# Enhancing Crime Prevention via Human Scream Detection with Deep Learning and Machine Learning

MSc Research Project
MSc Data Analytics

Ashraf Hussain Raheem Basha
Student ID: x23191899

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | | | |
|---|---|---|---|
| **Student Name:** | Ashraf Hussain Raheem Basha | | |
| **Student ID:** | x23191899 | | |
| **Programme:** | MSc Data Analytics | **Year:** | 2024 |
| **Module:** | MSc Research Project | | |
| **Supervisor:** | Vikas Tomer | | |
| **Submission Due Date:** | 12 December 2024 | | |
| **Project Title:** | Enhancing Crime Prevention via Human Scream Detection with Deep Learning and Machine Learning | | |
| **Word Count:** | 8883 | **Page Count:** | 23 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Ashraf Hussain Raheem Basha

**Date:** 12 December 2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Enhancing Crime Prevention via Human Scream Detection with Deep Learning and Machine Learning

Ashraf Hussain Raheem Basha
X23191899@student.ncirl.ie

**Abstract**

Scream recognition is one of the very important components of audio analysis with prospects to be used in cases such as crime prevention, emergency situations, and security. The current study examines the efficacy of machine learning classifiers: Support Vector Machines (SVM), Multilayer Perceptron (MLP), and ResNet-34 deep learning for scream detection. Features including the Mel-Frequency Cepstral Coefficients (MFCC) and spectrogram visualisations were applied when feeding the inputs for classification models.

Advanced data augmentation approaches were used to fix drawbacks such as noise and class skewness so that both clean and noisy data sets were created. Specific performance indicators like precision, recall, F1 score and noise robustness of each model were measures used in the evaluation process. However, when it came to clean well-controlled data, SVM and MLP gave a much better result than ResNet-34, but the latter did have overall higher accuracy, adaptability, and worked better in noisy inputs.

While this paper only provides evidence of the screams' detection, it demonstrates the usefulness of each model separately to help further evaluate their applicability in practice. The objective of the study is to identify the aspects that are needed in the improvement of audio-based safety systems while recommending them in the application of emergency response and public safety.

**Keywords:** Scream Detection, Machine Learning, Deep Learning, Resnet-34, Support Vector Machines, Multilayer Perceptron, Data Augmentation, Spectrograms, Public Safety.

# 1 Introduction

## 1.1 Background

Speech and more specifically human screams is considered to be an important sub-problem of audio processing for the detection of distress signals in various acoustic scenes. This capability is critically important in such applications as public safety, healthcare monitoring or monitoring the operations of a smart city as early detection is often a matter of life and death. Nevertheless, the problem of identifying screams in noisy and unpredicted contexts still remains an essential one that requires additional investigation and advancements. Earlier scream analysis techniques used only features that were manually extracted from the scream and they included the MFCC, chroma features and spectral contrast which was vital in the temporal and frequency behaviour. These were usually used in models such as SVM and MLP (Gao et al. 2023) (Mac et al. 2022). However, these approaches were less stable in noisy conditions and needed much more feature engineering. The latest developments have expanded the CNN framework with spectrograms and mel-spectrograms so that CNN can treat sound as an image. It has improved both the quality and reliability of the models in the applications (Gao et al. 2021) (Laffitte et al. 2016).

## 1.2 Motivation

The requirement for effective scream detection systems has risen due to the variability of the need for the same. In public safety, such systems can help identify incidents occurring in real-time in real-time thus improving the safety of cities (Saeed et al. 2021)(Zaheer et al. 2015). Similarly, in healthcare, they can pick signals of distressed screams from suffering patients, information critical to the attendants particularly for the young and the aged (Mac et al. 2022)(Handa & Vig 2020). In addition, implementations of scream detection into smart city configurations could potentially improve safety by both recognizing when certain conditions arise and by controlling responses to them appropriately (Weimin Huang et al. 2010)However, challenges persist. Environmental interference usually ends up in invalidations or invalid results thereby bringing down the reliability of the system. Also, the small size and heterogeneity of the available data sources hinder replicability and generalizability (Pandey et al. 2023)(Nandwana, Ziaei & Hansen 2015). The specific application of deep learning models is highly efficient but consumes a considerable number of computational resources; therefore, it has limitations if applied in small systems or embedded systems (Sharma & Kaul 2016). Overcoming these problems is crucial to achieving the optimum potential of scream detection technology.

## 1.3 Research Question and Objectives

How do traditional machine learning and deep learning techniques compare in terms of accuracy, robustness, and practicality for human scream detection in real-world scenarios such as crime prevention?

To address this question, the following research objectives have been defined:

- Investigate the state-of-the-art on human scream detection techniques with regard to feature extraction from the audio, data handling approaches and a comparison of the traditional ML and current DL models in this field.
- Design a more highly engineered framework which employs techniques such as data augmentation and spectrogram visualizations and analysis to deal with noise and variation in the environment which is not an ideal scenario for a real life setup of the designed system.
- Implement machine learning and deep learning models individually, by using Standard models such as Support Vector Machines (SVM) and Multilayer Perceptron (MLP) and deep learning architectures such as ResNet-34 in order to compare effectiveness in identifying human screams.
- Assess the performance of these models on balanced and augmented design datasets and use measurements such as precision, recall, F1-score and noise robustness to measure the efficiency and applicability of these models specifically for real-world applications including crime fighting and emergency responses.

## 1.4 Contribution

In this study, a detailed comparative investigation of the applicability of the original ML algorithms including SVM and MLP as well as DL approaches starting from ResNet-34 architecture is performed. This analysis offers substantial overviews of possible advantages and drawbacks of these approaches in relation to exactness, non-sensitivity to noise, and applicability for real-word usage. Second, the research proposes a new carefully designed and augmented dataset together with state-of-the-art techniques for feature extraction thus providing for a reliable and fair comparison of the detection models. These contributions target one of the most significant problems of scream detection, namely, data unbalanced, which increases model stability. Further, this work considers the noise robustness of detection systems as important for practical applications like security and live response to incidents. Thus, assessing model performance under noisy scenarios contributes to the development of the practical applications of scream detection technologies, as provided by the research. Moreover, the results shed light on the importance of developing lightweight and resourceful models capable of implementation on low-end devices which opens the way to expansive and inclusive solutions.

### 1.5 Structure of the Paper

The remainder of this paper is structured as follows:

**Section 2:** A detailed literature review covering the progression of scream detection methodologies, gaps in the literature, and emerging opportunities.

**Section 3:** Methodology describes the approach to experiment setup, dataset preparation and model training.

**Section 4:** Design Specification describes the fundamental outlines in this study such as the system and model architecture used as well as the evaluation framework used.

**Section 5:** Implementation involves showing the process of the machine learning and deep learning models, which include training, validation, and testing.

**Section 6:** The evaluation phase aims at determining the level of performance of the models that were implemented individually.

**Section 7:** In the Results and Discussion section where the model results are evaluated against relevant measures and the study findings are compared with prior research.

**Section 8:** Conclusion and Future Work to give an overview of the contributions of the thesis and to suggest possible research in the next investigations.

## 2 Related Work

This section focuses on the literature review of the methods involving the application of machine learning and deep learning for the identification of human screams. Scream detection is going to be the focus of the choice of publications where the emphasis will be placed on their methodology, outcomes, and shortcomings.

### 2.1 Feature Extraction, Applications and Model Architecture

Several recent works also highlight the necessity of feature extraction as one of the major parts of scream detection. (Gao et al. 2023)(Kalbag & Lerch 2022)(Nazir et al. 2018) suggest that basic techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and chroma features that capture spectral and pitch characteristics are still viable. These features improve the differentiation capability of basic machine learning algorithms such as Support Vector Machines (SVMs). MFCCs are still efficient for short-term analysis of signals separating the signal into frequency bands that resemble human ear perception, while chroma features are helpful for harmonic and tonal characteristics which are crucial for discriminating screams from the background noise. Unlike other works, (Gao et al. 2021) and (Laffitte et al. 2016) propose to use spectrogram-based visualization in order to feed audio into CNN as image data. The use of spectrograms is also advantageous since these represent the signals within the time-frequency domain and bring out features of acoustic significance. This transition, therefore, enhances classification performance notably when the environment is noisy but with a drawback of higher resource utilization. As revealed in the paper (Mac et al. 2022), ResNet-34 takes full advantage of residual connections to protect against vanishing gradients, but at the cost of being complex and computationally demanding. Residual connections in ResNet enable the network to hold accuracy in deeper architectures as it solves issues that older deep networks face. (Rashedi et al. 2024) discusses MLP for non-linear classification and elaborates shows how these models can be beneficial to find out the patterns in the audio signals, which can be challenging anyhow.

