computation

MDPI

Article

Multimodal Data Fusion for Depression Detection Approach

Mariia Nykoniuk [1], Oleh Basystiuk [1,*], Nataliya Shakhovska [1,2] and Nataliia Melnykova [1]

[1] Department of Artificial Intelligence, Lviv Polytechnic National University, Stepan Bandera 12, 79013 Lviv, Ukraine; mariia.nykoniuk.knm.2019@lpnu.ua (M.N.); nataliya.b.shakhovska@lpnu.ua (N.S.); nataliia.i.melnykova@lpnu.ua (N.M.)
[2] Department of Civil and Environmental Engineering, Brunel University of London, Uxbridge UB8 3PH, UK
* Correspondence: oleh.a.basystiuk@lpnu.ua

Abstract: Depression is one of the most common mental health disorders in the world, affecting millions of people. Early detection of depression is crucial for effective medical intervention. Multimodal networks can greatly assist in the detection of depression, especially in situations where in patients are not always aware of or able to express their symptoms. By analyzing text and audio data, such networks are able to automatically identify patterns in speech and behavior that indicate a depressive state. In this study, we propose two multimodal information fusion networks: early and late fusion. These networks were developed using convolutional neural network (CNN) layers to learn local patterns, a bidirectional LSTM (Bi-LSTM) to process sequences, and a self-attention mechanism to improve focus on key parts of the data. The DAIC-WOZ and EDAIC-WOZ datasets were used for the experiments. The experiments compared the precision, recall, f1-score, and accuracy metrics for the cases of using early and late multimodal data fusion and found that the early information fusion multimodal network achieved higher classification accuracy results. On the test dataset, this network achieved an f1-score of 0.79 and an overall classification accuracy of 0.86, indicating its effectiveness in detecting depression.

Keywords: depression detection; multimodal networks; early fusion; late fusion; mental health; deep learning

Academic Editors: Dmytro Chumachenko and Sergiy Yakovlev

Received: 28 November 2024
Revised: 22 December 2024
Accepted: 25 December 2024
Published: 2 January 2025

Citation: Nykoniuk, M.; Basystiuk, O.; Shakhovska, N.; Melnykova, N. Multimodal Data Fusion for Depression Detection Approach. *Computation* **2025**, *13*, 9. https://doi.org/10.3390/computation13010009

1. Introduction

Depression is one of the most common mental health disorders worldwide. According to the World Health Organization (WHO) [1], more than 280 million people, or approximately 5% of the adult population, suffer from this condition. The impact of depression goes beyond emotional distress, significantly affecting a person's ability to perform daily tasks, maintain personal relationships, and stay physically and mentally healthy. Symptoms of depression can include feelings of persistent sadness, loss of interest in usual activities, fatigue, sleep problems, changes in appetite, and concentration [1]. Identifying depression early is critical, as this mental disorder tends to worsen without proper treatment. Undiagnosed or untreated depression can lead to serious physical and mental complications, such as chronic illness, reduced quality of life, and even suicidal thoughts or actions. People suffering from depression often lose the ability to work productively, maintain healthy relationships, and even perform basic daily tasks.

Artificial intelligence can be of great help in detecting depression, especially in situations wherein patients are not always aware of or able to express their symptoms. Multimodal approaches combine different types of data, such as analyzing how a person speaks (audio) and what they say (text), to identify patterns related to depression. The

multimodal approach (combining audio and text) was chosen over unimodal analysis because it provides more complete data about the participant's emotional and mental state. While text-only analysis focuses on the content of what is said, it does not take into account important paralinguistic cues present in the audio, such as tone, pitch, and rhythm, which can be strong indicators of emotional distress. Similarly, audio-only analysis does not take into account linguistic patterns in spoken language, such as specific word usage or sentence structure, which can reveal cognitive signs of depression.

By integrating audio and textual modalities, the model can detect subtle clues that would otherwise be missed if it focused on only one type of data. This approach provides a more holistic view of the participant's state, significantly increasing the accuracy of depression detection, especially in cases wherein one modality alone may not be sufficient to detect depression.

The contribution of this work is highlighted as follows:

- This work outlines an effective data preprocessing pipeline, addressing the issue of class imbalance.
- The study proposes and develops two multimodal networks for depression detection—early fusion and late fusion models. These models combine both audio and text modalities to capture a more holistic set of features relevant to identifying depressive symptoms.
- The performance of the developed models is rigorously evaluated using multiple performance measures, including accuracy, recall, precision, and f1-score.
- A detailed interpretation of its results on both validation and test datasets, illustrating its effectiveness in detecting depression across various scenarios.

2. Related Work

Depression detection using artificial intelligence has been an active area of research in recent years. Various approaches have been applied, utilizing different types of data such as text, audio, and even visual information. Early works in this domain focused predominantly on single-modality data (either text or audio) for classification tasks.

This article could be improved by integrating a broader and more representative range of literature sources that cover the development and evaluation of depression detection models using artificial intelligence. The current review primarily references a limited selection of studies and does not fully capture the diversity of approaches in the field, particularly when it comes to multimodal detection, dataset balancing, and model performance evaluation. To demonstrate a more comprehensive understanding of the field, this article could incorporate references to key papers that explore recent advancements in both single-modality and multimodal approaches.

While studies such as [2–5] provide valuable insights into the use of speech-related features like log spectrograms, MFCC, COVAREP, and HOSA for detecting depression, this article overlooks some of the more recent methods in this area, such as newer feature extraction techniques or improved neural network architectures. Including references to works that explore deep learning models beyond CNNs and RNNs [6], such as attention-based mechanisms or self-supervised learning for speech-based depression detection, would broaden the discussion.

This article acknowledges the use of traditional NLP methods such as Bag of Words (BoW) and TF-IDF, as well as more recent deep learning models like LSTM for text-based depression detection. However, there is a growing body of work on pre-trained transformer models (e.g., BERT, RoBERTa, S-RoBERTa), which have shown significant improvements in understanding linguistic cues from textual data. Citing studies that explore these advanced

transformer models for better accuracy in depression detection through text analysis would enhance this article's coverage of text-based approaches.

