NATIONAL RESEARCH UNIVERSITY

AICS

Laboratory of Artificial Intelligence for Cognitive Sciences

Artificial intelligence to identify depression from audio information

Anna Kazachkova
HSE MS student
Dr. Soroosh Shalileh
Head of AICS and Research fellow at CLB

Regular Scientific Seminars of Laboratory of Artificial Intelligence for Cognitive Sciences (AICS), HSE University
29 May 2024, Moscow, Russia.

Contents

NATIONAL RESEARCH UNIVERSITY

AICS

Laboratory of Artificial Intelligence for Cognitive Sciences

Introduction

Problem

- **Depression:** a psychiatric disorder defined by feeling constantly despondent for at least two weeks, which can significantly deteriorate the quality of life. *(World Health Organization, 2023)*
- According to Word Health Organization report in the beginning of 2023, **4%** of the world population suffer from depression.
- Depressed voice is more likely to be lower, slower, hesitating, and monotonous. *(Kraepelin E., 1921)*
- Automatic voice-based diagnostics could be a reliable and affordable tool.

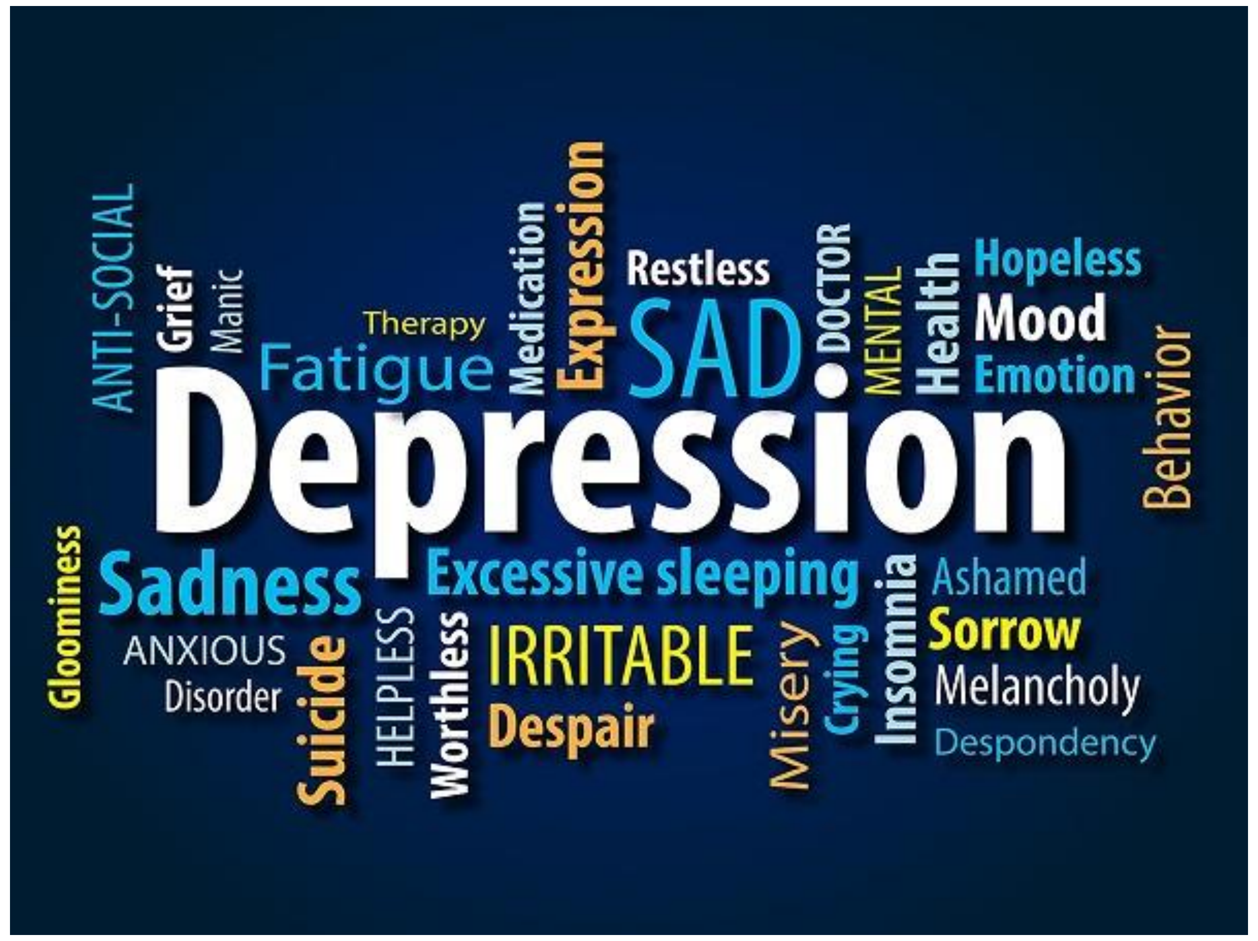


Image source

Introduction

Motivation and novelty

- **Goal:** study how accurately can be predicted on our exclusive dataset and what are the most sustainable models and data representations.
- **Novelty:** study of the established methods and new experiments on the exclusive dataset.
- This research addresses the following questions:

  Q1: The main research question is to investigate how accurately we can detect depression from audio recordings using DL models.

  Q2: Which method of extracting spectrograms and the acoustic features is more suitable for training AI models?

  Q3: Which DL model is the most effective solution to detect depression?

  Q4: Can transfer learning techniques improve the quality of the results, if yes, which of the two sub-techniques, i.e., the feature extraction or fine-tuning the weights, is more effictive?

  Q5: Is one-class classification more effective than binary classification for our main research question?

  Q6: Which depression assessment battery led to more stable and consistent results in identifying depression?

NATIONAL RESEARCH UNIVERSITY

AICS
Laboratory of Artificial Intelligence for Cognitive Sciences

Literature review

Previous methods

According to [1], two main groups of approaches to voice-based diagnostics are:

1. **ML algorithms for acoustic features**
   - Geneva Minimalistic Acoustic Parameter Set (GeMAPS) is one of the most common feature set.
   - Most common algorithms are Logistic Regression, Decision Tree, Random Forest, and others.
2. **DL algorithms for spectrograms**
