IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2023, accepted 4 September 2023, date of publication 13 September 2023,
date of current version 19 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3315243

APPLIED RESEARCH

Taking All the Factors We Need: A Multimodal Depression Classification With Uncertainty Approximation

SABBIR AHMED[1], (Member, IEEE), MOHAMMAD ABU YOUSUF[1], MUHAMMAD MOSTAFA MONOWAR[2], (Member, IEEE), MD. ABDUL HAMID[2], AND MADINI O. ALASSAFI[2]

[1]Institute of Information Technology, Jahangirnagar University, Dhaka 1342, Bangladesh
[2]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding authors: Sabbir Ahmed (sabbir.iit.ju@gmail.com) and Mohammad Abu Yousuf (yousuf@juniv.edu)

This work was supported in part by the Institutional Fund Projects funded by the Ministry of Education under Grant IFPIP:505-611-1443; and in part by the King Abdulaziz University Deanship of Scientific Research (DSR), Jeddah, Saudi Arabia.

ABSTRACT Depression and anxiety are prevalent mental illnesses that are frequently disregarded as disorders. It is estimated that more than 5% of the population suffers from depression or anxiety. Although there have been a number of studies in these fields, the majority of the research focuses on one or two factors for detection purposes, whereas these factors are not mutually inclusive and vary among studies. To mitigate these issues, we first consider all possible symptoms associated with depression and develop a multimodal diagnosis system that may take into account any number of patient-specific factors. If multiple factors can be addressed within a single learning model, it is advantageous for data collection and future development. To facilitate training with missing modalities, we propose an attention-based multimodal classifier with selective dropout and normalization, which can facilitate the training of various multimodal datasets on one neural network. We have experimented with three multimodal datasets with varying modalities to show the impact of combined training in one neural network and achieved an F1 score of 0.945. However, missing modalities in the model can create uncertainty in the prediction. For uncertainty approximation, the Monte Carlo dropout (MC dropout) and the spectral-normalized neural Gaussian process (SNGP) with the coefficient of variation and S1-Score metrics are implemented to provide important information about multimodal diagnosis processes. In the experiment, selective dropout with SNGP achieved a coefficient of variation in loss of 0.384 and an S1-score of 0.9374.

INDEX TERMS Deep learning, multi-modal neural network, uncertainty approximation, ensemble.

I. INTRODUCTION

According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), depression is characterized by the presence of one or more major depressive episodes that last at least two weeks and include symptoms such as a depressed mood, decreased interest in activities, and feelings of worthlessness or guilt [1]. Depression is also characterized as a mood disorder, with its primary manifestation being a state of temporary or persistent feelings of sadness, diminished enjoyment, and diminished self-esteem, as well as disturbances in sleep and eating patterns, concentration difficulties, and feelings of fatigue. These symptoms may persist over time, leading to chronic and recurring episodes that can hamper an individual's ability to engage in daily activities [2]. Depression is a prevalent mental health condition that impacts a substantial portion of the global population, with an estimated 280 million individuals, or approximately 5% of adults [3]. It has been identified as a possible precursor to suicide, and the number

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

of suicide-related deaths exceeds 700,000 per year [3]. A person's capacity to participate in professional, academic, and social activities can be hampered by depression. According to Dattini et al., depression causes a global loss or impairment of 50 million years of work per year [4]. Even though mental health services and treatment are not accessible to over 75% of individuals in low- and middle-income countries [3]. While the exact reason for depression still remains unknown, social, psychological, environmental, and medical conditions might be some factors in its development [5]. Thus, it is a multidimensional field where psychological, medical, and technical researchers are trying to correlate symptoms with plausible detection systems. This would allow early detection as well as a self-assessment system to mitigate possible risks. However, the cause and symptoms of the condition exhibit significant heterogeneity [1], posing challenges for conventional questionnaires and analytical methods given their complex characteristics.

In this regard, numerous studies have indicated that depression can be detected by observing and collecting data from subconscious states, which can be accomplished using a variety of methods and tools. Psychological methods [6], [7], [8], [9], [10] involve standardized questionnaires, interviews, or scales to evaluate the symptoms, severity, and ramifications of depression. Nevertheless, these methodologies exhibit certain limitations, including but not limited to subjectivity, bias, low sensitivity, and cultural dissimilarities. In contrast, machine learning (ML) approaches employ computational algorithms to examine diverse modalities such as facial expressions, speech, text, or physiological signals [11], [12], [13], [14], [15], [16], [17]. Common approaches for detection include feature extraction, feature selection, and classification using various algorithms. However, they still have drawbacks like data quality and availability, moral and privacy concerns, robustness and generalizability, or interpretability. Neuroimaging techniques such as electroencephalography (EEG), magnetic resonance imaging (MRI), or positron emission tomography (PET) are used to measure structural or functional changes in the brain associated with depression [18], [19], provide insights into the neurobiological mechanisms and biomarkers of depression. These approaches are limited by their high computational cost, invasiveness, low accessibility, or technical challenges. However, most of these methods rely on a single domain or modality of data, which may not capture the complexity and heterogeneity of depression. Moreover, different modalities may provide complementary or contradictory information about depression. While these approaches have their strengths and weaknesses, combining them in multimodal neural networks can provide a more comprehensive and accurate diagnosis of depression. Methods such as concatenation, weighting, and gating are employed to integrate multiple modalities at the input or feature level. The multimodal deep learning framework (MDLF), cross-modal attention network (CMAN), deep convolutional neural network (DCNN), and bi-directional long short term memory (BiLISTM) are examples of these methods [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Both feature (early) and dense (late) level concatenation with existing and custom algorithms are proposed by several researchers. One drawback shared by all of these methods is that they must be trained in specific modalities. If the model is trained on text and speech features, for instance, it may not be able to detect depression using images or physiological features. Quantifying and incorporating uncertainty into the classification procedure is a further obstacle for multimodal approaches. There are numerous sources of uncertainty, including noise, ambiguity, variability, and insufficient data. Uncertainty can affect the confidence and dependability of classification results, leading to misdiagnoses and inappropriate interventions. Thus, we propose the use of an attention-based multimodal classifier featuring selective dropping out in order to facilitate the training of various multimodal datasets on one particular neural network. The experimentation with classification approaches involves the utilization of multiple datasets where the modalities are mismatched with one another. Furthermore, the incorporation of uncertainty approximation or confidence in the predictions or representations has been implemented to ensure model training with missing modalities. Matrics such as the S1-Score and Coeffiicient of Variation are also used for uncertainty approximation. The contributions of this paper are the following:

- We have proposed a selective dropout layer compatible with TensorFlow, to drop unnecessary or not given modalities in the concatenation layer. Selective dropout, attention and normalization are used as a block to accommodate training with missing modalities
- A multimodal Neural network is proposed and trained on separate datasets with absent modalities where the model can omit specific modalities selectively during training while still effectively utilizing the available information.
- The method also incorporates uncertainty approximation techniques, such as Monte Carlo dropout, and spectral-normalized neural Gaussian process, to enhance the robustness and generalizability of depression detection.

In this manuscript, Section II presents a review of relevant literature. The algorithms and methodologies are described in Section III. Section IV provides the analyses, results and discussions. Finally, Section V summarizes the findings and presents the conclusions with the limitations of the study.

II. RELATED WORK

Depression is a well-studied subject, both in terms of psychological and technological perspectives. According to DSM-V, depression can be classified into multiple types: disruptive mood dysregulation disorder, premenstrual dysphoric depression, persistent or major depressive disorder, depressive disorder, depression due to loss or grief, and depression due to medical conditions, which spawn their own sets of symptoms and result in more serious psychological conditions [1]. This type of mental condition is also universal regarding age and gender [32].

The presence of depressive symptoms and other neuropsychiatric symptoms may have a negative influence on both patients' and caregivers' well-being [33]. Moreover, individuals who demonstrate persistent physical and emotional symptoms after getting therapy for depression seem to be more susceptible to relapse than those who do not exhibit such symptoms [34]. Humans are susceptible to their own thoughts and tend not to express these feelings towards diagnosis. There are several self-assessment questionnaires like PHQ-8 [7], PHQ-9 [6], CES-D [8], GDS [9] and HADS [10] that normally ask about the frequencies of depressive syndromes in a specific timeframe, normally daily, weekly, or monthly. However, the existence of human biases in self-filled questionnaires led researchers to develop tests that capture the subconscious thoughts of individuals.

Early detection of depression using machine learning is vital yet challenging owing to limitations in medical technology and expertise. Researchers have investigated several ways of identifying depression, including those based on social media, EEG [15], acoustic testing, and virtual reality. Lin et al. [11] have suggested social media-based depression detection systems that leverage a deep visual-textual multimodal learning technique to expose the psychological condition of social network users. The depression detection process may also include collecting posted images and tweets from users with and without depression on Twitter, extracting deep features using CNN-based classifiers and Bert from the text and images, combining the visual and textual features, and classifying users with depression and normal users using a neural network. In separate research, a hybrid model was used to predict sadness by analyzing Reddit user text postings. This model used BiLSTM with various word embedding methods and metadata characteristics [12]. Socially Mediated Patient Portal (SMPP) is a programme that uses a data-driven approach and machine learning classification algorithms to discover depression-related signals in Facebook users [13]. Govindasamy et al. [14] utilized machine learning algorithms to identify sadness through social media user postings. Twitter data is given to two distinct classifiers, Nave Bayes and NBTree, and the results are evaluated based on the greatest accuracy value to find the most effective algorithm for detecting depression. But these methods may also suffer from inefficiency and bias since people often showcase their positive sides through behaviour or social media. In addition, EEG and eye movement (EM) data have been frequently employed for depression identification owing to their noninvasiveness and ease of recording. Using EEG and EMs datasets, this study presents a content-based ensemble approach (CBEM) to improve depression identification accuracy [18].

