ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

Шерман Ксения Валерьевна

DIAGNOSIS OF THE SEVERITY OF MENTAL ILLNESS USING SPEECH RECORDING ANALYSIS

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ-201

Академический руководитель образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« » _____ 2018 г.

Научный руководитель

канд. технических наук, доц.
Д.И. Игнатов

_____

Москва 2024

Abstract

More than 3% of people worldwide experience depression. This diagnosis is established through interviews and clinical observations, which is a time and money demanding process. Additionally, there are a variety of symptoms associated with depression which are difficult to capture due to the limited capabilities of a human being. Many studies propose methods of automatic mental disorder recognition (MDR) using machine learning methods which are based on acoustic or linguistic feature extraction followed by a complex process of selection of the most suitable characteristics. Nevertheless, the data collecting process is difficult, thus, the solution for MDR must be able to handle limited data and avoid complicated and uninterpretable feature engineering processes. Hereby, we propose four methods based on the fine-tuned Wav2Vec-2.0 model. These approaches overcome the mentioned limitations since this transformer model is able to capture information from both acoustic and linguistic modalities and does not require a big collection of labeled data. Moreover, three of the proposed methods are novel approaches to long audio classification problems and allow us to evaluate the capabilities of acoustic transformer models to deal with long speech recordings.

Аннотация

Более 3% людей по всему миру сталкиваются с симптомами депрессии. Процесс определения этого диагноза включает в себя интервью с пациентом и медицинские обследования. Этот процесс требует много времени и денег. К тому же, существует множество симптомов, относящихся к депрессивному расстройству, которые сложно определить ввиду ограниченных способностей человека. Существуют исследования по автоматическому определению ментального состояния человека с использованием методов машинного обучения. Предлагаемые методы основаны на извлечении акустических и лингвистических характеристик речи пациента, которые после проходят нетривиальный процесс отбора признаков, наиболее подходящих для решения задачи. Тем не менее, процесс сбора данных включает в себя много сложных этапов, поэтому методы решения задачи определения ментальных расстройств должны иметь возможность обрабатывать ограниченное количество данных и избегать сложных и часто неинтерпретируемых процессов отбора признаков. Таким образом, мы предлагаем 4 метода для решения этой задачи на основе дообученной (*fine-tuned*) модели Wav2Vec-2.0. Предложенные подходы способны обойти упомянутые ограничения, так как данная модель способна улавливать информацию как из лингвистической, так и из акустической модальностей, и при этом не требует разметки большого количества данных. Более того, три метода из предложенных являются новыми подходами к

решению задачи классификации длинных аудио. Это позволило нам оценить возможности акустических моделей в обработке длинных записей речи человека.

Contents

1. Introduction

According to the World Health Organization, more than 280 million people worldwide suffer from depression (World Health Organization: WHO & World Health Organization: WHO, 2023). Moreover, depressive disorders are predicted to become one of the most common diseases by 2030 (Mathers, & Loncar, 2006). It can be described as a mental disorder which involves prolonged state of sadness, loss of pleasure from previously interesting activities etc. It affects all aspects of life and may lead to problems in communication with family members and friends, eating disorders or even suicide. Thus, it is important to diagnose depression in the early stages to prevent severe consequences of the illness. Early treatment makes it possible to recover from illness as fast as possible and with less effort. However, the process of establishing a diagnosis of a mental disorder can be difficult since it requires money, time and, most importantly, the desire of a person to receive treatment. Moreover, the reliability of a diagnosis depends on the qualification of a medical expert. Despite the professionalism of doctors, their decisions are not protected from subjectivity and bias. Besides, the limited capabilities of a human increase the possibility to miss important symptoms of the disease. Therefore, there is a need for a tool to enhance the objectivity of the diagnosis and prevent false or incomplete assessment.

Machine learning methods prove to be fast, reliable and accurate in many aspects of life: speech recognition, economic trends prediction, image generation etc. Various approaches have been proposed for the mental disorder recognition (MDR) using computer models. Many of these studies are based on either acoustic or linguistic features of patient speech. However, uni-modal data may not be sufficient for correct diagnosis since only the information obtained from both modalities will provide a complete picture of a person's mental state. In addition, most of the approaches require the excessive and usually uninterpretable extraction of hand-crafted features.

Our idea is to develop a framework which is multimodal and, at the same time, does not rely on complex feature engineering. Pre-trained transformers have demonstrated success in solving various tasks. They are better at capturing complex dependencies in data than simple machine learning models and also require less data because most of the information is learned during the pre-training. For raw speech data analysis state-of-the-art (SOTA) acoustic model Wav2Vec-2.0 was introduced (Baevski, Zhou, Mohamed & Auli, 2020). It can be used to solve

acoustic speech recognition (ASR), audio classification, speaker diarization tasks etc. Although there is a variety of studies for MDR, few studies have been done using acoustic transformer models (Khamdeeva, 2022).

In this paper we are going to propose novel approaches based on transfer learning of the Wav2Vec-2.0 model. The model has limitations on the length of an input audio, notably, it cannot process the audio of more than 40 seconds in length. Since the dataset for this work contains recordings of speech which are at least 30 seconds long, different ways of audio chunking are proposed to overcome the length problem. We define four fine-tuning approaches based on chunking from which three methods aim for sequence classification.

1. We predict label for each audio chunk and then aggregate the prediction for the whole audio using voting.
2. We obtain the embeddings for 8 chunks and concatenate them to obtain the prediction for all these chunks at once.
3. We obtain the embeddings for 8 chunks and pull over the prediction scores.
4. We obtain the embeddings for 8 chunks and feed them to LSTM model to obtain the prediction for sequence of sounds.

The attributions of this work are the following:

1. We propose a method which does not involve complex feature engineering and preprocessing. It works with raw audio wave.
2. This method is based on the most recent state-of-the-art technology in acoustic speech recognition.
3. It requires only one model to be trained since it is assumed that Wav2Vec model is multi-modal.
4. Finally, this thesis analyzes different approaches to overcome the limitations on long speech recording analysis.

