MULTIMODAL EVENT DETECTION IN BIG DATA USING MULTI-LEVEL FUSION CLASSIFIER

K Swapnika[1*]

[1*]Phd Scholar, Computer Science and Engineering,
JNTUH College of Engineering, JNT University, Hyderabad, 500085, India.
Email id : swapnika.griet@gmail.com

D Vasumathi[2]

[2]Head & Professor, Computer Science and Engineering Department
JNTUH College of Engineering, JNT University, Hyderabad, 500085, India.
Email id : vasukumar_devara@jntuh.ac.in

Abstract

Deep learning-based multimodal event detection using the Multi-Level Fusion Classifier (MLFC) model is proposed to overcome these complexities. The data are gathered from the multimodal data, including text, images and audio placed in the Hadoop platform for storage. The data are fed to MLFC, and the modals of text, image and audio are generated through various approaches. The text modal is generated through pre-processing, Enhanced Term frequency - Inverse document frequency (TF-IDF) and attention-based BiLSTM (Attn_BiLSTM). The image modal is generated through the Improved Capsule Network (I-CapsNet). The audio modal is generated by extracting low-level, mid-level and high-level features directed to Convolutional Neural Network _Opposition Salp swarm Algorithm (CNN_OSA). The extracted features are fused through Deep FF (Feature Fusion) strategy, and the various events are classified through the SoftMax classifier. The overall accuracy obtained in classifying the multimodal events is 98.26% which outperforms better when compared to the existing approaches.

***Keywords*: Multimodal events; deep learning; text-image-audio features; feature fusion; big data; softMax classifier.**

1. Introduction

Event detection (ED) is the procedure of exploring the event streams to determine different sets of real-world events and offers a clear understanding of social events. Multimodal ED investigates events from heterogenous vast data such as images, texts, and audio/video. Advanced image processing technology can detect different types of events automatically [Zhou et al. (2020)]. With the increasing growth of multimedia content on the internet and broadcast, creating unstructured multimedia data is searchable and approachable with high flexibility [Cao et al. (2018)]. Various technologies are developed to detect the events in varied scenarios like road traffic event detection [Alomari et al. (2020)], event detection in smart cities [Chen et al. (2021)], event detection in social media, sports event detection etc.

Event detection is mainly important for learning video semantic procedures for video summarization, retrieval and indexing purposes [Liu et al. (2017)]. Hence high research efforts have been committed to detecting the event for video analysis. Many of the previous event detection techniques depend upon videos and domain knowledge features and utilize labelled samples to train event detection models. The semantic gap between reduced level features and enlarged level events of varied types of videos, background clutter, unclear video cues and different alternations of camera motion etc., makes the video analysis more complex and obstructs the implementation process of event detection systems [Lu et al. (2018)].

Moreover, constructing a generic model for event detection in varied video domains like news, sports, surveillance, and movies is hard to classify because of the various domain knowledge in several video genres and the lack of training samples. To consider these problems, most of the recent techniques depend on supervised learning and video content based on labelled video clips for certain event classes [Lohithashva et al. (2020)]. Nowadays, event detection is also performed with the assistance of big data. In general, big data is a large set of data corresponding to a size improved exponentially over time. Compared with other traditional data management tools, big data can store large data and process it efficiently.

This big data is more helpful in event detection because of the large size of the storage space [Granat et al. (2020)]. Also, the past event detection approaches are highly concentrated on a single domain. Still, currently, multimodal event detection methods are developed to find events in huge heterogeneous data like images, texts and video clips. Event detection in a particular domain is a challenging problem for machine learning techniques [Qu et al. (2020)]. Thus, various deep learning (DL) techniques are introduced in many research works related to event detection [Dabiri and Heaslip (2019); Cakir et al. (2017); Pouyanfar and Chen, (2017); Rangasamy et al. (2020); Fernando et al. (2018); Chen and Jin (2019)]. The deep learning models have several processing layers to understand the data representation with numerous levels of abstraction. The existing deep learning models for event detection provide improved results than the traditional machine learning techniques.

Various research works are efforts to generate an effective technique for multimodal event detection in big data analysis. Most of the existing approaches are mainly suitable for single modality and do not support big data. Recently, the deep learning techniques provided better outcomes than the machine learning techniques for multimodal event detection. Also, the deep learning approaches are highly applicable for multi-modality and big data applications. The existing deep learning models bring better results only using text and image-based event detection. Moreover, deep learning with big data analytics provides improved classification accuracy. It inspires the author to design a deep learning framework for multimodal event detection (text, image and audio) in big data analysis.

The major objectives of the proposed multimodal event detection are:

- Introduced multimodal event detection using the new MLFC to accurately classify different types of events.
- The proposed MLFC model uses images, texts, and audio data to create a multi-level, fine-tuned DL framework that reduces the effect of low-quality data regarding image quality and labelling.
- Improving the event detection performance using the Deep FF strategy and SoftMax classification.
- To develop and execute the proposed deep learning model and evaluate the performance in terms of accuracy, recall, F1-score and precision.
- To evaluate the suggested approach's overall performance and prove its efficiency by comparing the obtained results with existing approaches.

The proposed research is structured into several sections. The literature review of recent research works regarding event detection is described in section 2. In section 3, the proposed work with different methodologies of multimodal event detection is discussed. Section 4 represents the results and performance analysis of the proposed classification model based on deep learning. Section 5 concludes the proposed work, followed by feasible future scope and references.

2. Related Works

Most researchers have implemented several works to detect the events based on various methodologies. Some of the leading event detection models adopted by the researchers are surveyed as follows.

[Sun et al. (2017)] developed a hybrid mechanism for abnormal event detection in video sequences using a one-class support vector machine (SVM) and a convolutional neural network (CNN). The features are powerfully captured using the CNN model by understanding the underlying large dimensional normal representations. The one class SVM categorizes the normal and abnormal classes, and also the parameters of the entire model are optimized by using the one-class SVM. This SVM model generates an accurate optimal solution for abnormal event detection. Also, the suggested model minimizes the cumbersome intermediate operation more effectively than the other techniques.

[Kidziński et al. (2019)] presented a DNN (deep neural network) for automatically detecting real-time gait events in children. This study utilizes a data-driven model for foot-contact prediction and foot-off events from marker and kinematic time series in children with a pathological and normal gait. Long short-term memory (LSTM) with artificial neural network (ANN) is introduced for the predictive model, and it detects the foot-off and foot contact events. In this, three methods, machine learning-based approach, coordinate-based, and velocity-based approach, are combined to detect the gait events. The developed model minimizes the costs and attains highly precise data processing. In this existing work, the foot-off detection is not as much better than the foot-contact detection. Thus, it becomes the disadvantage of this suggested model.

[Singhal et al. (2019)] introduce a Spot Fake multimodal approach model for detecting fake news in social media. This method cannot consider any of the subtasks, and it accurately detects the presented fake news from the input samples. The Spot Fake involves three modules: texture feature extractor, visual feature extractor, and multimodal fusion module. In the third module, the features obtained from the texture and visual feature extractor are combined using a concatenation method to attain the appropriate news representation. This news representation is fed via a fully connected neural network to classify fake news. This existing model is superior to the recent Weibo and Twitter datasets techniques for fake news detection.

[Tippaya et al. (2017)] developed a shot boundary detection (SBD) to determine the discontinuity signal behaviours of visual representation. The SBD method is designed to attain a perfect shot boundary detection in

which the process of detection directly determines the behaviour of transition. The entire shot boundaries are separated into two types such as gradual and cut. The feasible shot boundaries are gathered by performing candidate segment selection. The suggested SBD approach provides improved performance than the other methods in terms of shot boundary detection. This existing work also notices the inter-frame distance depending on the developed visual features. This inter-frame parameter disrupts the gradual shot detection's performance also, detection speed is increased in this model. Thus, an advanced technique is needed to detect the shot boundary detection with improved detection speed.

