Multimodal Emotion Recognition using Bi-LG-GCN for MELD Dataset

Hussein Farooq Tayeb Al-Saadawi, Resul Daş

Abstract—**Emotion recognition through multimodal data can significantly enhance human interactions. We introduce the Multimodal Emotion Lines Dataset (MELD) and present a novel method, Bi-LG-GNN, which capitalizes on diverse emotion labeling across text, audio, and visuals. This research emphasizes detecting concealed affective computing states within textual and audio data, contributing to emotion recognition and sentiment analysis. The quality and consistency of the data undergo improvements via meticulous pre-processing techniques, which encompass noise elimination, normalization, and linguistic adjustments. These interventions specifically aim to address linguistic discrepancies and reduce background noise in the discourse. To extract salient features, we employ the Kernel Principal Component Analysis (K-PCA), aiming to derive meaningful attributes from each modality and to encode labels for array values. We introduce a Bi-LG-GCN-based architecture meticulously designed for multimodal emotion recognition, which effectively amalgamates data from diverse modalities. This Bi-LG-GCN system interprets the enhanced multi-modal data representations, producing synthetic samples that capture multimodal relationships. As a result, it allows for precise emotion recognition and prediction on multimodal datasets. When tested on the MELD dataset, the results were remarkable, showing accuracy (80%), F1-score (81%), precision (81%), and recall (81%). The implemented pre-processing and feature extraction phases substantially elevate the quality and discrimination of input representations. Our Bi-LG-GCN approach, accentuated by its ability to synthesize multimodal data, surpasses existing methods, showcasing its significant practical value.**

Index Terms—**Bimodal emotion recognition, text and speech recognition, Multimodal Emotion Lines Dataset (MELD), Bilateral gradient graph convolutional network (Bi-LG-GCN), Affective computing identification.**

I. Introduction

Emotions are central to how humans interact. With more multimedia data available, our understanding of computer recognition of these emotions has greatly improved. Expression of emotions, whether written or spoken, deeply influences behavior and markedly affects decision-making, learning, and cognitive functions. In contemporary periods, the domain of artificial intelligence has strived to develop platforms proficient in comprehending and expressing emotions, progress majorly influenced by innovations in affective computing. Audio and written content integration has evolved, enhancing user experience and comprehension. As a result, dealing with singular modes of texts without this combination has become even more tedious and intricate [1]. The integration of audio and written content has improved, enhancing user comprehension and experience. Consequently, handling text without this combination has become more cumbersome and complex, underscoring the importance of bimodal content [2]. Challenges persist in communication via speech, including variability in the precision of emotional speech detection and a shortage of foundational understanding centered on emotions. Such dimensions are quintessential for articulating intent, interpreting social cues, and nurturing profound connections. Traditionally, emotion discernment predominantly relied upon facial lines [3]. Humans employ many means for emotional conveyance, encompassing gestures, postural indications, tonal variations, and facial nuances. A comprehensive approach that contemplates these diverse indicators is essential for precise emotion discernment and assimilation [4]. Stemming from this understanding, multimodal emotion recognition has emerged as a captivating field. It involves aggregating and examining emotional indicators from multiple channels. A more holistic and accurate representation of human emotions can be achieved by amalgamating information from varied sources, such as facial cues, vocal rhythms, physical gestures, and physiological responses. This interdisciplinary domain, converging research from affective computing, voice analytics, computer visualizations, and machine intelligence, endeavors to formulate algorithms and models adept at adeptly processing and interpreting multimodal emotional stimuli [5].

Improving human-computer interaction is an important factor behind multimodal emotion detection. To provide more individualized and interesting encounters, emotionally intelligent systems can modify their behavior and replies according to the user's emotional state [6]. Similarly, multimodal emotion recognition may be used in healthcare applications to assess a patient's emotional health and provide important insights for individualized care and treatment. Multimodal emotion recognition faces a variety of difficulties. It is a challenging effort to integrate and synchronize data from several modalities, each with distinct properties and noise sources [7]. Deep learning algorithms have been developed to mix data from many modalities efficiently. Because emotional displays can vary greatly across people, creating solid models that can generalize across diverse people and cultural backgrounds is another difficulty. Further obstacles come from the subjectivity and context-dependence of emotions, which individual experiences may impact.

While GCNs, or Graph Convolutional Networks, have proven

Hussein Farooq Tayeb Al-Saadawi is with the Department of Software Engineering, Technology Faculty, Firat University, Elazig, 23119 TÜRKİYE. Email: husseinftayeb@gmail.com

Resul Daş is with the Department of Software Engineering, Technology Faculty, Firat University, Elazig, 23119 TÜRKİYE. Email: rdas@frat.edu.tr

Manuscript received October 6, 2023; accepted October 16, 2023.
DOI: 10.17694/bajece.1372107

indispensable for visually understanding intricate systems through graphs, the importance of textual data in human interactions cannot be overlooked, especially in the realm of sentiment analysis. Graphs, as outlined by the rising popularity of infographics and knowledge graphs, can convey information efficiently[8]. On the other hand, textual data plays a significant role in sentiment analysis, as evidenced by the work of Mucahit Pinar Savci et al.[9]. Their research, which focused on evaluating deep learning and machine learning methods using e-commerce corpora in different languages, emphasized the power of pre-trained language models in discerning various sentiments. As for emotion extraction from speech, Pulatov et al., [10] present a notable advancement by employing dual feature extraction techniques, combining CNNs, MFCC, and Speech2Vec. While their methodology displays impressive accuracy enhancements on two renowned datasets, the omission of 10-fold cross-validation might prompt questions on generalizability. Nevertheless, their findings, suggesting the potential integration of diverse features and modalities, significantly enrich the current understanding of Speech Emotion Recognition research. Therefore, while graphical representation simplifies our comprehension of complex systems, textual data analysis remains crucial for understanding human emotions and intentions. Future studies might explore how integrating these two domains—graphical and textual—can bring more holistic insights across various sectors. Multimodal emotion identification has a lot of potential despite the difficulties in several areas. It may benefit from affective computing, human-robot interaction, virtual reality, gaming, market research, and psychology. Researchers and practitioners may design intelligent systems that are more sympathetic, intuitive, and responsive by analyzing and interpreting the intricate interplay of emotional signals [11]. There are many different ways that people might express and interpret their emotions. It tries to give a comprehensive knowledge of human emotions by including a variety of modalities. By developing emotionally intelligent systems, the discipline has the potential to transform human-computer interaction, healthcare, and other industries completely. We proposed a bi-lateral gradient graph convolutional network (Bi-LG-GCN) method to recognize the emotions of human beings.

A. *Contributions*

- We provide the MELD dataset, which was utilized for accurate emotion categorization on both textual and audio modalities.
- To improve the multimodal dataset's quality and consistency, we use pre-processing techniques (linguistic and normalization).
- We apply K-PCA as a feature extraction method to get specific characteristics from each modality.
- To assess the effectiveness of our suggested strategy, we carry out thorough metrics such as accuracy, precision, recall, and f1-score using the multimodal dataset.

