BioGecko

Vol 12 Issue 03 2023
ISSN NO: 2230-5807

Coalition Strife Clustering Algorithm-based CRNN framework for Multimodal Sound Event Detection in Smart Cities

Arun Kumar Singh[1], Dr. Anoop Kumar[2]
[1]Research Scholar, Banasthali Vidyapeeth, Tonk, Rajasthan, India
[2] Assistant Professor, Banasthali Vidyapeeth, Tonk, Rajasthan, India

ABSTRACT
A quite serious concern in contemporary civilization in smart cities is the exponential growth of counterfeit news on social Medias. Of course, it is generated by means of the handling of pictures, video, audio, and text. This suggests that a multi-modal system is necessary for the recognition of audio/sound events. One of the most crucial obstacles in smart cities is critical sound event detection, prior research does not focus on the precious detection and also attains a lack training of in multimodal sounds overlapping. Hence this paper efficiently introduces the Coalition Strife Clustering Algorithm-based Convolutional Recurrent Neural Network framework (CSCA_CRNN), which enhances the sound event detection growth by tackling the overlapping of multimodal sounds. Initially, the work introduces the CSCA-based feature extraction for extracting the three levels of features via employing librosa library with optimized weight. After that deep convolution fusion strategy is utilized to amalgamate the extracted features and fed into the exceptional CRNN network for classification. In addition, CRNN performs the productivity of Sound Conjuncture Revelation and determination of Superintendence of Accession for reducing the overlapping of sound events with enhances the detection accuracy.

Keywords: Deep learning, Deep Neural network, Multimodal sound detection, Sound event detection.

INTRODUCTION
In the smart generation, internet-of-things (IoT) is the most widely adopted model for deploying artificial intelligence (AI), particularly DNNs to take out usable data from sensors [1]. This cloud centric strategy, however, has numerous disadvantages, including higher latency caused by reliability, communiqué delays, and availability constrained by the communiqués, confidentiality concerns brought on by the streaming of sensitive data to a remote site, and high energy costs for transmission of data [2]. By bringing AI near to the sensors and transmitting only pertinent data and warnings, edge computing is a cutting-edge solution to these restrictions [3]. IoT end-nodes often run-on batteries and aim for long battery life; in a perfect world, they would operate on energy harvesters, but the energy they collect is insufficient to run powerful CPUs. Public security issues are among the many topics that are often covered in case of smart cities. Given the situation, using lower cost sensor-based observation capability deliberately placed to generate a virtual supervising layer is ultimately necessary.

The typical method, which merely uses camcorders, might be used to conduct such remote surveillance. Larger count of camcorders is needed to capture the surroundings continually in order to record the full area of interest [4]. The high costs of the infrastructure and equipment needed to implement this strategy, the complexity of protection, the higher energy rate of the procedure, the further deprivation caused by weather, the augmented complexity of stealth fitting, the higher demand to process power, the dependence on enormous storage capacity, and, worst of all, the need for human guards to monitor the images are some of the challenges. The use of ambient sound may be a more practical solution to these issues; instead of filming the sites, audio [5] recordings of the

areas under observation will be made. Building a scheme that involuntarily detect possible public security events and take deeds in line with the sort of crisis observed would be more convenient than using individuals who would constantly monitor the surroundings waiting for a potential occurrence. Several works demonstrate the viability of employing sound as a surveillance resource [6], however, there is still little research that expresses worry about practicality, scalability, and cost. The most current research typically uses expensive [7]-[10] DL

algorithms with deep and complicated NN topologies and supercomputers with strong GPUs to do such tasks [11]. The statistic most frequently used to measure classifier performance is accuracy. However, because less expensive methods are frequently used for embedded hardware [12] installations, it is moreover practical to analyze classifier [13] processing period for categorization to be done in the circumstance of smarter cities.

Due to the huge size of the storage space, this big data is much beneficial in detecting events [14]. Additionally, the methods for past event recognition are heavily focused on a particular domain. Multi-modal event recognition techniques are still being introduced to locate events in massive amounts of diverse data, such as photos, video and texts recordings. Machine learning systems have a difficult time detecting events in a certain area [15]. As a result, numerous research works on event detection [16] introduce various deep learning (DL) strategies. The deep learning models feature several processing layers that help them comprehend the various degrees of abstraction in the data representation. For event detection, the current deep learning models outperform more established machine learning methods.

