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