The reminder of the study is organized in the following manner: section 2 provides a review on the previous research in the field of mental disease recognition. The dataset and preprocessing steps are described in section 3. Section 4 shows the proposed methods to solve MDR task. The experiments and the results are illustrated in section 5, which is followed by the discussion of the results in section 6. The last section explores future work and summarizes the findings of the study.

The text of this work was partially taken from the text of my project proposal which was written in March, 2024.

2. Literature Review

2.1. Mental disorder diagnosis

The diagnosis of mental illness involves an interview which shows how the patient speaks and formulates thoughts. Different questionnaires were proposed to measure the severity of depression and other illnesses: for schizophrenia there are SAPS or SANS scales, for mood assessment – HDRS or YMRS. The studies show that language can be considered a biomarker for depression and other disorders. Among the markers are (Tarasova & Baĭkova, 2021):

- Plenty of first-person singular pronouns (Fineberg et al, 2016);
- Frequent use of negatively and positively polarized particles;
- Emotional vocabulary filled with metaphorical expressions etc.

Moreover, the differences between healthy and depressed people can be found in speech: depressed individuals show longer response time, longer pause time and slower speech rate (Yamamoto et al., 2020).

However, it must be considered that the language markers of depression may vary in different cultures (Loveys et al, 2018). For example, Asian or Pacific Islander residents tend to avoid negative emotion vocabulary while Latino people incline to express both negative and positive feelings. In addition, differences may be encountered in the vocabulary used or in the topics used to develop a discussion.

Thus, the process of depression diagnosis is complex and non-trivial because of the diversity of the possible symptoms that should be looked for.

2.2. Uni-modal approaches

The audio wave can be defined by its characteristics such as wavelength, pitch, mel-frequency cepstral coefficients (MFCC), prosody. First approaches were based on the extraction of such characteristics and discriminating classes based on them. The main problem of this method lies in selecting best suited features from a wide-range of audio wave characteristics. For example, MFCC features show perfect results after fitting a supported vector machine (SVM) to diagnose depression (Sharma, 2020). However, glottal descriptors are also believed to perform well in mental disorder prediction compared to vocal tract and prosodic features (Moore II, Clements, Peifer & Weisser, 2007).

Another way to represent a raw speech wave is to extract a spectrogram – an image of an audio which shows the changes of amplitude and frequency over time. This representation allows to apply convolutional neural networks (CNN) which initially were proposed to deal

with image classification tasks. The combination of raw audio features and spectrograms is used to diagnose depression severity (He & Cao, 2018). In this study CNN models output vector representations of raw audio, hand-crafted features and spectrograms, and their combination is used for depression recognition. The model is trained to solve regression task instead of classification, e.g. the degree of severity is predicted as non-discrete value and the answer is the nearest integer to a predicted value.

Nevertheless, traditional machine learning methods are believed to perform better for MDR (Shalileh, Koptseva, Shishkovskaya, Khudyakova & Dragoy, 2024). The authors compared eight traditional machine learning and deep learning algorithms to predict the severity of depression. The models are trained on the extended Geneva minimalistic acoustic parameter set (Eyben et al., 2015) which contains 88 parameters. K-nearest neighbors, Random Forest, and a Multilayer perceptron show the best results. Moreover, the study claim that the CNN models do not perform well on the used data.

Some methods are based on transcripts extraction. In (Bedi et al., 2015), (Corcoran et al., 2018) authors make use of latent semantic analysis which matches a word with a vector and then trains a classifier to discriminate people with clinical high-risk for psychosis from the ones without it. In the latter article also part of speech tagging was added as an additional syntactic feature.

Nevertheless, speech recordings are not the only data used for depression recognition. Texts from social media may also contain information about the mental state of a writer (Chiong, Budhi, Dhakal, & Chiong, 2021). The study compares methods of text preprocessing like stop word removal, tokenization etc., and applies various models to obtain great results on texts from different Internet resources. Moreover, some studies analyze electrical activities of brain to define the severity of depression (Hosseinifard, Moradi, & Rostami, 2013).

Finally, many approaches are based on non-linguistic questionaries which simplify the process of data collection and preparation since the data contain only numerical or categorical answers. Since there is no defined list of questions for depression identification, feature selection methods are needed. Such methods are Select K-Best Features (SelectKBest), Minimum Redundancy and Maximum Relevance (mRMR), and Boruta which are compared for depression recognition task (Zulfiker, Kabir, Biswas, Nazneen, & Uddin, 2021). All methods are based on statistical tests and correlation analysis. SelectKBest in combination with AdaBoost classifier leads to the best result. However, no work has been done on audio data with mentioned feature selection methods.

2.3. Multimodal approaches

All mentioned approaches are based only on either acoustic or textual features. However, both modalities can affect the diagnosis. Therefore, there is a need for multimodal methods. Overall, all such approaches include extracting acoustic and textual features and then combining them to make the prediction. In (Naderi, Soleimani & Matwin, 2019) authors obtain audio and textual embeddings of an audio segment, concatenate the representations and train a long short-term memory model (LSTM) to predict a class of an audio based on its segments. Similar method is used in (Al Hanai, Ghassemi & Glass, 2018), except that, instead of audio embeddings authors extract acoustic features of a recording such as vocal pitch, Mel-frequency cepstral coefficients etc. Visual data can also be used in addition to acoustic and textual one (Yang et al., 2017). Three deep learning models for each modality are trained to generate embeddings which are next fused in a fusion network.

The multimodal approach is also investigated with the use of Wav2Vec-2.0 (Khamdeeva, 2022). The study demonstrates that the combination of BERT and Wav2Vec models is excessive because it achieves the same results as the single Wav2Vec model. Authors conclude that the reason for such a behavior may be that Wav2Vec-2.0 model already knows the semantics and syntactic information from the speech, thus a distinct textual model is unnecessary.

