Scream Detection Amid Noisy Background to Reduce Crime Rate Using Machine Learning

1Kailash Chandra Bandhu, 2Arpit Deo, 3Ratnesh Litoriya, 4Rashi Gupta, 5Rashmeet Chugga, 6Rishabh Solanki

1kailashchandra.bandhu@gmail.com, 2 deo.arpit33@yahoo.com,  3 litoriya.ratnesh@gmail.com,

4 rashi99506gupta@gmail.com , 5 rashmeetchugga@gmail.com , 6 rishabhsolanki362@gmail.com

1,2,3,4,5,6Department of Computer Science and Engineering, Medi-Caps ,, Indore, India

**Abstract:** Each person’s security is a major concern in the increasing crime rate in today’s era, especially for women. There is need of better solutions to report the surrounding situation of people to their family members as well as law enforcement for timely prevention of crime. There are several solutions available for this issue but the proposed work offers an alternative approach with higher accuracy and efficiency. The proposed work implements a meticulous three phase approach employing SVM based classifiers and multilayer perceptron model was adopted to effectively segregate human distress sounds from surrounding noise and further differentiate screams from shouts. The core of the classification system involved training SVM based classifiers using Mel-frequency cepstral coefficients (MFCC) as feature vectors by filtering through a pool of 400 audio sets. The culmination of these methodologies and analyses revealed an impressive performance, showcasing the system's ability to achieve a commendable 99% accuracy rate.

**Keywords:** MFCC, SVM, Scream detection, Multilayer perceptron model, Perceptron, Distress sound.

1. **Introduction**

The rising crime rates globally have become a major concern for law enforcement agencies and civilians alike. Addressing this challenge requires a multifaceted approach that includes better situational awareness of crime-prone areas and improved response times. One area of focus is leveraging audio signals for enhanced situational responsiveness [1]. This involves understanding the critical role of audio signals in discerning situations, character traits, temporal context, and environmental settings.

The purpose behind the development of this project was a specialized investigation conducted to develop a distress audio event classification system capable of accurately categorizing various audio events such as ambient noise, screams, and shouts. The study started by defining a scream as a high-pitched vocalized sound lacking phonological structure, which formed the basis for the research objective. The ultimate goal was to detect crimes in real-time using machine learning and deep learning concepts, specifically targeting scream sounds captured by users' device microphones [2]. The methodology included a comprehensive selection process from 400 audio datasets. These datasets underwent a two-phase selection to determine the most relevant feature subsets for training and testing classifiers. The classifiers, based on Support Vector Machines (SVM), underwent rigorous analysis and testing across various feature subsets, utilizing both linear and radial basis function (RBF) kernels to decipher and predict incoming sound signals' nature [3] [4].

Recognizing the urgent need to address the escalating crime rates, the study emphasized leveraging cutting-edge technologies like machine learning and deep learning to bolster situational awareness and expedite reaction times. The use of Mel-frequency cepstral coefficients (MFCC) was pivotal in extracting pertinent features from audio data, enabling the discrimination and accurate identification of 'scream' sounds within the collected data.[5] Additionally, a Multilayer Perceptron Model was integrated to enhance classification accuracy by identifying intricate patterns within audio signals and adjusting weights to achieve more accurate outcomes [6].

The primary objective of this initiative was to facilitate an immediate and automated response to potential criminal incidents. Upon detecting a 'scream' sound, the system would swiftly analyze the situation, corroborate criminal activity likelihood, and relay pertinent information, including GPS coordinates, to predetermined contacts like law enforcement agencies or emergency services. This innovative approach aimed not only to enhance real-time crime detection but also to expedite emergency response times, contributing to a safer environment [7]. The core focus of the endeavor was identifying crucial auditory cues, specifically 'scream' sounds, within ambient noise captured by users' device microphones. The widespread use of smartphones and smart devices presented an unprecedented opportunity to transform these gadgets into powerful tools for crime prevention [8]. By harnessing machine learning algorithms grounded in deep learning principles, the goal was to sift through background noise and pinpoint distress instances indicative of potential criminal activity [9].

The study's significance lies in its potential to revolutionize public safety through technology integration. By combining advanced audio signal processing techniques with machine learning and deep learning algorithms, the system aimed to provide law enforcement and emergency services with real-time alerts and actionable data, enabling them to respond swiftly and effectively to criminal incidents [10]. This proactive approach not only enhances crime prevention but also fosters a sense of security and trust within communities.

Moving forward, continuous research and development in the field of audio-based crime detection are crucial. This includes refining algorithms, improving accuracy, expanding the range of detectable audio events, and integrating with existing emergency response systems. Additionally, addressing privacy and ethical considerations, such as data protection and consent, is paramount to ensure the responsible and effective deployment of such technologies in real-world scenarios [11]. In conclusion, the integration of machine learning and deep learning algorithms for real-time crime detection based on audio signals represents a significant step towards improving public safety. By harnessing the power of technology and innovative methodologies, we can create smarter, more responsive systems that contribute to a safer and more secure society for all.

1. **Literature Review**

# This work presented a novel approach to automatically detect human screams amidst noisy background sounds. They used unsupervised method that relies on advanced signal processing and machine learning techniques. Their approach demonstrated robustness in challenging acoustic settings, making it suitable for applications like emergency response systems and security monitoring. By effectively isolating human screams from a cacophony of sounds, this research contributed to enhancing the safety and efficiency of audio-based surveillance and detection systems [12].

# In this work the development of an autonomous CCTV surveillance system was outlined with a focus on improving audio recognition. It acknowledged the limitations of traditional visual surveillance and emphasized the importance of audio processing. The paper detailed the use of a dataset for training and testing, employing Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and Deep Boltzmann Machines (DBM) for classification. Experimental results showed promise in enhancing audio recognition precision compared to traditional methods like GMM kernel and SVM. The text concluded by suggesting future improvements, including dataset expansion and addressing multi-class recognition challenges, making it a valuable contribution to autonomous surveillance technology [13].

# This study focused on developing a robust system using computer vision and deep learning to automatically identify illicit activities and detect screams in video and audio data. It addressed the challenges of automation in this context. The methodology likely involved CNNs for visual tasks and RNNs for audio processing. The paper discussed the dataset used, its size, and sources. Results demonstrated the system's accuracy, reported through metrics like accuracy, precision, recall, and F1-score. The discussion highlighted potential applications and any limitations, while the conclusion underscored the significance of this work for real-world applications in security and surveillance [14].

# This study introduced an innovative audio-based video surveillance system designed to automatically identify and locate anomalous audio events like screams and gunshots in a public square. This system employed two parallel Gaussian Mixture Model (GMM) classifiers, each trained with different audio features, to discriminate these events from background noise. Acoustic source localization was achieved by calculating the time differences of signal arrivals at a microphone array and using a linear-correction least square algorithm. Experimental results demonstrated impressive precision, with a 93% detection accuracy and one-degree source direction estimation, even in challenging conditions with a Signal-to-Noise Ratio (SNR) of 10dB.The paper also mentioned plans for real time implementations of the system in a public square in Milan [15].

