

Deep Multimodal Fusion Convolutional Neural Network for Emotion Recognition

Shailesh Kulkarni^{1, 2}, S.S. Khot³, Yogesh Angal⁴

¹Department of Electronics and Telecommunication, JSPM'S Rajarshi Shahu College of Engineering, Tathawade, Pune, Maharashtra, India, kulkarnishaileshece@sanjivani.org.in

²Department of Electronics and Computer Engineering, Sanjivani College of Engineering, Kopergaon, Maharashtra, India

³Department of Electronics and Telecommunication, K J College of Engineering and Management Research, Pune, Maharashtra, India, drkhotss@gmail.com

⁴Department of Electronics and Telecommunication, JSPM'S Bhivarabai Sawant Institute of Technology and Research, Wagholi, Pune Maharashtra, India, ysangal_entc@jspmbsiotr.edu.in

Corresponding Author: Shailesh Kulkarni; kulkarnishailesh1001@gmail.com

How to cite this article: Shailesh Kulkarni, S.S. Khot, Yogesh Angal (2024) Deep Multimodal Fusion Convolutional Neural Network for Emotion Recognition. *Library Progress International*, 44(3), 12381-12391.

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 trial-and-error-based fusion (TE-DCNN) for emotion recognition. The proposed TE-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 a trial-and-error-based (TE) fusion method to fuse three data formats, which is highly feasible to avoid over-fitting problems. Here, the TE-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 TE-DCNN model can be evaluated by five metrics such as accuracy, specificity, precision, recall and F1 score, and achieved 94.33%, 94.58%, 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, trial-and-error based fusion, multi-modal, deep learning.

1. Introduction:

Emotion recognition is a nascent technology that is widely used to identify people's feelings such as happiness, anger, sadness, and so on automatically using either speech, text or image. During communication, emotion plays a crucial role, in which the person captures emotion from the face and judges based on the analysis. Human resource information (HRI) faces a major problem in accurately detecting the emotions of humans and this detection came across quite many challenges [16].

The wild emotions are a more unconstrained and more extensive approach which was quite challenging to detect and predict the emotions [6] using the traditional approaches. The traditional approaches showed inefficiency of emotion recognition, thus the deep learning approach overcame the limitations and trend in multimodal emotion recognition [5].

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 [12] [2]. The emotion recognition process reflects major fields like medicine,

social media platforms, organizations, communication, and so on [13] [3].

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 [14] [15].

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 (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) are more efficient methods for emotion recognition using speech signals [27]. Other DL techniques like three-dimensional CNN attention sliding recurrent neural network (ASRNN) methods [28] [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.

2. Literature Review

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.

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.

Ram Avtar Jaswal and Sunil Dhingra [7] developed a multimodal emotion recognition using CNN by observing the input from EEG and audio. The noise is reduced in this model by pre-processing the signal. This approach suffers from overfitting and requires more space to train. Jiahui Pan, et al [8] developed a deep learning-based multimodal emotion recognition (MER) called Deep - Emotion. In this research for face expression, a GhostNet network is used, for the speech branch LFCNN is employed, and for the EEG branch a tree-like LSTM (tLSTM) is used. However, effective learning requires more training data and computationally it is costly.

In this work, the TE-DCNN model is developed for easy emotion recognition, which effectively overcomes the limitations of traditional approaches and, therefore gains significant results for recognition.

The research for emotion recognition is organized according to the following sequence: Section 2 contains Proposed Methodology, section 3 contains working mechanism of the proposed model and the results and discussions are presented in Section 4. Finally, the conclusion takes part in section 5.

3. Proposed Methodology:

The research aims at recognizing the emotions of an individual with multi-features using a multimodal distributed architecture-based TE-DCNN. Initially, the data are gathered from the MELD [37] dataset, which only holds video and the collected data are pre-processed, to reduce unnecessary background noises, cleans, transforms, and integrates the video for better emotion recognition. The pre-processed video is classified into audio, visual, and text features which are extracted using the feature extraction process. In this process, the audio feature is extracted using statistical features, where the visual feature is extracted using LBP, LDP, and LOOP channels and lastly, the text feature is derived using a graph embedding process. The extracted features are deployed to the TE-DCNN model, which automatically trains and fuses the model effectively for emotion recognition. The TE-DCNN model reduces time-consuming and over-fitting issues and thus improves the recognition process. In addition, the model also claims the advantage of robust fusion with high-accuracy performance. The TE-DCNN methodology is schematically shown in Figure 1.

