Exploring Log-Mel Spectrogram Features for Depression Detection: A Comparative Study with EATD-Corpus

Ali Tabaraei
Computer Science Department
University of Milan
Milan, Italy
ali.tabaraei@studenti.unimi.it

Ntalampiras Stavros
Computer Science Department
University of Milan
Milan, Italy
stavros.ntalampiras@unimi.it

***Abstract*—Detecting depression from audio signals has gained significant interest due to its potential for non-intrusive and scalable screening. This study explores the effectiveness of Log-Mel spectrogram features extracted from the Emotional Audio-Textual Depression Corpus (EATD-Corpus) in a convolutional neural network (CNN) for automated depression detection, revealing insights and challenges that underscore the need for improved methodologies[1].**

While our approach showed promise, its overall performance fell short of expectations compared to the results reported by the authors of EATD-Corpus. Replicating the EATD-Corpus training process yielded different results compared to the ones reported in their paper, suggesting sensitivity to fold randomization in K-fold cross-validation.

Key observations include limitations in preprocessing, notably the loss of crucial information and the curse of dimensionality due to simplistic methods such as zero-padding. Data imbalance, with 132 non-depressed and 30 depressed individuals, further complicates the analysis, where the creators of EATD-Corpus utilized optimistic techniques like 3-fold cross-validation, data augmentation, and resampling to overcome this issue.

Our findings highlight the necessity for more advanced preprocessing and algorithmic frameworks tailored to the complex data structure. Future efforts should focus on refining preprocessing techniques and addressing data imbalance to fully recover the potential of Log-Mel features in identifying the depression status of individuals.

***Index Terms*—Depression Detection, Acoustic Features, Audio Classification**

I. Introduction

Depression presents a significant global health challenge, affecting millions of individuals worldwide with symptoms such as persistent low mood, loss of interest, and lack of energy [1], [2]. Despite its importance, treatment rates remain low due to barriers like cost, time, and patient reluctance to disclose their mental state during clinical assessments [3]. To address these challenges, automated depression detection systems have emerged to enable private self-assessment and encourage consultation with psychologists.

Previous studies have explored various approaches, including feature extraction methods and machine learning techniques [4]–[6], with recent advances focusing on deep learning approaches that integrate multi-modal features, such as deep convolutional neural network (CNN) [7], LSTM [8], and CNN-LSTM [9] networks. By integrating advanced feature extraction and deep learning techniques, researchers aim to develop robust systems for depression assessment that overcome the limitations of traditional diagnostic methods.

Throughout this study, we aim to use the Emotional Audio-Textual Depression Corpus (EATD-Corpus) [10] dataset to explore the possibility of detecting the depression status of the individuals using solely their acoustic features extracted from the audio recording from their interview and classify them as depressed and non-depressed groups. To achieve this, we will attempt to recreate the results reported in the original EATD-Corpus paper and compare them with our experiments. Our investigations include the utilization of a convolutional neural network (CNN) coupled with Log-Mel spectrogram features, a CNN-LSTM network employing extracted MFCC features, and a set of traditional classifiers trained on custom statistical features.

Section II contains all the details regarding the dataset, preprocessing, feature extraction methods, data augmentation and resampling, classification methods investigated, and the evaluation metrics. Section III will discuss the effectiveness of our experiments compared to those reported by EATD-Corpus. In Section IV, we will address the limitations of our study and propose potential improvements for future research. Finally, Section V summarizes our study to highlight the key findings.

II. Methods

A. Dataset

The project focuses on uncovering the depression status of individuals using EATD-Corpus [10], which contains audio and extracted transcripts of responses from depressed and non-depressed volunteers. For this study, we will only utilize the audio recordings for our analysis.

Each volunteer responded to three randomly selected interview questions and completed an SDS (Self-Rating Depression Scale) questionnaire which assesses common characteristics of depression. The raw SDS score obtained from this questionnaire is an indicator of depression severity. In Chinese individuals, the SDS score multiplied by 1.25 (i.e., raw SDS score × 1.25) suggests depression if it is greater than or equal to 53. Consequently, to form our classification problem we map the target classes to 0 (non-depressed) if their score is less than 53, and to 1 (depressed) otherwise.

