INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Emotionally Intelligent System for Multimodal Sentiment Classification

R Suganya[1]*, **M Narmatha**[1], **S Vengatesh Kumar**[1]

1 Department of Computer Science with Cyber Security, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, 641049, Tamil Nadu, India

OPEN ACCESS

Received: 21-07-2024

Accepted: 24-10-2024

Published: 20-11-2024

Citation: Suganya R, Narmatha M, Kumar SV (2024) An Emotionally Intelligent System for Multimodal Sentiment Classification. Indian Journal of Science and Technology 17(42): 4386-4394. https://doi.org/10.17485/IJST/v17i42.2349

* **Corresponding author**.

rsugu1983@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop a multimodal sentiment classification model by analyzing the impact of biological signals and examining the concatenation of various modalities in a marketing scenario. **Methods:** This paper proposes a new emotionally intelligent system for multimodal sentiment classification. Initially, a multimodal database is prepared by collecting text, speech, facial expression, posture, and biological signals for each individual in the user-machine interaction scenario. This database is preprocessed to remove unwanted noise or missing values. After preprocessing, the dataset is split into training and testing sets. The training set is then fed into the feature extraction phase which involves different methods to extract various types of features like texture, color, acoustic, linguistic, and biological signals. These features are independently trained by the Multi-Modal Deep Belief Network (MMDBN) model to generate the trained classifier. The trained classifier is later used to classify the test sets as different sentiment classes like positive, negative, or neutral. The effectiveness of the proposed MMDBN model is evaluated in MATLAB 2019b using Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) and Wearable Stress and Affect Detection (WESAD) datasets. Each dataset is divided into 70% for training and 30% for testing. A comparative study is conducted by applying existing sentiment classification models on the datasets to evaluate MMDBN prediction efficiency using metrics like accuracy, precision, recall, and F-Score. **Findings:** The MMDBN model achieves 80.28% and 81.17% accuracy on IEMOCAP and WESAD databases compared to the existing sentiment classification models like Deep Multi-View Attentive Network (DMVAN), Multi-Channel Multimodal Joint Learning Method (MCMJLM) and Multi-Level Textual-Visual Alignment and Fusion Network (MTVAF) and Attention-Based Multimodal Sentiment Analysis and Emotion Recognition (AMSAER). **Novelty:** This study introduces a novel emotional intelligence system for multimodal sentiment classification, integrating biological signals and other modalities to enhance accuracy in marketing by combining linguistic, acoustic, visual, and biological data.

Keywords: Sentiment classification; Biological signals; Multimodal database; Deep Learning; Deep belief network

1 Introduction

Sentiment defines a physiological tendency, i.e., a proclivity to have a specific form of emotional state (e.g., positive or negative)[1]. Sentiment categorization is the process of identifying the emotional tone of a text, such as whether it is good, negative, or neutral[2]. With the growing number of online platforms that include text, images, and videos, the current research has focused on the efficacy of multimodal categorization[3]. Many academics have suggested a variety of multimodal sentiment categorization models that use both Machine Learning (ML) and Deep Learning (DL) methods.

Filali et al.[4] developed a Meaningful Neural Network (MNN) for multimodal emotion recognition. This model extracts text and audio features and fuses them as bimodal features. These features were fed into MNN to learn each modality separately for sentiment prediction. Al-Tameemi et al.[5] ^{dev}eloped DMVAN for multimodal sentiment and emotion classification. It extracts multi-view visual and textual features using a multi-head attention mechanism and multi-layer perceptron for improved classification performance. Gong et al.[6] developed an MCMJLM for multimodal sentiment analysis. This model employs a multi-channel feature extraction strategy for image and text features and employs an adaptive multi-task fusion method for sentiment prediction. Liu et al.[7] introduced a Multimodal Fusion Network model using Transformer Encoder (MFNTE) to refine and unify features from different modalities. This model captures complementary features and integrates them with initial features for sentiment analysis through residual connections.