The following sections of the paper are written as follows: Section 2 contains reviews of the relevant literature. Section 3 explains the proposed approach, Section 4 discusses the findings, and Section 5 presents the conclusion.

II. Background

The field of research into understanding human emotions is rapidly growing and attracting the attention of both industries as it delves deep beyond facial expressions and body language to explore how emotions resonate within the nuances of our spoken words. Emotion, an intricate aspect of human communication, is pivotal in conveying sentiments, thoughts, and intentions. Our speech encodes information through elements like text meaning, speech rate, rhythm, pitch, volume, voice quality, articulation, duration, and inflection. Accordingly, the research objectives outlined of [12], a novel multi-modal approach for recognizing human emotions was introduced. The proposed method employed a deep learning architecture, specifically the “3D-Convolutional Neural Network (3D-CNN),” to extract spatiotemporal features from both electroencephalogram (EEG) signals and facial recordings. Then, a mix of data enhancement and ensemble learning approaches was presented to obtain the final fusion projections. The suggested scheme's multi-modality fusion was accomplished using data and score fusion approaches. The “3D-CNN output characteristics of the face chunks were then classified using the Support Vector Machine (SVM) classifier”. In a parallel vein, wi and Liang [13] proposed a fusion-based approach to detect emotions in affective speech. This method integrated acoustic-prosodic information and semantic labels with multiple classifiers.

The text serves as a cornerstone in emotion recognition, crucial for enhancing human interaction within affective computing systems across languages. Mucahit Soylu et al.,[14] delved into attitude markers (AMs) in academic writing by English and Turkish authors. They identified ’significance’ as a key functional category and ’adjective’ as the top form. Their innovative approach, incorporating Java for data cleaning and a radial knowledge graph for visualization, offers promising avenues for emotion recognition in diverse linguistic landscapes. A key component of their methodology was the MDT technique, which is responsible for selecting the most suitable classifier based on recognition confidence. This work, although focused primarily on audio and semantic information, highlighted the effectiveness of using multi-modal data and sophisticated classifiers for enhanced emotion recognition. They emphasized the synergistic effects of combining distinct data types, achieving an impressive recognition accuracy of up to 85.79% when considering the personality traits of speakers. Building on the exploration of acoustic features, Jin et al. [15] centered on the fusion of features from both acoustic and lexical levels to enhance emotion recognition in speech. At the acoustic level, a gamut of features, including intensity, F0, and others, was extracted, with novel representations such as Gaussian supervectors introduced. The lexical dimension saw the proposal of the ”emotion vector” (eVector) feature, which was predicated on emotion lexicons, giving words weights based on their emotion-expressive inclination. Applied to the USC-IEMOCAP database, the study demonstrated that the late fusion of both these feature realms culminated in a four-class emotion recognition accuracy of 69.2%. This further solidified the potential of combining various feature sets to hone the precision of emotion recognition, marking it as a noteworthy contribution to the field. The researchers in [16] introduced a new approach for multimodal emotion recognition utilizing raw waveforms and cross-modal attention mechanisms. By integrating raw audio processing through a one-dimensional convolutional model and establishing a cross-modal attention network between audio and text features, they optimized their emotion detection system. This approach showed that features extracted from raw audio, when coupled with a cross-modal attention mechanism, can effectively capture the interplay between audio and textual cues, enhancing emotion classification capabilities.

The investigators of [17] presented an innovative emotion detection system according to various expressions and other modalities of the face, Galvanic Skin Response (GSR), and electroencephalogram (EEG). Emotional Analysis Databases for variable numbers of emotional classes, four on a median, including angry, disgusted, fearful, joyful, neutral, sad, and confused [18]. In natural deceptive facial expressions, the suggested model demonstrated the benefit of detecting the proper state of emotions. In the study by [19], a method for multi-modal emotion detection was introduced, which harnesses speech-visual correlation features. This approach utilized two-dimensional convolutional neural networks (2D-CNN) to extract speech characteristics and 3D-CNN for visual attributes. During multi-modal fusion, a feature correlation analysis technique was employed to analyze both audio and visual data. Results from experiments performed on diverse datasets show that the approach is comparable to other cutting-edge algorithms in terms of recognition effectiveness. The focus is on deploying deep learning methodologies to discern underlying emotional tones in textual data. A notable advancement in this domain is the Emo2Vec model, which, when integrated with Logistic Regression and GloVe, demonstrates competitive performance in emotion detection. Ragheb et al. have made a significant contribution by developing a learning-based model that recognizes six emotions as described by Paul Ekman; their method employs a two-phase approach of encoding and classification, utilizing tokenization, encoders, and Bi-LSTM units trained via average stochastic gradient descent [20]

According to Lian et al., [21], they proposed “domain adversarial neural networks (DANN) for emotion detection. The main objective was to forecast emotion categories and develop a common representation where speaker’s identity could not be discriminated”. The depictions of different speakers were closer together thanks to their use of this technique. Using unlabeled data during the training phase, they reduced the effect of low-resource sample collections. The researchers of [22] introduced a deep learning (DL) based method for accessing and combining text and perceptual data to classify emotions. To extract acoustic characteristics from unprocessed audio, they added a DCNN layer after using a SincNet layer based on a customizable sinc value with band-pass filtering. When compared to performing convergence over the raw voice signal, the method develops filter banks modified for the recognition of emotions and provides superior aspects. To infer N-gram level association on hidden models derived from the Bi-RNN, they used two parallel streams for text analysis “(a DCNN and a Bidirectional RNN followed by a DCNN)” with cross attentiveness. The researchers of [23] demonstrated M3ER, a learning-based approach for emotion identification from several input modalities. Their system incorporated inputs from many co-occurring models and is more resistant to any of the sensors are noisy various approaches than previous methods. “M3ER model an innovative, data-driven multiplicative fusion strategy” for combining the models, which learns to enhance the extra accurate cues. M3ER was resistant to sensor noise by incorporating a check step that used Canonical Correlational Assessment to distinguish between ineffective and effective models.

Liu et al., in their study [24], proposed utilizing deep canonical correlation analysis (DCCA) for recognizing emotions across multiple modalities. DCCA primarily functions by independently transforming each modality and then correlating the varied modalities into a unified hyperspace, adhering to predefined canonical statistical constraints. The efficacy of DCCA was assessed across five different datasets. The research results revealed that DCCA obtained state-of-the-art detection accuracy rates across all datasets. Mittal et al. introduced ”EmotiCon,” a learning-based system designed for identifying perceived emotions from photos and video clips by considering context [25]. This approach incorporates three different perspectives of context for emotion detection. The initial interpretation was based on the use of various senses for emotion identification. In the second analysis, they obtained semantic information from the input image and used a self-attention-based CNN” to encode the information. At last, depth maps were employed to simulate the third interpretation, which was connected to socio-dynamic interactions and agent proximity. They showed the effectiveness of their network by running tests on datasets. The researchers [26] suggested an innovative cross-representation speech model for detecting emotions on “wav2vec 2.0 voice characteristics and also trained a CNN-based model to distinguish emotions from text data extracted with Transformer-based models. A score fusion method was used to merge the speech-based and text-based findings”.Authors of [27] developed a method for multimodal emotion identification named “deep generalized canonical correlation analysis with an attention mechanism (DGCCA-AM).” The model established multimodal adaptive fusion with a focus system, extending the usual “canonical correlation analysis (CCA) from two modalities” to arbitrarily numerous modalities. According to Zhang [28] suggested a deep automated encoder-based expression-EEG interaction multi-modal recognition of emotions approach. In the beginning, a decision tree was used as a feature-based selection approach. The solution vector values were then examined to establish the expression classification for the test sample based on the facial expression characteristics detected by sparse representation. The bimodal deep automated encoder was then used to combine the EEG. The 3rd layer of BDAE extracts characteristics for the supervised training phase. To complete the categorization, the LIBSVM classifier was employed.