Papers such as (Weimin Huang et al. 2010) discuss the applicability of scream detection in elder care and security in home automation systems with the potential of recording with microphones installed in home automated devices. Such systems can allow caregivers or emergency responders to provide help in cases of distress, quickly. Likewise, the paper (Laffitte et al. 2016) also shows the applicability of this model in public transport security with the implementation of scream identification in noisy subway surroundings for boosting the alertness of emergencies. The detection of emotional states based on pitch-based features is presented in (Nazir et al. 2018) as an application for healthcare. In other words, emotional classification is especially valuable when it comes to recognition of distress that may help a therapist grasp some of patient's actions. (Mac et al. 2022) also incorporates end-to-end learning to address patient distress level and demonstrates how efficient data processing pipelines can be trained and constructed even in noisy hospital contexts without having to perform any feature engineering. Screams and cries are dealt in the paper (Sharma & Kaul 2016) using SVM techniques with varying noise levels while using unsupervised methods for Detecting human screams in (Nandwana et al. 2015). These benefits not only impact safety but also pose problems for what private information will continually be recorded through the use of surveillance apps.

Managing data related issues is widely discussed. Supervised datasets present class imbalance issues such as screams that are addressed in (O'Donovan et al. 2020) by oversampling methods such as SMOTE, which synthesizes new samples within the same class in order to attain a balanced data distribution. However, oversampling guarantees sample sufficiency; augmentation methods, as described in (Ojha & Adarsh 2024), time-shifting, pitch-shifting and addition of noise interfere with datasets to mimic realistic scenarios. Time-shifting means the difference in the time relationship of the audio signals, while pitch-shifting involves changing the frequency and making the audio sound like voice or sound environment changes. Noise addition incorporates interference in the form of background noises, which are an accurate representation of acoustical difficulties in real-world settings. Nonetheless, over-augmentation is a dangerous aspect of increasing recall while reducing precision and therefore requires consistent procedures. Legal and ethical issues mostly regarding privacy, particularly in the investigations of scenarios like surveillance are evident in (Laffitte et al. 2016) and (Sharma & Kaul 2016) and Cries in Urban Environments where the authors call for the application of anonymity measures and meta-regulation relating to the GDPR. These issues show that there is a need to enshrine good gathering and utilization of data practices in order to prevent misuse of surveillance technologies.

## 2.2 Gaps and Opportunities

As a result of the numerous research carried out on human scream detection, some of the areas that have not been adequately filled in the current literature are as follows. The reviewed papers show that there are several shortcomings in the papers' approaches and implementations, alongside the depicted opportunities to deal with these issues and (Gao et al. 2023) work with high accuracy by using MFCCs. However, there is no evaluation of the algorithms in various and real, uncontrolled acoustical environments. This gap can be filled by building strong testing from across data sets of different acoustics environments and using noise adaptive procedures including dynamic filters or adaptive noise cancellation. Likewise, the paper (Kalbag & Lerch 2022) employs chroma features and spectral contrast but generalized appendices that may be less suitable for noisy or stochastic conditions as the emphasis is on ideal circumstances. Here, challenges include enhancing the model generalization by incorporating techniques of data augmentation, including adding synthetic noise or generative adversarial network (GAN) to mimic noisy conditions.

(Gao et al. 2021) enhances classification by means of spectrograms. Nevertheless, the problem of computational complexity, which arises when using the proposed algorithm in actual time, stays unsolved. A better architecture can be simply an optimized ResNet through pruning, quantization or a lightweight network such as the MobileNet. Further, (Laffitte et al. 2016) provide good results in public transport situations, but provides weak anonymization methods. It is unknown whether data privacy could be preserved with the same level of detection accuracy through differential privacy mechanisms or by adopting the federated learning method. In (Zaheer et al. 2015)(Rashedi et al. 2024), MLPs are presented for the non-linear relation, but no further architecture to improve the performance is provided. It was also found that single-layered architectures which incorporated MLPs alongside convolutional layers could capture both local and global audio features. However, in (Mac et al. 2022) restricted datasets where adjustments are less frequent and necessary, the model is quite rigid. Increasing data diversity sources of patient audios and incorporating techniques in unsupervised domain adaptation may enhance the generalizability of models across new settings. (Weimin Huang et al. 2010) explains home automation, but the paper does not discuss the aspect of scalability and interaction with the other smart systems. Versions of this that are easily integrated with present smart home platforms and Internet of Things gadgets would improve its performance. Likewise, (Handa & Vig 2020) proposes a new and effective way of categorizing emotions but the model needs larger sets of data to achieve reasonable cross-cultural and cross-linguistic validity. An adaptation of these models to cross-cultural datasets and transfer learning could provide the strategy for translating the models for different language and emotional sets. The (Sharma & Kaul 2016) does not factor in scalability and dynamic learning to cater for ever changing noise characteristics in urban environments. It is likely that reinforcement learning approaches can be used for real-time adjustment to the fluctuation in urban sound environments and noise profiles. On the contrary, (Nandwana et al. 2015) is capable of noise adaptation but sensitive to it, which hampers real-time use except for careful signal setting. To increase the flexibility and deployment of models, meta-learning strategies to automate calibration processes make models more suitable.

A lack of general data sets is a common problem, as mentioned by (Pandey et al. 2023). However, this work does not provide further recommendations on how to work with general data sets. Joint initiatives to develop new large-scale MP3 databases, incorporating a variety of acoustic environments and user-realized noise samples, are promising. However, (O’Donovan et al. 2020) gives a theoretical account of reinforcement learning and oversampling but doesn’t show how they may be put to work. As a result, specialized reinforcement learning algorithms enhanced specific data balancing and feature extraction approaches needed for scream detection could help fill this gap. While augmentation is presented in (Ojha & Adarsh 2024), the actual risk of over-augmentation that could lead to a worsening of the model is not explained. Perhaps that is why it is desirable to set such criteria for augmentation, in which both diversity and relevance are taken into account; and to use auto-augmentation to assess the best strategies to be used when facing certain risks. Spectrogram-based CNNs have large computational load as pointed out in (Laffitte et al. 2016), which makes their implementation unsuitable for resource-limited settings. Another approach could be the use of low rank factorization method, or we could switch to completely different architectures like EfficientNet. As seen in (Gao et al. 2021), transfer learning seems to be a viable strategy but its incorporation into real-world systems is not well realized. Such models aimed at defining a transfer learning approach, where the initial pre-training step was performed on a number of significantly large and diverse audio datasets consistent with the targeted domain-specific fine-tuning step would fill this gap. Another source of low replicability is calibration problems, as well as the absence of unified evaluation protocols in the (Nandwana et al. 2015) file taped in various soundscapes. It has been suggested that unified evaluation protocols should be proposed together with the development of automated calibration methods, which can provide the necessary scale.

## 2.3 Summary of the previous findings

The current literature on scream detection shows remarkable developments in the methodologies, applications, as well as the problems in the field of security, health, and smart cities. Two basic methods, which are basically used for the feature classification are still very relevant today namely Support Vector Machines and Hidden Marked Model, where Mel Frequency Cepstral Coefficients and Chroma features give greatly used features for scream detection. At the same time, such deep learning methods as ResNet-34 and CNNs, spectrogram visualization and automated feature learning brought to the considerably higher level the detection accuracy. Nonetheless, these aspects make many models highly promising but their usage of controlled datasets, and lack of secure, extensive evaluations hinders many of them.