Li et al.[8] developed the MTVAF for sentiment analysis. MTVAF converts images into content, face, and text descriptions and extracts the text and visual information for prediction. Also, the network aligns probability distributions between textual and combined textual-visual inputs were observed to reduce noise. Prashant[9] developed a White-Headed Bird (WHB) optimization and a Bidirectional Long Short-Term Memory (BiLSTM) classifier for multimodal sentiment and emotion classification. Aslam et al.[10] introduced the AMSAER framework, which classifies sentiment and emotions using intra-modal features and inter-modal associations across visual, audio, and textual domains. This framework extracts semantic words, image areas, and audio features.

Table 1 provides a summary of the existing multimodal sentiment classification models, highlighting their advantages and limitations.

Table 1. Summary of the Existing Multimodal Sentiment Classification Models

Author [Ref No]	**Methods**	**Advantages**	**Disadvantages**
Filali et al.[4]	MNN	Ability to create improved word representations for rare or unseen words.	Inability to manage intricate patterns in lengthy texts.
Al-Tameemi et al.[5]	DMVAN	Faster training times and reduced computational overhead	It generates high predictive errors.
Gong et al.[6]	MCMJLM	Finely Extracts potential syntactic and semantic features for sentimental analysis	Hyperparameters are not optimized well.
Liu et al.[7]	MFNTE	It can handle complex, nonlinear relationships between input features and output labels.	Limited clarity in the decision-making process.
Li et al.[8]	MTVAF	Initializes both local and global data to create word vectors for analyzing sentiment.	Incapacity to manage unfamiliar or out-of-vocabulary terms.
Prashant[9]	WHB - BiLSTM	Effectively navigates dialects and informal or slang words.	Failing to generalize new or varied datasets, needs domain adaptation for precise outputs.
Aslam et al.[10]	AMSAER	It regulates the flow of information entering and exiting the network.	Accuracy may be limited with small datasets that lack enough training examples

1.1 Research Gap

Existing frameworks for multimodal sentiment classification primarily utilize text, images, and videos, often neglecting the effective use of noticeable signals. While some studies have investigated biomedical signals for emotion recognition, the impact of short-duration signals in user-machine interactions remains unclear. Additionally, the integration of biomedical signals with other data types has not been thoroughly examined. To address these gaps, this article presents a novel emotionally intelligent system for multimodal sentiment classification. The model employs various modalities, including text transcriptions, speech utterances, video (capturing facial expressions and posture), and biological signals from individuals during interactions.

1.2 Major Contributions

This paper presents a novel emotionally intelligent system called MMDBN for multimodal sentiment categorization. The major contributions of this study are the following:

- A multimodal database is first established by gathering text, speech, facial expressions, posture, and biological signals from individuals in marketing contexts (e.g., sales and customer satisfaction).
- Next, the collected dataset undergoes pre-processing to eliminate noise and address missing values, enhancing model training.
- Various feature extraction methods are employed, including Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams for linguistic feature extraction, Electrodermal Activity (EDA) for physiological feature extraction, and the OpenSMILE and OpenFace tools for audiovisual feature extraction.
- The extracted features are then fed into the MMDBN model, which comprises multiple Deep Belief Networks (DBNs) to learn the features and create a trained classifier.
- Finally, the trained classifier is utilized to categorize the test data as positive, negative, or neutral.

Thus, the proposed model accurately classifies an individual's sentiment using text, speech, facial expressions, posture, and biological signals for multimodal sentiment prediction, making it useful for customer feedback analysis and marketing strategies.

2 Methodology

In this section, the presented multimodal sentiment classification model is explained briefly (as depicted in Figure 1). Initially, a multimodal database is created by collecting the different modalities such as text transcriptions, speech utterances, video (face expressions and posture), and biological signals for a group of individuals during conversion. The collected data are preprocessed to remove noise or missing values. Then, various feature extraction methods are performed to extract the linguistic, acoustic, visual, and physiological signal characteristics. Such extracted characteristics are independently learned by the multiple DBNs and the score from each DBN is fused to classify sentiment classes as positive, negative, or neutral.