The EATD-Corpus dataset comprises audio recordings derived from interviews conducted with 162 student volunteers who received counseling, particularly within the Chinese context to support research in depression detection. The training set contains data from 83 volunteers (19 depressed and 64 non-depressed), whereas the validation set contains data from 79 volunteers (11 depressed and 68 non-depressed). The total duration of response audios in the dataset amounts to approximately 2.26 hours, having a sample rate of 16KHz.

In the original EATD-Corpus paper, however, they do not respect the same data orientation through their learning process. Instead, the performance of their proposed method is further evaluated with 3-fold cross-validation. The training and validation samples in the dataset are all merged into a single dataset, and then divided into three groups where two of which are used for training and the other one for testing.

[1] The code is publicly available online at https://github.com/tabaraei/depression-detection

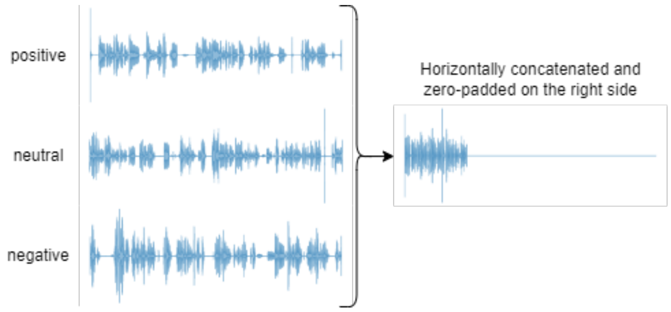

Fig. 1. Waveform Concatenation and Zero-Padding

This results in 108 training and 54 validation samples. The results reported are an average performance across the folds.

B. Preprocessing

Before delving into the classification algorithms, it is important to preprocess and cleanse the data to ensure optimal performance and reliability throughout the analysis.

EATD-Corpus also provides the preprocessed version of the audio files, where they mute audios shorter than one second, trim the silent sections at the beginning and end of each recording, and eliminate the background noise using RNNoise [11] with default parameters. We will attempt to use the same preprocessed files to ensure compatibility when comparing the algorithms.

For each sample, there are three different responses recorded in files named `positive_out.wav`, `neutral_out.wav`, and `negative_out.wav`. These recordings have different duration and vary from 2 seconds to a minute long.

To address this variability, in the original EATD-Corpus paper, the authors extracted the Mel spectrogram features from each of these responses, adopted NetVLAD [12] to generate audio embeddings of the same length from Mel spectrograms, and concatenated all these embeddings together.

In our approach, however, we separated our preprocessing step from the feature extraction phase. We first extracted and concatenated the waveforms from `positive_out.wav`, `neutral_out.wav`, and `negative_out.wav`. These concatenated waveforms were then directly used in the statistical feature extraction. However, to ensure uniform waveform lengths for other proposed feature extraction techniques, we applied zero-padding to the right side based on the maximum duration observed. This process is illustrated in Fig. 1.

C. Feature Extraction

Assuming a frame size of 512 and overlapping frames with a hop size of 256, we attempt to perform three different experiments with different feature selection methods: Log-Mel spectrogram features, Mel-frequency cepstral coefficients (MFCC), and custom statistical features.