LITERATURE REVIEW

Advanne et al [17] the suggested technique can link numerous DOAs to their corresponding event labels of sound and track linkage over time. The proposed method avoids any method- and array-specific feature extraction by using the magnitude and phase components of spectrograms generated on every auditory channel separately as the characteristic. On 5 Ambisonic datasets and 2 spherical array datasets with various overlapping noise occurrences in anechoic, reverberant, and real-world contexts, the approach is evaluated. Comparisons are made between the suggested approach and 2 SED, 3 DOA estimates, and 1 SELD baseline. The findings demonstrate the suggested technique generality and applicability to all array topologies, as well as its resilience to scenarios involving low SNR, reverberation, and undetected DOA values.

Sharath Adavanne et al [18] The localization side is also addressed by recent developments in sound event detection systems, but they do not test the capabilities of localization and detection separately; instead, they measure the joint performance of the system as a whole. This study suggests adding a detection-related condition to the localization metrics and, in the opposite direction, using location data to determine the true positives for detection. The behavior of such joint measures is demonstrated using a detailed evaluation case. The suggested joint metrics work coherently and logically and effectively to define both aspects, as evidenced by the assessment of detection only and localization only performances.

Trigoni et al [19] A unique method used by SoundDet treats the spatiotemporal sound event as an absolute "sound object" to be recognized and consumes the raw, multichannel waveform directly. A backbone NN and 2 similar heads for sequential recognition and spatial localization, in that order, make up SoundDet. The backbone system initially learnt a bank of frequency selective and phase sensitive filters to clearly preserve direction of arrival data while significantly more efficient over others, provided the high sampling rates of the raw wave. After that, a proposal map is built to address the difficulties of anticipating occurrences with wider sequential variations.

Singhal et al [20] The subtask has a significant impact on false news recognition, and with no subtask training, performance often declines by 10%. We offer SpotFake, a multi-modal framework for false news recognition for addressing this crisis. Without considering any additional subtasks, our suggested approach can identify bogus news. It makes use of an article's textual and graphic components. In particular, we used language models (such as BERT) to learn text features, while VGG19, which was pre-trained on the ImageNet dataset, was used to learn image features. On 2 publicly accessible datasets, namely Weibo and Twitter, all analysis is run.

Qu et al [21] provided a new approach for high resolution micro-seismic event recognition to address the noise problem of weak microseismic events. Fix sized segmentation with 2 wavelengths is deployed for dividing the data for preprocessing. The support vector machine (SVM) scheme was afterward trained using 191 features that had been retrieved and used as input data. These characteristics, which show the consistency, smoothness, and irregularity of the events/noise, contain 128 2D textural and 63 1D features.

Wakayama et al [22] the studies, researchers found that it is challenging to construct a SED model that is more accurate than CNN-Transformer in predicting the future. We suggest architecture termed a CNN & SAN-Transformer that keeps CNN in blocks and employs SAN in every other block, to construct a scheme with higher

forecast precision whilst accurately captivating the characteristics of audio signal. The anticipated technique might be a parameter efficient design, according to experimental results, which show that it has the same or greater prediction accuracy with fewer constraints and higher forecast precision with more parameters than the CNN-Transformer.

COALITION STRIFE CLUSTERING ALGORITHM-BASED CRNN FRAMEWORK

In consumer and smart city applications, Sound Conjuncture Revelation (SCR) is a hipster area of concern. Deep neural networks (DNN)-based methods now in use are quite successful, but they are also very demanding in terms of real-time constraints and reliable classification with exact module detection. Hence to overcome these crucial obstacles, the paper introduces the Coalition Strife Clustering Algorithm based CRNN (CSCA_CRNN) framework for efficiently extracting the audio signal features and classifying the sound signal. The CSCA_CRNN structure's basic block diagram is shown in Fig. 1.