[Roy et al. (2018)] reported an event detection for various PMUs performed by utilizing big data analysis. The analysis has been performed depending on the time-stamped realistic time data attained from the TTU PMU network system. R programming approach is utilized to find the duration and disturbances utilizing the fast detection method of the R library function. A fast algorithm is developed and verified for event identification in this existing work. This detection approach is suitable for all small events from the dataset. This analysis exhibits the relationship between violations in a grid and considers the effects of voltage and frequency data. Table 1 describes the review of recent event detection research works with respective techniques and contributions.

Table 1. Review of various event detection techniques

Author name and Reference	Techniques used	Dataset used	Objective	Merits	Demerits
[Sun et al. (2017)]	Hybrid combination of One class SVM and CNN	UCSD pedestrian	To recognize abnormal event detection in a video sequence.	Effective minimization of cumbersome intermediate operation.	Classification error is high.
[Kidziński et al. (2019)]	LSTM with ANN	Gillette Children's Specialty Healthcare Center	To detect the foot-off and foot contact events through the combination of three methods.	Cost minimization and accurate processing of data.	Foot-off detection efficiency is not better.
[Singhal et al. (2019)]	Spot Fake multimodal approach	Twitter and Weibo	To efficiently detect fake news in social media.	The probability of error occurrence is low.	Consumption of time is more.
[Tippaya et al. (2017)]	SBD	Golf video clips and TREC2001	To analyze the discontinuity signal behaviours of visual representation.	Improved performance can be attained.	Performance of gradual shot detection is disrupted, and detection speed is increased.
[Roy et al. (2018)]	TTU PMU network system.	Real-time data	To detect the events by utilizing big data analysis.	Highly suitable for small events.	Difficult to extract the optimal features.

Detection of an event is the problem of automatically finding particular incidents by analyzing some data. The event detection method helps determine whether the event occurred or not. Many existing techniques fail to detect the event with better classification accuracy because of the large size of data, worst transfer learning ability, hard to extract features etc. Also, some of the previous techniques require more time to detect the event and make the detection speed minimum. Moreover, the machine learning techniques fail to perform in large size data and generate more classification errors that disrupt the system's performance. To overcome these issues, the proposed work utilizes deep learning approaches for multimodal event detection with the help of big data analytics.

3. Proposed Methodology

Detection of events possesses a significant role in the modern era, especially in a huge quantity of real-life data targeting various events held on social media platforms. Accurate detection of events using various data, submerging error, and increasing detection efficiency are challenging scenarios in recent days. Hence, an efficient deep learning-based multimodal event detection approach is proposed in this research work for better classification outcomes. Initially, the multimodal data needed to carry out the proposed event detection model is gathered, consisting of text, image and audio based real-world actions for identifying the various events. The different steps involved in accurate multimodal event detection are mentioned as follows.

- Data acquisition
- Multimodel Generation

- Feature Fusion
- Event Classification

The gathered multimodal data, including text, images and audio, are accumulated in the Hadoop platform to store huge sized data and entire data processing in a distributed manner. The multimodal attributes are fed as the input to the proposed MLFC approach. In a multimodal generation, the text models are created through pre-processing, feature extraction and an *Enhanced TF-IDF* system. The image model is generated through I-Caps Net. The acoustic model is generated by feature extraction and the CNN_OSA approach. The extracted features from the multimodel are fused using the Deep FF strategy. The different events are classified through the SoftMax classifier.

The overall schematic representation of the proposed event detection framework is illustrated in Figure 1.

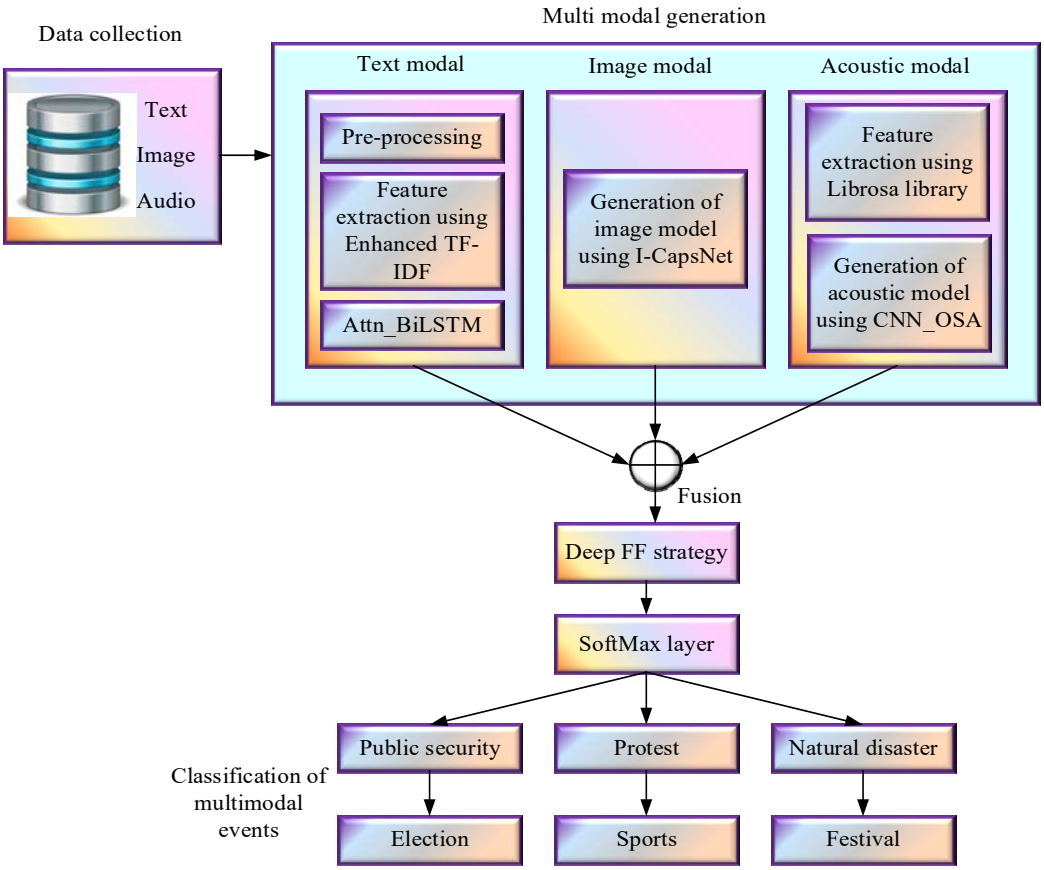

Fig. 1. Overall architecture of the proposed system

3.1. *Multimodal generation*

The multiple models are generated to text, image and audio for the accurate classification of different events based upon the acquired data. The description to generate each modal is mentioned as follows.

3.1.1. Text modal generation

After gathering text input, the generation of the text modal is carried out through the steps including pre-processing, textual feature extraction, and text modal creation.

Pre-processing: The text inputs are pre-processed initially by undergoing the several processes depicted in Figure 2.

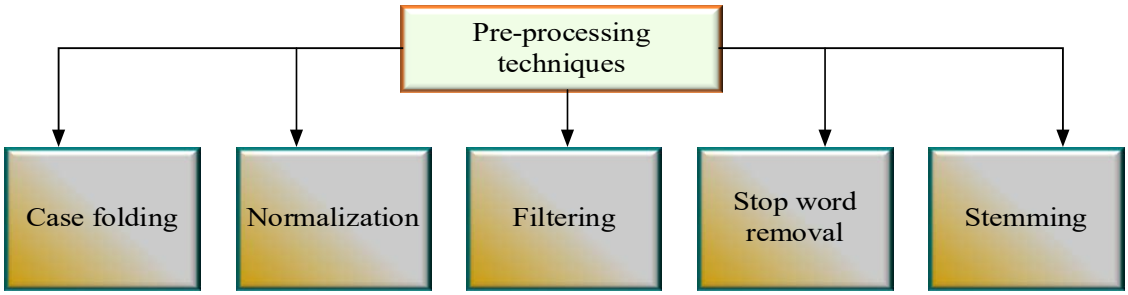

Fig. 2. Pre-processing techniques

Case folding is used as the initial process is used to convert the letters from the text documents to the respective lower or upper case. In text pre-processing, case-folding was adopted to convert letters to lowercase format. The next process of pre-processing is text normalization, in which the text transformation is carried out in a standard format through the removal of punctuation and lemmatization. For instance, the words 'goooood' and 'gud' can be transformed into good which denotes the standard form. The punctuation removal was done to neglect the unwanted information and consider only the valuable features.