2.4. Wav2Vec-2.0 model

Wav2Vec-2.0 (W2V) is a transformer model which was introduced in 2020 (Baevski et al., 2020). It consists of three parts: feature extractor, context network, and quantization module. First, audio is divided into segments with a stride of 20ms in the feature extractor part, then, segments are fed into the context network which modifies vectors of segments with context information from all parts. Finally, the quantization module matches each segment representation with some quantized representation which is learned during training. The goal of training is to solve a contrastive task, which is based on prediction of a masked audio segment. The model must choose the right quantized representation among the set of distractions, e.g., other learned representations. Since the model outputs the embeddings of audio segments, which are also called the hidden states, the embedding of an audio is defined as an average vector of all segments.

The main advantage of this model is that it can capture information from both acoustic and textual modalities (Shah, Singla, Chen & Shah, 2021). The study illustrates that W2V encodes speaker information, in particular, it defines different recordings of a single person.

This fact is also shown in other studies (Fan, Li, Zhou, & Xu, 2020). In addition, it learns the semantics and syntax of the language, for example, the tense of the main clause or number of nouns in a phrase. Therefore, the main attribution of this model is that it combines both modalities while other approaches require two or more models to be trained.

3. Dataset

The dataset for the research was provided by Center for Language and Brain, HSE. Overall, the data consist of the interviews of 346 people. Of all the participants, 136 people are the patients of a psychiatric clinic (PD) and were assessed by medical experts with clinical scales (PANSS, SAPS, HDRS). Other people make up the control group (PN) and completed online questionnaires (QIDS-16, ASRM, SCL-90-R). After the assessment each person was assigned two scores from a 4-point scale which represents the severity degree of depression and thought disorder (TD): 0 – no symptoms of an illness, 1- light depression or TD, 2 – moderate depression or TD, 3 – severe depression or TD. The distribution of people by the severity scores is shown in Table 1. The people from the control group may have symptoms of some psychiatric disorder as well, they are just not in hospital.

		Thought disorder symptoms		
		0	**1**	**2**
Depression symptoms	**0**	180	17	3
	1	86	7	0
	2	34	1	1
	3	16	1	0

Table 1. The distribution of people be the severity score

Each person recorded up to three oral tasks:

1. Picture description task. A person was given a short comics written by Herluf Bidstrup: “Superman”, “Discovery of the World”, “Wonderful Day”.
2. Picture-based instruction task. A person was asked to explain the process of assembling furniture from IKEA: a chair, a table, or a bench.
3. Story task. A person was asked to tell a story from his or her life: about a memorable gift, a trip or a party.

We decided to remove the recordings of the participants with thought disorder symptoms and solve only a depression recognition task. In total there are 97128 seconds of recordings. The distribution of length of audio data is illustrated in Table 2. It is shown than data distribution across 4 classes is uneven, therefore, we propose two solutions to balance the classes. First, we decide train the model to distinguish people based on the presence of the symptoms, namely, we merged the data with depression severity tags 1, 2, and 3. Second, we augment the data to balance the classes. The augmentation is done by random sampling from the training set. The new distribution after augmentation is shown in Table 3.

		Type of elicitation task			**Total, sec**
		Picture description	**Instruction**	**Personal story**	
Depression symptoms	**0**	16075	21981	13499	51555
	1	7873	10373	6693	24939
	2	4109	5882	4624	14615
	3	1965	2346	1702	6013
Total, sec		30022	26518	40582	97128

Table 2. The number of seconds of audio data depending on the depression severity scores and the type of elicitation task

Severity score	**0**	**1**	**2**	**3**	**Total**
Length, sec	40835	29638	37192	27371	135036

Table 3. The number of second of audio data depending on the depression severity scores with augmentations

Two studies have already used this dataset for the purpose of MDR (Khamdeeva, 2022; Shalileh et al., 2024). Both articles demonstrate high results but it is important to mention that these works do not take into account the fact that speech of the same person is similar from recording to recording. Consequently, the results obtained may reflect the model's ability to match the recordings of the same person. Therefore, we decided to group all the recordings of each person and use them only in either training, validation or test set. The 80% of data was training set, 10% - validation set, and 10% - test set. All divisions are made while preserving of the percentage of samples of each class.

3.1. Audio preprocessing

The W2V model expects the input audio to have sample rate of 16kHz. Sample rate is used to transform continuous time signal to discreet one by measuring the defined number of *samples* per second. Sample is thought as a value of the speech signal at some point of a time. The more samples are measured per second, the more accurate the discreet approximation of an audio. Therefore, all the recordings were resampled to 16kHz sample rate, which means that each second consists of 16000 values. For this purpose, librosa python library is used (McFee et al, 2015).

Secondly, the speech recordings were divided into a sequence of equally sized segments. The complexity of the attention layer is $O(n^2)$, meaning that its computational requirements grow as fast as a quadratic function. Therefore, to adjust the algorithms to the computational resources which we possess, we split the recordings into 5 second segments. Each audio was divided into small parts with 2 second stride – each segment has a 3 seconds overlap with the previous one. Thereby, we capture the context of the sound on left and right sides as much as possible.

Thirdly, we pad sequences which have less than 5 seconds length because all the segments in a batch have to be the same size. The input of the Wav2Vec model calculates as following: the length of an audio in seconds multiplied by the sample rate value. Thus, most of the audio segments representations are of the size 5*16000=80000 values. However, some chunks may be shorter due to differences in the length of recordings. For example, the fragment with the length of 3 seconds has the representation of the length of 3*16000=48000. To equalize dimensions all short segments are upsampled by padding to have the length of 80000.

The data is organized using Dataset class from transformers library (Wolf et al, 2020).