# This work presented an audio event detection system designed to automatically distinguish between ambient noise, screams, and gunshots in noisy environments. The system utilized two parallel Gaussian Mixture Model (GMM) classifiers for the classification task, with each classifier trained on different sets of audio features. These features, selected from a pool of 47 options, were determined through a two-step process that involved assembling feature subsets of increasing size and assessing their performance through classifier training and testing. This approach resulted in an optimal feature vector dimension, improving efficiency compared to wrapper feature selection methods. The paper's experiments, which involved a variety of gunshots and screams mixed with ambient noise at different Signal-to-Noise Ratios (SNRs), demonstrated the system's effectiveness, achieving a 90% precision rate at an 8% false rejection rate [16].

# This work explained a novel application of machine learning for detecting victim's screams in burning sites. The authors recognized the critical need for early detection of individuals in distress within burning sites, as prompt rescue operations are essential for saving lives. To address this challenge, they proposed a machine learning-based approach that utilizes audio data to identify screams, a key auditory cue in such emergencies. The paper outlined the methodology, including data collection and preprocessing, feature extraction, and the development of a machine learning model. While the paper did not provide specific results, it highlighted the importance of this research direction and set the stage for future work in the development of effective scream detection systems to improve emergency response in burning sites [17].

# Through this exploratory study, researchers aimed to identify tantrums by detecting screams in home audio recordings using transfer machine learning. They collected a dataset of audio clips which included different kinds of sounds but they emphasized tantrum-related screams, and applied transfer machine learning techniques to develop an automated system. This system compared features categorized for tantrums with each audio clip and made a decision based on the likeliness of the sound [18].

This study focused on the detection of sarcasm using a combination of machine learning classifiers and rule-based approaches. The objective was to developed a robust system for recognizing sarcastic statements in text data. To achieve this, the researchers collected a dataset containing both sarcastic and non-sarcastic text examples, which was used to train machine learning models. These models were designed to identify patterns and linguistic cues associated with sarcasm. In addition to machine learning techniques, a rule-based approach was implemented to capture specific linguistic features and context that often accompany sarcastic expressions. By combining the strengths of both approaches, the study aimed to improve the accuracy and reliability of sarcasm detection. Preliminary results showed promise, with the combined approach demonstrating the potential to effectively identify sarcasm in text. Such a system could have applications in sentiment analysis, social media monitoring, and chatbot development. Further research and fine-tuning are necessary to enhance the model's performance and applicability in real-world scenarios [19].

# This study investigated the application of machine learning algorithms for detecting sarcasm in Twitter data. It assessed a range of research papers and studies focused on the development and evaluation of sarcasm detection models specifically tailored to the unique context of Twitter. The review highlighted the diverse approaches and techniques employed in these studies, including the use of linguistic features, sentiment analysis, and contextual information to identify sarcasm. It discussed the challenges associated with Twitter's brevity, slang, and evolving language patterns, which necessitated specialized approaches. The findings of the systematic review emphasized the potential of machine learning algorithms in effectively detecting sarcasm in Twitter, making it relevant for applications like sentiment analysis and social media monitoring. However, the review also underscored the need for further research to improve the accuracy and robustness of these models, considering the dynamic and ever- changing nature of Twitter conversations [20].

# Scream Detection in heavy metal music is a specialized task aimed at identifying instances of aggressive or intense vocalizations within the genre This unique area of audio analysis involved the development of algorithms and models to automatically recognize screams, growls, or other high-energy vocal techniques commonly used in heavy metal music. The challenge lied in distinguishing these vocalizations from the overall audio context, including the accompanying instruments, which were often loud and chaotic. Researchers and music analysts utilized various techniques, including spectral analysis, pattern recognition, and machine learning, to pinpoint these vocal expressions accurately. Scream detection in heavy metal music has practical applications in music indexing, recommendation systems, and genre classification. It can also be a valuable tool for music enthusiasts, researchers, and performers looking to explore and understand the distinctive vocal elements in heavy metal compositions. While this area of research is niche, it showcases the versatility of audio analysis in music and highlights the unique characteristics of the heavy metal genre [21].

# The detection of emotions through text analysis using machine learning is a burgeoning field with various applications. This approach involved training algorithms to recognize and classify human emotions based on the content of written text, such as social media posts, emails, or customer reviews. Natural language processing (NLP) techniques were employed to extract features and sentiment from the text data. Machine learning models, including support vector machines, recurrent neural networks, and deep learning architectures, were used to identify emotional cues in text. These cues include keywords, sentence structure, and sentiment polarity. Researchers used labeled datasets to train and evaluate these models, with common emotions being happiness, anger, sadness, and fear. The practical applications of emotion detection through text analysis are diverse, including sentiment analysis for business intelligence, customer feedback analysis, mental health monitoring, and social media monitoring. As machine learning techniques continue to advance, this field holds promise for more accurate and widespread emotion detection in text data, contributing to improved decision-making and understanding of human sentiment [22]

# Forest fire detected using machine learning as a critical application of technology aimed at early and accurate identification of wildfires, reducing their devastating impact. This approach relied on a combination of sensors, image analysis, and machine learning algorithms to detect and respond to forest fires swiftly. Machine learning models were trained to analyze various data sources, such as satellite imagery, thermal imaging, and environmental sensors, to identify signs of a forest fire. These models could recognize factors like smoke, increased temperature, and rapid changes in vegetation, which are indicative of fire outbreaks. Once an anomaly was detected, the system could trigger alerts and response mechanisms, including notifying authorities and initiating firefighting efforts. The advantages of this technology were numerous, including faster response times, reduced environmental damage, and enhanced safety for both the environment and human lives. As machine learning continues to evolve, forest fire detection systems are becoming more reliable, making them a valuable tool in preventing and managing wildfires [23].

# Event detection for an audio-based surveillance system is a critical component in enhancing security and monitoring in various environments. This technology involved using audio sensors to capture and analyze sounds, identifying specific events or activities that may pose security concerns. Machine learning and signal processing techniques played a pivotal role in this process. These systems were trained to recognize predefined events, such as breaking glass, gunshots, verbal altercations, or other unusual noises, within audio data. They could operate in real-time or analyze stored audio recordings, providing immediate alerts and responses when suspicious events are detected. Audio-based surveillance systems have application in various security contexts, including smart cities, public transportation, and critical infrastructure protection. They contribute to rapid response times, increased situational awareness, and proactive threat mitigation. Continuous advancements in machine learning and audio processing technologies are expected to make these systems even more effective in identifying and responding to events of interest, thereby enhancing security and safety across a wide range of settings [24].

Noise-robust scream detection technique using band-limited spectral entropy is an advance technique in the field of audio analysis and machine learning designed to identify human screams accurately in noisy environments. This approach focused on extracting specific acoustic features, such as spectral entropy within specific frequency bands, to enhance the detection of screams while mitigating the impact of background noise. By isolating the spectral entropy within defined frequency ranges, this method improved the discrimination of scream-like patterns from other sounds, even in acoustically challenging surroundings. Machine learning models were trained to recognize these distinctive spectral characteristics, making the detection process more robust and accurate. Applications for noise-robust scream detection are extensive, including public safety and security systems, as well as smart home devices for emergency response. This technology holds the potential to swiftly alert authorities or initiate protective measures in response to distress calls, significantly enhancing the safety and well-being of individuals in diverse environments. Advances in machine learning and audio analysis continue to refine the effectiveness of this technique in real-world applications [25].