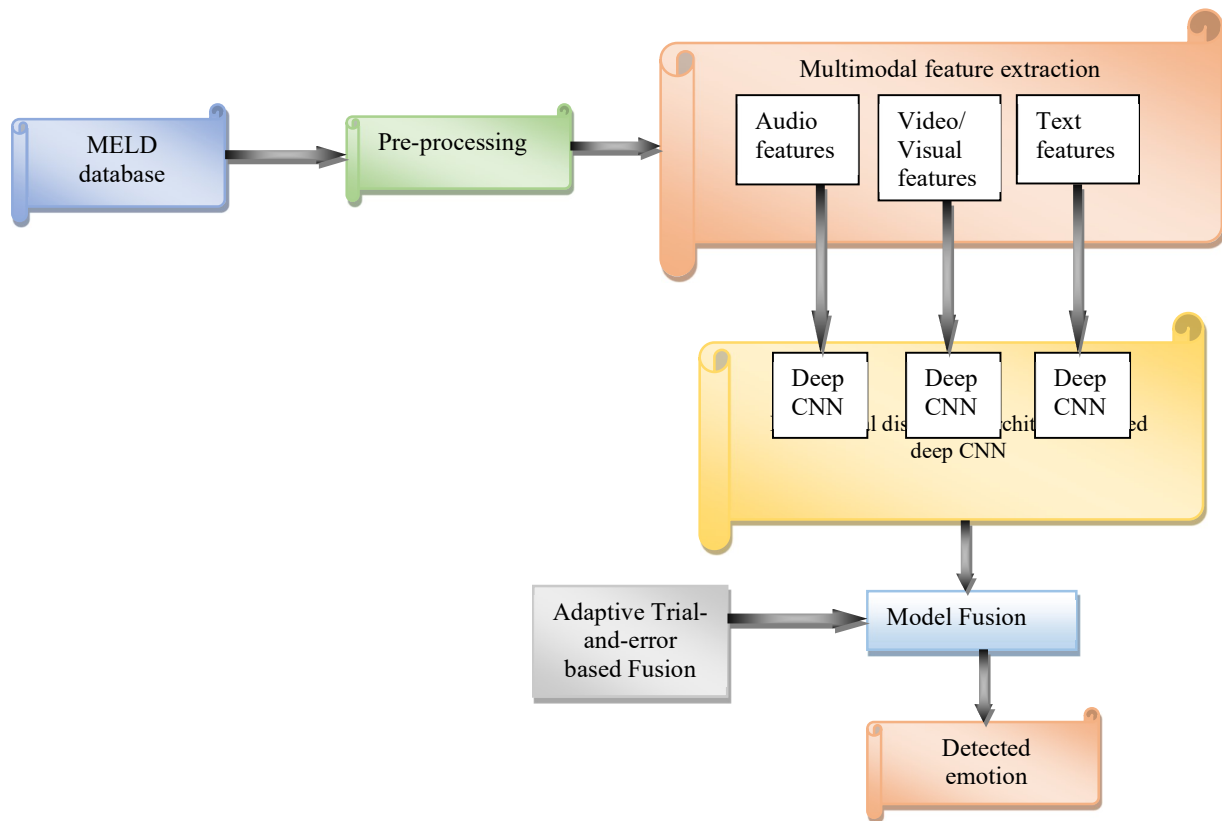


Figure 1: Proposed methodology for emotion recognition

In this research, the MELD dataset is taken, which holds video-based input for emotion recognition. The MELD dataset [19] 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 labeled with multiple emotions and can be defined as,

$$D = [V_1, V_2, V_3, \dots, V_j, \dots, V_n] \quad (1)$$

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 [21] 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.

$$D = [P_1, P_2, P_3, \dots, P_j, \dots, P_n] \quad (2)$$

the video frames are converted into audio files, and text format for emotion recognition using multiple statistical features.