Although the multimodal approach is common in depression detection, the majority of the existing research focuses on bi-modality or tri-modality. A review study by Arioz et al. [35] shows that of the 1095 existing studies, only 20 devised their methodology on more than two modalities. The prevalent modalities comprise acoustic characteristics and visual cues, primarily obtained from video recordings. Nevertheless, conducting comprehensive literary analyses of all existing methodologies is beyond the scope of this research. Therefore, in this section, priority is given to researchers who have closely examined our study or have frequently employed the datasets used in our research. Multimodal methodologies pose a challenge owing to the requirement of incorporating joint representation, alignment, and fusion mechanisms. Some of the solutions for these problems involve the utilization of BiGRU, BiLSTM [36], [37], and Hierarchical Attention Network (HAN) [28] architectures for text analysis. Other approaches involve the application of GPT2-medium language models to generate task-oriented embeddings [26]. However, integrating multiple modalities in feature states with convincing fusing algorithms and feature concatenation still poses a challenge. From a tri-modal perspective, Yang et al. [20], proposed audio, video, and text streams with handcrafted feature descriptors in a DCNN to acquire high-level global features and predict PHQ-8 scores. Yazdavar et al. [25] proposed identification of depressive symptoms from tweets utilizing statistical techniques to combine heterogeneous types of characteristics collected through the collection and analysis of visual, textual, and user-generated data. Similarly, Shimpi et al. [24] proposed customised ensemble methods and have subsequently expanded their research to encompass mobile applications and cloud development. Nonetheless, the clarity of this approach is limited, as the custom fusion is typically described as a series of BI-LSTM layers within the methodology. Mantri et al. [38] proposed a system that captures a combination of facial characteristics, speech properties, and brain waves to predict the severity of depression. The system employs a numeric conversion technique and a single fully connected classifier for this purpose. The approaches discussed suffer from a loss of multimodality due to the absence of feature-merging techniques and reliance on a single classifier for feature mixing. Arroz et al. [35] compared algorithms for unimodal, automatic, and multimodal classification conversations with LSTM and gated recurrent units (GRU). Alternative approaches to multimodal depression detection encompass the examination of various indicators such as the dynamics of acoustic, facial, and head movement [27], [39], behavioural and physiological signals [40], brain functional abnormalities, heart rate variability, hemodynamic parameters [41], and partially convergent structural features [23].

Despite recent progress, existing studies on the detection of depression through multiple modes of communication still face several limitations. A significant constraint pertains to the inadequacy of efficient feature fusion mechanisms within multi-modal neural networks. The integration of information from various modalities, including text, audio, and video, presents a significant challenge for researchers seeking to develop a more comprehensive understanding of a patient's mental condition. Furthermore, current research exhibits rigidity with regard to the modalities employed. But the major constraint pertains to the level of uncertainty involved in decision-making, which is an intrinsic aspect of predicting mental health conditions. The absence of uncertainty estimation in prior studies poses a challenge in determining the degree of confidence in the model's prognostications. Hence, these limitations present noteworthy obstacles to the advancement of dependable and precise multimodal depression detection mechanisms.

III. METHODOLOGY

In this study, we propose a methodology for multimodal depression classification with uncertainty approximation. The first step is preprocessing, which involves extracting features from different modalities. Various audio features are extraceded, including zero crossings, spectral centroids, spectral rolloff, mel-frequency cepstral coefficients (MFCC), and chroma short-time Fourier transform (STFT). The 68 facial landmarks are used to extract facial and image features, and text feature extraction is performed with the BERT encoder. A multimodal neural network is proposed that takes in video, audio, text, and EEG features as input. The model also incorporates multimodality with tolerance for N missing modalities by selectively dropping out one modality so the model can learn from missing modalities in the datasets. Additionally, we employ unimodal ensembling to improve the classification performance of individual modalities. Lastly, Monte Carlo dropout and spectral-normalized neural Gaussian process methods reduce uncertainty and biases, and the S1-Score with coefficient of variation quantification estimates the uncertainty in the model's predictions. The aim of the overall methodology is to provide better feature fusion among various datasets and modalities.

A. PREPROCESSING AND FEATURE EXTRACTION

The preprocessing procedures entail the consolidation of facial landmark video modalities from the Dvlog and DAIC-WOZ datasets, as well as the generation of corresponding eye scan paths. Additionally, audio feature extraction is performed on the DAIC-WOZ and MODMA datasets, while text encoding is carried out using the BERT encoder. Finally, EEG feature extraction is conducted on the MODMA dataset. The PHQ scales were converted into binary classifications (PHQ>7). The subsequent segment delineates the process of feature extraction.

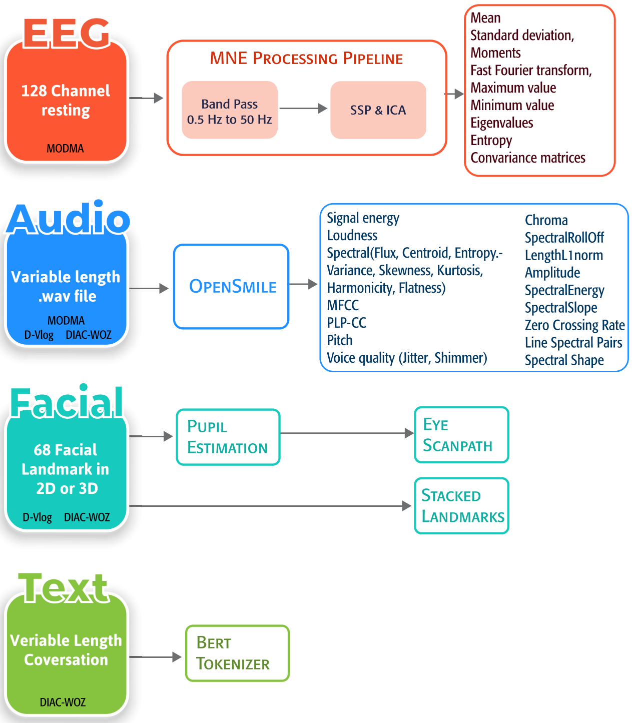


FIGURE 1. Preprocessing and feature extraction.

1) THE VIDEO FEATURE

The facial landmark for the video feature has been designated as a set of 68 2D points. Subsequently, the 68 points are arranged in the shape of 340 $*$ 340 $*$ 2, wherein each row contains 5 consecutive points. This encoding results in a total of 340 rows, which is equivalent to 28.333 minutes of data. The Dvlog (Dlib) and DAIC WoZ datasets provide video features at a rate of 1fps or one instance of facial features consisting of 68 points per second. The arrangement of video features is structured in a manner that enables the utilization of convolutional operations for the purpose of extracting features during runtime, thereby leading to a subsequent reduction and compression of the information.

2) EYE SCAN PATH

This study employs an estimated eye scan path due to the absence of pupil positions in the 68 landmark dataset. By utilising six designated points for each eye, the centre position can be determined and subsequently plotted onto a 256 $*$ 256 $*$ 3 image file. Nonetheless, the mean point of the visual organ merely estimates the location of the pupil. In future academic research endeavours pertaining to comparable multimodal investigations, it may be advantageous to utilise Facemesh, a technology developed by Google, as it offers superior resolution capabilities and the ability to track pupil coordinates in three dimensions. Therefore, it is recommended to employ comparable techniques such as Facemesh in forthcoming research endeavours.

3) AUDIO FEATURE

Since the Dvlog dataset is already provided as an Opensmile 25 feature detector, the DAIC WoZ and Modma datasets were also analysed with Opensmile [42], while 238 low-level descriptions (LLD) were considered for feature extraction, ensuring that all of the datasets have an identical auditory feature. LLD consists of mel frequency cepstral coefficients (MFCC), zero crossing, mean crossing, energy, intensity, linear predictive coefficient, and chroma characteristics. LLDs are extracted for the entire conversation length of the DAIC WoZ and Modma datasets, yielding 238 feature vectors.

4) EEG FEATURES

The MNE [43] is utilised for completing the EEG preprocessing and feature extraction. In the MODMA dataset, 128 channels of resting and event EEG data are provided. This data, however, comprises noise and a broad spectrum of frequencies. Consequently, a bandpass filter is used to remove EEG signals between 0.5 Hz and 50 Hz. This cutoff in low and high frequencies eliminates both signal drift and noise. Signal-space projection (SSP) and independent component analysis (ICA) are employed to remove any signal artefacts from the data. Then, independent attributes such as mean, standard deviation, moments, Fast Fourier transform, maximum, minimum, eigenvalues, and entropy are extracted from the EEG brainwave data using MNE [43]. The complete feature length is $128 * 21$.

5) TEXT PROCESSING

For text data in the DAIC Woz dataset, each conversation is punctuation-cleaned. BERT (Bidirectional Encoder Representations from Transformers) is then used to transform textual data into numerical representations. The BERT model tokenizes the text by separating it into individual words, or tokens, and assigning (encoding) unique numbers to each token.

