International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-5, January 2021

Violence Content Detection Based on Audio using Extreme Learning Machine

Mrunali D. Mahalle, Dinesh V. Rojatkar

Abstract: ***In this paper, we proposed an audio based violent scene detection system. As visual based approach has been widely used in identification of violent scenes from video data, audio-based approach; on the other hand, has not been explored as much as visual approach of the video data. In some applications such as video surveillance, visual scenes can be absent because of environmental situations. Also, in many approaches different systems are proposed for movies and real time videos. Therefore, we practiced the audio approach of video data to decide whether a video scene is violent or not from movies and real time videos. For this purpose, we propose an Extreme Learning Machine (ELM) method to detect video scenes as "violent" or "non-violent" using two types of datasets Standardized Media Eval VSD-2014 and other is Customized dataset for the same classifier. After successful training and testing, 85.7% accuracy is achieved by ELM for VSD-2014 dataset and 88.89% for Customized dataset respectively.***

Keywords**: Violent Scene Detection (VSD), Audio Based System, Extreme Learning Machine (ELM), VSD-2014 Dataset, Customized dataset.***

I. INTRODUCTION

During last few decades, videos, movies on internet, and other entertaining media, also videos in surveillance system have been increasing rapidly, resulting in 90% of the video traffic. As because of the fact that connection of internet has spread widely and accessing online content from devices like smart TVs, mobiles and tablets are currently a standard. Many video and audio tracks on internet and other digital media-streaming devices which contains violent scenes can have bad impact on all three generations, but especially on children and youths. It is therefore obvious that the need of protection of sensitive social groups (e.g. children) is imperative. The violent scene detection in videos also have its realistic significance in a number of applications, such as sensible surveillance, video retrieval, Internet filtering, film rating, toddler protection towards violent behaviour and so on [1][2]. As, movies significantly differ from video surveillance material because, movies are highly edited, including special visual and audio effects that are not present in video surveillance, and portray a large spectrum of violence which is rarely seen in real life [1]. We proposed a system for detection of violence in movies using Standardized Media Eval VSD-2014 dataset and real time videos using Customised dataset. As, visual based approach has been widely used in identification of violent scenes from video data, audio-based approach; on the other hand, has not been explored as much as visual approach of the video data. Also, in some scenarios such as video surveillance, visual scenes can be absent because of environmental situations. Audio track also contains much information that visual cues cannot represent. For example, screaming, explosions, words of abuse and even thrilling music or sound effects which can be violent. Also, computational cost and space required for audio features is less as compared to visual features [3]. Therefore, for the prevention of violence in the digital content material and for the security surveillance gadget is important, and this can be finished by means of violent scene detection system (VDS).

Manuscript received on January 03, 2021.
Revised Manuscript received on January 15, 2021.
Manuscript published on January 30, 2021.

Mrunali D. Mahalle*, Department of Electronics Engineering, Government College of Engineering, Amravati, Amravati, Maharashtra, India. Email: mmahalle27@gmail.com

Dinesh V. Rojatkar, Department of Electronics Engineering, Government College of Engineering, Amravati, Amravati, Maharashtra, India. Email: dinesh.rojatkar@gmail.com

II. RELATED WORK

C. Clavel. [4] proposed a surveillance system based on audio so as to detect violent events in crowded places where the environment is full of noise using Gaussian Mixture Model (GMM). For detection of abnormal situations, they considered event such as cries, shots and explosions. In order reduce false rejection (FR) and False Detection (FD) rate, in this approach many experimentations are carried out. Giannakopoulos T. [5] proposed a system by utilizing audio features with support vector machine (SVM) classifier by extracting audio segments from real movies for classification of violence content. Scenes such as shots, explosions, fights and screams, the violent audio segments were extracted whereas music and speech non-violent audio segments were extracted, together with which have similar characteristics to violent sounds are the non-violent sounds like fireworks. Giannakopoulos T. [6] further extended his work using multi-class classification algorithm, for audio segments recorded from movies for violence detection by. By the usage of Bayesian Network, for classification of audio segment into six classes 3 for violent which contains scenes such as shots, fights, screaming and 3 for non-violent which are music speech and other non-violent sounds, particularly by using 12D audio features the binary classification task was accomplished. Esra Acar. [7] proposed a violent scene detection (VSD) system for Hollywood movies by using midlevel audio features with Bag of audio words (BoAW) representation based on mel-frequency cepstral coefficient (MFCC) by using two different methods, namely the vector quantization-based (VQ-based) method and the sparse coding-based (SC-based) method. Md. Zaigham Zaheer et al.

[8] proposed a system for scream sound detection in surveillance system based on deep learning. In order to detect different screaming sounds which are produced in different situations, the Deep Boltzmann Machines (DBM) algorithm was used. The feature used is the MFCC an input to the system.