1) Log-Mel Spectrogram Features: Using the zero-padded waveforms and the windowing characteristics mentioned earlier, Mel-spectrogram features with 80 mel frequency bands were extracted by the `librosa` [13] library from the audio waveform data for time-frequency analysis. The resulting mel-spectrogram was then converted to a logarithmic scale to enhance its representation for further analysis. The dimensions of the final feature space resulted in samples of size `(1, 80, 18551)`, having mono audio channels. A sample extracted log-mel spectrogram can be seen in Fig. 2

Fig. 2. Log-Mel Spectrogram Features Extracted from Waveforms

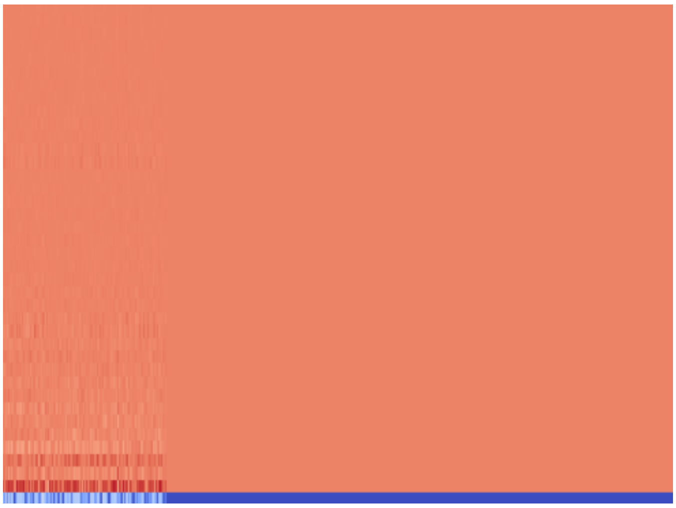

Fig. 3. MFCC Features Extracted from Waveforms

2) MFCC Features: Similar to the extraction of Log-Mel spectrogram features, the MFCC extraction involved calculating 13 coefficients over the specified frame and hop lengths (512 and 256, respectively). To capture dynamic changes in the MFCCs over time, first-order and second-order temporal derivatives were also computed and concatenated vertically on top of the MFCC features to form a comprehensive feature set for subsequent analysis. The resulting dimensions of the MFCCs formed samples of size `(1, 39, 18551)`. A sample extracted MFCC can be seen in Fig. 3

3) Statistical Features: We experimented with another set of simplified features in our settings which was robust to variation in audio lengths, not requiring the zero-padding. As discussed in [14], different acoustic features can be effective in identifying psychiatric disorders such as depression. For each frame of the given sample, using the same framing characteristics as previously described, we computed several variables: Energy (1 value per frame), Zero-Crossing Rate (1 value per frame), and the first 13 Mel-frequency cepstral coefficients (MFCCs) along with their corresponding first-order and second-order derivatives (39 values per frame in total). This process yielded a comprehensive list of features for each frame, which we subsequently subjected to statistical analyses including average, variance, minimum, and maximum calculations across all frames for each feature. This process results in acquiring samples of size `(164)`.

In order to reproduce the results of the EATD-Corpus, we will also attempt to store their log-mel features. Acoustic features extracted by the authors of EATD-Corpus have the dimensions of `(3, 256)` across all the samples.

TABLE I
SUMMARY OF CNN ARCHITECTURE UTILIZING LOG-MEL FEATURES

Layer (type)	Output Shape	Param #
Conv2d-1	[162, 6, 78, 18549]	60
BatchNorm2d-2	[162, 6, 78, 18549]	12
MaxPool2d-3	[162, 6, 39, 9274]	0
Conv2d-4	[162, 16, 37, 9272]	880
BatchNorm2d-5	[162, 16, 37, 9272]	32
MaxPool2d-6	[162, 16, 18, 4636]	0
Conv2d-7	[162, 32, 16, 4634]	4,640
BatchNorm2d-8	[162, 32, 16, 4634]	64
MaxPool2d-9	[162, 32, 8, 2317]	0
Linear-10	[162, 64]	37,961,792
Dropout-11	[162, 64]	0
Linear-12	[162, 1]	65

