WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Classification and Categorization of Digital Violence Content Using Acoustic Features and Supervised Learning

Yakaiah P[1]*, Manjunathachari K[2], SVS Prasad[3]

[1]Research Scholar, Department of ECE, Rayalaseema University, Kurnool, India
[2] Department of ECE, GITAM University, Sanga Reddy, Telangana, India
[3] Department of ECE, MLR Institute of Technology, Hyderabad, India
*** E-mail: potharaju.yakaiah@gmail.com**

ABSTRACT

Due to a substantial improvement in the technology, the most advanced media internet, broad band internet services, video streaming has gained a lot of popularity. Along with this popularity, the dissemination of objectionable content such as Violence has become uncontrollable. This content has a severe effect on the sensitive social communities like children because it has already proven that the media violence has too many negative effects on the attitude, behavior, and emotional state of children. Hence there is a need to develop an automatic violence detection system through which the multimedia content is filtered for violence content detection and removal. In an automatic violence detection system, the most important thing is to define the violence and in real time there are so many ways like audio, video and the combination of both. In the case of audio, the audio characteristics like screaming, sound, gunshots, music, speech different emotional states, Environmental sounds with abrupt changes in signal energy, can define the violence. this research work developed a knowledge insisted ontology based framework for violence detection from audio-visual datasets. This system is accomplished in two phases, one is based on ontological audio data and another is based on ontological visual data. In the first phase, the violence is defined through audio signals.

1.Introduction

Technological advancements have enabled the widespread adoption of next-generation media and entertainment services such as IPTV, video on demand, broadband internet, and video streaming platforms. However, this proliferation has also brought concerns regarding the prevalence of objectionable content, including pornography and violence. Multimedia content is becoming increasingly accessible to a broader audience, often without centralized regulation. Therefore, it is imperative to safeguard vulnerable populations, particularly children, from exposure to harmful media. Research has demonstrated the significant negative impact of media violence on children's emotional well-being, behavior, and attitudes.

Manual annotation of multimedia content for violence detection is challenging due to the sheer volume of data. Existing monitoring bodies, such as the Central Board of Film Certification, face limitations in effectively screening such vast amounts of content.

Violence detection is gaining importance not only in practical applications but also in scientific research, as it is a fundamental aspect of generic action recognition. The primary purpose of installing video surveillance systems in jails, schools, and mental health facilities is to notify the authorities of potentially harmful circumstances. There is a great need for an automated alert system because the human operators needed to achieve this are overworked due to the large number of video feeds and the slowness of manual procedures. Concurrently, there's a growing need for automated processing systems that handle the large- scale videos that websites host. Violence detection has become more important in today's research because of its critical role in providing security alerts in video surveillance systems, both at the scientific and application levels. There are certain specific differences between the general awareness of human actions and the detection of violence. For all of these reasons, there has been a steady

increase in the interest in violence detection studies. Additional psychological study on media violence has confirmed its detrimental impacts on children's emotional behavior, attitude, and other elements. By highlighting these consequences, new violence detection technologies that exclude this kind of content from the data are being developed.

The primary challenge lies in establishing a precise definition of violence based on audio-visual data. Subjectivity in characterizing violence poses difficulties in defining explicit criteria for its identification. Violence is typically defined as any action or scenario that poses a potential threat of physical or psychological harm to one or more individuals. Video clips containing such occurrences can be classified as violent conten[3].Previous research has identified well-established auditory and visual cinematic grammar within the film industry. These aspects have been leveraged in prior attempts to detect violence in video clips[4]-[5].

A portion of the research on violence detection has been completed, taking into account weapons, explosions, kicks, and first fighting as sources of violence. Owing to the current understanding of crime in the human mind, people are attempting to incite violence in a variety of ways. However, the most advanced automatic violence detection systems are unable to address these issues, necessitating the development of a novel automatic violent detection system. Furthermore, there is no one set definition for violence; instead, one must look at a variety of factors, including images, sounds, screams, and gunfire, when detecting violence. The detection system must be trained exclusively in that way and become more effective in the learning process in order to identify violence in such circumstances.

To address the challenges in violence detection, this research aims to develop a robust and efficient system. The specific objectives are:

a). Ontological Definition of Violence: Provide a comprehensive and inclusive definition of violence that encompasses the diverse perspectives of users.

b). Audio-Visual Representation of Violence: Utilize both audio and video modalities to capture a broader spectrum of violent scenes. Develop novel feature extraction techniques for audio signals to enhance violence discrimination. Employ histogram-based representation for efficient video representation.

c). Optimized Computational Approach: Implement a lightweight search mechanism to reduce computational overhead. Utilize Support Vector Machine (SVM) for audio-based classification and a spatio-temporal localization mechanism for video-based classification.

d). Performance Evaluation: Create a comprehensive database by extracting audio and video clips from Indian Hindi movies. Select movies with high violence content and diverse representations of violence for database construction.

2.RELATED WORKS

This chapter discusses the literature survey details regarding violence detection. Since the audio and video are the two important aspects in the movies through which the scene can be justified as violent, this survey focused over these two aspects only. Based on these aspects the literature survey is classified as audio signals based approaches and visual cues based approaches.

2.1 Audio based Approaches

Various approaches are developed to perform violence detection through auditory models. Geiger et al. propose energy entropy based acoustic scene classification in [6]. Energy entropy can be used to identify abrupt changes in the audio signal, which generally corresponds to the violence content. Some non-violent sounds, such as thunder, are also detected as violent, even though the energy entropy features are more effective at detecting violent content. In addition, a further technique for segmenting and categorizing audio signals based on signal entropy is developed in [7]. Next, in order to distinguish between the vocal and non-vocal portions of the songs, the authors of [8] created a brand-new technique based on machine learning and signal processing techniques. Artificial Neural Networks (ANN) were used to extract the features of the vocal and non-vocal segments.