3.1 Text format feature extraction

The key features of the text are extracted using the graph embedding technique, where the extracted features are mapped with their necessary features and hence converted into vector nodes with low dimensional vector space. The graph can be constructed using the Term Frequency-Inverse Document Frequency (TF-IDF) technique to

yield more information regarding the input text for emotion recognition.

The TF-IDF is a weight feature that, immensely collaborates with both the local parameter and the global parameter and is likely to convert text data into a vector space model (VSM) [24]. Thus the cosine similarity and the linear kernel graphs are taken from the TF-IDF for the graph embedding process and the graph embedding is shown in Figure 3.

The resemblance of the text vector interval can be calculated between $[-1, 1]$, 1 shows the cosine similarity of two constant ratios of the vector, 0 represents the cosine similarity between the orthogonal vector and lastly, -1 defines the similarity between the opposite vector.

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

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 [26].

$$L(Z_v, Z_w) = Z_v^T Z_w \quad (3)$$

here, v, w are the input vectors of the text features.

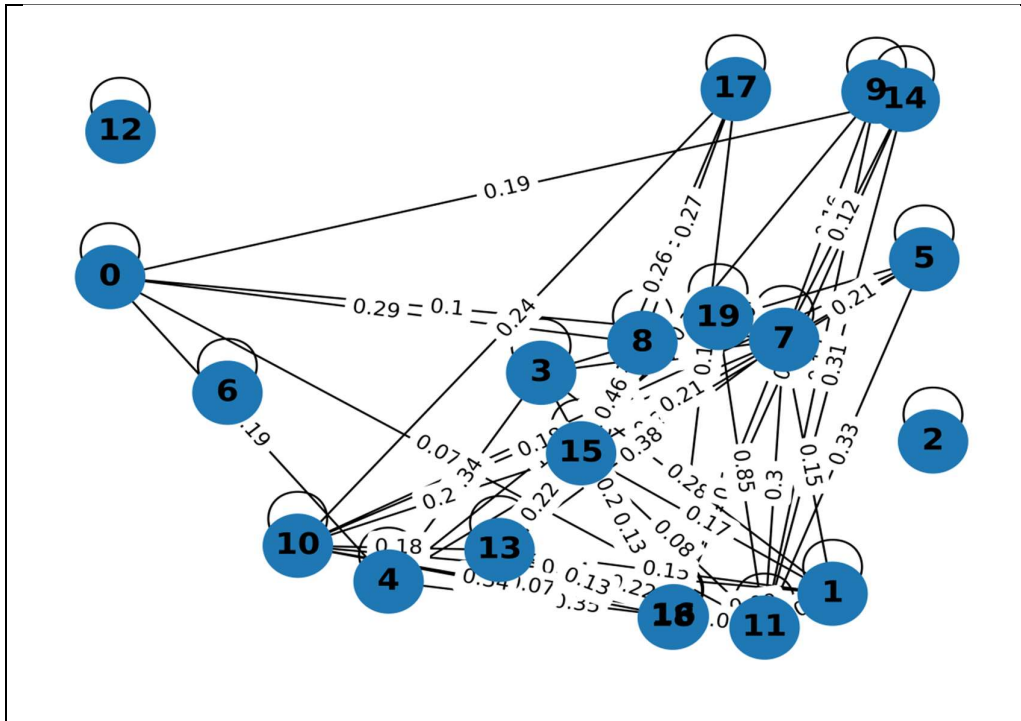


Figure 3: Graph embedding

4. Multimodal distributed architecture based on TE-DCNN

The multimodal distributed architecture is more effective and more viable than the single modal architecture for all kinds of emotion recognition including text-based, visual-based, and audio-based. The extracted features from visual, audio, and text are directly moved toward the three deployed DCNN models which consequently train the model. In this phase, the audio signals, visual, and text features can be extracted and effectively classified for the development and labeling of the model for emotion recognition. Each deep CNN consists of several processes including a 1D convolution layer for audio and text features, a 2D convolution for visual features, a pooling

process, as well as a fully connected (FC) layer, which is then fused using the adaptive TE-based Fusion method. The developed model enhanced the best solution for emotion recognition by proving the optimal values and therefore minimizes computational expenses. The proposed deep CNN comes up with high accuracy for emotion recognition, which is quite feasible for using large video sequence data, and reduces over-fitting problems. In addition, provides highly reliable performances and an easy recognition process. The architecture of the proposed model is shown in the below figure 4.