While studies like [7–9] are cited for their multimodal fusion approaches, this article could benefit from incorporating additional works that investigate the use of multimodal systems combining not only speech and text but also visual cues, such as facial expressions and body language, for more holistic depression detection. Including studies that explore early fusion or hybrid fusion models could expand the discussion on how different data modalities can be leveraged together.

This article rightly points out the issue of dataset imbalance in studies like [7,8], but it could strengthen this point by discussing techniques for addressing class imbalance, such as oversampling, undersampling, and data augmentation, which are commonly used in depression detection research. Additionally, incorporating literature on fairness and model generalizability would highlight the importance of ensuring that models are not biased toward over-represented classes and can effectively perform across diverse patient populations.

Some of the studies reviewed (e.g., [9]) achieve low accuracy, indicating that such models may not be reliable enough for practical use in clinical settings or everyday use.

Finally, this article could deepen its analysis of model performance, particularly with reference to accuracy and reliability in clinical or real-world settings. While studies like [9] report lower performance metrics, a more nuanced discussion of evaluation metrics beyond accuracy—such as precision, recall, f1-score, and area under the ROC curve—would provide a clearer picture of how these models perform in various contexts. By expanding the citation range and including these relevant works, this article would demonstrate a more thorough understanding of the state of the field and provide a more balanced overview of the challenges and opportunities in depression detection using artificial intelligence.

Thus, the main challenge in the field of depression diagnosis is to develop reliable, accurate, and fair models that can effectively detect signs of depression. Existing methodologies have several serious problems, including unbalanced datasets, dependence on a single data modality, and generally low prediction accuracy. Addressing these issues will significantly increase the reliability of depression detection tools, making them more effective for clinical and personal health monitoring.

3. Proposed Approach

3.1. Dataset

This study uses the DAIC-WOZ (Depression and Anxiety Interview Corpus—Wizard of Oz) dataset [10], which is one of the key datasets used for research in the field of detecting depression and anxiety disorders based on multimodal data.

DAIC-WOZ contains recordings of interviews between participants and a virtual agent named Ellie, who was programmed as an interlocutor to collect information about a person's mental state. The virtual agent asks participants a variety of questions about their mental health, emotional state, life events, etc.

The DAIC-WOZ dataset contains multimodal interview recordings, including the following:

- Textual data. Interview transcripts that provide a detailed record of verbal communication between the patient and the virtual interviewer. The transcripts contain the exact start and end time of each cue, which makes it possible to synchronize the audio with the text.
- Audio recordings. The set contains audio files that record the voice of patients during interviews.

The DAIC-WOZ contains several important features that make it unique for a study of depression detection:

- The dataset includes 189 interviews, which is a small dataset size for model training.
- The length of the interview audio recordings can vary from a few minutes to half an hour, depending on the pace of the conversation and the number of questions asked to the patient.
- Each text file has time annotations that allow you to synchronize text and audio. These annotations are important for analysis because they allow us to more accurately identify specific moments in the interview where the patient shows signs of depression. They also simplify the procedure of removing interviewer questions from the text and audio, as these annotations contain the beginning and end of each line.

One of the key problems with the DAIC-WOZ dataset is the imbalance of classes between patients with and without depression.

The extended dataset EDAIC-WOZ (Extended Depression and Anxiety Interview Corpus—Wizard of Oz) [10] is an additional resource to the basic DAIC-WOZ dataset. It contains more records of interviews with patients diagnosed with depression, which were not included in the original dataset (Figure 1).

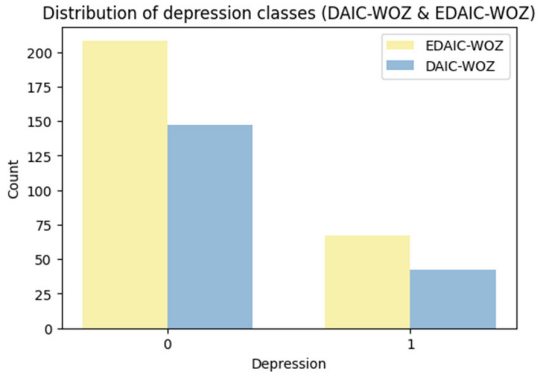

Figure 1. Comparative plot of depression class distribution in DAIC-WOZ and EDAIC-WOZ datasets.

The purpose of this extended dataset is to increase the amount of data available for depression analysis and provide researchers with more opportunities to train models on multimodal data.

At the same time, the extended EDAIC-WOZ dataset has some drawbacks related to transcript quality. Although the transcripts in this dataset contain the exact start and end times of each replica, there are several serious problems that make them difficult to process:

1. Lack of speaker separation. The EDAIC-WOZ transcripts do not clearly indicate whether a particular line belongs to the patient or the interviewer. This creates confusion, as sometimes the interviewer's question and the patient's answer may be combined in one text segment.
2. Inaccuracy of transcripts. Transcripts in the extended dataset may be inaccurate. This means that what is written in the transcript does not always accurately reflect what was actually said by the interviewee. This mismatch between the actual audio data and the text complicates the data preparation process, as it is necessary to manually double-check the transcripts against the audio. This is especially important for the accurate extraction of textual features, which can significantly affect the overall quality of the model.

For this study, only those recordings that are labeled as depression and do not belong to the regular DAIC-WOZ dataset were taken from the extended EDAIC-WOZ dataset (Figure 2).

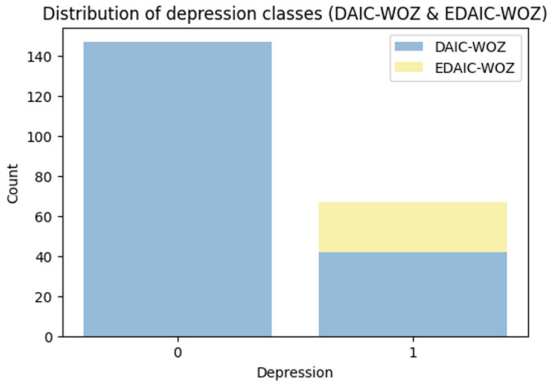

Figure 2. Graph of the final distribution of the depression classes.