Giannakopoulos et al. [9] have developed a new method for detecting violence in audio signals by extracting significant time domain and frequency domain features from the signal. Additionally, various statistical operations are carried out on the acquired feature set, and the SVM algorithm is used to train them. The segment is classified as violent or not by this SVM algorithm. Building on the approach presented in [9], a novel approach is put forth in [10] to categorize audio clips taken from motion pictures with the intention of identifying violent content in order to safeguard socially vulnerable populations, such as children. In the end, a total of twelve features were evaluated using the feature extraction method suggested in [10]. Using a one-versus-all approach, the Bayesian Network (BN) classifies the audio segment into Completely

Furthermore, Thaweesak [13] suggests a novel approach to determine a speaker's depression from their speech. This method involved computationally extracting the full-band and further sub-band entropies of eight evenly spaced frequency bands of 625 Hz from the female voiced segments, which were then utilized to create the parameter models for between-group classifications. In order to perform classification, a machine learning classifier is further implemented over the extracted feature set. M. Balede et al. propose in [14] a Gaussian Mixture Model (GMM) based audio signal detection and classification mechanism. In this case, a dictionary is created after modeling the sound spectrum using a mixture model. The best match is used as a result of further classification carried out through likelihood estimation.

A novel approach for detecting environmental sounds based on time-frequency audio features is put forth by M. Loughlin et al. [15] because there are various types of environmental sounds. Various features are suggested to distinguish distinct sounds from the audio signals, taking into account the semantics of the various audio signals. These features comprise the front end features of the spectrogram as well as the auditory image. Based on the signatures in the time domain, environmental sounds such as the sounds of rain, birds, and insects have a broad, flat spectrum that is similar to the noise spectrum [17]. S. Sameh and Z. Lachiri [16] are working on another study to use the analysis of spatiotemporal features to examine ambient sounds. Their primary focus was on differentiating polyphonic music from ambient sounds found in urban settings. Higher classification accuracy for ambient sounds is achieved by supplementing the MFCC features with a feature based on Log-Gabor filters. Nevertheless, because of their usual success, the use of Log-Gabor filters adds to the system's computational load.

Based on group learning, S. Saman et al. [18] classified the violent scene using auditory models. This method used the Random Forest algorithm to classify the audio signals by extracting their Zero Crossing Rate (ZCR) as a feature. Unfortunately, there isn't much discrimination between various audio classes that a single ZCR can offer. Furthermore, because of the hierarchical strategy, it becomes too difficult for the Ontological based classification. In

an attempt to identify violent scenes, Vu Lam et al. [19] took into account the low level multiple features of audio signals. In order to detect violence in real-time environments, Marta et al. [20] developed an additional method that takes into account a variety of features, including Mel- Frequency Cepstral Coefficients (MFCCs), Pitch [21], Harmonic Noise rate (HNR), Short time energy [22], ZCR, etc. Three additional classifiers are completed for classification: a Neural Network based detector with five hidden neurons, a Least Squares Linear Detector (LSLD), and a simplified version of the Least Squares Quadratic Detector (LSQD).

2.1.1 Emotion Based

Furthermore, a person's feelings when speaking can also be used to define violence. Anger-related emotions are characterized as violent, while other emotions are characterized as non-violent [23], [24]. Sadness, happiness, joy, neutrality, etc., are all regarded as non-violent emotions, whereas anger is the only one that is classified as violent. Defining emotions universally remains challenging due to their highly subjective nature. This subjectivity has led to the development of numerous models in psychological literature[33].

One model classifies emotions into discrete categories known as basic emotions. Paul Ekman proposed seven basic emotions: neutral, disgust, boredom, anxiety, anger, sadness, and happiness. These emotions are considered fundamental, with more complex emotions arising through their combinations[34].

Many methods have been developed in the past to identify emotion from audio signals [25]. S. Demircan and H. Kahramanl [26] created an emotion recognition system with unsupervised learning using MFCCs of speech signals. The K-Nearest Neighbor algorithm was utilized in this method to categorize the emotions. The unsupervised learning algorithm, however, adds another layer of complexity to the recognition system. More supervised learning- based methods, such as neural networks [28] and recurrent neural networks [29], are being suggested in an effort to more precisely extract emotion information from speech signals.

An improved model of brain emotional learning (BEL) that combines the Multilayer Perceptron (MLP) and Adaptive Neuro-Fuzzy Inference System (ANFIS) for speech emotion recognition was presented by Sara et al. [27]. The primary drawback of the MLP algorithm is its inability to ensure that the minimum it stops at during training is the global minima due to the way it is trained. The user must determine the number of Hidden Neurons, which is another restriction on the MLP algorithm. If this number is set too high, the MLP model may over fit, and if it is set too low, the MLP model may under fit. Diana et al. have developed a sparse coding framework [30] to identify emotions in speech signals. A hierarchical sparse coding (HSC) scheme is proposed and features from audio are automatically represented using the sparse coding framework.

An emotion classification system was proposed and evaluated by F. Chenchah and Z. Lachiri [31], with an emphasis on the distinctions between acted and spontaneous emotional speech. They looked at wavelet packet energy and entropy features applied to the Mel, Bark, and ERB scale using the Hidden Markov Model (HMM) as a classification system in order to complete this work.

