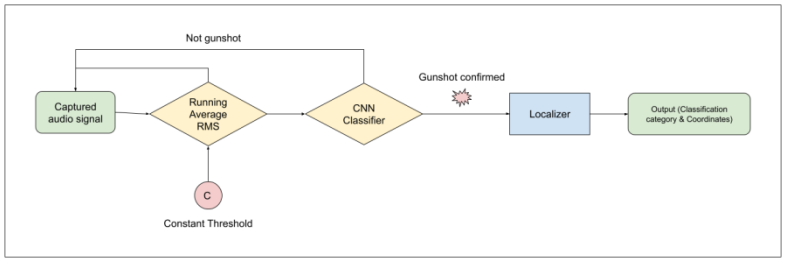

Figure 2. Overview of the RPI System

3.2. Mobile APP

To demonstrate the potential of our mass shooting alert and response system for users, we developed a native iOS mobile app client to receive the classifier and localizer's output information in real time [14]. We installed a prototype gunshot classification and localization system in Santa Margarita High School located in California. Any faculty or student of a school where the system is in place, employees of a company, or general public traveling through an airport will all have access to real time information in case of an incident. When an acoustic sensor detects an impulsive signal and classifies it as a gunshot, the localization module instantly computes the location of the gunfire before instantly sending it to a backend server. We chose Firebase Realtime Database as the cloud infrastructure due to its ease of development for prototyping and realtime callback implementation, allowing the client app to constantly listen for new database changes. This architecture allows our system to be installed with extreme ease and low cost, fully utilizing existing AC (alternating current) power and internet infrastructures. The acoustic sensor's Raspberry Pi single board computer can be powered by an AC plug or a battery pack, while all the data pipeline travels through the HTTPS network infrastructure. All the sensor information is analyzed and processed almost instantly and directly related to victims of a mass shooting through a mobile client, without requiring any human monitoring or intervention.

4. EXPERIMENT

4.1. Experiment 1

In this experiment, a comparison between our CNN model and YAMNet pre-trained model was performed to evaluate the accuracy of our CNN model. YAMNet is a pre-trained neural network that employs the MobileNetV1 depth wise-separable convolution architecture. It can use an audio waveform as input and make independent predictions for each of the 521 audio events from the Audio Set corpus.Internally, the model extracts "frames" from the audio signal and processes batches of these frames [2].

To evaluate the accuracy of the CNN approach, 200 audio recorders were randomly picked from UrbanSound8K dataset in which 96 audio recorders were from diverse sound categories (air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, jackhammer, siren and street_music) and 104 audio recorders were from gunshot sounds [1]. Next, the audio features were extracted from each audio file and performed the prediction of the audio data with both models. Finally, confusion matrices were used to visualize the result. For the YAMNet prediction, before visualizing the confusion matrices, an extra step was performed to match the YAMNet classes with our model classes (no gunshot and gunshot). If the sound prediction matches any firearms sounds, which are Explosion, Gunshot, gunfire, Machine gun, Fusillade, Artillery fire, or Cap gun, then the prediction is classified to " gunshot" otherwise it is classified to "no gun shot".

The experiment result shows that the Our CNN model achieves better results than the YAMNet model. In Figure 3, we can observe that confusion matrices show that 96 gunshot sounds were predicted as a "gunshot" and 8 gunshot audio were predicted "no gunshot". On the other hands, for the other urban sounds (air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, jackhammer, siren and street_music), the result shows that 95 urban sounds were predicted to be "no gunshot" while one urban sound was predicted to be "gunshot". Thus, the accuracy of the CNN model was 95 %.
In Figure 4, we can observe that confusion matrices show that 56 gunshot sounds were predicted as a "gunshot" and 48 gunshot audio were predicted "no gunshot". While the other urban sounds prediction shows that 96 urban sounds were predicted to be "no gunshot".