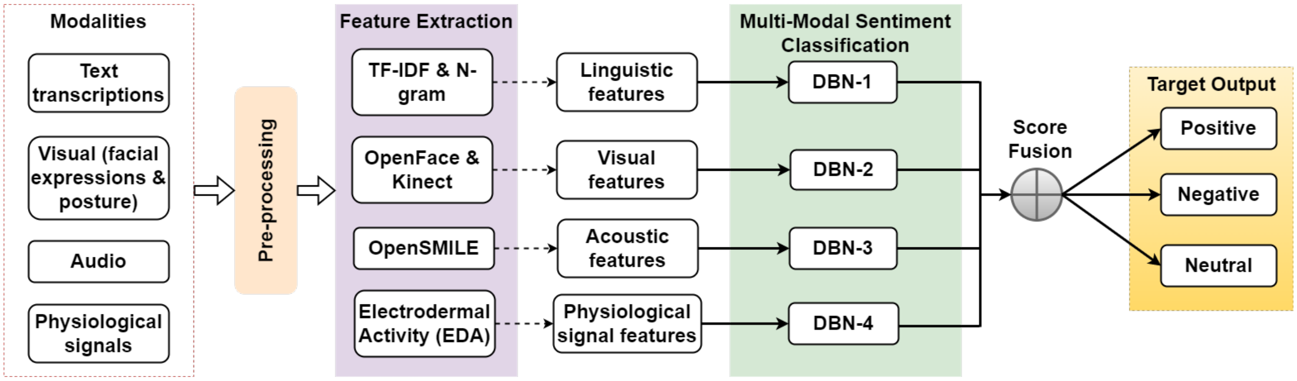

Fig 1. An Overview of Presented Multimodal Sentiment Classification System

2.1 Dataset Collection

In this study, a database containing different modalities is obtained from the available resources which are listed below.

IEMOCAP [11] :It is obtained from the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). It contains about 12 hours of audiovisual information such as video, speech, face, and text transcriptions. Ten actors in dyadic sessions with markers on the face, head, and hands recorded this corpus, providing comprehensive information on their facial expressions and hand motions in both planned and unscripted conversation situations. These actors executed specific sentimental drafts and also unscripted hypothetical situations developed to find certain classes of sentiments (positive [Joy, Happy], negative [Depressed, Sad, Angry, and Frustration], and neutral conditions).

WESAD [12] :It is a dataset created to allow study into emotion and stress detection using physiological inputs from wearable devices. The dataset contains a variety of physiological parameters, including ECG, EMG, EDA, heart rate, blood volume pressure, and triaxial acceleration signals sampled at 700Hz, which were gathered from people throughout diverse emotional states.

2.2 Data Pre-processing

After obtaining each database, different preprocessing techniques are applied to eliminate noises and address the missing values for improving the model's training in multimodal sentiment prediction. For noise cancellation, median filtering techniques are applied to retain the crucial features. Also, median imputation models are used to address the missing values. These pre-processing models are critical to improving data quality and guaranteeing accurate analysis in both IEMOCAP and WESAD datasets.

2.3 Feature Extraction

After preprocessing, various feature extraction techniques are applied to extract the linguistic features, physiological features, and audiovisual features.

2.3.1 Extraction of Linguistic Features

To extract the linguistic features, the Term Frequency-Inverse Document Frequency (TF-IDF) and N-gram feature extraction methods are used. TF-IDF is a popular scheme to analyze the significance of a word in a database. The Term Frequency (TF) of a specific term (t) is determined by dividing the number of occurrences of the term in a document by the total number of words in the document. IDF (Inverse Document Frequency) is employed to quantify the significance of a particular term. The IDF is determined as IDF(t)=log(N/DF), where N is the number of transcriptions and DF is the number of transcriptions with t. Similarly, N-gram can extract the features of a text. This is a series of n tokens extracted from the provided text.

2.3.2 Extraction of Physiological Features

The time series physiological signals are separated into the Skin Conductance (SC) level and the SC response. Polynomial fitting is used to determine standard SC, which represents the overall function of the sweat glands as a result of the ambient temperature. Then, PeakUtils8 was used to find the GSR. Finally, a physiological feature called GSR number per exchange is retrieved. Furthermore, various statistical measures such as the mean, standard deviation, maximum and minimum values, mean of the first and second differences, range, skewness, kurtosis, slope and intercept of the linear approximation, and values for the 25th and 75th percentiles of the SC in each exchange were computed and employed as physiological attributes. So, physiological features are extracted and those are standardized using min-max regularization.