4. Methods

The main goal of this study is to develop a framework for mental disorder severity prediction based on Russian speech using the advantages of transfer learning of a pre-trained Wav2Vec model. Firstly, the input of the model is raw audio which eliminates the necessity to extract and select features from a recording. Secondly, the model has already learned most of the information about the speech during the pre-training, thus, it can be fine-tuned even on limited data. Finally, we have reasons to claim that the Wav2Vec is multimodal (Shah et al., 2021), therefore, no other models are needed.

We use W2V which has been pre-trained on multilingual data (Conneau, Baevski, Collobert, Mohamed & Auli, 2020) including Russian language. Overall, 53 languages were used for the training. No fine-tuning has been done to this model. The size of the output of W2V model is [250×1024], where 250 is the time dimension, the number of segments to which the audio chunk was divided, and 1024 is the size of hidden space. Before feeding the representations of chunks into a *classification head*, we take an average through all the segments, namely, through the time dimension, obtaining a vector of the size 1024.

The model is downloaded from Hugging Face repository.

4.1. Architectures of a classifier

All the proposed models consist of two modules: Wav2Vec model to obtain sound embeddings and classification head which predicts label of an audio based on the output of acoustic model.

4.1.1. Classification head

All architectures have a fully-connected neural network as classifier except LSTM one. The model architecture alternates several modules: linear layer, dropout and activation function. Linear layer is a simple function which applies linear transformation to the data which converts a vector space from a one dimension to another. Dropout is a technique to reduce the chance of overfitting by zeroing some elements of an input vector randomly (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012). The overfitting is a plausible result due to the relatively small amount of data used for training. Finally, activation function is used to introduce a non-linearity. Since the sequence of linear layers is as complex as one full-connected layer, non-linear function should be included to provide a model the ability to learn complex dependencies in data. For the classifier ReLU function was chosen as an activation layer. It zeroes all the negative numbers in an input vector while leaving all the positive numbers as they were. The architecture of the classifier for *basic model* is as follows:

1. Dropout with 0.2 probability of an element to be zeroed.
2. Linear layer which transforms size of a vector from 1024 to 64.
3. Activation function ReLU.
4. Dropout with 0.1 probability of an element to be zeroed.
5. Final linear layer which transforms size of a vector from 256 to number of labels.

Different number of linear layers in classification head were tested for *pooled over* and *batched models*: two and three. The results of the experiments showed that classifier with three

linear layers is better at learning information about the depression symptoms. The classifier for these architectures is organized as follows:

1. Dropout with 0.2 probability of an element to be zeroed.
2. Linear layer which transforms size of a vector from 1024 to 256.
3. Activation function ReLU.
4. Linear layer which transforms size of a vector from 256 to 64.
5. Activation function ReLU.
6. Dropout with 0.1 probability of an element to be zeroed.
7. Final linear layer which transforms size of a vector from 256 to number of labels.

For the last approach we decide to apply long short-term memory (LSTM) module (Hochreiter, & Schmidhuber, 1997) as a classification head. LSTM is a recurrent neural network which is able to capture information from long sequences of information. During training this module uses information from previous outputs as inputs. Since the result of Wav2Vec model is a set of sequential sound embeddings, LSTM is assumed to be able to combine information from all the vectors in a set. Other solutions may lose information because of an averaging operation. The architecture of LSTM classifier is based on 2 bidirectional layers and is trained with 0.1 dropout.

The classification head outputs the probability for each class. Thereby, when solving binary classification task, the module returns two scores, whereas for multiclass task the output of the classifier is of the size 4: from 0 to 3. The sum of the scores for each output vector is equal to 1.

4.1.2. *Basic model*

First approach is to classify each segment of an audio and then aggregate all the results to obtain the answer. The aggregation method is based on voting. There are two voting options: hard and soft voting. For the hard voting the class of an audio is claimed to be the most frequent prediction among the chunks. For the soft voting the average score for each class is taken among all the predicted scores for an audio and then the label with the highest value becomes the answer.

The possible problem of this approach is that the model makes an assumption based on a short segment independently from the other audio fragments, which may not be representative. Second, third and fourth methods are supposed to solve this problem.

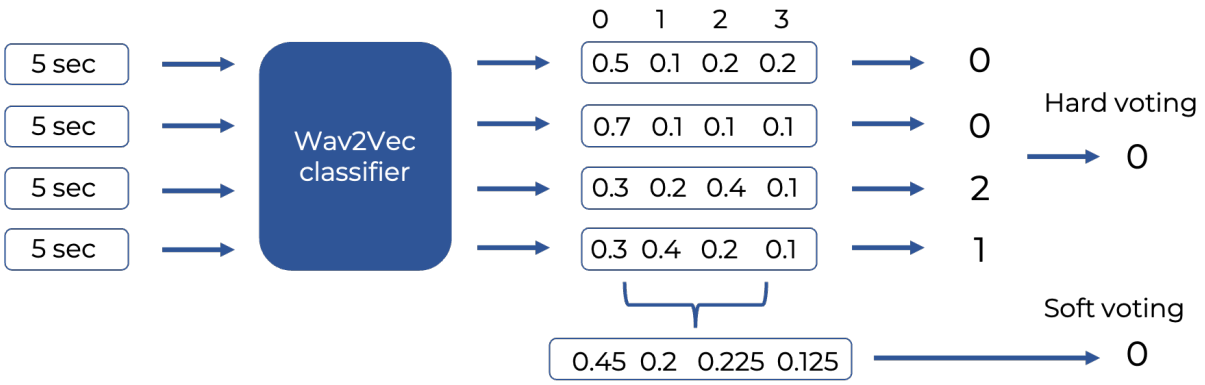


Figure 1. The process of audio label prediction using *basic model*.