Tian Han et al. conducted a simulation experiment [32] to identify violence in school environments. They employed the 'Consecutive Elimination Process (CEP)' for emotion recognition. Optimal parameters were extracted from vocal signals, and a Support Vector Machine (SVM) was utilized for emotion classification. The algorithm was validated using the Berlin Adults' Emotional Speech Database. Additionally, Florian B. Pokorny et al. [35] presented a low-complexity and robust emotion recognition system for detecting negative emotions in speech signals. This approach leveraged a Bag-of-Words (BoW) feature extraction technique inspired by document retrieval applications. The system was evaluated on a novel database created with children's speech.

Q. Jin et al. [36] suggested using both acoustic and lexical levels to identify emotions like happy, sad, angry, and neutral. First, extracts low-level features from the acoustic level,

ncluding jitter, shimmer, intensity, F0, and spectral contours, among others. Next, based on these low-level features, several representations of acoustic features are generated: a new representation derived from a set of low-level acoustic code words, a new representation derived from Gaussian Super vectors, and statistics over these features. A novel feature representation at the lexical level called emotion vector (eVector) is put forth and makes use of the conventional Bag-of-Words (BoW) feature.

3.1 Proposed Violence Detection Framework

The suggested framework for detecting violence is completed in two steps, specifically:

1) Feature extraction and
2) Classification.

The necessary set of features is extracted from the audio signals during the feature extraction phase, and a supervised learning algorithm is used to process the extracted feature set for classification during the classification phase. Initially, the system is trained through different set of signals and these signals are extracted from different Hindi movies. For every audio clip, a set of features are extracted and they are trained through SVM algorithm. In the testing phase, one audio clip is given for testing and the same features are extracted and then processed for classification through SVM classifier.

In this method, the binary SVM classifier was used for classification purpose. Classifying the audio signals into deeper classes requires multiple attempts due to the SVM classifier's binary classification nature. This section begins with an explanation of the ontological structure of violence, followed by an

examination of the audio features that were taken into consideration for extraction. Lastly, a representation of the Multi-class SVM architecture is given. Figure 1 depicts the general architecture of the suggested violence detection system.

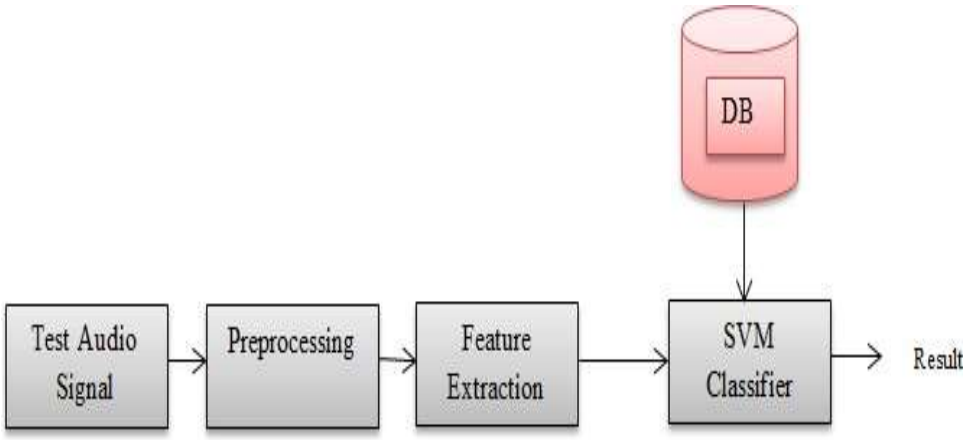

Figure.1 Generalized Architecture of audio features based violence detection system

3.3 Semantic audio for violence

A simple factor cannot provide a conclusion about violence because there are different definitions of violence from different points of view. Sounds such as screams and loud, abrupt noises can be classified as violent in the context of audio data. Thus, this method was successful in defining violence in relation to an ontological framework. Furthermore, since this chapter solely addressed audio data, the definition of violence only takes into account audio signals. In terms of violence, the extra auditory cues aid in improving the accuracy of the system for detecting violence.

The person's scream and speech determine the Person Related Sound. Emotional and neutral sounds are additional categories for speech-related sounds. Ultimately, the audio signals are categorized as angry, sad, happy, or surprised based on their emotional state. The sound associated with a weapon varies according to the weapon—such as a gun, sword, bottle, or other object—that is used to inflict harm. The abruptness and smoothness of the sounds in the environment are used to define violence. Screams, speech, gunshots, sharp and smooth environmental sounds, and fights (beatings) are therefore the audio classes of interest. Figure 2 shows the hierarchy of Violence ontology obtained from the audio modality.