An approach to identifying age, gender, and state of mind from audio was developed by Zaman et al. [29]. All audio recordings in our system were transformed into 20 statistical attributes, and the transformed numerical dataset was utilized to build several prediction models in order to achieve the goal. “Kneatest neighbors (KNN), XGBoost, AdaBoost, and Decision Tree, Artificial neural networks (ANN), Naive Bayes, and Support vector machine (SVM).” The authors of I will schedule some time for us to connect.[30] suggested an innovative emotion-relevant crucial subnetwork identification method and examined 3 EEG functional connectivity network features. On three open emotion EEG databases, the identifying capacity of the EEG connectivity characteristics in emotion detection was examined. By single-channel analysis, the strength feature surpassed the state-of-the-art differential entropy feature in the accuracy of classification. The findings from the study demonstrated that each of the five emotions had different functional conjunction characteristics. The researchers [31] proposed a deep multi-task learning system that analyzed emotions simultaneously. A video’s multi-modal inputs provide unique and distinct information, and they typically do not contribute equally to decision-making. They suggested a context-level inter-modal attention method for evaluating an utterance’s sentiment and conveying emotions at the same time. They tested the suggested method for multi-modal sentiment and evaluation of emotions using the CMU-MOSEI database.

According to Nemati et al. [32] they suggested an integrated multimodal data fusion approach in which the audio and visual models were fused using a latent space linear map, and their presented characteristics into the cross-modal space were fused with the textual modality using a Dempster-Shafer (DS) theory-based evidential fusion approach..” The suggested technique outperformed “both decision-level and non-latent space fusion approaches when tested on DEAP dataset.” The HED dataset was created by Fang et al. of [33] as a sizable multimodal emotion dataset to aid in the emotion detection challenge. The method for multimodal emotion recognition was then suggested. The ”HED” dataset was far bigger than previous datasets and comprises emotion-aligned face, body, and text samples that express many emotions, including happiness, sadness, disgust, anger, and fear. The researchers of [34] utilized Muse-CaR, a vehicle review database, to make continuous emotional predictions. To do this, they initially extract handmade characteristics and complex depictions from several modes. They next used the “Long Short-Term Memory (LSTM) recurrent neural network and the self-attention system to represent the sequence’s complicated temporal relationships. The Concordance Correlation Coefficient (CCC) loss was used to direct the model’s learning of local fluctuations as well as the global trend of emotion”. At last, two fusion procedures, early fusion and late fusion, were used to improve the model’s performance even more by complementing information from various modalities.

To enhance the efficiency of the emotional identification framework, researchers presented a multimodal emotion detection model based on voice and text [35]. Additionally, the study confirmed that the intersection of sentiment analysis of textual data and the domain of artificial intelligence yields effective outcomes for classification and prediction. Various neural network architectures, including CNN, LSTM, and GRU, have demonstrated their efficacy in this regard [36]. To learn acoustic emotion features, ”CNN and LSTM (long short-term memory) were combined in the form of binary channels; in the meantime, an effective Bi-LSTM (bidirectional long short-term memory) network” was used to record textual elements. In addition, a Deep Neural Network (DNN) was employed to classify and gain knowledge of the fusion characteristics. The results of both voice and text emotion analysis were used to establish the final emotional state. The authors of [37] introduced the Visual Aspects Attention Network, or VistaNet, which integrates text and visual elements. They demonstrated that, in many situations, visuals support text in emotion identification, emphasizing significant elements of an entity rather than conveying emotions independent of the text. As a result, rather than using visual data as features, VistaNet uses visual information as alignment to highlight the relevant phrases in a text using attention.

III. Methodology

Leveraging bimodal datasets in affective computing is gaining traction due to its ability to enhance human interactions and multiple use cases. Discerning emotions with multimodal information significantly elevates the caliber of human engagements. This approach is drawing widespread attention and is under exploration in various domains. By amalgamating details from sources such as facial cues, vocal nuances, gestures, and physiological indicators, multimodal emotion detection offers a deeper understanding of emotional states. This is particularly salient in areas like healthcare, gaming, virtual reality, and human-machine interfaces [38]. We compiled the MELD dataset for this specific emotion detection purpose. Data preprocessing was executed with normalization techniques and linguistic methodologies. Kernel Principal Component Analysis (K-PCA) was introduced for feature derivation. Subsequently, we employed the proposed Bi-Lateral Gradient Graphical Conventional Network (Bi-LG-GCN) algorithm for emotion discernment. The results were assessed based on metrics including accuracy, precision, recall, f1-score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The comprehensive methodological approach is illustrated in Figure 1.

A. Data collection

In our daily interactions, emotions are frequently communicated and understood through a mix of signals from various modalities. Figure 2 provides a glimpse into multimodal