Batched model

One of the solutions is to batch eight consecutive segments together, feed them to Wav2Vec model and concatenate the sets of sound embeddings to get the representation of a larger audio segment which is then used to obtain the prediction from classification head. Since the neighboring segments have overlaps, we should take an average embedding for each overlapping pair of sound embeddings, instead of a simple concatenation. For better understanding, one second of an audio consists of 50 sound embeddings in the output of Wav2Vec model. Therefore, last 150 embeddings of a first chunk and first 150 embeddings of a second chunk are the vectors of the same 3 seconds of a recording and make up the overlapping segment. Each audio segment is represented by a set of 250 sound embeddings.

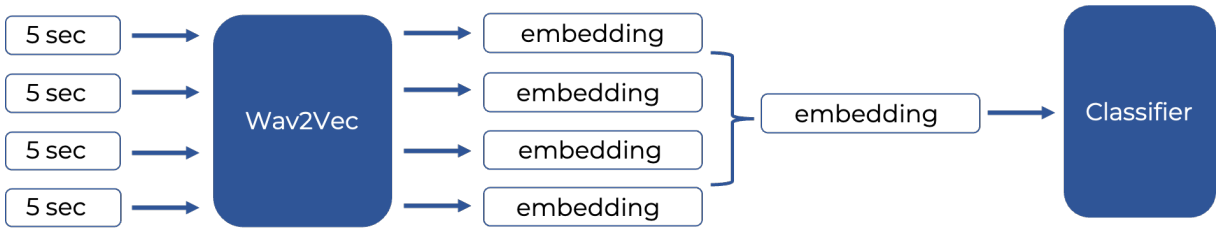


Figure 2. The process of audio label prediction using *batched model*.

The resulting set, which is obtained after the concatenation of 8 sets of embeddings, consists of 950 sound segments, which is equivalent to 19 seconds of speech. The result on the whole audio is calculated via voting as well because the most of the audio recordings are too long to fit into the memory in one batch. The batches are obtained in the following manner: the first batch consists on 8 segments, the second one consists of the last 4 segments from the previous batch and new 4 segments etc.

4.1.3. *Pooled over model*

For the third solution the chunks of an audio are batched together as for the previous architecture, and the predictions are pooled over the chunks by taking the mean prediction. This fine-tuning strategy is proposed for long text sequence classification using BERT (Stremmel et al., 2022) but it has not been implemented for long audio classification yet.

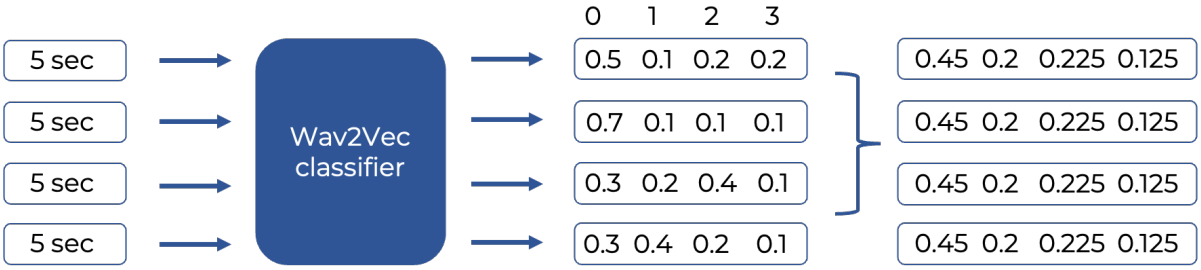


Figure 3. The process of audio label prediction using *pooled over model*.

4.1.4. *LSTM model*

For the last method W2V is used to extract the sets of sound embeddings from Wav2Vec model, preprocess the overlapped sections, as is done with *Batched model,* and then feed the new set of embeddings to LSTM classifier.

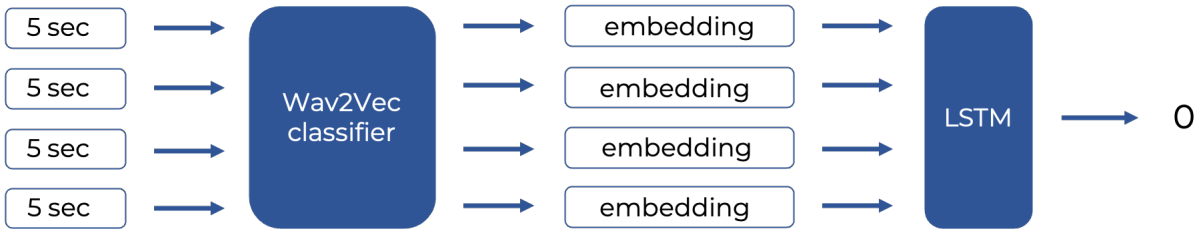


Figure 4. The process of audio label prediction using *LSTM model*.

All these methods are novel for long audio classification. The set of hyperparameters is provided in Table 4. Various learning rates, number of epochs and optimizers were tested. More precisely, we experimented with learning rates for Wav2Vec model and classification head from the set: *1e-5, 1e-4, 1e-3, 2e-3, 1e-2*, number of epochs: 2, 3 or 4 epochs, and optimizers: Adam and AdamW. The size of a batch is set to 8. The objective of the training is cross entropy: $H(p,q) = -\sum p(x) \cdot \log q(x)$, where p is true probability distribution, q is predicted probability distribution.

The validation set was used to compare the results with different hyperparameters. The obtained results are calculated on the test set. The training is done using pytorch library (Paszke

et al, 2019), version 2.1.2, and tranformers library, version 4.39.3 All the procedures were done in Kaggle with GPU P100. The training process took 6-10 hours depending on the architecture. The code used for data preprocessing and training is in GitHub repository: https://github.com/ShermanKsenia/thesis-code .

Learning rate, Wav2Vec	Learning rate, classification head	Optimizer	Warmup steps	Dropout	Number of epochs
3e-4	1e-3	AdamW: β_1=0.99, β_2=0.999, ε=1e-8	10% of total number of steps	Hidden – 0.2 Final – 0.1	2

Table 4. The hyperparameters used for training all the architectures.