   - Architectures mostly consist of CNN elements, in some cases also LSTM elements and attention mechanism are applied.
   - More advanced approaches may imply combination of some acoustic features and spectrograms.

The values of ROC-AUC achieved 0.79-0.85, although they were not often reported.

[1]: Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. Information Fusion, 80:56–86, 2022.

Literature review
Previous methods

- The majority of the works on voice-based recognition were prepared in the light of yearly **Audio/Visual Emotion Challenge** (AVEC) [2].

  The most commonly exploited datasets are:
  - Distress Analysis Interview Corpus - Wizard of Oz (**DAIC-WOZ**) [3],
  - Multi-modal Open Dataset for Mental-disorder Analysis (**MODMA**) [4].
- Other examples of data to reveal depression on:
  - Electroencephalography signal,
  - Brain imaging,
  - Facial data.

[2] Fabien Ringeval, Bj¨orn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop, pages 3–12, 2019.

[3] https://dcapswoz.ict.usc.edu/

[4] https://modma.lzu.edu.cn/data/index/

Literature review
Previous methods

Refer to:

- **[5]** for implementations of ML algorithms on acoustic features on the subset of the exploited in the current research data.
- **[1]** for review of DL methods for depression diagnostics, including audio modality, and
- **[6]** for more details on exploited ML algorithms for acoustic features.

[1] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. Information Fusion, 80:56–86, 2022.

[5] Shalileh, S., et al. "An explained artificial intelligence-based solution to identify depression severity symptoms using acoustic features."Doklady Mathematics. Vol. 108. No. Suppl 2. Moscow: Pleiades Publishing, 2023.

[6] Pingping Wu, Ruihao Wang, Han Lin, Fanlong Zhang, Juan Tu, and Miao Sun. Automatic depression recognition by intelligent speech signal processing: A systematic survey. CAAI Transactions on Intelligence Technology, 8(3):701–711, 2023.

NATIONAL RESEARCH UNIVERSITY

AICS
Laboratory of Artificial Intelligence for Cognitive Sciences

Data

Dataset overview

- An extended version of **Discourse diversity database (3D)** [7].
- Up to **3 audio recordings** for each of 346 participants aged from 16 to 82 years.

  Each audio relates to one of the incentives:

  1. **Picture-elicited narratives** (characterize one of three possible comics by Herluf Bidstrup)
  2. **Personal stories** (share one of three proposed memorable events in private life)
  3. **Picture-based instructions** (describe one of three available IKEA self-assembly furniture manuals).
- Depression symptoms of participants were assessed according to either **HDRS** or **QIDS scales**.
- People with thought disorders were excluded from the current research.