B. UNIMODAL ENSEMBLE CLASSIFIER

Five deep learning models are designed for detecting depression using five different input modalities: facial landmarks, the eye scan path as an image, feature-extracted audio, text, and EEG data (similar to a multimodal model). The unimodal classification method employs distinct neural networks for each modality, namely CNN for landmarks and eye scan paths, BiLSTM-CNN for EEG, BiLSTM network for audio, and BilSTM with Bert Tokenizer for text. Figure 3 depicts a network with a similar architecture to multimodal networks; however, the characteristics of each modality are transmitted to distinct classification-dense layers. Again, some of the layers and neurons were modified for fine-tuning in order to maximize the statistical results (accuracy, F1 score, and AUC) of each network. Each model is trained on the corresponding dataset modality. The prediction was then soft ensembled, that is, the prediction of each model on a specific modality of the dataset was averaged to determine the final binary depression classification probability.

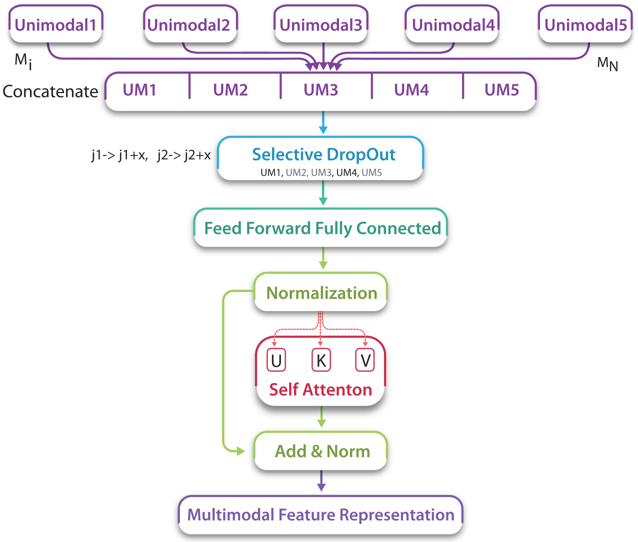


FIGURE 2. Illustration of selective dropout with attention.

C. SELECTIVE DROPOUT WITH NORMALIZATION AND ATTENTION

Modality mismatches are the fundamental issue while training multimodal neural networks on current datasets. For instance, the DAICWoz dataset includes raw audio, text, and extracted video data, while the Dvlog dataset only includes extracted audio and video characteristics. Similarly, the Modma dataset includes audio and ECG data. Challenges in neural network architecture arise from training a single model with all of the datasets. However, it is possible to resolve this problem by building a neural network that incorporates all of the modalities and then handling the modality that is missing from the classification process. We propose selective dropout with normalization and attention (fig. 2) to integrate tolerating missing modality and training one model with varied modality datasets. Similar to the regular dropout layer, the selective dropout layer allows just a certain set of nodes or each modality to be deleted by specifying a predetermined range. Once more, if there are enough modalities available, we can randomise by choosing to leave any of the offered modalities.

The integration of information from various modalities is made possible through the combination of multiple unimodal networks in a multimodal neural network. Let the set of N modalities is denoted by $\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \ldots, \boldsymbol{M}^{(N)}$, where each modality is characterized by a distinct set of input attributes. The unimodal networks are placed in isolation for each modality, where the output of the ith unimodal network is represented as $\boldsymbol{UM}^{(i)}$. The process of creating a multimodal representation $\boldsymbol{UM}^{(N+1)}$ involves merging the

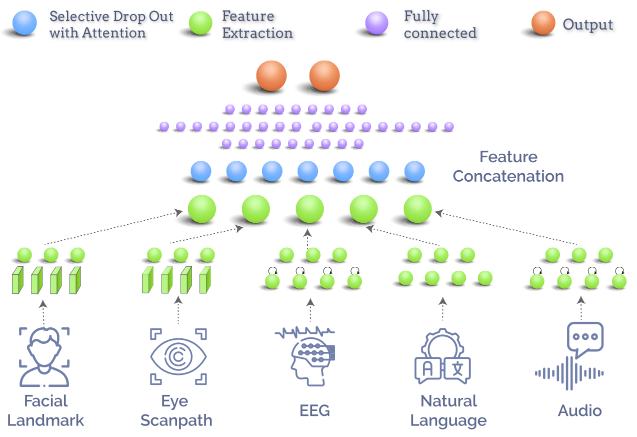


FIGURE 3. **Proposed multimodal classifier accommodating various modalities for depression detection.**

outputs of the unimodal networks through concatenation, that is $\boldsymbol{UM}^{(N+1)} = [\boldsymbol{UM}^{(1)}, \boldsymbol{UM}^{(2)}, \ldots, \boldsymbol{UM}^{(N)}]$. In order to address the potential consequences of absent modalities, a selective dropout layer has been implemented to selectively target the combined representation $\boldsymbol{UM}^{(N+1)}$. The dropout layer is designed to randomly eliminate a subset of nodes or modalities, with the specific subset being determined by a pre-established range $J_1 to J_1 + l$ as given in figure 2. The resulting amalgamated representation is subsequently fed into a fully connected feed-forward layer, which is equipped with weights $\boldsymbol{W}$ and bias $\boldsymbol{b}$. This process yields an intermediary representation $\boldsymbol{M}^{(N+2)}$ in eq 1.

$$\boldsymbol{M}^{(N+2)} = \boldsymbol{W} \sum \boldsymbol{UM}^{(i+1)} + \boldsymbol{b} \qquad (1)$$

A layer normalization (eq. 2) is placed after the dense layer to normalise the results and lessen the impact of missing modalities. Where M represents the input variable, and μ and σ are the mean and variance computed over the feature dimension of x. To prevent division by zero, the constant ϵ is introduced as a small value. Additionally, the parameters γ and β are incorporated as learnable scaling and shifting factors, respectively.

$$\text{Layer Normalization}(M) = \frac{M - \mu}{\sqrt{\sigma^2 + \epsilon}}\gamma + \beta \qquad (2)$$

Let the preceding neuron's maximum value be 5, to further illustrate the process. The preceding layer is used to build a weighted sum in a dense layer. We may therefore obtain a maximum of 25 modalities from 5 modalities (if the bias is zero and all weights are 1). However, by applying normalization both values will be 1. Thus, this process effectively mitigates the effect of missing modalities.

Subsequently, the Self-attention mechanism is employed to calculate the significance of each neuron by computing a weighted sum of values $\boldsymbol{U}$ using the similarity between keys or modalities $\boldsymbol{K}$ and a query vector $\boldsymbol{Q}$. Subsequent to the self-attention process, the resulting output is subjected to normalization via a scaling factor of $\frac{1}{\sqrt{d_k}}$, wherein d_k denotes the dimensionality of the key vectors as delineated in equation 3. The function 'softmax' is utilized to calculate the attention scores or weights.

$$\text{Self-Attention}(U, K, V) = \text{softmax}\left(\frac{MK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

Following that, the attention layer inputs and the previous normalization layer inputs are combined and subjected to batch normalization to generate a proposed multimodal fusion feature representation that can accommodate the absence of modalities during runtime.

Figure 3 illustrates the comprehensive model incorporating all modalities present in the dataset, utilizing the proposed techniques of selective dropout with attention and normalization. Nonetheless, this information could prove valuable in future instances. The datasets referenced in the previous section do not contain all of the modalities. For example, the DAIC-WOZ dataset does not contain any EEG information, thus it was dropped using selective dropout. The input modality is a necessary requirement for the implementation of a neural network using TensorFlow. Thus, we have provided a set of inputs consisting entirely of zeros. Nevertheless, this is inconsequential as the information is discarded during the phase of selective dropout. Implementation of the proposed selective dropout layer is presented as a class in Appendix A.

D. PROPOSED MULTIMODAL NEURAL NETWORK

In this section, we aim to construct a deep learning model for multimodal depression classification that incorporates modalities like facial landmarks, eye scan paths, EEG features, text features from the BERT tokenizer, and audio features from OpenSmile. The model is intended to exploit the unique characteristics of each modality in order to improve depression classification performance (fig 4). The model extracts features from previously mentioned modalities using a combination of convolutional neural networks (CNNs), long-short-term memory (LSTM) networks, and attention mechanisms. The proposed multimodal neural network comprises five input layers for each modality, namely face, eye, EEG, text, and audio. The dimensions of the face input layer are (340, 340, 3), eye scan path input shape is (256, 256, 3). The input layer for EEG has a shape of (1, 128, 21), whereas the input layer for text possesses a shape of (1200, 1) and has a shape of (238, 1).

Residual-inception-style convolution blocks are utilized for the purpose of image feature extraction. The convolutional blocks are composed of three parallel blocks of convolutional layers that are responsible for extracting features. The first (leftmost) part comprises two convolutional layers with a filter size of Fl, 3×3 kernel, and a rectified linear unit (ReLU) activation function, followed by a max-pooling layer with a pool size of 2*2 and a dropout layer with a rate of 0.25 to remove some inputs to avoid overfitting. The second part is composed of convolutional layer with kernel sizes of 1*1 and

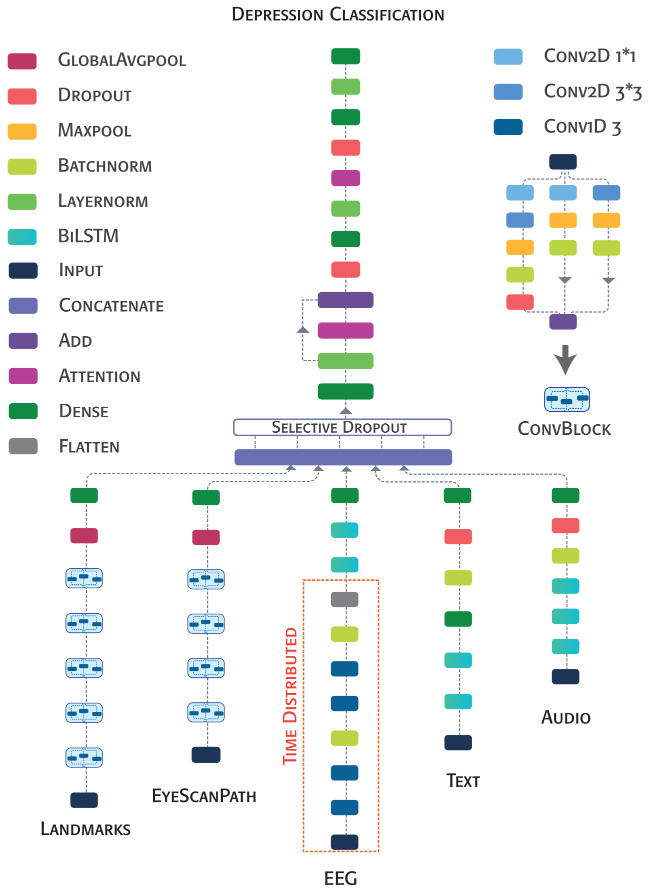


FIGURE 4. Proposed multimodal neural network.