4.2. Baseline models

We decide to train three baseline models on extracted features from speech recordings to compare their results with the result obtained using our approaches. As proposed in (Shalileh et al, 2024) we extracted eGeMAPS from each audio. It consists of 88 audio features, 16 of which are related to loudness level. However, some of the recordings may be noisy and, consequently, the loudness parameters may negatively affect the prediction. Therefore, these 16 features were excluded from the set. All the hyperparameters for the baseline models were selected using grid search.

First approach is k-nearest neighbors (KNN) which computes the prediction on test data based on the target values of its K nearest neighbors. The neighbors are calculated using different distance formulas. KNN is trained with K = 3, and the nearest values are calculated using the Manhattan distance metric: $d(x, y) = \sum_{i=1}^{k} |x_i - y_i|$.

Second baseline model is SVM, whose task is to find the optimal hyperplane which separates the data points in different classes. The best hyperplane is the farthest from the closest points of different classes and it strives to minimize the classification errors. The hyperplane is described by the following equation: $< w, x > -b = 0$, where w and b are weights and bias parameters, x is a feature vector. When predicting a label of some object, a sign of a resulting number is taken: $sign(< w, x > -b)$. For multiclass classification new SVM model is trained for each pair of labels and then the most frequent prediction is taken as a result.

The last method is random prediction which assigns random class from a set of values.

5. Results

5.1. Binary task

The results are illustrated in Table 5. The main metric for comparison is macro f1-score because it is more sensitive to unbalanced data. Thus, the better all the classes are predicted, the higher the macro f1-score. Soft voting was chosen for prediction calculation to address the problem of equality of votes for different classes.

Basic model, which was trained to predict a label for a small segment, show the best performance. *Batched model* accuracy is equal to *baseline models* score, but macro f1-score is higher, which means that *batched* model is better at capturing the peculiarities of speech in different classes. Other proposed methods turned out to perform worse than baseline models KNN and SVM.

The differences in results are noticeable when comparing prediction scores for each class. *Basic model* outputs 0.9 probability of predicted class, which shows that this model is more confident at its predictions, whereas *batched model*, for example, is sure in the predicted class with 0.85 probability. The *LSTM model* illustrates the lowest scores of 0.55 meaning that it is difficult for this architecture to learn necessary information for MDR.

Another observation worthy of attention is that the model may treat various recordings of one person differently. For example, for a person “PD-144” the instruction task recording shows no depression symptoms while the story and picture tasks reveal that the person may have some issues with his or her mental state. Moreover, even segments of one recording may receive distinct labels (Figure 5).

Model	Accuracy	Macro f1-score	Weighted f1-score
Baseline models			
Random	0.43	0.43	0.43
KNN	0.67	0.62	0.64
SVM	0.67	0.62	0.64
Proposed methods			
Basic model	**0.70**	**0.69**	**0.70**
Batched model	0.67	0.64	0.65
Pooled over model	0.63	0.58	0.60
LSTM model	0.59	0.58	0.59

Table 5. The results of baseline models with comparison to proposed architectures

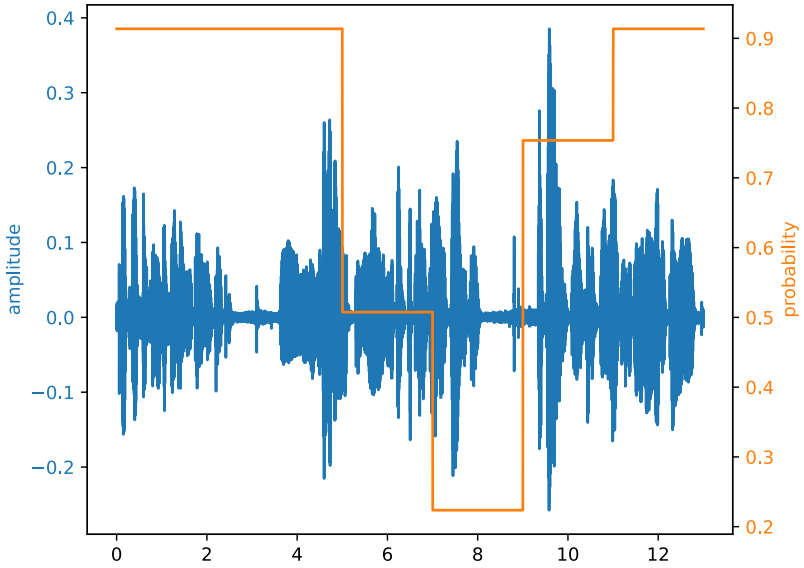


Figure 5. The segments' prediction of basic model. The blue line represents the amplitude of the sound wave, the orange line illustrates the probability that the speaker has no depression symptoms

5.2. Multiclass task

Since *the basic model* show the best performance, we decide to train it to solve multiclass task, namely, to predict the depression severity score from 0 to 3. We conducted two experiments: with initial imbalanced data and with augmented data. First experiment did not show any promising results because the model was not able to learn all the classes – it always predicted the most frequent class. The results of the second experiment are provided in Table 6 with the scores of baseline models. The section "Experiment with thresholds" in the table will be discussed in the next chapter.

Although the KNN and SVM models show better accuracy on a test set, these models fail at predicting all the classes – they predict only two classes, "no depression" and "light depression". Contrariwise, *basic model* has the highest score because it is able to learn more classes during training. The error matrix is illustrated in Figure 6 (a). The matrix shows that the model, even though it is able to learn predict all the classes, makes a great number of mistakes – half of the test set size. Moreover, it confuses people with light and severe depression with ones which do not have any symptoms.