By the proposed system 100% accuracy is achieved, by using its self-recorded scream dataset. Vivek P. [9] proposed a system using multiple instance learning (MIL) approach based on audio features for binary classification of news. Acoustic features like MFCC and Perceptual Linear Prediction (PLP) from News based on audio signals were extracted from each instance which were segmented. Bags are formed of Features of the instances having same audio files which are grouped together. Bags and instances which were fed to classifiers, proper labels were assigned to them. mi-Graph and mi-SVM were two methods which were used as a MIL classifier. Guankun Mu [10] proposed a system based on deep audio features for violent scene detection using convolutional neural network (CNN). They used CNN in two ways as a deep audio feature extractor and as a classifier. S. Sarman [11] proposed a system based on audio using ensemble learning approach for violent scene classification. Medial Eval VSD-2014 dataset were used in which 31 extremely violent and non-violent movies are present, in order to detect audio based violent scenes. Acoustic features, such as MFCC from time domain and zero crossing rate (ZCR) from frequency domain were used. SVM with ZCR showed better performance as compared to others, out of three classifiers such as SVM, Random Forest, and Bagging data which were used for classification.

III. PROPOSED WORK

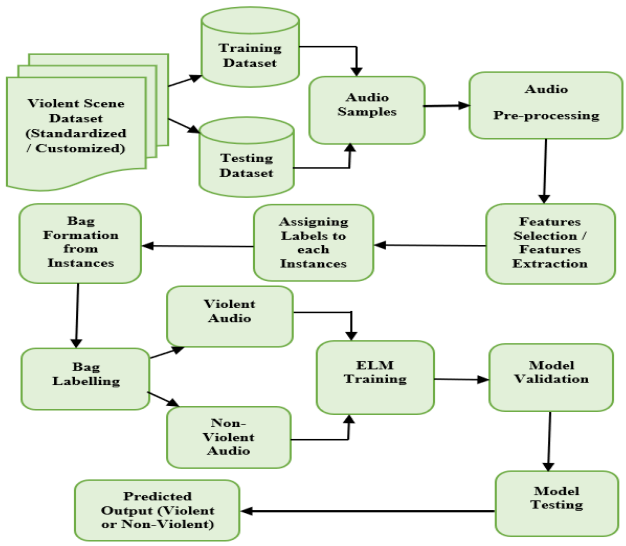


Fig.1. Block diagram of proposed work

The block diagram in "fig. 1". gives a view of the proposed work:

A.Data Generation

1) Media Eval VSD-2014 Standardized Dataset:

In this proposed work, we use Media Eval VSD-2014 dataset. Annotations of 31 movies from which 24 movies are used in the Hollywood: Development and 7 in the Hollywood: Test set are present in this dataset. Three types of content are included in VSD2014 dataset they are movies/videos (and metadata), features, and annotations. Each movie present in both training and test are having different levels of violence [1] [16]. In this experimentation we only used audio modality from movies as each frame are annotated with 3 different audio perceptions, they are Gunshot, Explosion and Screaming Sound.

2) Customised Dataset:

We manually, collected real time data from You Tube videos which includes some violent and non-violent videos. These videos contain violent scenes like Gunshots, Screaming, Explosions and Arguments from real world. From these videos, MP3 audio file are extracted, for a particular duration of 30 seconds with a sampling rate of 48kHz and bitrate of 256 bit/sec. We divided the dataset into non-overlapping sets in the proportion of 70% and 30% which contains total 60 MP3 audio files. The 42 files are for training i.e. 70% of data and 18 files are for testing purpose i.e. 30% of data.

B. Pre-processing

The aim of audio pre-processing is to develop a dataset compatible with a violence detection system. Audio signal has to be transformed to a digital format, before mathematical operations can be applied to it. The data can be processed digitally, when the data is transformed from an analog to a digital signal. Hence, this digital signal is converted into a more usable format. "Signal analysis transforms a signal from one domain to another, for example from the time domain to the frequency domain. By transforming the signal, the intent is to emphasize information in the signal and cast it into a form that is easier to extract". The wave or mp3 input files are transformed into the frequency domain because this is a representation of the audio input over which machine learning models can efficiently generalize in most of the machine learning applications. The data has to be processed in small frames, in order to transform the data from the time domain to the time-frequency domain. All the audio files were resampled to a sampling rate of respective frequency in Hz and converted to mono channel in order to have them with the same characteristics. Although there are certain floating-point file formats that contain values greater than (-1,1), the amplitude range of most of the audio files is (-1,1). However, this was observed just for very few of the examples. It is required that all inputs files have the same duration, as the network architectures considered in the present study do not accept inputs of different sizes. The duration was chosen to be thirty seconds. Therefore, a last step in the processing of the audio files was to zero-pad all the files that did not reach the thirty seconds duration. In proposed scheme we extend and adjust the audio files. We appended the same file as many times as needed until reaching the desired seconds file duration.

C.Feature Selection/Extraction

Features of audio samples, which are utilized to train a classifier, plays a significant role for correct detection. To improve the execution of the system, it is perhaps the main factors that are likely to be used. In case, of scenes from videos, such as firework and explosion have similar sound effects, may lead to misclassification. Audio signals also contains different sound effects such as screaming, gunshots, explosions, etc.

This will produce the difficulty for the system to detect violent scenes. In order to reduce the data dimensionality by extracting the most important features from audio samples, is the main objective of using feature extraction. A set of features can be useful for representing the characteristics of audio samples, when the feature vector dimensionality is smaller.