Assessment scale	All	0	1	2	3
HDRS	106	71	34	1	0
QIDS	210	109	52	33	16

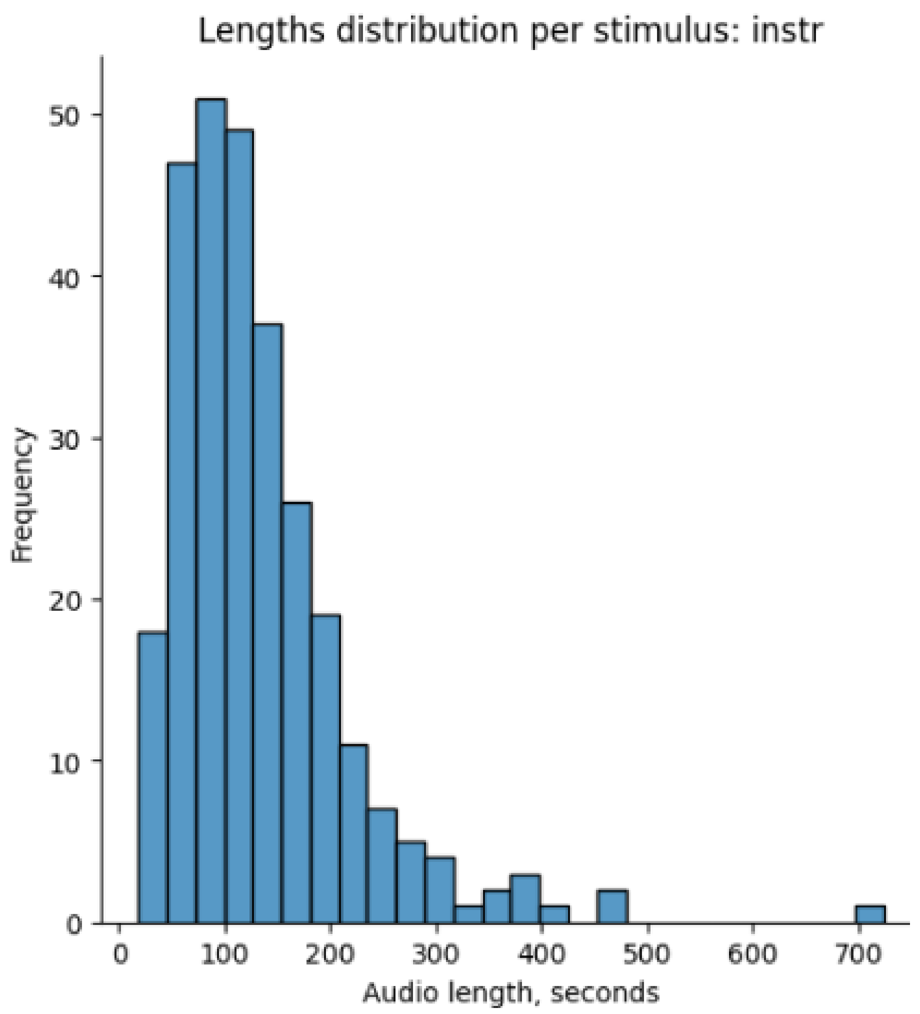
Table 1: Number of participants in terms of different assessment scales and depression symptoms severity. Rows: assessment batteries. Columns: depression symptoms severity out of 3.

