Diagnosis of Depression Based on Four-Stream Model of Bi-LSTM and CNN From Audio and Text Information

A-HYEON JO AND KEUN-CHANG KWAK, (Member, IEEE)
Department of Electronic Engineering, Chosun University, Gwangju 501759, South Korea

ABSTRACT Recent development trends in artificial intelligence applications have seen increasing interest in the design of automated systems for depression detection and diagnosis among the affective computing community. Particularly, active research has been conducted in depression diagnosis, based on multi-modal approaches in deep learning technology, which enable utilization of various information through fusion of varied data types. This study proposes a four-stream-based depression diagnosis model consisting of Bidirectional Long Short-Term Memory (Bi-LSTM) and convolutional neural networks (CNN), using speech and text data. One-dimensional features of audio signals are extracted using Mel Frequency Cepstral Coefficients and Gammatone Cepstral Coefficients, and two-dimensional features are extracted from Bark, equivalent rectangular bandwidth, and Log-Mel spectrograms, based on time-frequency transform. The extracted features are applied to Bi-LSTM and CNN-based transfer learning models. Word encoding was used for mapping of text to sequences with numeric indices, and word embedding used for representation of all words in numeric dense vectors. These were applied to Bi-LSTM and n-gram-based CNN models. Finally, an ensemble of the softmax values output from the four deep learning models was used to perform depression diagnosis, based on the proposed four-stream model. Using the proposed model, experiments were performed with the Extended Distress Analysis Interview Corpus Wizard of Oz depression database and other datasets. Experimental results showed improved performance by 10.7% to 11.9% over two-stream-based state-of-the-art methods. This demonstrates that the proposed model is effective for depression diagnosis.

INDEX TERMS Artificial intelligence, depression diagnosis, multi-modal, four-stream, bidirectional long short-term memory, convolutional neural networks.

I. INTRODUCTION

During the COVID-19 pandemic, more than half of the world population experienced isolation and the impact of the virus led to the declaration of a health emergency state, with implementation of unprecedented nationwide measures, such as social distancing strategies, in many countries across the globe. The pandemic disrupted normal patterns of daily life for large numbers of people, resulting in a remarkable increase in the number of people suffering from symptoms of anxiety and depression, and the global incidence of depression has more than doubled over time [1], [2]. Depression generally refers to a mental disorder characterized by persistent decrease in mental functions, including lack of interest or pleasure in activities, and disturbances in sleep and appetite, which may lead to adverse effects on normal day-to-day activities [3]. Although depression is a serious medical disorder, accurate and timely detection and diagnosis of depression remain a challenge, which adversely affects public health.
For diagnosis and treatment of depression, a patient either visits a hospital to talk about his/her symptoms, and a diagnosis is made based on the subjective judgment of a mental health professional after consultation and use of a questionnaire related to depression diagnosis. However, these methods have limitations because the process of diagnosis does not involve objective measures, making accurate diagnosis difficult. Additionally, according to the estimates of the World Health Organization (WHO), there is a world-wide shortage of adequate numbers of environmental and mental health professionals for the treatment of depression cases. With an increase in the number of patients with depression, the burden on healthcare professionals increase, and adds to the difficulties of accurate diagnosis of depression with objective and effective methods [4]. Therefore, to achieve early diagnosis of depression and timely intervention, there is an urgent need for objective and accurate methods of depression diagnoses. To this end, studies have been actively carried out to provide an objective solution for depression diagnosis with application of artificial intelligence (AI) technology.

With ongoing development of AI technology, in recent times, it has found application in many different fields. It has brought about significant changes in the field of medicine where it has been utilized as means to acquire detailed information. One of the useful applications of AI in mental health is in the accurate detection and diagnosis of a variety of mental disorders, including depression. Several previous studies exist with noteworthy achievements in the analysis of depression data and diagnosis of depression, based on different AI technologies and deep learning algorithms [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. These achievements contributed to early detection of depression in individual patients, which ultimately proved useful in the treatment process by reducing more serious development of and damage from the illness. Thus, there have been continuous reports and publications on the depression diagnosis system using AI. Nevertheless, to achieve more objective and accurate diagnosis of depression, further studies and improvements are needed.

Research on depression diagnosis with application of AI technology has generally been carried out with several different types of data, such as face image, electroencephalogram (EEG), human voice, behavior, and text. Linguistic data, such as speech and text, especially reveal the characteristics of patients with depression. This is because persons with symptoms of depressive disorder exhibit speech characteristics, such as reduced vocal intensity, reduced pitch range, and slower speech [5], [6]. Also, with the development of social media, there are large volumes of texts available on various social media, showing users’ own writings with expressions reflective of their feelings, indicating that it is now possible to define characteristics in text and speech that can be used for diagnosis of depression. While these speech and text data can be used to analyze the characteristics of depression, since depression is a complex mental disorder, analyzing a specific aspect using a single mode of data may not be sufficient for effective evaluation. Several studies have demonstrated that the fusion of various modalities of data significantly increases the accuracy of depression diagnosis [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Therefore, research is needed to enable accurate diagnosis of depression considering multiple types of data and their characteristics.

The motivation of this study is as follows. First, we explore efficient features and models to improve depression estimation/detection performance. Second, since the quality of the data can affect model performance, we improve the quality of the data through preprocessing, and we address the problem of data class imbalance for depression diagnosis models by augmentation of depression data. Third, speech and text data were analyzed for depression data to develop the diagnosis of depression, using a modal approach to the options of multiple modalities. This study aims to design a deep learning model containing various characteristics and information of patients with depression by combining data, and to effectively diagnose depression.

II. RELATED WORK

In recent years, there has been active research on development of objective AI instruments for automatic depression diagnosis by analyzing physiological indicators such as facial expressions, audio features, EEG, and text data. There has been increasing interest in multi-modal analysis since the analysis enables acquisition of comprehensive and robust features, leading to a shift of research from single-mode to multi-mode analysis for depression diagnosis. Since depression is a complex and multifaceted disorder, a multi-modal approach for depression diagnosis and consideration of multiple modalities of information are essential in developing automatic depression diagnosis. Some of the existing literature on multi-modal approaches developed for depression diagnosis are outlined as follows.

Ay et al. [7] proposed a deep hybrid model developed using Convolutional Neural Network (CNN) and Long-Short-Term Memory (LSTM) architectures to detect depression using EEG signals. In deep models, the temporal properties of signals are learned by the CNN layer and the sequence learning process is provided by the LSTM layer. In this work, they used EEG signals from the left and right hemispheres of the brain. They provided 99.12% and 97.66% classification accuracy for right and left hemisphere EEG signals, respectively. Therefore, it can be concluded that the developed CNN-LSTM model is accurate and fast in detecting depression using EEG signals.

Zhu et al. [8] proposed a method to distinguish between mild depression and normal controls using two feature fusion strategies (feature fusion and hidden layer fusion) for the fusion of EEG and EM signals based on the multimodal denoising autoencoder. They used the electroencephalography(EEG)-eye movement(EM) synchronous acquisition network to verify that the simultaneous recorded EEG and EM data during the experiment were
synchronized with millisecond precision, which is the basis for meaningful analysis of EEG and EM data. Also, extracted 14 features (12 nonlinear and 2 linear features) for each band, Delta (0–0.2 Hz), Theta (0.2–0.4 Hz), Alpha (0.4–0.6 Hz), Beta (0.6–0.8 Hz), and Gamma (0.8–1 Hz). Finally, in relation to the multimodal autoencoder, their studied two feature fusion methods (Feature Fusion and Hidden Layer Fusion) to achieve fusion of EEG and EM, and compared the classification performance differences of the two fusion methods, the effective and powerful fusion method was Hidden Layer Fusion method.

In the study of Zhang et al. [9], explored simultaneously from a physiological and behavioral perspectives and fused pervasive electroencephalography (EEG) and audio signals to make depression more objective, effective, and convenient to detect. After extracting several effective features for these two types of signals, we trained six representative classifiers for each form, then used the co-determination tensor to correlate with the diversity of decisions of different classifiers, and combine these decisions into ultimate classification results using multiple agents. Experiments on 170 (81 depression patients and 89 normal controls) demonstrated that the proposed multi-mode depression detection system outperforms single-mode classifiers or other common late fusion strategies in accuracy, f1-scores, and sensitivity.