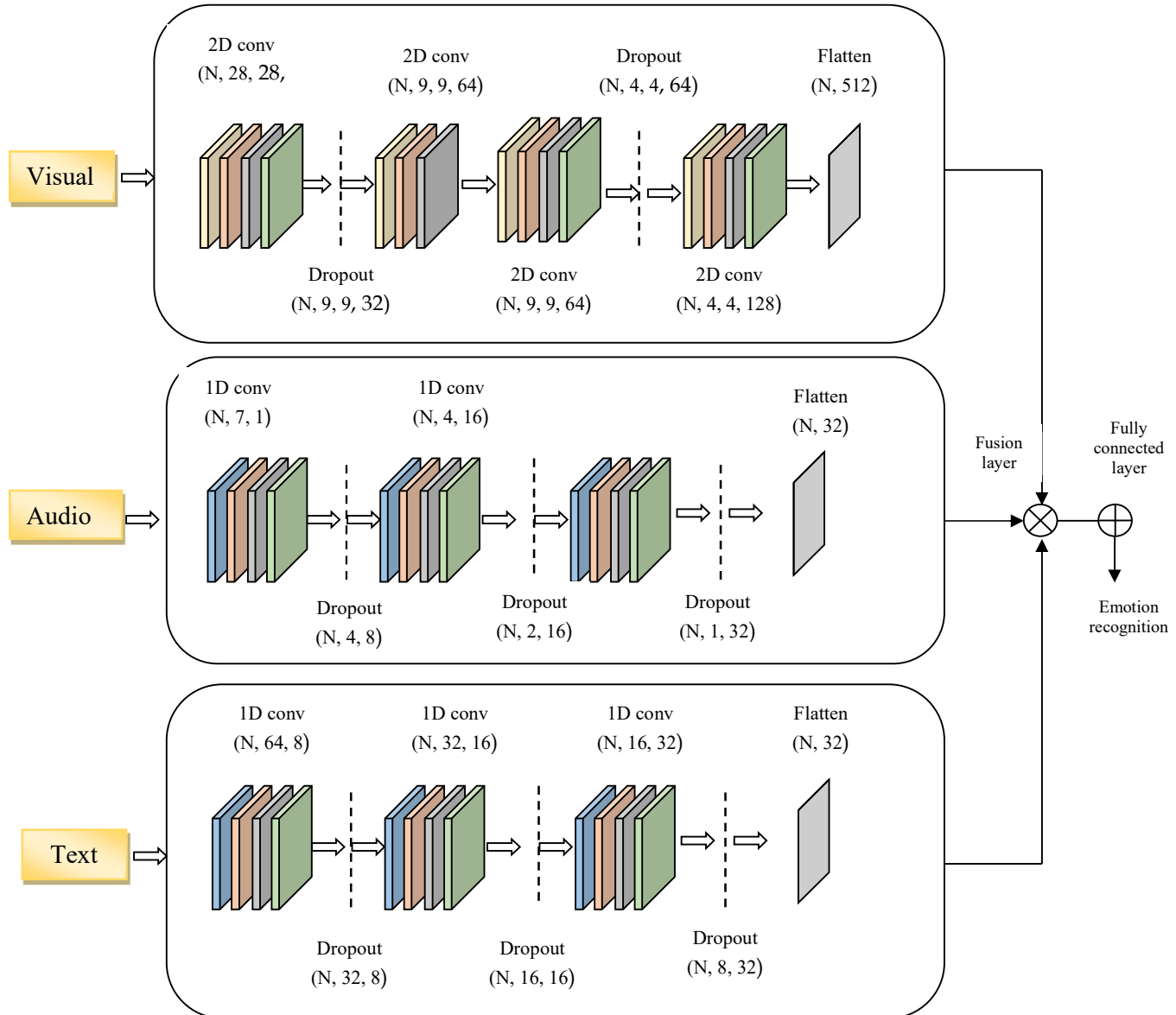


Figure 4: Architecture of the proposed model

5. Analysis and Interpretation

This experiment provided enhanced emotion recognition using the TE-DCNN model and achieved high accuracy compared to other traditional models the discussion of the existing methods is described thoroughly in the below sections.