Moreover, if there is no large training data available is available, feature extraction can play an important role for obtaining high accuracies in violent scene detection systems. Several features utilized for violent scene detection are discussed below.

1. Temporal Feature

a. Zero Crossing Rate (ZCR):

ZCR represents the number of zero crossings in time domain per frame present in a given signal. ZCR is mathematically expressed as:

The zero-crossing rate of a stationary signal can be defined by the "(1)".

$$ZCR = \sum_{i=-\infty}^{\infty} |sgn(t(n)) - sgn(t(n-1))| \quad (1)$$

Where sgn() is a signum function and is defined as by the "(2)".

$$sgn(t(n)) = \begin{cases} 1 & if\ t(n) \geq 0 \\ -1 & if\ t(n) < 0 \end{cases} \quad (2)$$

The "(1)" can be modified for non-stationary signal like speech which is known as short term ZCR by the "(3)" that computes ZCR for the i^{th}analysis frame of length N of the speech signal.

$$ZCR(n) = \frac{1}{2N} \sum_{m=0}^{N-1} t(m)x(n-m) \quad (3)$$

The factor 2 (by symmetric feature of speech signal) comes to take care from the fact that one cycle of a signal gives two zero crossings values.

b. Amplitude Envelope (AE)

Amplitude envelope defined as the variations in the amplitude of a sound over time, and is a significant property as it affects our perception of timbre. This is an important property of sound, because it is what allows us to easily classify sounds, and distinctively separate them from other sounds [12].

c. Short Time Energy (STE):

The Short Time Energy (STE) is the energy associated within the short audio signal is time varying in nature. The loudness of a speech signal is the most prominent characteristics according to human aural perception. There are several interchangeable terms like volume, energy, intensity etc. which are commonly used to describe the loudness of speech signals. That is why short time energy is computed from a speech signal for describing the loudness. STE is mathematically expressed as:

$$E_T = \sum_{n=-\infty}^{\infty} t^2(n) \quad (4)$$

d. Root Mean Square Energy (RMS):

The square of the amplitude represented by the waveform is defined as the energy of a signal. The energy transmitted per unit time (second) is the power of a sound. As a result, power is the mean-square of a signal. Hence, the root of power (root-mean-square, RMS) is useful for feature extraction.

2. Frequency Feature

a. Spectral Flux (SF):

The Spectral Flux measures how rapidly the spectrum of a signal is changing. It is calculated by computing the difference between the present spectrum which of the previous frame.

SF is mathematically expressed as:

$$SF_n = \sum_{K=0}^{M/2} (|X_{Kn}| - |X_{K(n-1)}|)^2 \quad (5)$$

b. Spectral Centroid (SC):

Spectral Centroid is defined as the spectral centre of gravity of the magnitude spectrum. It is used as an indicator of the brightness of sound in a given audio signal. The SC determines the point in the spectrum where most of the energy is concentrated and is linked with the dominant frequency of the signal.

SC is mathematically expressed as:

$$SC_n = \frac{\sum_{K=0}^{\frac{M}{2}} K \cdot |X_{Kn}|}{\sum_{K=0}^{\frac{M}{2}} \cdot |X_{Kn}|} \quad (6)$$

c. Bandwidth (BW):

Bandwidth is mainly defined as the magnitude-weighted average of the differences between the spectral components and the spectral centroid.

d. Spectral Rolloff (SR):

The spectral rolloff is the frequency $L_M(n)$ below which M% percentile of the power spectral distribution, where M is normally between 85% or 95%. The rolloff point is the frequency below which M% of the magnitude distribution of STFT coefficients are concentrated for the n^{th} frame. As the bandwidth of a signal increases it also increases. It denotes the skewness of the spectral shape. SR is mathematically expressed as:

$$\sum_{K=0}^{K_M(n)} |X_{Kn}| = \frac{M}{100 \sum_{K=0}^{\frac{N}{2}} |X_{Kn}|} \quad (7)$$

e. Band Energy Ratio (BER):

Band Energy Ratio is defined as the ratio of the energy in a specific frequency-band to the total energy [13].

f. Energy Entropy (EE):

The Energy Entropy (EE) defines the abrupt changes in the energy level of the audio signal. When the rapid changes occur in the tone of voice this feature is useful for detecting violence in the audio signal. Apart from the mean and the standard deviation (SD), statistics applied to the energy entropy are the ratios of maximum to mean and maximum to median values [6]. To evaluate this measurement, each time frame of M samples is divided into N blocks, and the energy of each block is then measured. So, EE for the n^{th} time frame can be calculated using:

$$EE_n = -\sum_{n=1}^{N} \sigma_{nm}^2 \log_2 \sigma_{nm}^2 \quad (8)$$

where σ_{nm}^2 is the normalised energy calculated for the n^{th} block of the m^{th} frame, n = 1,...,N.

g. Mel-Frequency Cepstral Coefficient (MFCC):

MFCC is one of the most popular feature extraction techniques used in automatic speech or speaker recognition system using the Mel scale which is based on the human ear scale. The Mel-Frequency Cepstral Coefficients (MFCCs), which are a set of perceptual parameters commonly used in speech recognition, calculated from the spectrum. It provides a compact representation of the spectral envelope.