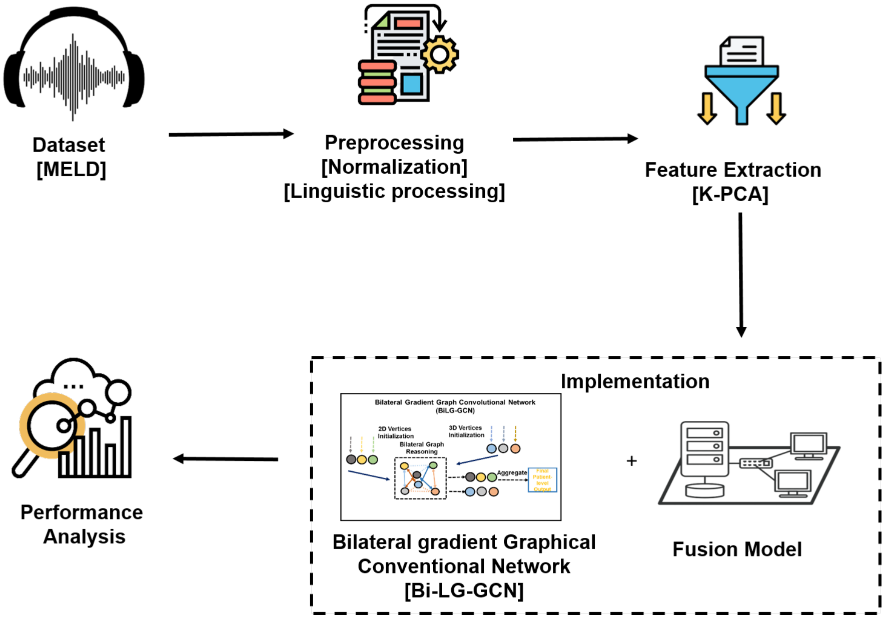


Fig. 1. Multimodal Emotion Recognition using Bi-LG-GCN system

databases that accumulate both auditory and textual information. Given the significance of reliable data, many researchers have dedicated efforts to identify nuanced emotions, be it manifestly or in more subtle forms. These feelings typically originate from three main channels: text, voice, and visuals. As we embark on this data-gathering phase, it's pivotal to pinpoint the specific type of data to collect and the sources from which they'll be procured. The method of data collection can differ based on the application's nature [8]. A prime illustration of this is the MELD dataset, which was crafted by enhancing the core affective computing dataset.

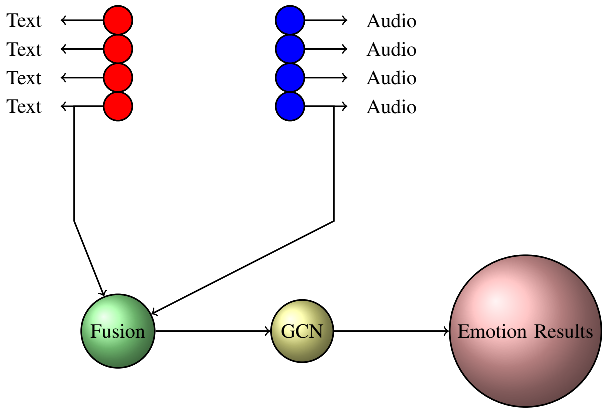


Fig. 2. Design for Combining Multiple Text and Audio Datasets

The MELD was developed by improving and expanding upon the original Emotion Lines dataset. This multimodal dataset includes audio and visual media and text and consists of the same conversation examples found in Emotion Lines. For the purpose of multimodal emotion identification, we only employ the text and audio modalities. The dataset makes available more than 1400 conversations and 13000 utterances from the Friends TV series. The discussions included a number of speakers. Each remark in a discourse is given an identity according to one of the seven emotions: fear, anger, surprise, sorrow, neutral, disgust, and joy. Additionally, the dataset annotates each utterance's emotion (positive, negative, and neutral). The dataset was collected from (https://www.kaggle.com/datasets/zaber666/meld-dataset).

B. Data preprocessing

Once the data has been collected, the primary task is to refine it for training the Bi-LG-GCN. This refinement may require feature extraction or modifications to account for variations in speaker attributes or recording quality. Such adjustments are crucial, particularly when dealing with datasets comprising both text and audio, to guarantee consistent normalization and effective language processing.

1) Normalization: Normalization describes standardizing and modifying input data from numerous modalities (text input, voice) to a single scale or range. This approach tries to eliminate differences and disparities between modalities, allowing for optimal emotional information integration and comparison. By normalizing the data, information acquired from diverse modalities may be treated similarly and merged in a meaningful way. This normalization process improves the accuracy and reliability of the emotion detection system, allowing for more robust and precise recognition and interpretation of human emotions across many modalities.

2) Linguistic processing: Linguistic processing is the computerized examination and interpretation of linguistic information with the purpose of identifying and detecting emotions in a multimodal environment. Enhancing the durability and precision of identifying emotional systems entails synthesizing emotional information from many language modalities. To evaluate linguistic elements like emotions, semantic content, and structural patterns, this discipline utilizes techniques from Machine learning, AI, and the processing of natural language. The objective is to enhance human emotion interpretation and processing across several modalities, enabling more precise and thorough emotion detection[39].

C. Feature extraction using Kernel Principle Component Analysis (K-PCA)

Feature extraction is the process of extracting significant features from large amounts of data. It involves evaluating many modalities and translating raw input signals into compact and useful summaries that incorporate relevant emotional cues. These retrieved properties are input for machine learning algorithms to classify emotions. In our work, we proposed Kernel Principle Component Analysis (K-PCA) for feature extraction. For reducing the dimension of vast quantities of data, PCA is a popular multimodal estimate technique. Typically, dimensionality reduction is done by randomly selecting the linear relationship between the parameters. However, as was previously said, conventional PCA offers linear dimensional reduction. For mapping a non-linear process in a data set, KPCA is still a more effective method. Contrary to other non-linear techniques, the ability of the kernel algorithm to run without any non-linear optimization is essential. This approach includes altering the input variables and using them as independent PC parameters. Kaiser-Meyer-Olkin (KMO) [40] is one of the most often used statistics for determining the amount of data in any factor analysis (FA), a quick explanation of the procedure to establish KPCA for feature extraction is provided with Eq. 1.

$$KMO = \frac{\sum\sum q_{ji}^2}{\sum\sum q_{ji}^2 + \sum\sum b_{ji}^2} \tag{1}$$

Where q_{ji} is the correlation value between variables j and i, and b_{ji} is their partial correlation value. Considering that the non-linear change $\phi(w)$ from the initial in sample covariance matrix D in F space should meet the Eq. 3, the projected novel characteristics have zero mean Eq. 2:

$$\frac{1}{M}\sum_{j=1}^{M} \Phi\left(\Phi(W_j)\right) = 0 \tag{2}$$

$$D = \frac{1}{M}\sum_{j=1}^{M} (\phi(W_j) \cdot (\phi(W_j))^S) \tag{3}$$

If the kernel function is formatted as follows Eq. 4:

$$l(W_j, W_i) = \varnothing(W_j)^S \varnothing W_i \tag{4}$$

$$L_{\text{bl}}^2 = \lambda_l M L_{\text{bl}} \tag{5}$$

$$L_{ji} = l(W_j, W_i) \tag{6}$$

Where $\mathbf{b}_l$ is an N-dimensional column vector and $\mathbf{b}_{lj}^{\prime s}$ as follows in the Eq. 7:

$$\mathbf{b}_l = [b_{l1}, b_{l2}, \ldots, b_{lM}]^S \tag{7}$$

The theory b_{lis} is solvable by the Eq. 8

$$L_{bl} = \lambda_l M_{b_l}, \tag{8}$$

and the corresponding kernel primary components can be determined by the Eq. 9

$$z_l(W) = \varnothing(W)^S \quad u_l = \sum_{j=1}^{M} b_{lj} \cdot \ell(W, W_j) \tag{9}$$

If the predicted dataset $\{\varphi(w_j)\}$ does not have a zero mean, the kernel matrix L can be replaced with the Gram matrix $\tilde{L}$. The Gram matrix is denoted by Eq. 10 :

$$\widetilde{L} = L{-}1_M L - L1_M + 1_M LM \tag{10}$$

Where 1_M is the $M \times M$ matrix with all components equivalent to $\frac{1}{M}$. The positive aspect of kernel approaches is that it is not essential to calculate $\{\varphi(w_j)\}$ explicitly; the kernel matrix may be rapidly established from the training dataset $\{w_j\}$ in Eq. 11.

$$k(W, Z) = \exp\left(-\frac{1}{2\sigma^2}\|W - Z\|^2\right) \tag{11}$$

In our work, KPCA analyzes and understands human emotions by combining numerous sources of data. KPCA is a dimensionality reduction approach that converts the data provided into a lower-dimensional space while retaining vital details. KPCA can manage nonlinear interactions between modalities by utilizing a kernel function. This method attempts to find the fundamental trends and patterns in multimodal data, enabling more accurate and robust emotion identification across several modalities improving human emotion comprehension and interpretation.