The final data distribution consists of 120 non-depression cases from the DAIC-WOZ dataset, 40 depression cases from the DAIC-WOZ dataset, and 20 depression cases from the EDAIC-WOZ dataset. This approach effectively increases the number of examples in the depression class without duplicating existing data from the original DAIC-WOZ set. By including additional depression cases from the extended EDAIC-WOZ dataset, this method addresses the problem of class imbalance, where the depression class is underrepresented compared to non-depression cases. The adjusted distribution ensures a more balanced representation of both classes for analysis.

This will effectively increase the number of examples of the class "depression" without duplicating existing data from the original set. This approach can help to address the problem of imbalanced classes, wherein data with depression are significantly less represented compared to data without depression.

3.2. Data Processing

Audio. When working with audio data in depression detection tasks, it is important to properly prepare the data to ensure the accuracy of the model. The audio preprocessing process includes several important steps, each of which is aimed at improving data quality and increasing the efficiency of further feature analysis (Figure 3).

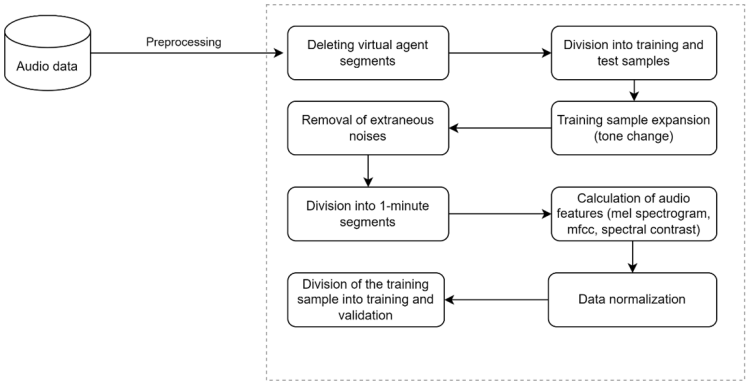

Figure 3. Audio data preprocessing algorithm.

According to the diagram above, the stages of audio data preprocessing include the following:

1. Removal of interviewer segments. First of all, all segments where the interviewer speaks are removed. This is important for analyzing only the participant's speech, as we are interested in identifying signs of depression in the patients' responses.
2. Dividing into training and test samples. In the next step, the data are divided into training and test samples for training and evaluation of the model in the ratio of 80:20.

The training set is used to train the model, and the test set is used to check its accuracy on new data.

3. Training sample augmentation (tone change). To increase the amount of data in the training set and balance the classes, audio data are augmented by changing the tone of speech, which allows you to create new variations of data without losing meaning. In this case, the tone is changed by shifting the signal frequencies by 1.5 semitones. This means that the pitch of the participant's voice was lowered by 1.5 semitones. It is important to note that changing the pitch within a small range, in this case by 1.5 semitones, allows us to preserve the naturalness of the voice while providing new audio variations.
4. Removal of extraneous noise. At this stage, background noise is removed from the audio. This includes clearing the audio of unwanted sounds such as noise, hum, or extraneous conversations.
5. Divide into segments of 1 min duration. Audio recordings are divided into 1 min segments. Segments that are too short (e.g., 5–10 s) may not contain enough information about the emotional state or tone of the respondent. When broken down into one-minute segments, the context of the conversation is preserved, which is important for understanding changes in the participant's tone and mood during the conversation.
6. Calculating audio features (Mel spectrograms, MFCC, spectral contrast). For each segment, audio features such as Mel spectrograms, MFCC coefficients (Mel frequency cepstral coefficients), and spectral contrast are calculated.
    - A Mel spectrogram is a spectrum of audio frequencies calculated on the Mel scale [11]. The Mel spectrogram shows the distribution of energy across frequencies over time, reflecting changes in voice and intonation in speech. It is well suited for analyzing temporal changes in audio.
    - MFCC are coefficients derived from the Mel spectrogram, which is a numerical representation of the sound spectrum on the Mel scale [12]. They allow for even more accurate characterization of audio signals. MFCCs are well suited for analyzing speech data because they capture the features of the speaker's articulation. They are able to distinguish tone, rhythm, intonation, and other speech aspects.
    - Spectral contrast is a method that measures the difference between the maximum and minimum amplitudes in different frequency bands [13]. In the case of speech analysis for depression, it is useful for recognizing changes in the voice, such as monotony or lack of expression, which are typical signs of depression.

These features make it possible to identify voice and speech characteristics that may indicate a depressive state. These features will be used as input for training the model.

7. Normalize the data. This step transforms the data so that they have a mean value of 0 and a standard deviation of 1. As a result, all features are brought to the same scale, which is critical for some machine learning models.
8. Dividing the training sample into training and validation samples. The training sample is further divided into training and validation samples in the ratio of 80:20. This makes it possible to evaluate the model performance on the validation data during the training process in order to adjust the model parameters.

Thus, this sequence of steps helps to prepare the audio data properly, which ensures high-quality training and subsequent classification of depression signs.

Text. Text data preprocessing (Figure 4) is an important step as it prepares textual cues for further feature extraction and removes redundant or uninformative information. This

helps the model to focus on the relevant data, which improves the quality of training and the accuracy of the results.

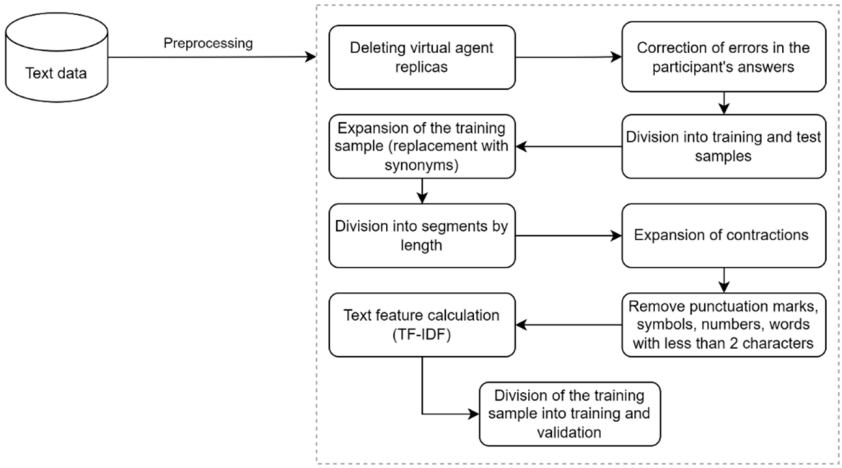

Figure 4. Text data preprocessing algorithm.