It is based on the non-linear human perception of the frequency of sounds. These coefficients represent audio based on perception. They are derived from the Mel frequency cepstrum.

The relation between real frequency (Hz) and Mel-frequency is given by the following Equation:

$$F_{mel} = 2595 \log(1 + \frac{F}{700}) \quad (9)$$

The Mel-frequency Cepstral Coefficients (MFCC) C_i of the i^{th} frame is mathematically expressed as:

$$C_i = \sqrt{\frac{2}{M}} \sum_{K=1}^{M} m_K \cos(\frac{\pi i}{M} (K - 0.5)) \quad (10)$$

D. Assigning Label to Each Instances

By extracting features so as to separate different types of sound, for example speech, music, environmental sounds, screaming, silence, and combination of these sounds i.e. normal and violent sounds, instance will be generated from each audio segments. Label assigning is done after all the instances are created. Bags are formed when, all the feature set i.e. group of instances belonging to the same audio sounds files are grouped together. Violent labels will be assigned to the bag in which violent sounds instances are present, and non-violent labels will be assigned to bag in which normal sounds instances are present. For efficient detection, all the bags along with their labels are assigned into an Extreme learning machine (ELM) classifier in order to train the classifier.

E. Classification

We use ELM algorithm for purpose of classification and the experimentation is carried out by giving input of two types of dataset they are Medial Eval 2014 and manually collected dataset. The violent and non-violent bags created are fed to ELM in order to train the model. When the ELM model is trained with the perfect validation accuracy the testing is performed from which outputs are predicted whether the given data is violent or non-violent. ELM is said to be a single layer feedforward neural network (SLFN). The speed of learning of ELM is extremely fast because it avoids multiple iteration. As compared to other deep learning (DL) and machine learning (ML) algorithms it is almost free from human interference. Models available for compression, feature learning, clustering, regression and classification are all homogenous. ELM is easy for small scale real-time learning and control, up to thousand times faster, efficient for multichannel data fusion and potential for decision synchronization then DL. All these features make ELM algorithm capable for classification and detection [14] [15]. Given a single hidden layer of ELM, suppose that the output function of the i^{th} hidden node is $h_i(x) = G(a_i, b_i.x)$, where a_i and b_i are the parameters of the i^{th} hidden node. The output function of the ELM for SLFNs with L hidden nodes is:

$$f_L(x) = \sum_{i=0}^{L} \beta_i \, G(a_i, b_i, x) \quad (11)$$

$$f_L(x) = \sum_{i=0}^{L} \beta_i \, h_i(x) \quad (12)$$

Where, β_i is the output weight of the i^{th} hidden node.
$h(x) = (h_i(x), \dots, h_L(x)$ is the hidden layer output mapping of ELM.

IV. EXPERIMENTAL RESULTS

For verifying the accuracy and effectiveness of the system ELM algorithm with different appropriate features are utilised in order to detect violence in movies and real time videos. The proposed technique has been implemented by using the MATLAB (2019b) environment on Intel Core i5, processor speed 4.0 GHz with 8GB RAM having 64-bit Windows-10 operating system. Proposed method focused on audio based violent scene detection by ELM algorithm by using two datasets as mentioned above.

In order to evaluate the effectiveness of proposed system, this section provides Sensitivity, Specificity, F1-Score and Accuracy values of mentioned algorithm and showed how some features with our algorithm showed best results. Four parameters which are used to analyse performances are:

True Positive (TP) - Correctly detected Violent Scenes

True Negative (TN) – Correctly detected Non-Violent Scenes

False Positive (FP) – Number of Non-Violent Scenes detected wrongly as violent Scenes.

False Negative (FN) – Number of Violent Scenes detected wrongly as non-violent scenes.

Hence, by utilizing these parameters Sensitivity, Specificity, F1-Score and Accuracy can be measured as:

$$Sensitivity = \frac{TP}{TP + FN} \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

$$F1 - Score = \frac{2TP}{TP + TN + FP + FN} \quad (15)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

Sensitivity measures how likely the test is positive when the scene is violent. Specificity measures how likely the test is positive when the scene is non-violent. F1-Score measures model accuracy on dataset. Accuracy measures the number of violent and non-violent scenes detected.

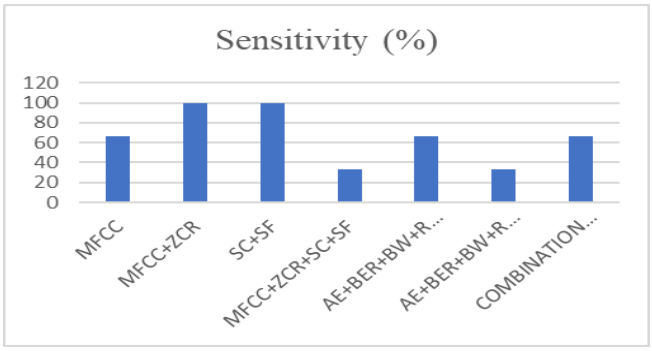
A. Experimentation Results for Standardized Media Eval VSD-2014 Dataset.