5.1 Experimental setup

The experiment for emotion recognition is conducted using Windows 11 OS, RAM-16 GB, ROM-100 GB, and employed in the Python 3.7.6 platform. The hyperparameter details are depicted in Table 1.

Table 1: Hyper Parameter Details


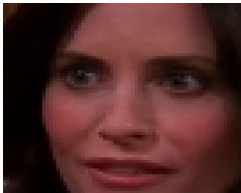



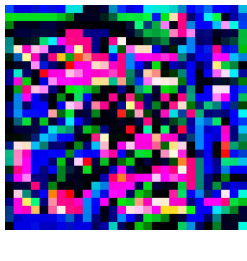
| Hyper Parameter Details | |
|-------------------------|-------------------------------|
| Optimizer | ADAM |
| Drop Out Rate | 0.5 |
| Learning Rate | 0.01 |
| Batch Size | 32 |
| Loss Function | Mean Square Error |
| Activation Function | ReLU |
| Metrics | Accuracy, Mean Absolute Error |

5.2 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 labeled 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

5.3 Experimental analysis

The experiment analysis shows the results of visual extraction from the MELD dataset for emotion recognition using the proposed TE-DCNN model which is illustrated in Figure 5.

| Method vs. dataset | MELD dataset | | |
|------------------------------------|---|--|---|
| Input | Video input | | |
| Visual extraction |  |  |  |
| Hybrid multi-level ternary pattern |  |  |  |

5.4 Performance metrics

The performance of the proposed model is evaluated using the following metrics.

a) Accuracy: The accuracy performance of the TE-DCNN model can be calculated using the below equation.

$$A_{cc} = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (4)$$

b) Sensitivity: The sensitivity of the TE-DCNN model can be evaluated using the below equation.

$$S_{en} = \frac{Tp}{Tp + Fn} \quad (5)$$

c) **Specificity:** The specificity of the TE-DCNN model can be measured using the below equation.

$$S_{pe} = \frac{Tn}{Tn + Fp} \quad (6)$$

here, Tp, Tn, Fp, Fn are truly positive, true negative, false positive, and false negative respectively. The above equation is used for performance analysis and comparative analysis to find the better performance of the models.

5.5 Performance analysis

The TE-DCNN model performance can be analyzed using the several above-mentioned metrics and shows how the TE-DCNN model is efficient for emotion recognition and provides high accuracy, which are discussed in the below context. The performances are analyzed using the MELD dataset with a maximum training percentage of 90.

5.5.1 Performance analysis of the accuracy of the proposed model

In this analysis, the performance of the TE-DCNN model using the MELD dataset shows quite high accuracy and achieved better performances for emotion recognition which is represented in Table 2. In this context, the performance of the TE-DCNN with a maximum training percentage of 90 and with varying epochs 100, 200, 300, 400, and 500 provide 88.81%, 91.39%, 93.35%, 93.91%, and 94.33% of accuracy. At the same time, the specificity is 86.54%, 91.36%, 94.19%, 94.23%, and 94.58%, similarly, the precision of the TE-DCNN model is 88.67%, 90.61%, 91.97%, 92.09% and 93.80%, furthermore, the recall of the TE-DCNN model is 91.07%, 91.43%, 92.51%, 93.58%, and 94.08% and the F1-Score of the TE-DCNN model is 89.86%, 91.01%, 92.24%, 92.83%, and 93.94%. This analysis proves that the TE-DCNN model achieved higher performance than the other models, using only the image-based or audio-based dataset. This research reduces the over-fitting problem because of the proposed unique fusion technique and increases the convergence speed for effective recognition. In addition, this technique is highly feasible and robust to train the model.