2.3.3 Extraction of Audiovisual Features

Using the widely popular OpenSMILE tool, the voice features are retrieved from each utterance's sound automatically. Various elementary attributes and associated statistical data are extracted as features using the OpenSMILE tool. The extraction process also includes characteristics like amplitude mean, arithmetic mean, quadratic mean, standard deviation, flatness, skewness, kurtosis, and quartiles. Those derived features comprise the different acoustic sub-features: prosodic features, energy metrics, voice likelihoods, spectral attributes, and cepstral characteristics Meanwhile, speed and acceleration at each location are estimated for facial feature extraction using the OpenFace library to identify facial landmarks surrounding the eyes, mouth, and brows. The Microsoft Kinect sensor's dynamic metrics for the hands, shoulders, and head are also taken into account, determined speed and acceleration are used as dynamic attributes. So, the visual features are also derived where all audio-visual features are standardized by the Z-score regularization scheme.

Thus, the features from various modalities are extracted and concatenated to learn by the different classifiers and classify them into specific sentiment classes.

2.4 Classification

In this stage, the concatenated features of multiple modalities are learned by the different classifiers for sentiment analysis. Such classification processes are described below.

SVM Classifier: The multimodal concatenated features are directly fed to SVM, which is constructed as a hyperplane in a multi-dimensional space. Appropriate sentiment classification is accomplished by the hyperplane that owns the primary space for the nearby learning features of some label i.e., functional margin. The learning set is denoted as a collection of feature-label pairs (x_i, y_i) , $i = 1, \ldots, n$, $x_i \in R^n$, $y_i \in \{-1, +1\}$, where x_i refers to the concatenated feature vectors and y_i refers to labels.

For various labels, solving the below-unconstrained optimization problem yields the optimal hyperplane with the maximum margin:

$$\underset{w}{min}\frac{1}{2}w^T w+F\sum_{i=1}^{n}\xi\left(w;x_i,y_i\right) \tag{1}$$

In Equation (1), F>0 refers to the penalty parameter and w refers to the weight of learning features x_i . By handling this optimization issue, emotions are classified into positive, negative, and neutral.

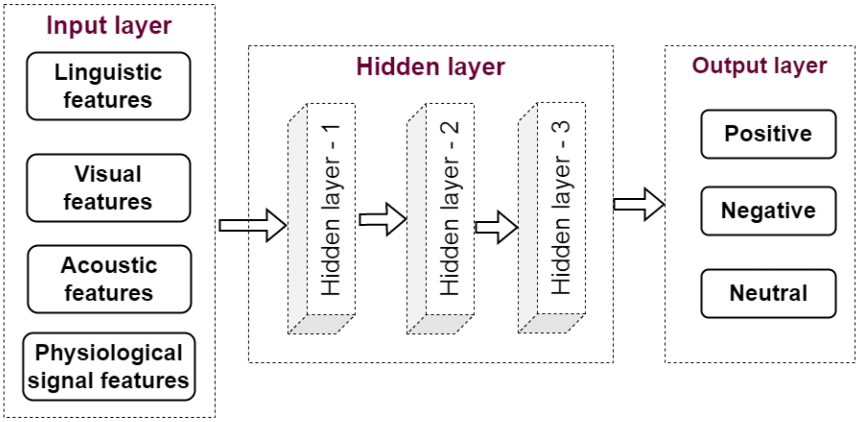

Fig 2. Structure of the DNN Model

DNN Classifier: The multimodal concatenated features are directly fed to the DNN, which contains an input layer, hidden layers, and an output layer (portrayed in Figure 2) to classify emotional states.