D. Implementation of Bilateral gradient graphical Conventional Network (Bi-LG-GCN)

An innovative method for multimodal emotion identification is the bi-lateral gradient graph convolutional network (Bi-LG-GCN). Emotion recognition is the process of identifying and labeling the emotions that are sent by humans through a variety of modalities such as audio and text. The Bi-LG-GCN needs the advantage of graph convolutional networks (GCNs) to detect the complex relationships between multiple modalities and increase recognition of emotional performance. Figure 3 shows the GCN architecture.

The input data to the Bi-LG-GCN consists of various modalities, each represented as a graph. The nodes in each graph represent the individual samples in the dataset, while the edges among the nodes record the relationships between them. In the context of multimodal emotion detection, the nodes may represent distinct face, voice characteristics, or textual information, and the edges could reflect similarities between them. The Bi-LG-GCN's primary concept is to efficiently capture multimodal interactions by combining local and global information from graphs. A bilateral gradient approach that makes use of the advantages of both local and global graph convolution processes achieves this. While the global graph

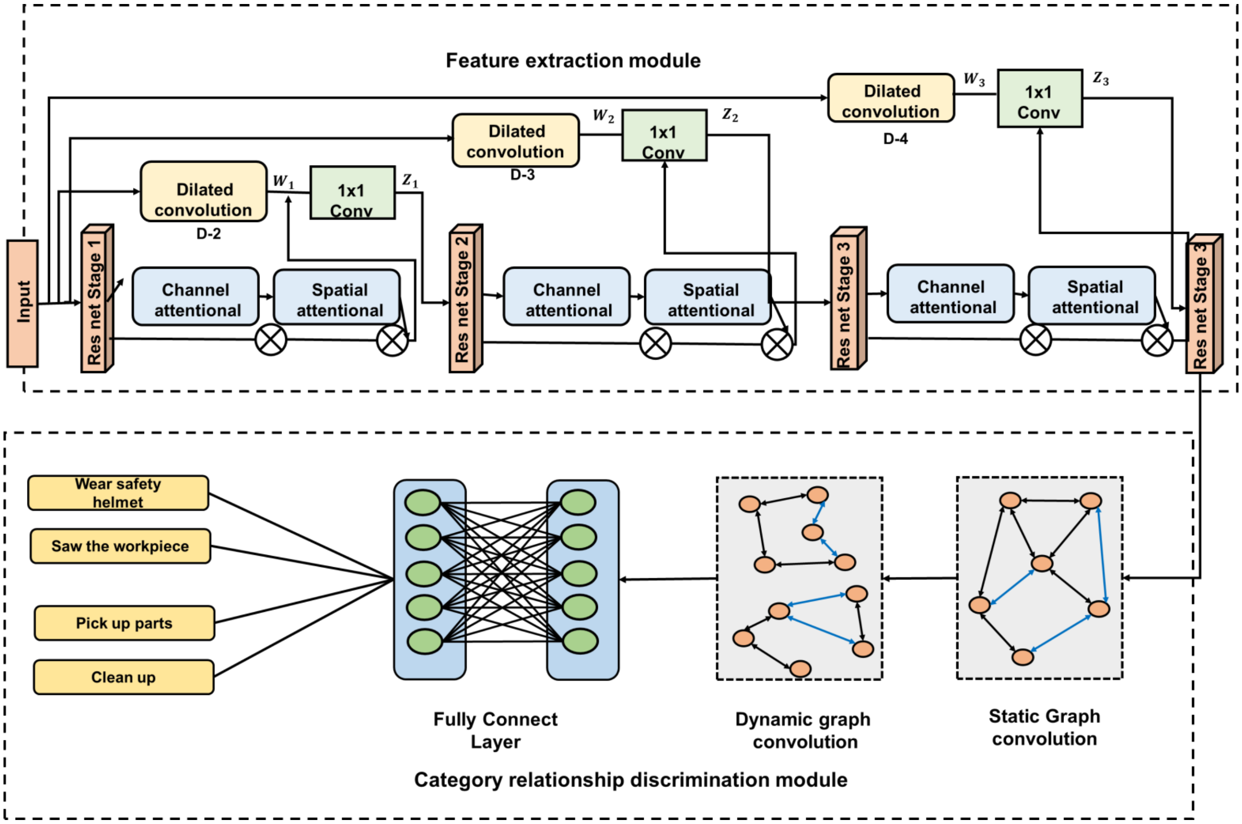


Fig. 3. GCN Architecture

convolution covers the general framework of the network and captures global interactions, the local graph convolution concentrates on gathering information about each node's nearest area.

Bi-LG-GCN can be represented as follows:

Local graph convolution in the Eq. 12:

$$g_j k = \text{ReLU}\left(V_k \cdot Y_j + \sum_{j \in M_j} \alpha_{ji} \cdot Y_i\right) \tag{12}$$

A hidden instance of the i^{th} node after the l^{th} local convolution of the graph layer is denoted by the symbol g_j^k in this equation. The k^{th} local convolution layer's weight matrix is denoted by V_k, while the i^{th} node's surrounding nodes are represented by M_j. The attention-related elements that control the information flow between the nodes are illustrated by the α_{ji}. Global graph convolution in the Eq. 13:

$$g_j h = \text{ReLU}\left(V_h \cdot g_j k + \sum_{j \in w} \beta_{ji} \cdot g_i k\right) \tag{13}$$

After global graph convolution, g_j delivers the hidden representation of the i^{th} node. W is for the set of all nodes in the graph, V_h stands for the weight matrix of the global convolution layer, and β_{ji} stands for the attention coefficients that capture the global connections between nodes. The bilateral gradient technique combines the local and global representations to generate the final node representations. It is achieved using an iterative optimization procedure that updates the attention coefficients α_{ij} and β_{ij} while minimizing the loss function.

Overall, the Bi-LG-GCN offers a strong foundation for multimodal emotion recognition by effectively capturing the intricate interactions between various modalities. The bi-lateral gradient approach combines local neighborhood information and the global graph structure to combine local and global information, improving the accuracy of recognized emotion. By adopting this methodology, not only can the accuracy of emotion classification tools be enhanced, but the broader domain of multimodal emotion recognition could also witness significant advancements.