Shen et al. [10] proposed a multimodal-based deep learning model for automatic depression diagnosis using EATD-Corpus, a publicly available Chinese depression dataset consisting of audio and text records extracted from 162 volunteers’ interviews, and DAIC-WoZ dataset, an interview-type clinical depression data. The proposed model fuses audio and text features using a Gate Recurrent Unit (GRU) model and a Bi-LSTM model with an attention layer. Text features are extracted by projecting sentences into high-dimensional sentence embeddings using ELMo. For audio features, Mel spectrogram is extracted from audio. Experimental results show that the proposed method is very effective.

Guo et al. [11] proposed a detection framework for detecting student mental health named educational data fusion for mental health detection (CASTLE). This framework is largely divided into three parts. First, using Presentation Learning, we fuse data such as social life, academic performance, and appearance. The name Multi-View Social Network Embedding (MOON), an algorithm, is proposed to effectively fuse students’ disparate social relationships to express their social lives comprehensively. Second, the Synthetic Minority Oversampling Technique (SMOTE) algorithm is applied to the label imbalance problem. Finally, we use the Deep Neural Network (DNN) model for final detection. Extensive results show the promising performance of the proposed method compared to a wide range of state-of-the-art baselines.

Park et al. [12] proposed a multi-modal data-based attention-mechanism depression diagnosis model to improve the low accuracy problem of depression detection model based on single mode data. The proposed model was composed of a bidirectional encoder representations from transformers-convolutional neural network (BERT-CNN) fusion for natural language analysis, a CNN-Bidirectional Long Short-Term Memory (CNN-BiLSTM) for voice signal processing, multi-modal analysis, and a fusion model for depression diagnosis. The model used audio and text data of the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) database, converted speech data into log-mel spectrograms, extracted features through the CNN model, and the extracted features were learned through the Bi-LSTM, with application of an attention mechanism. The text data was converted into embedding vectors using a BERT tokenizer, and the pre-trained BERT-CNN model was fine-tuned and trained to extract feature vectors. The proposed model resolved the problem of rapid loss increase due to the use of single mode data, and showed improved accuracy.

Xiao et al. [13] proposed a novel approach for automatic depression detection based on audio-text sequences. They proposed, as a new model for depression analysis, an attention mechanism to casual-CNN (Attention-C-CNN) for audio feature extraction. For text feature extraction, BERT, a pre-trained model proposed by Google in 2018, was used. Furthermore, a new co-attention encoder that allows improved fusion of audio and text features was applied by exchanging key-value pairs of the multi-head attention mechanism in the co-attention transformation layer. As a result of experiments using the DAIC-WOZ dataset, the proposed model showed more competitive performance than single-mode data and state-of-the-art multi-modal data-based methods.

Niu et al. [14] proposed a hierarchical context-aware graph (HCAG)-based attention model for depression diagnosis in which a graph attention network (GAT) [15] was applied to text/audio modality data for effective identification and integration of context information between relational interview questions. Specifically, the hierarchical context-aware structure enables the grasping of important information within the answers to clinical interviews, and the GAT network allows collection of sufficient relational information between interview questions and logical information. The experimental results showed that the proposed HCAG was more robust and outperforms existing state-of-the-art models in all of five evaluation metrics.

In the study of L. Yang et al. [16], a multi-model fusion framework consisting of deep convolutional neural network (DCNN) and deep neural network (DNN) models was proposed. Additionally, new feature descriptors for text and video stream data were proposed. The proposed model considers audio, video and text data. For each data, hand-crafted feature descriptors are input to the DCNN to learn high-level global features with compact dynamic information. The learned features are then fed to the DNN to obtain the predicted scores. For multimodal fusion, the PHQ-8 scores estimated from the three modalities are integrated into the DNN to obtain the final PHQ-8 score. For text descriptors,
answers associated with words such as sleep disorder were selected, and Paragraph Vector (PV) unsupervised learning algorithm was used to learn distributed vector representations of variable-length text pieces, such as sentences. For video descriptors, a new global descriptor, Histogram of Displacement Range (HDR), calculated directly from facial landmarks to measure their displacement and speed, was proposed. Experimental results on the AVEC 2017 depression dataset showed that the proposed multi-modal hybrid structure with fusion of DCNN and DNN models obtained promising accuracy.

Alhanai et al. [17] proposed a Long Short-Term Memory (LSTM) model, which represents multi-modal interactions with audio and text features. The combination of audio and text modalities contained complementary information as well as temporally varying and discriminative information about the state of a depressed person. The proposed model was composed of two LSTM branches. For audio data, the collaborative voice analysis repository for speech technologies (COVAREP) features were used, and for learning the features, the model had LSTM layers. For text data, features were extracted using Doc2Vec, and the model was composed of two LSTM layers for learning the features. The outputs were merged into a final feedforward network. The branches were composed of different topologies and optimized for the different characteristics and information content of the respective modalities.

Lin et al. [18] proposed a novel automatic depression diagnosis method by combining audio and text information from patient interviews. Specifically, the proposed method has a network architecture integrating the outputs of a Bi-LSTM network with an attention layer for processing text content and a 1D-CNN model for processing speech signals, and these are fed into two fully connected (FC) networks to perform classification. This method diagnoses the presence of depression and assesses the severity of depression symptoms. As a result of evaluating the proposed model using two publicly available datasets, the DAIC-WOZ and Audio-Visual Depressive Language Corpus (AVisD-Corpus) datasets, the proposed method achieved high performance in the task of depression detection.

Lam et al. [19] proposed a novel method integrating context-aware analysis and a data-driven approach for depression diagnosis. The method incorporated a data augmentation procedure based on topic modeling, and showed its effectiveness for training multi-modal (audio and text) deep learning models. A deep 1D CNN and transformer model achieved good performance as a result of training on audio and text data, respectively, whereas the proposed multi-modal results showed improved performance. In this study, data augmentation based on topic modeling was performed on the training dataset and a two-stream-based depression diagnosis model was proposed by fusion of audio and text data.

These previous studies have used either one-dimensional features of audio data or two-dimensional spectrogram features based on time-frequency transformation, and there has been no research with simultaneous considerations of features such as Mel Frequency Cepstral Coefficient (MFCC) and Gammatone Cepstral Coefficient (GTCC) that allow derivation of harmonic structures from the spectrum of speech signals, and time-varying frequency features of speech with nonlinear characteristics. Moreover, text data is mainly applied to the LSTM model, which is characterized by sequential processing of input values, thus preserving the local information of sentences. This limits the reflection of the information on the order of words or expressions, and thereby limits accurate diagnosis of depression.

Therefore, in this study, use of depression data consisting of audio and text, the late score fusion method is applied to the Bi-LSTM network and CNN models, which are deep learning models, to design a four-stream-based depression diagnosis model. To address the limitation of single-mode data, different types of data, speech and text data, that is, multimodal data, are used for training of the proposed depression diagnosis model. Among the two modes of data, speech data is denoised through preprocessing and data augmentation is applied to depression data to resolve the problem of class imbalance. One-dimensional features of audio signals are then extracted through MFCC and GTCC and training of the Bi-LSTM model is performed. In addition, time-frequency transformation-based two-dimensional features such as Bark, equivalent rectangular bandwidth (ERB), and Log-Mel spectrograms are extracted, and the extracted features are learned through a CNN-based transfer learning model to perform comparative analysis on model performance. For text data, a word embedding layer, in which text is converted to numeric sequences and mapped to words, is included in the neural network and the performance of the depression diagnosis model based on Bi-LSTM and CNN is compared. The multi-modal data obtained by data fusion of the probability values output from the softmax layers of the four models proposed above, through the late score fusion method, was used to design a four-stream deep learning model-based depression diagnosis model. Through noise removal and augmentation of depression data, the problems of data quality and class imbalance were addressed to improve the performance of a single model, and this had a positive effect on the final result of the model performance. Finally, by applying the late score fusion method for the four models used for audio and text data learning, it was shown that the performance of the proposed four-stream-based deep learning model improved compared to the model using single mode data and two-stream-based state-of-the-art methods.

The deep learning algorithm proposed through this study is expected to be able to predict or diagnose abnormalities such as mental health-related diseases, anxiety, and suicide signs as well as depression. Therefore, the diagnosis of depression through audio and text information analysis based on deep learning 4-stream can early detect not only depression but also mental health-related diseases and anomalies. Also, it can be recommended appropriate treatment, and it is expected to be prevented through active management.
The remainder of this paper is organized as follows. Chapter 3 describes the methodologies of deep learning model design using audio signals. The basic concept of the depression diagnosis model using depression data composed of interview-type speech and the design of a deep learning model are presented. This includes a description of the feature extraction method from one-dimensional data and the models used in this study. Also, a method of two-dimensional time-frequency transform of one-dimensional signals and a CNN-based transfer learning model used for depression diagnosis are described in this chapter. Chapter 4 presents a depression diagnosis model using text data. Chapter 5 proposes and describes a four-stream-based depression diagnosis model using audio and text data. Chapter 6 presents the comparison of experimental results using publicly available depression datasets to verify the performance of the proposed four-stream-based depression diagnosis model. The final Chapter 7 and Chapter 8, respectively, present the discussion of this study and the final conclusion of this study and future research plans.