If the input feature x_i is given to DNN, the output of the hidden layer is defined as the tan-sigmoid transfer function:

$$f\left(x_i\right)=\frac{2}{1+e^{-2x_i}}-1 \tag{2}$$

The adder function takes in all of the features as input and uses their weight values $(w_1,w_2,\ldots,w_n)$ to calculate the weighted sum:

$$u=\sum_{i=1}^{n}w_i x_i \tag{3}$$

The output layer of DNN is defined as:

$$y=f\left(\sum_{i=1}^{n}w_i x_i+b_i\right) \tag{4}$$

In Equation (4), y denotes output neuron value; $f(\bullet)$ denotes transfer function, w_i denotes weight values, x_i refers to input features and b_i denotes the bias value. According to y, the given features are learned and classified into different emotional states: positive, negative, or neutral. The hyperparameters assigned for the DNN classifier are listed in Table 2.

Table 2. DNN Hyperparameters

Parameters	Values
Input layer neurons	4
First hidden layer neurons	50
Second hidden layer neurons	50
Third hidden layer neurons	20
Output layer neurons	1
Learning rate	0.001
Transfer function	Tan-Sigmoid
Maximum number of iteration	100

MMDBN Classifier: It is a probabilistic network made up of several Restricted Boltzmann Machines (RBMs). There are three layers in the network: output, hidden, and visible. Weights connect the visible and hidden layers, and each neuron has an offset that specifies its weight. Figure 3 illustrates the DBN structure with n hidden layers.

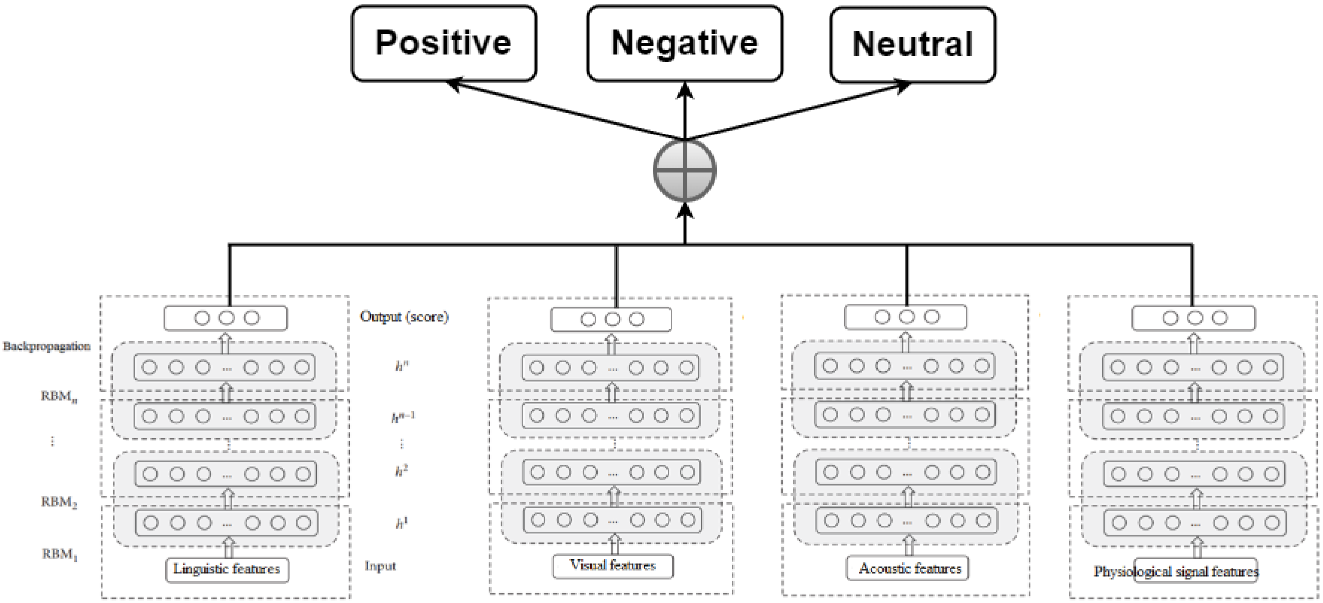

Fig 3. Structure of MMDBN Model

To change the key parameters of the hidden layer, the output and the previously hidden layers build a backpropagation NN. For DBN learning, data input is taken from the base layer and sent to hidden layers. The coefficients of connection and offsets among RBM layers are set before learning using an unsupervised greedy layer-by-layer strategy. After that, each RBM layer is individually taught from the base to the top. Let RBM be a power-produced Bernoulli model; the state's power (v, h) is:

$$\begin{cases} E(v,h;\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H}\omega_{ij}v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j, \\ \theta = \{\omega, a, b\} \end{cases} \tag{5}$$

