Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Empirical-Based Fusion Deep Convolutional Neural Network for Multimodal Emotion Recognition

Shailesh Kulkarni[1, 2], S.S. Khot[3], Yogesh Angal[4]
[1]*Department of Electronics and Telecommunication, JSPM'S Rajarshi Shahu College of Engineering, Tathawade, Pune, Maharashtra, India, Email: kulkarnishaileshece@sanjivani.org.in*
[2]*Department of Electronics and Computer Engineering, Sanjivani College of Engineering, Kopargaon, Maharashtra, India*
[3]*Department of Electronics and Telecommunication, K J College of Engineering and Management Research, Pune, Maharashtra, India, Email: drkhotss@gmail.com*
[4]*Department of Electronics and Telecommunication, JSPM'S Bhivarabai Sawant Institute of Technology and Research, Wagholi, Pune Maharashtra, India, Email: ysangal_entc@jspmbsiotr.edu.in*
Corresponding Author: Shailesh Kulkarni; Email: kulkarnishailesh189@gmail.com

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Emotion recognition plays an effective and efficient role in identifying a person's feelings. The performances of using either one feature provide no accurate recognition, in case the format is vague. This research develops a new model, a deep convolutional neural network with empirical approach-based fusion (EBF-DCNN) for emotion recognition. The proposed EBF-DCNN model extracts the audio, visual, and text formats to enhance the emotion recognition process. In this approach, three DCNN models are trained using either format, which consequently reduces the time dependencies and recognition is much faster than the other methods. The model adopts an empirical approach-based fusion method to fuse three data formats, which is highly feasible to avoid over-fitting problems. Here, the DCNN model outperformed with better results and also minimized the computational complexity. Moreover, the model is quite flexible and scalable to recognize the emotions of humans. The performance of the EBF-DCNN model can be evaluated by four metrics such as accuracy, precision, recall and F1 score, and achieved 94.33%, 93.80%, 94.08, and 93.94% for emotion recognition compared to other state-of-the-art methods.

Keywords: Emotion recognition, deep convolutional neural network, empirical approach-based fusion, multi-modal, deep learning.

1. INTRODUCTION:

Emotion recognition is a critical area of research in human-computer interaction, affective computing, and artificial intelligence. Automating the detection and interpretation of human emotions enhances user experience and improves communication between humans and machines. This is crucial for applications such as mental health monitoring, customer service automation, and interactive entertainment systems. However, accurately identifying human emotions is challenging due to the complexity and subtlety of emotional expressions, which are conveyed through multiple modalities. Relying on a single modality for emotion recognition often leads to vague interpretations and suboptimal performance.

Recent advancements in deep learning, especially convolutional neural networks (CNNs), have demonstrated significant potential in addressing the challenges of emotion recognition. CNNs can automatically learn complex representations from data without manual feature extraction. However, models trained on a single modality often struggle with time dependencies, data noise, and the risk of overfitting, particularly with limited datasets. This limitation results in less accurate emotion recognition, which can notably impact the performance of real-time systems.

To address the challenges of multimodal emotion recognition, a novel Empirical-Based Fusion Deep Convolutional Neural Network (EBF-DCNN) has been proposed. This approach integrates audio, visual, and textual features to improve recognition accuracy by capturing the complementary nature of emotional expressions. The empirical-based

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

fusion method reduces the risk of overfitting and enhances generalization by structurally fusing the modalities. Leveraging information from multiple sources, the model effectively addresses ambiguities arising from using a single modality and reduces time dependencies, enabling faster and more efficient emotion recognition compared to existing methods.

The primary objective of this research is to develop an efficient and scalable multimodal emotion recognition model that combines the strengths of different data formats. The proposed empirical-based fusion DCNN model integrates audio, visual, and text data at different stages of the recognition process to improve the overall performance. This fusion strategy is designed to maximize the contribution of each modality while minimizing noise and redundancy. Additionally, the model is designed to be computationally efficient, making it feasible for real-time emotion recognition in practical applications.

In summary, the paper presents three significant contributions. Firstly, it introduces an innovative empirical-based fusion approach that integrates multiple modalities (audio, visual, and textual) to enhance emotion recognition. Secondly, it proposes a scalable and flexible deep convolutional neural network architecture capable of achieving superior performance in multimodal emotion recognition tasks. Finally, it provides a comprehensive performance analysis, demonstrating the efficiency of the EBF-DCNN model in terms of recognition accuracy and computational time.