III. DEEP LEARNING MODEL DESIGN WITH AUDIO SIGNALS

A. FEATURE EXTRACTION METHODS FOR 1D AUDIO SIGNALS

For development of an interface using one-dimensional audio signals, the most important technique is effective extraction of useful features from the speech data. The most used feature extraction methods include Linear Predictive Coding (LPC) cepstrum, Perceptual Linear Predictive (PLP) cepstrum, MFCC, GTCC, and filter bank energy. The MFCC extraction method extracts features termed MFCCs from audio signals. An MFCC refers to a number representing the unique features of sound. The MFCC is one of the most effective feature extraction methods generally used in speech recognition because it reflects the characteristics of the frequency range of human hearing [20]. The calculation process of MFCC is illustrated in Fig. 1.

GTCC, another Fast Fourier Transform (FFT)-based speech feature extraction method, is based on a Gammatone filter bank of Equivalent Rectangular Bandwidth (ERB) that measures the width of auditory filters at each point. The impulse responses of the Gammatone filter bank are very similar to the size characteristics of human auditory filters [21]. The calculation process of GTCC is shown in Fig. 2.

The calculation process of GTCC is similar to the MFCC feature extraction method. First, the audio signals are divided into short frames of 10-50 ms. Then, a Fast Fourier Transform (FFT) operation is performed on the audio signals to convert them from the time domain to the frequency domain, which passes through the Gammatone filter bank to visually emphasize the significant frequencies of the audio signals. Finally, a log function and a discrete cosine transform are applied to decorrelate the log-compressed filter output and generate better energy compression to obtain GTCC features. Equation 1 presents the formula for calculating GTCC.

$$GTCC_s = \sqrt{\frac{2}{M} \sum_{m=1}^{M} \log(F_m) \times \cos\left[\frac{\pi m}{M} \left(s - \frac{1}{2}\right)\right]} 1 \leq s \leq K$$  (1)

where $F_m$ denotes energy of the signal in the m-th spectral band, $M$ is the number of Gammatone filters, and $K$ indicates the number of GTCC. In this study, the Hamming window was used to extract the MFCC and GTCC features of one-dimensional audio signals, and the length of the window was set to 480.

B. 2D TIME-FREQUENCY TRANSFORM-BASED FEATURE EXTRACTION METHOD

To train a two-dimensional CNN model for speech signals, the features of Bark spectrograms, ERB spectrograms, and log-mel spectrograms were extracted based on a two-dimensional time-frequency transform. First, the Bark spectrogram is based on the Bark Frequency Scale, a psychoacoustical scale proposed by Eberhard Zwicker, a German acoustics scientist in 1961. It is named after Heinrich Barkhausen, who first proposed a subjective measurement of
Humans can distinguish characteristics such as intensity, pitch, duration, and timbre of sound by using the auditory organ; therefore, to express the characteristics of sound specifically, scales to distinguish different aspects of sound are needed. Consequently, based on the results of numerous psychoacoustic experiments, the Bark scale was defined so that the human auditory critical band has the width of one Bark scale each. There are values between 1 and 24 on the scale, and these 24 values are related to the 24 critical bands. The scale is effective in representing important features in a specific band in speech data.

Based on the Bark frequency scale, the Bark filter bank can be designed by specifying the frequency range to design the auditory filter bank, the FFT length, which is the number of points used for computing discrete Fourier transform (DFT), and the number of bandpass filters. The Bark spectrogram can be obtained by applying Short Time Fourier Transform (STFT) to audio signals to convert to a spectrogram and then multiplying it with a previously designed filter bank. Fig. 3 is a visualization of the STFT-based spectrograms for depression and non-depression data, respectively, and Fig. 4 is a visualization of the Bark scale-based spectrogram for depression and non-depression data, respectively. As can be seen from the figures, the features of the spectrogram using the Bark filter bank are more prominent than those of the spectrogram without the filter bank and only with STFT.

The last spectrogram used in this paper is a log-mel spectrogram. In human recognition of speech signals, the signals are not recognized linearly according to frequency but with

$$ERB(f) = 24.7 + 0.108f$$

Based on the ERB concept as described above, the ERB auditory filter bank is designed, and the product operation between this filter bank and the spectrogram is performed. Then, the ERB spectrogram features used in this study are extracted. Fig. 5 shows the ERB spectrogram features for depression and non-depression audio data.
a Mel scale, which is a method of converting real frequency information into a mathematical form based on the human auditory system. Since frequency information can be used differently according to need, it is widely applied in the field of speech signal processing.

A log-mel spectrogram is obtained based on spectrogram features. A spectrogram is a visual representation of the spectrum of sound frequencies obtained by dividing audio signals into frames, converting them from the time domain to the frequency domain, and then stacking them horizontally. By converting audio signals into spectrograms, even highly complex audio signals can be analyzed at each frequency. Therefore, the log Mel spectrogram is obtained by converting audio signals to the Mel scale, applying STFT to extract Mel spectrogram features, and taking the logarithmic transform. The spectrogram features obtained by taking the log transform are useful in terms of representing features of audio signals, thus a log-mel spectrogram is widely used in the field of audio processing and shows high performance.

As with the Bark and ERB spectrograms described above, the log Mel spectrogram features can be extracted by designing a Mel filter bank and applying STFT to obtain a spectrogram, multiplying it with the Mel filter bank, and taking the logarithmic transform. In Fig. 6, log-mel spectrogram features for depression and non-depression data respectively are illustrated.

From the network architecture of Bi-LSTM shown in Fig. 7, one can see an orange block forward LSTM and a yellow block backward LSTM in parallel arrangement. In Bi-LSTM, an additional LSTM layer is added with backward flow of information. That is, the input sequence flows backwards in the additional LSTM layer. Then the outputs of the two LSTM layers are merged in several ways, such as averaging, summing, multiplying, or concatenating. Bi-LSTM is generally used when specifying order is required in the process. This type of network can be used for text classification, speech recognition and prediction models.

We propose a model for depression diagnosis by extracting the features of one-dimensional audio signals, performing learning based on a Bi-LSTM model, and classifying the state of depression. The architecture of the proposed Bi-LSTM network is presented in Fig. 8. First, one-dimensional features are obtained from audio signals using MFCC and GTCC feature extraction methods and converted into sequence data. Then, a network is built by stacking two Bi-LSTM layers, and training with data is performed by adding a fully connected layer, softmax layer, and classification layer. At this time, a dropout layer with a probability of 30% is added between the Bi-LSTM layers to prevent overfitting. Finally, using the probability output from the softmax layer, the state of depression is determined as depression or non-depression.

D. 2D TIME-FREQUENCY TRANSFORM-BASED CNN TRANSFER LEARNING MODEL

In this study, image features extracted by converting audio signals into two-dimensional time-frequency representations, such as Bark, ERB, and Log-Mel spectrograms, are learned using CNN-based transfer learning models, VGGish, YAMNet, and OpenL3 models. In general, a large amount of data is needed to properly train a CNN-based deep learning model. However, it is not easy to build a large dataset because of time and cost constraints. One of the solutions to insufficient data is transfer learning, which refers to a method of using a deep learning model pre-trained on a large dataset, usually for large-scale image classification tasks. Transfer learning method uses the pre-trained model as it is, or initializes the
weights of the model for recalibration according to the task at hand. The transfer learning model is effective as it enables the performance of deep learning with a relatively small amount of data, and provides high accuracy and fast learning rate.

The VGGish model, one of the CNN-based transfer learning models, is a deep learning neural network proposed by S. Hershey [25] for classification of audio classes by using audio signals in a large-scale video database in YouTube and training a neural network. A neural network training was performed using audio content consisting of over 2 million YouTube videos with 527 audio classes. The 527 audio classes include voices of adult men and women, baby utterances, and animal sounds [26]. The VGGish model has a network architecture based on Visual Geometry Group (VGG), which has been widely used in computer vision for image classification. Fig. 9 shows a simplified VGGish network architecture. In the VGGish model, spectrogram-based features of size 96 by 64 by 1 of an audio clip are input, and the network consists of 4 convolution blocks. Each block includes a 2D-based convolution layer that functions as a feature extractor, a rectified linear unit (ReLU), an activation function, and a Max Pooling layer that retains image features and reduces dimensions. Next, two fully connected layers functioning as classifiers, an embedding layer, and a regression output layer are included in the back-end modules of the network.