In Equation (5), θ is the RBM parameter, ω_{ij} is the link weight between the visible and hidden layer neurons, V and H are the number of visible and hidden units, v_i and h_j are the node states of the visible and hidden layers, b_i and b_j are the offsets of the visible and hidden layers. In order to preserve sparsity, the visible layer offset b_j is initialized to $lb(\hat{p}_i/(1-\hat{p}_i))$, where $\hat{p}_i$ represents the probability of $v_i = 1$. The maximum positive number is used to initialize the hidden layer offset a_j while the minimum arbitrary value is used to initialize ω_{ij}. Additionally, the model's joint probability is ascertained as:

$$\begin{cases} P(v,h) = \frac{1}{Z}e^{-E(v,h)}, \\ Z = \sum_{v,h} e^{-E(v,h)} \end{cases} \tag{6}$$

In Equation (6), Z refers to the regularization factor. The probability of v_i and h_j is independent due to the absence of a link between the peer nodes:

$p(v_i = 1|h;\theta) = \sigma\left(\sum_{j=1}^{H}\omega_{ij}h_j + b_j\right),$

$$\begin{cases} P(v_i = 1|h;\theta) = \sigma\left(\sum_{j=1}^{H}\omega_{ij}v_i + a_j\right), \\ \sigma(x) = \frac{1}{1+e^{x}} \end{cases} \tag{7}$$

In Equation (7), $\sigma(x)$ denotes the sigmoid function. The boundary distribution of $p(v,h;\theta)$ to h is discovered as:

$$p(v;\theta) = \frac{1}{Z}\sum_{h} exp(-E(v,h;\theta)) \tag{8}$$

In Equation (8), θ is obtained by solving the maximum log probability prediction function on the training set and RBM parameter tuning criteria are obtained by contrast divergence scheme:

$$\begin{cases} \triangle b_i = \varepsilon(\langle v_i\rangle_{data} - \langle v_i\rangle_k), \\ \triangle a_j = \varepsilon(\langle h_j\rangle_{data} - \langle h_j\rangle_k), \\ \triangle b_i = \varepsilon(\langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_k), \end{cases} \tag{9}$$

In Equation (9), ε refers to the training efficiency, $\langle . \rangle_{data}$ and $\langle . \rangle_k$ are the expected values of the distribution that the present and the reconstructed framework represent. Parameters learned via pre-learning do not have optimal primary values since it is unsupervised training. At this point, the backpropagation NN is combined with the label for modifying the parameters to solve high-outcome errors.

This backpropagation NN is located in the output layer of DBN and top-down execution is used for supervised learning. In order to maximize DBN's classification ability, the link parameters among all layers are adjusted according to Equation (6). Thus, multiple DBNs, i.e. four independent DBNs for four distinct features are trained and their scores are fused to classify the classes of emotions. Table 3 presents the hyperparameters used for constructing the DBN classifier.

Table 3. Design of DBN Framework and Training Parameters

Parameters	Range
Number of neurons in the Input Layer	4
Number of Neurons in RBM1	100
Number of Neurons in RBM2	100
Number of Neurons in RBM3	100
Number of Neurons in Output Layer	1
Batch Size	15
Learning Rate	0.01
Momentum Value	0.05
Maximum Iteration	100

3 Results and discussion

This section evaluates the effectiveness of the proposed sentiment classification method using the MMDBN model, implemented in MATLAB 2019b. The experiment includes a brief overview of the IEMOCAP and WESAD datasets, as detailed in Section 2.1. In this analysis, 70% of the data is allocated for training, and the remaining 30% is designated for testing. Additionally, a comparative analysis is conducted using various existing classifier models, including DMVAN(5) MCMJLM(6), MTVAF(8), and AMSAER(10) on the collected dataset. This comparison aims to assess prediction performance based on metrics such as accuracy, precision, recall, and F-measure.