2. RELATED WORK:

Emotion recognition is a technology used to automatically identify people's feelings like happiness, anger, and sadness using speech, text, or image. Traditional HRI methods struggle with accurately detecting emotions, especially wild emotions [11][6]. Deep learning approaches, such as Long Short-Term Memory (LSTM) networks and Deep Convolutional Neural Networks (DCNNs), have overcome these limitations and trends in multimodal emotion recognition, enabling more accurate and efficient emotion recognition [5]. These models have proven effective in capturing and judging emotions based on facial expressions.

The motivation behind the detection and recognition process of emotions made it possible to develop research that integrates multi-modality data and provides complementary data between the modalities. The recent research is prominently achieved using deep learning techniques which grow rapidly and increase the power efficiency for automated emotion recognition [7] [2]. The emotion recognition process reflects major fields like medicine, social media platforms, organizations, communication, and so on [8] [3].

In recent, the deep learning techniques-based approach, one-dimension deep CNN method was developed for emotion recognition due to its simple and general approach [1]. Deep neural networks (EBF-DCNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) are more efficient methods for emotion recognition using speech signals [17]. Other DL techniques like three-dimensional CNN attention sliding recurrent neural network (ASRNN) methods [18] [5] were also developed for effective emotion recognition, which was highly used for extracting the local features from the dataset and effectively training the model. Moreover, these applications are not highly reliable, scalable, and computationally expensive for emotion recognition. Emotion recognition is useful in quite many tasks such as identifying customer satisfaction, e-learning, criminal activities, security monitoring, smart card applications, social robots, and so on [9] [10].

H. M. Shahzad et al. [1] developed the Multimodal CNN features for recognizing the emotions of humans, the approach utilized a standard dataset for identifying facial and vocal emotion expressions. This developed method understands the complex relationship between masks and vocals. Alireza Sohail Masood Bhatti et al. [2] introduced a multimodal-based deep learning approach for emotion recognition, more methods are implemented to solve the overfitting. Moreover, it requires more time to train the data. Shuai wang et al, [3] introduced a deep learning model for multimodal emotion recognition based on the fusion of EEG and facial expressions. The training process of the model is stable but also the model requires more time for the operation. Dilnoza Mamieva et al. [4] used a new attention-based approach for multimodal emotion recognition; two datasets are used in this approach. The overfitting issue has been improved. However, the model faces challenges in misclassification.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Minjie Ren et al. [5] developed a Multimodal Adversarial Learning Network (MALN) for conversational emotion recognition. However, it suffers from misclassification in some emotional cases. Bogdan Mocanu et al [6] combine the spatial, channel, and temporal attention mechanism into a visual (3D-CNN) and temporal attention mechanism into an audio (2D-CNN) for the identification of emotion. The developed model can identify the primary emotions only. However, it suffers from detecting the secondary emotion state.

2. PROPOSED METHODOLOGY:

The research focuses on utilizing a multimodal distributed architecture-based EBF-DCNN to identify individual emotions using multi-features. Data from the MELD dataset is pre-processed to enhance emotion recognition by reducing background noise, cleaning, transforming, and integrating the video.