Deep neural networks for the automatic detection of screams and shouted speech in subway trains represent a cutting-edge approach to ensure the safety and security of public transportation systems. This technology was designed to identify and differentiate emergency situations involving loud vocalizations from typical subway noise, offering timely responses to potential threats. These deep neural networks leveraged extensive training datasets to recognize the acoustic features of screams and shouts, which often exhibit unique pitch and intensity characteristics. They could operate in real time, continuously monitoring audio streams from onboard microphones. When a scream or shouted speech was detected, the system triggers were alerted, enabling quick responses by security personnel or emergency services. The application of deep neural networks in this context was instrumental in enhancing passenger safety, deterring unruly behavior, and swiftly addressing emergencies. It played a crucial role in maintaining a secure and efficient public transportation system, making it a valuable addition to urban infrastructure. Ongoing advancements in machine learning and neural networks contribute to the refinement of these systems, making them more accurate and reliable over time [26].

This work discussed the creation of an affordable, portable device that could accept real-time audio distress signals and process it using a multi-headed CNN architecture. This model outperformed other models and achieved a recognition rate of 79.67%,. While future work may involve further architectural exploration and IoT integration ,this innovation had promising implications for enhancing urban safety and emergency response systems [27].

This study highlighted the importance of audio analytics in surveillance systems, emphasizing its cost-effectiveness and lower data volume compared to video analytics. It introduced a deep learning model, combining CNN and RNN, for identifying critical audio events in urban environments. The model achieved impressive accuracy (81% to 95%) in classifying incidents like gunshots, explosions, and sirens. The study showcased the value of audio analytics in both open and closed security settings, including the use of smartphones for sound detection and emergency notifications. This research underscored the potential of deep learning, particularly the CNN-RNN model, in enhancing urban security by recognizing dangerous urban sounds and complementing video surveillance effectively [28].

This work addressed the growing concern of violent crimes in public spaces, emphasizing the (GMM) for scream detection, with SVM exhibiting a low False Acceptance Rate and GMM showing sensitivity to scream sounds but with a higher False Rejection Rate. The research highlighted the distinct merits of both methods and suggests the potential for improved security systems by combining them. This paper contributed to enhancing safety in public areas through innovative audio-based surveillance technologies [29].

**Table 1: Literature Summary**

| **S. No.** | **Work Done** | **Technologies Used** | **Result** | **Limitations** | **Research Gap** |
| --- | --- | --- | --- | --- | --- |
| 1 | The paper introduces an unsupervised method using signal processing and machine learning to robustly detect human screams in noisy environments, benefiting safety and surveillance systems [12] | Machine Learning, Noise Reduction, Feature Extraction and Dimensionality Reduction. | The ability to detect screams in continuous audio streams over a range of noise type have been shown up to 10dB SNR. | Limited Applicability to Low SNR Environments.  Limited Diversity of Noisy Environments Computational Complexity. | Unsupervised Approach Limitations.  Lack of Validation with Real Screams. |
| 2 | The paper focuses on autonomous CCTV surveillance, emphasizing audio recognition using MFCC and DBM. It enhances precision over traditional methods, suggesting future improvements for multi-class recognition [13] | MFCC feature extraction, Deep Boltzmann Machines classification, GMM Kernel, SVM. | Average accuracy using GMM kernel and SVM is 91.14%. | Dataset Computational Complexity and User Friendliness. | Real-World Deployment Model Explainability. |
| 3 | The paper develops a robust system using computer vision and deep learning to automate the detection of illicit activities and screams in video and audio data, highlighting its real- world security applications [14] | CNN and RNN. | VGGNet19 gave best training accuracy results with the average accuracy of 0.92, recall 0.91 and F1-score of 0.93. | Dependency on Pre-trained Models. False Positives and Negatives Data Quality and Diversity. | Privacy and Ethical Concerns Response Time |
| 4 | The paper presents an audio-based surveillance system that identifies screams and gunshots using dual GMM classifiers and precise acoustic source localization, achieving high detection accuracy in challenging conditions [15]. | GMM Classifier. | 93% detection accuracy, signal to noise ratio of 10dB | Generalization to Different Environments and Noise Robustness False Positives. | Real-Time Implementation Challenges and Real-World False Alarm Mitigation |
| 5 | The paper presents an audio event detection system differentiating ambient noise, screams, and gunshots using dual GMM classifiers and optimized feature selection [16]. | GMM classifier. | 90% precision rate at an 8% false rejection rate | Limited Evaluation Metrics False Positives | Hardware and Cost Considerations Scalability |
| 6 | The paper introduces a machine learning- based approach for detecting victim's screams in burning sites, emphasizing the importance of early distress detection in emergencies [17]. | Machine learning and Feature extraction. | - | Lack of Specific Results and Generalization to Real-World Scenarios | Scalability Real-Time Processing |
| 7 | The paper explores scream detection to identify tantrums in home audio recordings, using transfer machine learning with a dataset focusing on tantrum- related screams [18]. | Transfer machine learning. | - | Data Collection and Quality Scream Detection Accuracy | Integration with Support Systems Cost and Maintenance Considerations |
| 8 | The study combines machine learning and rule-based methods to detect sarcasm in text, aiming to enhance accuracy for sentiment analysis and social media applications [19]. | Machine learning techniques and rule- based approach. | - | - | - |
| 9 | The systematic review delves into machine learning algorithms for detecting sarcasm in Twitter, assessing studies with approaches like linguistic features and sentiment analysis. Despite highlighting the potential for applications like sentiment analysis, it stresses the need for further research to improve accuracy, considering Twitter's dynamic and evolving language[20]. | Machine learning techniques, focusing on approaches like linguistic features and sentiment analysis. | - | Contextual Challenges: Twitter's brevity and the potential for ambiguous language make it challenging to capture nuanced contextual cues, impacting the accuracy of sarcasm detection models. | Improved Model Robustness: There may be a need for research aimed at developing more robust models that can adapt to the rapidly changing linguistic landscape of Twitter, ensuring sustained accuracy over time. |
| 10 | Specialized in heavy metal music, scream detection involves algorithms recognizing intense vocalizations. Challenges include distinguishing screams amid chaotic instrumentals. Techniques like spectral analysis and machine learning are used. Applications include music indexing and genre classification [21]. | Spectral analysis and machine learning are used for accurate detection. | - | The challenge of accurately distinguishing screams amid chaotic instrumentals. The paper lacks specific performance metrics. | A research gap exists in quantifying the accuracy of scream detection models within the complex context of heavy metal music, leaving room for further exploration and refinement. |
| 11 | Text-based emotion detection, using machine learning and NLP, categorizes human emotions in written content. Applications include business intelligence, mental health monitoring, and social media analysis [22]. | Machine learning and NLP, training algorithms Techniques like support vector machines, recurrent neural networks, and deep learning identify emotional cues. | - | Limitations may involve cultural context challenges. | A research gap in real-time emotion detection accuracy. |
| 12 | Forest fire detection using machine learning combines sensors and algorithms to swiftly identify wildfires through data analysis [23]. | Technologies used involve satellite imagery, thermal imaging, and environmental sensors. | - | Limitations is false positives | Refining real- time detection accuracy. |
| 13 | Scream and gunshot detection using audio analysis and machine learning enhances public safety in noisy environments. These models rely on features like pitch and spectral characteristics. Specific results are not provided [24]. | Technologies used involve real-time audio analysis. | - | Limitations may include false positive. | A research gap exists in refining detection accuracy, and ongoing advancements in machine learning are expected to contribute to system improvements. |
| 14 | The technique employs band-limited spectral entropy for noise-robust scream detection, enhancing accuracy in noisy environments. It isolates spectral features, improving discrimination. Specific results are not provided [25]. | The technology used involves advanced audio analysis and machine learning. | - | Limitations may include potential false positives. | A research gap exists in addressing real-world deployment challenges. |
| 15 | Deep neural networks for scream and shouted speech detection in subway trains use extensive training datasets for recognizing unique acoustic features [26]. | Deep neural networks and ongoing advancements in machine learning. | - | Privacy Concerns: Continuous audio monitoring raises privacy concerns, necessitating careful consideration of ethical implications and legal frameworks to protect passengers' privacy. | A research gap exists in adapting the technology to diverse subway environments. |
| 16 | The paper introduces an affordable real-time distress signal recognition system for urban emergencies, using deep learning on a Raspberry Pi [27]. | deep learning, feature extraction (MFCCs, Mel spectrograms), and a Raspberry Pi for real-time processing. | The proposed model achieved a recognition rate of 79.67%, emphasizing cost- effectiveness. | Limitations may include noise sensitivity. | Future work involves architectural exploration and IoT integration. |
| 17 | The paper underscores the cost-effectiveness of audio analytics, presenting a CNN- RNN deep learning model achieving high accuracy (81% to 95%) in identifying critical urban audio events [28]. | CNN-RNN deep learning model. | CNN-RNN deep learning model achieving high accuracy (81% to 95%). | Noise Sensitivity: The model may be sensitive to background noise, potentially leading to false positives or reduced accuracy in acoustically challenging urban environments. | A research gap may include scalability and real-world deployment challenges. |
| 18 | This research focuses on leveraging unique acoustic properties of human screams, including high pitch and irregular contours, to enhance speaker recognition systems [29]. | Acoustic Analysis Techniques, Feature Extraction Algorithms and Machine Learning or Pattern Recognition Models. | - | Scope of Acoustic Features: The research may not cover all possible acoustic features of screams, and the effectiveness of the identified features in diverse contexts might be a limitation. | Comprehensive Analysis of Acoustic Features: Further research could explore additional acoustic features of screams, ensuring a more comprehensive understanding for broader applications. |