3*3, then concatenated in a max-pooling layer with a pool size of 2*2. The third part comprises a convolutional layer with a kernel size of 3*3, followed by a two-dimensional max-pooling layer. The add layer is used to aggregate the output of each block, which is subsequently passed as input to the succeeding block.

The first modality is a convolutional network that receives facial inputs with the shape (340, 340, 3) and implements a series of convolution operations, which were previously referred to as a convolution block. Following this, five convolutional blocks with filter sizes of 32, 64, 128, 256, and 512 are applied. Then, a global average pooling layer is applied, which reduces the output's spatial dimensions, followed by a dense layer with a size of 128 and a ReLU activation function applied to the output.

The second modality is similar to the first modality, although with the distinction of receiving eye scan path inputs with dimensions of (256, 256, 3). This modality comprises four convolutional blocks that incorporate filter sizes of 32, 64, 128, and 256, respectively, in addition to a global average pooling layer. The resultant of the global average pooling layer is passed through a dense layer that has a dimension of 128 and is activated by the Rectified Linear Unit (ReLU) function. The third modality comprises time-distributed convolutional layers, which are used to process electroencephalogram (EEG) inputs with a shape of (1, 128, 21). The aforementioned layer implements a filter of 3, utilizing a kernel size of 32, a ReLU activation function, and a stride value of 2. Subsequently, two convolutional layers that are temporally distributed are used, accompanied by batch normalization with axis value −1. The output is processed by a flattened layer followed by two LSTM layers, each with a size of 128 and utilizing a GELU activation function. The output is subjected to a ReLU activation function and a dense layer of 128 units.

The fourth modality comprises of Long Short-Term Memory (LSTM) layers, which are used to process textual inputs with a shape of (1200, 1). The architecture of the layer consists of two LSTM layers with dimensions of 128 and 256, respectively. Additionally, the layer incorporates a dropout rate of 0.1. The output is subjected to a ReLU activation function and a dense layer of 128 units. The produced output is sent to batch normalization, followed by dropout with a rate of 0.25, and subsequently, another dense layer with a size of 128 and a ReLU activation function.

The fifth modality also uses LSTM layers that are similar to those used to process sequential data but with audio inputs that have the shape (73, 1) instead. The neural network architecture comprises three LSTM layers, each with a size of 128 and 256 and a dropout rate of 0.1. The output is subjected to batch normalization, dropout regularization with a rate of 0.25, a fully connected layer with a dimension of 128 and a rectified linear unit activation function.

The resultant output of each modality is merged together and subsequently subjected to selective dropout, multiple dense layers, layer normalization, and attention mechanisms. The utilization of the attention mechanism will facilitate the model in selectively directing its focus towards the most pertinent characteristics across various modalities. The technique of selective dropout is employed to handle missing modalities through the selective removal of features. The output dense layer has a sigmoid activation function for binary classification. The training of the model is conducted by utilizing the binary cross-entropy loss function with the Adam optimizer. The learning rate is 0.001, and 80 percent of the data is used for training and 20 percent for testing.

E. UNCERTAINTY

Monte Carlo Dropout (MC Dropout) and Sparse Gaussian Process (SGNP) are frequently employed methodologies in the field of machine learning to enhance model resilience and approximate uncertainty. These methods can be applied to the previously developed classification model for multimodal depression in order to gain a deeper understanding of the model's predictions and possibly enhance its accuracy. The advantage of these methods is that they can provide insight into the significance of any out-of-distribution data and the position of a prediction within a distribution.

The fundamental concept of SNGP is to enhance the distance understanding of a deep learning classifier through the implementation of uncomplicated modifications to the model. The distance understanding of a neural network refers to its ability to accurately estimate the probability of an outcome based on the distance within the training and testing data. As described by Liu et al. [44], spectral normalization can be applied to the last epoch of training in any neural network to reduce the covariance shift. For incorporating SNGP into the existing multimodal neural network described in the previous section, instead of batch normalization, spectral normalization is used with successive Gaussian processes (GP) in the dense layer after the concatenation. More formally, spectral normalization [45] controls the Lipschitz Constantin in multimodal neural network f by limiting the spectral values of each layer $x : M_{in} \rightarrow M_{out}$. Where $\sigma(x)$ is the spectral norm of previous layer as matrix X (L_2 normalization of X) given in equation 4

$$\sigma(x) = \max_{M:M\neq 0} \frac{\|XM\|_2}{\|M\|_2} = \max_{\|M\|_2 \leq 1} \|XM\|_2 \tag{4}$$

For the expected output probability $E(p(x))$, Monte Carlo sampling as in equation 5 can be used. However, according to Lu et al. [46] it is latency-prone thus we have also used the mean-variance method as described in equation 6. In both of these equations $Logit_m(X)$ is the posterior mean of layer X.

$$E(p(x)) = \sum_{m=1}^{M} \text{softmax}(Logit_m(x)) \tag{5}$$

$$E(p(x)) \approx \text{softmax}\left(\frac{Logit_m(x)}{\sqrt{1+\lambda\sigma^2(x)}}\right) \tag{6}$$

Monte Carlo Dropout has also been employed during the inference stage to approximate the model's level of uncertainty. The application of dropout to the input layer and the execution of multiple forward passes through the network with distinct dropout masks can lead to the attainment of the desired outcome. Subsequent to the network's output, an averaging process is conducted across the various runs, thereby providing an improved approximation of the model's uncertainty as equation 7. Where $p(\mathbf{x_i})$ is the predicted output, $\mathcal{D}$ is the training data, θ is the neural network parameters and $p(\theta|\mathcal{D})$ is the posterior distribution over the weights.

$$p(\mathbf{x_i}) = \int p(\mathbf{x_i}, \theta)p(\theta|\mathcal{D})d\boldsymbol{\theta} \tag{7}$$

However, the complete layered equation with previously proposed selective dropout incorporates both resilience for missing modalities and uncertainty. The output of these processes is summarized in the equation 8.

$$\mathbf{M}^{i+1} = \mathbf{M}^{i} + LNorm(SA(SNGP(\text{SD}(\mathbf{x}_i)))) \tag{8}$$

where $\mathbf{M}^i$ is the unimodal features, extracted from each modalities, $\mathbf{M}^{i+1}$ is the concatenated state, *SNGP* is the spectral normalized Gaussian process layer, *SD* is the selective dropout, *SA* is the self-attention layer, and *LNorm* is the normalization layer.

IV. EXPERIMENTS AND EVALUATION

A. EXPERIMENTAL SETUP

The experiment was carried out on several machines, including a system comprises of both dual NVidia T4 GPU and RTX 4090 GPU. The Keras framework, using the TensorFlow library as its backend, is implemented using the Python programming language. Subsequently, the neural network stated in the methodology has been implemented correspondingly. The optimization of multiple parameters and layers was carried out to attain optimal classification performance.

B. DATASET

There are numerous datasets for the detection of depression using different modalities, including text, EEG, video, and audio. To show the effectiveness of the proposed system three openly available datasets were taken into account.

1) D-vlog DATASET

D-vlog [47] is a multimodal vlog dataset composed of 961 YouTube vlogs. The entire duration of the dataset is approximately 160 hours, with an output label of "depressed" or "not depressed". There are a total of 555 depressed data points and 406 non-depressed data points, with average durations of 640 seconds and 536 seconds, respectively. 25 acoustic features (OpenSmile [42] LLD features) and 136 visual features (68 facial landmarks) are recorded in the feature-extracted dataset for each second (1 FPS) of the vlog. The acoustic characteristics include zero crossing, the sum of the auditory spectrum, Mel-Frequency Cepstral Coefficients (MFCC), Root Mean Square (RMS), spectral roll-off, and spectral flux, among others.

2) DAIC-WOZ

The Distress Analysis Interview Corpus: Wizard of Oz (DAIC-WOZ) [48], [49] dataset consists of text, video, and acoustic conversations between participants and an automated interviewer. The interview queries are derived from depression's medical symptoms. Until now, two versions of this dataset have been made available, and for the purposes of this study, either the extended DAIC-WOZ is used. The dataset contains a total of 189 sessions; however, due to technical issues (mentioned in the dataset documentation), seven sessions have been removed, resulting in a duration of 182 sessions, with 88 depressed sessions and 94 non-depressed sessions. The dataset includes 68 facial landmarks as video features with 2D and 3D points, unprocessed audio files, audio features converted using the Collaborative Voice Analysis Repository (COVAREP), and text of variable duration.