The filtering process is done through tokenizing, where the sentences are separated into tokens through space detection. Text filtering is a data recognizing process in which the sentences are chosen from a dynamic text stream to gather the relevant information. For instance, the email addresses can be detected by considering the symbol'@'. After text filtering, stop word removal is the next step undertaken. By adopting the Natural language toolkit, the process of stop word removal can be easily carried out. The process of eradicating the words across all the text documents in common is called stop word removal. Generally, the texts holding pronouns and articles are considered stop words that are removed.

The final process of pre-processing is stemming, which helps reduce a word to its corresponding word stem that affixes to the roots of words. In the process of information retrieval, stemming possess two significant functions. The first function is to enhance the capability of the information retrieval system to choose the relevant data, and the next function minimizes the vocabulary size through variant mapping based on root words.

Textual feature extraction: The valuable textual features are extracted from the pre-processed text by adopting the Enhanced TF-IDF methodology. The original value of TF is used directly with the assumption that a high TF value is more significant when compared to a low TF value. Due to the ignorance of collection frequency, the capacity of TF is too low, and so IDF is employed together to address the issue.

Even though when they are combined, there are some limitations like imbalanced text documents, size variations among different categories, and larger values of IDF in case of low document frequency compared to others. These forms of issues are addressed by considering deviated IDF value in the process of term weighting using the Enhanced TF-IDF approach.

The collection frequency factor is modified by adding ADF (Average Document Frequency), the variance between the DF value and the average values of all DF. The average of all DF values can be expressed as,

$$DF_{avg} = \frac{\sum DF(V,T)}{m} \tag{1}$$

$$A_{DF}(V,T) = \frac{\left(DF(V,T) - DF_{avg}\right)^2}{m} \tag{2}$$

Where the average of all DF values is represented as DF_{avg}, the ADF value of a term V in a text document T is denoted as $A_{DF}(V,T)$, and the number of terms is given by m. The ADF is an extended form of DF, and it paves the simplest way in optimizing IDF. The enhanced formula of collection frequency is represented as follows.

$$IADF(V,T) = \log \frac{|T|+1}{A_{DF}(V,T)+1} \tag{3}$$

The ADF is utilized for minimizing the term weights with very high or low DF values.

$$IADF'(V,T) = \log \frac{|T|+1}{DF(V,T)+1} * \frac{1}{\log(A_{DF}(V,T)+1)+1} \tag{4}$$

To overcome the variance tending to be too small or large, the above equation is optimized by ADF normalization to minimize the effect generated by the extreme term values. The value of $A_{DF}(V,T)$ is modified and can be expressed as,

$$A_{DF}'(V,T)=\log\frac{1}{\left(A_{DF}(V,T)+1\right)}+1 \tag{5}$$

Through the adoption of the normalization formula, the expression can be given as,

$$A_{DF}'(V,T)=\frac{A_{DF}'(V,T)-\min\left(A_{DF}'(V,T)\right)}{\max\left(A_{DF}'(V,T)+1\right))-\min\left(A_{DF}'(V,T)\right)} \tag{6}$$

On this basis of A_{DF}'', two novel formulas are generated,

$$IADF_n(V,T)=\log\frac{|T|+1}{A_{DF}''(V,T)+1} \tag{7}$$

$$IADF_n'(V,T)=\log\frac{|T|+1}{DF(V,T)+1}*\left(A_{DF}''(V,T)*\delta\right) \tag{8}$$

where the default value of δ is one and is used as the optional weight proportion to adjust the significance of A_{DF}'' in various cases. Dependent upon the four proposed formulas of collection frequency, four novel term weighting representations are given as,

$$TF-IADF(V,t,T)=TF(V,t)*IADF(V,T) \tag{9}$$

$$TF-IADF'(V,t,T)=TF(V,t)*IADF'(V,T) \tag{10}$$

$$TF-IADF_n(V,t,T)=TF(V,t)*IADF_n(V,T) \tag{11}$$

$$TF-IADF_n'(V,t,T)=TF(V,t)*IADF_n'(V,T) \tag{12}$$

The valuable textual features are effectively extracted through an enhanced TF-IDF approach, whereas the unbalanced text collection is enhanced. The extracted features are fed into the Attention-based BiLSTM (Attn_BiLSTM) Network for text model generation. Normally, the Attn_BiLSTM model includes an attention mechanism over the top of the BiLSTM layer to apply attention weighting. The purpose of the attention mechanism is to distinguish every text's weight and enable the entire sequence to obtain valuable information very easily. The structure of the Attn_BiLSTM framework is illustrated in Figure 3.

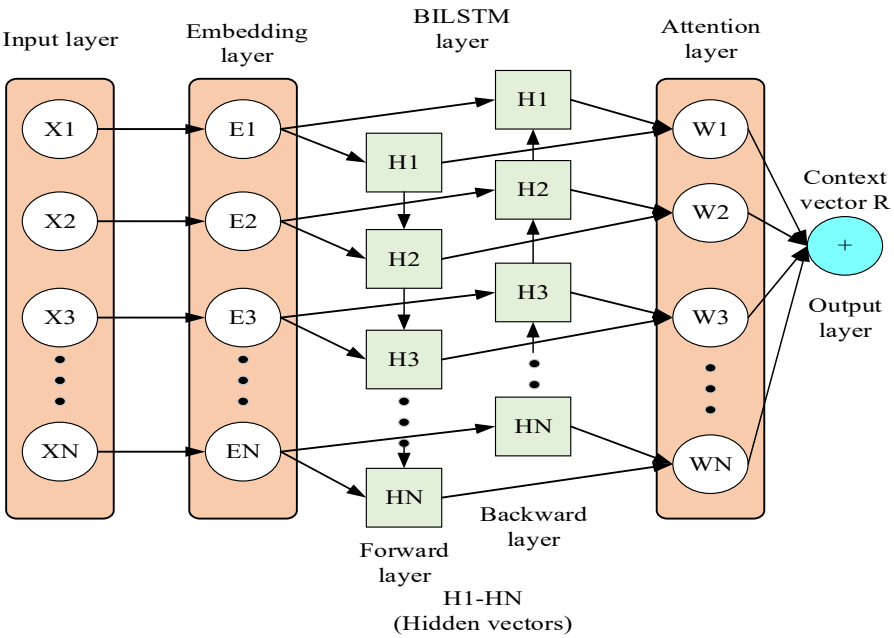

Fig. 3. Framework of Attn_BiLSTM

The Attn_BiLSTM approach consists of five layers: the input layer, embedding layer, Bi-LSTM layer, attention layer, and output layer. In this model, the encoder acts as an Attn_BiLSTM, and the decoder has a BiLSTM. The input layer acquires the input and is fed to the embedding layer, which maps every text in a low latitude space. The BiLSTM utilizes a bidirectional LTSM for the advanced feature extraction process and generates a weight vector. BILSTM layer generates the sum up information of both the forward and backward layer. An attention layer is used to create sentence-level features by combining weight vectors and lexical features. Finally, an image modelling vector is a generator in the output layer.

An attention mechanism is used in the efficient generation of texts, whereas a context vector R is trained over a text sequence N . Based on the annotation sequence, the context vector is dependent on which the encoder maps the input sequence. Every annotation holds the information regarding the entire input sequence. The context vector is evaluated through the weighted sum of these annotations as,

$$C_u = \sum_{v=1}^{N} \beta_{uv} h_v \qquad (13)$$

From the above equation, β_{uv} denotes the weight and h_v denotes the hidden vector. The probability of generating a more precise text outcome can be maximized and enhance the correlation between the summarized and source text. Thus, an efficient generation of the text model is effective done through these processes.