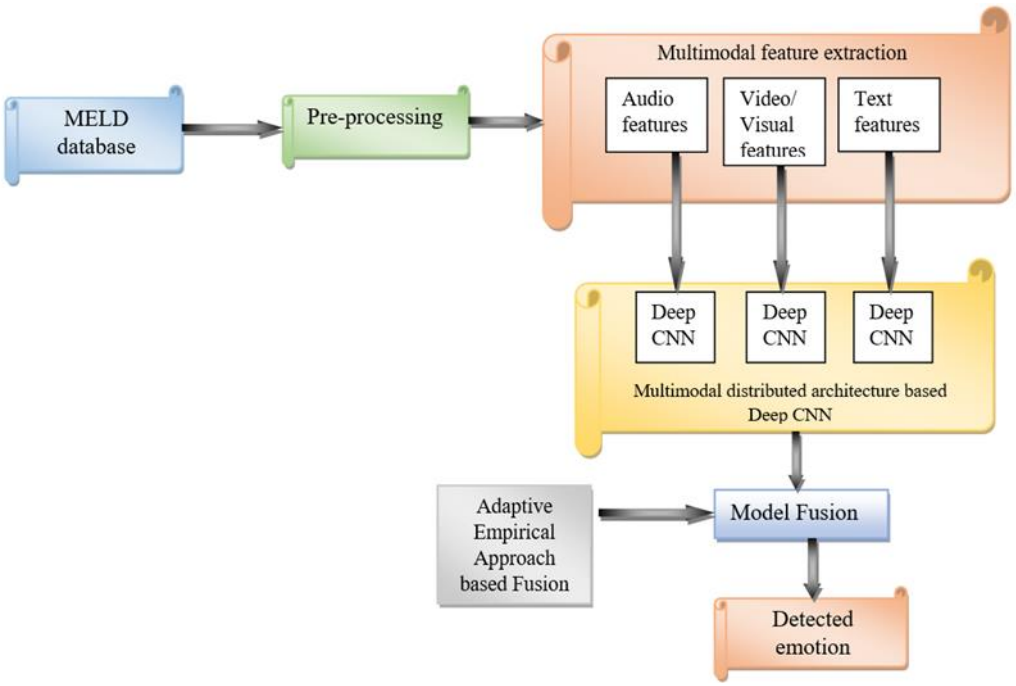

Figure 1: Proposed methodology for emotion recognition

The pre-processed video is classified into audio, visual, and text features extracted using statistical features, LBP, LDP, LOOP channels, and a graph embedding process. These features are deployed to the EBF-DCNN model, effectively training and fusing the model for emotion recognition. The EBF-DCNN model addresses time-consuming and over-fitting issues, enhancing the recognition process.

In this research, the MELD dataset is taken, which holds video-based input for emotion recognition. The MELD dataset [12] holds multiple corresponding video-based data for emotion recognition in which, the audio, text, and images can be extracted from the video. This dataset contributes over and above 1400 dialogues with 13000 sounds and expressions like anger, happiness, sadness, and so on from the very popular Friends TV show. Each sequence of the video is labelled with multiple emotions and can be defined as,

$$D = \left[V_1, V_2, V_3, \ldots V_j, \ldots V_n \right] \tag{1}$$

Here, D denotes the video dataset with the number of videos V , then V_j represents the j^{th} video and V_n denotes the total number of video data in the dataset. Initially, the video input undergoes pre-processing using the viola-Jones algorithm, which is an automatic face detector [13] that effectively cropped the face without unwanted background noise from the video frames and helps to detect the necessary information for further processing. The pre-processed video data can be denoted as P which is represented in the below equation. The video frames are converted into audio files, and text format for emotion recognition using multiple statistical features.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$D=\left[P_1,P_2,P_3,...,P_j,...,P_n\right] \tag{2}$$

2.1. Impact of Each Modality:

A. Audio format feature extraction:

Audio files are extracted from video frames for accurate emotion recognition by analyzing sound waves using statistical features. Key measures include the mean (Equation 3) for the average signal, standard deviation (Equation 4) for signal variation, and skewness (Equation 5) to assess data asymmetry (Equation 6) all contributing to a detailed analysis for enhanced emotion recognition.

$$\varphi^2=\frac{\sum_{j=1}^{P_n}\left(P_j-\overline{P}\right)^2}{P_n} \tag{3}$$

$$\psi=\sqrt{\frac{\sum_{j=1}^{P_n}\left(P_j-\overline{P}\right)^2}{P_n-1}} \tag{4}$$

$$\mu_3=\sum^{P_n}\frac{\left(P_j-\overline{P}\right)}{\left(P_n-1\right)\times\psi^3} \tag{5}$$

$$\xi=\sum\frac{\left(P_i-\overline{P}\right)^4}{P_n\psi^4} \tag{6}$$

By utilizing the above-mentioned factors, the audio from the video data is efficiently extracted and can be denoted as $A=\left\|\overline{P}\right\|\left\|\varphi^2\right\|\left\|\psi\right\|\left\|\mu_3\right\|\left\|\xi\right\|$ emotion recognition.