| **Authors** | **Year** | **Datasets Used** | **Methodology** | **Model Used** | **Metrics** | **Value** | **Limitations** | **Future Work** |
|---|---|---|---|---|---|---|---|---|
| (Kalbag & Lerch 2022) | 2022 | 280 minutes of annotated heavy metal songs | Feature extraction (cepstral, spectral) | SVM | Accuracy | 90% | Limited to heavy metal vocals | Expand to other scream contexts |
| (Nazir et al. 2018) | 2018 | Various sources (e.g., Freesound) | Non-speech scream classification | SVM, HMM | Precision, Recall | Precision: 88%, Recall: 85% | Lack of real-world noisy testing | Apply in surveillance and healthcare settings |
| (Gao et al. 2021) | 2021 | Wood defect images, Transfer Learning | Feature extraction, ResNet-34 | ResNet-34 | Accuracy, Recall | Accuracy: 98.4%, Recall: 98.5% | Requires large image datasets | Transfer learning for other domains |
| (Saeed et al. 2021) | 2021 | Custom dataset (fire emergency sounds) | Audio feature extraction, ML classification | SVM, LSTM | Accuracy, Sensitivity | Accuracy: 93%, Sensitivity: 90% | Limited computational power in real-time | Implementation of NVIDIA Jetson for deployment |
| (Pandey et al. 2023) | 2023 | 25+ articles review on scream detection | Taxonomy of scream classification techniques | Various ML models | Comparative analysis | Overview of various ML models' accuracies | Predominantly theoretical, lacks experiments | Practical testing across varied environments |
| (Mac et al. 2022) | 2022 | Spirometry data (medical domain) | Deep learning with multilayer perception | Multilayer Perceptron | Accuracy | Accuracy: 95% | Focused on respiratory sounds only | Expand DL for non-speech detection |
| (Laffitte et al. 2016) | 2016 | Subway sounds | Feature extraction with MFCC and DNN | Deep Neural Networks | Accuracy, F1-Score | Accuracy: 91%, F1-Score: 89% | Limited to subway environments | Extend to other noisy urban environments |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| (Handa & Vig 2020) | 2020 | High-pitch signals (distress and joyful screams) | Acoustic signal analysis for emotion | SVM | Specificity | Specificity: 92% | Focused on high pitch sounds | Broaden to other emotions and contexts |
| (Ojha & Adarsh 2024) | 2024 | Sound database of aggressive situations | MFCC feature extraction, ML and DL models | GMM, HMM, CNN | ROC-AUC, Accuracy | ROC-AUC: 0.98, Accuracy: 98% | Limited robustness for low SNR | Develop integration for real-time applications |
| (Gao et al. 2023) | 2023 | Small fresco image dataset | ResNet-34 with transfer learning | ResNet-34 | Accuracy, Recall | Accuracy: 98.4%, Recall: 98.5% | Requires high-quality image datasets | Use TL models for other cultural artifacts |
| (Zaheer et al. 2015) | 2015 | Self-recorded scream dataset | MFCC feature extraction, Deep Boltzmann Machine | Deep Boltzmann Machine | Accuracy | Accuracy: 100% in controlled conditions | Controlled environment, lacks real noise | Apply in varied real-world environments |
| (O'Donovan et al. 2020) | 2020 | AudioSet, TV show "Supernanny" | Transfer learning for scream detection | Gradient Boosted Tree | ROC-AUC, Precision | ROC-AUC: 0.86, Precision: 42% | Limited accuracy in noisy environments | Validate on diverse home and clinical datasets |
| (Nandwana, Ziaei & Hansen 2015) | 2015 | Custom dataset, noisy environments | Unsupervised detection with CompSeg, T2-BIC | MFCC, PMVDR | Detection accuracy | High accuracy at +10 dB SNR | Performance declines in low SNR | Improve for extreme low SNR environments |
| (Weimin Huang et al. 2010) | 2010 | Self-collected home sound samples | Log energy and high-pitch detection, SVM | SVM | Detection accuracy | Accuracy: 88% | Limited portability to varied settings | Integrate with multi-sensor home care systems |
| (Sharma & Kaul 2016) | 2016 | 100+ hours smartphone-recorded urban audio | Two-stage supervised learning | SVM | Detection rate, FAR | DR: 93.16% at 20 dB, FAR: 4.76% | Limited low-SNR accuracy | Extend to low-SNR urban noise scenarios |
| (Rashedi et al. 2024) | 2023 | Saudi Arabia Stock Returns | Classification and outlier detection | MLP | Accuracy | 92% | Focused on financial data, lacks direct audio application testing | Adapt methodology to audio signals like scream detection |

**Table 1 –** Summary of key findings in Scream Detection Literature

# 3 Research Methodology

This research uses a systematic approach to enhancing scream detection as an area of ML and DL. The methodology includes data gathering, cleaning, feature engineering, processing, experimenting, and assessing, to make the system scalable to the actual business environment. This workflow can be described as sequential and interlinked and is illustrated in Figure 1.

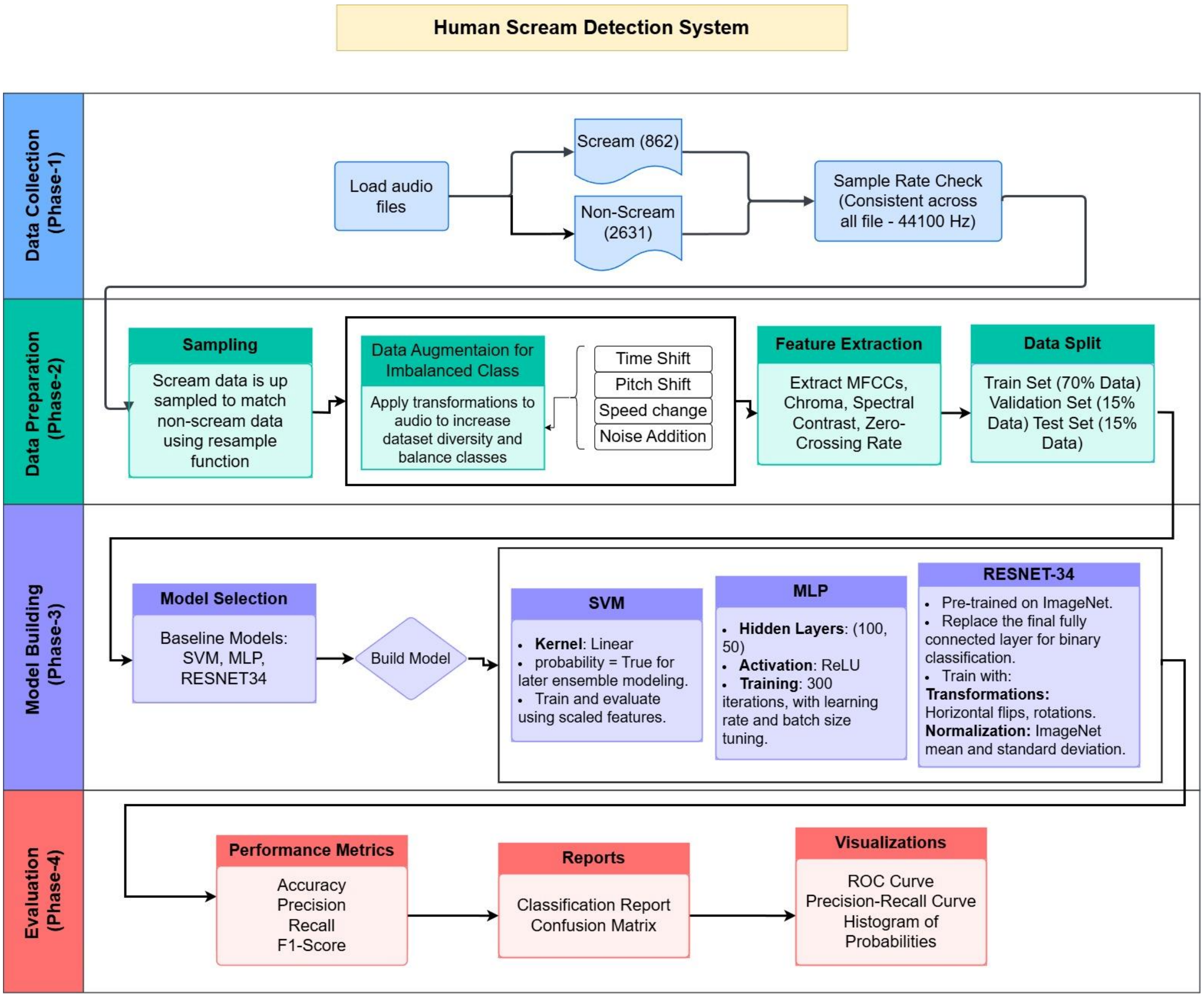
Human Scream Detection System
Data Collection (Phase-1)
Load audio files
Scream (862)
Non-Scream (2631)
Sample Rate Check (Consistent across all file - 44100 Hz)
Data Preparation (Phase-2)
Sampling
Scream data is up sampled to match non-scream data using resample function
Data Augmentaion for Imbalanced Class
Apply transformations to audio to increase dataset diversity and balance classes
Time Shift
Pitch Shift
Speed change
Noise Addition
Feature Extraction
Extract MFCCs, Chroma, Spectral Contrast, Zero-Crossing Rate
Data Split
Train Set (70% Data) Validation Set (15% Data) Test Set (15% Data)
Model Building (Phase-3)
Model Selection
Baseline Models: SVM, MLP, RESNET34
Build Model
SVM
Kernel: Linear
probability = True for later ensemble modeling.
Train and evaluate using scaled features.
MLP
Hidden Layers: (100, 50)
Activation: ReLU
Training: 300 iterations, with learning rate and batch size tuning.
RESNET-34
Pre-trained on ImageNet.
Replace the final fully connected layer for binary classification.
Train with:
Transformations: Horizontal flips, rotations.
Normalization: ImageNet mean and standard deviation.
Evaluation (Phase-4)
Performance Metrics
Accuracy
Precision
Recall
F1-Score
Reports
Classification Report
Confusion Matrix
Visualizations
ROC Curve
Precision-Recall Curve
Histogram of Probabilities

**Figure 1** – Workflow of Human Scream Detection

## 3.1 Data Collection

In this study, data collection is the first and a critical phase in the research process. The audio samples were collected from publicly available archives including AudioSet and Freesound which have a rich variety of audio contents. These sources were selected as they encapsulate a wide variety of real-life audio situations and may comprise of environmental interferences and different voice tones thereby making them suitable for scream recognition tasks. The dataset was classified into two categories: ‘scream’ and ‘non-scream’. The scream class comprised 862 samples, and the non-scream class comprised 2,631 samples. For the purpose of standardisation in processing, all the audio files were brought to same sampling rate at 44,100 Hz. It does this in a way that makes swallowing whole into feature extraction and modelling pipelines easy.