Table 2: Performance analysis of the TE-DCNN model

| | Accuracy % | Specificity % | Precision % | Recall % | F1 Score % |
|------------------------|---------------|------------------|----------------|-------------|---------------|
| TE-DCNN at Epoch = 100 | 88.81 | 86.54 | 88.67 | 91.07 | 89.86 |
| TE-DCNN at Epoch = 200 | 91.39 | 91.36 | 90.61 | 91.43 | 91.01 |
| TE-DCNN at Epoch = 300 | 93.35 | 94.19 | 91.97 | 92.51 | 92.24 |
| TE-DCNN at Epoch = 400 | 93.91 | 94.23 | 92.09 | 93.58 | 92.83 |
| TE-DCNN at Epoch = 500 | 94.33 | 94.58 | 93.80 | 94.08 | 93.94 |

5.5.2 Comparative analysis for Audio of existing methods with the proposed method.

The existing methods with an audio format for emotion recognition are gathered to compare with the TE-DCNN model with constant training percentages of 90. The improved accuracy percentage of the TE-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 improved specificities are 11.91%, 19.40%, 21.45%, 17.04%, 6.29% whereas, the precision is 18.26%, 14.02%, 15.95%, 18.61%, 8.60% similarly, the improved recall are 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 differences are graphically represented in Figure 6. This comparison gradually signifies the TE-DCNN model's efficiency over recognition.