1. **Proposed Methodology**

To optimize the performance of our models for detecting human scream voices, we will carefully tune the parameters rather than relying solely on default settings. This adjustment is crucial for ensuring the reliability and effectiveness of our detection model.

Our methodology involves several key steps. Firstly, we will process the dataset for training purposes. This involves converting the audio samples into Mel-frequency cepstral coefficient (MFCC) feature vectors, which serve as the input for our models. To obtain the final MFCC features, we will utilize Discrete Cosine Transform (DCT) on the energy (Ek) values. The formula for DCT is as follows:

(1)

Where N is number of triangular bandpass filters and L is number of MFCC’s.

Once the dataset is prepared, we will proceed with training and testing our supervised machine learning models. Our approach involves using two different models for detection: a three-stage supervised learning approach employing a Support Vector Machine (SVM) based classifier initially, followed by utilizing a Multilayer Perceptron model.

In the SVM-based classifier, we will employ two types of Support Vector Classifiers (SVCs): one with a linear kernel and another with an RBF (radial basis function) kernel. For the RBF kernel, the formula used is \(exp(-\gamma \times (x - x')^2)\), where (**gamma**) must be greater than 0. By following this approach, we aim to develop a robust and optimized model capable of accurately detecting human scream and shout voices.

**3.1. Algorithm**

phase\_one\_prediction\_svm = phase\_one\_svm\_predict([audio])[0]

phase\_one\_prediction\_mlp = phase\_one\_mlp\_predict([audio])[0]

if phase\_one\_prediction\_svm == "Noise":

return "Noise"

if phase\_one\_prediction\_svm == "Speech":

return "Speech"

if phase\_one\_prediction\_svm == "Scream":

return "Scream"

if phase\_one\_prediction\_svm == "Shout":

return "Shout"

if phase\_one\_prediction\_svm in [2, 3, 4]:

phase\_two\_prediction\_svm = phase\_two\_svm\_predict([audio])[0]

if phase\_two\_prediction\_svm in [2, 3, 4]:

phase\_three\_prediction\_svm = phase\_three\_svm\_predict([audio])[0]

if phase\_three\_prediction\_svm in [2, 3]:

svm\_combined\_prediction = phase\_one\_prediction\_svm == phase\_two\_prediction\_svm == phase\_three\_prediction\_svm

mlp\_combined\_prediction = phase\_one\_prediction\_mlp == phase\_two\_prediction\_mlp == phase\_three\_prediction\_mlp

final\_prediction = svm\_combined\_prediction and mlp\_combined\_prediction

if final\_prediction:

return "Scream"

else:

return "Shout"

else:

return "Shout"

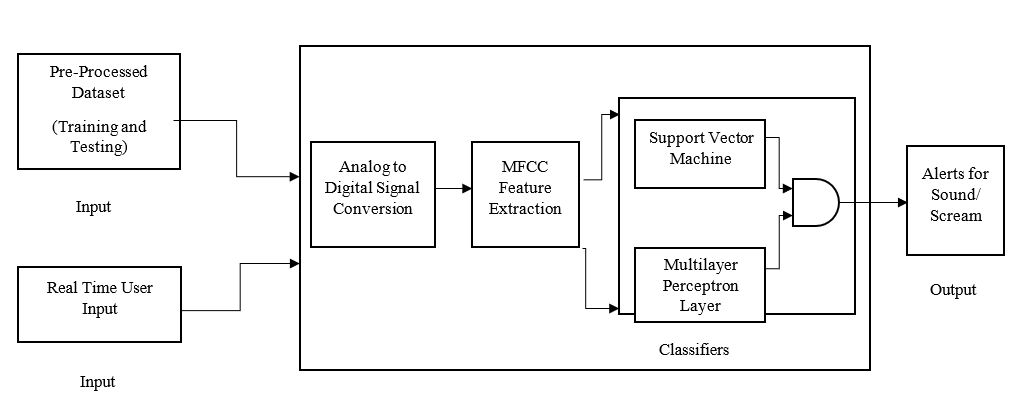
else:

return "Speech"

else:

return "Noise"

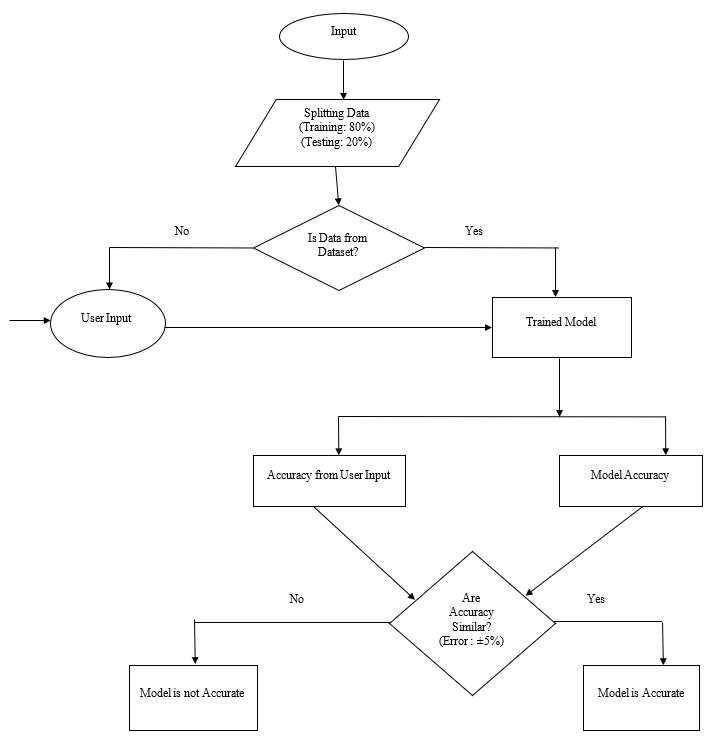
**3.2. Architecture**

****

**Figure 1: Proposed Architecture**

Figure 1 depicts the complete architecture that this study is based on. The audio input, when fed into the ML model goes through conversions and multiple classifiers to finally detect the output as scream or sound

**3.3. Flowchart of Proposed Work**

****

**Figure 2: Flowchart**

Figure 2 describes the flow control of data in the ML model. It shows the two diverging paths of training and testing data and how they are ultimately compared to determine the accuracy of the prediction model.