## 3.2 Data Preparation

The dataset consists of spectrogram images obtained from the audio signals classified as ‘yes’ (scream) or ‘no’ (otherwise). The dataset was resampled in order to balance the classes and reduce the problem of class imbalance. Mel-spectrograms were used for spectrograms generation because the analysis of time-frequency information is essential for scream signal detection when the background noise is present. Other data augmentation methods such as time shifting, pitch variation and noise inclusion were also used to increase data variety and reduce dataset vulnerability.

### 3.2.1 Sampling

Due to the skewed nature of the distribution in criteria that maps the two classes, oversampling techniques were used to increase the number of instances of the scream class. Using the resample() function allowed for balancing of the classes in the dataset. This approach reduces the risk of model bias towards the large class and brings the classification accuracy of the instances in the small class closer to baseline.

### 3.2.2 Data Augmentation

To address the variability and scarcity of audio samples, data augmentation techniques were applied:

- **Time Shifting:** Adding temporal variations in audio signals by simply moving the waveform slightly forward or backward in time. It also helps to mimic real-life situations in which the emergence of screams can happen in different ways.
- **Pitch Shifting:** Pitch shifting of the audio samples removing the speed factor for mimicking different vocal range, thus testing resilience of the model against different speakers.
- **Speed Changes:** The speed of playback of audio information is changed periodically to replicate normal intonation fluctuations in human speech.
- **Noise Addition:** Adding artificial ambient noise to the scenes and adding fake noise from the car traffic or crowded places to improve the model when working under noisy environment.

In the process, the changes improved the heterogeneity and firmness of the dataset, ensuring the models are ready to respond to any acoustic conditions.

### 3.2.3 Feature Extraction

Feature extraction also extracts audio data into numerical data for better analysis. This study utilized the following features:

- **Mel-Frequency Cepstral Coefficients (MFCCs):** Obtain the short-term power spectrum of sound which is considered to be an essential feature in any audio analysis.
- **Chroma Features:** Used to denote pitch content of audio signals are important when analyzing the scream for the tonal content.
- **Spectral Contrast:** For noise separation and vocal signal isolation, point out the difference between spectral peaks and valleys.
- **Zero-Crossing Rate:** Estimate the rate in which the sign of the waveform fluctuations, common with the voiced and unvoiced segments.

For deep learning models, spectrograms and mel-spectrograms were computed to build image-based representations for the audio signals. These spectrograms were subsequently resized and normalized for input into the residual network of 34 layers. Spectrograms selected in the current study were visually inspected to ensure that the samples were clean and could be used for training the model.

### 3.2.4 Data Splitting

The preprocessed dataset was divided randomly into three partitions: training (80%), validation (10%) and test (10%). This split allows having reliable evaluation metrics for each subset as well as avoiding data leakage between them. The larger training set gives broad data to ensure robust model learning while the validation and test set bias free.

### 3.2.5 Standardization

Standardize features by subtracting the mean of each feature and then dividing by the standard deviation to a standard variance of 1. This process assures that each of the features considered makes a proportionate input to the overall model.

## 3.3 Model Training and Evaluation

The used methodology involves training a ResNet-34 which is formed out of residual connections, hence addressing the vanishing gradient issue associated with CNN architectures with deep learning architectures. CrossEntropyLoss was chosen because of its fitness for the situations where the input is in softmax space, and it is a binary classification problem. The use of the Adam optimizer was due to its ability to learn the rate at which the parameters should be updated. The results included precision, recall, F1-score, and confusion matrices as ways of assessing the test performance of the best model chosen.

# 4 Design Specification

## 4.1 System Architecture

The proposed framework for performing scream detection that integrates conventional machine learning approaches with deep learning techniques, thereby solving the challenges involved. Normalization, augmentation, and feature extraction are the initial stages of preprocessing audio data. These are used to capture important acoustic properties, including: Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast and zero-crossing rates. Spectrograms and mel-spectrograms are utilized in deep learning models to convert audio signals into visual representation, to improve classification accuracy in high-intensity noisy environments. ResNet-34, SVM, and MLP are among the models included in the framework, which were chosen for their ability to tackle specific scream detection tasks. Using feature engineering as a combination tool with multiple approaches to modeling, the framework is designed to test how different techniques work under various circumstances.

## 4.2 Model Architecture

In the proposed framework, several models were selected based on their applicability and effectiveness for scream detection task. The architecture of each model was then considered to meet the purpose of this study and handle the requirements of the dataset.

Support Vector Machine (SVM) was considered suitable for the binary classification task and is usually applied to large sets of numerical features such as MFCCs, chroma values, spectral contrast, and zero-crossing rates. SVM performs a binary classification with the support vectors of the highest margin for discriminating features between scream and non-scream classes. A linear kernel was used, which meant that this data set could be handled computationally in a very simple and efficient manner. SVM was preferred over traditional machine learning models due to its good scalability when data contains noise and when it is mapped into high feature space. The computational proficiency on small sets of data and its probabilistic results makes it a suitable reference for evaluating scream detection benchmark against other machine learning strategies against deep learning.

Multilayer Perceptron (MLP) is a feedforward type of artificial neural network which performs well in cases of non-linear features like scream detection therefore being ideal for this kind of classification. In doing so, MLP should be able to model complex interactions between features because of the interconnection of neurons in its network. The model takes the input features through several hidden layers, which allows a network between, it means that the network can then identify hierarchical patterns from the data. Because of these characteristics, MLP was chosen for this framework, especially if it has been trained carefully on relevant audio features, Mel-frequency cepstral coefficients, chroma features, or spectral contrast features. These features give a good set of characteristics of scream and non-scream classes and the structure of MLP is effective in extracting valuable information from such numerical data. Furthermore, MLP also has a scalable procedure to compute its results as opposed to the deep convolutional neural networks and is also capable of outperforming the approaches such as logistic regression in different model dimensions.

Out of various deep learning techniques of categorizing data, ResNet-34 is chosen for its effectiveness in processing image-related information including spectrograms of audio signals. Its residual connections help to solve another problem, namely vanishing gradient, allowing for training deeper structures. ResNet-34 was pre-trained when building on ImageNet and thus uses transfer learning to reduce training time while still being accurate for scream detection in noisy conditions. Resized and normalized spectrograms were used as data inputs to the model since these features enabled the model to learn higher order patterns that could be obscured by numerical vectors. ResNet-34 was chosen over other models such as VGGNet and Inception because of the model's depth as well as computational efficiency with sound as its input, while also showing high generalization capacity across different acoustic conditions. The proposed model is suitable for this study because of its ability to handle spectrogram data and a variety of scream detection issues.

This framework leverages the capabilities of both conventional machine learning and deep learning techniques to provide an in-depth analysis of their performance across acoustic environments.

## 4.3 Evaluation Framework

As for the evaluation of the proposed scream detection framework, much attention has been paid to the fact and the methodology of each model to avoid mistakes. The dataset is split into training, validation, and test subsets in an 80:10:10 ratio to control the load of training and to avoid influencing decision making. This division provides enough data for the formation of a solid base model whilst maintaining reasonable subsets for validation and testing.

The assessment here uses Precision, Recall, F1-Score and Accuracy as the assessment measures since they are sensitive to imbalanced data sets and real-world cases. The noise robustness examination is conducted by applying different types of acoustic distortions to the test dataset, allowing for understanding the performance of the model in adverse conditions. Realistic noise types and their levels are presented in this step with the purpose of testing the applicability of the method. Misclassification is analyzed and compared using confusion matrices, and the relationship between sensitivity and specificity is depicted by ROC and Precision-Recall curves. These visual aids enhance a richer interpretation of model performance. The assessment is based on the comparison between SVM, MLP, and ResNet-34 networks in order to reveal that traditional machine learning algorithms have their advantages and disadvantages when compared to deep learning strategies. Real world variability is incorporated with the augmented data and the framework is tested to check the performance of the framework even with variability added to the data. In addition to providing confirmation for the effectiveness of the discussed models, this evaluation framework reveals potential directions for further enhancement, including increasing model generality, handling specific exceptional cases, and enhancing computational efficiency. Thus, these findings make a theoretical and methodological contribution to the study of scream detection and its potential use.

# 5 Implementation

In this section, detailed information about the tools, models and techniques with specialties of the roles they play in this framework are given are explained.

## 5.1 Tools and Libraries

The main reason for selecting Python as the primary programming language is that the language provides enormous libraries on machine learning and data analysis. Key tools used in this project include:

- **Librosa & Soundfile:** For audio analysis as well as the feature extraction process. It introduces methods for computing features like MFCCs, chroma features and spectrogram.
- **Matplotlib and Seaborn:** Used for charting some material, such as spectrograms or performance charts such as the ROC and precision-recall curves.
- **Torch, Torchvision and Pytorch:** Some of the important libraries for deep learning model building and training especially for the ResNet-34 architecture is given below.
- **Scikit-learn:** Used in traditional machine learning algorithms like SVM and in performance metrics computations.
- **Pandas and Numpy:** The data structure used in data processing and also used for performing numerical operations.
- **TQDM:** A utility library used in repetitive processes when it is necessary to show a progress bar in data preprocessing or model training.