3.1.2. Image modal generation

The capsule networks (CapsNet) possess the new classifier generation with more advantages, including high robustness and efficient detection of overlapping images. CapsNet is equivalent to the Convolutional Neural Network (CNN), which aims to extract low-level features. The upcoming layers are varied in CapsNet as the neurons are grouped into vectors called capsules. Even though CapsNet performs better in most criteria, it also holds certain drawbacks like computationally expensive, preferable only over small-scale datasets and deterioration in accuracy. Hence an Improved CapsNet (I-CapsNet) is proposed for the efficient generation of image modal.

Here Convolutional Fully Connected (CFC) layer is used as an improved architecture to the conventional CapsNet. I-CapsNet is a more effective network in which training and testing can be established rapidly, whereas slightly increased accuracy can be obtained compared to the conventional CapsNet. I-CapsNet comprises fewer parameters (Training weights), and hence it is highly efficient for memory utilization. The I-CapsNet is entirely different from the conventional CapsNet as it performs an optimized solution of the feature extraction process. The architecture of I-CapsNet for extracting the features in image modal generation is depicted in Figure 4.

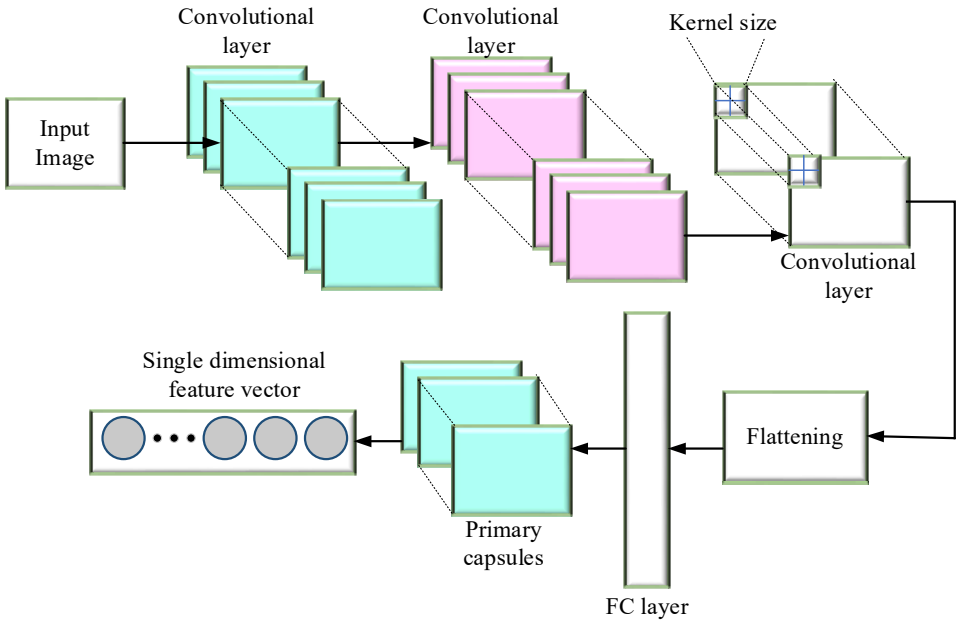

Fig. 4. Architecture of I-CapsNet

Two convolutional layers are adopted in the CapsNet to extract features. These features are generated from the outcome activation reformed to the initial set of samples referred to as primary capsules. The I-CapsNet includes a new layer called CFC that is accountable for representing the output activation into vectors. The CFC layer is added after the second convolutional block of CapsNet, which means the reshaping operation is replaced with this layer. Rather than the reshaping operation, a limited number of primary capsules are generated by the CFC layer with fewer parameters and provide a faster network.

The convolutional layer shares the equivalent weights over various regions of the input. The process of weight sharing using a single kernel helps minimize the number of parameters. The CFC layer possesses a similar functionality to the convolutional layers, but it does not undergo weight sharing. This layer is responsible for minimizing the spatial size related to the input that is similar to the convolutional layer, and

every element in the output feature map is generated through the utilization of a separate kernel. It is considered to integrate the convolutional and fully connected (FC) layer.

The figure clearly states that every element in a spatially correlated section of the activation function is flattened. These outcomes are fed into the FC layer, which creates a single vector or capsule. Two hyperparameters present in the CFC layer are given,

- The capsule dimensionality indicates the number of neurons present in the output of every FC layer.
- Kernel size is equivalent to the size of convolutional layers.

The FC layer summarizes the vectors from the feature maps, and the CFC layer transforms the extracted features into primary capsules. The advantages of I-CapsNet are minimization in the number of parameters and increased generalization capability. I-CapsNet is considerably faster than the CapsNet as a limited number of primary capsules are generated.

3.1.3. Audio model generation

The audio features are extracted through the Librosa library in audio model generation. The Librosa library is used to downsample the original audio files to extract valuable acoustic features. Different types of audio features are extracted as high-level, mid-level and low-level features. The low-level features, including amplitude envelope, energy and zero-crossing rate, are extracted. The High-level features, including pitch, beat related descriptors, Mel Frequency Cepstral Coefficients (MFCC) and mid-level features like melody, are extracted.

These extracted features are fed into the CNN_OSA (Convolutional Neural Network _Opposition Salp swarm Algorithm). A typical CNN comprises single or numerous convolutional blocks and subsampling layers. The convolutional layers follow the fully connected layers and output layer. Figure 5 illustrates the basic block diagram of the CNN_OSA structure.

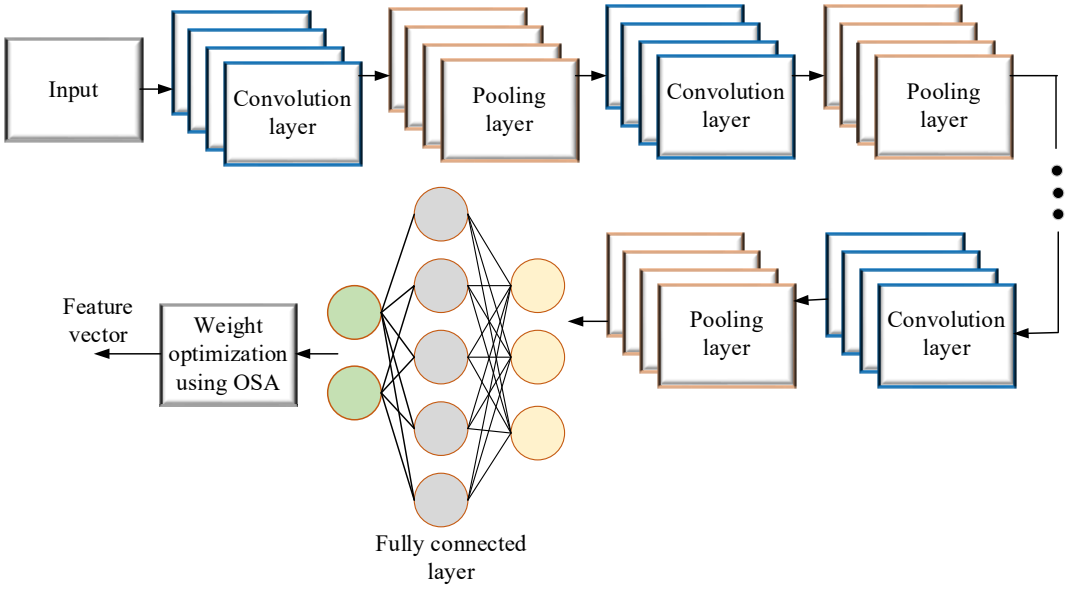

Fig. 5. Structure of CNN_OSA

Convolutional layer: The CNN layer is the central part of the overall CNN structure, and the audio features are stationary in nature. It indicates the formation of one section of audio is similar to the other section. Hence, a feature learned in one section can match an equivalent pattern in another region. A small part is considered and passed as the input from the larger section of features. These features are convolved into a single position of output. A compressed form of signal is attained, and the filtered form is transferred to the next layer.

Sub-Sampling or pooling layer: Pooling is the process of downsampling the signal, and it considers the small portion of convolutional output as the input that sub-samples to generate a single output. The pooling layer minimizes the size of the audio features gradually in order to decrease the number of parameters, computational complexity and to maintain the overfitting problems. The pooling layer compresses the features available in the region produced by a convolution layer, and also it controls the resolution of features to improve the steadiness.