# Working Process: The project runs as a background process, continuously monitoring audio input for potential human scream. The kivy framework is utilized to create a user-friendly interface that can run seamlessly in the background. This application captures real time audio input, either from a microphone or an audio file. For the preprocessing, librosa library is used , it is a python package for audio analysis and can be used to extract relevant features from the audio data. Features such as mel frequency cepstral coefficient and common characteristics that are used in scream detection. A pre-trained SVM model is employed for classifying audio samples as screams or non-screams. The SVM model is trained using labeled data, distinguishing between positive and negative instances When a potential scream is detected by an SVM model, an alert is generated, the alert may include details about the detected scream and its probability score. The application use location services such as GPS to determine the user's location . The location information is included in the alert message to help the police to reach crime scene quickly

1. **Design and Implementation**
   1. **Code Snippets of SVM classifier and Multilayer Perceptron**

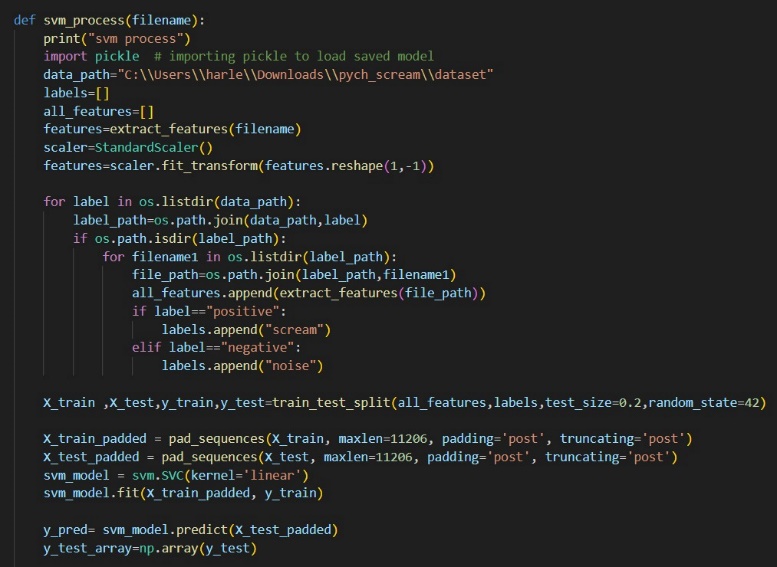


Figure 3: Code Snippet of SVM classifier

Figure 3 depicts the code snippet for the SVM classifier used in this ML model. The data is trained and tested using the SVM classifier and it is subsequently given a tag of either noise, speech, scream or shout. After this, the result is passed to the next stage, that is, the Multilayer Perceptron Model.

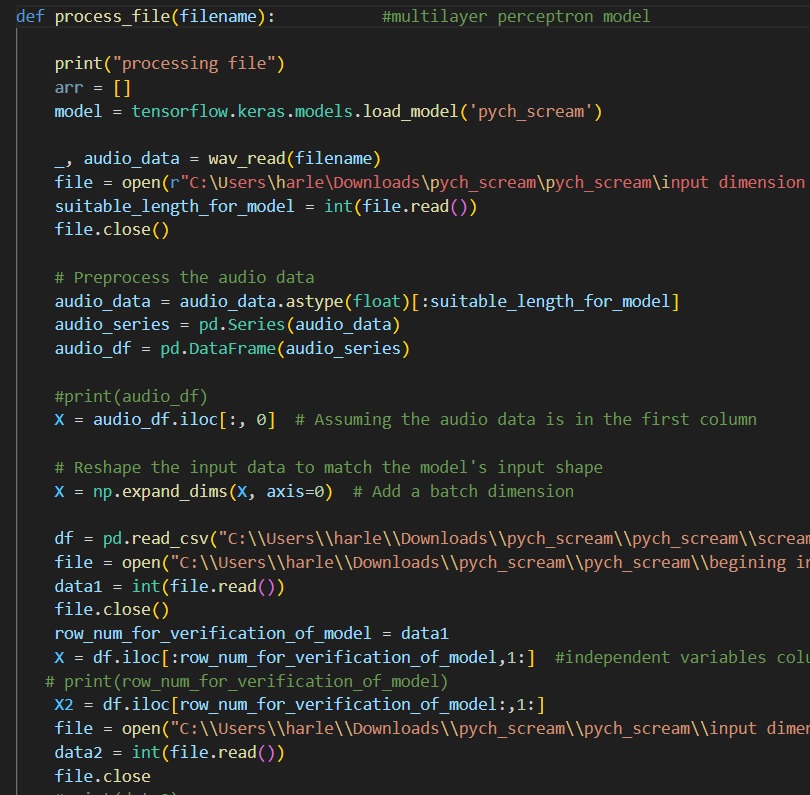


Figure 4: Code snippet of Multilayer Perceptron Model

Figure 4 depicts the code snippet of the Multilayer Perceptron model that is used after SVM to process the data. The data is again labelled in different categories of noise, speech, scream or shout. The label results from both models are then compared and the final output is decided after taking an AND product. Multiple parameters of the Multilayer Perceptron model are then hyper tuned to achieve optimal accuracy.

**4.1** **Input :**

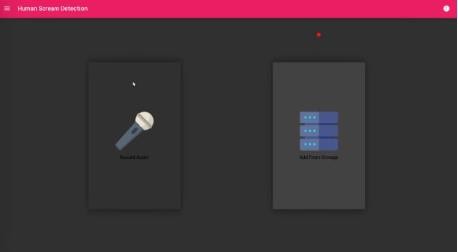


Figure 5: Input by the User

Figure 5 shows the user interface of the model which allows the user to feed either real-time audio into the system through a microphone or test an audio saved in the storage.



Figure 6: Voice Recording

Figure 6 shows the interface visible to the user while the input audio is being processed by the system. The input audio clip duration is set at 10 seconds.