Audio signals are converted into time-frequency two-dimensional features and training is performed using a CNN-based transfer learning model to perform diagnosis of depression. To determine the state of depression, a model with binary classification (depression/non-depression) is needed. Therefore, a new fully connected layer is added following the original fully connected layer and the number of classes for classification increases to two. Additionally, a classification model for depression data was constructed by replacing the regression output layer with a classification output layer.

The YAMNet model is an acoustic detection model trained on audio dataset with 521 classes, such as laughter, dog barking, and siren sounds, from an AudioSet of over 2 million YouTube videos. The YAMNet model, developed by Ellis and Chowdhry, is a computationally efficient model for classifying audio events for the AudioSet corpus. The drawback of the VGGish model is that it is computationally complex, using more than 72 million parameters, whereas the YAMNet model uses only 4.7 million parameters; so it is more computationally efficient. Using convolution kernels that are depth-wise separable, a lightweight model was developed for utilization in the field of computer vision, and the YAMNet model was finally developed based on the MobileNet architecture proposed by H. A. Andrew et al. [27]. The YAMNet model receives spectrogram images with a size of 96*64*1 as inputs and consists of 14 convolution layer blocks. All layers, except the first layer, are based on convolution kernels by depth. A fully connected layer was added at the back end of the last convolution layer, the number of classes was re-set, and the transfer learning model was fine-tuned by replacing the classification layer.

Apart from the VGGish and YAMNet models trained to recognize human voice among various types of transfer learning models, the OpenL3 model, optimized for identification of music and environmental sounds, is used for training on speech depression data. The reason for using this model is to
investigate whether a model trained to recognize music and environmental sounds can be used to recognize the features of a parallel speech corpus in depression data. In J. Cramer et al. [28], the OpenL3 model was developed based on the L3 (Look, Listen, Learn) concept proposed by R. Arandjelovic and A. Zisserman [29]. In addition, various network architecture types were investigated for STFT-based spectrograms or Mel-scale frequency-based spectrograms, and different sizes and types of network for deep acoustic embedding. It is important to note that although the videos used to train the OpenL3 model are selected within the AudioSet corpus, the model has far fewer classification classes than the VGGish and YAMNet models. The network architecture of OpenL3 model is composed of 4 convolution layer blocks used as feature extractors by receiving spectrogram inputs with a size of 128*199*1. The operation of the MaxPooling layer is performed on the feature extractor output using the option to generate embeddings with sizes from 512 to 6,144.

IV. DEEP LEARNING MODEL DESIGN WITH TEXT DATA

A. BI-LSTM-BASED DEPRESSION DIAGNOSIS MODEL

The first deep learning model for text classification is a depression diagnosis model based on Bi-LSTM. The Bi-LSTM network is a recurrent neural network (RNN) that can learn long-term dependencies between time steps in sequence data, and since text data is essentially sequential data, it can be processed using Bi-LSTM. To use text data as an input of Bi-LSTM, text data must first be converted into a numeric sequence. Therefore, text data was mapped to a sequence with numeric index using word encoding. The performance of the neural network varies depending on the method of expressing the word. In this study, to obtain better performance, a word embedding layer was added to the neural network, for mapping all words in the collection of words to numeric vectors rather than scalar indices. In this way, semantic information of words is collected so that words with similar meanings have similar vectors, and the relationship between words can be modeled through vector operations.

The architecture of the proposed Bi-LSTM-based depression diagnosis model using text data is shown in Fig. 10. First, the text data is imported and the text obtained through preprocessing is converted into a numeric sequence through word encoding. After the sequence input layer, a word embedding layer, two Bi-LSTM layers with 10 hidden units are stacked, and a dropout layer with a probability of 30% is placed to design the Bi-LSTM neural network, followed by a fully connected layer and a softmax layer for depression of diagnosis using the probability output from the softmax layer.

B. CNN-BASED DEPRESSION DIAGNOSIS MODEL

The second deep learning model for classifying text data is a model based on CNN. To classify text data using CNN, it is necessary to use a one-dimensional convolutional layer that performs convolution operation on the time dimension of the input data. In the proposed model, the network is trained using one-dimensional convolution filters with various widths. Each filter width of the convolutional layer corresponds to the number of words (n-gram length) that the filter can see. An n-gram consists of n consecutive words, and is a method of expressing a word as a vector using numbers. It analyzes word sequences of one or more words and determines how many words the word sequence is composed of according to the number “n” at the beginning of the n-gram. Cases of n=1, 2, and 3 are called Uni-gram, Bi-gram, and Tri-gram, respectively. Fig. 11 shows how the word sequences are constructed. The n-gram has the advantage of predicting the next word, detecting typos, and addressing the limitations of the bag-of-words method that does not take into account the order of words.

Using the n-gram method as described above, a network was constructed, and the network of the proposed CNN-based text classification model is illustrated in Fig. 12. First, the text is imported from the transcript, the text word obtained through preprocessing is converted into a numeric sequence through word encoding. After the sequence input layer, a word embedding layer, two Bi-LSTM layers with 10 hidden units are stacked, and a dropout layer with a probability of 30% is placed to design the Bi-LSTM neural network, followed by a fully connected layer and a softmax layer for depression of diagnosis using the probability output from the softmax layer.
and the sequence is input to the sequence input layer and the word embedding layer. Then, the convolutional network is divided into two and uses the length of n-grams with values of 2 and 3, respectively. In each convolutional network, a convolution layer, a batch normalization layer, a ReLu layer as an activation function, a dropout layer with a probability of 20%, and a global max pooling layer are arranged in that order. Finally, using the fully connected layer and the softmax layer, depression state classification is performed.

V. DESIGN OF FOUR-STREAM ENSEMBLE-BASED DEEP LEARNING MODEL AND DEPRESSION DIAGNOSIS WITH AUDIO AND TEXT DATA

A. LATE SCORE FUSION METHOD

In machine learning/deep learning, late score fusion is the most common and simple data fusion method. This approach uses deep learning models that can be trained independently to obtain final values by fusion of common values output from different modalities [30]. In this method, different types of data are used for respective training of deep learning models, and the score of the softmax layer, the final layer, is fused into one value of classification. The scores of the softmax layer are added or multiplied and the maximum or average score of the values is used to obtain the final output value, which is the value used for classification into classes. The late score fusion method is simply illustrated in Fig. 13.

B. FOUR-STREAM-BASED DEEP LEARNING MODEL FOR DEPRESSION DIAGNOSIS FROM AUDIO AND TEXT DATA

In this study, a four-stream-based deep learning model is designed in which, among multi-modal data, audio and text data are learned through Bi-LSTM and CNN models, and late fusion is performed on their softmax scores to diagnose depression. One-dimensional audio data is trained through Bi-LSTM model by extracting features such as MFCC and GTCC so that the harmonic structure of the spectrum can be derived. Also, to consider the changing frequency characteristics over time, two-dimensional features such as Bark, ERB, and Log-Mel spectrograms are extracted and trained through CNN-based transfer learning models such as VGGish, YAMNet, and OpenL3. Then each softmax score is obtained from the respective methods. For text data, the text is converted into a numeric sequence through the Bi-LSTM network, and a word embedding layer is added for the mapping of words to vectors. In the CNN model, one dimension-based CNN network is created with the sequence vectors as the input, and the respective softmax scores are obtained. Finally, for depression diagnosis, the softmax values obtained by training four deep learning models on audio and text data are summed or multiplied using the late score fusion method to take the maximum value. Then, the final output value and classification performance are obtained, and the state of depression is diagnosed by binary classification into depression or non-depression. Fig. 14 illustrates the structure of the proposed four-stream-based depression diagnosis model with Bi-LSTM and CNN and, the pseudocode, which clearly shows the flow of the proposed system, can be found in Table 1.

VI. EXPERIMENTS AND RESULTS

A. EDAIC-WOZ DEPRESSION DATABASE

The Extended Distress Analysis Interview Corpus-Wizard of Oz (EDAIC-WOZ) depression dataset is an extended version of the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) dataset. This data is a set of depression data used in the 2019 Audio/Visual Emotional Challenge (AVEC 2019) [31]. The DAIC-WOZ dataset is part of a larger dataset, the Distress Analysis Interview Corpus...
FIGURE 14. Structure of the proposed, four-stream-based depression diagnosis model of Bi-LSTM and CNN using speech and text data.