3) MODMA

The Multi-modal Open Dataset for Mental-Disorder Analysis (MODMA) [50] is a dataset that has been labelled using the PHQ-9 scale. There are 24 participants with depression diagnoses and 29 participants in good mental health. Major depressive disorder (MDD) patients were chosen from Lanzhou University Second Hospital in Gansu, China, based on the direction of at least one psychiatrist. The dataset includes an aligned electroencephalogram (EEG) obtained from 128 channels of recording, along with the corresponding raw audio data.

C. EVALUATION METRICS

Performance measures adopted for this work are accuracy, precision, recall and F1 score. These measures ultimately prove the performance reliability of the proposed multimodal technique. A brief explanation of the evaluation metrics used for this work are as follows:

1) PRECISION

Precision is the fraction of correct prediction (TP) in all positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

TP and FP in equation 9 are True Positive and False Positive respectively.

2) RECALL

Recall is the fraction of correct prediction (TP) in all relevant predictions.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

TP and FN in equation 10 are True Positive and False Negative respectively.

3) F1-SCORE

F1-score is the weighted average of Precision and Recall (eq 11. The overall performance of the proposed architecture can be evaluated better by the F1-score as found in the course of study of this work.

$$\text{F1-score} = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \tag{11}$$

4) S1-SCORE

Given a model m, uncertainty quantifier m and test set I as defined by Weiss et al. [51], the $S1$Score is defined as equation 12.

$$S_1(m, u, I) = \frac{2}{\frac{obj^{+} - obj^{-}}{obj^{(}Im) - obj^{-}} * \Delta(I)^{-1}} \tag{12}$$

where $obj-$ and $obj+$ are the lower and upper bounds used for normalizing the objective function. It also measures the supervisors' acceptance rate.

5) COEFFICIENT OF VARIATION

The use of the coefficient of variation (CV) is beneficial in estimating uncertainty in neural networks as it offers a means to quantify the relative variability of the anticipated values in relation to their mean. Stated differently, the method quantifies the level of unpredictability in the forecast through the computation of the proportion between the predicted values' standard deviation and their average. An elevated coefficient of variation (CV) is indicative of an increased degree of unpredictability or fluctuation in the predicted, whereas a reduced CV suggests a more assured or consistent prediction.

$$\text{CV}(y_{\text{true}}, y_{\text{pred}}) = \frac{\sqrt{\text{variance}|(y_{true} - y_{pred}|)}}{\text{mean}(|y_{true} - y_{pred}|)} \tag{13}$$

The variables y_{true} and y_{pred} represent the actual and predicted values, respectively. The aforementioned function partitions the subtraction of those variables into their respective mean and variance components along the second axis. The computation of the standard deviation involves taking the square root of the variance. Ultimately, the coefficient of variation is determined by dividing the standard deviation by the absolute mean. The output of the process is the average coefficient of variation across the batch as given in equation 13. Appendix B presents the Python program for the metrics.

D. EXPERIMENTS AND COMPARISONS

1) EPOCH-WISE RESULTS OF THE PROPOSED METHOD WITH SNGP

In this section, we demonstrate the improved performance of the proposed model 5, SNPG + Selected Dropout and Attention (SD), on three distinct datasets: dvlog, DAICWoz, and modma. Each dataset's epoch-wise loss and F1 Score are represented as line graphs in six figures illustrating the experimental findings.

Figure 5a depicts the test-train loss for the dvlog dataset, while Figure 5b depicts the test-train F1-Score. The loss and F1 Score values are comparatively high in the initial epochs, indicating suboptimal performance. Nonetheless, as the training continues, we observe a substantial increase in both metrics. By the 100th epoch, the loss has been reduced to 0.0029 and the F1 Score has reached 0.9496, demonstrating the efficacy of the proposed model. Figure 5c depicts the test-train loss and Figure 5d depicts the test-train F1 Score for the modma dataset. Over the training epochs, we observe a gradual decline in the loss and an increase in the F1 Score. At the 100th epoch, the loss is minimized to 0.1305 and the F1 Score is maximized at 0.9468, implying that the proposed SNPG + Selected Dropout and Attention model is effective.

Figure 5e depicts the test-train loss, while Figure 5f depicts the test-train F1 Score for the DAICWoz dataset. Similar to the dvlog dataset, we observe a progressive decrease in loss and an increase in F1 Score during the training procedure. Notably, by the 100th epoch, the loss has been reduced to

0.0052 and the F1 Score has increased to 0.9455, indicating a significant advance over the initial epochs.

2) COMPARISION OF UNIMODAL AND MULTIMODAL CLASSIFICATION

The presented table showcases the effects of selective dropout (SD) with attention in comparison to unimodal and multimodal methodologies across three distinct datasets, Dvlog, DAIC WOZ, and MODMA. The evaluation of each model's test performance is based on several metrics, including accuracy, recall, precision, F1 score, S1 score, and CV.

In Dvlog dataset, the unimodal ensemble approach yields a testing accuracy of 80.19%, F1 score of 84.26%, and S1 score of 79.56%. Meanwhile, the multimodal individual training approach achieves an accuracy of 81.02%, F1 score of 84.23%, and S1 score of 81.97%. Since the multimodal approaches incorporate prior training in all datasets, it achieves better results than other methods

On the other hand, in DAIC WOZ dataset, the unimodal ensemble approach achieves an testing accuracy of 80.54%, recall of 78.13%, precision of 80.96%, F1 score of 79.52%, and S1 score of 79.91%. The multimodal individual training approach attains an accuracy of 81.99%, recall of 81.22%, precision of 82.01%, F1 score of 81.61%, and S1 score of 80.99%. The aforementioned outcome suggests that the utilization of various modalities can considerably enhance the efficacy of the model concerning this particular dataset.

On the MODMA dataset, the unimodal ensemble approach achieves an testing accuracy of 79.15%, recall of 81.52%, precision of 80.11%, F1 score of 80.81%, and S1 score of 78.82% while the multimodal individual training approach attains an accuracy rate of 82.79%, recall rate of 84.79%, precision rate of 83.75%, F1 score of 84.27%, and S1 score of 97.69%. Multiple evaluation metrics, including accuracy, recall, precision, F1 score, and S1 score, demonstrate that the multimodal approach outperforms the unimodal ensemble approach. The findings signify that the use of multiple modalities can significantly improve the model's performance by combining knowledge of datasets.

The multimodal approach is enhanced through incorporating selective dropout and transfer learning techniques. This approach tested on the Dvlog dataset results in an accuracy of 89.37%, recall of 86.77%, precision of 88.58%, F1 score of 87.67%, and S1 score of 89.78%. The approach applied on the DAIC WOZ dataset yields an accuracy of 87.55%, a recall of 89.17%, a precision of 89.26%, an F1 score of 89.21%, and an S1 score of 87.87%. Finally, the approach using the MODMA dataset attains an accuracy of 88.68%, recall of 87.59%, precision of 89.47%, F1 score of 88.52%, and S1 score of 88.54%. Thus, incorporating selective dropout and transfer learning into the multimodal approach may improve the model's performance on the aforementioned datasets. Utilizing a multimodal approach in conjunction with selective dropout and transfer learning produces the best results across all three datasets, highlighting the importance of the proposed methodology.