According to the diagram below, the stages of text data preprocessing include the following:

1. Removal of interviewer's remarks. The first step is to remove all questions asked by the interviewer. For data from the extended dataset (EDAIC-WOZ), where the annotations are inaccurate, the interviewer's remarks are not separated and completely removed. In this case, the removal of interviewer questions is conducted manually to ensure consistency between text and audio and to focus the analysis on the participants' responses.
2. Correction of errors in the participant's answers. Transcripts from the extended dataset sometimes contain errors or inconsistencies between what was said and what was recorded. At this stage, a manual check is performed to correct these errors and bring the text in line with the actual answers. This provides a more accurate analysis of the text data.
3. Dividing into training and test samples. To ensure the reliability of the results, the data are divided into training and test samples in the ratio of 80:20. This helps to check how the model will work on new data and prevents overfitting.
4. Training sample augmentation. Text augmentation is used to increase the amount of data and improve model accuracy. In this case, replacing words with synonyms makes it possible to generate new text variants while maintaining the meaning. Each text fragment is analyzed to identify words that can be replaced with synonyms. For this purpose, the WordNet lexical database is used, which contains synonyms for various words. In addition, not all words can be replaced. The text is processed in parts of 100 words, and the probability of replacing words with their synonyms is 15%. The 15% probability was chosen as a compromise that allows for adding variability to the data without changing the text too radically. A higher percentage can change the semantics of the text, which is undesirable, especially when analyzing mental states, where every detail is important. In addition, stop words (e.g., "I", "you", "depression") are specifically excluded from the list of replacement words. This is done to avoid changing keywords that may have a significant impact on the model or are critical to the analysis of depression. Augmentation is applied to all data with the depression class in the training set.
5. Split into segments by length. Based on the number of segments when splitting audio data, the text is divided into the same number of segments by word count.

6. Expanding contractions. During the interview, contractions may be used that need to be expanded to better understand the model.
7. Removal of punctuation marks, symbols, numbers, and words with less than two characters. This step involves removing non-informative characters such as punctuation, numbers, and words that are too short and do not have a meaningful impact on the analysis.
8. Calculation of textual features (TF-IDF). At this stage, textual features are calculated by increasing the weight of words related to depression and applying the TF-IDF (term frequency–inverse document frequency) method, which measures the frequency of each word in the text and its importance relative to other texts.

To increase the weight of words related to depression, a set of keywords (symptoms) that are directly related to the depressive state and pre-trained GloVe vectors (Global Vectors for Word Representation) are used, which represent words as numerical vectors in a multidimensional space. For each keyword, GloVe vectors are used to find semantically similar words. This is achieved by calculating the cosine similarity between the word vectors, which allows us to determine how similar two words are in their contexts in the texts. After that, for each keyword, the top 10 most similar words are determined and added to the main list.

When the list of words related to depression is supplemented with similar words, these words receive increased weight during text processing. This is achieved by repeating the word several times so that the model gives them more weight compared to other words.

After the weights of important words are increased, TF-IDF is used to vectorize the text.

9. Dividing the training sample into training and validation samples. After all the stages of text data preprocessing are completed, the training data are divided into training and validation samples in the ratio of 80:20. The training set is used to train the model, while the validation set is used to evaluate its performance and adjust the hyperparameters.

After the text data preprocessing is completed, the data are ready for further use in the model training process. It is important to ensure that the quality of the cleaned data is high, as even minor errors at this stage can lead to distortion of the results during modelling and reduce the accuracy of predictions.

3.3. A Multimodal Network for Depression Detection

Multimodal networks are neural architectures that combine different types of data to obtain deeper analysis and make predictions with increased accuracy [14,15]. In the context of detecting participants' depression, a multimodal network is a tool that can combine data from different formats, such as text interviews and audio recordings. This combination allows for capturing not only the content of what was said (through textual data), but also how it was said (through audio), which is especially important in the diagnosis of depression. In particular, textual data can contain important depressive keywords or phrases, while audio can provide information about the patient's emotional state through tone, speech rate, pauses, etc. An important aspect of multimodal networks is that different data sources, such as audio and textual data, can interact with each other, providing the model with more context for decision making.

There are two main approaches to data fusion in multimodal neural networks: early and late fusion. These methods determine at what stage of processing features from different modalities, such as audio and text, are combined.

3.3.1. Early Fusion Model

The architecture of the early fusion model shown in Figure 5 combines audio and textual features at an early stage, allowing for the efficient fusion of different types of data for more accurate depression classification.

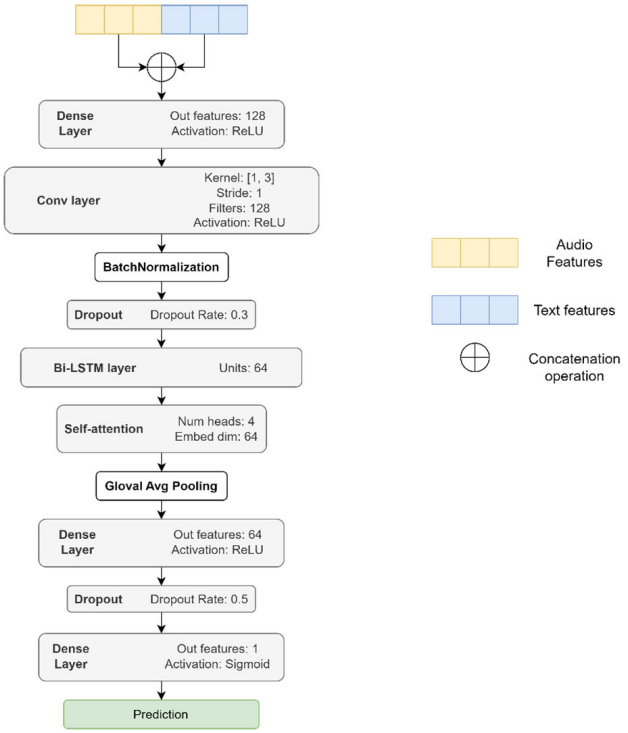

Figure 5. Early fusion multimodal network architecture.