TABLE 1. Experimental Results Obtained for ELM of Violence Detection for VSD-2014 Dataset.

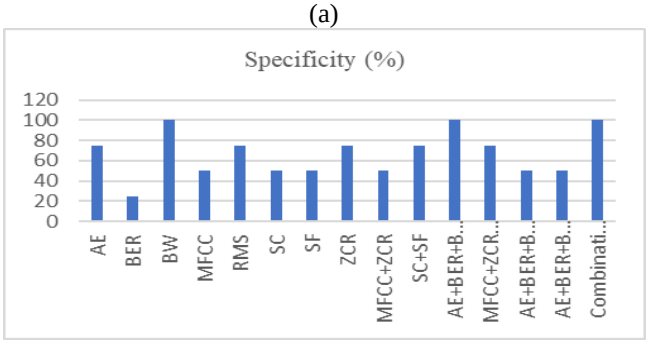
Feature Combination	Parameters			
	Sensitivity (%)	*Specificity (%)*	*F1-Score (%)*	*Accuracy (%)*
AE	0	75	0	42.9
BER	0	25	0	14.3
BW	0	10	0	57.1
MFCC	66.7	50	52.63	57.1
RMS	0	75	0	42.9
SC	0	50	0	28.6
SF	0	50	0	28.6
ZCR	0	75	0	42.9
MFCC+ZCR	100	50	65.22	71.4
SC+SF	100	75	78.95	85.7
AE+BER+BW+RMS	0	100	0	57.1

(MFCC +ZCR) +(SC+SF)	33.33	75	45.45	57.1
(AE+BER+BW+RMS) + (SC+SF)	66.67	50	52.63	57.1
(AE+BER+BW+RMS) + (MFCC+ZCR)	33.33	50	33.33	42.9
Combination of all 8 Features	66.67	100	90.91	85.7

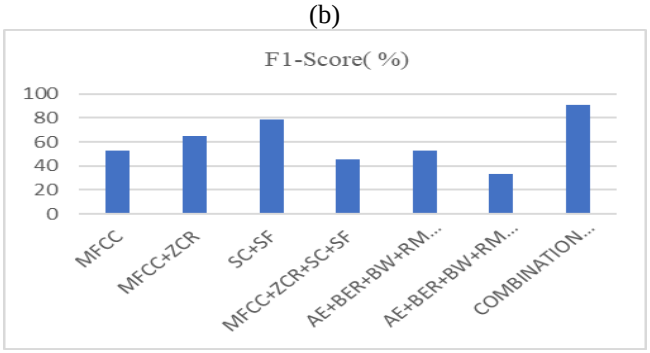
The results are being tabulated in Table I. Table I shows values of Sensitivity, Specificity, F1-Score and accuracy for Media Eval VSD-2014 dataset using various features they are AE, BER, BW, MFCC, RMS, SC, SF, and ZCR which are present in the dataset. The processing time is increased by using feature combinations but accuracy given is high as compared to single features used. The accuracy given by feature combination such as MFCC and ZCR is 71.4%, SC and SF is 85.7%. As MFCC can detect violent speech signals of human and ZCR can capture abrupt changes of audio signal in Violent scene. As SC and SF can detect thrilled violent music from movies as SF can be used for speech/music discrimination [17]. The Overall sensitivity, specificity, F1-score and accuracy obtained by the experimentation using ELM for Media Eval VSD-2014 dataset is 66.67%, 100%, 90.91% and 85.7 % for combination of all eight features. "Fig. 2" shows graphical representations of all values shown in Table I.



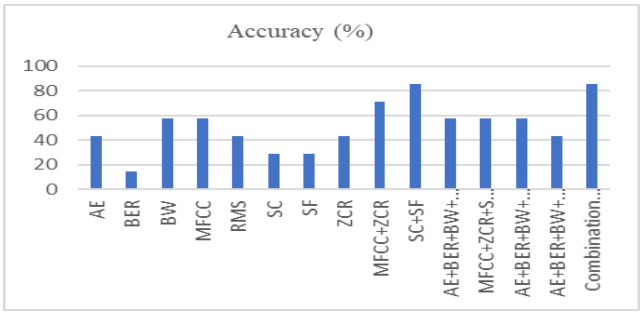
(a)



(b)



(c)



(d)

Fig. 2 Shows results after features integration. (a) Sensitivity. (b) Specificity. (c) F1- Score. (d) Accuracy.

As in [11] the experimentation performed based on audio approach, we prefer using as a baseline to our study. Table II consists of comparison of our work and previous work using same VSD-2014 dataset. Best results in the reference paper is given by Average Precision (AP) at 20 gathered using different algorithms such as Support Vector Machine (SVM), Bagging and Random Forest (RF).