6. Discussion

There are few possible explanations for the results on the binary classification task. One of them is that the architectures with lower scores tend to overfit because the segments are batched together and require only one label to predict. Thus, 37733 segments were used for training *basic model* and only 9584 batched samples – for training *batched, pooled over* and

Model	Accuracy	Macro f1-score	Weighted f1-score
Baseline models			
Random	0.30	0.25	0.32
KNN	**0.57**	0.23	0.46
SVM	**0.57**	0.18	0.41
Proposed methods			
Basic model	0.50	**0.34**	**0.47**
Experiments with thresholds			
Basic model + [0.3, 0.35, 0.2]	0.50	**0.37**	**0.48**
Basic model + [0.1, 0.35, 0.2]	0.37	0.32	0.33

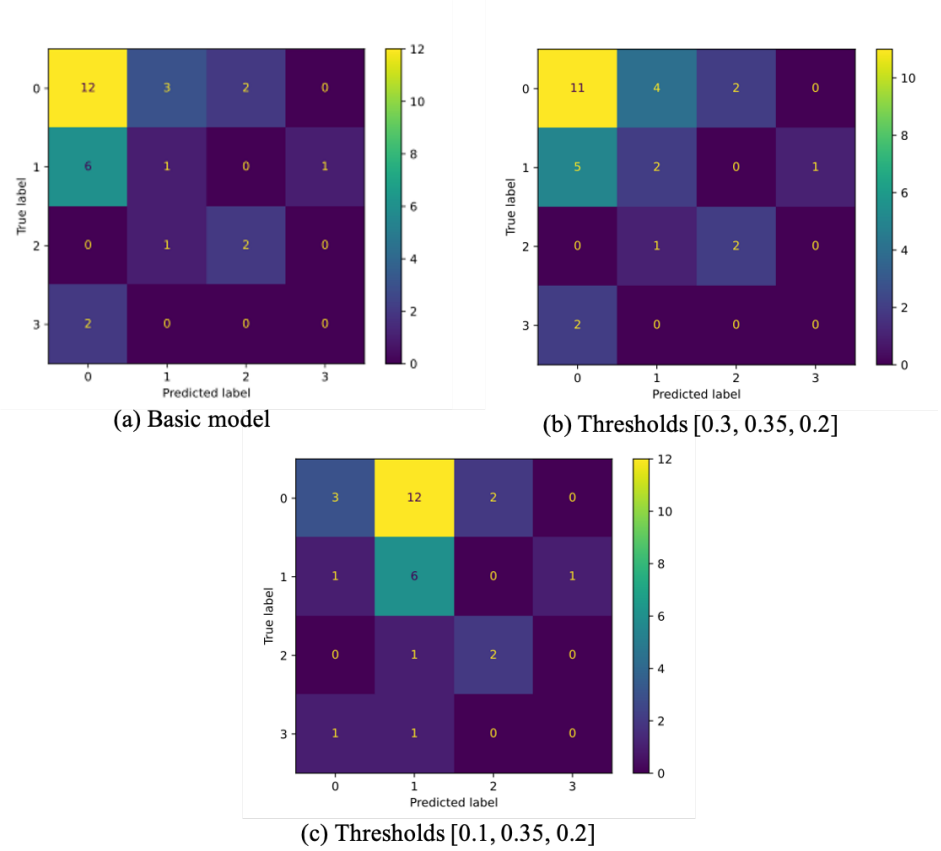
Table 6. The results of baseline models with comparison to proposed architectures on multiclass task.

LSTM models. The latter architectures may lack the data to be able to generalize the information for predicting the unseen data.

Batched model concatenates hidden states which are outputted by W2V module to get the representation of a larger audio segment making the set of sound embeddings much larger: from 250 representations to 950. After that the operation of taking the average is used which usually leads to the important information vanishing. Although, *basic model* uses the same operation, it receives less embeddings at the output of the Wav2Vec model and, therefore, the chance of the information loss is lower.

LSTM model is known as RNN model which is able to capture information from long sequences. However, it has difficulties in handling very long sequences (more than 400 steps). Such a behavior is represented in (Wen, Zhang, Luo, & Wang, 2016): the model's performance increases until the length of sequence exceeds 400 samples. Since the sequence of sound embeddings is approximately 950 embeddings, the LSTM may not have ability to process the information from all the representations due to vanishing gradients. Moreover, unlike other approaches, it is difficult to apply dropout to LSTM model which may lead to overfitting and, therefore, low results on a test set.

Nevertheless, the highest score on *basic model* confirms that small segments are enough for depression severity classification leading to the conclusion that even 5 seconds of the speech may contain an important information about the mood, mental state of a person. However, as it



(a) Basic model (b) Thresholds [0.3, 0.35, 0.2]

(c) Thresholds [0.1, 0.35, 0.2]

Figure 6. The confusion matrices of basic model with or without thresholds

is shown in Figure 5, different segments may be related to different classes and, thus, all segment are still necessary for the prediction.

For the multiclass task, during the analysis, it was revealed that the model is less confident when predicting the recordings with depression severity class 1, meaning that the model has learned some information about differences between 0 and 1 classes but this information is not enough to distinct classes. The possible explanation to the confidence problem is that the model did not have enough time or data to learn that differences.

Nonertheless, the model mostly predicts that a person does not have any depression symptoms. The zeroth class is the largest and the most diverse class since no augmentation has been added to this set of data, which suggests that this class was easier to generalize. Contrariwise, other classes are less likely to be predicted because the model have seen many of repetitive segments due to random sampling from the train set and may overfit.

Since it is more important to detect people with any depression symptoms than the ones without them, we can enhance the prediction system and make it more aware of the probabilities score on the 1-3 classes. In other words, we can set some threshold on the classes 1-3. If the predicted score is more than value of this threshold, than the person will be assigned the

corresponding depression severity score. We decide to test different thresholds to obtain better results. The appropriate thresholds were selected depending on the results on the validation set. The best values for 1, 2, 3 classes are 0.3, 0.35 and 0.2, respectively. These values increase the accuracy to 0.5, the macro f1-score to 0.37 and weighted f1-score to 0.48. The error matrix is shown in Figure 6 (b). The difference with the initial prediction is that one person from the zeroth class is mistakenly predicted to have light depression while one person from the first class is predicted right.