Input data:

- Audio: the model is fed with extracted features from the audio, such as Mel spectrograms, MFCC (Mel frequency cepstral coefficients), and spectral contrast. These features reflect the acoustic characteristics of the signal.
- Text: Text data are represented as feature vectors extracted using TF-IDF.

The model integrates audio and text features through a concatenation layer, allowing for the combined processing of both data types. A dense layer with 128 neurons and ReLU activation reduces the complexity of the input, capturing nonlinear feature relationships. Next, a convolutional layer extracts important patterns, especially from audio data, with batch normalization helping stabilize the learning process. Dropout layers (30% and 50%) prevent overfitting, while the Bi-LSTM layer with 64 neurons captures sequential dependencies in both directions. A four-headed self-attention mechanism helps the model focus on key aspects of the input. Global average pooling reduces data dimensionality, followed by a final dense layer with 64 neurons, which prepares the data for the binary classification output via the Sigmoid function, determining the presence of depression.

This architecture combines the power of convolutional networks to learn local patterns, Bi-LSTM for contextual sequence processing, and a self-attention mechanism to improve focus on key parts of the data.

3.3.2. Late Fusion Model

The architecture of the late fusion model consists of two separate branches for processing audio and text features, which are combined at a later stage (Figure 6). This makes it possible to first extract specific features from each modality (audio and text), which

improves the quality of processing for each data type, and then combine these features for joint analysis.

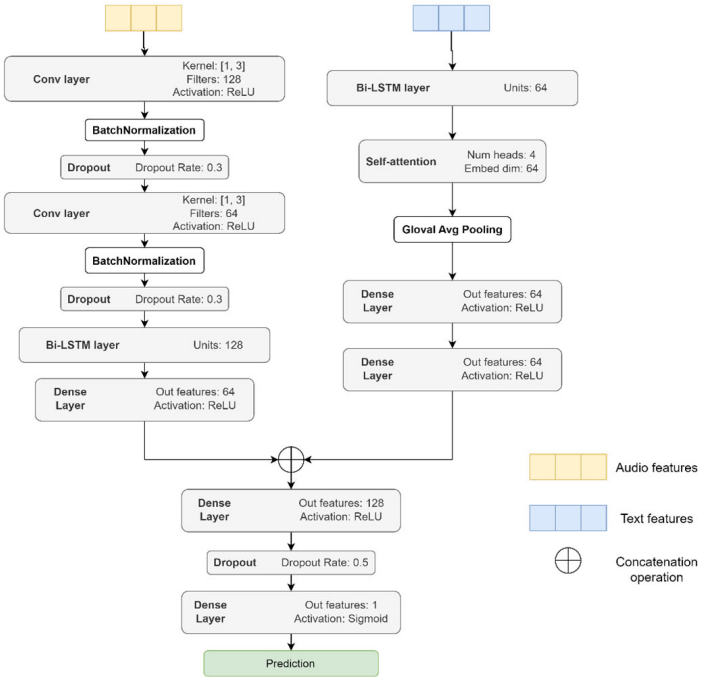

Figure 6. Late fusion multimodal network architecture.

The audio data are processed using a series of convolutional layers (CNNs) to extract local features, followed by Bi-LSTM to capture long-term dependencies in the sequence. The initial convolutional layers apply filters to detect patterns in the audio, with batch normalization and dropout layers to maintain stability and prevent overfitting. The Bi-LSTM layer, with 128 units, processes the sequential nature of audio data, extracting temporal relationships from different segments. A final dense layer further reduces the dimensionality of extracted features, preparing them for combination with text features.

For text data, a Bi-LSTM layer with 64 units is used to capture context in both forward and backward directions. The self-attention mechanism then focuses on the most relevant parts of the text, making it easier for the model to identify crucial indicators of depression. Global average pooling reduces the number of features, retaining the essential information for classification.

After separate processing, the audio and text features are combined using a concatenation layer in a late fusion stage. The resulting feature set is then passed through dense layers with ReLU activation to refine the combined features. A final output layer with a Sigmoid activation function is used to classify the combined representation as indicating the presence or absence of depression.

The chosen architecture makes it possible to first maximize the use of specific audio and text features and then combine them efficiently, which can contribute to better overall model performance.

4. Results

To implement multimodal data fusion models, the Google Colab environment was used, which is a powerful tool for conducting experiments and training neural networks. In this work, a T4 GPU with 15 GB of GPU RAM was used to train the models. Python 3.8 was chosen as the programming language for the project. The following libraries were

used in the process: NumPy, Pandas, nltk (Natural Language Toolkit), nlpaug, noisereduce, pydub, librosa, matplotlib, seaborn, scikit-learn, tensorflow, keras.

4.1. Performance of a Multimodal Early Fusion Network

The early fusion model was trained using the binary_crossentropy loss function for binary classification (depression/non-depression) and the accuracy metric for evaluation of model. The model was trained over 25 epochs. In addition, to ensure a stable learning process, the Adam optimizer was used with a predefined learning rate set at 0.00001.

The plots (Figure 7) show the accuracy and loss of the model on the training and validation data over the training epochs.

- Graph of accuracy: The first graph shows how the accuracy of the model increases over the epochs. The training set has a steady increase in accuracy to over 95%. The validation sample also shows a high level of accuracy, exceeding 90%, which indicates a good generalization of the model.
- Graph of losses: The second graph shows the reduction in model loss on the training and validation data. There is a significant decrease in loss during the first epochs, which is an indicator that the model is learning well. The loss of the validation data also decreases, indicating that there is no overfitting.