Librosa and CNN-CSCA-based feature extraction

At first feature, extraction is based on rising level, half level, and minimal level. In that, the Librosa library is used in the audio model development process to extract the audio features. To extract useful acoustic elements from the original audio files, the Librosa library is employed and audio feature types such as rising, half, and minimal features are all different kinds are extracted effectively. The zero-crossing rate, energy, and rising-level characteristics such as amplitude envelope are extracted. “Pitch, beat-related descriptors, Met Harmonic Cepstral Coefficients (MFCC), and half-level features” like melody are extracted as features of minimal-level. These collected features are used as input to CNN CSCA (CNN- Coalition Strife Clustering Algorithm). A classic CNN consists of solitary or plentiful sub-sampling layers and convolution blocks. Following the output layer and fully linked layers are the convolution layers. The details about those layers are described in the below section and illustrated in Fig. 2.

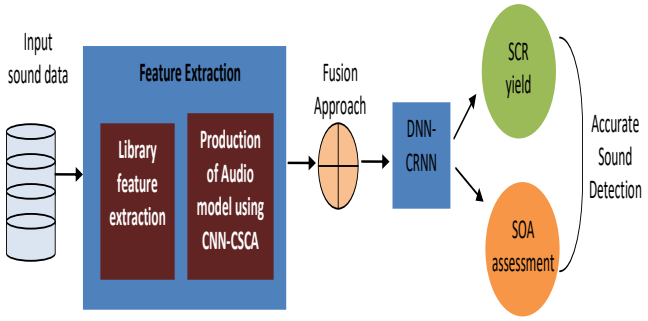

Fig. 1 Proposed CSCA-CRNN Framework

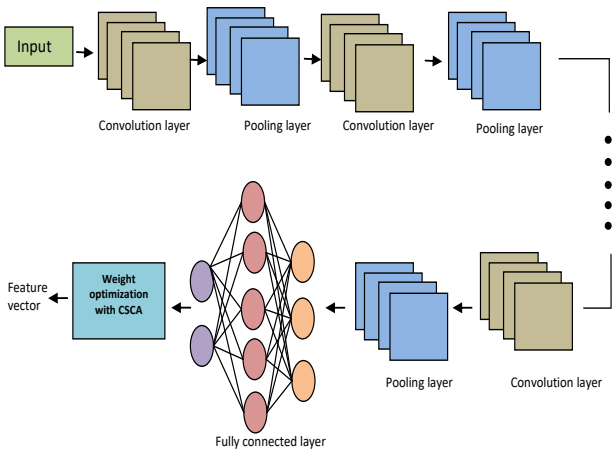

Fig. 2 2- CNN-CSCA structure

Convolutional layer: The audio features are stationary in nature, and the CNN is main component of the overall CNN structure. It implies that one audio section's composition is comparable to another's. As a result, a feature discovered in one area can correspond to an identical pattern discovered in another. From the greater section of features, a small portion is taken into account and transmitted as input. These features are combined into one output position. The signal is converted to a compressed form, and the filtered form is then passed on to the following layer.

Sub-Sampling or pooling layer: It is a down-sampling technique that uses a tiny amount of convolution output as an input. For reducing parameter count, and the computing difficulty, and to preserve the over fitting issues, this layer gradually reduces the size of the audio features. The pooling layer regulates the resolution of features to increase stability while also compressing the features that are present in the region created by the convolution layer.

Fully connected layer: This layer collects input from neurons and processes each neuron to create the output. The output of every part is intended to the activated unit of following layer, which is a crucial type of feed forward NN. The completely linked layer produces an output feature vector, however, owing to the existence of cross entropy losses, there is a potential that the accuracy of audio creation will suffer.

In order to improve accuracy, the weights are changed for optimizing the loss using the CSCA technique. The Strife Clustering algorithm incorporates strife-based learning owing to SCA. The strife numbers are essential to the initiative of this learning, and if the upper and lower bounds are defined by V and M , then the opposite form of an authentic quantity T is shown as (1),

$$P = V + M - T \tag{1}$$

The opposite of T is represented by P , which quickens convergence. The ideal audio feature vector is chosen based on the objective function. The expressions of leaders and followers in CSCA, which are expressed as follows, are updated through the audio features (2) and (3).