Combined, these tools provided the necessary foundation to enable smooth workflow integration of both conventional as well as DL approaches.

## 5.2 Exploratory Data Analysis (EDA)

EDA was vital in determining the attributes of the dataset and any inherent challenges that might affect the analysis.

### 5.2.1 Class Distribution

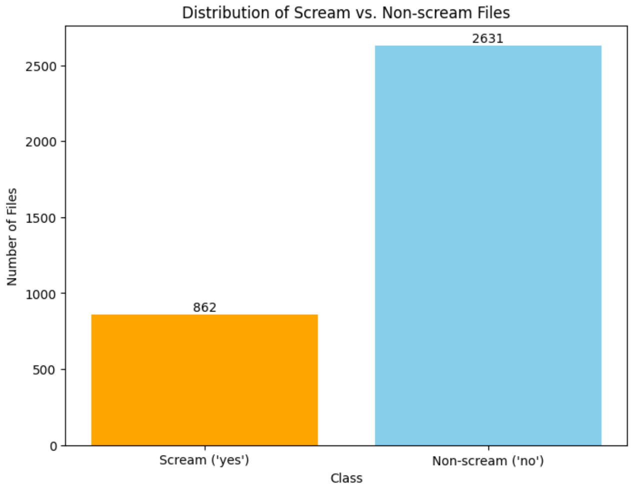

**Figure 2** – Distribution of unbalanced data

While determining the dataset there was a problem of imbalance with 862 samples in scream and 2631 in the non- scream class. It means that the classifier might be trained in such way that it will favour the majority class and ignore the minority class. To counter this, resampling techniques were adopted with an aim of having an equal number of instances for scream and non-scream classes.

It was important to balance the data as it helped in minimizing the biasing of model to the majority class during training and thus making the models more flexible to unseen data and making efficient identification of the scream signals. This is especially important for real-life situations when screams, being minority class, denote as events of high relevance, such as accidents or thefts.

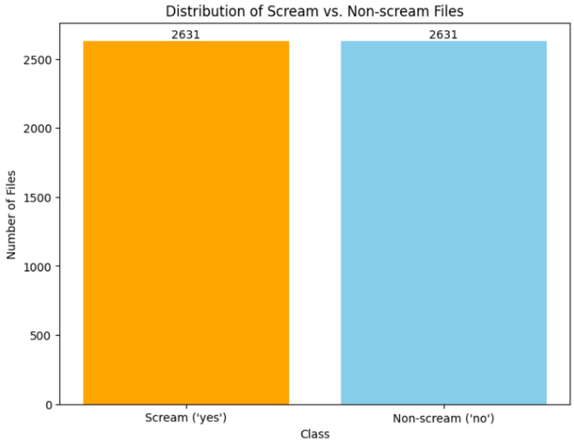

**Figure 3** – Distribution of balanced data

## 5.2.2 Raw Audio Waveforms

Raw waveforms for 'scream' and 'non-scream' classes were plotted. Scream waveforms can be described as having impulses or amplitude spikes, which means they consist of sudden and powerful signals. The non-scream waveforms, however, may have different but relatively smooth and regular shapes without so many fluctuations. Within classes, energy distribution patterns were compared and analyzed to find that scream samples have one or more intervals of high energy density while non-scream samples have a similar or lower energy density spread out over more intervals.

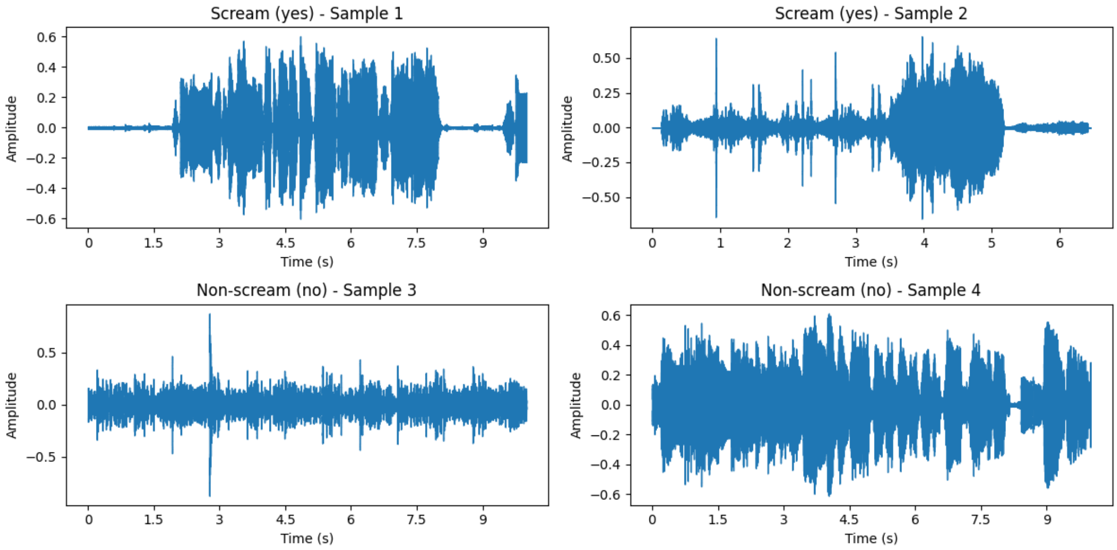

**Figure 4** – Raw Audio Waveforms

### 5.2.3 Spectrograms using Short-Time Fourier Transform (STFT)

Spectrogram can be described as an image that represents the frequency of a signal as the intensity varies with time. The STFT was used as it converts the audio signals into the time-frequency domain, thus extrapolating the audio data layer. The STFT transforms the audio signal into a set of frames whereby the frames overlap and compute Fourier transform. This captures the frequency content over time hence useful for analyzing signals that are not stationary as the scream. Mel-spectrograms obtained from spectrograms which are calculated by mapping the frequencies to the Mel scale were computed for each sample. Mel scales frequency identification with the human ear uptake, stressing important octave and fractional Octave bands.

Spectrograms for ‘scream’ samples contained dense, compact energy blobs within high frequencies, while ‘non-scream’ samples feature more or less equal energy distribution across the frequency spectrum. These patterns showed some unique features as to the acoustic properties of each class.

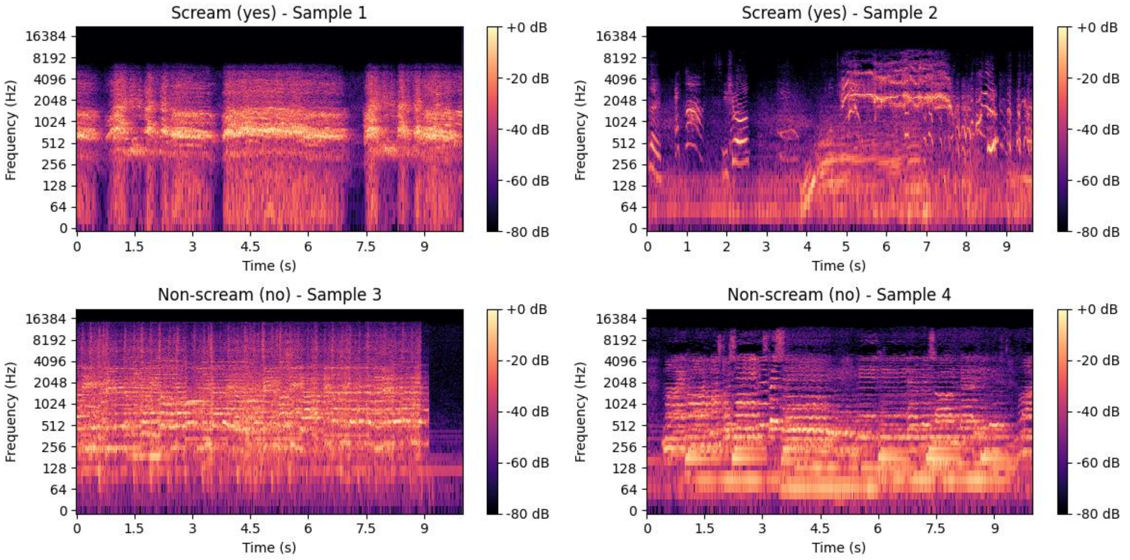

**Figure 5** – Spectrograms of audio files using STFT

### 5.2.4 Feature Analysis

The scatterplot on the figure below is proved to represent the concept of feature clustering. Yellow circles refer to one class (“scream”), whereas purple points refer to the other class (“non-scream”). The clusters suggest that particular characteristics can successfully partition the two classes. Part of the space separates out regions that are closely associated with the two classes suggesting that some features work well in discriminating between scream and non-scream. Such qualities could be comprised of high energy flashes or other frequencies within scream signals. The existence of the overlap indicates where the features are less selective probably because of the noise or similarities in the acoustic attributes of some scream and non-scream samples. The overlap at the different points in the feature space shows where the feature representation fails to distinguish between classes of objects.