E. COMPARISON OF SNGP AND MC DROPOUT ON THREE DATASETS WITH UNCERTAINTY APPROXIMATION METRICS

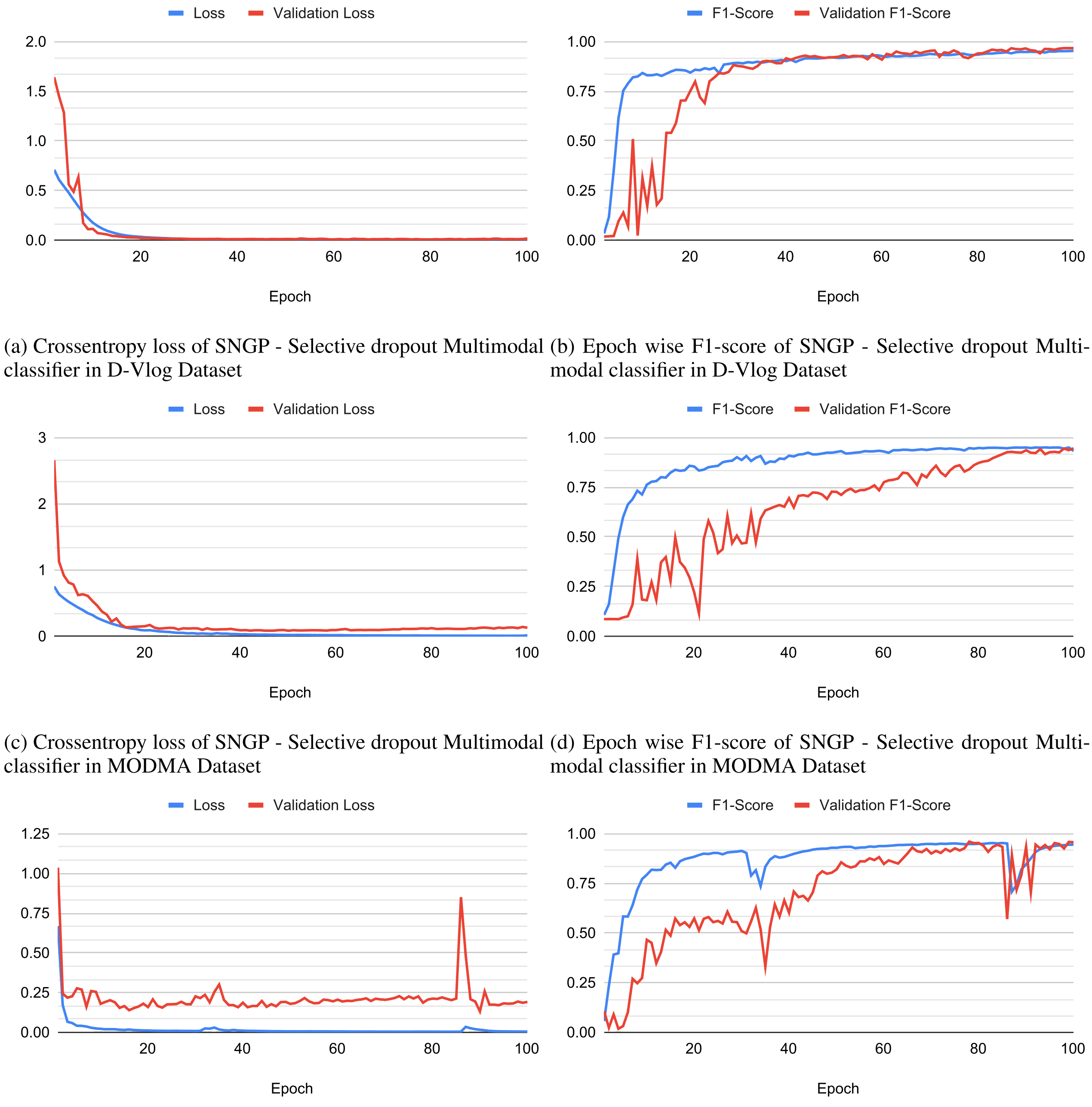
Table 1 displays the use of selective dropout (SD) alongside MC dropout and SNGP in the multimodal model with transfer learning. The impact of these techniques can be observed by comparing the results to the data-discussed previous models. The S1 Score is a metric that assesses the precision of the model's predictions regarding the minority class. CV assesses the variation in the model's predictions. The lower the CV, the more consistent the model's predictions will be. When SD is combined with MC dropout in the multimodal model with transfer learning, the S1 Score and CV of the model are improved across all three datasets. On the DIAC WOZ data set, for instance, the S1 Score for multimodal model without SD rises from 0.8786 to 0.9299 when SD is combined with MC dropout in the multimodal model. In a similar fashion, the CV falls from 0.0751 to 0.0589. These enhancements suggest that the model is now more capable of consistently predicting the minority class. When SD is combined with SNGP in a multimodal model with transfer learning, the S1 Score and CV of the model is improved further when SD and MC dropout are used. On the DAIC WOZ dataset, for instance, the S1 Score rises from 0.9299 to 0.9257, while the CV falls from 0.0589 to 0.0484. The SNGP method improves the estimation of uncertainty in model predictions. Consequently, the model is able to predict the minority class with greater precision.

The results demonstrate that the use of selective dropout in conjunction with either MC dropout or SNGP can enhance the performance of the multimodal model with transfer learning. Particularly, the SNGP technique can further improve the model's performance by enhancing the estimation of uncertainty in the model's predictions. Overall, the enhancements to the S1 Score and CV metrics indicate that the models are now able to predict the minority class with greater consistency, a crucial factor in a variety of practical applications.

F. DISCUSSION

Table 2 compares the performance of the various modalities in the DAIC WOZ dataset (Text, Audio, Eye Scan Path, and Facial Landmark). With an accuracy of 0.9643, F1-Score of 0.9430, and S1 Score of 0.9499, the findings show that the combination of Text and Eye Scan Path with Facial Landmark (T + E + F) produces the greatest overall performance. In contrast to integrating audio with facial landmarks (A + F), which has lower accuracy and precision values, this combination demonstrates that the integration of eye scan data with face landmarks improves the model's ability to identify depression. Interestingly, the T + E + F combination outperforms the T + A + E + F combination, indicating that adding Audio may have a detrimental effect on the results.

On the other hand, Table 3 represents a comparison of Recall, Precision, and F1 Score with existing studies on



(a) Crossentropy loss of SNGP - Selective dropout Multimodal classifier in D-Vlog Dataset

(b) Epoch wise F1-score of SNGP - Selective dropout Multimodal classifier in D-Vlog Dataset

(c) Crossentropy loss of SNGP - Selective dropout Multimodal classifier in MODMA Dataset

(d) Epoch wise F1-score of SNGP - Selective dropout Multimodal classifier in MODMA Dataset

(e) Crossentropy loss of SNGP - Selective dropout Multimodal classifier in DAIC-WOZ Dataset

(f) Epoch wise F1-score of SNGP - Selective dropout Multimodal classifier in DAIC-WOZ Dataset

FIGURE 5. Epoch-wise F1Score and cross-entropy loss in D-VLog, MODMA and DAIC-WOZ datasets.

DAIC WOZ and MODMA datasets. The proposed approach acquires an accuracy of 0.9286, a recall of 0.9535, and an F1 score of 0.9409 on the DAIC-WOZ dataset. It should be noted that Du et al. [52] method achieves an F1 score of 0.746, which is due to the prediction of PHQ scale. However, F1 scores of 0.931 and 0.920, respectively, were achieved by Zhao et al. [53] and Niu et al. [54], which are equivalent to the proposed approach. Both Lam et al. [55] and Zhao et al. [56] obtained F1 scores of 0.8895 and 0.916, which are comparable to the performance of the proposed method. Again, the proposed approach performs equivalent to the existing approaches on the MODMA dataset, with a precision of 0.9578, recall of 0.9564, and F1 score of 0.9453. Zhang et al. [58] obtained an F1 score of 0.9157, which is lower than the output of the proposed models. On the other hand, Zhao et al. [53] outperformed the proposed approach

TABLE 1. Model performance on various datasets.

Model	Dataset	Accuracy	Recall	Precision	F1 Score	S1 Score	CV
Unimodal Ensemble	DVlog	0.8019	0.8187	0.8679	0.8426	0.7956	0.1177
	DAIC WOZ	0.8054	0.7813	0.8096	0.7952	0.7991	0.1441
	MODMA	0.7915	0.8152	0.8011	0.8081	0.7882	0.1167
Multimodal individual training	DVlog	0.8102	0.8387	0.8459	0.8423	0.8198	0.1196
	DAIC WOZ	0.8199	0.8122	0.8201	0.8161	0.8099	0.1328
	MODMA	0.8279	0.8479	0.8375	0.8427	0.9769	0.1120
Multimodal + SD+ transfer learning	DVlog	0.8937	0.8677	0.8858	0.8767	0.8978	0.0739
	DAIC WOZ	0.8755	0.8917	0.8926	0.8921	0.8787	0.0751
	MODMA	0.8868	0.8759	0.8947	0.8852	0.8854	0.0704
Multimodal + SD-Norm-Att+ transfer learning + MC dropout	DVlog	0.9430	0.9201	0.9469	0.9333	0.9395	**0.0416**
	DAIC WOZ	0.9381	0.9211	0.9482	0.9345	0.9299	0.0590
	MODMA	0.9510	0.9399	0.**9432**	0.9415	**0.9413**	0.0411
Multimodal + SD-Norm-Att+ transfer learning + SNGP	DVlog	**0.9498**	**0.9241**	0.**9524**	**0.9381**	**0.9397**	0.0425
	DAIC WOZ	**0.9507**	**0.9286**	0.**9535**	**0.9409**	**0.9258**	**0.0484**
	MODMA	**0.9578**	**0.9564**	0.9345	**0.9453**	0.9374	**0.0384**

TABLE 2. Performance comparison of different modalities [Text (T), Audio (A), Eye Scan Path (E), and Facial Landmark (F)] in the DAIC WOZ dataset.

Modality	Accuracy	Recall	Precision	F1-Score	S1 Score
T + A	0.9413	0.8891	0.9503	0.9187	0.9249
T + E + F	0.9643	0.9210	0.9662	0.9430	0.9499
A + F	0.9174	0.9068	0.8727	0.8894	0.9124
T + A + E + F	0.9507	0.9286	0.9535	0.9409	0.9258

TABLE 3. Comparison of proposed model with existing methods on DAIC-WOZ and MODMA Datasets.

Dataset	Model	Precision	Recall	F1 Score
DAIC-WOZ	Proposed	0.9286	0.9535	**0.9409**
	Du (2023) [52]	-	-	0.746
	Zhao (2021) [53]	0.930	0.938	0.931
	Niu (2021) [54]	0.920	0.920	0.920
	Lam (2019) [55]	0.870	0.910	0.8895
	Zhao (2019) [56]	0.912	0.920	0.916
MODMA	Proposed	0.9578	0.9564	0.9453
	Du (2023) [52]	-	-	0.857
	Deng (2022) [57]	0.9507	0.9366	0.944
	Zhang (2022) [58]	0.9394	0.8932	0.9157
	Zhao (2021) [53]	0.981	0.974	**0.977**
	Zhao (2019) [56]	0.929	0.935	0.932

in this particular dataset with a high F1 score of 0.977. This fluctuation in performance can be further explained by two factors. Firstly, many of the state-of-the-art methods treat depression detection as a PHQ predictor problem. However, to simplify and show the impact of multimodal learning with selective dropout, we have converted to the problem of binary classification, where a PHQ score > 7 is treated as 1, otherwise 0. This reduction in output class overall simplifies the problem for the models as well as improves their accuracy. Since each model is first trained on three datasets and tested all together in the end, this transfer of learning impacts the model's overall performance. The model is sequentially trained on each dataset, with the weights saved after each training. Then, the saved model is used as a starting point for training on the next dataset. This way, the model accumulates knowledge from all datasets, and during testing, it leverages this collective knowledge for improved performance on individual datasets. This improvement could improve the overall performance of the experiments; however, this approach may cause uncertainty and overfit. Thus, we have put more emphasis on learning about the uncertainty of transfer learning with different input sets. However, this process is common in the literature to use a pre-trained model for completely new input types or even modalities like brain tumours, lung cancer, anomaly detection, etc. By incorporating uncertainty reduction and measuring techniques, a pre-trained model with a similar feature convergence technique might improve classification in other fields of study.