- **Accuracy:** It represents the ratio of correct predictions to the total number of data points analyzed.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + False\ Negative\ (FN)} \tag{10}$$

In Equation (10), the count of positive labels accurately predicted as positive is represented, while TN denotes the number of negative labels accurately predicted as negative. Conversely, FP refers to the count of positive labels incorrectly classified as negative, and FN represents the count of negative labels incorrectly classified as positive.

- **Precision:** It evaluates the proportion of accurately predicted labels among the total predicted positives, considering both TP and FP rates.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

- **Recall:** It reflects the proportion of correctly predicted labels out of the total actual positives, accounting for both TP and FN rates.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

- **F-score:** This is the harmonic mean of precision and recall, providing a balanced measure of both metrics.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

From the analyses provided in Figure 4, it is indicated that the MMDBN classifier on IEMOCAP databases achieves better efficiency compared to the other classifier models for classifying the emotion states from multiple modalities. The precision of the MMDBN classifier is 40.24%, 32.58%, 22.76%, and 11.79% higher than the DMVAN, MCMJLM, MTVAF and AMSAER models. The recall of the MMDBN model is 40.34% higher than the DMVAN, 32.7% higher than the MCMJLM, 22.67% higher than the MTVAF, and 11.34% higher than the AMSAER classifiers. Also, the f-score of the MMDBN classifier is 40.3%, 32.65%, 22.71%, and 11.56% increased than the DMVAN, MCMJLM, MTVAF, and AMSAER models. Similarly, the accuracy of the MMDBN model is 39.35% higher than the DMVAN, 31.93% higher than the MCMJLM, 20.71% higher than the MTVAF, and 10.78% higher than the AMSAER classifiers for sentiment classification.

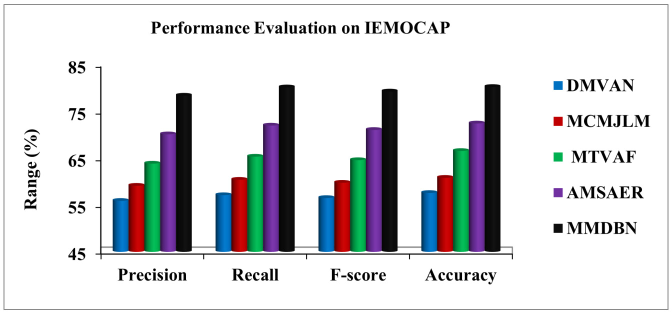

Fig 4. Comparison of Different Multimodal Sentiment Classification Algorithms on IEMOCAP database

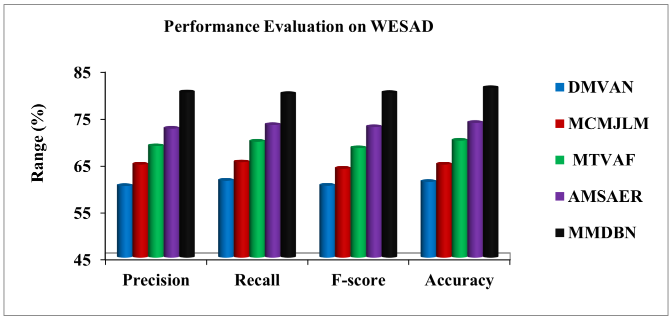

Fig 5. Comparison of Different Multimodal Sentiment Classification Algorithms on the WESAD database

From the analyses provided in Figure 5, it is indicated that the MMDBN classifier on WESAD databases achieves better efficiency compared to the other classifier models for classifying the emotion states from multiple modalities. Additionally, the precision of the MMDBN model is increased by 11.4%, 7.58%, 5.7%, and 2.6% in comparison to the DMVAN, MCMJLM, MTVAF, and AMSAER models respectively. Similarly, the recall of the MMDBN model is 11.06%, 7.66%, 5.5%, and 2.7% higher than the DMVAN, MCMJLM, MTVAF, and AMSAER models respectively. The F-score of MMDBN model is 11.23%, 7.62%, 5.62%, and 2.65% higher than the DMVAN, MCMJLM, MTVAF and AMSAER models respectively Furthermore, MMDBN model demonstrates higher accuracy compared to the DMVAN, MCMJLM, MTVAF and AMSAER with improvements of 10.49%, 7.76%, 5.17%, and 2.69% respectively.