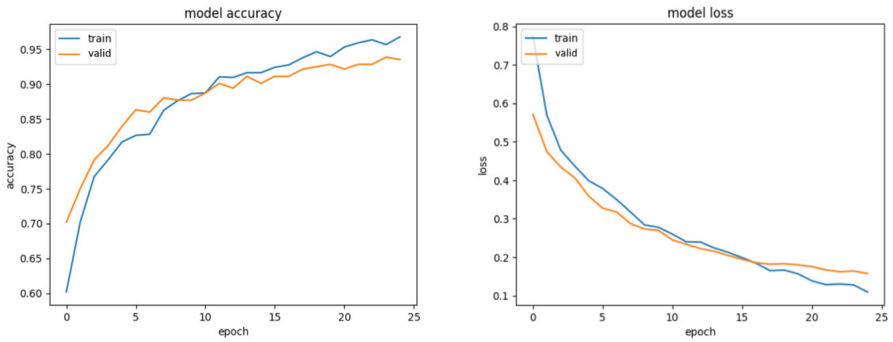

Figure 7. Training progress plots for the multimodal early fusion model.

Considering the confusion matrix for the validation data (Figure 8), we can see that the model correctly predicted 138 samples from class 0 (without depression) and made 10 false predictions. For class 1 (with depression), the model correctly classified 135 samples, and nine samples were classified incorrectly.

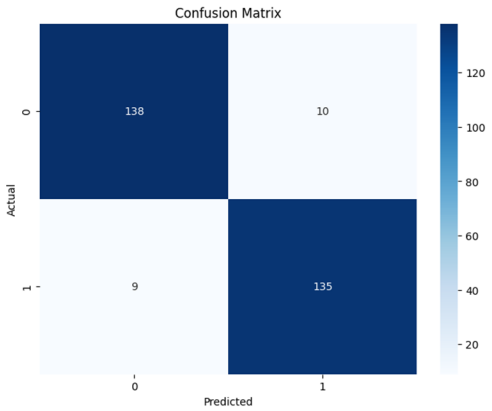

Figure 8. Confusion matrix for validation data (early fusion model).

Table 1 shows the obtained values of the metrics for evaluating the early fusion model performance on the validation data. The model has an accuracy of 0.94 for class 0 and 0.93

for class 1, which means that the model makes few errors among the predictions. For both classes, the model demonstrates high recall values (0.93 and 0.94), which shows that most samples in each class were correctly classified. The high f1-score values for both classes (0.94 and 0.93) indicate a balanced model that performs well in both precision and recall. The area under the curve (AUC) score of 0.90 highlights moderate discriminative power and strong correlation based on the Matthews correlation coefficient (MCC) of 0.87.

Table 1. Report on early fusion model performance metrics on validation data.

Metrics	Class	
	Non-Depression	Depression
Precision	0.94	0.93
Recall	0.93	0.94
F1-score	0.94	0.93
Accuracy	0.93	
AUC	0.90	
MCC	0.87	

In general, the model demonstrates high classification accuracy for both classes, which indicates its effectiveness in solving this task.

4.2. Performance of a Multimodal Late Fusion Network

The late fusion model was trained for 50 epochs. The Adam optimizer was used to train the model with the learning rate set to 0.01, which was configured to train faster than the previous early fusion model. In addition, the binary_crossentropy loss function, which is widely used for binary classification, was applied. The accuracy metric was used to evaluate the model's performance.

Figure 9 shows the improvement in accuracy and reduction in error over 50 epochs of training.

- Accuracy: The accuracy graph shows an increase for both training and validation data, although there are fluctuations, especially in the validation sample, which may indicate some problems in generalizing the model. The final accuracy of the model for the validation data is about 85%.
- Loss: The loss plot shows a gradual decrease for both training and validation data, which is a positive signal. However, again, there are some fluctuations, which may indicate the complexity of the data or the insufficient number of epochs for full stabilization.

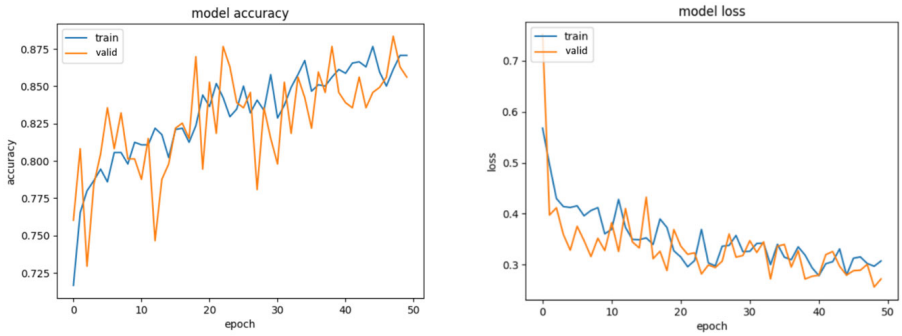

Figure 9. Training progress plots for the multimodal late fusion model.

Thus, both graphs show that the model has gradually learned to distinguish between depression, and although there are some fluctuations, the overall trend indicates good progress and satisfactory accuracy.

At the same time, the training results of the late fusion model are somewhat worse compared to the early fusion model. Although the validation accuracy of the late fusion model reaches about 85%, it is inferior to the accuracy of the early fusion model, which has shown much more stable results.

The confusion matrix (Figure 10) shows the distribution of model results on the validation sample. The model correctly classified 130 samples into class 0 (no depression) and 118 samples into class 1 (depression). However, there were errors: 18 samples in class 0 were misclassified as 1, and 26 samples in class 1 were misclassified as 0.

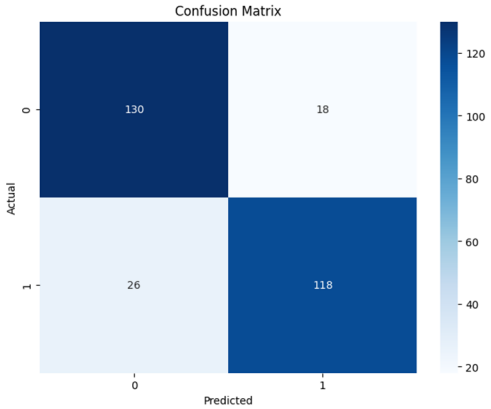

Figure 10. Confusion matrix for validation data (late fusion model).

These results indicate some difficulties in correctly classifying depression, but overall the model showed good accuracy in detecting both classes.