B. Visual format feature extraction:

The hybrid multilevel ternary pattern method identifies image patterns using blue, green, and red channels, utilizing texture feature extraction processes like LBP, LDP, and LOOP, and is used for emotion recognition texture analysis.

$$\zeta_{s,r}\left(a_C,b_C\right)=\sum_{j=1}^{s-1}2^s q\left(k_s-k_C\right) \tag{7}$$

Here, k_s,k_C represent the neighbourhood and center pixel values, respectively, with r as the radius, q is the gradient, and s is the sampling point

The Local Directional Pattern (LDP) enhances texture analysis by reducing noise and detecting fixed patterns [15], calculated using Equation 8.

$$\chi\left(a_C,b_C\right)=\sum_{n=0}^{s-1}q\left(u_n-u_f\right)\cdot 2^n \qquad \begin{cases}1 & if\ q\geq 0\\ 0 & otherwise\end{cases} \tag{8}$$

here, u_n denoted the kirsch masks, which $\left(a_C,b_C\right)$ represent the center pixels and u_f denote the f^{th} kirsch activation for visual extraction.

The Local Oriented Patterns (LOOP) technique, which captures both local and global texture variations, uses Equation 9 to categorize patterns efficiently [14].

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\gamma(a_C, b_C) = \sum_{n=0}^{s-1} q(t_n - t_C) \cdot 2^{e_n} \qquad \begin{cases} 1 & if\ q \geq 0 \\ 0 & otherwise \end{cases} \tag{9}$$

Here, t_C denotes the intensity of an image and e_n denotes the exponential of each pixel. The above-extracted features from LBP, LDP, and LOOP techniques are combined to form a hybrid multilevel ternary pattern (HMTP), which provides an efficient and effective vector image for further classification.

C. Text format feature extraction

The text modality adds a distinct dimension to emotion recognition, especially in analyzing the sentiment conveyed through spoken language or written communication. This textual analysis is particularly important when emotions are not explicitly expressed through audio or visual cues.

$$C_S(X, Y) = \cos\theta = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{j=0}^{n} X_j Y_j}{\sqrt{\sum_{j=0}^{n} X_j^2} \cdot \sqrt{\sum_{j=1}^{n} Y_j^2}} \tag{10}$$

Here, C_S denotes the cosine similarity of (X, Y) vectors attribute with n dimension and $X_j and Y_j$ represents the j^{th} components of cosine similarity vectors. The linear kernel graph reduces the complexity and lowers the computational cost for effective emotion recognition [16]. Furthermore here, v, w are the input vectors of the text features.

$$L(Z_v, Z_w) = Z_v^T Z_w \tag{11}$$

2.2. Multimodal distributed architecture based on EBF-DCNN

The multifaceted distributed architecture, which includes text, visual, and audio features, has shown great results in emotion recognition. It employs 1D convolution and pooling layers for audio and a 2D layer for text and visual data, extracting deep video representations using a large-scale DCNN. This model reduces overfitting and improves attentional region performance as shown in Figure 2.

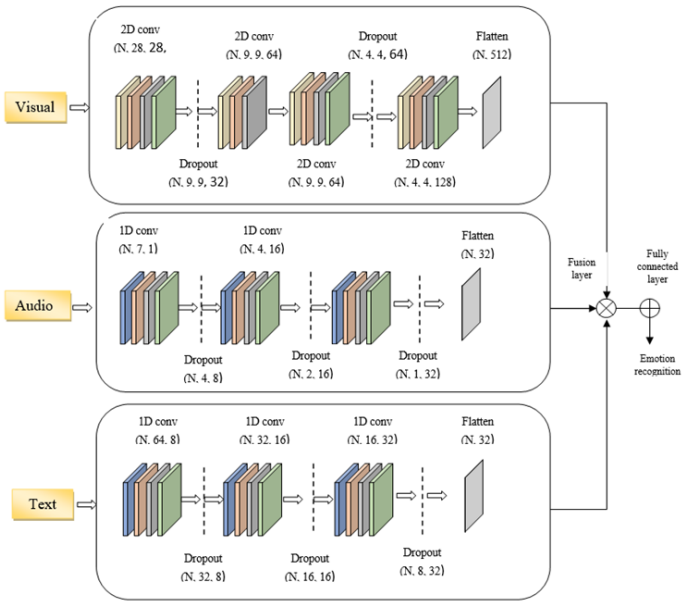

Figure 2: Architecture of the proposed model

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. EXPERIMENTAL SETUP AND RESULTS:

This experiment provided enhanced emotion recognition by evaluating the EBF-DCNN model, which achieved the highest accuracy compared to traditional models. The model's performance was fine-tuned using key hyperparameters. The ADAM optimizer was employed for efficient optimization, while a dropout rate of 0.5 was set to prevent overfitting. A learning rate of 0.01 ensured steady convergence, and a batch size of 32 helped manage the model's training data. The mean square error (MSE) was used as the loss function, with ReLU as the activation function for non-linearity. The model's performance was evaluated based on accuracy and mean absolute error metrics. The results of this setup, along with a discussion of existing methods, are thoroughly described in the following sections

3.1 Dataset Description

Multimodal EmotionLines Dataset (MELD): By increasing and extending the EmotionLines dataset, the MELD dataset is created. MELD and EmotionLines contain the same dialogues but the MELD dataset encompasses audio and visual modality along with text. MELD has more than 1400 dialogues and 13000 utterances. Each utterance in a dialogue is labelled by seven emotions. The total number of samples used in this research is 9812, where 4630 samples are neutral emotions, 1707 samples for joy, 673 for sadness, 1085 samples for anger, 1187 samples for surprise, 260 for fear, and 270 samples for disgust emotion

The experiment analysis shows the results of visual extraction from the MELD dataset for emotion recognition using the proposed EBF-DCNN model which is illustrated in Figure 3.

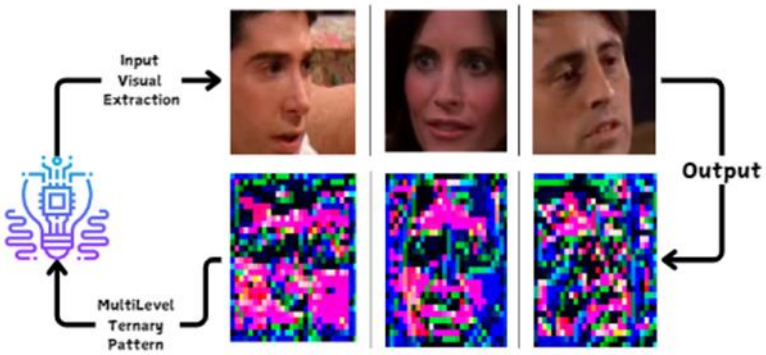

Figure 3: Experiment analysis of visual extraction

3.2 Performance analysis

In this analysis, the performance of the EBF-DCNN model using the MELD dataset shows quite high accuracy and achieved better performances for emotion recognition which is represented in Table 1. In this context, the performance of the EBF-DCNN with a maximum training percentage of 90 and with varying epochs. This analysis proves that the EBF-DCNN model achieved higher performance than the other models in terms of accuracy, precision, recall, and F1-score. The proposed fusion technique effectively mitigates overfitting and accelerates convergence, making the model more robust and efficient in training.

EBF-DCNN Epoch Values	Accuracy %	Specificity %	Precision %	Recall %	F1 Score %
Epoch = 100	88.81	86.54	88.67	91.07	89.86
Epoch = 200	91.39	91.36	90.61	91.43	91.01
Epoch = 300	93.35	94.19	91.97	92.51	92.24

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Epoch = 400	93.91	94.23	92.09	93.58	92.83
Epoch = 500	94.33	94.58	93.80	94.08	93.94

Table 1: Performance analysis of the EBF-DCNN model

3.3 Comparative analysis

The EBF-DCNN model outperforms existing methods such as such as support vector machine (SVM) [19], Xgboost [20], decision tree [22] [23], LSTM [21], and DCNN by achieving maximum accuracy for emotion recognition. Unlike other methods, the EBF-DCNN model incorporates audio, visual, and text parameters in a single model, reducing over-fitting and demonstrating robustness in training. The upgraded model proves to be more efficient for emotion recognition, as indicated by comparative performance measures. The detailed comparison using specific metrics at a constant training percentage of 90 is provided in the subsequent context.