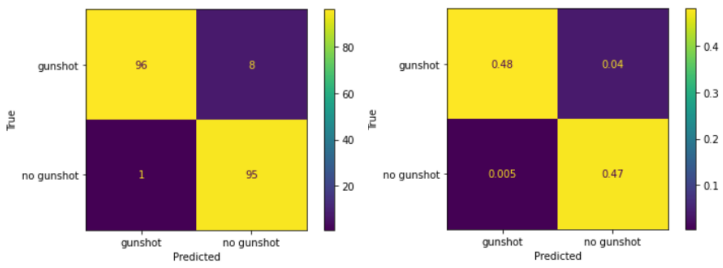

Figure 3. Sound Classification - CNN Model

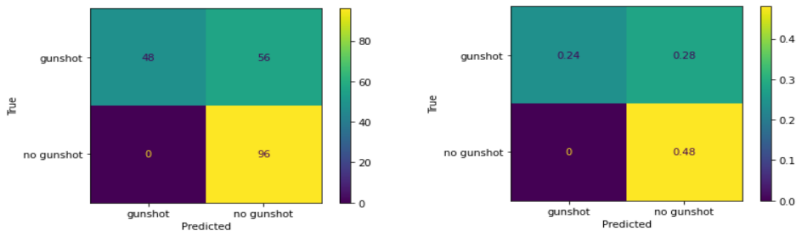

Figure 4. Sound Classification - YAMNet model

4.2. Experiment 2

To evaluate the accuracy of the CNN model, 120 similar sounds, which are 40 fireworks, 40 clapping and 40 thunderstorms, were used to observe if the sounds are predicted to be “gunshot” or “no gunshot”. For the dataset we used the ESC-50 dataset. The ESC-50 dataset is a labeled collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification [3]. The dataset consists of 5-second-long recordings organized into 50 semantic classes (with 40 examples per class) loosely arranged into 5 major categories: Animals, Natural soundscapes and water sounds, Human, non-speech sounds, domestic sounds, urban noises. Similarly, to experiment 1, the sound features were extracted from each sound file and the CNN model was utilized to predict and classify the sound.

The result shows that the accuracy of thunderstorm sounds is higher in comparison to the firework and clapping sound. (see Figure 5) For the thunderstorm sound, the study shows that 33 sounds were predicted to be “no gun shot” while 7 sounds were predicted to be “gun_shot”. The clapping sound result shows that 30 sounds were predicted to be “no gun shot” while 10 sounds were predicted to be “gun_shot”. Finally, the firework sound outcome shows that 17 sounds were predicted to be “no gun shot” while 23 sounds were predicted to be “gun_shot”. Thus, the accuracy of our CNN model using similar sounds is 67%. We believe that the accuracy of our model is a little bit low since when we trained the model, these sound categories were not included in our dataset, so the model does not have enough data to differentiate between the “gunshot” and “no gunshot” sound. Even Though, the accuracy was 67%, we concluded that the model performs well since it did utilize similar sounds in its training.

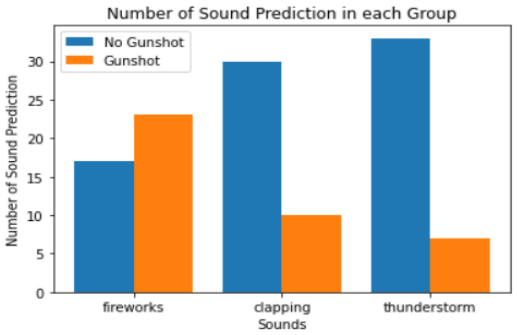

Figure 5. Sound Gunshot Classification using fireworks, clapping and thunderstorm sounds

5. RELATED WORK

In "Evaluation of Gunshot Detection Algorithms", the paper compares and evaluates many popular gunshot detection preprocessing algorithms with an emphasize on feasibility of the algorithm running on a low powered VLSI (Very Large Scale Integration) circuit [4]. While the paper concludes that correlation against a gunshot template as a highly accurate gunshot classification method when compared against other preprocessors, it is important to note the experiment setting and the acoustic environment. The research was conducted in an outdoor setting with the system intended to prevent illegal hunting, and thus a very low noise floor and absence of other impulsive signals can be assumed. In their classification system, the preprocessed signal is directly fed into a simple RMS threshold classifier, using the loudness level of the audio after preprocessing as a discriminator for gunfire. However, this is largely inapplicable to a much more sophisticated urban setting when many noises share similar temporal power characteristics as gunshot. Our MFCC combined with deep learning network classifier takes into account the spectral characteristics of sound and appears to be better suited for gunshot detection in urban settings.