Fully connected layer: The final section of CNN is the fully connected layer that captures input from all the neurons in the previous layer and undergoes processing with every neuron in the present layer to establish the output. The fully connected layer is an essential kind of feed-forward neural network, and the output of each section is directed to the energized unit of the next layer. An output feature vector is generated from the fully

connected layer, but there are chances of the degradation of audio generation accuracy due to the presence of cross-entropy loss.

Hence, the weights are updated to optimize the loss function through the OSA approach to enhance accuracy. OSA introduces opposition-based learning in the Salp Swarm algorithm (SSA). The concept of opposition-based learning is dependent on the opposite numbers, and if the upper, lower bound is defined by U and L, then the opposite form of a real number R is represented as,

$$O = U + L - R \tag{14}$$

The opposite form of R is represented as O that increases the convergence speed. Based on the objective function, the best audio feature vector is selected. The audio features are updated through the expressions of leaders and followers in SSA, represented as follows.

$$Z_v^1 = \begin{cases} F_v + A_1((UB_v - LB_v)A_2 + LB_v) & A3 \geq 0 \\ F_v - A_1((UB_v - LB_v)A_2 + LB_v) & A3 < 0 \end{cases} \tag{15}$$

$$Z_v^i = (Z_v^i + Z_v^{i-1})/2 \tag{16}$$

The above expressions, UB_v and LB_v denotes the upper and lower bound of j^{th} dimension, respectively. A_2 and A_3 are the uniform random numbers in the range [0,1]. Z_v^i and Z_v^1 describes the leader salp position and i^{th} follower salp in the v^{th} dimension. F_v denotes the food source position in the j^{th} dimension. The value of A_1 is evaluated by,

$$A_1 = 2e^{-(4l/L)^2} \tag{17}$$

where, the current iteration number is denoted as l and the maximum number of iterations is denoted as L. An effective audio feature vector is created in the audio model generation.

3.2. *Deep feature fusion strategy*

The process of deep feature fusion considers the detailed information and considers highly semantic information to enhance the performance of an event detection process. The textual features are denoted as $T = \{t_1, t_2, ...t_n\}$, the image features are denoted as $I = \{i_1, i_2, ...i_n\}$ and the audio features are represented as $A = \{a_1, a_2, ...a_n\}$. These extracted features from the multimodal generation are fused to form a new vector that can be represented as $V = \{v_1, v_2, ...v_n\}$. Figure 6 represents the structure of the deep feature fusion strategy.

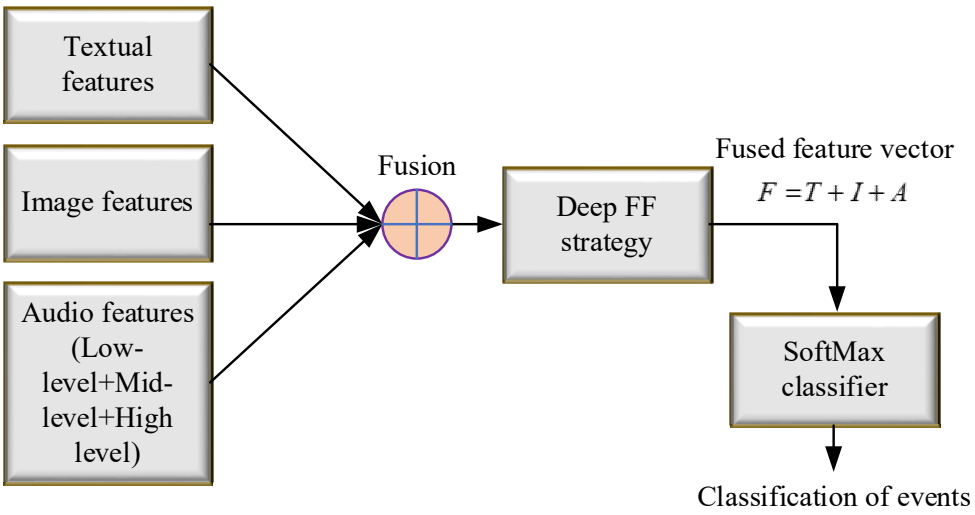

Fig. 6. Deep feature fusion strategy

The following equation expresses the fused features,

$$F = T + I + A \tag{18}$$

From the above equation, $A = L + M + H$ where, $L = \{l_1, l_2, ...l_n\}$, $M = \{m_1, m_2, ...m_n\}$ and $H = \{h_1, h_2, ...h_n\}$. L, M, H represents the low level, mid-level and high-level features of audio. The deep feature fusion strategy merges the feature vectors of multimodal generation output into a single vector. The feature vector is pooled on the pooling layer that compresses the input vector. The complexity of the proposed

approach is highly minimized, and valuable outcomes are attained. The generated feature vector is denoted as $G=\{g_1, g_2, ...g_n\}$ where, G can be computed by,

$$G_r = \max_{s=1}^{k}(F_r), r = 1,2,....n \quad (19)$$

where, k resembles the size of the pooling region.

3.3. *Event classification*

The different text, image and audio events are precisely classified in the proposed approach, dependent on the fused feature vector. The fused features obtained from the deep feature fusion strategy are fed into the SoftMax classifier. The SoftMax classifier renders the probabilities of each class label, and the different events like public security, protest, natural disaster, election, sports and festivals can be classified precisely. By implementing this proposed event detection model, the classification accuracy is highly improved, the complexities are minimized with minimum error, and the convergence is higher. The overall performance of the event detection model is enhanced due to the consideration of a limited number of features.

4. Results and discussion

The experimental outcomes of the proposed deep learning-based event detection model are described in this section. The performances of the proposed work are evaluated through the utilization of the PYTHON simulation tool. In order to estimate the proposed performance of event detection, several existing methodologies are compared. The description of data, explanation of different performance metrics, analysis and comparison, are provided in the sub-sections.

4.1. *Dataset description*

The data used for analyzing the performance of multimodal event classification through the MLFC classifier model is gathered from Multi-domain and Multi-modality Event Dataset (MMED). The performances are assessed by separating the data set into 80% for training and 20% for testing. Here 25,165 textual articles of news were acquired from hundreds of data sources from the online domain of news media involving yahoo, Google, NBC, New York Times, Fox, NBC news and so on. A count of 76,516 Flickr image posts is collected, which are shared by 4,473 social media users of Flickr. The audio samples are collected from the social sites to generate the whole database as multimodal data. The multimodal data is fed as the input to the proposed classifier model for effectively classifying different real-world events, including public security, protest, natural disaster, election, sports and festivals.

4.2. *Performance metrics*

In order to estimate the performance of the proposed event detection model, various performance metrics, including Accuracy, Precision, Recall, F1 score and AUC, are considered. The description of the various metric is explained with its mathematical expressions as follows.

Accuracy: The entire number of correct predictions divided by the overall count of predictions is termed the metric called accuracy. The mathematical expression for accuracy is described as,

$$Acc = \frac{G+H}{G+H+I+J} \quad (20)$$

From the above equation, G signifies true positive, H denotes true negative, I means false positive and J symbolizes false negative.

Precision: The availability of predicted positives that are certainly positive is called precision. It is also termed a Positive predictive value (PPV). The precision can be represented as,

$$P = \frac{G}{G+I} \quad (21)$$

Recall: The count of positive outcomes over the total count of truly positive samples is also termed sensitivity. The recall can be mathematically denoted as,

$$R = \frac{G}{G+J} \quad (22)$$

F1 score: The harmonic means of PPV or precision and True positive rate (TPR) or recall is called to be F1 score. It can be described as,

$$F1\ score = 2\ \frac{PPV \times TPR}{PPV + TPR} \quad (23)$$

AUC: AUC determines the capability of the technique to differentiate between the aimed classes. It is also called to be the area under the receiver operating curve. The performance of AUC is assessed by mapping the graph for true positive rate (TPR) over False positive rate (FPR). It can be represented as,

$$AUC = \frac{Sensitivity + Specificity}{2} \quad (24)$$

4.3. *Performance analysis and comparison*

The proposed MLFC classifier model is compared with certain existing approaches like Recurrent neural network (RNN), Convolutional neural network (CNN), Deep Convolutional neural network (DCNN) and Deep belief network (DBN) for analyzing the performance of multimodal event detection. The performance outcomes of the proposed model are evaluated in terms of accuracy, precision, recall and F1 score. Implementing the proposed model using the PYTHON simulation tool has attained better results in classification accuracy. The better capability of transfer learning can be attained with minimal time. Through this research, the classification error is highly minimized. Table 2 represents the performance analysis of proposed various multimodal events.