TABLE II. RESULTS COMPARED TO [11]

Parameters / Authors	Ref [11]			Our method		
	Support Vector Machines with MFCC	*Bagging with ZCR*	*Random Forests with ZCR*	*ELM with SC+SF*	*ELM with MFCC + ZCR*	*ELM combination of all 8 features*
Accuracy (%)	-	-	-	85.7	71.4	85.7
Average Precision (%)	0.654	0.596	0.688	0.75	0.6	1
Sensitivity (%)	-	-	-	100	100	66.67
Specificity (%)	-	-	-	75	50	100
F1-Score (%)	-	-	-	78.95	65.22	90.91

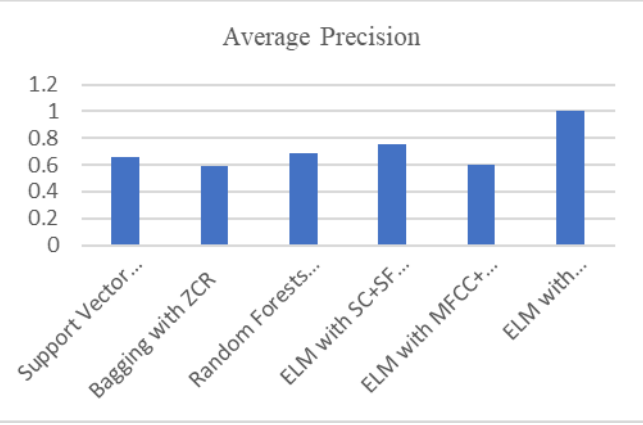


Fig. 3 Shows Best Results for Average Precision.

"Fig. 3" shows Average Precision results [11] proposed methods using Support Vector Machines, Bagging and Random Forests with our proposed methods using Extreme Learning Machine. This figure shows that all of our methods perform better for Average Precision at 1 used in the paper.

The AP predicted is better with feature combinations such SC and SF, MFCC and ZCR, and lastly with combination of all eight features as compared to all other feature combinations, computed by our approach using the Media Eval Violent Scene Detection Task performance dataset.

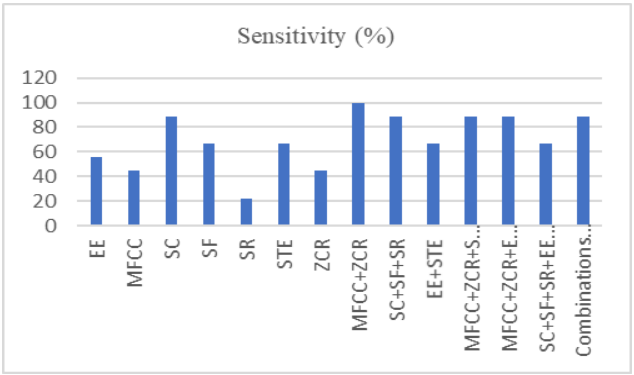
B. Experimentation Results Manually Collected Customized Dataset.

The results are being tabulated in Table III. Table III shows values of Sensitivity, Specificity, F1-Score and accuracy for Manually Collected dataset using various features different from some features given in dataset as movies are significantly different from real time videos, they are EE, MFCC, SC, SF, SR, STE and ZCR. "Fig. 4" shows graphical representation of all parameters given in Table III. Feature Combinations such as MFCC and ZCR given the accuracy of 83.33%, SC, SF and SR given the accuracy of 88.89%. The overall sensitivity, specificity, F1-score and accuracy obtained by the experimentation using ELM for manually collected customised dataset is 88.89 % using all combinations of seven features.

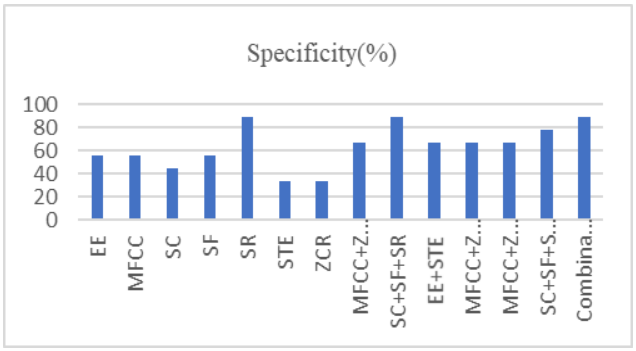
"Fig. 5" shows binary classification confusion matrix for Media Eval VSD-2014 dataset for combination of all eight features and "Fig.6" shows binary classification confusion matrix for Manually Collected dataset for combination of all seven features.

TABLE III. Experimental Results Obtained for ELM of Violence Detection for Customized Dataset.