In "Gunshot Detection in Noisy Environments", the paper concludes that the correlation against a template technique is superior in performance while being much cheaper in computational cost when compared against advanced algorithms like MFCC fed into a HMM (Hidden Markov Model) [5]. It is surprising that the paper concludes that the correlation method is robust, even when faced with similar impulsive sounds like balloon pops and claps and even with extreme levels of background noise added. While template correlation initially appears to be an appealing algorithm for our use case, simple testing on a large gunshot dataset quickly revealed its deficiencies. The caliper of the weapon, the weapon type, the use of accessories like a muzzle suppressor, and the firing mode can all significantly affect the temporal domain of the signal. When faced with a large dataset assorted with different weapons fired in different characteristics, the template correlation technique's ability to discriminate significantly decreases. When applied on gunshot audio recordings in 2048 sample windows with 256 sample overlap and computing correlation using Pearson's Correlation Coefficient, the template correlation method frequently fails to discriminate between the gunshot and other impulsively characterized urban noises. In our system, we deemed a properly optimized MFCC algorithm sufficiently efficient and used a CNN (Convolutional Neural Network) rather than HMM. The CNN as a deep learning model far exceeds the learning capabilities of HMM, a traditional statistical model based on states.

In An empirical evaluation: gunshot detection system and its effectiveness on police practices, this study analyzed the effectiveness of a gunshot detection system in a southeastern Massachusetts city with a high incidence of violent crime [6]. Data was collected from the police dispatch log and analyzed using a quasi-experimental design to determine the impact of the system on police ability to identify, investigate, and prosecute gun-related crimes.

The paper concludes that gunshot detection systems like ShotSpotter, are indeed effective in reducing crime rate by improving response times, decreasing dispatch time, and improving case action outcome. When compared against the commercial ShotSpotter system, our mass shooting classification and localization system provides numerous improvements. Our mass distributed system can be deployed at a much larger scale than ShotSpotter while remaining economically feasible to governments and businesses. Our system also does not require constant 24/7 human monitoring to confirm the classification like ShotSpotter does. By using state-of-the-art audio feature extraction techniques and classification models like MFCC and a trained CNN model, we are able to achieve extremely high sensitivity while keeping a low specificity. We also achieved better localization accuracy than ShotSpotter's advertised 20-30 meter radius, even in highly reverberant indoor environments.

6. CONCLUSIONS

One approach for improving mass shooting survivability is through the use of machine learning and audio analysis [15]. By analyzing audio recordings of mass shooting events in real-time, it is possible to identify and classify different sounds, such as gunshots and screams, and use this information to alert individuals to the presence of a shooter and take appropriate action. Additionally, by using source localization techniques, it is possible to identify the location of the shooter within the event space, enabling individuals to take cover or evacuate the area.

In this paper, we present a real-world deployable system that not only incorporates acoustic classification and localization, but also data aggregation to broadcast the threat to end users through a mobile application in real time. Our approach involves the use of Raspberry Pi-based microphone sensors mounted throughout a public space, which are capable of live processing of audio and sending notification alerts to a mobile application in real-time when a gunshot is detected.

Overall, our work represents a significant step forward in the development of systems for improving mass shooting survivability, and has the potential to save many lives in the future. The use of Raspberry Pi-based microphone sensors allows for the deployment of our system at a low cost, making it accessible to a wide range of organizations and individuals. Additionally, the real-time notification alerts provided by our system allow individuals to take immediate action in the event of a mass shooting, increasing their chances of survival. The effectiveness of our approach through a series of qualitative analysis on real-world mass shooting scenarios show that our system is able to accurately classify different sounds and locate the shooter with high precision, leading to improved survivability for individuals caught in mass shooting situations.