Table 2. Performance analysis of proposed multimodal events

Performance	**Multimodal events**					
	Public security	Protest	Natural disaster	Election	Sports	Festival
Accuracy	98.99	98.15	97.98	98.15	98.48	97.81
Precision	100	95.57	92.15	95.07	92.30	88.46
Recall	95.27	94.73	95.91	97.12	93.75	86.79
F1 score	97.58	95.15	94.00	96.08	93.02	87.61

From the above table, it can be clearly analyzed that the proposed MLFC classifier model has attained better results in terms of accuracy, precision, recall and F1 score. The six various events, including public security, protest, natural disaster, election, sports and festivals, are effectively classified. The accuracy of 98.99% is obtained in classifying the public security events, 98.15% in protest events, 97.98% in natural disaster events, 98.15% in election events, 98.48% in sports events and 97.81% in case of festival events. The higher accuracy denotes the effectiveness of the overall system. Higher precision is attained in classifying the protest events, and higher recall performance is obtained in classifying natural disaster events. Table 3 describes the performance comparison of proposed and existing techniques.

Table 3. Comparison of proposed and existing techniques

Techniques	**Accuracy**	**Recall**	**Precision**	**F1 score**	**AUC**
CNN	97.64	91.92	91.86	91.89	79.59
RNN	97.08	90.14	90.24	90.19	84.09
DCNN	96.24	87.63	87.52	87.58	93.09
DBN	95.63	85.60	85.65	85.63	87.59
Proposed	98.26	98.93	97.92	98.42	95.59

An obvious analysis can be made by comparing proposed and existing approaches that the proposed MLFC classifier model has obtained better performance in terms of accuracy, precision, recall, F1 score and AUC. The overall accuracy obtained in classifying the multimodal events is 98.26%, whereas the existing methods like RNN, CNN, DCNN and DBN have obtained 97.64%, 97.08%, 96.24% and 95.63% in the classification process. On comparing the accuracy performance between proposed and existing models, the proposed model outperforms with a better accuracy rate. Comparatively, the performance of precision, recall, F1 score and AUC tends to be lower than the proposed model. Figure 7 describes the graphical representation of performance analysis in terms of accuracy, precision, recall and F1 score for various events, including public security, protest, natural disaster, election, sports and festivals.

The graphical representation of accuracy, precision, recall and F1 score obtained for the models of multimodal events expose the performance outcomes. From the figure, it is identified that the accuracy of the proposed model in classifying the public security event is 98.99%, 98.15% for protest events, 97.98% for natural disasters, 98.15% for election events, 98.48% for sports events and 97.81% of festival events which tends to be more accurate. This is achieved because of the effective utilization of the multimodal generation models. The multimodal event data is provided as the input to the classifier that the proposed model well adapts, and the

classifications are made more accurate. The proposed model identifies the most relevant data samples and merges them for effective classification. The precision performance of the proposed model in terms of various events like public security, protest, natural disaster, election, sports and festivals are found to be 100%, 95.57%, 92.15%, 95.07%, 92.30% and 88.46%. In contrast, the Recall performance is found to be 95.27%, 94.73%, 95.91%, 97.12%, 93.75% and 86.79%. The F1 score performance in terms of multimodal events has obtained 97.58%, 95.15%, 94.00%, 96.08%, 93.02% and 87.61%. greater performance is attained due to the utilization of redundant features.

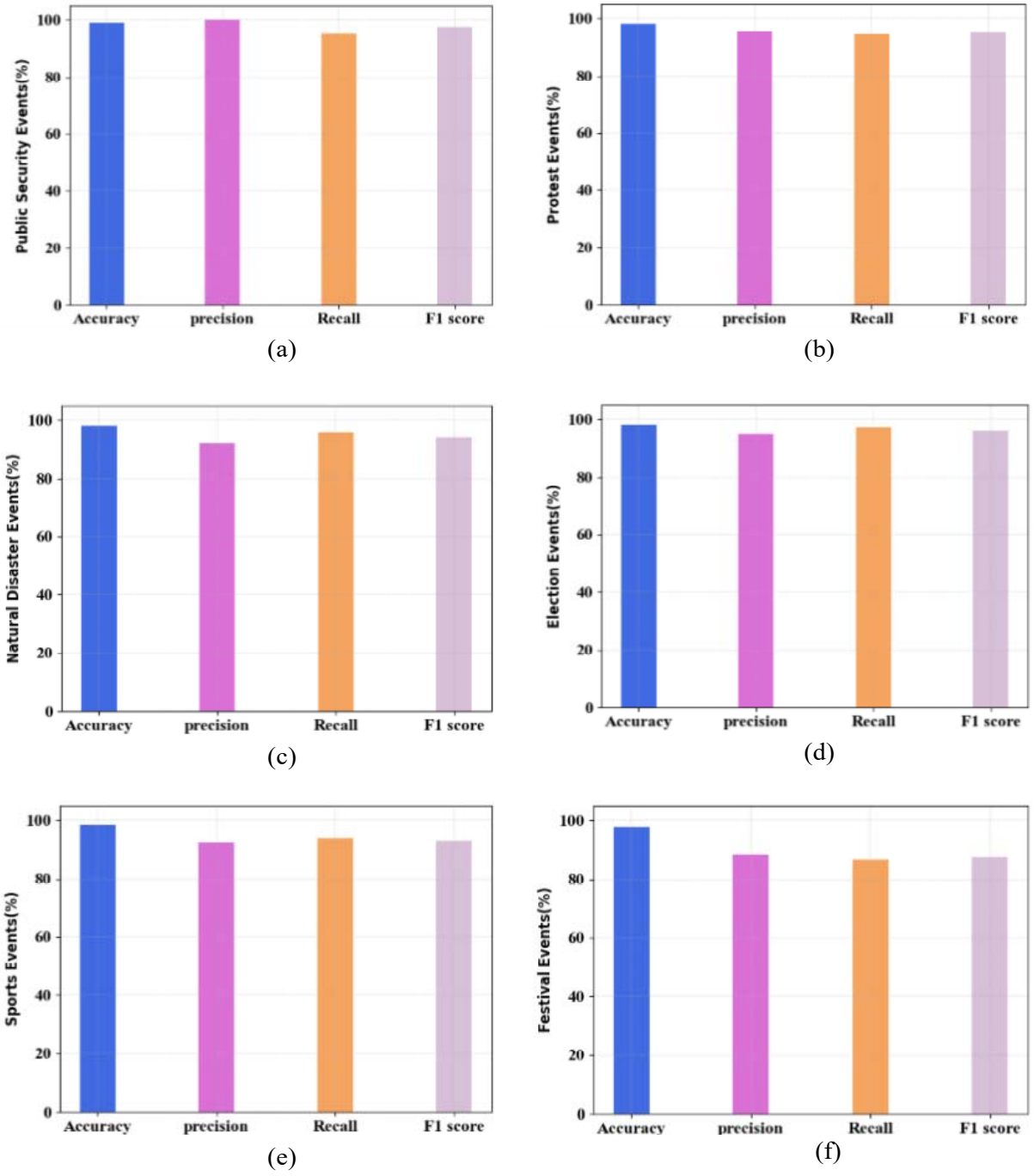

Fig. 7. Performance of Accuracy, precision, Recall and F1 score for different events (a) Public security (b) Protest (c)Natural disaster (d) Election (e) Sports (f) Festival

Figure 8 represents the performance comparison of proposed and existing techniques in terms of accuracy. It can be clearly observed that the proposed MLFC classifier model obtains better accuracy in multimodal event classification when compared to the existing approaches. The overall classification accuracy for the proposed

model is 98.26%, while the existing methods like RNN, CNN, DCNN and DBN have achieved 97.08%, 97.64%, 96.24%, 95.63% of accuracy. The existing methods obtained only less rate of accuracy due to a larger accumulation of datasets and increased classification error. Hence, the proposed method holds more tendency to accurately classify multimodal events.