IV. Result Evaluation

Result evaluation in our context relies on accuracy, precision-recall, and the F1-score metrics. While accuracy offers a general overview, precision and recall prioritize positive predictions. The F1-score strikes a balance. The method employs Bi-LG-GCN for multimodal emotion recognition, assessing performance with these metrics to ensure precise emotion detection across diverse data sources.

A. Experimental setup

We designed our experimental setup based on a combination of both hardware and software configurations tailored for efficient and taxing machine learning applications. The experiments were performed on a Windows 11 operating system.

The chosen programming environment is Python 3.11, which supports the deployment of PyTorch 2.0. We also incorporated the use of Colab Google software to complement the setup. The underlying hardware and software infrastructure ensures the seamless implementation and execution of our technique.

TABLE I
HARDWARE AND SOFTWARE CONFIGURATION

Component	Specification
OS	Windows 11
Processor	11th Gen Intel(R) Core(TM) i7-11370H @ 3.30GHz
Memory (RAM)	16.0 GB
Storage	SSD
Graphics Card	GPU 0 Intel(R) Iris(R) Xe Graphics
	GPU 1 NVIDIA GeForce RTX 3050 Ti Laptop GPU
Network Adapter	Ethernet (Turk Telecom) TTNET 24Mbps
Software	Pytorch 2.0 Python 3.11 Colab Google
Total Processing Time	(h, m, s): 00, 21, 31

B. Existing methods

A crucial aspect of classification involves evaluating the models to gauge their precision and dependability in classifying data. Examining the performance metrics helps pinpoint the optimal model for classification assignments. Insights from model training and testing play a pivotal role in evaluating their efficacy. For the performance assessment of classification models, the confusion matrix is a standard tool. This matrix encompasses four components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Using the confusion matrix, we can derive various performance metrics like accuracy, precision, recall, F1-score, etc., as illustrated in the color-coded axis above. Model predictions undergo testing and training processes, categorizing outcomes into the following segments: **True Positive** - A successful classification of an actual positive outcome, **False Negative** - An erroneous classification where a positive outcome is predicted as negative, **False Positive** - A misclassification where a negative outcome is predicted as positive, and **True Negative** - A successful classification of an actual negative outcome [10]. To provide a clearer visualization of these categorizations, the following Table II shows how TP, FP, TN, and FN are represented.

TABLE II
REPRESENTATION OF TP, FP, TN, AND FN EXPLANATION MATRIX

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Positive (FP)
Actual Negative	False Negative (FN)	True Negative (TN)

The accuracy, f1-score, precision, and recall metrics are used in this section to assess the performance of the suggested technique Bi-LG-GCN. For calculating accuracy and f1-score, we compare the Bi-LG-GCN with Dialogue Contextual Reasoning Networks (DialogueCRN) [41], Multimodal fused Graph Convolutional Networks (MMGCN), and Graph network-based Multimodal Fusion Technique [41]. For calculating precision and recall, we compare the Bi-LG-GCN with Support Vector Machine (SVM) [33], CentralNet, and Low-rank Multimodal Fusion (LMF).

C. Comparison phase

The degree of accuracy reflects how effectively these systems can identify and categorize emotions. It assesses how well the system can correlate the observed multimodal input data with the relevant emotional states, indicating how well it can decode and identify human emotions. Table III and Figure 4 show the accuracy comparison of the suggested Bi-LG-GCN and alternative procedures. Whereas Dialogue CRN, MMGCN, and Graph MFT only succeed with 65.31, 65.56, and 67.90% accuracy, respectively, the suggested technique Bi-LG-GCN achieves 80% accuracy. The Bi-LG-GCN achieved higher accuracy than the other methods. The accuracy value is expressed in the following Eq. 14.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Instances}} \tag{14}$$

TABLE III
COMPARISON OF ACCURACY

Methods	Accuracy (%)
DialogueCRN	65.31
MMGCN	65.56
GraphMFT	67.9
Bi-LG-GCN	80

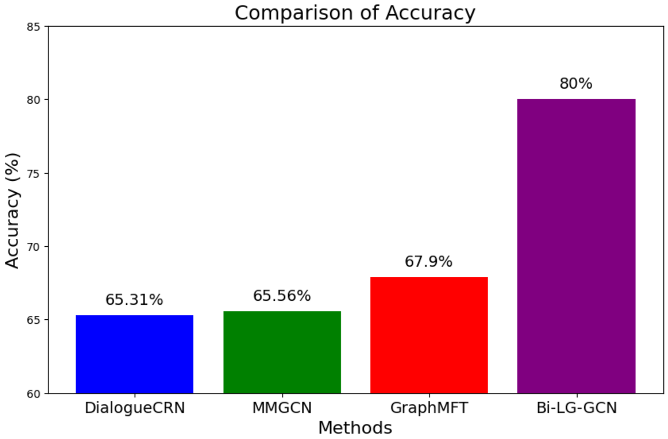


Fig. 4. Accuracy comparison of Bi-LG-GCN with existing methods

The effectiveness of a model that attempts to identify and categorize emotions utilizing different forms of media combines recall and precision into a single measurement that considers the precision and comprehensiveness of emotion forecasts. The overall ability to capture emotions across several modalities is greater when the F1 score is higher. The F1 score comparison of the suggested and other approaches is displayed in Table IV and Figure 5. Whereas DialogueCRN, MMGCN, and GraphMFT only succeed at 65.34, 65.71, and 68.07%, respectively, the suggested technique Bi-LG-GCN achieves 81% accuracy. Comparing the Bi-LG-GCN method to other conventional approaches provides a higher F1 score. The F1-score with the Eqn. 15 is described below.

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{15}$$

TABLE IV
COMPARISON OF F1-SCORE

Methods	F1-Score (%)
DialogueCRN	65.34
MMGCN	65.71
GraphMFT	68.07
Bi-LG-GCN	81

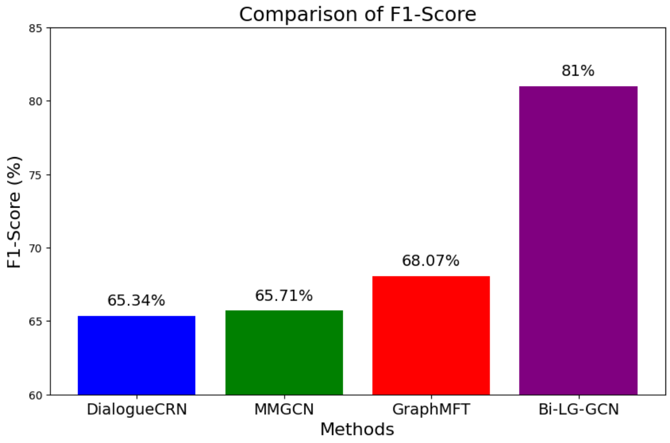


Fig. 5. F1-score comparison of Bi-LG-GCN with existing methods

Precision is the ability to reliably and accurately recognize particular emotions from various signals; it calculates the percentage of emotions that were properly categorized out of all the emotions a system predicts. The system is more accurate at recognizing the desired emotions when the precision is higher since a lower percentage of false positives is shown. Precision is a key indicator of how well multimodal emotion recognition systems execute. Table V and Figure 6 demonstrate the precision comparison between the suggested and alternative procedures. A precision value of 81% is achieved by the recommended technique, Bi-LG-GCN, as opposed to merely 66%, 80%, and 73% for SVM, CentralNet, and LMF. The Bi-LG-GCN concept offers a high precision value when compared to earlier techniques. The following equation is used to determine the precision value with the Eqn. 16.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