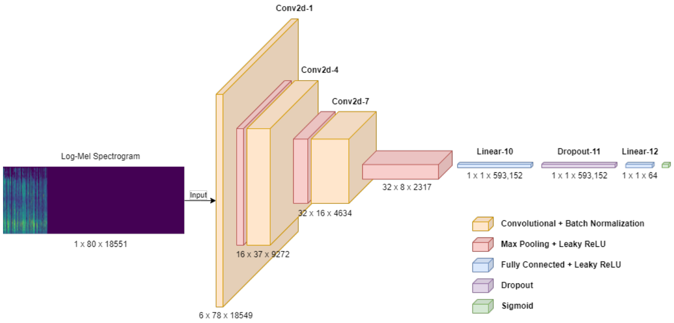

Fig. 4. Proposed CNN Architecture for Log-Mel Features

TABLE II
SUMMARY OF CNN-LSTM ARCHITECTURE UTILIZING MFCC FEATURES

Layer (type)	Output Shape	Param #
Conv2d-1	[162, 6, 37, 18549]	60
BatchNorm2d-2	[162, 6, 37, 18549]	12
MaxPool2d-3	[162, 6, 18, 9274]	–
Conv2d-4	[162, 16, 16, 9272]	880
BatchNorm2d-5	[162, 16, 16, 9272]	32
MaxPool2d-6	[162, 16, 8, 4636]	–
Conv2d-7	[162, 32, 6, 4634]	4,640
BatchNorm2d-8	[162, 32, 6, 4634]	64
MaxPool2d-9	[162, 32, 3, 2317]	–
LSTM-10	[162, 2317, 64]	74,752
Linear-11	[162, 12]	1,779,468
Dropout-12	[162, 12]	–
Linear-13	[162, 4]	52
Linear-14	[162, 1]	5

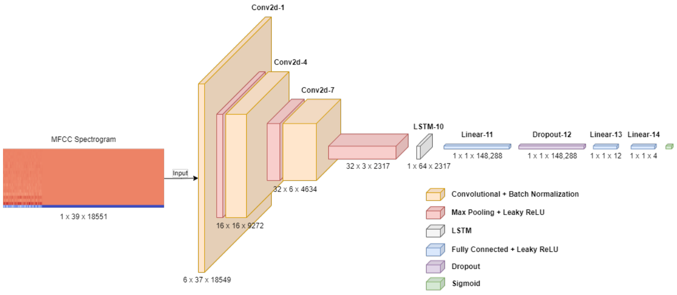

Fig. 5. Proposed CNN-LSTM Architecture for MFCC Features

D. Data Augmentation and Resampling

Depression datasets often suffer from significant data imbalance, where the number of samples in the non-depressed class dominates over the depressed class. This imbalance can lead to a bias towards the majority class during model training. To overcome this issue, the authors of EATD-Corpus used an augmentation and resampling technique to rearrange the order of the volunteers' three responses, and then resample these rearranged samples to increase the size of the depressed class and create new training samples. Since there are 6 ways of response rearrangement for each individual, the size of the depressed class can be enlarged by a factor of 6.

However, in our experiment, we do not follow their methodology to overcome the data imbalance. We believe that this approach naively introduces fake and augmented samples in the dataset, preventing us from reporting realistic evaluation measurements on the dataset.

E. Classification Methods

Three different experiments were investigated to be compared with the reproduced results from the EATD-Corpus. In order to achieve this, we first utilized the exact code reported by the EATD-Corpus GitHub repository [15] and used the same experiment setup to replicate their results. Then, we implemented our own set of classifiers with specialized feature extraction techniques, which are summarized below:

1) CNN Architecture for Log-Mel Features: We will input the log-mel features extracted from the audio files into a CNN network specifically designed for depression detection. Throughout the network, convolution layers with 6, 16, and 32 channels are used respectively, all with a kernel size of `(3x3)`. In order to perform downsampling and dimensionality reduction, max-pooling layers with kernel size of `(2x2)` and stride of 2 are employed. To speed up the training process and improve the generalization capability, batch normalization is adopted. Leaky ReLU is utilized as the activation function of these convolutional layers. Then, the resulting feature maps are flattened and fed into a fully connected (FC) layer (with Leaky ReLU activation) with only one hidden layer of 64 neurons, through which a Dropout layer is used to enhance the regularization and prevent overfitting. Finally, at the last layer, a single neuron goes through a Sigmoid activation function reflecting the likelihood of the individual experiencing depression.