One of the main limitations in our current classification system is the tendency to predict false positives, especially when the microphone is overloaded beyond the maximum sound pressure level. Even non-impulsive signals like conversational voice that barely resembles the temporal and spectral characteristics of a gunshot can be distorted to a near impulsive signal form in such cases. This can likely be mitigated using both hardware and software approaches, either by using a pair of low and high sensitivity microphones, or using a multi-layered classifier using ensemble learning to reduce false positives. Another limitation exists in our sound localization system that still lacks enough resistance to specific extremely reverberant indoor situations. The cross correlation algorithm performance will be significantly reduced in cases where multi path sound reverberation is present. This limitation will likely be amplified in narrow hallways constructed with reverberant materials like concrete.

A mass shooting response system that effectively classifies and localizes audio can bring tremendous value for saving many innocent lives and reducing economic damage. Numerous unique challenges remain for gunshot classification and localization, from the ambiguity of differentiating between impulsive signals to difficulties of finding TDOA when multipath reverberation is present. These limitations likely will require further research and experimentation with the software algorithm combined with optimizing microphone sensor engineering to better capture the audio information. Further investigation into the practicality and real-world effectiveness of such systems.

REFERENCES

[1] Rahman, Aamer Abdul, and J. Angel Arul Jothi. "Classification of urbansound8k: A study using convolutional neural network and multiple data augmentation techniques." International Conference on Soft Computing and its Engineering Applications. Springer, Singapore, 2021.

[2] Paper, David, and David Paper. "Simple Transfer Learning with TensorFlow Hub." State-of-the-Art Deep Learning Models in TensorFlow: Modern Machine Learning in the Google Colab Ecosystem (2021): 153-169.

[3] Piczak, Karol J. "ESC: Dataset for environmental sound classification." Proceedings of the 23rd ACM international conference on Multimedia. 2015.

[4] Chacon-Rodriguez, Alfonso, et al. "Evaluation of gunshot detection algorithms." IEEE Transactions on Circuits and Systems I: Regular Papers 58.2 (2010): 363-373.

[5] Freire, Izabela L., and José A. Apolinário Jr. "Gunshot detection in noisy environments." Proceeding of the 7th International Telecommunications Symposium, Manaus, Brazil. Vol. 1. No. 4. 2010.

[6] Choi, Kyung-Shick, Mitch Librett, and Taylor J. Collins. "An empirical evaluation: gunshot detection system and its effectiveness on police practices." Police Practice and Research 15.1 (2014): 48-61.

[7] Fox, James Alan, and Monica J. DeLateur. "Mass shootings in America: moving beyond Newtown." Homicide studies 18.1 (2014): 125-145.

[8] Star, Susan Leigh, and Geoffrey C. Bowker. "How to infrastructure." Handbook of new media: Social shaping and social consequences of ICTs (2006): 230-245.

[9] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. "A novel approach for MFCC feature extraction." 2010 4th International Conference on Signal Processing and Communication Systems. IEEE, 2010.

[10] Shultz, James M., et al. "Multiple vantage points on the mental health effects of mass shootings." Current psychiatry reports 16.9 (2014): 1-17.

[11] Lu, Lie, Hong-Jiang Zhang, and Hao Jiang. "Content analysis for audio classification and segmentation." IEEE Transactions on speech and audio processing 10.7 (2002): 504-516.

[12] Morehead, Alex, et al. "Low cost gunshot detection using deep learning on the raspberry pi." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.

[13] Gustafsson, Fredrik, and Fredrik Gunnarsson. "Positioning using time-difference of arrival measurements." 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).. Vol. 6. IEEE, 2003.

[14] Joorabchi, Mona Erfani, and Ali Mesbah. "Reverse engineering iOS mobile applications." 2012 19th Working Conference on Reverse Engineering. IEEE, 2012.

[15] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245 (2015): 255-260.