$$A_w^1 = \begin{cases} G_w + B_1((VC_w - MC_w)B_2 + MC_w) & B_3 \geq 0 \\ G_w - B_1((VC_w - MC_w)B_2 + MC_w) & B_3 < 0 \end{cases} \tag{2}$$

$$A_w^j = (A_w^j + A_w^{j-1})/2 \tag{3}$$

From the (4) expressions, VC_w and MC_w described a maximum and minimum bound of k^{th} aspect, correspondingly. Here consistent arbitrary numerical are B_2 and B_3 in the interval [0, 1]. A_w^1 and A_w^j denotes the head strife point and j^{th} follower strife in the w^{th} face. G_w described the provisions cause point in the k^{th} aspect. The rate of B_1 is assessed by,

$$B_1 = 2e^{-(4m/M)^2} \tag{4}$$

Where, the maximum number of iterations is M and the current iteration number is m. The audio model generation produces an efficient audio feature vector. Hence based on that feature vector, the work insists on the fusion strategy based on deep convolution that is elaborated in the following section.

Approach for deep convolution Fusion

To enhance the effectiveness of an incident detection procedure, deep feature fusion takes into account extremely contextual information as well as technical description. The audio features are described as a rising feature as $R = \{r_1, r_2 ..... r_n\}$, half feature as $H = \{h_1, h_2 ..... h_n\}$, and minimal feature as $M = \{m_1, m_2 ..... m_n\}$. Such heterogeneous synthesis features that have been extracted are combined to generate a unique vector that is denoted by the letter $W = \{w_1, w_2 ..... w_n\}$. The deep feature fusion strategy's organizational structure is shown in Fig. 3.

The upcoming derivations are described as the fused feature vectors that are (5)

$$FF = R + H + M \tag{5}$$

The heterogeneous generating output's feature vectors are combined data into a single vector using the deep feature fusion approach. Mostly on the pooling stage, which compresses the input vector, the feature vector is gathered. The suggested approach has a very low complexity level, and it produces worthwhile results. Then the generating outputs feature vector is $G_{OF} = \{G_{OF1}, G_{OF2}, ... G_{OFn}\}$ where, G_{OF} can be defined by (6),

$$G_m = \max_{t=1}^{l}(FF_m), m = 1,2,...n \tag{6}$$

Where, l bears a resemblance to the capacity of the pooling stage. Correspondingly the paper performs the detection of sound events based on extracted and fused features, thus the details follow.

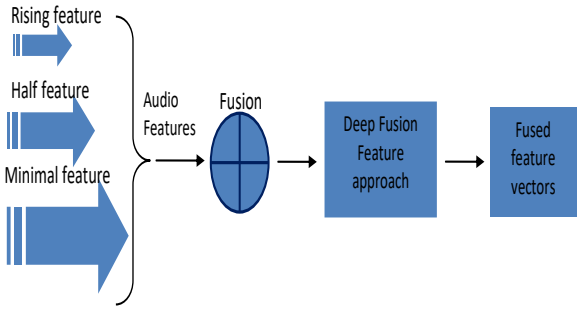

Fig. 3 Approach for deep convolution Fusion strategy

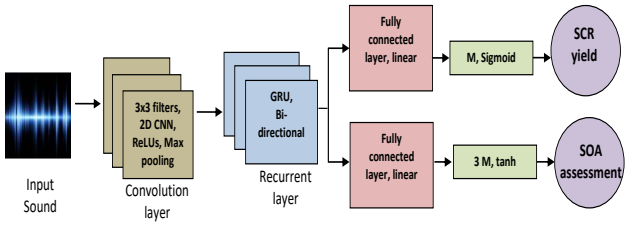

Fig. 4 Unique CRNN network structure

CRNN based SCR with SOA Detection

Phase and magnitude spectrograms are features that are extracted using the suggested approach from multi-channel audio. First, a number of features are fed into continuous spectrum frames using this method, which then go on to forecast the events with active sound and their spatial positions inside every segment. In order to provide two outputs, this approach processes a number of features using CRNN. Two branch jobs make use of these two outputs. For each time segment in the first branch, Sound Conjuncture revelation (SCR) can approximate numerous over lapping sound event and attain multi label categorization. In the second branch, every sound event set is linked with 3 regresses, and the Superintendence of accession (SOA) is estimated using 3-D Cartesian coordinates.