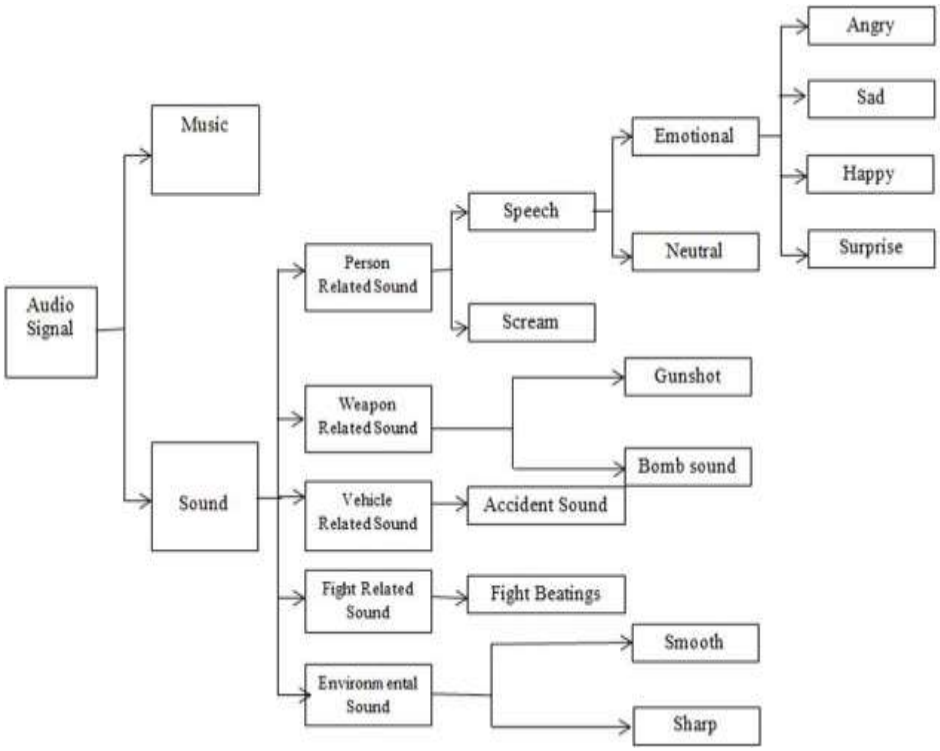

Figure.2. Hierarchy of Violence ontology obtained from the audio modality.

A total of twelve features are extracted from each audio segment in accordance with the suggested feature extraction methodology. To achieve this, each audio signal was first divided into a few non-overlapping segments according to the amount of time that had passed. Because each segment contains 12 distinct features, each segment yields 12 features in total, which together represent all of the information contained in that segment. The collected features are then measured for each audio segment using a standard statistic, such as the average value or standard deviation, to create a 12-D vector that contains all of the information.

Each segmented audio clip that is extracted from the audio sequence is represented using various low-level features, such as 12 MFCCs,Root Mean Square Frame Energy (RMSFE),Pitch, Harmonic Noise Ratio (HNR) and Zero Cross Rate (ZCR).

3.2 MFCCs extraction

For a given audio signal, the MFCCs are derived as follows;

Step 1: Frame the signal into short frames.

Step 2: For each frame calculate the periodogram estimate of the power spectrum. To take the Discrete Fourier Transform of the frame, perform the following:

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \quad (2)$$

Step 3: Apply the Mel filter bank to the power spectra, sum the energy in each filter. The formula for converting from frequency to Mel scale is:

(f) = 1125 ln(1 + f/7 00) (3)

To go back from Mel to frequency:

$$M^{-1}(m) = 700\,(\exp\left(\frac{m}{1125}\right) - 1) \quad (4)$$

Step 4.Evaluate the log for every energy value obtained at step 3 which results in a 26 log filter bank energies.

Step 5.Apply Discrete Cosine Transform (DCT) over the 26 log filter bank energies to obtain 26 Cepstral Coefficients. Among the 26 values, only the first 12-13 coefficients are kept and remaining is discarded.

3.4 RMSFE

In this phase of feature extraction, the RMSFE method extracts the RMS power of every 20ms frame. Further, a local mean is measured for the obtained RMS power at the last second of data. In every frame, after the first second of data, the frame is counted as a low-energy frame, if

it‘s RMS power is lower than the 50% of local mean. Finally, the RMSFE feature is obtained as the ratio of total number of low-energy frames to the total number of frames of every audio sample.

3.5 Multi-Class SVM

Given that hierarchical definitions of violence are taken into account in the audio signals, a hierarchy-based approach to classification is also required. Furthermore, this method also took the SVM algorithm into consideration for classification purposes because of its robustness and efficiency in detecting necessary events. But because it is a binary classifier, the SVM can only classify two classes at once. Therefore, in order to implement the suggested ontological violence detection through SVM classifier, the SVM algorithm is also carried out at several hierarchical instances. When the SVM is used for classification, the three most well-known classification strategies are one-versus-one, one-versus-all, and binary trees. This method approaches the binary tree SVM classifier, taking into account the computational load and additional number of SVM classifiers. When using a binary tree SVM classifier, the count is only "k-1." In contrast, the number of SVM classifiers needed in the one-versus-one and on-versus-all scenarios is observed to be k(k-1)/2 and "k," respectively. To categorize the ontological violence events from the audio signal, the suggested classification model thus just used the binary tree SVM. Sound and music classes are created from the signal during the first stage of classification. In this instance, sound is viewed as violent, while music is viewed as nonviolent.

Only the furious class—of these four—is regarded as violent; the other classes are non- violent. Subsequently, within the category of weapon-related noises, the sounds of a gunshot and a bomb are classified as violent. Only accident sounds are categorized as violent when it comes to vehicle sounds; all other sounds are considered non-violent. The audio signal that depicts beatings is regarded as violent when it comes to fight-related sound. Lastly, depending on their smoothness, ambient noises are also regarded as violent.