The graphical representation of precision in terms of proposed and existing approaches is described in Figure 9. Precision is one of the significant aspects to be considered for gathering the effectiveness of outcomes. The proposed MLFC classifier model has achieved 97.92% precision and shown a better outcome in minimizing the false detection rate. While the existing classifier methods like CNN, RNN, DCNN and DBN have obtained lower results as 91.86%, 90.24%, 87.52 and 85.65%. Finally, from the figure, it can be analyzed that the proposed method outperforms better when compared to the existing approaches.

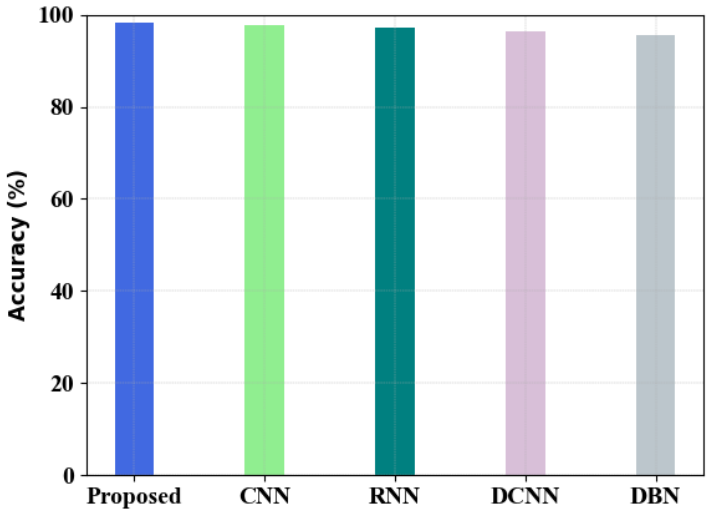

Fig. 8. Accuracy performance comparison

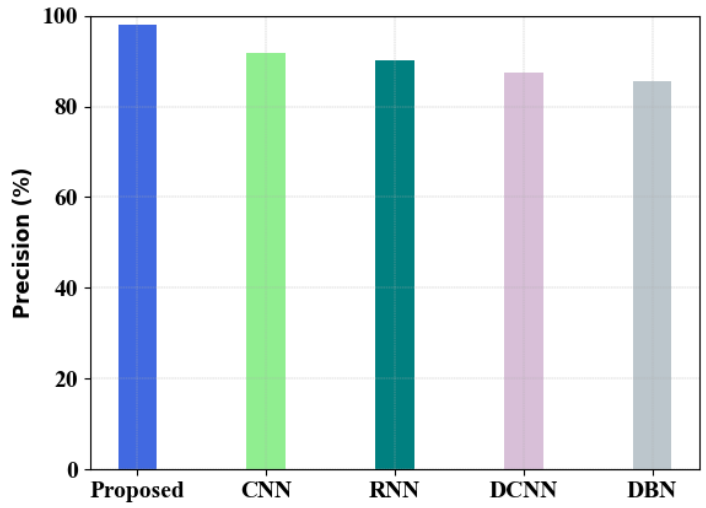

Fig. 9. Precision performance comparison

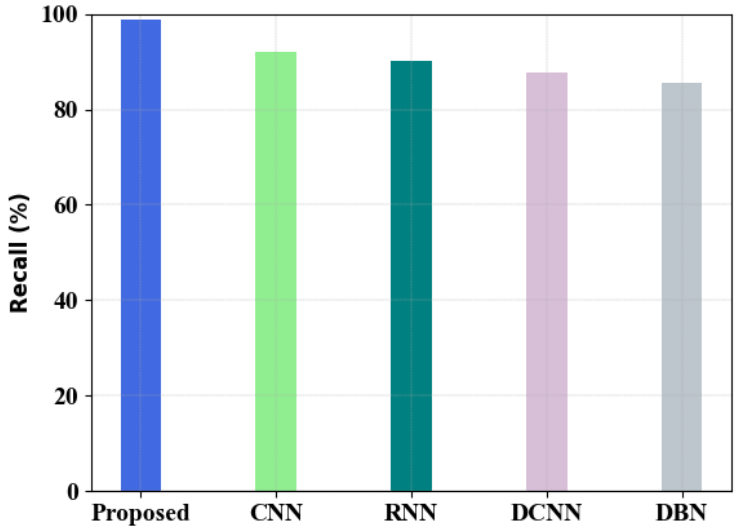

Fig. 10. Recall performance comparison

Figure 10 illustrates the graphical representation of recall with respect to proposed and existing approaches. 98.93% of recall is obtained while assessing the performance of the proposed method in contrast to the existing approaches. The existing classifier approaches in terms of the recall have attained 91.92%, 90.14%, 87.63% and 85.60% with respect to CNN, RNN, DCNN and DBN. Due to increased complexities of time and storage, the existing classifier models tend to offer lower performance when compared to the proposed method. Hence, a clear justification can be made from the figure that the proposed MLFC model performs better in classifying multimodal events.

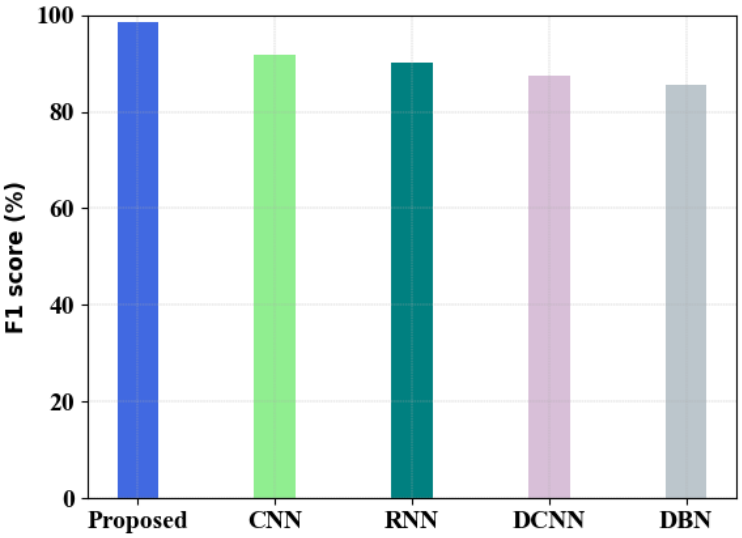

Fig. 11. F1 score performance comparison

The graphical representation of the F1 score with respect to proposed and existing techniques is depicted in Figure 11. It is obvious that the proposed model holds more capability to classify the multimodal events like public security, protest, natural disaster, election, sports, and festivals based on the input parameters than the existing techniques. The value of the F1 measure for the proposed MLFC model is 98.42%, whereas the F1 score values of existing models like RNN have obtained 90.19%, CNN as 91.89%, and DCNN as 87.58% and DBN as 85.63% in multimodal event classification. Hence, the F1 score shows better results in the classification process in the proposed method.