A. Comparative analysis for Audio, Video and Text of existing methods with the proposed method.

The existing methods with an audio format for emotion recognition are gathered to compare with the EBF-DCNN model with constant training percentages of 90. The improved accuracy percentage of the EBF-DCNN model with the traditional methods of SVM, Xgboost, decision tree, LSTM, and DCNN is 14.46%, 19.15%, 17.58%, 18.00%, 5.96% and the precision is 18.26%, 14.02%, 15.95%, 18.61%, 8.60% similarly, the improved recall is 17.01%, 18.89%, 13.69%, 18.96%, 5.62% and the enhanced F1- score are 17.64%, 16.52%, 14.83%, 18.78% and 7.14%. The existing methods with visual format for emotion recognition are gathered to compare with the EBF-DCNN model with constant training percentages of 90. The improved accuracy percentage of the EBF-DCNN model with the existing methods of SVM, Xgboost, decision tree, LSTM, and DCNN is 14.85%, 10.21%, 14.78%, 10.78%, and 1.95% and the precision is 5.26%, 11.39%, 10.32%, 15.06%, 2.18% similarly, the improved recall is 12.89%, 14.09%, 13.36%, 15.31%, 2.00% and the enhanced F1- score are 9.23%, 12.76%, 11.87%, 15.18% and 2.09%. In this context, the comparative analysis of previous methods used only text features for emotion recognition, and the performance is calculated with constant training percentages of 90. The improved accuracy percentage of the EBF-DCNN model with the existing methods of SVM, Xgboost, decision tree, LSTM, and DCNN is 9.78%, 10.53%, 9.71%, 10.23%, 4.04%, and the precision is 6.05%, 15.40%, 11.26%, 8.03%, 4.60% similarly, the improved recall is 6.82%, 16.71%, 8.28%, 6.20%, 4.58% and the enhanced F1- score are 6.44%, 16.06%, 9.80%, 7.12% and 4.59%.

B. Time Complexity Analysis

In the time complexity analysis, the developed EBF-DCNN model demonstrates superior computational efficiency compared to traditional methods. With a computational time of 24.51 ms at 100 iterations, EBF-DCNN outperforms other approaches such as SVM **(**24.83 ms**)**, XGBoost **(**24.53 ms**)**, Decision Tree and LSTM (24.76 ms each), and DCNN **(**24.81 ms**).** This reduction in computational time highlights the EBF-DCNN model's effectiveness, making it the most efficient option in terms of time consumption, as summarized as shown in figure 3.

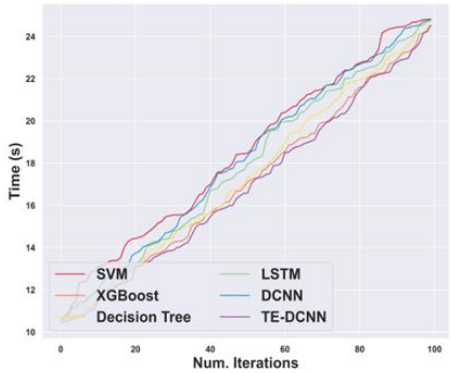

Figure 3. Time Complexity Analysis

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.6 Comparative discussion

The discussion of the EBF-DCNN model compared to several traditional models highlights its superior performance in emotion recognition, as shown in Table 2. The EBF-DCNN demonstrated higher accuracy and efficiency using the MELD dataset. This was largely due to the use of the TE-based fusion approach, which significantly enhanced the model's capability for easy emotion recognition. The fusion of audio, visual, and text features allowed for the selection of the least error values, further improving performance. Additionally, the model effectively addressed overfitting issues, reduced computational time, and achieved fast, accurate emotion detection.

Table 2: Comparative discussion table

Average of existing and proposed methods				
Methods vs. metrics	**Accuracy %**	**Precision %**	**Recall %**	**F1- Score %**
SVM	82.03	84.56	82.56	83.51
Xgboost	81.79	81.04	78.45	79.75
Decision tree	81.10	82.06	82.99	82.51
LSTM	82.06	80.77	81.39	81.07
DCNN	90.57	88.99	90.25	89.61
EBF-DCNN	94.33	93.80	94.08	93.94

5. CONCLUSION

Emotion recognition is a rising technology for recognizing human feelings, mostly adopted in healthcare research. This research significantly develops a multi-model-based EBF-DCNN method for emotion recognition. This model beats up the limitations of traditional methods by utilizing three features from the video such as visual, audio, and text. The feature extraction techniques for emotion recognition are a massive approach to wrench out the beneficial features and also provide an effective recognition process. The proposed EBF-DCNN model ensemble with DCNN as well as the empirical-based fusion (EBF) technique trains and fuses the three formats effectively. Here, EBF technique is a remarkable approach for active merging and highlighting the emotion recognition task. The performance of the proposed method solely provides high accuracy, precision, recall, and F1-Score of 94.33%, 93.80%, 94.08%, and 93.94% compared to other state-of-the-art methods. The advantages of the proposed model are highly reliable, reduces over-fitting problems, and reduces the computational cost and time. In the future, this work will be deployed in other efficient deep-learning models for better recognition.