(DAIC), which was designed to help diagnose conditions of psychological distress in people such as anxiety, depression, and Post-Traumatic Stress Disorder (PTSD). This dataset was collected as a part of a larger effort to develop a computer agent that identifies verbal and nonverbal indicators of mental illness through clinical interviews with participants [31], [32]. The clinical interview contains audio, video and text (script) files of the participants.

In the DAIC-WOZ dataset, clinical interviews are conducted with participants and a virtual interviewer named ‘Ellie’, which is an animated character, and a controller of the virtual interviewer. Both of those with and without depression symptoms were included as participants. The EDAIC-WOZ dataset not only includes an interview with ‘Ellie’, a virtual interviewer, as in DAIC-WOZ, but the data is also collected using an AI-controlled agent. The agent operates fully autonomously using a variety of automated recognition and action generation modules. The data of 189 people in sessions with IDs in the range [P300-P492] were collected through interviews with virtual interviewer ‘Ellie’, and the data of 86 people in sessions with IDs in the range [P600-P718] were collected through an AI-controlled agent [33].

Figs. 15 and 16 simply show how to collect EDAIC-WOZ dataset using the virtual interviewer ‘Ellie’ and AI-controlled agent. Prior to data collection, PHQ-8, one of the depression surveys, was used to conduct a preliminary survey with the participants to screen their depression status. The PHQ-8 questionnaire presents a standardized scale of depression with which survey participants rate their scores based on the number of days they experienced eight out of nine symptoms of Diagnostic and Statistical Manual of Mental Disorders (DSM) over the last two weeks [34]. The score on each item is summed to calculate the total score in the range of 0-24 points. A total score of 0-4 indicates none or minimal symptoms of depression, 5-9 mild, 10-14 moderate, 15-19 moderately severe, and 20-24 indicates severe symptoms of depression. In this study, to determine the state of depression by means of binary classification, when the score of PHQ-8 < 10, the participant was diagnosed as not depressed (labeled as non-depression), and when the score ≥ 10, the participant was diagnosed as depressed (labeled as depression).

As shown in Fig. 15, 189 participants with identifications (IDs) in the range [P300-P492] will be interviewed by the virtual interviewer ‘Ellie’ after completing the PHQ-8 survey. ‘Ellie’ is controlled by a third party, and during the interview, the pose and face data of the participants are collected through a camera and speech and text data are collected through a microphone. As shown in Fig. 16, pose, face, speech, and text data of 86 participants with IDs in the range [P600-P718]...
were collected through an interview with an AI-controlled agent.

To build the dataset, a total of 275 people participated, and the interview was conducted for an average duration of 16 minutes, ranging from 7 to 33 minutes per participant. In this study, audio files of participants and text (script) files created based on the interviews were used, and the audio files were recorded at 16 kHz and in the .wav file format. For text (script) data, the transcripts of conversations during the interviews are saved as Microsoft (MS) Excel files. The typical forms of data are presented in Fig. 17 and Fig. 18.

The dataset is divided into training, validation, and test data in consideration of age, gender, and PHQ-8 scores. Table 2 shows the distribution of the number of EDAIC-WOZ data by gender. Among male participants (n=170), 35 had depression symptoms and 135 had no symptoms. Among female participants (n=105), 31 people had depression symptoms and 74 people had no symptoms. Male participants accounted for a higher proportion than female participants. In summary, the dataset consists of data from a total of 275 participants (66 participants with symptoms of depression and 209 participants without symptoms). In Table 3, we can see the size of participants’ data used in training, validation, and test, respectively. The training data consisted of data from 37 participants with depression and 126 participants with non-depression, making up a total of 163 patients; the validation data consisted of data from 12 participants with depression and 44 participants with non-depression, making up a total 56 participants; finally, the test data consisted of data from 17 participants with depression and 39 participants with non-depression, making up a total of 56 participants. The training and validation data are a mixture of an interview with ‘Ellie’ and an AI-controlled agent, but the test data consists only of an interview with an AI-controlled agent. In the study, we did not use a dataset that was divided originally, but the data of 220 participants (depression: 53, non-depression: 167), which is 80% of the total data, were used as the training data, and the data of 55 participants (depression: 13, non-depression: 42), which is the remaining 20% of the total data, were used as validation data.

B. PREPROCESSING OF SPEECH AND TEXT DATA

In the speech data, various types of noise such as background noise or silence may be included. The quality of the data can affect the model performance, and thus it is necessary
In the first method, preprocessing was performed for various cases of performance comparisons by dividing it into one of the following three outlined methods. Participants was used as data. Next, a deep learning model for the virtual interviewer was removed, and only the text of the interview are saved as MS Excel files, all of the text of the conversation between the virtual interviewer and the participant during the EDAIC-WOZ dataset. First, in the transcript file, all the conversations that are less noise are also removed. Finally, sentences that are less than two words or more than fifteen words are removed.

### C. AUDIO AND TEXT DATA AUGMENTATION

In general, there is much less amount of available data from people with illnesses than data from people without illnesses. In the EDAIC-WOZ data, there are 66 depression data and 209 non-depression data; the difference in the amount of data between the two classes is about three times, resulting in a class imbalance problem. Data imbalance can cause a problem of overfitting, and the model tends to assign more weights to the class with more data and the model prediction also includes more classes from higher weights. This may increase the prediction accuracy, but also may cause a problem in that precision and recall for a class with a small amount of data are decreased. Therefore, to tackle this problem, the amount of data in each class was adjusted through a method of data augmentation. Data augmentation increased each file of depression data three-fold by using methods such as pitch shifting to shift the pitch of the speech left/right, time shifting to shift the time of the speech to the left/right, and adding noise to add white noise to audio signals. Data augmentation solved the class imbalance problem by increasing the size of the depression data class to 199, whereas it was 209 for the non-depression data class.

As in the case of audio data, the text scripts also had a class imbalance problem, with 66 depression data and 209 non-depression data. To resolve this problem, data augmentation was carried out for the text data of participants with depression using the Easy Data Augmentation (EDA) concept. EDA is a text data augmentation technique proposed by J. Wei [36], which can help improve the performance of text classification. The EDA method consists of operations such as Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD); SR is to randomly select n words that are not stop words from a sentence, and replace these words with randomly selected synonyms; RI is to find a random synonym of a random word in a sentence that is not a stop word for n times, and insert the synonym at the end of the sentence. In the order of tokenization, lemmatization, removing punctuation marks, removing unnecessary words, and removing overly short or long sentences. Text data is loaded, tokenization is performed, and lemmatization is carried out. Then, punctuation marks are removed and stop words that can be considered noise are also removed from the text data. Finally, sentences that are less than two words or more than fifteen words are removed.

### TABLE 3. EDAIC-WOZ data classification.

| Class          | Train | Validation | Test |
|----------------|-------|------------|------|
| Depression     | 37    | 12         | 17   |
| (M:18, W:19)   | (M:8, W:4) | (M:9, W:8) |     |
| Non-depression | 126   | 44         | 39   |
| (M:75, W:51)   | (M:26, W:18) | (M:34, W:5) |     |
| Total          | 163   | 56         | 56   |
a random position in the sentence; RS is to randomly select two words and swap their positions for n times in a sentence; and RD is random removal for each word in the sentence with probability p. By changing the data values of using these methods, depression data was increased three-fold for each file. In this way, the class imbalance problem was resolved with the result of 199 depression data and 209 non-depression data.

TABLE 4. Experimental environment.

| Division          | Use                                      |
|-------------------|------------------------------------------|
| **Hardware**      |                                          |
| CPU               | Intel Core i9 10900K @ 3.70GHz           |
| GPU               | NVIDIA GeForce RTX 2080 SUPER           |
| RAM               | 128GB                                    |
| **Software**      |                                          |
| OS                | Windows 10                               |
| Programming Language | Python, Matlab2022a                      |

D. PERFORMANCE EVALUATION

In this section, we carry out comparative analysis of the depression diagnosis performance of the proposed four-stream-based deep learning model using Bi-LSTM and CNN from audio and text signals. Information on the experimental environment and the software used can be found in Table 4, and the experimental process of this section is as follows:

[Step 1] The experiment with audio data is divided into the cases of data before and after preprocessing (noise removal from the speech data and data augmentation) for comparative analysis of the performance of the depression diagnosis model using the Bi-LSTM and CNN-based transfer learning models.

[Step 2] The experiment with text data is divided into the cases of data before and after data augmentation for comparative analysis of the depression diagnosis performance of the Bi-LSTM and CNN model.