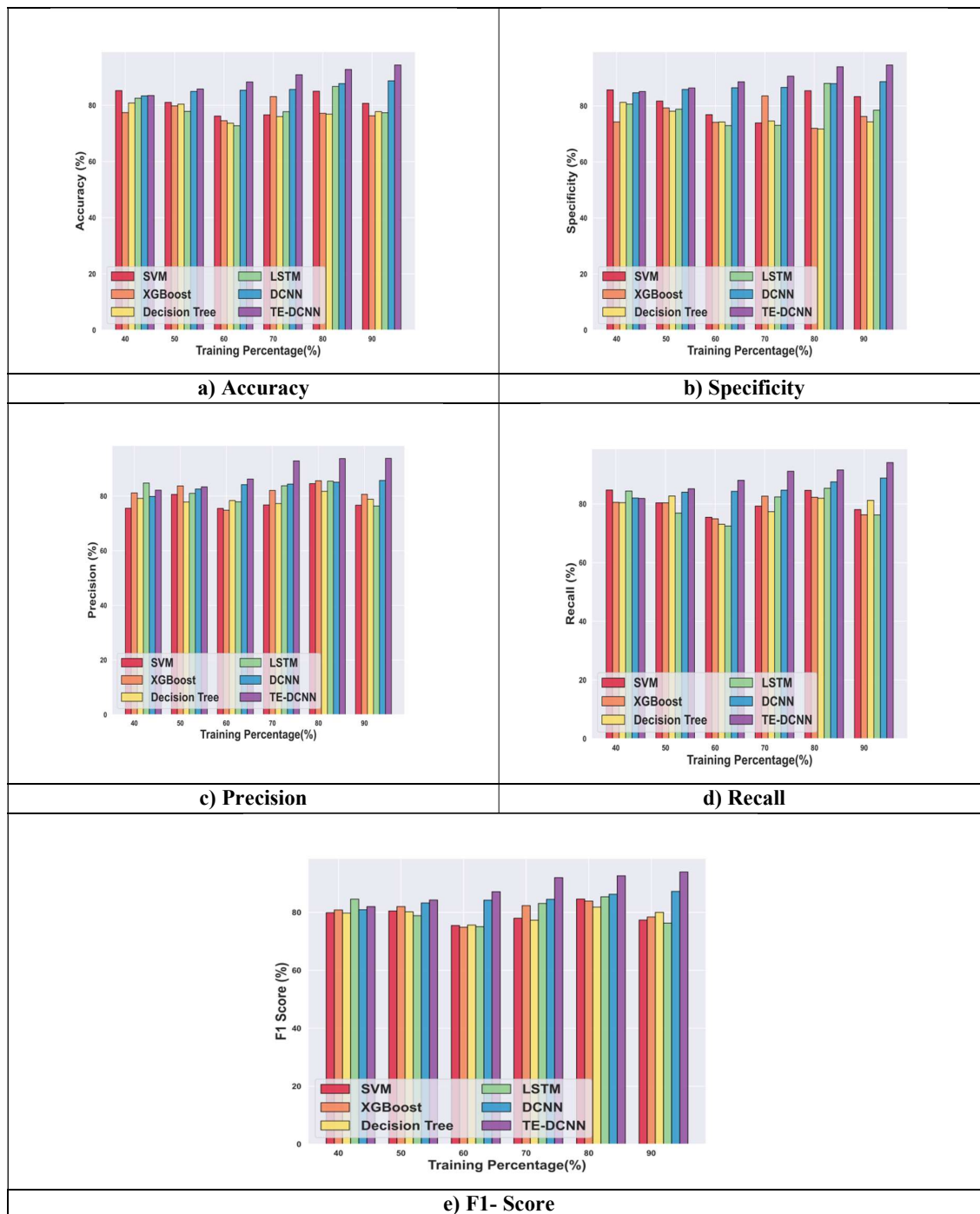


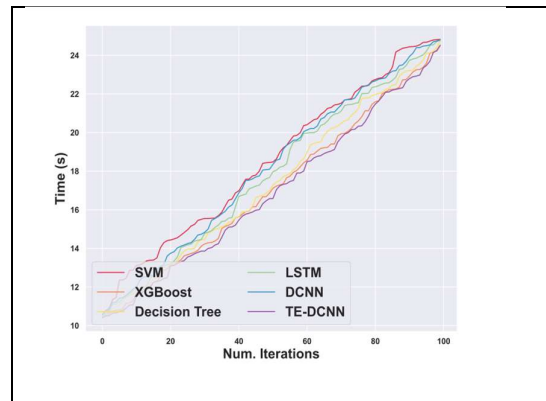
Figure 6: Graphical representation for comparative analysis

5.3 Time Complexity Analysis

The computational time comparison between the developed TE-DCNN with the other existing methods is analyzed with different iterations to show the effectiveness of the TE-DCNN method. The results show the method's computational efficiency, it consistently requires a reduction in time compared with other existing methods. While comparing the suggested method with the existing approaches, the developed model has the lowest computing time of 24.51 ms at iteration 100. Table 3 includes the information and the computational time analysis method is depicted in Figure 9.

Table 3: Comparative Analysis based on computational time

| Methods | Computational Time |
|---------------|--------------------|
| SVM | 24.83 |
| XG Boost | 24.53 |
| Decision Tree | 24.76 |
| LSTM | 24.76 |
| DCNN | 24.81 |
| TE-DCNN | 24.51 |


Figure 9: Time Complexity Analysis

5.4 Comparative discussion

The discussion of the TE-DCNN model along with several traditional models is illustrated in this context. The discussion made that the proposed TE-DCNN model showed quite high performance for easy emotion recognition with increased accuracy compared to other traditional approaches. The TE-DCNN model achieved 94.33% of accuracy, 94.58% of specificity, 93.80% of precision, 94.08% of recall, and 93.94% of F1 Score using the MELD dataset. Besides, the model signifies the advantage of using a TE-based fusion approach that highly intensifies the model efficiency as well as makes it possible for easy recognition. Moreover, the TE-DCNN gains estimation of three features namely audio, visual, and text therefore selects two least error values for emotion recognition. This approach is highly feasible, more modern, and provides quite easy recognition. In addition, the model reduces the over-fitting issues and increases computational speed to reduce time consumption. Moreover, the TE-DCNN model effectively recognizes the emotions of humans and also represented in Table 4.

Table 4: Comparative discussion table

| Average of existing and proposed methods | | | | | |
|--|------------|---------------|-------------|----------|-------------|
| Methods vs. metrics | Accuracy % | Specificity % | Precision % | Recall % | F1- Score % |
| SVM | 82.03 | 81.52 | 84.56 | 82.56 | 83.51 |
| Xgboost | 81.79 | 85.08 | 81.04 | 78.45 | 79.75 |
| Decision tree | 81.10 | 79.20 | 82.06 | 82.99 | 82.51 |
| LSTM | 82.06 | 82.73 | 80.77 | 81.39 | 81.07 |
| DCNN | 90.57 | 90.89 | 88.99 | 90.25 | 89.61 |
| TE-DCNN | 94.33 | 94.58 | 93.80 | 94.08 | 93.94 |