In order to minimize the error and fit the training set, the Binary Cross Entropy loss function in conjunction with the Adam optimizer. The learning rate was set to `1e-6`, and the model was trained across 40 epochs using a batch size of 8.

Table I represents the proposed CNN architecture having inputs of size `(1, 80, 18551)`. Fig. 4 visualizes the network architecture for better intuition.

2) CNN-LSTM Architecture for MFCC Features: LSTM models have been shown to be effective in the domain of depression detection [8]. Using another architecture, we will attempt to train a CNN-LSTM network on the extracted MFCC features to classify the individuals into depressed and non-depressed classes. The convolutional layers preserve the same characteristics as the previous network, having the same dimensionality, pooling layer, batch normalization, and activation function. Then, an LSTM layer with 64 hidden dimensions accepts the sequences flowing from the previous layer. The outputs of the LSTM layer are then flattened and fed to a fully connected (FC) network with two hidden layers, having 12 and 4 hidden neurons respectively. The Dropout layer is also utilized, and the last single neuron represents the depression probability similar to the previous network through a Sigmoid activation function.

Similar to the previous network, we utilized the Binary Cross Entropy loss function along with the Adam optimizer to minimize error and fit the training set. The learning rate was set to `5e-6`, and training was conducted over 50 epochs with a batch size of 8.

Table II represents the CNN-LSTM architecture having inputs of size `(1, 39, 18551)`. Fig. 5 demonstrates this architecture in more detail.

3) Traditional Classifiers: Inspired by the effective classical models used in the domain reported in [16], we use the custom statistical features extracted from the audio files and adopt different traditional classifiers to be able to evaluate the effectiveness of our approach.

- K-Nearest Neighbors (KNN) classifier is used with the following hyper-parameter settings: {`algorithm:"auto", leaf_size:10, n_neighbors:5, p:1, weights:"uniform"`}
- Decision Tree classifier is used with the following hyper-parameter settings: {`criterion:"gini", max_depth:15, max_features:"log2", min_samples_leaf:1, min_samples_split:2`}
- Random Forest classifier is used with the following hyper-parameter settings: {`criterion:"gini", max_depth:5, min_samples_leaf:1, n_estimators:20`}
- AdaBoost classifier is used with the following hyper-parameter settings: {`algorithm:"SAMME.R", learning_rate:1.0, n_estimators:50`}
- Multi-Layer Perceptron (MLP) classifier is used with the following hyper-parameter settings: {`activation:"relu", alpha:0.0001, hidden_layer_sizes:[60], learning_rate:"adaptive", max_iter:1000, solver:"adam"`}

F. Evaluation Metrics

In order to provide a comparison between the different classification methods discussed, we will use the following metrics:

- Accuracy: The accuracy metric measures the proportion of correctly classified instances (both depressed and non-depressed) among the total number of instances in the dataset. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Precision: Precision measures the proportion of correctly predicted depressed instances among all predicted depressed instances. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- Recall: Recall (Sensitivity or True Positive Rate) measures the proportion of correctly predicted depressed instances among all actual depressed instances. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- F1-Measure: The F1 measure (F1 score) is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance. It is computed as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP (True Positive) is the number of correctly predicted depressed instances, TN (True Negative) represents the number of correctly predicted non-depressed instances, FP (False Positive) denotes the number of non-depressed instances incorrectly predicted as depressed, and FN (False Negative) demonstrates the number of depressed instances incorrectly predicted as non-depressed.