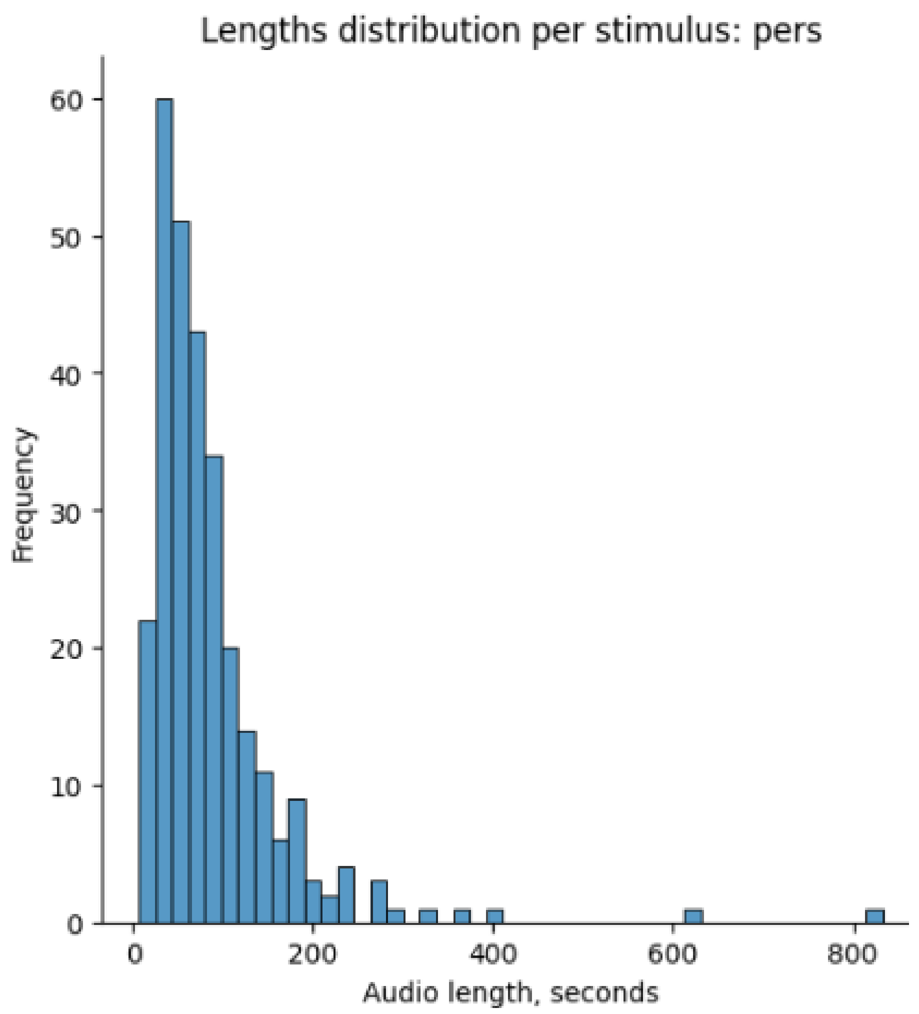
[7] Khudyakova M. et al. Discourse diversity database (3D) for clinical linguistics research: Design, development, and analysis //Bakhtiniana: Revista de Estudos do Discurso. – 2022. – T. 18. – C. 32-57.