CNN and FC neural networks were both utilized as classifiers by early researchers. This study, combine CNNs and RNN layers to give the network the ability to learn long-term temporal information. Figure 4 displays the unique CRNN network structure used in this study. Three different neural network layer components make up the CRNN:

1) 1. Convolutional layers: The CNN layers receive the retrieved characteristics as input. The feature output is passed to the activation function after passing through the 2D convolution filters on the CNN layers. This paper employs ReLU for the activation function and max pooling for dimensionality reduction.
2) 2. Recurrent layers: By feeding a bidirectional RNN the output of a CNN, more contextual sequence information can be efficiently learned. Q nodes with GRU activated by the tanh function are present in each RNN layer.
3) 3. Fully connected layers: For SCR and SOA estimation, the FC layers have two parallel branches. The first FC layer in the SCR branch includes N linearly activated nodes, and the second FC layer has M sigmoid activated nodes, allowing for the simultaneous activation of several classes. The M nodes match the M sound event classes that need to be found. There are 3M nodes with tanh activation in the second FC layer of the SOA branch. Three nodes a, b, and c positions of sound events in 3-dimensional Cartesian coordinate, are used to characterize each class of sound event.

To assess SCR and SOA estimations, the work employs independent metrics. The paper employs the SCR metrics, F score, and joint evaluation of error rate $E_r R_a$, which is derived from the collective amount of all test set segments.

The following is the formula for calculating the F-score (7).

$$F_s = \frac{2\sum_{j=1}^{O} SE_j}{2\sum_{j=1}^{O} SE_j + \sum_{j=1}^{O} FPO_j + \sum_{j=1}^{O} FNE_j} \quad (7)$$

In addition to being active in ground truth, SE_j is the count of sound events that were picked up in the j-th 1 sec segment. The quantity of false positive is known as FPO_j. The term "false negatives" FNE_j refers to the number of

times a network output misrepresents an event as being active while, in reality, it is inactive for the j-th segment of the one-second segment.

Error rates depends upon the segment and it is the other evaluation metric. The active sound event in a 1 second segment is shown by the notation NE_j. The precise formula for the error rate depends on the deletions T_j, insertions E_j, and substitutions J_j. The E_rR_a metric is computed as follows (8),

$$E_rR_a = \frac{\sum_{j=1}^{O} T_j + \sum_{j=1}^{O} E_j + \sum_{j=1}^{O} J_j}{\sum_{j=1}^{O} O_j} \quad (8)$$

The combined evaluation of SOA error and frame recall is used in the SOA estimate method. The reference of the dataset's synthesis (a_H, b_H, c_H) and the anticipated SOA estimations (a_F, b_F, c_F) form the central angle, which is provided by (9),

$$\lambda = 2.\arcsin\left(\frac{\sqrt{(a_H - a_F)^2 + (b_H - b_F)^2 + (c_H - c_F)^2}}{2}\right).\frac{180}{\alpha} \quad (9)$$

Then to estimate the SOA_{error} as (10),

$$SOA_{error} = \frac{1}{L}.\sum_{l=1}^{L} \lambda\left(\left(a_H^l, b_H^l, c_H^l\right), \left(a_F^l, b_F^l, c_F^l\right)\right) \quad (10)$$

Where, L is the whole determination of SOA_{error} and $\lambda\left(\left(a_H^l, b_H^l, c_H^l\right), \left(a_F^l, b_F^l, c_F^l\right)\right)$ is the angle between l-th predictable and orientation SOAs.

The frame recall rate was also measured as a % of TP / (FN + TP). Where FN implies count of frames with predicted and reference SOAs that are not equal, and TP implies count of predicted SOAs that is equivalent to whole count of references.

The work also incorporates determining the SCLR score during the training procedure to assess as follows (11),

$$SCLR_{score} = (SCR_{score} + SOA_{score})/2 \quad (11)$$

Where (12) and (13),

$$SCR_{score} = (FS + (1 - G))/2 \quad (12)$$

$$SOA_{score} = \left(SOA_{error}/180 + (1 - framerecall)\right)/2 \quad (13)$$

The optimum result in the suggested procedure has an SCLR score of zero. The estimated frame recall rate of SOA is used to describe the SCLR score. When compared to SOA estimation, the SCLR score can provide a more accurate assessment of SCR performance.