TABLE III
CLASSIFICATION RESULTS FOR DIFFERENT METRICS
(AVERAGED ACROSS THE 3 FOLDS)

Model	Metric			
	Accuracy	F1-Measure	Precision	Recall
GRU (EATD-Corpus)	0.5488	**0.5112**	0.5255	**0.5137**
CNN (Log-Mel)	**0.8148**	0.2137	0.4667	0.1558
CNN-LSTM (MFCC)	0.5679	0.0765	**0.7099**	0.3333
KNN (Custom)	0.8025	0.0889	0.5000	0.0606
DecisionTree (Custom)	0.7654	0.2702	0.3444	0.2417
RandomForest (Custom)	0.7963	0.0444	0.4444	0.0278
AdaBoost (Custom)	0.7346	0.2129	0.2492	0.2312
MLP (Custom)	0.7531	0.1486	0.4833	0.1558

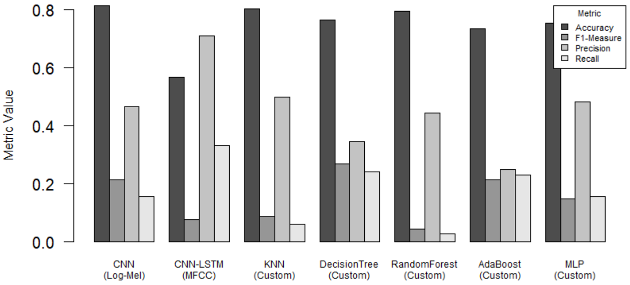

Fig. 6. Performance of Classifiers (by classification method)

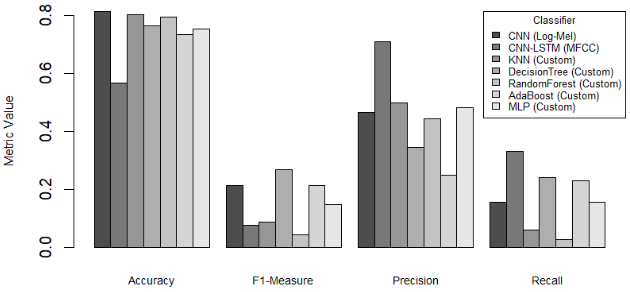

Fig. 7. Performance of Classifiers (by evaluation metric)

III. RESULTS

Now that we have different classification results obtained from CNN, CNN-LSTM, and traditional classifiers along with different choices of feature selection, we can proceed with comparing the capability of these models to detect depression against the GRU model reported by EATD-Corpus for acoustic features.

We attempted to replicate the evaluation findings of the EATD-Corpus using an identical experimental configuration. In their original study, they reported achieving an F1-Measure of `0.66`, precision of `0.57`, and recall of `0.78`. However, our replicated results yielded an F1-Measure of `0.5112`, precision of `0.5255`, and recall of `0.5137`, indicating discrepancies compared to the reported values in the paper. This sudden drop in the reproduced results suggests that the performance of their algorithm is dependent on the random split of the folds across the K-fold cross-validation.

To ensure methodological consistency, we employed a 3-fold cross-validation approach similar to theirs for evaluating the dataset. Subsequently, we aggregated the metrics across the three folds for each classifier and calculated the average. Table III illustrates the varied metrics obtained for each classifier. We proceed to visualize the results for better intuition in Fig. 6 and Fig. 7.

Analyzing the results of our experiments compared to the replicated results of EATD-Corpus, the following observations can be made:

- Considering the EATD-Corpus replicated results, it can be seen that the GRU model proposed by EATD-Corpus achieves more stable results on the Mel spectrogram features extracted from the audio files.
- The CNN network on the Log-Mel spectrogram features yields more trustable performance compared to the CNN-LSTM network on the MFCC features. While the precision and recall in the CNN-LSTM are shown to be higher, they are not trustworthy since their aggregated F1-Measure is noticeably lower than that of CNN.
- Although traditional machine learning classifiers applied to our custom-extracted statistical features in some cases may yield some encouraging results, overall, they struggle to accurately capture the characteristics required for effective depression detection in classifying individuals.
- Overall, the CNN network with Log-Mel features can be seen as a possible competitor of the EATD-Corpus. It should be reminded that the authors of EATD-Corpus utilize NetVLAD to perform dimensionality reduction and ensure fixed-length features, and they augment and resample the data to overcome data imbalance, which highly affects the evaluation results.