6. Conclusion

Emotion recognition is a rising technology for recognizing human feelings, mostly adopted in healthcare research. This research significantly develops a multi-model-based TE-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 TE-DCNN model ensemble with DCNN

as well as the TE-based fusion technique trains and fuses the three formats effectively. Here, the TE-based fusion technique is a remarkable approach for active merging and highlighting the emotion recognition task. The performance of the proposed method solely provides high accuracy, specificity, precision, recall, and F1-Score of 94.33%, 94.58%, 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] Jaswal, R. A., & Dhingra, S. (2023). Empirical analysis of multiple modalities for emotion recognition using convolutional neural network. *Measurement: Sensors*, 26, 100716.
- [8] Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z., & Wang, S. (2023). Multimodal emotion recognition based on facial expressions, speech, and EEG. *IEEE Open Journal of Engineering in Medicine and Biology*.
- [9] Abdullah, Sharmeen M. Saleem Abdullah, Siddeeq Y. Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. "Multimodal emotion recognition using deep learning." *Journal of Applied Science and Technology Trends* 2, no. 02 (2021): 52-58.
- [10] A.M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 International Conference on Platform Technology and Service (PlatCon), IEEE, 2017, pp. 1–5.
- [11] E. Sariyanidi, H. Gunes and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113-1133, Jun. 2015.
- [12] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv:1804.08348*, Oct. 2018.
- [13] 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.
- [14] Kołakowska A, Landowska A, Szwoch M, Szwoch W, Wrobel MR (2014) Emotion recognition and its applications. In: *Human computers systems interaction: backgrounds and applications*, pp 51–62.
- [15] Dubey M, Singh L (2016) Automatic emotion recognition using facial expression: a review. *Int Res J Eng Technol (IRJET)* 3:488.
- [16] 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 .
- [17] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial Expression Recognition in Videos Using Dynamic Kernels," *IEEE Transactions on Image Processing*, vol. 29, pp. 8316-8325, 2020.
- [18] S. Bateman and S. Ameen, "Comparison of algorithms for use in adaptive adjustment of digital data receivers," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 137, pp. 85-96, 1990.
- [19] 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.
- [20] Ravi R, Yadukrishna SV. A face expression recognition using CNN & LBP. In *2020 fourth international conference on computing methodologies and communication (ICCMC)* 2020 Mar 11 (pp. 684-689). IEEE.
- [21] 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.
- [22] 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.
- [23] Jabid T, Kabir MH, Chae O. Local directional pattern (LDP) for face recognition. In 2010 digest of technical papers international conference on consumer electronics (ICCE) 2010 Jan 9 (pp. 329-330). IEEE.
- [24] Umadevi M. Document comparison based on tf-idf metric. International Research Journal of Engineering and Technology (IRJET). 2020;7(02):1546-50.
- [25] Gholamalinezhad H, Khosravi H. Pooling methods in deep neural networks, a review. arXiv preprint arXiv:2009.07485. 2020 Sep 16.
- [26] Goel A, Srivastava SK. Role of kernel parameters in performance evaluation of SVM. In 2016 Second international conference on computational intelligence & communication technology (CICT) 2016 Feb 12 (pp. 166-169). IEEE.
- [27] 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.
- [28] Peng Z et al. “Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends.” IEEE. Access 2020;8:16560–72.
- [29] Sarvakar K, Senkamalavalli R, Raghavendra S, Kumar JS, Manjunath R, Jaiswal S. Facial emotion recognition using convolutional neural networks. Materials Today: Proceedings. 2023 Jan 1;80:3560-4.
- [30] Singh P, Sahidullah M, Saha G. Modulation spectral features for speech emotion recognition using deep neural networks. Speech Communication. 2023 Jan 1;146:53-69.
- [31] 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.
- [32] Parui S, Bajiya AK, Samanta D, Chakravorty N. Emotion recognition from EEG signal using XGBoost algorithm. In 2019 IEEE 16th India Council International Conference (INDICON) 2019 Dec 13 (pp. 1-4). IEEE.
- [33] Fernandes B, Mannepalli K. Speech Emotion Recognition Using Deep Learning LSTM for Tamil Language. Pertanika Journal of Science & Technology. 2021 Jul 1;29(3).
- [34] 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.
- [35] Salmam FZ, Madani A, Kissi M. Facial expression recognition using decision trees. In 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV) 2016 Mar 29 (pp. 125-130). IEEE.
- [36] Mayya V, Pai RM, Pai MM. Automatic facial expression recognition using DCNN. Procedia Computer Science. 2016 Jan 1;93:453-61.
- [37] MELD dataset: <https://www.kaggle.com/datasets/zaber666/meld-dataset/data> (Accessed on Ma