For training SCLR with Adam optimizer, a weighted mixture of MSE and cross-entropy loss is used. The cross-entropy estimation (14) is:

$$I_{z'}(z) = -\sum_j z'_j \log(z_j) \quad (14)$$

From the above considerations, the work clearly described the proficiency of the proposed work in multimodal sound event detection in smart cities via tackling the major obstacles such as overlapping of sounds and accurate detection. To prove the competence of the work, the paper attains the experimental analysis that is elaborated in the following section.

Simulation result

The analysis was done using matlab. The performance of developed model was examined over CNN, RNN and

LSTM in terms of accuracy, error, F1-score, recall and precision.

Accuracy

Accuracy is "defined as one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right". It can be represented as in (15), wherein, NCP refers to count of correct predictions and TNP refers to total count of predictions.

$$Accuracy = \frac{NCP}{TNP} \quad (15)$$

The accuracy is shown in Fig. 5. The accuracy should be better for enhanced system performance. As stated, the accuracy attained by proposed method is high over the compared schemes like CNN, RNN and LSTM. The LSTM has gained an accuracy of around 75, while, RNN has gained an accuracy of around 78 and CNN has gained an accuracy of around 89, whereas, our developed model shows a higher accuracy of around 91. This betterment is due to the incorporated improvisations and optimization concept in our work.

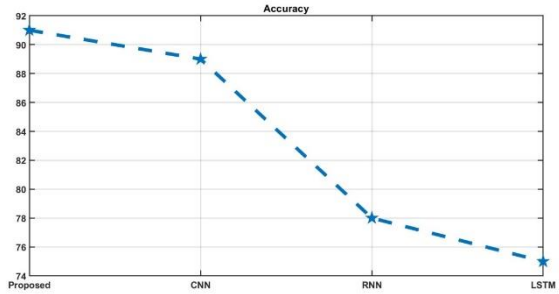

Fig. 5 Accuracy of developed model over conventional schemes

Error

"Error (also known as the out-of-sample error or the risk) is defined as a measure of how accurately an algorithm is able to predict outcome values for previously unseen data". The error should be less for better system performance. The error is plotted in Fig. 6. From Fig. 6, the error attained by proposed method is less over the compared schemes like CNN, RNN and LSTM. The LSTM has acquired a high error of around 25, while, RNN has acquired an error of around 21 and CNN has acquired an error of around 10, whereas, our developed model shows a much lesser error of around 8. This minimal error is due to the incorporated optimization and improvisations done in our work.

F1-score

The F1-score attained by proposed method is high over the compared schemes like CNN, RNN and LSTM as shown in Fig. 7. "F1-score is defined as one of the most important evaluation metrics in machine learning. It elegantly sums up the predictive performance of a model by combining two otherwise competing metrics — precision and recall". Since the metrics precision and recall comprises F1-score, definitely it should be high for better system performance. Our proposed method has gained high F1-score than others. The LSTM has acquired a relatively lesser F1-score of around 72, while, RNN has acquired a F1-score of around 77 and CNN has acquired an F1-score of around 87, whereas, our developed model shows a much higher F1-score of around 90. This high F1-score value attained by proposed method is due to the enhancements and techniques adopted in this research.