[Step 3] From the fusion of audio and text data, the performance of the deep learning four-stream-based depression diagnosis model is evaluated and compared with the performance of an existing two-stream model.

The first experiment shows the experimental results of Bi-LSTM-based depression diagnosis using audio data from the EDAIC-WOZ dataset. The experiment was conducted by dividing audio data into before and after speech data noise removal and data augmentation, the feature extraction methods of one-dimensional audio signals were divided into three cases: MFCC, GTCC, and MFCC+GTCC, and the number of hidden units in the Bi-LSTM layer was respectively changed to 10, 50, and 100 to carry out the experiment. Table 5 shows the training parameter values of the Bi-LSTM model. The training was conducted by fixing parameters with values as shown in Table 5.

TABLE 5. Training parameter of Bi-LSTM model(audio).

| Parameter | Optimization function | Gradient Decay Factor | Mini Batch Size | Epoch |
|-----------|-----------------------|-----------------------|-----------------|-------|
| Values    | Adam                  | 2                     | 512             | 10    |

Fig. 19 presents a graph of the classification performance of the Bi-LSTM model before speech data noise removal and data augmentation. As shown in Fig. 19, when audio data is not subjected to any preprocessing, the performance is generally similar between different cases. When the features of MFCC and GTCC are extracted together and the number of hidden units is 100, the accuracy of Bi-LSTM is 80%, showing the highest performance.

Fig. 20 shows a graphic presentation of the classification performance of the Bi-LSTM model after data denoising and augmentation. As shown in Fig. 20, the overall performance was improved compared to the performance before data preprocessing. When the number of hidden units in the Bi-LSTM layer is 10 and MFCC and GTCC features are extracted together, the accuracy is 96.34%, showing the highest performance. This is the highest performance in Bi-LSTM-based depression diagnosis using audio data. When compared to the performance before data preprocessing under the same conditions, the performance was improved by about 18.16%.
blue (RGB) images. The transfer learning model used in the experiment receives B&W images as input. Therefore, when using RGB images as inputs, the input end of the model and the first convolution layer were tuned according to the size of the input. The input size of each transfer learning model is presented in Table 6. Table 7 shows the training parameter values of the transfer learning model. The training was conducted by fixing the parameters with the values shown in Table 7.

**TABLE 6.** Size of input image for each CNN-based transfer learning model.

| Input image          | Pre-trained model | VGGish | YAMNet | OpenL3 |
|----------------------|-------------------|--------|--------|--------|
| Black-and-white image| 96*64*1           | 96*64*1| 128*199*1 |
| RGB image            | 96*64*3           | 96*64*3| 128*199*3 |

**TABLE 7.** Training parameter of transfer model based on CNN (audio).

| Training Options | Optimization function | Gradient DecayFactor | Mini BatchSize | Epoch |
|------------------|-----------------------|----------------------|----------------|-------|
| Parameter values | Adam                  | 2                    | 256            | 5     |

Figs. 21 and 22 show the classification performance of the transfer learning models with different feature extraction methods when B&W image and RGB image are input before speech data noise removal and data augmentation, respectively. As shown in Fig. 21, when the B&W image of the 2D time-frequency transform-based features of audio signals is input to the transfer learning model before preprocessing, the performance is generally similar between different cases. Among these cases, the model showing the highest performance in all features is the VGGish model with an accuracy of 76.36%.

Similarly, as can be seen in Fig. 22, similar performance is exhibited when RGB image features are entered as input data. As a result, before preprocessing, the CNN-based transfer learning model was trained more intensively for the non-depression class which has a large amount of data compared to that of depression class, showing that the classification performance of depression data was not high, on the whole.

Figs. 23 and 24 show graphic presentation of the classification performance for the transfer learning models with different feature extraction methods after speech data noise removal and data augmentation when B&W and RGB image features are entered as inputs, respectively. As shown in Figure 23, when model training is performed using the Bark and Log-Mel spectrogram features, the performance is generally good. By contrast, the ERB spectrogram showed that the features were not presented clearly in the B&W image, indicating that the data could not be properly classified. When the OpenL3 model was trained using the Log-Mel spectrogram...
features, it showed the highest classification performance, with an accuracy of 95.12%.

In Fig. 24, it can be seen that overall performance improved compared to the result with B&W image as the input. In the ERB spectrogram, when RGB images were input to the transfer learning model, the two-dimensional features of the audio data were exhibited clearly, showing a good performance of classification by class, compared to the case when B&W images were entered as inputs. In this case, each transfer learning model showed the highest performance with Log-Mel spectrogram as input. In particular, the classification performance of the OpenL3 model was 96.34%, showing the highest performance.

As can be observed from the above results, in the first experiment, the classification performance of the Bi-LSTM and CNN-based transfer learning model was improved by improving the data quality and resolving the class imbalance issue after preprocessing of speech data for noise removal and depression data augmentation. In this case, the classification performance of the BI-LSTM model showed the highest performance of 96.34% when the number of hidden units in the Bi-LSTM layer is 10 and MFCC and GTCC features are extracted together. The performance of the CNN-based transfer learning model showed the highest performance in depression diagnosis with a classification accuracy of the OpenL3 transfer learning model of 96.34% when the log-mel spectrogram features were used with the input of RGB images.

The second experiment was performed to show the results of BI-LSTM and CNN-based depression diagnosis experiments using text script data from the EDAIC-WOZ dataset. The experiments were divided into cases of before and after depression data augmentation, and as presented in Table 8, the preprocessing methods of text data were divided into three types and the dimensions of the embedding layer were changed to 100, 200, and 500 to conduct experiments. The training options of the BI-LSTM and CNN models were used for model training by specifying parameter values as shown in Table 9.

### TABLE 8. Text preprocessing method.

| Preprocessing                          |
|---------------------------------------|
| method 1                              |
| Tokenization  →  convert text to lowercase  →  remove punctuation marks |
| method 2                              |
| Tokenization                           |
| method 3                              |
| Tokenization  →  topic word extraction  →  remove punctuation marks  →  remove unnecessary words  →  remove short or long sentences |

### TABLE 9. Training parameter of Bi-LSTM and transfer model based on CNN (text).

| Parameter values | Adam | 2 | 512 | 10 |
|------------------|------|---|-----|----|
| Optimization function | Gradient DecayFactor | Mini BatchSize | Epoch |

First, we discuss the experimental results of depression diagnosis using the BI-LSTM-based transfer learning model with text data. Table 10 presents the classification performance of Bi-LSTM before and after text data augmentation. From the table, before data augmentation, the overall performance is similar regardless of the values of embedding dimensions and preprocessing method. Among them, the highest performance is 78.18%, when only tokenization
preprocessing of text data is applied with embedding dimensions at 100.

From the table, it can be seen that depression data could not be properly classified before data augmentation. However, as the class imbalance problem was resolved after augmentation of the depression text data using EAD, although the accuracy was still not high, the result showed a balanced classification between depression and non-depression classes. In the experiment, with method 2 (text data tokenization) applied as the preprocessing method, after data augmentation with the embedding dimensions set to 200, the model accuracy is 76.83%, showing the highest performance.

TABLE 10. Performance of BI-LSTM Model with text data.

| Preprocessing method | 100   | 200   | 500   |
|----------------------|-------|-------|-------|
| Before data          |       |       |       |
| augmentation         |       |       |       |
| method 1             | 76.36%| 76.36%| 70.91%|
| method 2             | 78.18%| 76.36%| 76.36%|
| method 3             | 70.91%| 69.89%| 76.36%|
| After data           |       |       |       |
| augmentation         |       |       |       |
| method 1             | 70.73%| 64.63%| 74.39%|
| method 2             | 70.73%| 76.83%| 64.63%|
| method 3             | 73.17%| 70.73%| 69.51%|

Next, we discuss the experimental results of depression diagnosis using a CNN-based transfer learning model with text data. Table 11 presents the classification performance of the CNN model before and after text data augmentation. From the table, it can be seen that the overall performance is similar before data augmentation. Among the results of before augmentation, when the first preprocessing method (method 1) is used and the embedding dimensions are 100, the model shows the highest performance at 78.18%. After data augmentation, when method 2 (text data tokenization) is applied as the preprocessing method, the CNN model shows good overall performance. In particular, when training was carried out by specifying the dimension of the embedding layer as 100, the model showed the best performance with an accuracy of 81.71%. This result confirms that the performance is improved by 4.88% compared to that of the Bi-LSTM model.