* 1. **Output**

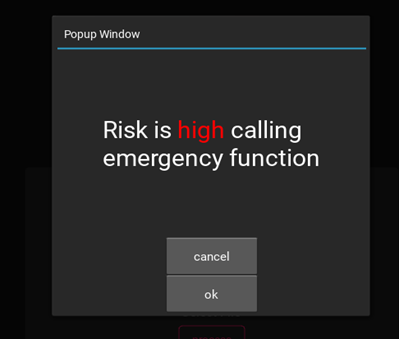


Figure 7: Model Output Case I

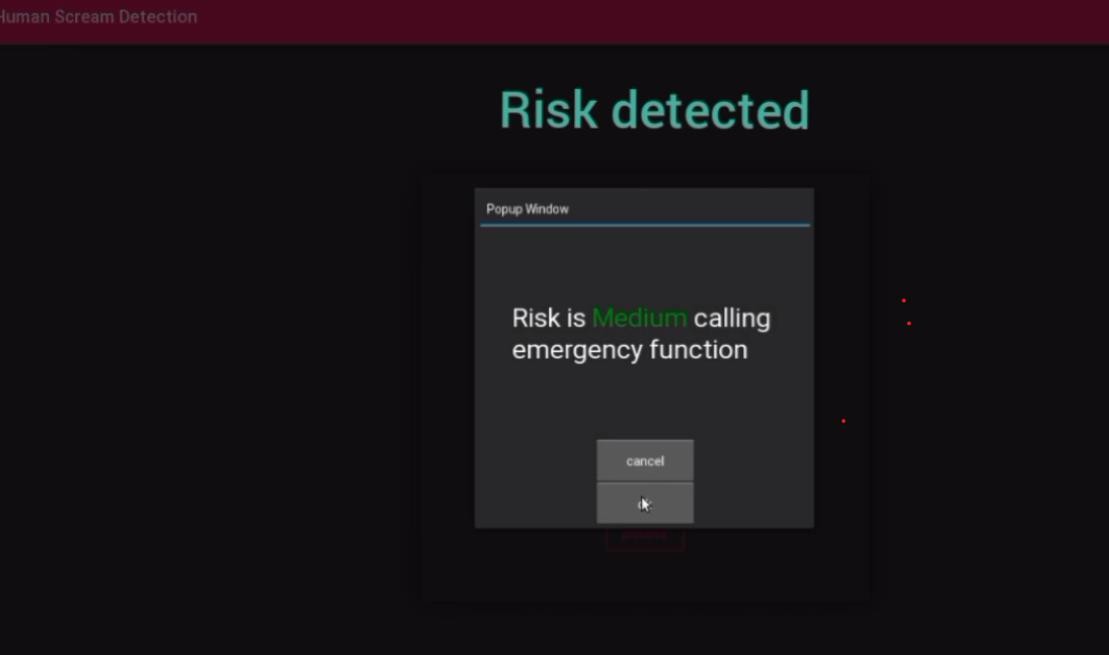


Figure 8: Model Output Case II

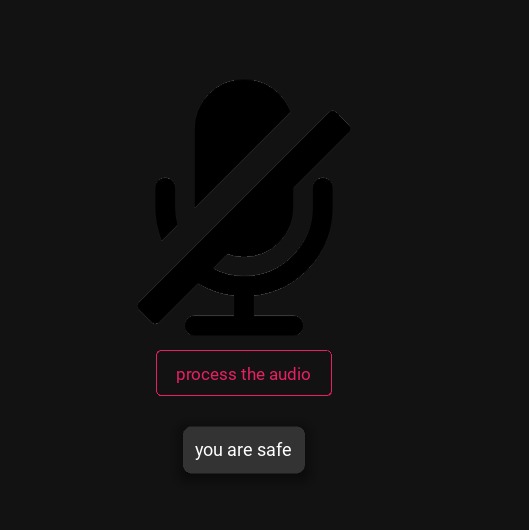


Figure 9: Model Output Case III

# Figure 7, 8 and 9 show the three different variations of the output possible by the ML model. The output can show a message of "you are safe" in case of background noise. The message can be "Medium Risk" in case of mild disturbances and far-away shouts. The message would be "High Risk" in case of high-pitched noise in close proximity.

1. **Result and Analysis**

After conducting data processing through both the Support Vector Machine (SVM) classifier and the Multilayer Perceptron Model, our next objective is to enhance the overall accuracy of the model.

This will be achieved through the systematic tuning of various hyperparameters within the Multilayer Perceptron Model until an optimal accuracy level is attained. The hyperparameters slated for tuning include the minimum and maximum values, validation split ratio, batch size, and activation function state.

The tuning process involves evaluating the model's performance under different configurations of these hyperparameters. By systematically varying each hyperparameter and assessing its impact on the model's accuracy, we aim to identify the most effective combination that maximizes predictive performance.

**Table 2: Hyperparameter Tuning – Minimum Value and Maximum Value**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No.** | **Minimum Value** | **Maximum Value** | **Accuracy (%)** |
| 1 | 20 | 64 | 89.51 |
| 2 | 10 | 64 | 90.91 |
| 3 | 30 | 64 | 88.81 |
| 4 | 10 | 60 | 95.10 |
| 5 | 10 | 50 | 88.81 |

**Best (10, 60) = 95.10%**

Table 2 depicts the different accuracy values calculated for the model on changing the minimum and maximum value range. On comparison, it is found that minimum value of 10 and maximum value of 60 gives the best accuracy of 95.10 so these values are made constant. These hyperparameters are often used in preprocessing steps such as feature scaling or normalization. Setting appropriate minimum and maximum values helps in scaling the input features to a range that the model can effectively learn from. Incorrect scaling can lead to numerical instability or slow convergence**.**

**Table 3: Hyperparameter Tuning – Validation Split**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Validation Split** | **Accuracy (%)** |
| 1 | 0.2 | 92.31 |
| 2 | 0.1 | 94.41 |
| 3 | 0.05 | 97.20 |
| 4 | 0.02 | 98.60 |
| 5 | 0.02 | 99.30 |

**Best (0.02) = 99.30%**

Table 3 depicts the different accuracy values calculated for the model on changing the values of the validation split. On comparison, it is found that a validation split of 0.02 gives the best accuracy of 99.30 so it is made constant**.** This hyperparameter determines the portion of the data used for validation during training. A larger validation split means more data is used for validation, which can lead to a more accurate estimation of model performance. However, it reduces the amount of data available for training, potentially leading to overfitting if not carefully chosen**.**

**Table 4: Hyperparameter Tuning – Batch Size**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Batch Size** | **Accuracy (%)** |
| 1 | 70 | 97.90 |
| 2 | 80 | 97.90 |
| 3 | 40 | 98.60 |
| 4 | 50 | 99.30 |
| 5 | 30 | 99.30 |

**Best (30) = 99.30%**

Table 4 depicts the different accuracy values calculated for the model on changing the values of the batch size. On comparison, it is found that a batch size of 30 maintains the best accuracy of 99.30 so it is made constant. The batch size determines the number of training examples utilized in one iteration. A smaller batch size might lead to faster convergence but with more noise in the updates, while a larger batch size might lead to more stable updates but slower convergence. The optimal batch size depends on the dataset and the model architecture.

**Table 5: Hyperparameter Tuning – Activation State**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Activation State** | **Accuracy (%)** |
| 1 | Linear | 47.56 |
| 2 | Elu | 52.45 |
| 3 | Relu | 55.94 |
| 4 | Tanh | 92.31 |
| 5 | Sigmoid | 98.60 |

**Best (Sigmoid) = 98.60**

Table 5 depicts the different accuracy values calculated for the model on changing the values of the activation state. On comparison, it is found that the activation state sigmoid gives the best accuracy of 98.60 so it is made constant. The activation function introduces non-linearity to the model, enabling it to learn complex patterns in the data. Common activation functions include ReLU, sigmoid, and tanh. The choice of activation function can greatly impact the model's ability to learn and generalize from the data. For example, ReLU is often preferred in hidden layers due to its simplicity and effectiveness in preventing the vanishing gradient problem. Optimizing these hyperparameters through tuning can lead to improved model performance and generalization on unseen data.