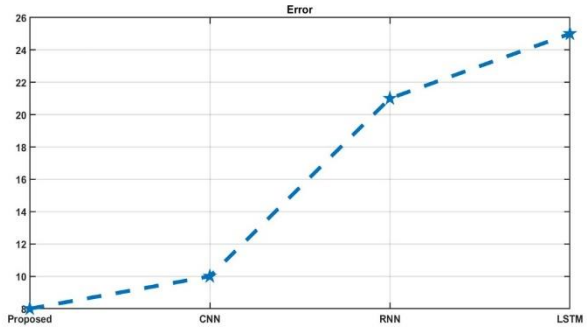

Fig. 6 Error of developed model over conventional schemes

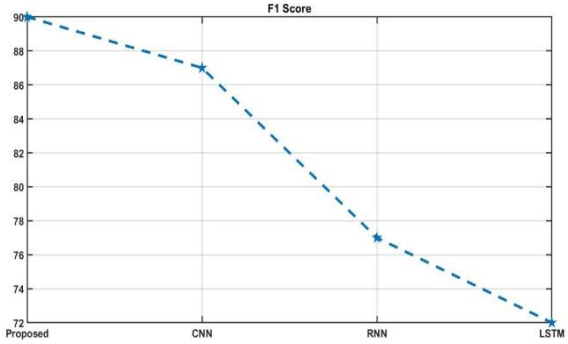

Fig. 7 F1-score of developed model over conventional schemes

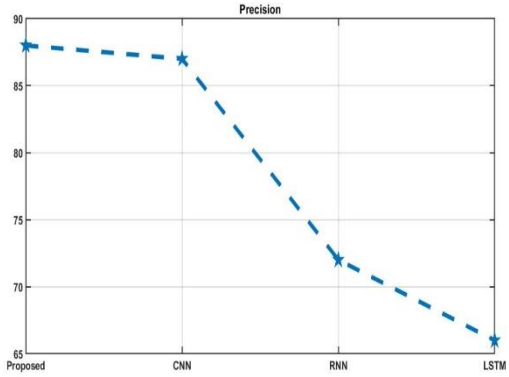

Fig. 8 Precision of developed model over conventional schemes

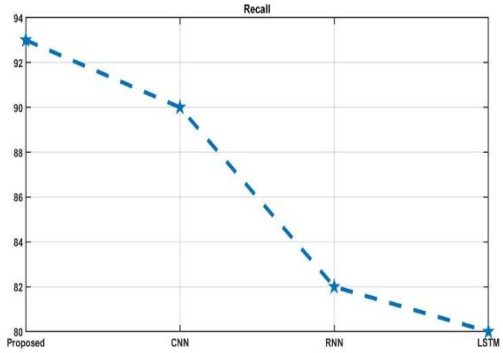

Fig. 9 Recall of developed model over conventional schemes

Precision

"Precision is defined as the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives)". The precision metric is almost equal to accuracy and therefore, the values should be higher for better performance of the system. The Precision graph is seen in Fig. 8. In Fig 8, it can be observed that the proposed method acquired a high precision value of 88, while, compared schemes like CNN acquired a precision value of 87, RNN acquired a precision value of 72 and LSTM acquired a minimal precision value of 66. Thus, the performance of our developed system is proven to be better over others.

Recall

"Recall is defined as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples". The recall for the developed system should be higher than others. The recall graph is plotted in Fig. 9. In Fig. 9, it can be observed that LSTM scores less recall of 80, RNN scores a recall of 82; CNN scores a recall of 90 and our developed model scores higher recall of 93. Thus, our proposed model is proven to be a better one over the compared approaches like RNN, CNN and LSTM.

The overall plot of proposed model for varied metrics like accuracy, error, F1-score, precision and recall are shown in Fig. 10.

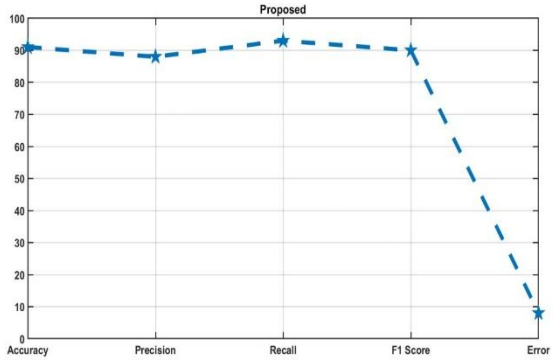

Fig. 10 Performance of developed model for varied metrics

CONCLUSION

This paper efficiently introduced the CSCA_CRNN) which enhanced the sound event detection growth by tackling the overlapping of multimodal sounds. Initially the work introduced the CSCA-based feature extraction for extracting the three levels of features via employing librosa library with optimized weight. After that deep convolution fusion strategy was utilized to amalgamate the extracted features that were fed into the exceptional CRNN network for classification. In addition, CRNN performed the productivity of Sound Conjuncture Revelation and determination of Superintendence of Accession for reducing the overlapping of sound events with

enhanced the detection accuracy. Finally, analysis was done to prove the enhancement of our work. From the analysis, our proposed method has gained high F1-score than others. The LSTM has acquired a relatively lesser F1-score of around 72, while, RNN has acquired a F1-score of around 77 and CNN has acquired an F1-score of around 87, whereas, our developed model shows a much higher F1-score of around 90. Moreover, it can be observed that LSTM scored less recall of 80, RNN scored a recall of 82; CNN scored a recall of 90 and our developed model scored higher recall of 93. Thus, the enhanced performance of our suggested scheme was proven from the outcomes.