Analyzing the metrics report (Table 2), we can note the following:

For class 0 (without depression), the precision value is 0.83, which indicates a slightly higher number of false positive predictions. In class 1 (with depression), the precision value is higher—0.87, i.e., the model predicts depression somewhat better without too many false positive predictions. The recall value for class 0 is 0.88, which means that the model recognizes most patients without depression quite well. For class 1, the recall value is slightly lower at 0.82, indicating that there are some cases of depression that the model could not detect. For class 0, the f1-score value is 0.86, and for class 1, it is 0.84. This means that the model performs slightly better in detecting the absence of depression than its presence. The overall accuracy of the model is 0.85, which is a fairly good result given the complexity of the depression classification task. The area under the curve (AUC) score of 0.78 highlights moderate discriminative power and strong correlation based on the Matthews correlation coefficient (MCC) of 0.70.

Table 2. Report on late fusion model performance metrics on validation data.

Metrics	**Class**	
	Non-Depression	**Depression**
Precision	0.83	0.87
Recall	0.88	0.82
F1-score	0.86	0.84
Accuracy	0.85	
AUC	0.78	
MCC	0.70	

In general, the model performs well, although there is a slight difference in accuracy between the two classes.

Thus, comparing the training results of the two developed models—early and late fusion—we can see that the early fusion model shows better results in terms of metrics. It demonstrates higher precision, recall, and f1-Score, which indicates that it is more effective in detecting depression.

4.3. Interpretation of the Results of the Best Model

In order to interpret the results of the best model, the multimodal early fusion network, it is necessary to conduct testing on a test dataset. Since the data of each participant were divided into segments, the results for each participant must be summarized to obtain a single answer regarding their status (presence or absence of depression).

The process of summarizing the results for each participant is performed using the method of aggregating segment probabilities. In this case, the average value of segment probabilities for each participant is used. This allows us to determine a single forecast for the participant based on individual segments. If the average probability exceeds the threshold of 0.5, the conclusion is made that depression is present. Below is an example of the functions used for this purpose:

1. Aggregate results function: for each participant, the average probability value is calculated based on the forecasts for each segment.
2. Model evaluation function: after prediction on the segments of the test sample, the results are aggregated to obtain a prediction at the participant level, after which the model is evaluated by metrics such as accuracy, error matrix, and classification report.

This approach allows us to obtain more generalized results and correctly evaluate the model's performance at the participant level rather than at the individual segment level.

The confusion matrix (Figure 11) of the model testing on unknown data shows that the model correctly classified 26 participants without depression and 11 participants with depression. At the same time, the model made four errors when participants without depression were classified as having depression and two errors when participants with depression were classified as healthy.

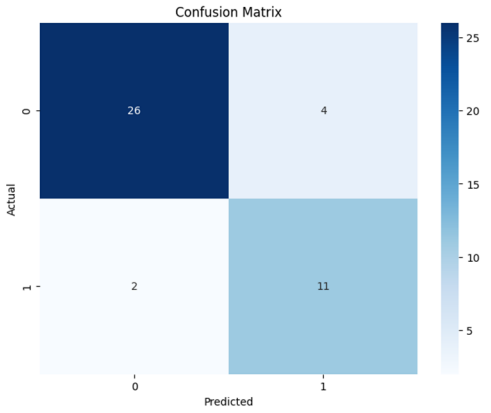

Figure 11. Confusion matrix for test data.

Based on the classification report (Table 3) from testing on unknown data, it can be seen that the model achieved an overall accuracy of 85%. Furthermore, the area under the curve (AUC) score of 0.73 highlights moderate discriminative power, suggesting that the model can distinguish between the two classes better than random guessing but leaves room for improvement. The Matthews correlation coefficient (MCC) of 0.68 reflects a

substantial level of correlation between the predicted and actual classifications, providing a robust evaluation of model performance even in cases of imbalanced datasets.

Table 3. Report on model performance metrics on test data.

Metrics	Class	
	Non-Depression	**Depression**
Precision	0.83	0.87
Recall	0.88	0.82
F1-score	0.86	0.84
Accuracy	0.85	
AUC	0.73	
MCC	0.68	

The results by class demonstrate the following metrics:

For class 0 (no depression):

- Precision: 0.93, which means that the model does a good job of correctly classifying examples without depression.
- Recall: 0.87, which means that the model detects most examples of this class.
- F1-score: 0.90, which indicates a good balance between precision and recall.

For class 1 (with depression):

- Precision: 0.73, which indicates some errors in predicting this class (possibly due to the smaller amount of data).
- Recall: 0.85, the model is good at detecting examples of depression.
- F1-score: 0.79, which still indicates relatively high efficiency.

Thus, the model demonstrates stable results with high accuracy for both classes, although the accuracy for class 1 is slightly lower, which may indicate difficulties with assessing the classification of depression due to the smaller amount of data for this class. Nevertheless, the overall performance of the model on the test data is quite high, indicating its potential in detecting depression based on multimodal information (audio and text).

5. Discussion

The experimental results of this research have demonstrated the potential of multimodal networks for depression detection. Specifically, the early fusion model has outperformed the late fusion model in terms of multiple evaluation metrics, including accuracy, precision, recall, and f1-score. The early fusion model achieved an f1-score of 0.93 on the validation set and 0.79 on the test set, compared to the late fusion model's f1-score of 0.84 on validation data.

One of the key reasons for the superior performance of the early fusion model is the immediate integration of audio and text data at the feature extraction stage. By combining these modalities early, the model can learn joint feature representations, allowing it to capture correlations between the audio and text modalities that might be missed when processing them independently. This integration leads to a more holistic understanding of the data, where both verbal (text) and non-verbal (audio) cues contribute to depression detection. In addition, an important advantage of the early fusion model is the smaller number of parameters, which reduces computational costs and training time, making it more optimal for practical implementation.

However, the early fusion approach also has certain limitations. The early fusion model demands more sophisticated handling of data preprocessing and input alignment since both modalities must be preprocessed and fed into the network simultaneously.

On the other hand, the late fusion model separates audio and text processing, allowing for more specialized feature extraction for each modality. This modular approach offers flexibility in learning modality-specific patterns but at the cost of not leveraging cross-modality correlations as effectively as early fusion. While late fusion ensures that audio and text features are treated independently, it loses valuable information that comes from integrating them earlier in the process.

The practical implementation of depression detection models, especially those that use multimodal approaches such as early and late fusion, must take into account a number of ethical and privacy issues. In practice, the implementation of these models requires robust safeguards to ensure patient confidentiality and avoid stigmatization.