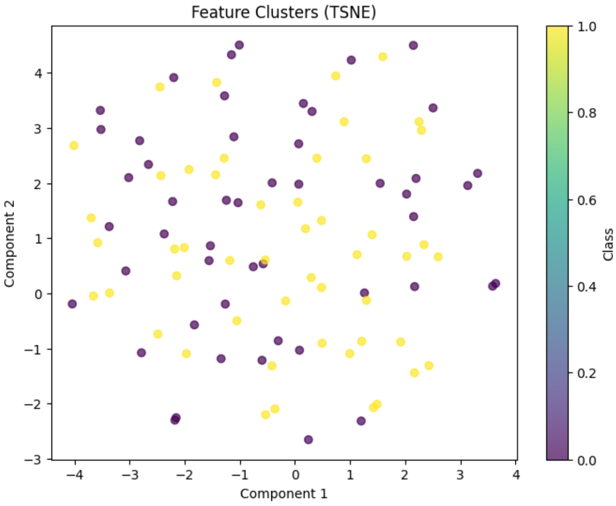

**Figure 6** – Feature Distribution

## 5.3 Data Preprocessing

As highlighted in the methodology section 3.2, the collected dataset went through processes of resampling, data augmentation and feature extraction. These steps helped to maintain a working balance between every class and produce reliable features for the models. For these transformations to be applied, preprocessing pipelines were constructed to automatically apply and process them in a structured way across experiments. Furthermore, the extracted spectrograms were normalized and resized to the input size ResNet expected which is 224x224 pixels.

## 5.4 Model Implementation

### 5.4.1 Support Vector Machines (SVM)

The Support Vector Machine (SVM) model was trained using the extracted features of the Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast and zero crossing rates features. The features were normalized to make all the features in the same range as SVMs are sensitive to the scale of input data. A linear kernel was used due to its ease and efficiency in performing a binary classification task particularly to distinguish the two classes of scream and non-scream. The model optimized the hyperplane that offers the greatest margin between the two.

When training the model, the probability parameter was set “True” to have a probabilistic output on the evaluation which is necessary to perform evaluation and interpretation of the result. The validation dataset was used in order to achieve the results with no sign of overfitting the model. Compared to the deep learning models that require multiple iterations during the training process, the training of the SVM was fast, and more significantly, the validation loss soon began to level off at a fairly low value, which means that the current model has reached the accurate fireworks clustering.

### 5.4.2 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) was trained with the same feature set as in case of classifier SVM. The architecture had two hidden layers with 100 and 50 neurons in each layer. Indeed, each hidden layer implemented the Rectified Linear Unit (ReLU) as the activation function to introduce non-linearity so that it can derive peculiarities from the data.

The training of the model used the Adam optimizer which adapts the learning rate for every parameter with respect to the two moments of gradient. This optimizer was selected given its computational efficiency and the related property of fast convergence. The batch size was used in a way that it acquired a representative training sample for each training step.

To overcome the risk of overfitting some dropout layers were added – this removes randomly selected neurons during the training session. These dropout rates were chosen to ensure that the dropout provided a form of regularization but did not damage the networks capacity. During the training phase, more than one epoch were used and validation loss was also included with the intent of stopping the training, if the model predict a high loss value.

### 5.4.3 ResNet-34

Training input images for the ResNet-34 model included spectrograms and mel-spectrograms. These spectrogram images have been resized to 224×224 pixels, and they have been normalized such that the input format matches the ResNet-34 model architecture. Transfer learning was used while the program was trained with a pre-trained ResNet-34 model with weights started from the ImageNet dataset. This approach enabled the model to exploit pre-learned features, which in turn removed the need for large, annotated datasets and possess a shorter time of training. The final fully connected layer of the network was removed and substituted by a custom classifier layer which includes two output neurons corresponding to two classes with softmax activation function. The optimization of probabilistic prediction for binary classification was done using cross entropy loss function. Mini batches where used during training and when it was time to update the weights of a model, the Adam optimizer was used. Through each epoch, the training data set was forward pass through the network and the gradients were propagated back in order to update the weights. The original learning rate was set equal to 0.001 and decreased by half if validation loss did not increase. During training, data augmentation including random flips and rotations were used to increase model’s ability to generalize with unseen data.

To assess the overfitting the performance of the model was checked at the end of every epoch. If validation loss started increasing while the training loss continued to reduce, then, early stopping was done. The ResNet-34 model was chosen since it had a very deep structure; however, this choice led to high computational complexity; but, when using spectrogram images, the model successfully extracts hierarchical features.

# 6 Evaluation

The evaluation part of this study focuses on assessing the performance of the three implemented models.

## 6.1 Support Vector Machine (SVM)

The model was trained and validated on the proposed dataset with 80% accuracy in both.. This shows that there was an optimum learning as the model was able to reduce the learning capability into a summary but without compromising with the amount of detail it was able to capture from the data.

The test accuracy of the SVM was also 80% confirming the consistency in the training, validation, and test phases of the execution of the model. Additional metrics highlight the model's balanced performance:

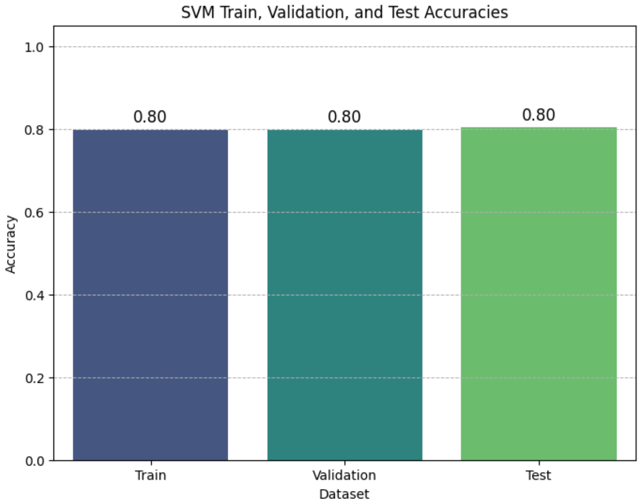

**Figure 7** – Comparison of accuracies for SVM

An 80% precision means that when the model is identifying scream samples amid all the positive classifications, it is right 80% of the time. The recall of 80% thus demonstrates its high capacity for scream sample identification out of the dataset. Likewise, the F1 score stands at 80 %, which shows that precision and recall ratios have been accurately attained and balance the reliability of the automatic classification. In addition, the 81% specificity indicates the ability of the model in identifying non-scream samples in the correct classification. Finally, a mean accuracy of 80% also corroborates the model's reliability and general performance in discriminating scream from other non-scream classes of sounds.

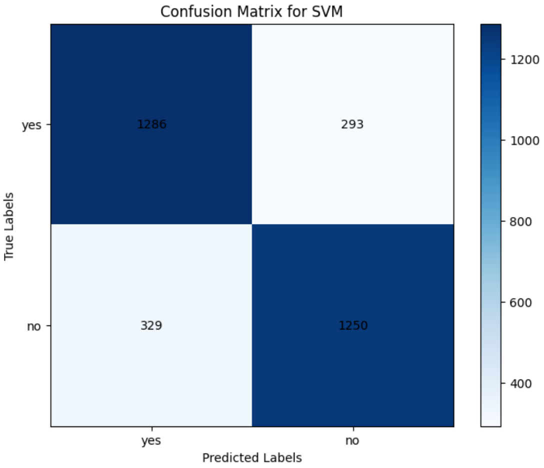

**Figure 8** – Confusion Matrix for SVM

Above image reveals that the SVM model arrived at a nearly equal classification of 1286 true positives and 1250 true negative cases. The following tables show the misclassification list comprising of 293 false positives and 329 false negatives, this results from difficulty of differentiating between a scream or not, especially due to linear kernel used.