V. CONCLUSION

The proposed methodology focuses on the detection of multimodal depression using multiple datasets. Depression is a prevalent mental illness that has a significant impact on the lives of individuals and on society as a whole. Depression symptoms can manifest differentially across various modalities, including text, EEG, audio, and visual signals, highlighting the significance of multimodality in detecting depression. However, when using multiple multimodal datasets to train a single neural network, the prevalence of absent modalities may hamper training and impact the model's overall performance. In order to resolve this difficulty, we present a novel strategy that incorporates selective dropout, attention mechanisms, and normalization techniques. The proposed method permits the training of multimodal datasets with absent modalities by selectively omitting specific modalities during training while still effectively utilizing the available information. This method increases the model's performance by enhancing its capacity to capture complex relationships between different modalities. In addition, we employ SNGP and MC (Monte Carlo) dropout algorithms to reduce uncertainty during the training procedure. By approximating the uncertainty associated with the missing data, these techniques enable the model to manage missing modalities more robustly. As demonstrated by the experimental findings, reducing uncertainty improves the model's precision, F1 score, and S1 score. This research contributes to the advancement of multimodal deep-learning methods and has the potential to improve depression detection and assessment of mental health in practical applications.

APPENDIX. SUPPLEMENTARY PYTHON CODES

A. SELECTIVE DROPOUT

```

class Selective_drops(tf.keras.layers.Layer):
  def __init__(self, selection):
    super(Selective_drops, self).__init__()
    self.selection = selection
    self.one_unit= tf.ones_initializer()
    self.zero_unit= tf.zeros_initializer()

  def build(self,input_shape):

    self.shapes=input_shape[-1]
    self.kernel = tf.Variable(initial_value=self.one_unit(shape=(input_shape[-1]),dtype=’float32’), trainable = False)

  def update(self):
    self.kernel = self.kernel.assign(self.one_unit(shape=(self.shapes),dtype=’float32’))
    selection= np.array(self.selection)
    rand=np.random.randint(0,selection.shape[0]-1)
    selected_low = selection[rand]
    selected_high = selection[rand+1]-1
    self.kernel= self.kernel[selected_low:selected_high].assign(self.zero_unit(shape=selected_high-selected_low,dtype=’float32’))

    #tf.print(self.kernel,summarize=-1)

  def call(self,inputs):
     self.update()
     new_tensor= tf.convert_to_tensor(self.kernel)
     return tf.math.multiply(inputs, new_tensor)
```

Listing. 1. **Python example of selective dropout.**

B. COEFFICIENT OF VARIATION

```

def coefficient_of_variation(y_true, y_pred):
    mean = np.mean(np.abs(y_true,y_pred))
    variance = np.var(np.abs(y_true,y_pred))
    std_dev = tf.math.sqrt(variance)
    cv = std_dev / tf.abs(mean)
    return tf.reduce_mean(cv)
```

Listing. 2. **Python example of coefficient of variation.**

REFERENCES

[1] *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, vol. 5, no. 5, Amer. Psychiatric Assoc., Washington, DC, USA, 2013.

[2] J. E. R. Bernard, "Depression: A review of its definition," *MOJ Addiction Med. Therapy*, vol. 5, no. 1, pp. 6–7, Jan. 2018.

[3] (Mar. 31, 2023). *Depressive Disorder (Depression)*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[4] S. Dattani, H. Ritchie, and M. Roser, 2021, "Mental health," Our World in Data. [Online]. Available: https://ourworldindata.org/mental-health

[5] R. Rajendram, V. B. Patel, and V. R. Preedy, "Recommended resources on the neuroscience of depression: Genetics, cell biology, neurology, behavior, and diet," in *The Neuroscience of Depression*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 531–537.

[6] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Internal Med.*, vol. 16, no. 9, pp. 606–613, Sep. 2001.

[7] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disorders*, vol. 114, nos. 1–3, pp. 163–173, Apr. 2009.

[8] L. S. Radloff, "The CES-D scale: A self-report depression scale for research in the general population," *Appl. Psychol. Meas.*, vol. 1, no. 3, pp. 385–401, Jun. 1977.

[9] J. I. Sheikh and J. A. Yesavage, "Geriatric depression scale (GDS): Recent evidence and development of a shorter version," *Clin. Gerontologist, J. Aging Mental Health*, vol. 5, nos. 1–2, pp. 165–173, 1986.

[10] P. Spinhoven, J. Ormel, P. P. A. Sloekers, G. I. J. M. Kempen, A. E. M. Speckens, and A. M. V. Hemert, "A validation study of the hospital anxiety and depression scale (HADS) in different groups of Dutch subjects," *Psychol. Med.*, vol. 27, no. 2, pp. 363–370, Mar. 1997.

[11] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, "SenseMood: Depression detection on social media," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 1–15.

[12] F. M. Shah, F. Ahmed, S. K. Saha Joy, S. Ahmed, S. Sadek, R. Shil, and Md. H. Kabir, "Early depression detection from social network using deep learning techniques," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jun. 2020, pp. 823–826.

[13] J. Hussain, F. A. Satti, M. Afzal, W. A. Khan, H. S. M. Bilal, M. Z. Ansaar, H. F. Ahmad, T. Hur, J. Bang, J.-I. Kim, G. H. Park, H. Seung, and S. Lee, "Exploring the dominant features of social media for depression detection," *J. Inf. Sci.*, vol. 46, no. 6, pp. 739–759, Dec. 2020.

[14] K. A. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on Twitter data," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2021, pp. 960–966.

[15] A. Zafar and S. Chitnis, "Survey of depression detection using social networking sites via data mining," in *Proc. 10th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2020, pp. 88–93.

[16] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella, and S. V. Alluri, "Real-time acoustic based depression detection using machine learning techniques," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (ic-ETITE)*, Feb. 2020, pp. 1–6.

[17] S. Alghowinem, T. Gedeon, R. Goecke, J. F. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 133–152, Jan. 2023.

[18] J. Zhu, Z. Wang, T. Gong, S. Zeng, X. Li, B. Hu, J. Li, S. Sun, and L. Zhang, "An improved classification model for depression detection using EEG and eye tracking data," *IEEE Trans. Nanobiosci.*, vol. 19, no. 3, pp. 527–537, Jul. 2020.

[19] H. Akbari, M. T. Sadiq, M. Payan, S. S. Esmaili, H. Baghri, and H. Bagheri, "Depression detection based on geometrical features extracted from SODP shape of EEG signals and binary PSO," *Traitement du Signal*, vol. 38, no. 1, pp. 13–26, Feb. 2021.

[20] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 1–8.

[21] L. Yang, "Multi-modal depression detection and estimation," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2019, pp. 26–30.

[22] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *The Handbook of Multimodal-Multisensor Interfaces*, vol. 2. New York, NY, USA: Association for Computing Machinery, 2018.

[23] J. P. Gray, V. I. Müller, S. B. Eickhoff, and P. T. Fox, "Multimodal abnormalities of brain structure and function in major depressive disorder: A meta-analysis of neuroimaging studies," *Amer. J. Psychiatry*, vol. 177, no. 5, pp. 422–434, May 2020.

[24] S. Shimpi, S. Thombre, S. Reddy, R. Sharma, and S. Singh, ''Multimodal depression severity prediction from medical bio-markers using machine learning tools and technologies,'' 2020, *arXiv:2009.05651*.

[25] A. H. Yazdavar, M. S. Mahdavinejad, G. Bajaj, W. Romine, A. Sheth, A. H. Monadjemi, K. Thirunarayan, J. M. Meddar, A. Myers, J. Pathak, and P. Hitzler, ''Multimodal mental health analysis in social media,'' *PLoS ONE*, vol. 15, no. 4, Apr. 2020, Art. no. e0226248.

[26] S. Rasipuram, J. H. Bhat, A. Maitra, B. Shaw, and S. Saha, ''Multimodal depression detection using task-oriented transformer-based embedding,'' in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2022, pp. 1–4.

[27] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, ''Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors,'' *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 478–490, Oct. 2018.

[28] N. Seneviratne and C. Espy-Wilson, ''Multimodal depression severity score prediction using articulatory coordination features and hierarchical attention based text embeddings,'' in *Proc. Interspeech*, Sep. 2022, pp. 3353–3357.

[29] T. R. Mim, M. Amatullah, S. Afreen, M. A. Yousuf, S. Uddin, S. A. Alyami, K. F. Hasan, and M. A. Moni, ''GRU-INC: An inception-attention based approach using GRU for human activity recognition,'' *Exp. Syst. Appl.*, vol. 216, Apr. 2023, Art. no. 119419.

[30] M. M. Hossain, M. M. Hasan, M. A. Rahim, M. M. Rahman, M. A. Yousuf, S. Al-Ashhab, H. F. Akhdar, S. A. Alyami, A. Azad, and M. A. Moni, ''Particle swarm optimized fuzzy CNN with quantitative feature fusion for ultrasound image quality identification,'' *IEEE J. Translational Eng. Health Med.*, vol. 10, pp. 1–12, 2022.

[31] N. F. Aurna, M. A. Yousuf, K. A. Taher, A. K. M. Azad, and M. A. Moni, ''A classification of MRI brain tumor based on two stage feature level ensemble of deep CNN models,'' *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105539.