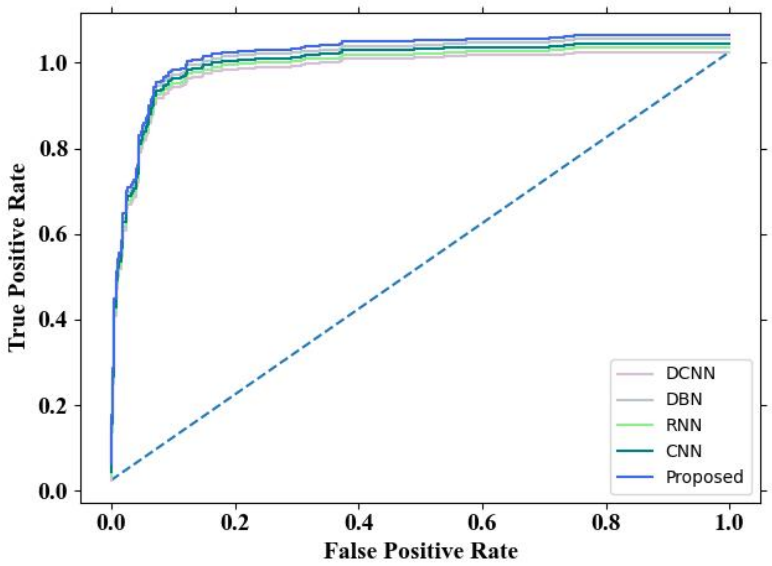

Fig. 12. Comparison of AUC analysis

The graphical representation of AUC comparison in terms of proposed and existing approaches is depicted in Figure 12. The proposed method holds the better capability to differentiate between the target classes. The graph has plotted between FPR and TPR to determine the AUC value. The proposed method has attained the AUC value of 95.59%, superior to the existing approaches like RNN, CNN, DCNN and DBN, which have obtained 84.09%, 79.59%, 93.09% and 87.59% of AUC values. The proposed model has enhanced the ability in multimodal event classification depending on the input parameters.

5. Conclusion

The precise analysis of multimodal events possesses a significant requirement in recent days to gather valuable information. An effective way to fulfil these requirements is to employ a prominent event detection model. Hence, the deep learning-based multimodal event detection using the Multi-Level Fusion Classifier (MLFC) model is proposed. Different processes were involved in the proposed approach, such as Data collection, Multi-model Generation, Feature Fusion and Event Classification. The big data are collected from the multimodal data, including text, images, and audio, located in the Hadoop platform for storage purposes. The gathered data from different models are directed to the MLFC approach to generate text, image and audio models. The steps involved in generating text are pre-processing, IF-IDF and Attn_BiLSTM. I-CapsNet is utilized to generate the image model. The audio features with respect to low-level, mid-level and high-level are extracted that are directed to CNN_OSA for generating the audio model. The extracted features from multimodal generation are fused through the Deep FF strategy. The different events like public security, protest, natural disaster, election, sports and festivals are classified through the SoftMax classifier. The overall classification accuracy is 98.26% by the proposed MLFC classifier model. The performances of the proposed approach are analyzed through the PYTHON simulation tool. In the future, the proposed work can be extended by investigating the interaction between visual paths and audio at different levels of the classification framework.

Acknowledgments

None

Conflicts of interest

The authors have no conflicts of interest to declare.

Funding

No funders were contributed for the preparation of the manuscript.

References

[1] Alomari, E.; Katib, I.; Mehmood, R. (2020). Iktishaf: A Big Data Road-Traffic Event Detection Tool Using Twitter and Spark Machine Learning. Mobile Networks and Applications, pp. 1-16.
[2] Cakir, E.; Parascandolo, G.; Heittola, T.; Huttunen, H.; Virtanen, T. (2017). Convolutional Recurrent Neural networks for Polyphonic Sound Event Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, **25**(6), pp. 1291-1303.
[3] Cao, J.; Xia, R.; Guo, Y.; Ma, Z. (2018). Collusion-Aware Detection of Review Spammers in Location Based Social Networks. World Wide Web, **22**(6), pp. 2921-2951.

[4] Chen, Q.; Wang, W.; Huang, K.; De, S.; Coenen, F. (2021). Multi-Modal Generative Adversarial Networks for Traffic Event Detection in Smart Cities. Expert Systems with Applications, **177**, pp. 114939.
[5] Chen, Y.; Jin, H. (2019). Rare Sound Event Detection Using Deep Learning and Data Augmentation. In Interspeech pp. 619-623.
[6] Dabiri, S.; Heaslip, K. (2019). Developing A Twitter-Based Traffic Event Detection Model Using Deep Learning Architectures. Expert Systems with Applications, **118**, pp. 425-439.
[7] Fernando, T.; Denman, S.; Sridharan, S.; Fookes, C. (2018). Soft + Hardwired Attention: An LSTM Framework for Human Trajectory Prediction and Abnormal Event Detection. Neural Networks, **108**, 466-478.
[8] Granat, J.; Batalla, J.; Mavromoustakis, C.; Mastorakis, G. (2020). Big Data Analytics for Event Detection in the Iot-Multicriteria Approach. IEEE Internet of Things Journal, **7**(5), pp. 4418-4430.
[9] Kidziński, Ł.; Delp, S.; Schwartz, M. (2019). Automatic Real-Time Gait Event Detection in Children Using Deep Neural Networks. PLOS ONE, **14**(1), e0211466.
[10] Liu, A.; Shao, Z.; Wong, Y.; Li, J.; Su, Y.; Kankanhalli, M. (2017). LSTM-Based Multi-Label Video Event Detection. Multimedia Tools and Applications, **78**(1), pp. 677-695.
[11] Lohithashva, B.; Manjunath Aradhya, V.; Guru, D. (2020). Violent Video Event Detection Based On Integrated LBP And GLCM Texture Features. Revue d'Intelligence Artificielle, **34**(2), pp. 179-187.
[12] Lu, C.; Shi, J.; Wang, W.; Jia, J. (2018). Fast Abnormal Event Detection. International Journal of Computer Vision, **127**(8), pp. 993-1011.
[13] Pouyanfar, S.; Chen, S. (2017). Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning. International Journal of Semantic Computing, **11**(01), pp. 85-109.
[14] Qu, S.; Guan, Z.; Verschuur, E.; Chen, Y. (2020). Automatic High-Resolution Microseismic Event Detection via Supervised Machine Learning. Geophysical Journal International, **222**(3), pp. 1881-1895.
[15] Rangasamy, K.; As'ari, M.; Rahmad, N.; Ghazali, N. (2020). Hockey Activity Recognition Using Pre-Trained Deep Learning Model. ICT Express, **6**(3), 170-174.
[16] Roy, V.; Noureen, S.S.; Bayne, S.B.; Bilbao, A.; Giesselmann, M. (2018). Event detection from pmu generated big data using r programming. In 2018 IEEE Conference on Technologies for Sustainability (SusTech) pp. 1-6.
[17] Singhal, S,; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S.I. (2019). Spotfake: A multi-modal framework for fake news detection. In 2019 IEEE fifth international conference on multimedia big data (BigMM) pp. 39-47.
[18] Sun, J.; Shao, J.; He, C. (2017). Abnormal Event Detection for Video Surveillance Using Deep One-Class Learning. Multimedia Tools and Applications, **78**(3), pp. 3633-3647.
[19] Tippaya, S.; Sitjongsataporn, S.; Tan, T.; Khan, M.; Chamnongthai, K. (2017). Multi-Modal Visual Features-Based Video Shot Boundary Detection. IEEE Access, **5**, pp. 12563-12575.
[20] Zhou, H.; Yin, H.; Zheng, H.; Li, Y. (2020). A Survey on Multi-Modal Social Event Detection. Knowledge-Based Systems, **195**, pp. 105695.

Authors Profile

K Swapnika pursuing Ph. D in Data Mining and Information Retrieval Systems stream at Jawaharlal Nehru Technological University, Hyderabad Hyderabad, she completed M. Tech in Software Engineering from Jawaharlal Nehru Technological University Hyderabad and has 8 years of academic experience. Her Research Interest includes Information Retrieval Systems and Bigdata Analytics.

Dr. D. Vasumathi is a Professor and HOD of Computer Science and Engineering Department in JNTUH College of Engineering, J. N. T. University, and Hyderabad. She has completed her PhD at J. N. T. University, Hyderabad in 2011 and has more than 20 years of experience in Teaching and Research. She is guiding 09 PhD scholars in Computer Science and Engineering and she is also Vice-President of National ST Employees and Officers Welfare Association (NST & OWA), and General Secretary for Teaching Association (NECTA), in JNTUH college of Engineering. She is Finance Secretary for both TS & AP States Tribal Development Association, TDA- Hyderabad.