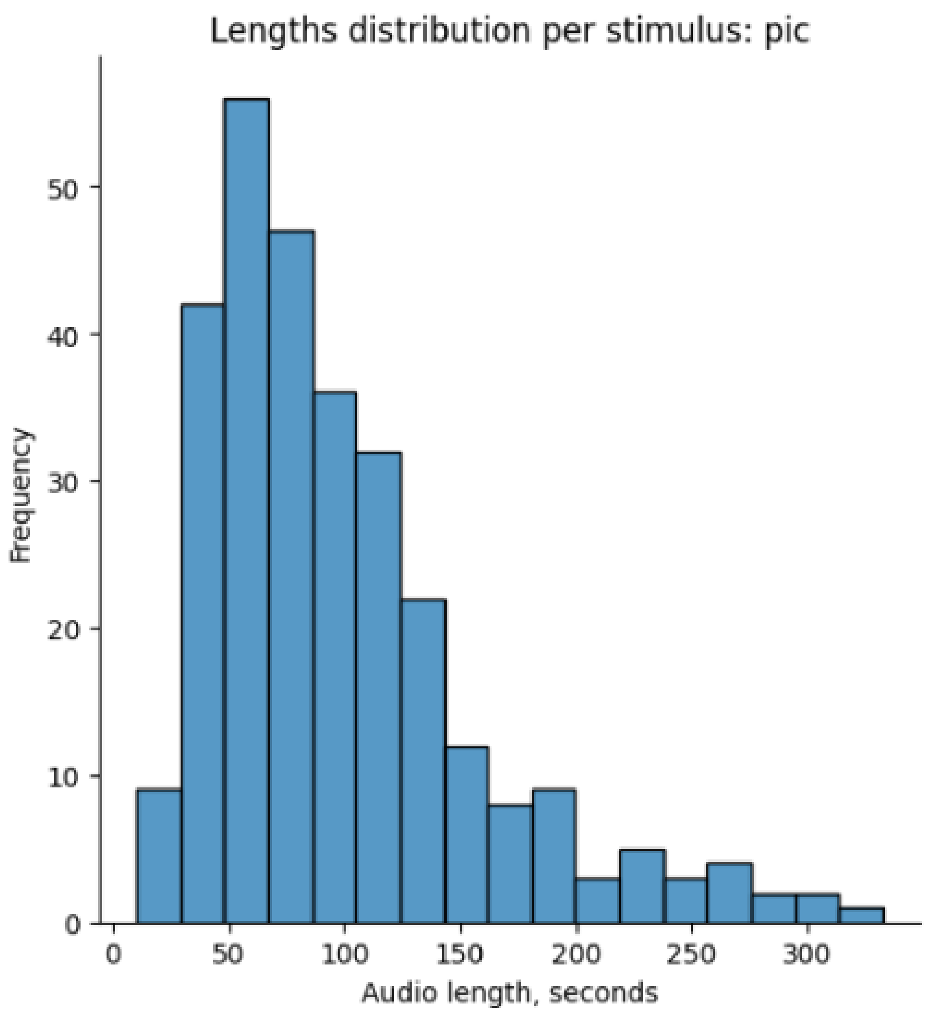
Data

Preprocessing

- **Sample rate**
  - **95%** of the files in the dataset were recorded with a sample rate of **48kHz** or **44.1kHz**. Other files were recorded with smaller rate and were not included in the research
  - All the included files were resampled to **44.1kHz**.
- **Audio lengths** were restricted by **1 minute**.







Data
Data representations

1. Acoustic features were computed based on **eGeMAPS** [8].

For instance, they include:

- Pitch
- Jitter
- Loudness
- Mel-scale Frequency Cepstral Coefficients

[8] Eyben F. et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing //IEEE transactions on affective computing. – 2015. – T. 7. – №. 2. – C. 190-202.

Data

Data representations

2. Spectrograms reflect the density of audio frequencies over time.

Steps of extracting spectrograms:

1. Application of **short-time Fourier transform** (STFT) to the audio slices of **5.8 ms** with 50% overlap,
2. Application of modulo and logarithm operations to the received embeddings,
3. Optionally, application of normalization and pseudo-coloring operations,
4. Converting embeddings to images.