REFERENCES

K. R. Coffey, R. G. Marx, and J F. Neumaier, "DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations," Neuropsychopharmacology, vol. 44, no. 6, pp.859-68, Apr. 2019.
Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," Artificial Intelligence Review, Vol. 53, no. 4, pp.2313-39, Apr. 2020.
J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," In2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp. 126-130, Mar. 2017.
J. Andreu-Perez, H. Perez-Espinosa, E. Timonet, M. Kiani, M. I. Girón-Pérez, A. B. Benitez-Trinidad, D. Jarchi, A. Rosales-Pérez, N. Gatzoulis, O. F. Reyes-Galaviz, and A. Torres-García, "A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels," IEEE Transactions on Services Computing, Vol. 15, no. 3, pp. 1220-32, Feb. 2021.
Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," Methods in Ecology and Evolution, vol. 10, no. 3, pp. 368-80, Mar. 2019.
J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," Pattern recognition letters, vol. 1, pp. 119:3-11, Mar. 2019.
Y. Bao, Z. Tang, H. Li, and Y. Zhang, "Computer vision and deep learning–based data anomaly detection method for structural health monitoring," Structural Health Monitoring, vol. 18, no. 2, pp. 401-21, Mar. 2019.
X. W. Ye, T. Jin, and C. B. Yun, "A review on deep learning-based structural health monitoring of civil infrastructures," Smart Struct. Syst. Vol. 24, no. 5, pp. 567-85, Nov. 2019.
Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," In2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF) pp. 1-7, Oct. 2019.
Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," InInterspeech pp. 82-86, Aug. 2017.
O. Mac Aodha, R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, and I. Pandourski, "Bat detective—Deep learning tools for bat acoustic signal detection," PLoS computational biology. Vol. 14, no. 3, pp. e1005995, Mar. 2018.
N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images," IEEE transactions on industrial informatics. Vol. 14, no. 12, pp. 5530-8, Oct. 2018.
J. Yang, S. Li, Z. Wang, H. Dong, J. Wang, and S. Tang, "Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges," Materials, vol. 13, no. 24, pp. 5755, Dec. 2020.
Y. Y. Zheng, J. L. Kong, X. B. Jin, X. Y. Wang, T. L. Su, and M. Zuo, "CropDeep: the crop vision dataset for deep-learning-based classification and detection in precision agriculture," Sensors, vol. 19, no. 5, pp. 1058, Mar. 2019.
Z. Wang, J. Ren, D. Zhang, M. Sun, and J. Jiang, "A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos," Neurocomputing, vol. 26, pp. 287:68-83, Apr. 2018s.
P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," The Journal of the Acoustical Society of America, vol. 152, no. 1, 107-51, Jul. 2022.

S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” IEEE Journal of Selected Topics in Signal Processing. Vol. 13, no. 1, pp 34-48, Dec. 2018.
S. Adavanne, A. Politis, and T. Virtanen, “TAU Moving Sound Events 2019-Ambisonic, Anechoic, Synthetic IR and Moving Source Dataset [Data set],” Zenodo, 2019.
Y. He, N. Trigoni, and A. Markham, “SoundDet: polyphonic moving sound event detection and localization from raw waveform,” InInternational Conference on Machine Learning, vol. 1, pp. 4160-4170, Jul. 2021.
S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. I. Satoh, “Spotfake: A multi-modal framework for fake news detection,” In2019 IEEE fifth international conference on multimedia big data (BigMM), vol. 11, pp. 39-47, Sep. 2019.
S. Qu, Z. Guan, E. Verschuur, and Y. Chen. “Automatic high-resolution microseismic event detection via supervised machine learning,” Geophysical Journal International, vol. 222, no. 3, pp. 1881-95, Sep. 2020.
K. Wakayama, and S. Saito, “CNN-Transformer with Self-Attention Network for Sound Event Detection,” InICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 23, pp. 806-810, May 2022.