Another possible set of thresholds is [0.1, 0.35 and 0.2]. The accuracy, macro f1-score and weighted f1-score become: 0.37, 0.32 and 0.33. The results are presented in Figure 6 (c). The people with light depression are now predicted much better, and one person with severe depression is at least estimated to have some symtoms of depression. However, these results are obtained to the detriment of the zeroth class where most of the people are now predicted to have light depression.

7. Conclusion

The process of establishing a diagnosis of mental disorder may be difficult, time-consuming and expensive. Machine learning methods may eliminate this problem helping the medical expert to obtain a reliable and accurate diagnosis much faster. The purpose of this work was to develop a framework for depression recognition based on a person's speech without any excessive data preprocessing. Many studies on MDR rely on complex extraction of hand-crafted features of linguistic or acoustic features. We propose four methods based on the pre-trained Wav2Vec-2.0 transformer model.

The results of the study show that the *basic model*, which predicts the label of each audio segment separately, outperforms all the proposed architectures and baseline models. The *batched model* achieves slightly higher scores than the *baseline models,* while last two methods turned out to perform worse than classic machine learning methods. These results illustrate that if there is a need to work with limited audio data, training on the short segments may be enough. However, this is not entirely true because this approach does not show good results on multiclass classification, probably, due to the amount of data of different classes.

For the future work, we propose several steps to obtain better results:

1. Conduct new experiments with different architectures of the models: more layers in *classification head,* max operation instead of the operation of taking the average at the hidden states aggregation step.

2. Add different augmentation functions to enlarge the dataset: making the sound louder of quieter, adding noise, changing tone or speed etc.
3. Change the segments characteristics: make the segments longer than 5 seconds, experiment with different sizes of batches for *batched, pooled over* and *LSTM models,* change the stride value.
4. Alter the training process: solve a regression problem instead of classification, experiment with schedulers.
5. Compare with other speech transformers such as HuBERT or Whisper.
6. Concatenate meta information to the sound embeddings, such as age, sex, education etc.

All in all, transformer models show great results in solving various types of tasks. In this work we show that the approaches based on Wav2Vec transformer model have potential in solving depression recognition task on long-audio recordings but currently require more work and research to get better and more accurate results. In addition, we show that, although Wav2Vec-2.0 has already learned some information about acoustic features, the amount of data we possess may not be enough for its fine-tuning process.

References

Al Hanai, T., Ghassemi, M. M., & Glass, J. R. (2018, September). Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Interspeech* (pp. 1716-1720).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, *1*(1), 1-7.

Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135, 104499.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67-75.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2015). The Geneva minimalistic

acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2), 190-202.

Fan, Z., Li, M., Zhou, S., & Xu, B. (2020). Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv*:2012.06185.

Fineberg, S. K., Leavitt, J., Deutsch-Link, S., Dealy, S., Landry, C. D., Pirruccio, K., Shea, S., Trent, S., Cecchi, G., & Corlett, P. R. (2016). Self-reference in psychosis and depression: a language marker of illness. *Psychological medicine*, 46(12), 2605-2615.

He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, *83*, 103-111.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv*:1207.0580.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Hosseinifard, B., Moradi, M. H., & Rostami, R. (2013). Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer methods and programs in biomedicine*, 109(3), 339-345.

Khamdeeva, D. (2022). *Diagnosing neurocognitive and mental diseases by speech analyses*. [Bachelor's thesis, Higher School of Economics]. HSE website. https://www.hse.ru/en/ba/ami/students/diplomas/634002125

Loveys, K., Torrez, J., Fine, A., Moriarty, G., & Coppersmith, G. (2018, June). Cross-cultural differences in language markers of depression online. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic* (pp. 78-87).

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11), e442.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In *SciPy* (pp. 18-24).

Moore II, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2007). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, *55*(1), 96-107.

Naderi, H., Soleimani, B. H., & Matwin, S. (2019). Multimodal deep learning for mental disorders prediction from audio speech samples. *arXiv preprint arXiv:1909.01067*.

Paszke A. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shah, J., Singla, Y. K., Chen, C., & Shah, R. R. (2021). What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *arXiv preprint arXiv:2101.00387*.

Shalileh, S., Koptseva, A. O., Shishkovskaya, T. Y. I., Khudyakova, M. V., & Dragoy, O. G. V. (2024, February). An explained artificial intelligence-based solution to identify depression severity symptoms using acoustic features. In *Doklady Mathematics* (pp. 1-8). Moscow: Pleiades Publishing.

Sharma, U. (2020, March). Detection of depression from speech signal through linear svm using mfcc feature. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

Stremmel, J., Hill, B. L., Hertzberg, J., Murillo, J., Allotey, L., & Halperin, E. (2022, November). Extend and Explain: Interpreting Very Long Language Models. In *Machine Learning for Health* (pp. 218-258). PMLR.

Tarasova, L. V., & Baĭkova, O. V. (2021). Yazȳkovȳe markerȳ sostoyaniya depressii. *Vestnik Shadrinskogo gosudarstvennogo pedagogicheskogo universiteta,* (2 (50)), 278-281.

Wen, Y., Zhang, W., Luo, R., & Wang, J. (2016). Learning text representation using recurrent convolutional neural network with highway layers. *arXiv preprint arXiv*:1606.06905.

Wolf T. et al. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

World Health Organization: WHO & World Health Organization: WHO. (2023, March 31). *Depressive disorder (depression).* https://www.who.int/news-room/fact-sheets/detail/depression

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K. C., Fujita, T., Mimura, M., & Kishimoto, T. (2020). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PloS one*, *15*(9), e0238726.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017, October). Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge* (pp. 53-59).

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression. *Current research in behavioral sciences*, 2, 100044.