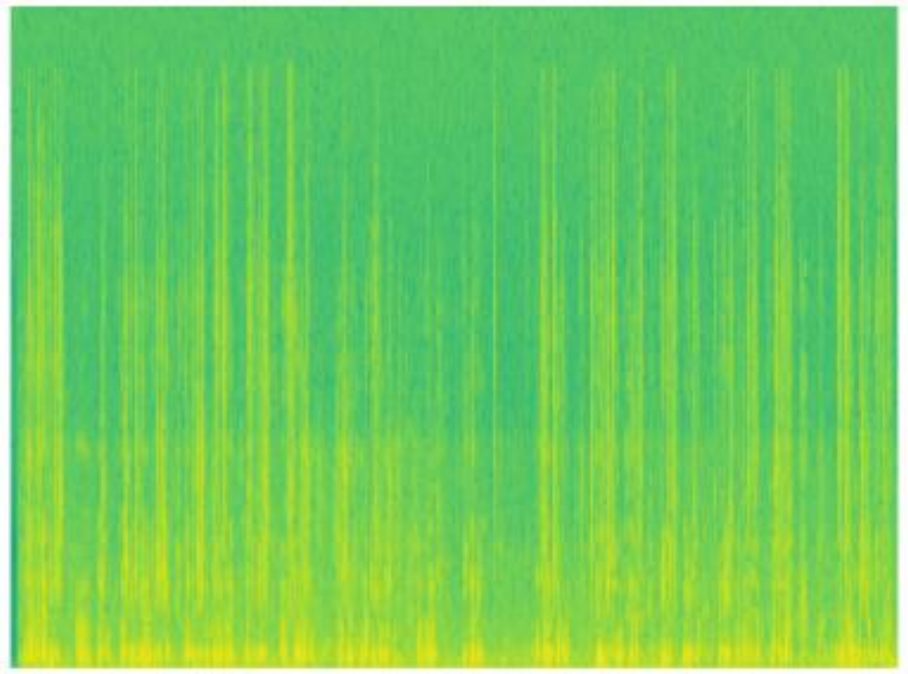


Figure 4: PD-003: default spectrogram

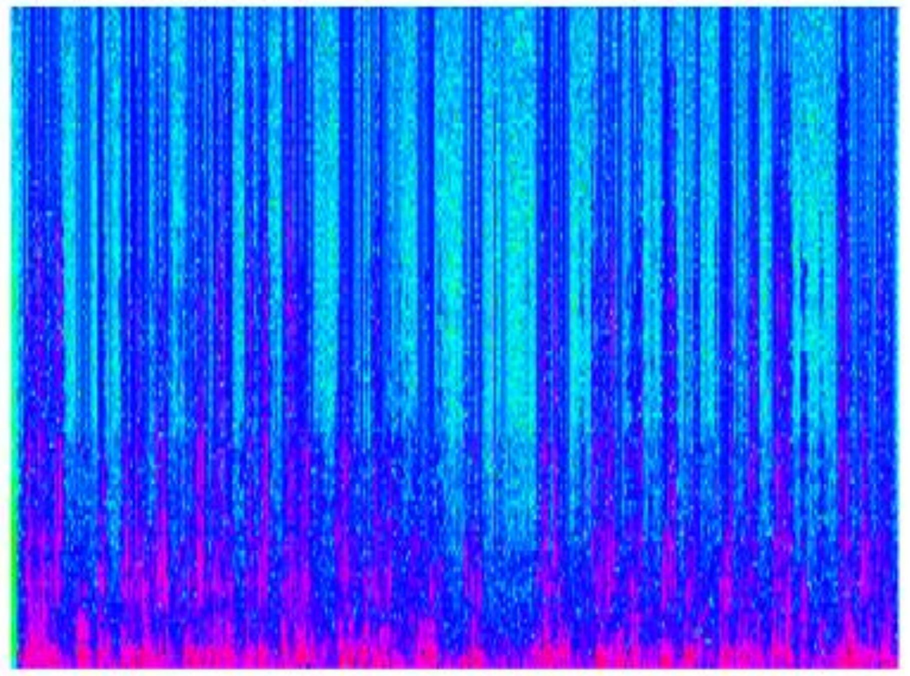


Figure 5: PD-003: HSV-based normalized spectrogram

NATIONAL RESEARCH UNIVERSITY

Methodology

Computational settings and hyperparameters tuning

AICS

Laboratory of Artificial Intelligence for Cognitive Sciences

Each model's evaluation consisted of:

1. **Hyper-parameters tuning**:
   - Bayesian optimization on ~10% of the dataset.
2. **Training and testing on stratified 10-fold cross-validation**
   - Data was split by people.
   - 5% of training data went to validation.
   - Model performance was assessed as *mean ± std* across all splits for the corresponding metric set.

Classification method	**Data representation**	**Scale**	$lr \in [1e-6, 1e-3]$	$units \in [256, 384, \ldots, 1024]$
Inception	№1	HDRS	7.2e-05	512
		QIDS	7.2e-05	512
	№2	HDRS	7.2e-05	512
		QIDS	7.2e-05	512
ResNet	№1	HDRS	0.000192	1024
		QIDS	9.3e-05	384
	№2	HDRS	9.3e-05	384
		QIDS	9.3e-05	384
InceptionResNet	№1	HDRS	0.000192	1024
		QIDS	7.2e-05	512
	№2	HDRS	9.0e-06	256
		QIDS	3.1e-05	640
ViT	№1	HDRS	2.6e-05	384
		QIDS	2.6e-05	384
	№2	HDRS	0.000782	512
		QIDS	0.000163	768

Table 3: The search domain of the hyperparameters and fine-tuned values of the CNN-based architectures and ViT fine-tuned models in predicting depression in the context of different scales and data representations.

Methodology

Problem formulation

4 approaches to formulate the given problem: (a) binary classification, (b) multi-class classification, (c) regression, and, additionally, (d) one-class classification.

- Main focus was on **binary classification** with additional experimenting with **one-class classification**.
- Multi-class classification and regression problem formulation did not lead to the acceptable results.