**Table 6: Accuracy Comparison**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Model** | **Model Accuracy (%)** |
| 1 | Proposed Support Vector Machine | 100.00 |
| 2 | Proposed Multilayer Perceptron Model | 98.60 |
| 3 | Support Vector Machine | 92.14 [3] |
| 4 | Multilayer Perceptron Model | 82.33 [3] |
| 5 | Logistic Regression | 71.01 [3] |
| 6 | K-NN Algorithm | 80.53 [3] |

Table 6 presents a comprehensive comparison of accuracy rates across various classifiers for distress audio sound prediction. Notably, the proposed SVM classifier demonstrates an exceptional accuracy prediction level of 100%. This marks a substantial improvement over previously published results, where the SVM classifier typically achieved an accuracy rate of around 92%.

These findings underscore the efficacy of the proposed SVM classifier in accurately predicting distress audio sounds. The significant increase in accuracy compared to alternative classifiers highlights its superiority in handling the complexities and nuances inherent in such predictive tasks.

Given the diverse range of cases and input types encountered in predictive tasks, a single classifier may not suffice to accurately predict all scenarios. Therefore, an additional model capable of handling more complex test cases was deemed necessary. Table 6 also provides an in-depth analysis of accuracy rates achieved by different classifiers, including SVM, when applied to this multifaceted model.

The proposed multilayer perceptron model emerges as a standout performer with an impressive accuracy score of 98%. This signifies a notable enhancement over previously published results, where the multilayer perceptron model achieved an accuracy of 82%. In comparison, logistic regression and the K-NN model yielded accuracy scores of 71% and 80%, respectively.

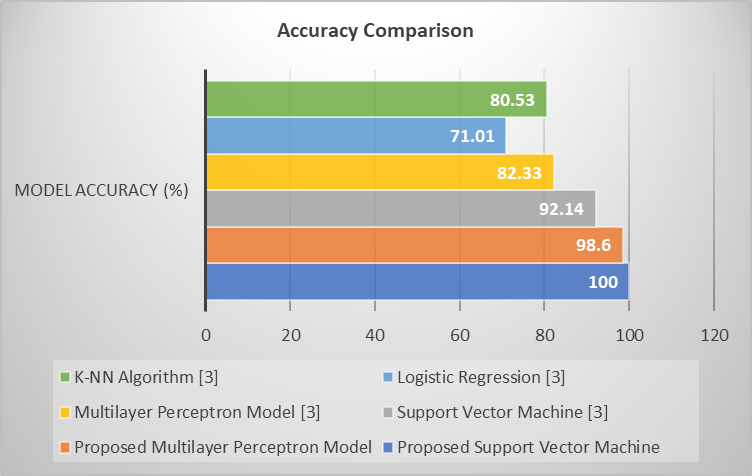


Figure 10: Accuracy Comparison

Figure 10 shows the accuracy comparison of different classifiers and models. This figure shows the accuracy of 100% in Proposed Support Vector Machine and 98.6% in Proposed Multilayer Perceptron Model. While among the previously published work , this graph shows an accuracy 92.14% in Support Vector Machine, 82.33% Multilayer Perceptron Model, 71% in Logistic Regression and 80% in K-NN Algorithm. These results make it clear that the Proposed Support Vector Machine and Proposed Multilayer Perceptron Model outperformed all the other classifiers There ability to handle complex test cases and achieve a significantly higher accuracy rate compared to alternative classifiers makes them a compelling choice for such predictive tasks.

**Table 7: Truth Table**

|  |  |  |
| --- | --- | --- |
| **Proposed SVM** | **Proposed MLP** | **Result** |
| If it predicts human scream (Shout) | If it predicts human scream (Shout) | Human Scream (High Risk) |
| If it predicts human scream (Shout) | If it predicts human scream | Human Scream (Medium Risk) |
| If it predicts human scream | If it predicts human scream | Human Scream(Medium Risk) |
| If it predicts human scream(Shout) | If it predicts background noise | Human Scream(Medium Risk) |
| If it predicts background noise | If it predicts human scream | Speech (Low Risk) |
| If it predicts background noise | If it predicts background noise | Background Noise (No Risk) |

This table depicts a decision-making framework for classifying audio data using two machine learning models: a Support Vector Machine (SVM) and a Multilayer Perceptron (MLP). The input conditions are whether the models predict a human scream (Shout) or background noise. The output categories are "Human Scream (High Risk)," "Human Scream (Medium Risk)," "Speech (Low Risk)," and "Background Noise (No Risk)." The table illustrates how the combination of predictions from both models leads to different classifications. For instance, when both SVM and MLP predict a human scream, the result is classified as "Human Scream (Medium Risk)." Conversely, if the SVM predicts background noise while MLP predicts a human scream, the classification shifts to "Speech (Low Risk)." This table highlights the nuanced decision-making process based on the predictions of the SVM and MLP models, enabling the categorization of audio data into distinct risk levels and types of sounds.

1. **Conclusion**

The culmination of this extensive research suggests a promising avenue for addressing scenarios demanding robust crime detection through audio signals. The study underscores the efficacy of employing a dual-model approach comprising both supervised machine learning, specifically SVM-based classification, and deep learning models based on neurons, such as perceptrons. The synergistic fusion of these models emerges as a formidable strategy in achieving optimal results while minimizing the likelihood of erroneous predictions.

The integration of SVM-based classification and neural network-based deep learning models presents a comprehensive solution to enhance the accuracy and reliability of audio-based crime detection systems. While the SVM classifier provided a perfect accuracy score of 100%, the Multilayer Perceptron Model also gave an exceptional accuracy score of 98.6%. By leveraging the strengths of both approaches, this study advocates for a holistic methodology that harnesses the precision of supervised learning alongside the intricate pattern recognition capabilities inherent in deep neural networks. The significance of this approach lies in its potential to mitigate the risks associated with misclassifications and false predictions. The collaboration between the SVM-based classification, adept at discerning intricate audio patterns, and the neural network-driven deep learning model, capable of sophisticated feature extraction and nuanced decision-making, establishes a robust framework for crime detection from audio data.

Practical implementation of this research could be facilitated through well-known Python libraries such as Librosa, specifically for extracting Mel-frequency cepstral coefficients (MFCC) crucial in audio feature extraction. Additionally, leveraging libraries like Keras, renowned for its versatility in building neural networks, would streamline the implementation of the dual-model architecture proposed in this study. The modular nature of these Python libraries, coupled with their extensive functionalities and ease of integration, could expedite the deployment of the proposed audio-based crime detection system. By utilizing these tools in tandem with the methodology outlined in this research, stakeholders in law enforcement, security systems, and related domains could benefit from a sophisticated and reliable framework for crime detection and prevention through audio signal analysis.

1. **Future Scope**

With the current model achieving an impressive accuracy of 98%, there are several avenues for enhancing its capabilities further. One area of focus is improving the accuracy of detecting negative and abusive language, thereby strengthening the system's ability to prevent potential incidents such as kidnapping or threats of violence. To achieve this, investment in advanced natural language processing (NLP) techniques is warranted, along with a continuous process of updating the model with fresh data to ensure its effectiveness over time.