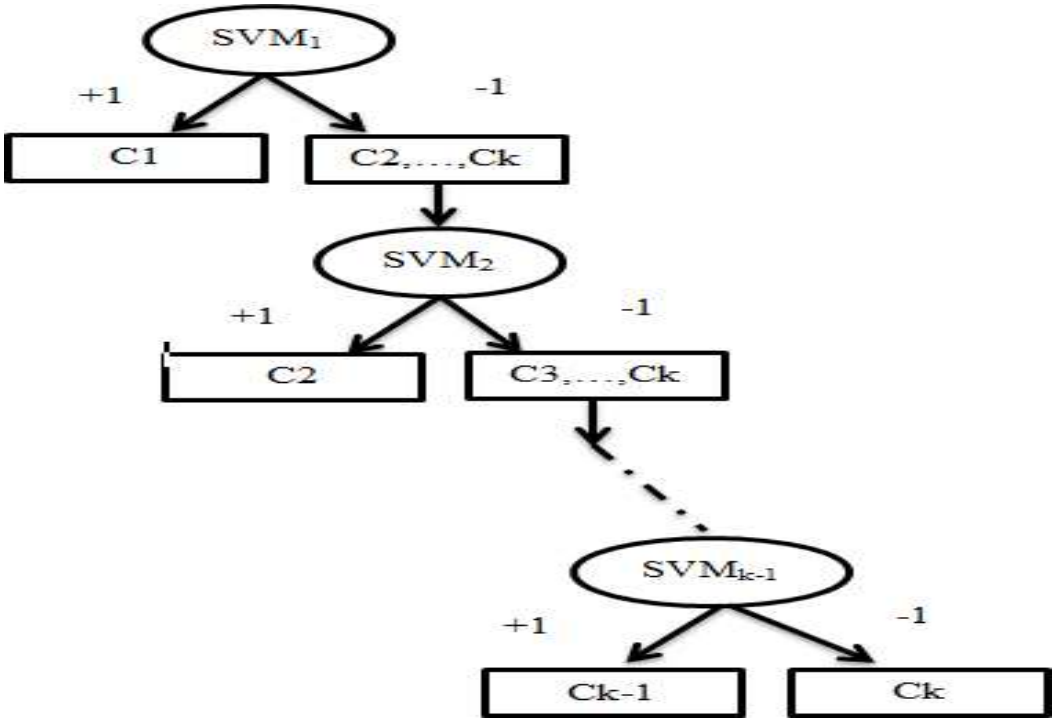

Figure.3 A hierarchy for detecting violence using Binary Tree Multi Class SVM

4. Simulation Experiments

The audio signal is classified as violent if there is a noticeable shift in the ambient sound; otherwise, it is classified as non-violent. In this way, the provided audio test signal is categorized as violent or non-violent, and the performance is checked manually. In this way, a total of 28 features are extracted from an audio sequence and trained into the classifier. The table 1 provides a description of the low-level feature details as well as the statistical functions that are applied to them.

Table.1. Acoustic features and statistical functions

Raw Features	**Statistical functions**
Pitch	Mean, Standard deviation (SD), Kurtosis
Root Mean Square Frame Energy (RMSFE)	Skewness, minimum and maximum value
Zero Cross rate (ZCR)	relative position, Ranges
Harmonic to noise ratio (HNR)	Two linear regression coefficients with their
12 Mel-frequency Cepstral coefficients (MFCC)	Mean Square Error (MSE) of regression coefficients

In this phase, the Zero-Crossing Rate (Total number of zero crossings per sample) is measured over every frame of 20ms time span. Then the local variance of ZCR is measured over every second of data (Ex. 50 frames of data per second). Finally the mean of the local variances is measured and it is considered as the ZCR of that particular audio sample. The samples of ZCR representations of a music and speech samples are shown in figure.4 and 5 respectively.