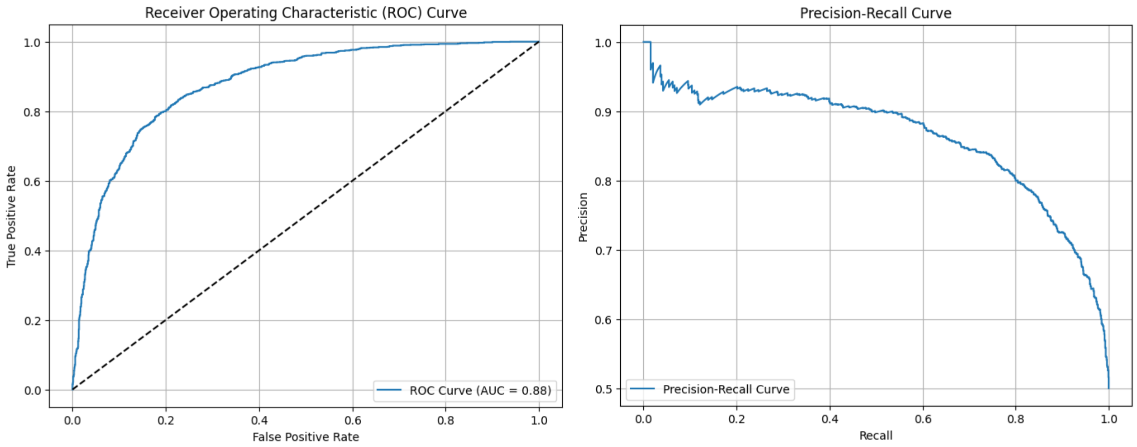

**Figure 9** – ROC and Precision-Recall Curve for SVM

The overall performance of the proposed system was quantified based on the Area Under the Curve (AUC) score received from Receiver Operating Characteristic (ROC) which was 0.88 when the system was tested for discrimination of scream and non-scream classes. The flattened curve on the ROC means that the model has constant true positive rate as it varies the thresholds. The Precision-Recall curve perfectly supported the findings of the SVM with high precision for moderate recall rates. When recall got to unreasonable levels, precision was seen to degrade, a signal of the compromise where all scream samples could be recalled at the cost of accuracy.

## 6.2 Multilayer Perceptron (MLP)

During the training and validation processes the MLP model performed very well. It managed to obtain training accuracy of 1.00 which trained the model on training data to maximum effectiveness. The value of 97% at validation accuracy proves that the model fitted the validation data without the overfitting problem. The high validation performance showcased in the algorithm proves the effectiveness of the MLP model in managing the complexity of the features and the potential of overfitting which is a normal problem for neural networks.

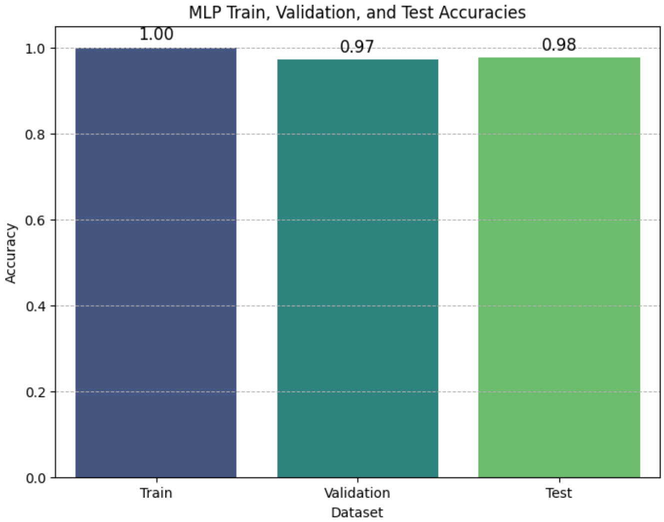

**Figure 10 –** Comparative analysis of accuracies for MLP

On the test dataset, while MLP model achieved good accuracy of 98%, it was satisfactory for a model to generalize well on the new data. The choice of the evaluation metrics used in

MLP model supports its reliability in scream detection. An accuracy level of 96% means it is effective in capturing scream signals without many unnecessary alerts. The recall, which is 99%, shows how the model can identify nearly all the scream samples, even at the worst of times. The F1 score of 98% also signifies that the model is an excellent one in terms of the degree to which it recalls the necessary features; at the same time, it embodies the measure of precision. Furthermore, with an accuracy of 96%, it also shows good performance in detecting the non-scream samples and the balanced accuracy of 98% reassures that the proposed model performs well in both scream and non-scream classes.

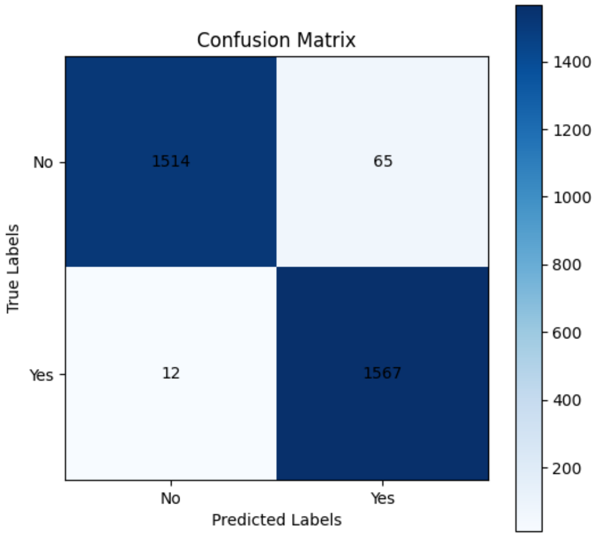

**Figure 11** – Confusion Matrix for RESNET

The above image shows good performance, and this has reduced the number of misclassified instances. There are more true positives (1567) and true negatives (1514) than false positives (65) and false negatives (12). This shows that MLP with this set of features is able to generalize well.

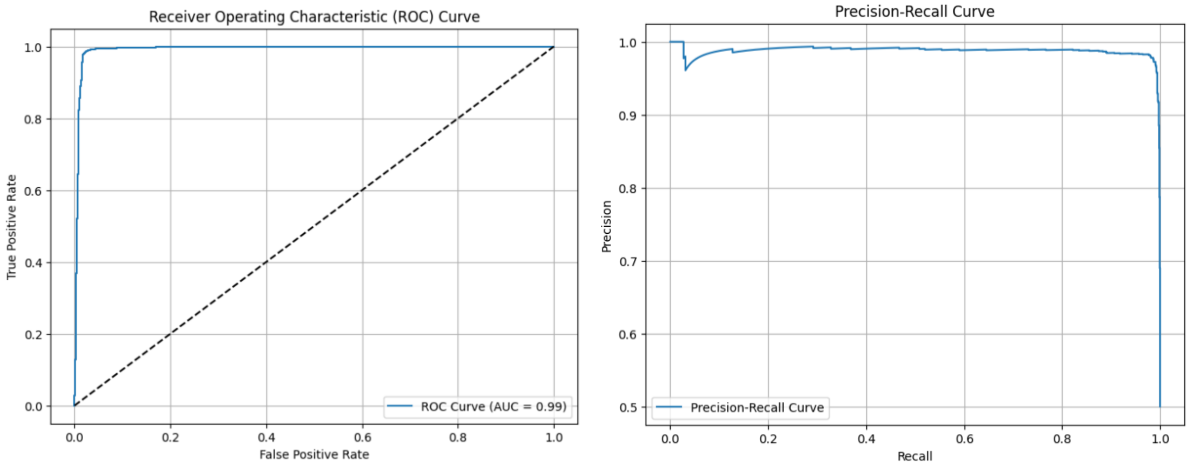

**Figure 12** – ROC and Precision-Recall Curve for RESNET

Analyzing the ROC curve of the MLP model, we can state that the model has very good discriminative capabilities. When speaking of the classification of the ‘scream’ and ‘non-scream’ classes, the utmost classification metric of AUC was registered as 0.99, meaning that almost all scream samples achieved higher classification scores than the non-scream samples at all thresholds. The operation of the Precision-Recall curve highlights the robustness of the

MLP across the imbalanced data set performance. The curve is quite smooth with precision almost equal to 1 even at lower recall confirming low false positive rates at high recall.

MLP model was found to be efficient in terms of classification for scream detection. The generalization proficiency of its training, validation, and test datasets is confirmed by high accuracy, precision, and recall rates.

## 6.3 ResNet-34

The model achieved a training accuracy of 98%, signifying that it had also learnt the pattern from the training data set perfectly. Through achieving high validation accuracy equalled to 94%, it can be inferred that the model did not over-learn on the training set such that the unseen data in the training phase were well predicted.

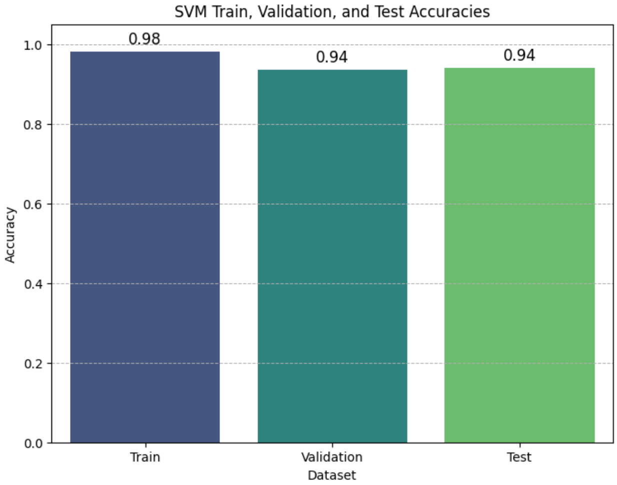

**Figure 13** – Comparative analysis of accuracies

On the test dataset, ResNet-34 estimated an accuracy of 94%. While the recall and the F1 score of 94%, confirming the model had a high capability to accurately predict the scream from negative samples without many false positives. Non-scream recognition performance was 95% in terms of specificity, and 93% in the balanced accuracy, explaining the ability to work with imbalanced datasets.