Since the target labels are expressed in terms of an SDS score (categorized using a threshold of 53 to differentiate between depressed and non-depressed), we also attempted to reframe the problem as a regression task and apply a threshold to the final numerical value instead. However, this approach yielded unsatisfactory results, leading us to exclude it from our report.

IV. Discussion

The findings from our analysis reveal numerous insights into the complexities and potential advancements in utilizing Log-Mel spectrogram features for depression detection. While our approach exhibited promising potential, the overall performance fell short of expectations. We highlight the following observations:

1) Limitations of preprocessing: Firstly, the preprocessing steps applied to the waveform data (Fig. 1) may have been overly simplistic, leading to the loss of important information and introducing noise into the analysis. As seen in the sample Log-Mel spectrogram (Fig. 2) or MFCC (Fig. 3) feature maps, a majority of the feature map is simply empty due to zero-padding. The loss of data continues to get worse as we go deepen through our CNN network by downsampling and applying max-pooling. Naive data preprocessing methods may not adequately capture the complexities inherent in the dataset, thereby limiting the effectiveness of subsequent classification algorithms.

2) Data Imbalance: With 132 non-depressed and 30 depressed individuals in the dataset, there is a notable data imbalance. The authors of EATD-Corpus attempted a 3-fold cross-validation instead of adhering to their constrained train/validation sets to achieve more optimistic results. They also implemented data augmentation and resampling, which may distance us from reality by altering the informativeness of the audio samples for distinguishing depression detection. Consequently, it becomes challenging to assess the true effectiveness of their model.

3) Classification results: It was noted that when replicating the training process of the EATD-Corpus using the same experimental setup, we obtained different outcomes. In their original study, they reported achieving an F1-Measure of `0.66`, precision of `0.57`, and recall of `0.78`. However, our replicated results showed an F1-Measure of `0.5112`, precision of `0.5255`, and recall of `0.5137`, which diverged from the reported values in the paper. It can be concluded that the performance of their algorithm may be influenced by the random splitting of folds during K-fold cross-validation.

V. Conclusion

In conclusion, while our study on the EATD-Corpus provides valuable insights into the potential of depression detection using acoustic features, the modest performance of our approach underscores the need for more sophisticated preprocessing techniques and algorithmic frameworks tailored to the complexities of underlying data. Future research efforts should focus on developing robust methods for data preprocessing, feature selection, and data imbalance to unlock the full potential of EATD-Corpus in identifying potentially depressed individuals.

References

[1] R. Peveler, A. Carson, and G. Rodin, "Depression in medical patients," *Bmj*, vol. 325, no. 7356, pp. 149–152, 2002.

[2] C.-S. Wu, C.-J. Kuo, C.-H. Su, S.-H. Wang, and H.-J. Dai, "Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records," *Journal of affective disorders*, vol. 260, pp. 617–623, 2020.

[3] R. C. Kessler, "The costs of depression," *Psychiatric Clinics*, vol. 35, no. 1, pp. 1–14, 2012.

[4] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, 2017, pp. 61–68.

[5] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 89–96.

[6] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, 2017, pp. 69–76.

[7] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, 2017, pp. 53–59.

[8] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[9] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.

[10] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6247–6251.

[11] J.-M. Valin, “A hybrid dsp/deep learning approach to real-time full-band speech enhancement,” in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python.” in *SciPy*, 2015, pp. 18–24.

[14] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[15] Y. Shen, H. Yang, and L. Lin, “ICASSP2022-Depression,” https://github.com/speechandlanguageprocessing/ICASSP2022-Depression, 2022.

[16] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, “Automatic depression recognition by intelligent speech signal processing: A systematic survey,” *CAAI Transactions on Intelligence Technology*, vol. 8, no. 3, pp. 701–711, 2023.