From the above analysis, it is observed that the proposed MMDBN models outperform other classification models. This is due to the probabilistic network, made up of multiple RBMs with visible, hidden, and output layers which connect the visible and hidden layers through weights, while neuron biases influence these weights. This enhances the model's efficiency in multimodal sentiment classification

4 Conclusion

This study presents a new emotionally intelligent paradigm for classifying the emotional states from multiple modalities. A multimodal dataset was preprocessed to remove noise and null values. Linguistic, visual, acoustic, and physiological features were then extracted and fed into the MMDBN model to classify unknown emotion data into positive, negative, and neutral categories. The experimental results proved that the presented system achieves 80.28% and 81.17% accuracy on IEMOCAP and WESAD databases compared to the other sentiment classification models. This model can be applied in various fields, particularly in marketing scenarios, to improve the accuracy of sentiment classification by leveraging the combination of linguistic, acoustic, visual, and biological data. This major limitation is that these model struggles to learn the contextual emotion features resulting in low accuracy. Future work will incorporate pre-trained networks to capture contextual semantics and long-dependency relations in the text from the multimodal database.

References

1) Bordoloi M, Biswas SK. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial intelligence review*. 2023;56(11):12505–12560. Available from: https://doi.org/10.1007/s10462-023-10442-2.
2) Lai S, Hu, Xu H, Ren Z, Liu Z. Multimodal sentiment analysis: A survey. *Displays*. 2023;80:1–15. Available from: https://doi.org/10.1080/08839514.2024.2371712.
3) Pandey A, Vishwakarma DK. Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Applied Soft Computing*. 2024;152:111206–111206. Available from: https://doi.org/10.1016/j.asoc.2023.111206.
4) Filali H, Riffi J, Boulealam C, Mahraz MA, Tairi H. Multimodal emotional classification based on meaningful learning. *Big Data and Cognitive Computing*. 2022;6:95–95. Available from: https://doi.org/10.3390/bdcc6030095.
5) Feizi-Derakhshi ATI, Pashazadeh MR, Asadpour S, M. Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data. *IEEE Access*. 2023. Available from: https://doi.org/10.1109/ACCESS.2023.3307716.
6) Gong L, He X, Yang J. An Image-Text Sentiment Analysis Method Using Multi-Channel Multi-Modal Joint Learning. *Applied Artificial Intelligence*. 2024;38(1):1–20. Available from: https://doi.org/10.1080/08839514.2024.2371712.
7) Liu C, Wang Y, Yang J. A transformer-encoder-based multimodal multi-attention fusion network for sentiment analysis. *Applied Intelligence*. 2024;54(17):8415–8441. Available from: https://doi.org/10.1007/s10489-024-05623-7.
8) Li Y, Ding H, Lin Y, Feng X, Chang L. Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. *Artificial Intelligence Review*. 2024;57(4):1–26. Available from: https://doi.org/10.1007/s10462-023-10685-z.
9) Huddar MG, Sannakki SS, Rajpurohit VS. Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. . *Multimedia Tools and Applications*. 2021;80:13059–13076. Available from: https://doi.org/10.1007/s11042-020-10285.
10) Aslam A, Sargano AB, Habib Z. Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. *Applied Soft Computing*. 2023;144:110494–110494. Available from: https://doi.org/10.1016/j.asoc.2023.110494.
11) Guo H, Ga Z. Multimodal sentiment recognition based on Bi-LSTM and fusion mechanism. *Academic Journal of Computing & Information Science*. 2023;6(6):127–132. Available from: https://doi.org/10.25236/AJCIS.2023.060620.
12) Pradhan A, Srivastava S. Hybrid DenseNet with long short-term memory model for multi-modal emotion recognition from physiological signals. . *Multimedia Tools and Applications*. 2024;83:35221–35251. Available from: https://doi.org/10.1007/s11042-023-16933-2.