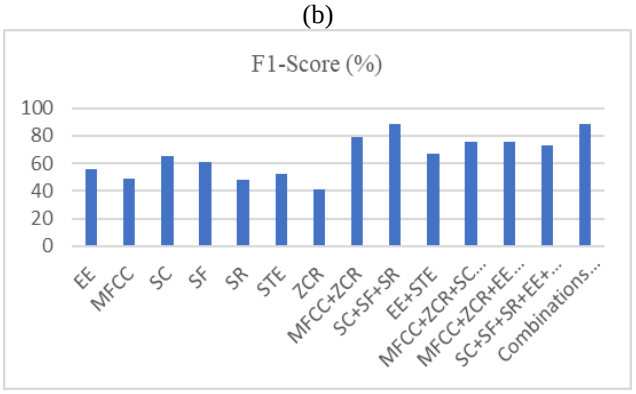
Feature Combinations	Parameters			
	Sensitivity (%)	*Specificity (%)*	*F1-Score (%)*	*Accuracy (%)*
EE	55.56	55.56	55.56	55.56
MFCC	44.44	55.56	48.78	50
SC	88.89	44.44	65.57	66.67
SF	66.67	55.56	61.23	61.11
SR	22.22	88.89	47.62	55.56
STE	66.67	33.33	52.63	50
ZCR	44.44	33.33	40.82	38.89
MFCC+ZCR	100	66.67	78.95	83.33
SC+SF+SR	88.89	88.89	88.89	88.89
EE+STE	66.67	66.67	66.67	66.67
MFCC+ZCR +SC+SF+SR	88.89	66.67	75.47	77.78
MFCC+ZCR +EE+STE	88.89	66.67	75.47	77.78
SC+SF+EE+ STE	66.67	77.78	73.17	72.22
Combination of all 7 Features	88.89	88.89	88.89	88.89



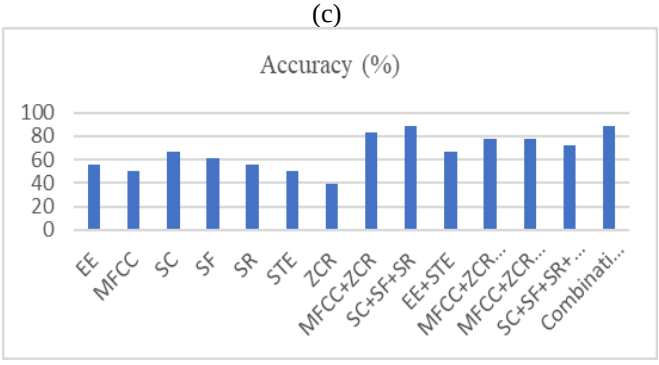
(a)



(b)



(c)



(d)

Fig. 4 Shows results after features integration. (a) Sensitivity. (b) Specificity. (c) F1- Score. (d) Accuracy.

V. CONCLUSION AND FUTURE WORD

In this paper, we proposed an audio-based system by practicing the same ELM classifier for violence identification by utilizing two different datasets which are Standardised Media Eval VSD-2014 and manually collected dataset. In our method, we have utilised some of most popularly used features for detecting violent scenes from audio signals with an ELM as classifier. The system is first tested using Media Eval VSD-2014 dataset with ELM classifier which contains annotation from movies from which we used audio modality and the accuracy obtained is 85.7%. Secondly, we tested the system using manually collected customised dataset using audio segments extracted from You Tube videos. For this, we considered some different audio features then the features present in standardized dataset. The accuracy obtained for customised dataset using ELM classifier is 88.89%. From these results it is clear that our system based on audio modality using ELM can give high accuracy for violence detection in movies and real time videos using appropriate acoustic features.

In future new audio features can be examined and used so to increase the performance of the system. Finally, in addition to this audio based system visual based approach could be combined, so as to increase the classification performance. In order to detect violent text present in videos audio visual with text based approach could be used in future.