TABLE V
COMPARISON OF PRECISION

Methods	Precision (%)
SVM	66
CentralNet	80
LMF	73
Bi-LG-GCN	81

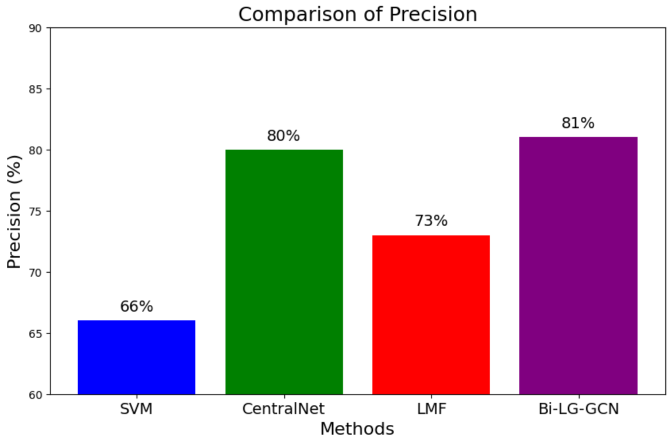


Fig. 6. Precision comparison of Bi-LG-GCN with existing methods

Recall is a characteristic of a system that enables it to identify and classify emotions through multiple modalities consistently. The efficacy of the system in accurately discerning emotions from diverse data inputs is quantified by the ratio of genuinely positive emotions it has identified. Table VI and Figure 7 illustrate the recall distinctions between the proposed and other methods. The suggested method, Bi-LG-GCN, attains a recall score of 81%, which surpasses the scores of 65%, 79%, and 69% achieved by SVM, CentralNet, and LMF respectively. Consequently, Bi-LG-GCN outperforms conventional methods in terms of recall with the Eqn. 17. The ensuing formula determines the recall metric.

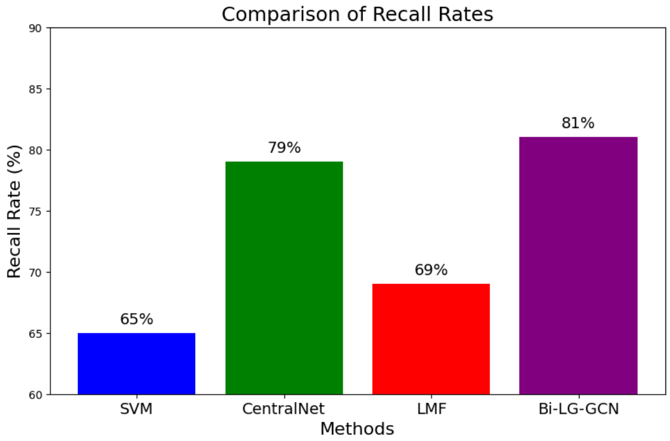


Fig. 7. Recall comparison of Bi-LG-GCN with existing methods

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

TABLE VI
COMPARISON OF RECALL

Technique	**Recall Rate (%)**
SVM	65
CentralNet	79
LMF	69
Bi-LG-GCN	81

In comparing the results, it becomes evident that Bi-LG-GCN demonstrates superior recall capabilities over other evaluated techniques. This dominance in performance reinforces the potential of Bi-LG-GCN for applications that prioritize recall. Furthermore, the detailed evaluation as per the provided formula offers a comprehensive perspective on the recall differences among these methods.

V. CONCLUSION

In our study, we proposed a novel method, a bi-lateral gradient graph convolutional network (Bi-LG-GCN) for multimodal emotion recognition. To recognize multimodal emotions using Bi-LG-GNN, we collected the MELD dataset using textual and audio modalities. And applied linguistic and normalization pre-processing methods to improve the standard and standardization of both datasets. As a feature extraction technique, we used K-PCA to extract certain traits from each modality. The experiment is done on various parameters such as ”accuracy (80%), precision (81%), recall (81%), F1-score (81%)” using Bi-LG-GCN for the MELD dataset. Several existing methods were used in the comparing phase. The experiment showed our proposed Bi-LG-GCN method performed efficiently when compared to earlier techniques. Its performance may be impacted by the quantity and caliber of multimodal input, which may result in unreliable emotion predictions. The approach may also have trouble generalizing to various cultural and linguistic environments, which makes it challenging to use in real-world situations involving a range of emotional expressions. For better user experiences, more studies can examine real-time emotion identification, cross-domain generalization, and its incorporation into virtual assistants and emotion-aware systems.

REFERENCES

[1] P. Savci and B. Das, “Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML,” *Heliyon*, vol. 9, no. 5, p. e15670, 2023-05-01. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023028773

[2] M. Aydogan, “A hybrid deep neural network-based automated diagnosis system using x-ray images and clinical findings,” *International Journal of Imaging Systems and Technology*, vol. 33, no. 4, pp. 1368–1382, 2023, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ima.22856. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22856

[3] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown, “A performance comparison of eight commercially available automatic classifiers for facial affect recognition,” *PLOS ONE*, vol. 15, no. 4, p. e0231968, 2020, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231968

[4] E. Cameron and M. Green, *Making Sense of Change Management: A Complete Guide to the Models, Tools and Techniques of Organizational Change*. Kogan Page Publishers, 2019. [Online]. Available: https://www.example.com/your-book-url

[5] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, “Cross corpus multi-lingual speech emotion recognition using ensemble learning,” *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021. [Online]. Available: https://doi.org/10.1007/s40747-020-00250-4

[6] A survey of emotion recognition methods with emphasis on e-learning environments | journal of network and computer applications. [Online]. Available: https://dl.acm.org/doi/10.1016/j.jnca.2019.102423

[7] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, “A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions,” *Knowledge-Based Systems*, vol. 223, p. 106970, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121002331

[8] R. Das and M. Soylu, “A key review on graph data science: The power of graphs in scientific studies,” *Chemometrics and Intelligent Laboratory Systems*, vol. 240, p. 104896, 2023-09-15. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743923001466

[9] P. Savci and B. Das, “Prediction of the customers’ interests using sentiment analysis in e-commerce data for comparison of arabic, english, and turkish languages,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 3, pp. 227–237, 2023-03-01. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S131915782300054X

[10] I. Pulatov, R. Oteniyazov, F. Makhmudov, and Y.-I. Cho, “Enhancing speech emotion recognition using dual feature extraction encoders,” *Sensors*, vol. 23, no. 14, p. 6640, 2023-01, number: 14 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/23/14/6640