From the results above, in the second experiment, the classification performance of the deep learning model did not improve significantly after data augmentation. However, by resolving the class imbalance problem, it was possible to prevent the problem of model learning bias toward a class with a large amount of data. The Bi-LSTM model shows the highest performance of 76.83% when dimensions of the embedding layer are set at 200 and only the data tokenization is applied. In addition, for the CNN model, when the embedding dimension is set to 100 and only the data tokenization is applied, an accuracy of 81.71% is achieved, indicating that these conditions were the most optimized conditions for depression diagnosis of the model using text data.

In the last experiment, we show the depression diagnosis performance of the four-stream deep learning model proposed in this study with late score fusion method applied, using multi-modal data. Table 12 presents three cases of feature extraction and preprocessing methods for audio and text data showing the highest performance in the first and second experiments for late score fusion. To evaluate the depression diagnosis performance of the four-stream deep learning model, fixed values are used for each case in the table. Finally, after speech data noise removal and data augmentation, the classification performance of the four-stream deep learning model is evaluated by applying the late score fusion method with the fixed values summarized in Table 12, based on the performance of the CNN-based transfer learning model.

FIGURE 25. Graphic representation of performance of four-stream-based deep learning model with late score sum method.

In Fig. 25, the performance of the four-stream model for each case is illustrated in graphs when all four deep learning
model softmax values are added using the late score sum method. As shown in Fig. 25, the accuracy when using the VGGish and OpenL3 models in case 2 is 97.56%, thus indicating that the model performance improved by 1.22% compared to the highest performance of a single model of case 2 in Table 12.

Fig. 26 shows the confusion matrix for the OpenL3_four-stream model, which has the highest performance among the four-stream-based deep learning models using the late score sum method. Fig. 26 (a) and (b) show the confusion matrix of the depression diagnosis model using audio data. From the confusion matrix, it can be seen that classification is performed relatively uniformly between depression and non-depression classes with high probability. Fig. 26 (c) and (d) show the confusion matrix of the depression diagnosis model using text data. Although the probability is not high, it can be seen that the classification result is not biased toward only one class, but is relatively uniform between the two classes. Fig. 26 (e) shows the confusion matrix for the late score

![Confusion Matrix Validation Data](image)
sum-based 4-stream model. We can see from the confusion matrix that although one data for each class could not be classified, classification was properly performed for the two classes of non-depression and depression.

Fig. 27 shows the performance (in graphs) of the four-stream model for each case when all four deep learning model softmax values are multiplied using the late score product method. As shown in Fig. 27, in case 1, the accuracy with the use of the YAMNet model is 97.56%, indicating an improvement in the model performance by 1.22%, compared to the highest performance of a single model in case 1 in Fig. 25. Also, in case 2, the accuracy with the use of the OpenL3 model is 98.78%, which indicates an improvement by 2.44% compared to the highest performance of the single model in case 2 in Table 12. Finally, in case 3, it can be seen that the performance of all models is improved compared to the highest performance of the single model in case 3 in Table 12.

Fig. 28 shows the confusion matrix for the OpenL3_four-stream model, which has the highest performance among the four-stream-based deep learning models using the late score product method. The confusion matrices shown in Fig. 26 (a), (c), and (d) show the same values as the confusion matrices shown in Fig. 26 (a), (c), and (d). Fig. 28 (b) shows the confusion matrix when the Log-Mel spectrogram features are applied to the OpenL3 model as RGB image inputs. From the result, it can be seen that classification into depression and non-depression is performed without imbalance, and with high probability. Fig. 28 (e) shows the confusion matrix for the late score product-based four-stream model. The result shows that all the data except for one depression data are properly classified.

**FIGURE 28.** Confusion matrix of four-stream-based deep learning model with late score product method (case3, OpenL3_4-stream model).
As a result, when all softmax probability values are multiplied, the four-stream models with OpenL3 in case 2 and with YAMNet and OpenL3 in case 3 have the highest classification accuracy of 98.78%. Therefore, the accuracy of the four-stream model using multi-modal data was improved by 2.44% compared to the use of single-mode data, demonstrating that the performance of the deep learning model for depression diagnosis was improved by the proposed method.

### TABLE 12. Feature extraction method, preprocessing method and accuracy of data for late score fusion.

| Data | Feature extraction and preprocessing method | Accuracy (%) |
|------|---------------------------------------------|--------------|
| case 1 | audio | 1D : hidden units 10, MFCC + GTCC | 96.34 |
| | | 2D : Bark Spectrum(RGB) | 85.37/86.59/84.15 |
| | text | Bi-LSTM : Embedding dimensions 200, tokenization | 76.83 |
| | | CNN : Embedding dimensions 100, tokenization | 81.71 |
| case 2 | audio | 1D : hidden units 10, MFCC + GTCC | 96.34 |
| | | 2D : ERB Spectrum(RGB) | 87.80/89.02/93.90 |
| | text | Bi-LSTM : Embedding dimensions 200, tokenization | 76.83 |
| | | CNN : Embedding dimensions 100, tokenization | 81.71 |
| case 3 | audio | 1D : hidden units 10, MFCC + GTCC | 96.34 |
| | | 2D : Log-mel Spectrum(RGB) | 96.90/95.12/96.34 |
| | text | Bi-LSTM : Embedding dimensions 200, tokenization | 76.83 |
| | | CNN : Embedding dimensions 100, tokenization | 81.71 |

In the study, for validation of the effective performance of the proposed four-stream-based model for depression diagnosis, the performance of state-of-the-art methods and that of the proposed model was comparatively analyzed. The database used in the previous studies was the DAIC-WOZ database, whereas the database used in this study was the EDAIC-WOZ database. To compare the performances under the same conditions, an experiment was conducted based on the model with the best performance using the DAIC-WOZ data. Table 13 shows the performance of the proposed model using DAIC-WOZ data. As can be seen in the table, the case of using DAIC-WOZ data also shows the best performance at 96.67% when late score product method was applied with multi-modal data rather than with single mode data.

For performance comparison with the state-of-the-art methods, performance indicators such as precision, recall, and F1-Score were used. Precision is the ratio of data that is actually true among the data predicted by the model, recall is the ratio of data that the model correctly predicted that is actually true, and F1-Score refers to the harmonic mean of precision and recall. Precision, recall, and F1-Score all have values in the range between 0 and 1, and the closer the value is to 1, the better the performance.

### TABLE 13. Performance of the proposed model using DAIC-WOZ data.

| Data | Feature extraction and preprocessing method | Deep Learning Model | Accuracy |
|------|---------------------------------------------|---------------------|----------|
| audio | MFCC+GTCC | Bi-LSTM | 90.00% |
| text | log-mel spectrogram | OpenL3 | 93.33% |
| audio and text | sum | four-stream model | 91.67% |

### TABLE 14. Comparison of performance between the state-of-the-art methods and the proposed four-stream model.

| Data | F1-score | Precision | Recall |
|------|----------|-----------|--------|
| Albanai et al. [17] | DAIC-WOZ | 0.77 | 0.71 | 0.83 |
| Lin et al. [18] | DAIC-WOZ | 0.85 | 0.79 | 0.92 |
| Lam et al. [19] | DAIC-WOZ | 0.87 | 0.91 | 0.83 |
| proposed four-stream model | DAIC-WOZ | 0.97 | 0.97 | 0.97 |
| proposed four-stream model | EDAIC-WOZ | 0.99 | 1.00 | 0.98 |

As a result of comparing the performance using these metrics, as can be seen in Table 14, when the DAIC-WOZ database is used, the precision of the proposed model is 0.97, recall is 0.97, and F1-score has values of 0.97. It can be confirmed that the performance of the proposed model is higher than that of the state-of-the-art methods under the same conditions. Also, when using the EDAIC-WOZ database, precision is 1.00, recall is 0.98, and F1-score is 0.99. This can be more easily seen from the graphs in Fig. 29. This result indicates that the proposed model has the best performance compared to the state-of-the-art methods, and finally, confirms that the proposed four-stream-based deep learning model of Bi-LSTM and CNN model
using the audio and text data is effective for depression diagnosis.

VII. DISCUSSION

This study aims to design a deep learning model that can analyze various characteristics and information of depression patients by combining data, and diagnosing depression effectively. This study demonstrates to solve the problems of data quality and class imbalance through noise removal and augmentation of depression data and positively influences the results by fusing various features of multi-modal data. Consequently, we prove that the proposed 4-stream deep learning model has the best performance and is more effective in diagnosing depression compared to previous studies [17], [18], [19] when voice and text data are used and various features are combined. However, there is a limitation in that the performance of models using text data is lower than that of models using voice data, and performance improvement is needed. Therefore, for a more objective depression diagnosis, additional research on how to accurately identify the characteristics of depression in text data and develop a deep learning model that has good performance is needed.