When working with medical data, it is also important to pay attention to the implementation of secure data collection and storage protocols, anonymization of patient data, and compliance with regulations such as GDPR or HIPAA. Additionally, it is worth noting that the model results should serve as an advisory function and are not a full-fledged diagnosis, i.e., they need additional validation by specialists (doctors) to prevent fault or misinterpretation that may increase stigma.

Further improvements to the model can be made in several areas. First, additional methods of data augmentation can be considered. This may help to reduce the problem of limited data size. Secondly, to improve the model's performance, more sophisticated self-attention mechanisms, such as transformer or attention-based models, which can better account for the interaction between text and audio data, should be integrated. We could also consider using more advanced audio processing techniques, such as calculating additional features that would better convey the emotional state of the speaker.

6. Conclusions

This research has demonstrated the effectiveness of multimodal networks for depression detection, specifically through the development and comparison of early and late fusion models. The early fusion model, which integrates audio and text data at the feature extraction stage, outperformed the late fusion model with higher accuracy and f1-score, showcasing the advantage of early cross-modal interaction. Both models were designed using advanced neural network techniques, including CNNs, Bi-LSTM layers, and attention mechanisms, to capture relevant patterns from the audio and text modalities.

The key findings of this research highlight the importance of combining multiple modalities to improve the robustness and accuracy of depression detection. The early fusion model achieved an f1-score of 0.93 on validation data, while the late fusion model scored 0.84. On the test set, the best-performing model (early fusion) achieved an f1-score of 0.79 and an accuracy of 0.86.

The implications of this study extend beyond the immediate technical contributions:

1. Clinical Applications: The developed model provides a foundation for automated depression screening systems. By identifying depressive signs from speech and text data, it has the potential to assist clinicians in diagnosing depressive disorders early and more efficiently. This can be particularly valuable in resource-constrained settings where mental health professionals are limited.
2. Early Intervention: The system could empower individuals to monitor their psycho-emotional states, encouraging them to seek help earlier when depressive symptoms emerge. This aligns with public health goals of early intervention to mitigate the progression of mental health disorders.
3. Integration into Broader Systems: Beyond clinical use, the model can be integrated into wellness applications, virtual assistants, or workplace mental health monitoring systems, fostering a more proactive approach to mental health care.

Future work could focus on expanding the dataset size to include more diverse populations and exploring different fusion architectures. The developed model can serve as a basis for further development of automated depression screening systems that can detect signs of depressive disorders in patients based on their speech and voice, helping doctors and people themselves diagnose the disease at early stages. In addition, the system can be used as part of integrated solutions for monitoring the psycho-emotional state of people.

Author Contributions: Conceptualization, M.N. and N.M.; methodology, O.B.; software, M.N.; validation, M.N. and N.S.; formal analysis, O.B.; investigation, M.N.; resources, O.B.; data curation, N.M.; writing—original draft preparation, M.N.; writing—review and editing, O.B.; visualization, M.N.; supervision, N.S.; project administration, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Dataset available on request from https://dcapswoz.ict.usc.edu/ (accessed on 28 November 2024).

Acknowledgments: The authors would like to thank the Armed Forces of Ukraine for providing the security to perform this work. This work was only possible because of the resilience and courage of the Ukrainian Army. The third author would like to acknowledge the financial support from the British Academy for this research (RaR\100727).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Depressive Disorder (Depression). 2023. Available online: https://www.who.int/news-room/fact-sheets/detail/depression (accessed on 22 December 2024).
2. Vázquez-Romero, A.; Gallardo-Antolín, A. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy* **2020**, *22*, 688. [CrossRef] [PubMed]
3. Yin, F.; Du, J.; Xu, X.; Zhao, L. Depression detection in speech using transformer and parallel convolutional neural networks. *Electronics* **2023**, *12*, 328. [CrossRef]
4. Miao, X.; Li, Y.; Wen, M.; Liu, Y.; Julian, I.N.; Guo, H. Fusing features of speech for depression classification based on higher-order spectral analysis. *Speech Commun.* **2022**, *143*, 46–56. [CrossRef]
5. Zhao, Y.; Liang, Z.; Du, J.; Zhang, L.; Liu, C.; Zhao, L. Multi-head attention-based long short-term memory for depression detection from speech. *Front. Neurorobotics* **2021**, *15*, 684037. [CrossRef]
6. Milintsevich, K.; Sirts, K.; Dias, G. Towards automatic text-based estimation of depression through symptom prediction. *Brain Inform.* **2023**, *10*, 4. [CrossRef] [PubMed]
7. Park, J.; Moon, N. Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability* **2022**, *14*, 3569. [CrossRef]
8. Mao, K.; Zhang, W.; Wang, D.B.; Li, A.; Jiao, R.; Zhu, Y.; Wu, B.; Zheng, T.; Qian, L.; Lyu, W.; et al. Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2251–2265. [CrossRef]
9. Iyortsuun, N.K.; Kim, S.-H.; Yang, H.-J.; Kim, S.-W.; Jhon, M. Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features. *IEEE Access* **2024**, *12*, 20479–20489. [CrossRef]
10. DAIC-WOZ Database. Available online: https://dcapswoz.ict.usc.edu/ (accessed on 22 December 2024).
11. Roberts, L. Understanding the Mel Spectrogram. 2020. Available online: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53 (accessed on 22 December 2024).
12. Wikipedia Contributors. Mel-Frequency Cepstrum. 2003. Available online: https://en.wikipedia.org/w/index.php?title=Mel-frequency_cepstrum&oldid=1233509682 (accessed on 22 December 2024).
13. Yang, J.; Luo, F.-L.; Nehorai, A. Spectral contrast enhancement: Algorithms and comparisons. *Speech Commun.* **2003**, *39*, 33–46. [CrossRef]

14. Basystiuk, O.; Melnykova, N.; Rybchak, Z. Multimodal Learning Analytics: An Overview of the Data Collection Methodology. In Proceedings of the 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 19–21 October 2023. [CrossRef]
15. Jaafar, N.; Lachiri, Z. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Syst. Appl.* **2022**, *211*, 118523. [CrossRef]