[11] M. Egger, M. Ley, and S. Hanke, “Emotion recognition from physiological signal analysis: A review,” *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157106611930009X

[12] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. W. Shalaby, “A 3d-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition,” *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 167–176, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866520301389

[13] C.-H. Wu and W.-B. Liang, “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels,” *T. Affective Computing*, vol. 2, pp. 10–21, 2011. [Online]. Available: https://ieeexplore.ieee.org/document/5674019

[14] M. Soylu, A. Soylu, and R. Das, “A new approach to recognizing the use of attitude markers by authors of academic journal articles,” *Expert Systems with Applications*, vol. 230, p. 120538, 2023-11. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417423010400

[15] Speech emotion recognition with acoustic and lexical features. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7178872/

[16] K. D. N. and A. Patil, “Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks,” in *Interspeech 2020*. ISCA, 2020, pp. 4243–4247. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/n20_interspeech.html

[17] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, “Cross-subject multimodal emotion recognition based on hybrid fusion,” *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/document/9195813

[18] T. Dalgleish and M. Power, *Handbook of Cognition and Emotion*. John Wiley & Sons, 2000-11-21, google-Books-ID: vsLvrhohXhAC. [Online]. Available: https://www.google.com.tr/books/edition/Handbook_of_Cognition_and_Emotion/vsLvrhohXhAC?hl=en&gbpv=1&dq=isbn:9780470842218&printsec=frontcover&pli=1

[19] C. Guanghui and Z. Xiaoping, “Multi-modal emotion recognition by fusing correlation features of speech-visual,” *IEEE Signal Processing Letters*, vol. 28, pp. 533–537, 2021, conference Name: IEEE Signal Processing Letters. [Online]. Available: https://ieeexplore.ieee.org/document/9340264

[20] S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, and A. Mahmoud, “Text-based emotion recognition using

deep learning approach," *Computational Intelligence and Neuroscience*, vol. 2022, p. e2645381, 2022, publisher: Hindawi. [Online]. Available: https://www.hindawi.com/journals/cin/2022/2645381/

[21] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Context-dependent domain adversarial neural network for multimodal emotion recognition." in *Interspeech*, 2020, pp. 394–398. [Online]. Available: https://www.iscaspeech.org/archive/interspeech_2020/lian20b_interspeech.html

[22] D. Priyasad, T. Fernando, S. Denman, C. Fookes, and S. Sridharan, "Attention driven fusion for multi-modal emotion recognition." [Online]. Available: http://arxiv.org/abs/2009.10991

[23] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 1359–1367, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1911.05659

[24] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep canonical correlation analysis." [Online]. Available: http://arxiv.org/abs/1908.05349

[25] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9156904

[26] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features." [Online]. Available: http://arxiv.org/abs/2111.10202

[27] Y.-T. Lan, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020-07, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9207625/

[28] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep AutoEncoder," *IEEE Access*, vol. 8, pp. 164 130–164 143, 2020, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/document/9187342

[29] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One source to detect them all: Gender, age, and emotion detection from voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 338–343, ISSN: 0730-3157. [Online]. Available: https://ieeexplore.ieee.org/document/9529731

[30] X. Wu, W.-L. Zheng, and B.-L. Lu, "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition." [Online]. Available: http://arxiv.org/abs/2004.01973

[31] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 370–379. [Online]. Available: https://aclanthology.org/N19-1034

[32] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172 948–172 964, 2019, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/document/8911364

[33] Z. Fang, A. He, Q. Yu, B. Gao, W. Ding, T. Zhang, and L. Ma, "FAF: A novel multimodal emotion recognition approach integrating face, body and text." [Online]. Available: http://arxiv.org/abs/2211.15425

[34] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, ser. MuSe'20. Association for Computing Machinery, 2020-10-15, pp. 27–34. [Online]. Available: https://doi.org/10.1145/3423327.3423672

[35] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-textual emotion recognition based on improved neural networks," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–9, 2019. [Online]. Available: https://www.hindawi.com/journals/mpe/2019/2593036/

[36] M. Aydoğan and A. Karci, "Improving the accuracy using pre-trained word embeddings on deep neural networks for turkish text classification," *Physica A: Statistical Mechanics and its Applications*, vol. 541, p. 123288, 2020-03. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378437119318436

[37] Q.-T. Truong and H. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 305–312, 2019-07-17. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.3301305

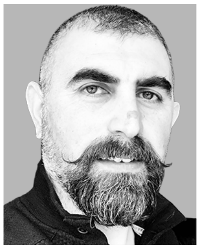
[38] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667305322001089

[39] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023-03-01. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522001634

[40] A. Solgi, A. Pourhaghi, R. Bahmani, and H. Zarei, "Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD)," *Ecohydrology & Hydrobiology*, vol. 17, no. 2, pp. 164–175, 2017-04-01. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1642359316300672

[41] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation." [Online]. Available: http://arxiv.org/abs/2208.00339

Hussein Farooq Tayeb Alsaadawi is currently pursuing a Ph.D. at Firat University's Technology Faculty in the Department of Software Engineering, located in Elazig, Turkey. He completed his M.Sc. in the same department from 2015 to 2017, further solidifying his expertise in the field. Earlier, he earned his B.Sc. degree in Computer Science from Cihan University in Erbil, Iraq, between 2007 and 2011, demonstrating his early passion for the subject. In addition to his academic achievements, He holds a Technician Diploma in Electronic Engineering from Erbil Technical Institute in Iraq, which he obtained from 2000 to 2002. Following the completion of his diploma, he embarked on a successful career as a Software Technician in Iraq. He has published academic papers and international conference proceedings based on his valuable insights and research. His passion for software development, combined with his expertise in artificial intelligence, drives him to uncover groundbreaking solutions that will shape the industry's future.



Resul Das is a full professor in the Department of Software Engineering, Technology Faculty, Firat University. He graduated with B.Sc. and M.Sc. degrees from the Department of Computer Science at Firat University in 1999 and 2002, respectively. Then he received a Ph.D. degree at the Department of Electrical-Electronics Engineering at the same university in 2008. He served as both a lecturer and network administrator at the Department of Informatics at Firat University from 2000 to 2011. In addition, he has been an instructor of CCNA and CCNP and the coordinator of the Cisco Networking Academy Program since 2002. He worked between September 2017 and June 2018 as a visiting scholar at the Department of Computing Science at the University of Alberta, Edmonton, Canada supported by TÜBİTAK-BIDEB 2219 Post-Doctoral Fellowship. He has many journal papers and international conference proceedings. He served as Associate Editor for the Journal of IEEE Access and the Turkish Journal of Electrical Engineering and Computer Science-TUBITAK from 2018 to 2021. He became the 2% of the "World's Most Influential Scientists" list published by Stanford University researchers from 2020 to 2022. His current research areas include computer networks and security, cyber-security, software architectures, software testing, IoT/M2M applications, complex networks, graph visualization, knowledge discovery, and data fusion.