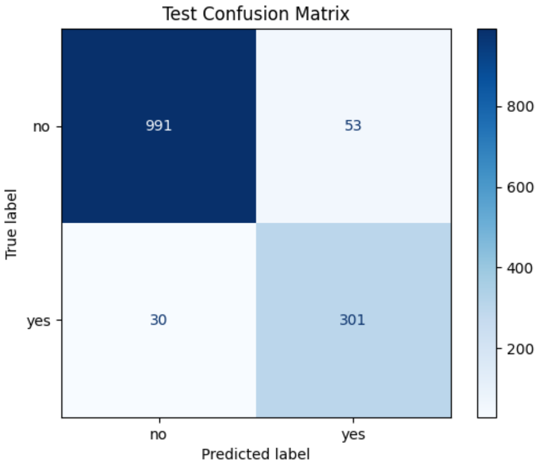

**Figure 14** – Cofusion Matrix for MLP

The results reinforced its fit and hold on accuracy standards and constraints, specifically for errors in class predictions on positive and negative classes alike.

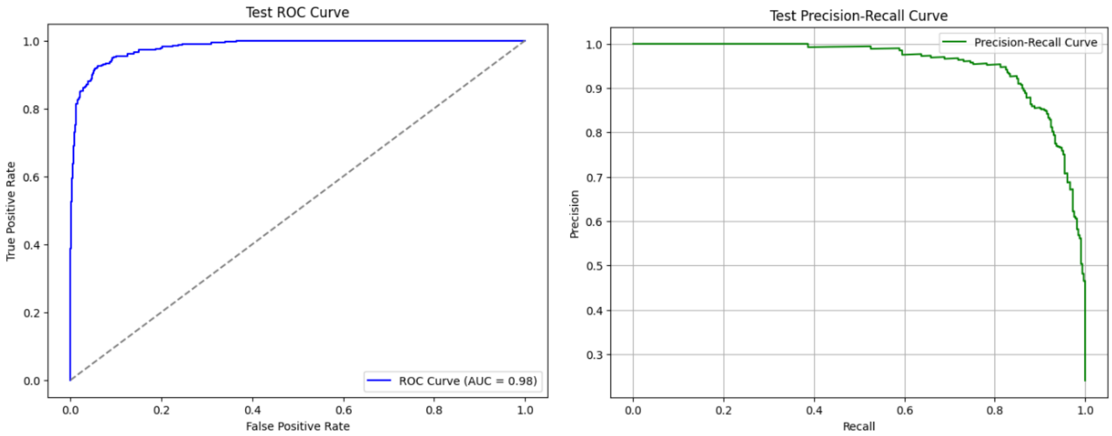

**Figure 15** – ROC and Precision-Recall Curve

The above image shows an AUC of 0.98. This high AUC value stipulates that the model performs equally well for various threshold levels, guaranteeing optimum efficiency for various FPR levels. From the Precision-Recall curve, it was seen that ResNet-34 had high precision at smaller recall values, which means that it could easily capture almost all the positive instances, which in this case was the scream.

Gathered, all these metrics and visualizations do confirm ResNet-34's capabilities in applying spectrogram-based features for audio classification tasks. Its good generalization performance and, therefore, high precision-recall rates show that it is well suited for scream detection in complex acoustic conditions.

# 7 Results and Discussion

The results show how effective the proposed scream detection framework is. All models were able to detect human screams in noisy and weak signal datasets, showing suitability to the task. The SVM had an 80% accuracy in training, validation, and test data sets using MFCCs, chroma, spectral contrast, and zero-crossing rates for consistent classification. Its performance under different conditions was confirmed by the ROC curve (AUC = 0.88).

The MLP got a test accuracy of 98% using features MFCCs and chroma values which are non-linear data for modelling. To address overfitting, regularization techniques such as dropout were used, and it was sampled with 97% validation accuracy. On the ResNet-34, the test accuracy was achieved to be about 94%, and higher-order acoustic features were learned through spectrogram-based representations. Training with ImageNet weights also brought advantages in terms of time as well as accuracy, which in return was illustrated by the ROC curve of 0.98. Methods of data augmentation like pitch shift and adding noises also improved its performance in real world environments. Data balancing and feature extraction were important for enhancing the model's performance despite the presence of minor bias in borderline cases.

# 8 Conclusion and Future Work

This research focused on scream detection using a strong framework that involves both machine learning and deep learning. The use of data augmentation successfully was reflected in the noisy data and the imbalanced data set performance of the proposed methodology pointed out further enhancement of feature extraction. The models properly learned the important features of scream signals and can be applied to real-life applications such as emergency, health care systems and smart cities. The research accomplished its goals effectively of achieving accurate and highly robust scream detection. The integration of formal and informal approaches allowed the framework to be effective in different scenarios. Nevertheless, as the model also inherits some drawbacks that include the database biases and large computational costs for performing the computations, there are certain aspects that need further enhancement.

Future work should address the testing of the framework in natural and unconstrained settings with the purpose of checking for its transferability across different scenarios. A number of techniques, like pruning the architecture, or performing quantization, are effective in reducing the computations and thereby can improve the applicability of deep learning models in embedded systems and low energy devices. To solve cross-cultural and acoustic variability, large scale and diverse datasets will be developed to improve the model performance. Furthermore, it is also possible to include techniques, like federated learning that would uphold privacy during the deployment in surveillance and monitoring areas. Such improvements shall render the potential for scream detection more efficient along with its uses in safety-critical applications broader.

# References

Gao, L., Zhang, X., Yang, T., Wang, B. & Li, J., 2023, ‘The Application of ResNet-34 Model Integrating Transfer Learning in the Recognition and Classification of Overseas Chinese Frescoes’, Electronics, 12(17), 3677.

Gao, M., Qi, D., Mu, H. & Chen, J., 2021, ‘A Transfer Residual Neural Network Based on ResNet-34 for Detection of Wood Knot Defects’, Forests, 12(2), 212.

Handa, D. & Vig, R., 2020, Distress Screaming vs Joyful Screaming: An Experimental Analysis on Both the High Pitch Acoustic Signals to Trace Differences and Similarities, 2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), 190–193, IEEE.

Kalbag, V. & Lerch, A., 2022, ‘Scream Detection in Heavy Metal Music’.

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

Mac, A., Xu, T., Wu, J.K.Y., Belousova, N., Kitazawa, H., Vozoris, N., Rozenberg, D., Ryan, C.M., Valaee, S. & Chow, C.-W., 2022, ‘Deep learning using multilayer perception improves the diagnostic acumen of spirometry: a single-centre Canadian study’, BMJ Open Respiratory Research, 9(1), e001396.

Nandwana, M.K., Ziaei, A. & Hansen, J.H.L., 2015, Robust unsupervised detection of human screams in noisy acoustic environments, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 161–165, IEEE.

Nazir, S., Awais, M., Malik, S. & Nazir, F., 2018, ‘A Review on Scream Classification for Situation Understanding’, International Journal of Advanced Computer Science and Applications, 9(8).

O’Donovan, R., Sezgin, E., Bambach, S., Butter, E. & Lin, S., 2020, ‘Detecting Screams From Home Audio Recordings to Identify Tantrums: Exploratory Study Using Transfer Machine Learning’, JMIR Formative Research, 4(6), e18279.

Ojha, A. & Adarsh, 2024, ‘Human Scream Detection and Analysis for Controlling Crime Rate using Machine Learning and Deep Learning’, International Journal for Research in Applied Science and Engineering Technology, 12(3), 928–932.

Pandey, H., Lakhi, J.R., Goel, V. & Bansal, D., 2023, ‘A Review on Scream Detection and Classification’, INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 07(12), 1–11.

Rashedi, K.A., Ismail, M.T., Wadi, S. Al, Serroukh, A., Alshammari, T.S. & Jaber, J.J., 2024, ‘Multi-Layer Perceptron-Based Classification with Application to Outlier Detection in Saudi Arabia Stock Returns’, Journal of Risk and Financial Management, 17(2), 69.

Saeed, F.S., Bashit, A. Al, Viswanathan, V. & Valles, D., 2021, ‘An Initial Machine Learning-Based Victim’s Scream Detection Analysis for Burning Sites’, Applied Sciences, 11(18), 8425.

Sharma, A. & Kaul, S., 2016, ‘Two-Stage Supervised Learning-Based Method to Detect Screams and Cries in Urban Environments’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(2), 290–299.

Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shiang Kok & Jit Biswas, 2010, Scream detection for home applications, 2010 5th IEEE Conference on Industrial Electronics and Applications, 2115–2120, IEEE.

Zaheer, Md.Z., Kim, J.Y., Kim, H.-G. & Na, S.Y., 2015, A Preliminary Study on Deep-Learning Based Screaming Sound *Detection, 2015 5th International Conference on IT Convergence and Security (ICITCS)*, 1–4, IEEE.