FIGURE 29. Illustration of performance comparison between the state-of-the-art methods and the proposed four-stream model.

VIII. CONCLUSION

In this study, we designed a four-stream-based depression diagnosis model using the late score fusion method for Bi-LSTM and CNN models based on the use of audio and text multi-modal data and compared the performance of the proposed model. Depression diagnosis using multi-modal data allows acquiring more information for depression, a complex mental health condition, and thus is effective for diagnosis. Among multi-modal data, linguistic features such as speech and text reveal the characteristics of people with symptoms of depressive disorder.

Audio signals may contain noise depending on the environment or equipment used for acquisition of the signals. Since the quality of data can affect the model performance, the audio signals were denoised using pyAudioAnalysis. The database used in this study is EDAIC-WOZ, a depression database, which includes audio and text data from a total of 275 participants, including 66 people with symptoms of depression and 209 people without symptoms. In this database, there is a problem of class imbalance due to the size of the non-depression data being about three times that of depression data. To resolve this problem, data augmentation was performed for each file of depression data to increase by three, and the augmented data was then used for model training. The preprocessed audio signals were divided into 1D and 2D for feature extraction and the extracted features were applied to Bi-LSTM and CNN-based transfer learning models, respectively.

Text data needs to undergo preprocessing through data cleansing processes such as tokenization, change of case (from upper to lower), and deletion of special characters. Furthermore, to use the preprocessed data as inputs to the deep learning models, text must be converted into numeric sequences. As in the audio data, data augmentation was performed with the depression data to increase each data file by three using the EDA method to resolve the class imbalance problem. The preprocessed text data was applied to the Bi-LSTM model including the word embedding layer and the n-gram-based CNN model.

Experiments were conducted using 80% of the total data as training data and the remaining 20% as validation data. For audio signals, one-dimensional features were extracted using MFCC and GTCC feature extraction techniques, and Bark, ERB, and Log-Mel spectrogram features were extracted based on two-dimensional time-frequency transform. By applying the extracted one-dimensional features to the Bi-LSTM model, the best performing methods and conditions after denoising and augmentation of the speech data were evaluated. The proposed model showed the highest performance with accuracy of 96.34% when the number of hidden units was 10 and MFCC and GTCC features were used together. By applying the extracted two-dimensional features to the CNN-based transfer learning models, VGGish, YAMNet, and OpenL3, the model performance improved after denoising and augmentation of the speech data. When RGB images with Log-mel spectrogram features were entered as inputs, the OpenL3 model showed the highest performance with an accuracy of 96.34%. In experiments using text data, in the case of Bi-LSTM, model performance of 76.83% was obtained when only tokenization was applied as a preprocessing method after data augmentation and the dimension of the embedding layer was set to 200. In the case of CNN models, the highest performance of 81.71% was achieved when only tokenization was applied as a preprocessing method after data augmentation and the embedding dimensions were set to 100.

Based on the audio and text data, the performance of the four-stream-based deep learning model using the late score fusion method was 98.78%, which was improved by
Y. Shen, H. Yang, and L. Lin, “Automatic depression detection: An emotional-audio-textual corpus and a GRU/BiLSTM-based model,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 6247–6251.

T. Guo, W. Zhao, M. Alrashoud, A. Tolba, S. Firmin, and F. Xia, “Multimodal educational data fusion for students’ mental health detection,” IEEE Access, vol. 10, pp. 70370–70382, 2022.

J. H. Park and N. M. Moon, “Design and implementation of attention depression detection model based on multimodal analysis,” Sustainability, vol. 14, no. 6, pp. 1–15, 2022.

J. Xiao, Y. Huang, G. Zhang, and W. Liu, “A deep learning method on audio and text sequences for automatic depression detection,” in Proc. 3rd Int. Conf. Appl. Mach. Learn. (ICAML), Jul. 2021, pp. 388–392.

M. Niu, K. Chen, Q. Chen, and L. Yang, “HACG: 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.

P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in Proc. Int. Conf. Learn. Represent. (ICLR), Feb. 2018, pp. 1–12.

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. 53–59.

T. Alhanai, M. Ghassemi, and J. Glass, “Detecting depression with audio/text sequence modeling of interviews,” in Proc. Interspeech, pp. 1716–1720, Sep. 2018.

L. Lin, X. Chen, Y. Shen, and L. Zhang, “Towards automatic depression detection: A BiLSTM/1D CNN-based model,” Appl. Sci., vol. 10, no. 23, pp. 1–20, Dec. 2020.

G. Lam, H. Dongyan, and W. Lin, “Context-aware deep learning for multimodal depression detection,” Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2019, pp. 3946–3950.

S. Chu, S. Narayanan, and C.-C.-J. Kuo, “Environmental sound recognition with time–frequency audio features,” IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

J.-M. Liu, M. You, G.-Z. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An, and S. Chen, “Cough signal recognition with gammatone cepstral coefficients,” in Proc. IEEE China Summit Int. Conf. Signal Inf. Process., Jul. 2013, pp. 160–164.

E. Zwickler, “Subdivision of the audible frequency range into critical bands,” J. Acoust. Soc. Amer., vol. 33, no. 2, p. 248, Feb. 1961.

P. Torben, “Acoustic communication. Hearing and speech. Version 2.0,” Acoust. Commun.-DTU, Lect. Note 31230-5, Aug. 2005, pp. 1–94.

B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” Hearing Res., vol. 47, nos. 1–2, pp. 103–138, Aug. 1990.

S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gennake, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “CNN architectures for large-scale audio classification,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Signal Process., Mar. 2017, pp. 131–135.

Z. S. Syed, S. Ali, and A. Latif, “Deep acoustic embeddings for identifying parkinsonian speech,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 10, pp. 726–734, 2020.

A. G. Howard, M. Zhu, B. Chen, D. Kalchenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861.

J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 1–5.

R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 609–617.

A. Cosimini, P. Rohoska, M. Kristoffersen, S. Shepstone, and Z.-H. Tan, “Subjective annotations for vision-based attention level estimation,” in Proc. 14th Int. Joint Conf. Comput. Vis., Imaging, Comput. Graph, Theory Appl., 2019, pp. 249–256.

F. Ringeval, É.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, M. Pantic, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, and S. Amirjanzadeh, “AVEC 2019 workshop and challenge: State-of-mind, detecting depression with A and cross-cultural affect recognition,” in Proc. 9th Int. Audio/Visual Emotion Challenge Workshop (AVEC), 2019, pp. 3–12.
A.-H. Jo, K.-C. Kwak: Diagnosis of Depression Based on Four-Stream Model of Bi-LSTM and CNN

[32] 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. LREC*, May 2014, pp. 3123–3128.

[33] D. DeVault et al., “SimSensei kiosk: A virtual human interviewer for healthcare decision support,” in *Proc. 13th Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, Jan. 2014, pp. 1061–1068.

[34] S. S. Dhingra, K. Kroenke, M. M. Zack, T. W. Strine, and L. S. Balluz, “PHQ-8 days: A measurement option for DSM-5 major depressive disorder (MDD) severity,” *Population Health Metrics*, vol. 9, no. 1, pp. 1–8, Apr. 2011.

[35] T. Giannakopoulos, “PyAudioAnalysis: An open-source Python library for audio signal analysis,” *PLoS ONE*, vol. 10, pp. 1–17, Dec. 2015.

[36] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in *Proc. Conf Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6382–6388.

A.-HYEON JO received the B.S. and M.S. degrees in electronic engineering from Chosun University, Gwangju, South Korea, in 2021 and 2022, respectively, where she is currently pursuing the Ph.D. degree in electronic engineering majoring in IT-bio convergence system. Her research interests include artificial intelligence, speech recognition, and deep learning.

KEUN-CHANG KWAK (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Electronics Engineering, Chungbuk National University, Cheongju, South Korea, in 1996, 1998, and 2002, respectively. He was a Postdoctoral Fellow at the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, from 2003 to 2005. He was a Senior Researcher at the Human–Robot Interaction Team, Intelligent Robot Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, from 2005 to 2007. He was also a Visiting Professor at the Department of Computer Science, California State University, Fullerton, CA, USA, from 2014 to 2015. He is currently a Professor with the Department of Control and Instrumentation Engineering, Chosun University, Gwangju, South Korea. His current research interests include computational intelligence, human–robot interaction, and biometrics.