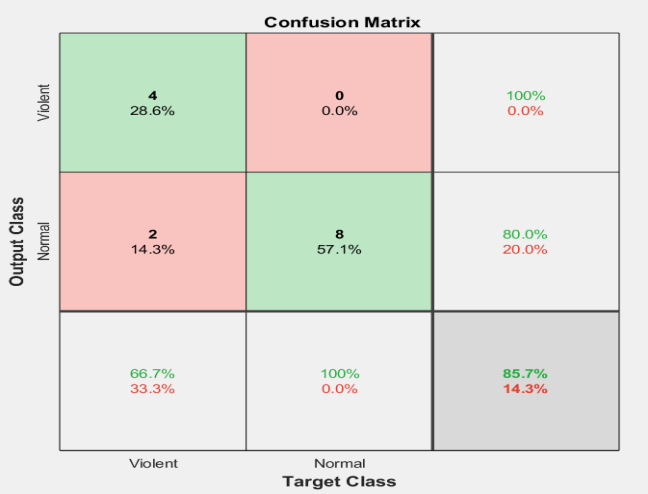


Fig. 5 Confusion Matrix for Media Eval 2014 Dataset

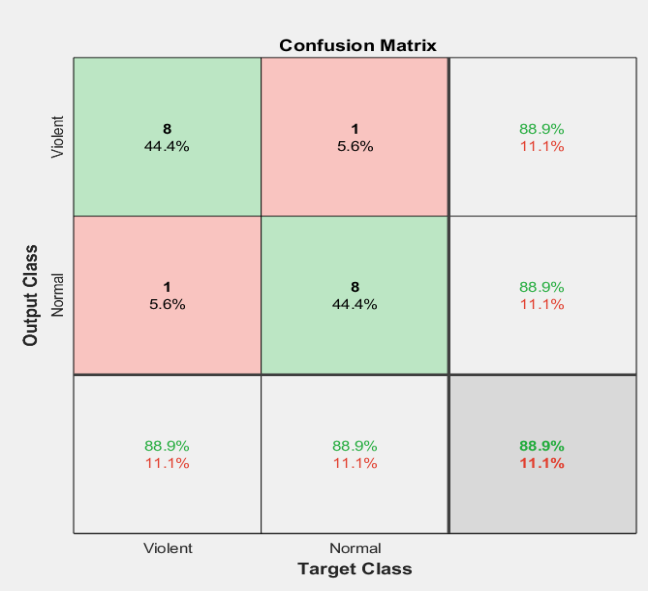


Fig. 6 Confusion Matrix for Customized Dataset

REFERENCES

1. Claire-H´ el`ene Demarty, C´edric Penet, Mohammad Soleymani and Guillaume Gravier. “VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation,” © Springer Science+Business Media New York 2014.
2. Rajeswari Natarajan and Chandrakala.S. “Audio-Based Event Detection in Videos - a Comprehensive Survey”. International Journal of Engineering and Technology (IJET) Vol 6 No 4 Aug-Sep 2014.
3. Marta Bautista Duran, Joaquin Garcia-Gomez, Roberto Gil-Pita, Inma Mohino-Herranz, and Manue Rosa-Zurera. “Energy Efficient Acoustic Violent Detector For Smart City”. International Journal of Computational Intelligence Systems, Vol. 10 (2017).
4. C. Clavel, T. Ehrette and G. Richard, "Events Detection for an Audio-Based Surveillance System," 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, 2005, pp. 1306-1309 (2005).
5. Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. “Violence Content Classification Using Audio Features”.in Hellenic Conference on Artificial Intelligence, 2006, pp. 502–507.
6. Theodoros Giannakopoulos, Aggelos Pikrakis and Sergios Theodoridis. “A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks”. IEEE 9th Workshop on Multimedia Signal Processing (2007).
7. Esra Acar, Frank Hopfgartner, and Sahin Albayrak. “Detecting Violent Content in Hollywood Movies by Mid-level Audio Representations”. IEEE 11th International Workshop on Content-Based Multimedia Indexing (CBMI) (2013).
8. Md. Zaigham Zaheer, Jin Young Kim, Hyoung-Gook Kim, Seung You Na “A Preliminary Study on Deep-Leaning Based Screaming Sound Detection”. International Conference on IT Convergence and Security (ICITCS) 978-1-4673-6537-6/15/$31.00 ©2015 IEEE.
9. Vivek P, Kumar Rajamani, and Lajish V L. “Effective News Video Classification Based On Audio Content: A Multiple Instance Learning Approach”. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (6), 2016.
10. Mu, H. Cao and Q. Jin, “Violent Scene Detection Using Convolutional Neural Networks and Deep Audio Features” Pattern Recognition. CCPR 2016. Communications in Computer and Information Science, vol 663. Springer, Singapore
11. S. Sarman and M. Sert, "Audio based violent scene classification using ensemble learning," *2018* 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-5.
12. https://maplelab.net/overview/amplitude envelope/#:~:text=Amplitude envelope refers to the,distinguish them from other sounds.(27/11/2020)
13. Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo, “Environmental Sound Recognition With Time–Frequency Audio Features”, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17, No. 6, August 2009.
14. Guang-Bin Huang, Qin-Yu Zhu and Chee-Kheong Siew, “Extreme learning machine: Theory and applications”, Neurocomputing, Volume 70, Issues 1–3, 2006, Pages 489-501, ISSN 0925-2312, 2006 Elsevier.
15. Shifei Ding, Han Zhao, Yanan Zhang, Xinzheng Xu and Ru Nie. “Extreme learning machine: algorithm, theory and applications”. © Springer Science + Business Media Dordrecht 2013.
16. M. Sj¨ oberg, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty. The MediaEval 2014 affect task: Violent scenes detection. In Working Notes Proceedings of the MediaEval 2014 Workshop, October 2014.
17. Dalibor Mitrović, Matthias Zeppelzauer, Christian Breiteneder, “Chapter 3 -Features for Content-Based Audio Retrieval, Advances in Computers, Elsevier, Volume 78, 2010, Pages 71-150, ISSN 0065-2458,ISBN 9780123810199,

AUTHORS PROFILE

Mrunali D. Mahalle received her B.E. and M.Tech Degrees from the University of Amravati, in 2018 and 2020, respectively. Her research interests include Artificial Intelligence, Content Classification, and Machine learning. The Author published her research paper in IJRASET journal related to Audio Based Violent Scene Detection.

Dinesh V. Rojatkar received his M.E. degree in Electronics Engineering from the University of SGBAU, Amravati, India, and his Ph.D. degree in Engineering and Technology from the University of SGBAU, Amravati, India, in 2003 and 2013, respectively. From 1999 to 2018 he was at the University of Amravati and Nagpur, as Assistant Professor. He is currently an Associate Professor in the Department of Electronics Engineering, the University of SGBAU, and Amravati, India.