REFERENCES:

[1] Shahzad, H. M., Bhatti, S. M., Jaffar, A., Rashid, M., & Akram, S. (2023). Multi-Modal CNN Features Fusion for Emotion Recognition: A Modified Xception Model. IEEE Access.
[2] Shahzad, H. M., Bhatti, S. M., Jaffar, A., & Rashid, M. (2023). A multi-modal deep learning approach for emotion recognition. Intell. At. Soft Comput, 36, 1561-1570.
[3] Wang, S., Qu, J., Zhang, Y., & Zhang, Y. (2023). Multimodal emotion recognition from EEG signals and facial expressions. IEEE Access, 11, 33061-33068.
[4] Mamieva, D., Abdusalomov, A. B., Kutlimuratov, A., Muminov, B., & Whangbo, T. K. (2023). Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. Sensors, 23(12), 5475.
[5] Ren, M., Huang, X., Liu, J., Liu, M., Li, X., & Liu, A. A. (2023). MALN: multimodal adversarial learning network for conversational emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology.
[6] Mocanu, B., Tapu, R., & Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image and Vision Computing, 133, 104676.
[7] S. Li and W. Deng, "Deep facial expression recognition: A survey,"arXiv:1804.08348, Oct. 2018.
[8] Y. Wang, L. Guan, An investigation of speech-based human emotion recognition, in: IEEE 6th Workshop on Multimedia Signal Processing, 2004, October, IEEE, 2004, pp. 15–18.
[9] Kołakowska A, Landowska A, Szwoch M, Szwoch W, Wrobel MR (2014) Emotion recognition and its applications. In: Human computer systems interaction: backgrounds and applications, pp 51–62.

Research Article

[10] Dubey M, Singh L (2016) Automatic emotion recognition using facial expression: a review. Int Res J Eng Technol (IRJET) 3:488.
[11] Z. Zeng , M. Pantic , G.I. Roisman , et al. , A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. In- tell. 31 (1) (2009) 39–58 .
[12] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508. 2018 Oct 5.
[13] Abebe HB, Hwang CL. RGB-D face recognition using LBP with suitable feature dimension of depth image. IET Cyber-Physical Systems: Theory & Applications. 2019 Sep;4(3):189-97.
[14] Chakraborti T, McCane B, Mills S, Pal U. LOOP descriptor: local optimal-oriented pattern. IEEE Signal Processing Letters. 2018 Mar 19;25(5):635-9.
[15] Jabid T, Kabir MH, Chae O. Local directional pattern (LDP) for face recognition. In2010 digest of technical papers international conference on consumer electronics (ICCE) 2010 Jan 9 (pp. 329-330). IEEE.
[16] Goel A, Srivastava SK. Role of kernel parameters in performance evaluation of SVM. In2016 Second international conference on computational intelligence & communication technology (CICT) 2016 Feb 12 (pp. 166-169). IEEE.
[17] Yao Z et al. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Commun 2020;120:11–9.
[18] Peng Z et al. “Speech emotion recognition using 3d convolutions and attentionbased sliding recurrent networks with auditory front-ends.” IEEE. Access 2020;8:16560–72.
[19] Jain M, Narayan S, Balaji P, Bhowmick A, Muthu RK. Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590. 2020 Feb 3.
[20] Parui S, Bajiya AK, Samanta D, Chakravorty N. Emotion recognition from EEG signal using XGBoost algorithm. In2019 IEEE 16th India Council International Conference (INDICON) 2019 Dec 13 (pp. 1-4). IEEE.
[21] Fernandes B, Mannepalli K. Speech Emotion Recognition Using Deep Learning LSTM for Tamil Language. Pertanika Journal of Science & Technology. 2021 Jul 1;29(3).
[22] Noroozi F, Sapiński T, Kamińska D, Anbarjafari G. Vocal-based emotion recognition using random forests and decision tree. International Journal of Speech Technology. 2017 Jun;20(2):239-46.
[23] Salmam FZ, Madani A, Kissi M. Facial expression recognition using decision trees. In2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV) 2016 Mar 29 (pp. 125-130). IEEE.
[24] MELD dataset: https://www.kaggle.com/datasets/zaber666/meld-dataset/data (Accessed on May 2024).