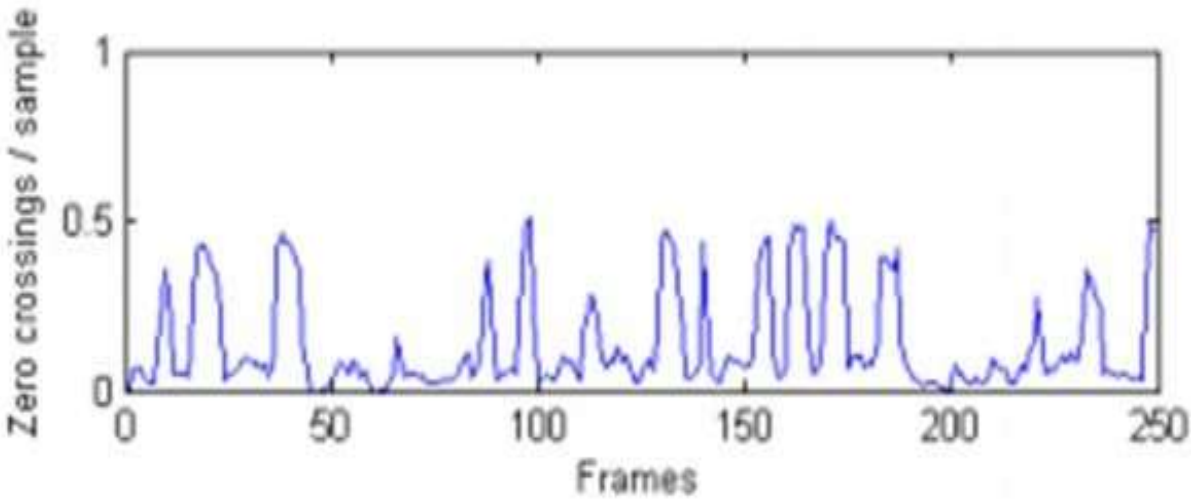

Figure 4 ZCR of a Music sample

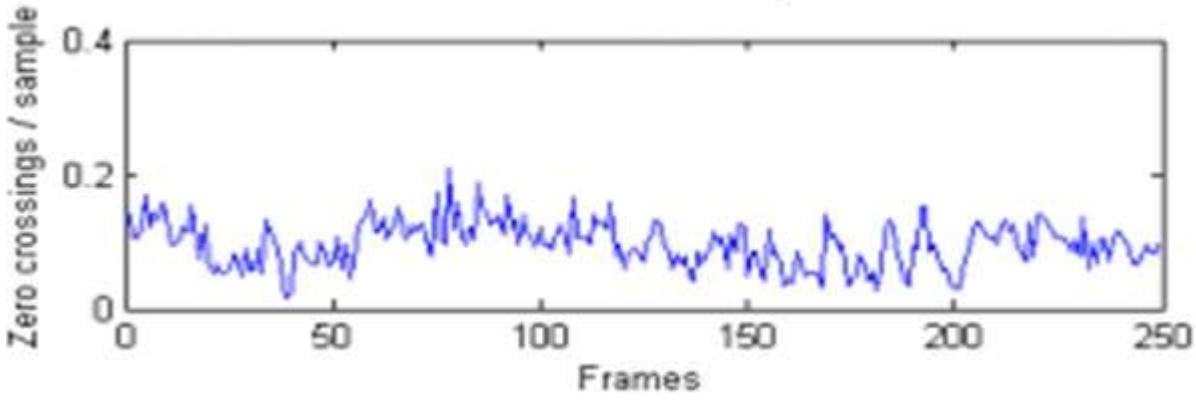

Figure 5 ZCR of a speech sample

In order to assess the developed method, a new dataset sourced from Indian movies is subjected to extensive simulations. Additional performance metrics, such as F-Measure, Accuracy, Recall, Precision, and False Positive Rate (FPR), are used to gauge performance.

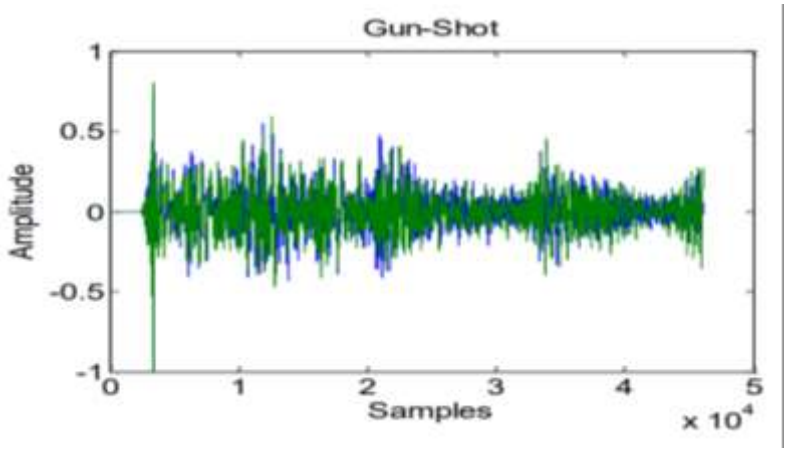

Figure 6 Gun shot signal

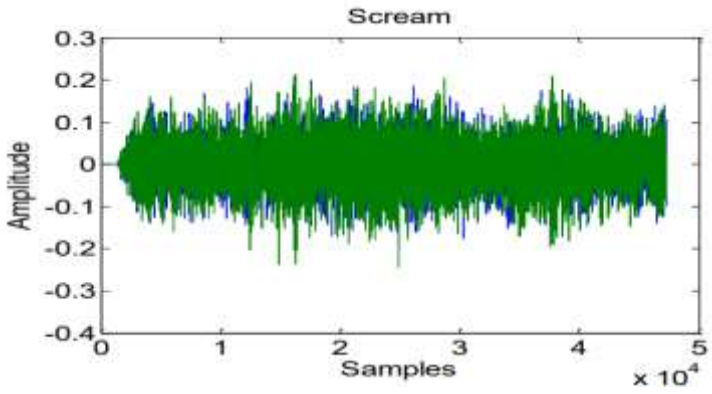

Figure 7 Scream signal

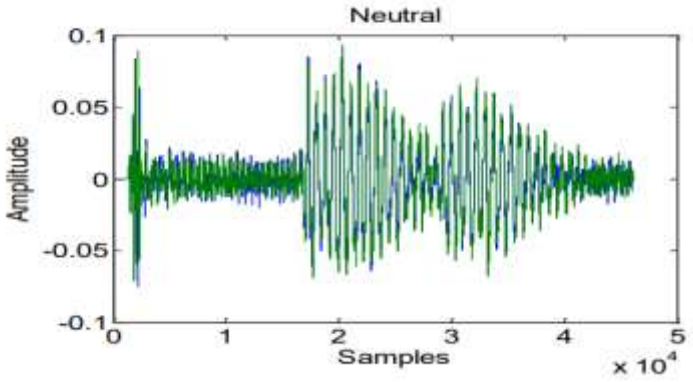

Figure 8 Neutral signal

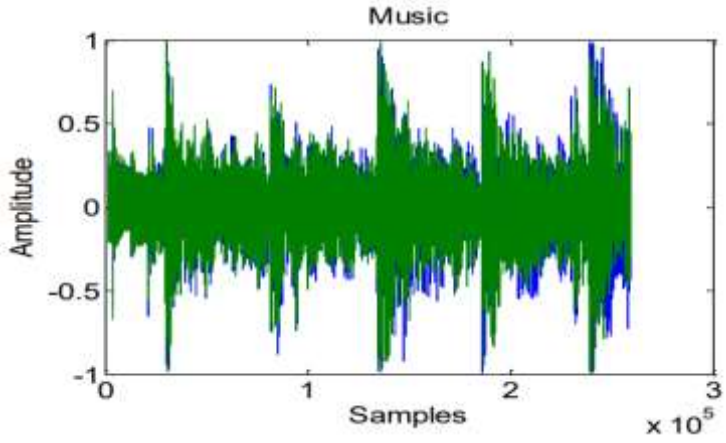

Figure 9 Music signal

Based on the above sample confusion matrix, the detected signals are formulated in the following tables 2 and 3 respectively.