Moreover, there's potential to extend the application beyond desktops to mobile platforms, making the scream detection system more accessible and versatile. Additionally, enhancing the system to operate continuously, 24\*7, could be beneficial, possibly by integrating additional characteristics and algorithms. Beyond merely detecting screams, the system could evolve to predict potential incidents based on the context of the detected sounds. By analyzing the tone and content of conversations, it could provide valuable insights and serve as an additional layer of preventive measures against various threats.

Furthermore, collaboration with law enforcement agencies could be explored to integrate the scream detection system into their emergency response systems. This collaboration could lead to faster and more coordinated responses to critical situations, ultimately enhancing public safety and security.

1. **References**

[1] A. Sharma and S. Kaul, “Two-Stage Supervised Learning-Based Method to Detect Screams and Cries in Urban Environments,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 24, no. 2, pp. 290–299, Feb. 2016, doi: 10.1109/TASLP.2015.2506264.

[2] S. Nazir, M. Awais, S. Malik, and F. Nazir, “A Review on Scream Classification for Situation Understanding,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, 2018, doi: 10.14569/IJACSA.2018.090809.

[3] A. Shankhdhar, Rachit, V. Kumar, and Y. Mathur, “Human Scream Detection Through Three-Stage Supervised Learning and Deep Learning,” 2021, pp. 379–390. doi: 10.1007/978-981-16-1395-1\_28.

[4] H. D. Tran and H. Li, “Sound Event Recognition With Probabilistic Distance SVMs,” *IEEE Trans Audio Speech Lang Process*, vol. 19, no. 6, pp. 1556–1568, Aug. 2011, doi: 10.1109/TASL.2010.2093519.

[5] B. Vimal, M. Surya, Darshan, V. S. Sridhar, and A. Ashok, “MFCC Based Audio Classification Using Machine Learning,” in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2021, pp. 1–4. doi: 10.1109/ICCCNT51525.2021.9579881.

[6] Y. Li *et al.*, “Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection,” *Multimed Tools Appl*, vol. 77, no. 1, pp. 897–916, Jan. 2018, doi: 10.1007/s11042-016-4332-z.

[7] F. Cruciani *et al.*, “Feature learning for Human Activity Recognition using Convolutional Neural Networks,” *CCF Transactions on Pervasive Computing and Interaction*, vol. 2, no. 1, pp. 18–32, Mar. 2020, doi: 10.1007/s42486-020-00026-2.

[8] Y. Choi, J. Park, J. Lee, and H. Kim, “Exploring Diverse Feature Extractions for Adversarial Audio Detection,” *IEEE Access*, vol. 11, pp. 2351–2360, 2023, doi: 10.1109/ACCESS.2023.3234110.

[9] V. K. Mittal and B. Yegnanarayana, “Production features for detection of shouted speech,” in *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, IEEE, Jan. 2013, pp. 106–111. doi: 10.1109/CCNC.2013.6488433.

[10] A. Hamza *et al.*, “Deepfake Audio Detection via MFCC Features Using Machine Learning,” *IEEE Access*, vol. 10, pp. 134018–134028, 2022, doi: 10.1109/ACCESS.2022.3231480.

[11] J. Pohjalainen, P. Alku, and T. Kinnunen, *Shout detection in noise*. IEEE, 2011. doi: 10.1109/ICASSP.2011.5947471.

[12] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, “Robust unsupervised detection of human screams in noisy acoustic environments,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2015, pp. 161–165. doi: 10.1109/ICASSP.2015.7177952.

[13] Md. 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, Aug. 2015, pp. 1–4. doi: 10.1109/ICITCS.2015.7292925.

[14] R. Mathur, T. Chintala, and D. Rajeswari, “Identification of Illicit Activities &amp; Scream Detection using Computer Vision &amp; Deep Learning,” in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, May 2022, pp. 1243–1250. doi: 10.1109/ICICCS53718.2022.9787991.

[15] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, IEEE, Sep. 2007, pp. 21–26. doi: 10.1109/AVSS.2007.4425280.

[16] Gerosa L, Valenzise G, Tagliasacchi M, Antonacci F, and Sarti A, “SCREAM AND GUNSHOT DETECTION IN NOISY ENVIRONMENT,” *2007 15th European Signal Processing Conference*, 2007.

[17] F. S. Saeed, A. Al 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, Sep. 2021, doi: 10.3390/app11188425.

[18] R. O’Donovan, E. Sezgin, S. Bambach, E. Butter, and S. Lin, “Detecting Screams From Home Audio Recordings to Identify Tantrums: Exploratory Study Using Transfer Machine Learning,” *JMIR Form Res*, vol. 4, no. 6, p. e18279, Jun. 2020, doi: 10.2196/18279.

[19] K. Sentamilselvan, P. Suresh, G. K. Kamalam, S. Mahendran, and D. Aneri, “Detection on sarcasm using machine learning classifiers and rule based approach,” *IOP Conf Ser Mater Sci Eng*, vol. 1055, no. 1, p. 012105, Feb. 2021, doi: 10.1088/1757-899X/1055/1/012105.

[20] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, “Sarcasm detection using machine learning algorithms in Twitter: A systematic review,” *International Journal of Market Research*, vol. 62, no. 5, pp. 578–598, Sep. 2020, doi: 10.1177/1470785320921779.

[21] V. Kalbag and A. Lerch, “Scream Detection in Heavy Metal Music,” May 2022.

[22] K. Machová, M. Szabóova, J. Paralič, and J. Mičko, “Detection of emotion by text analysis using machine learning,” *Front Psychol*, vol. 14, Sep. 2023, doi: 10.3389/fpsyg.2023.1190326.

[23] K. G. Madhwaraj, V. Asha, A. Vignesh, and S. Akshay Shinde, “Forest Fire Detection using Machine Learning,” in *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal: IEEE, Apr. 2023, pp. 191–196. doi: 10.1109/CSNT57126.2023.10134684.

[24] C. Clavel, T. Ehrette, and G. Richard, “Events Detection for an Audio-Based Surveillance System,” in *2005 IEEE International Conference on Multimedia and Expo*, IEEE, 2005, pp. 1306–1309. doi: 10.1109/ICME.2005.1521669.

[25] N. Hayasaka, A. Kawamura, and N. Sasaoka, “Noise-robust scream detection using band-limited spectral entropy,” *AEU - International Journal of Electronics and Communications*, vol. 76, pp. 117–124, Jun. 2017, doi: 10.1016/j.aeue.2017.04.002.

[26] 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, Mar. 2016, pp. 6460–6464. doi: 10.1109/ICASSP.2016.7472921.

[27] J. F. Gaviria *et al.*, “Deep Learning-Based Portable Device for Audio Distress Signal Recognition in Urban Areas,” *Applied Sciences*, vol. 10, no. 21, p. 7448, Oct. 2020, doi: 10.3390/app10217448.

[28] Z. Momynkulov *et al.*, “Fast Detection and Classification of Dangerous Urban Sounds Using Deep Learning,” *Computers, Materials & Continua*, vol. 75, no. 1, pp. 2191–2208, 2023, doi: 10.32604/cmc.2023.036205.

[29] Suk-Hwan Jung and Yong-Joo Chung, “Scream sound detection based on SVM and GMM,” in *May 2017 Thailand International Conferences*, EAP, May 2017. doi: 10.17758/EAP.AE0517126.