[32] M. Hamilton, ''Frequency of symptoms in melancholia (depressive illness),'' *Brit. J. Psychiatry*, vol. 154, no. 2, pp. 201–206, Feb. 1989.

[33] M. Baquero and N. Martín, ''Depressive symptoms in neurodegenerative diseases,'' *World J. Clin. Cases*, vol. 3, no. 8, pp. 682–693, 2015.

[34] J. F. Greden, ''Physical symptoms of depression: Unmet needs,'' *J. Clin. Psychiatry*, vol. 64, no. 7, pp. 5–11, 2003.

[35] U. Arioz, U. Smrke, N. Plohl, and I. Mlakar, ''Scoping review on the multimodal classification of depression and experimental study on existing multimodal models,'' *Diagnostics*, vol. 12, no. 11, p. 2683, Nov. 2022.

[36] B. Yin, H. Xu, and C. Zhao, ''Research on multimodal depression detection method based on BiGRU and BiLSTM,'' in *Proc. 5th Int. Conf. Mechatronics Comput. Technol. Eng. (MCTE)*, Dec. 2022, pp. 1–9.

[37] R. Singhal, S. Srivatsan, and P. Panda, ''A novel multimodal method for depression identification,'' *J. Trends Comput. Sci. Smart Technol.*, vol. 4, no. 4, pp. 215–225, Nov. 2022.

[38] S. T. Mantri, D. D. Patil, P. Agrawal, and V. Wadhai, ''Real time multimodal depression analysis,'' *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, no. 9, pp. 1–7, 2019.

[39] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, ''Dynamic multimodal measurement of depression severity using deep autoencoding,'' *IEEE J. Biomed. Health Informat.*, vol. 22, no. 2, pp. 525–536, Mar. 2018.

[40] R. Li, X. Lu, T. Pan, D. Shi, Y. Liu, and J. Yuan, ''Design on modeling of multimodal depression aided diagnosis from psychological perspective,'' in *Proc. 8th Int. Conf. Orange Technol. (ICOT)*, Dec. 2020, pp. 1–5.

[41] S. Ketelhut, E. Wehlan, G. Bayer, and R. G. Ketelhut, ''Influence of initial severity of depression on the effectiveness of a multimodal therapy on depressive score, heart rate variability, and hemodynamic parameters,'' *Int. J. Environ. Res. Public Health*, vol. 19, no. 16, p. 9836, Aug. 2022.

[42] F. Eyben, M. Wöllmer, and B. Schuller, ''Opensmile: The Munich versatile and fast open-source audio feature extractor,'' in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[43] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, and L. Parkkonen, ''MEG and EEG data analysis with MNE-Python,'' *Frontiers Neurosci.*, vol. 7, p. 267, Mar. 2013.

[44] J. Liu, Z. Lin, S. Padhy, D. Tran, T. B. Weiss, and B. Lakshminarayanan, ''Simple and principled uncertainty estimation with deterministic deep learning via distance awareness,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7498–7512.

[45] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, ''Spectral normalization for generative adversarial networks,'' 2018, *arXiv:1802.05957*.

[46] Z. Lu, E. Ie, and F. Sha, ''Mean-field approximation to Gaussian-softmax integral with application to uncertainty estimation,'' 2020, *arXiv:2006.07584*.

[47] J. Yoon, C. Kang, S. Kim, and J. Han, ''D-Vlog: Multimodal vlog dataset for depression detection,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, 2022, pp. 12226–12234.

[48] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, ''The distress analysis interview corpus of human and computer interviews,'' in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 3123–3128. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

[49] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, and E.-M. Messner, ''AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition,'' in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, 2019, pp. 3–12.

[50] H. Cai, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, and Q. Zhao, ''MODMA dataset: A multi-modal open dataset for mental-disorder analysis,'' 2020, *arXiv:2002.09283*.

[51] M. Weiss and P. Tonella, ''Fail-safe execution of deep learning based systems through uncertainty monitoring,'' in *Proc. 14th IEEE Conf. Softw. Test., Verification Validation (ICST)*, Apr. 2021, pp. 24–35.

[52] M. Du, S. Liu, T. Wang, W. Zhang, Y. Ke, L. Chen, and D. Ming, ''Depression recognition using a proposed speech chain model fusing speech production and perception features,'' *J. Affect. Disorders*, vol. 323, pp. 299–308, Feb. 2023.

[53] Y. Zhao, Z. Liang, J. Du, L. Zhang, C. Liu, and L. Zhao, ''Multi-head attention-based long short-term memory for depression detection from speech,'' *Frontiers Neurorobotics*, vol. 15, Aug. 2021, Art. no. 684037.

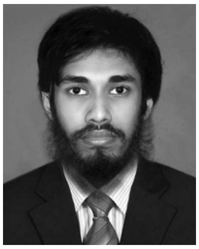
[54] M. Niu, K. Chen, Q. Chen, and L. Yang, ''HCAG: A hierarchical context-aware graph attention model for depression detection,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4235–4239.

[55] G. Lam, H. Dongyan, and W. Lin, ''Context-aware deep learning for multi-modal depression detection,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3946–3950.

[56] J. Zhao, X. Mao, and L. Chen, ''Speech emotion recognition using deep 1D & 2D CNN LSTM networks,'' *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.

[57] X. Deng, X. Fan, X. Lv, and K. Sun, ''SparNet: A convolutional neural network for EEG space-frequency feature learning and depression discrimination,'' *Frontiers Neuroinform.*, vol. 16, Jun. 2022, Art. no. 914823.

[58] B. Zhang, H. Cai, Y. Song, L. Tao, and Y. Li, ''Computer-aided recognition based on decision-level multimodal fusion for depression,'' *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 3466–3477, Jul. 2022.



SABBIR AHMED (Member, IEEE) received the B.Sc. degree in information technology from Jahangirnagar University. He is currently a Teaching Assistant and a Research Assistant with Jahangirnagar University. He is also the Chair of the IEEE Student Branch of Jahangirnagar University. He was with the DiversAsia Project and The University of Texas at Dallas as a Research Assistant. His research interests include algorithms, data structures, software engineering, graphic design, machine learning, security and blockchain, the Internet of Things, and robotics. He has published more than ten research papers in different international journals and conferences.

MOHAMMAD ABU YOUSUF received the B.Sc. (Engineering) degree in computer science and engineering from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 1999, the Master of Engineering degree in biomedical engineering from Kyung Hee University, South Korea, in 2009, and the Ph.D. degree in science and engineering from Saitama University, Japan, in 2013. In 2003, he joined as a Lecturer with the Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. In 2014, he moved to the Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh, where he is currently a Professor with the Institute of Information Technology. His research interests include bio medical image processing, human–robot interaction, computer vision, and natural language processing. He has published more than 100 research papers in different international journals and conferences.

MUHAMMAD MOSTAFA MONOWAR (Member, IEEE) received the B.Sc. degree in computer science and information technology from the Islamic University of Technology (IUT), Bangladesh, in 2003, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2011. He was a Faculty Member of the Department of Computer Science and Engineering, University of Chittagong, Bangladesh. He is currently a Professor with the Department of Information Technology, King Abdulaziz University, Saudi Arabia. His research interests include wireless networks, mostly ad-hoc, sensor, and mesh networks, including routing protocols, MAC mechanisms, IP and transport layer issues, cross-layer design, and QoS provisioning, security and privacy issues, and natural language processing. He has served as a program committee member for several international conferences/workshops. He served as an Editor for a couple of books published by CRC Press and Taylor & Francis Group. He also served as a guest editor for several journals.

MD. ABDUL HAMID was born in Sonatola, Pabna, Bangladesh. He received the Graduate degree from the Rajshahi Cadet College, Bangladesh, in 1995, the Bachelor of Engineering degree in computer and information engineering from International Islamic University Malaysia (IIUM), in 2001, and the Master-Ph.D. degree majoring in information communication from the Computer Engineering Department, Kyung Hee University, South Korea, in August 2009. He has been in the teaching profession throughout his life, which also spans over different parts of the globe. From 2002 to 2004, he was a Lecturer with the Computer Science and Engineering Department, Asian University of Bangladesh, Dhaka. From 2009 to 2012, he was an Assistant Professor with the Department of Information and Communications Engineering, Hankuk University of Foreign Studies (HUFS), South Korea. From 2012 to 2013, he was an Assistant Professor with the Department of Computer Science and Engineering, Green University of Bangladesh. From 2013 to 2016, he was an Assistant Professor with the Department of Computer Engineering, Taibah University, Madinah, Saudi Arabia. From 2016 to 2017, he was an Associate Professor with the Department of Computer Science, Faculty of Science and Information Technology, American International University-Bangladesh, Dhaka, Bangladesh. From 2017 to 2019, he was an Associate Professor and a Professor with the Department of Computer Science and Engineering, University of Asia Pacific, Dhaka. Since 2019, he has been a Professor with the Department of Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include networks/cyber-security, natural language processing, machine learning, wireless communications, and networking protocols. He has served as a program committee member contributing for the curriculum development and new program development in undergraduate and graduate disciplines.

MADINI O. ALASSAFI received the B.S. degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2006, the M.S. degree in computer science from California Lutheran University, USA, in 2013, and the Ph.D. degree in security cloud computing from the University of Southampton, Southampton, U.K., in 2018. He is currently an Associate Professor and the Vice Dean of the Faculty of Computing and Information Technology, King Abdulaziz University. He has published numerous conference papers, journal articles, and book chapters. His research interests include cloud computing and security, distributed systems, the Internet of Things (IoT) security issues, cloud security adoption and risks, cloud migration project management, the cloud of things, and security threats.