Table 2 Confusion matrix for sound and music

		Predicted		**Total**
		Sound	**Music**	
Actual	**Sound**	**1258**	542	1800
	Music	494	**1162**	1656
Total		1752	1704	**3456**

The provided dataset comprises 1800 test sound signals, of which 1258 were accurately identified as sound signals, while 542 were misclassified as music signals. Conversely, out of 1656 test music signals, 1162 were correctly identified as music signals, and 494 were incorrectly classified as sound signals.

Table 3 Confusion matrix for sound related classes

	PRS	WRS	VRS	FRS	ERS	Total
PRS	**719**	88	48	85	60	1000
WRS	15	**138**	13	16	18	200
VRS	08	18	**144**	13	17	200
FRS	11	21	12	**130**	26	200
ERS	14	18	22	21	**125**	200
Total	767	283	239	265	246	**1800**

As shown in the above table, out of 200 angry emotional speech signals, only 182 signals are declared as angry and remaining 18 (6-Happy, 3-Sad, and 9-Surprise) are detected as other.

5. Conclusion

The significance of machine detection of all aspects has increased due to the swift advancement of technology and its impact on human life. As a result, this strategy also concentrated on creating an automated system for detecting violence using machine learning algorithms and audio data. Determining the violence in a limited number of orientations does not increase the detection system's robustness because violence has multiple definitions. As a result, this method attempted to cover all potential definitions of violence while also considering an ontological definition of violence. In order to achieve this, the suggested system extracted 2000 audio signals from ten distinct Indian Bollywood films and used them as a test set. The detection system measures and processes a set of features for each audio signal. Following the simulation, the effectiveness is assessed using ROC metrics, and it is noted that the suggested system has performed better in identifying all forms of violence. The accuracy improvement that the suggested approach yielded, on average, was 2.2421% and 1.8919% higher than that of the conventional approaches, as reported by Sara et al.[27] and Vu Lam et al. [19]. Next, it is noted that the false positive rate has decreased by 0.0628% and 0.0413%, respectively, according to Sara et al. [27] and Vu Lam et al. [19]. This method classified a movie scene as violent or non-violent based solely on auditory cues. Video semantics can be taken into consideration for violence detection in addition to audio semantics, which will make it more effective and allow for future extensions of this work.

REFERENCES

[1] Ángel Vidal, M., Clemente, M., & Espinosa, P., ―Types of media violence and degree of acceptance in under-18s. Aggressive Behavior‖, Vol.29, No.5, (2003) 381–392.

[2] Kevin D Browne, C. H. G., ―The influence of violent media on children and adolescents: A public-health approach‖. Lancet, Vol.365, (2005), 702–710.

[3] J. Nam, M. Alghoniemy, and A. Tewfik., ―Audio-visual content-based violent scene characterization‖, In IEEE International Conference on Image Processing, pp. 353–357, 1998.

[4] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao., ―Detecting Violent Scenes in Movies by Auditory and Visual cues‖, In Springer Advances in Multimedia Information Processing, pp.317–326, 2008.

[5] L. H. Chen, H. W. Hsu, L. Y. Wang, and C. W. Su., ―Violence Detection in Movies. In IEEE International Conference on Computer Graphics‖, Imaging and Visualization, pp. 119–124, 2011.

[6] J. T. Geiger, B. Schuller, and G. Rigoll, G, ―Large-Scale Audio Feature Extraction and SVM for Acoustic Scene Classification‖, In: *Proc. Of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* New Paltz, NY,. USA, pp.1-4, 2013.

[7] Muthumari Arumugam, Mala Kaliappan, ―An efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals‖, *Circuits and Systems*, Vol.7, No.5, pp.255-279, 2016.

[8] Y. Srinivasa Murthy, and S. G. Koolagudi, ―Classification of Vocal and Non-Vocal Regions from Audio Songs Using Spectral Features and Pitch Variations‖, In: *Proc. of IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE),* Halifax, pp.1271- 1276, 2015.

[9] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, ―Violence content classification using audio features‖, In: *Proc. of the 4th Hellenic Conference on Artificial Intelligence*, Crete, Greece, pp: 502–507, 2006.

[10] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, ―A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks‖, In: *Proc. Of IEEE International Workshop On Multimedia Signal Processing*, Crete, Greece, pp: 90–93, 2007.

[11] J. LudeñaChoez, and A. Gallardo-Antolín, ―Feature Extraction Based on the High-Pass Filtering of Audio Signals for Acoustic Event Classification‖, *Computer Speech & Language*, Vol.30, No.1, pp.32-42, 2015.

[12] Y. Zhang, D. J. Lv, and H. S.Wang, ―The application of multiple classifier system for environmental audio classification‖, *Applied Mechanics and Materials*, Vol. 462-463, No.11, pp. 225–229, 2014.

[13] Y. Thaweesak, ―Spectral Entropy in Speech for Classification of Depressed Speakers‖, In: *Proc. of International Conference on Signal-Image Technology & Internet-Based Systems*, Naples, Italy, pp.1-6, 2016.

[14] M. Baelde, C. Biernacki, and R. Greff. ―A mixture model-based real-time audio sources classification method‖, In: *Proc. of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP*), New Orleans, United States, pp.52-59, 2017.

[15] McLoughlin, H. M. Zhang, Z. P. Xie, Y. Song, and W. Xiao, ―Robust Sound Event Classification using Deep Neural Networks‖, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.23, No.3, pp.540–552, 2015.

[16] SouliSameh ,ZiedLachiri, ―Using Spectro-Temporal Features for Environmental Sounds Recognition‖, *American Journal of Circuits, Systems and Signal Processing*, Vol. 1, No. 3, 2015, pp. 60-68.

[17] S. Chachada, and C. C. J. Kuo, ―Environmental sound recognition: A survey‖, In: *Proc. of International Conference on Signal and Information Processing*, Kaohsiung, Taiwan, pp.45-50, 2013.

[18] S. Sarman and M.Sert, ―Audio based Violet Scene Classification using ensemble learning‖, In: *Proc. of International Symposium on Digital Forensic and Security (ISDFS)*,Antalya, Turkey, pp.1-6, 2018.

[19] V. Lam, S. Phan, and D. DInh, ―Evaluation of Multiple features for violent scenes detection‖, Journal of *Multimedia Tools and Applications*, Vo.76, No.5, pp.7041-7065, 2017.

[20] M. B. Duran, R. G. Pita, H. S. Hevia, ―Acoustic Detection of Violence in Real and Fictional Environments‖, In: *Proc. of International Conference on Pattern Recognition Applications and Methods*, Porto, Portugal, pp.456-462, 2017.

[21] R. Gil Pita, B. Lopez Garrido, and M. Rosa Zurera, ―Tailored MFCCs for sound environment classification in hearing aids‖, *Advanced Computer and Communication Engineering Technology, Lecture Notes in electrical engineering*, Vol.315, No.3, pp. 1037– 1048, 2015

[22] M. Jalil, F. A. Butt, and A. Malik, ―Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals‖, In: *Proc. of international Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE*), Konya, Turkey, pp. 208–212, 2013.

[23] B. Schuller and A. Batliner, ―Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing‖, New York, NY, USA: Wiley, Nov. 2013.

[24] M. A. Quiros Ramirez, T. Onisawa, ―Considering cross-cultural context in the automatic recognition of emotions‖, *International Journal of Machine Learning Cybernetics*, Vol. 6, No.1, pp. 119–127, 2015.

[25] T. S. Gunawan, M. F. Algifhari, M. A. Morshidi, Mira KArtiwi, ―A review on emotion recognition algorithms using speech analysis‖, *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, Vol. 6, No. 1, pp. 12-20, 2018.

[26] S.Demircan and H.Kahramanl, ―Feature Extraction from speech data for Emotion recognition‖, Journal of advances in computer networks, Vol.2, No.1, 2014, pp.28-30.

[27] M. Sara, S. Saeed, and A. Rabiee, ―Speech emotion Recognition Based on a Modified Brain Emotional Learning Model‖, Biologically Inspired Cognitive Architectures, Vol.19, No.1, pp.32–38, 2017.

[28] P. Sathit, ―Improvement of Speech Emotion Recognition with Neural Network Classifier by Using Speech Spectrogram‖, In: *Proc. Of International Conference on Systems, Signals and Image Processing (IWSSIP)*, London, UK, pp.1-8, 2015

[29] W. Lim, D. Jang, and T. Lee ―Speech Emotion Recognition using Convolutional and Recurrent Neural Networks‖, In: *Proc. Of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, South Korea, pp.1-8, 2017.

[30] D. Torres Boza, ―Hierarchical sparse coding framework for speech emotion recognition‖, Journal of *Speech communication*, Vol. 99, No.5, pp. 80-89, 2018.

[31] F. Chenchah and Z. Lachiri, ―Speech emotion recognition in acted and spontaneous context‖, *6th International conference on Intelligent Human Computer Interaction*, IHCI 2014.

[32] Tian Han, Jincheng Zhang, Zhu Zhang, Guobing Sun, Liang Ye, HanyFerdinando, EskoAlasaarela, TapioSeppänen, Xiaoyang Yu and Shuchang Yang, ―Emotion recognition and school violence detection from children speech‖, *EURASIP Journal on Wireless Communications and Networking*, 2018, 99.1-10.

[33] H. Saarimaki, A. Gotsopoulos, I.P. Jaaskelainen, Discrete neural signatures of basic emotions. Cereb. Cortex 26(6), 2563–2573 (2016).

[34] H. Schlosberg, Three dimensions of emotion. Psychol. Rev. 61(2), 81–88(1954).

[35] Florian B. Pokorny, Franz Graf, Franz Pernkop, Bj¨orn W. Schuller, ―Detection of Negative Emotions in Speech Signals Using Bags-of-Audio-Wordsl, *International Conference on Affective Computing and Intelligent Interaction (ACII),* Xian, China, 2015.

[36] Q. Jin, C. Li, and S. Chen. Speech emotion recognition with acoustic and lexical features. *PhD Proposal*, 1:4749–4753, 2015.

[37] D. J. France and R. G. Shiavi. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[38] S. Yildirim, M. Bulut, and C. Lee.An acoustic study of emotions expressed in speech.*Proceedings of Inter Speech*, pages 2193–2196,2004.

[39] Perriere, G.; Thioulouse, J. (2003), "Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins", *Computer Methods and Programs in Biomedicine*. **70** (2): 99–105.

[40] Cohen et al. Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences 3rd ed. (2003). Taylor & Francis Group.

[41] C. M. Lee and S. S. Narayanan,Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*,13(2):293–303, 2005.

[42] Abdi. H. & Williams, L.J. (2010)< "Principal component analysis". *Wiley Interdisciplinary Reviews: Computational Statistics*. **2** (4): 433–459