For one-class classification, an advanced modification was exploited, in particular, **Brute-Force one-plus-epsilon algorithm (BOPE)**. Considering x^{+}_{i} as normal data and x^{-}_{i} as abnormality, BOPE determines an optimization step as:

$$x_i^0 \sim U[\Omega]$$

$$\nabla L^{+} = -\sum_i \nabla_\theta \log f_\theta(x_i^{+}),$$

$$\nabla L^{-} = -\sum_i \nabla_\theta \log(1 - f_\theta(x_i^{-}))$$

$$\nabla L^{0} = -\sum_i \nabla_\theta \log(1 - f_\theta(x_i^{0})),$$

$$\theta \leftarrow Adam(\nabla L^{+} + \varphi \nabla L^{-} + (1-\epsilon)\cdot \nabla L^{0}),$$

- $\mathbf{x^{+}_{i}}$ – normal data
- $\mathbf{x^{-}_{i}}$ – abnormal data
- $\mathbf{\Omega}$ – bounding box of actual data
- $\mathbf{x^{0}_{i}}$ – uniformly sampled
- data
- $\boldsymbol{\varphi}$ – a ratio of abnormal and normal classes
- $\boldsymbol{\epsilon}$ – a hyper-parameter of the method, which determines the strength of regularization
- **Adam** – corresponding optimization method

NATIONAL RESEARCH UNIVERSITY

Methodology
Methods overview

AICS
Laboratory of Artificial Intelligence for Cognitive Sciences

- **Deep Learning models for spectrograms:**
  - **Convolutional neural networks (CNN)**: is a fundamental architecture in computer vision, which provides specific feature extraction from images, owing to which various spatial dependencies are considered.
    - **Basic CNN**: 3 convolutional blocks and 2 dense blocks.
    - **Deeper CNN**: 10 convolutional blocks and 3 dense blocks.
    - **ResNet**: CNN-based model, which addresses the issue of vanishing gradients by introducing the concept of residual learning.
    - **Inception:** CNN-based model, which exploits the idea of strong correlation of neighboring pixels and tries to avoid a significant reduction in the number of parameters between neighboring layers.
    - **InceptionResNet**: CNN-based model, which exploits Inception architecture adding residual learning.
  - **Vision Transformer (ViT):** implementation of the original transformer architecture to images, adding as few modifications as possible.
  - Models, pre-trained on speech data: Audio Spectrogram Transformer (AST), yet another Audio Mobilenet Network (YaMNET), and Whisper did not lead to competitive results.
- **Classical machine learning models for acoustic features:**
  - K-Neighbors, Random Forest, Gradient Boosting, and AdaBoost.

NATIONAL RESEARCH UNIVERSITY

AICS
Laboratory of Artificial Intelligence for Cognitive Sciences

Methodology

Methods overview

Transfer learning:

- For Inception, ResNet, InceptionResNet, and ViT transfer learning approach was applied.
- Two options:
  1. **Feature extraction** – training only final classifier.
  2. **Fine-tuning** – training also last several pre-trained layers.
- We used models, which had been pre-trained on the task of images classifications. Some studies demonstrated, that employing such pre-trained weights is better, than starting training with randomly initialized weights, and it may result in close accuracy as in the case of using more specific pre-trains.
- Additional experiments with removing several last pre-trained layers to use more high-level features did not improve results.
- Pre-trained on audio data instances also did not provide competitive results, which may be partially explained by the distinct preprocessing from ours.

Methodology

Classification evaluation metrics

$$precision = \frac{|D_{pos} \cap \hat{D}_{pos}|}{|\hat{D}_{pos}|}$$

$$recall = \frac{|D_{pos} \cap \hat{D}_{pos}|}{|D_{pos}|}$$

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

where D is a vector of tuples (an object, corresponding true label) and $\widehat{D}$ is a vector of tuples (an object, corresponding predicted label); they both split by positive and negative labels both in the ground truth, returning corresponding notations of D_{pos}, D_{neg}, $\widehat{D}_{pos}$, and $\widehat{D}_{neg}$.

- **ROC-AUC** evaluates how accurately a model distinguishes between positive and negative objects taking into account predicted probabilities of classes. It was the main metric of the research.
- Values of all metrics range from 0 to 1, and the higher they, the better.
- For multi-classification, metrics were calculated in one-versus-all manner and weighted to receive mean.

Experimental results

Acoustic features

Classification methods	Scale	ROC-AUC	Precision	Recall	F1-Score
Nearest Neighbor	HDRS	0.5102 ± 0.1093	0.3186 ± 0.1930	0.2816 ± 0.1740	0.2946 ± 0.1757
	QIDS	0.5465 ± 0.0734	0.5508 ± 0.0779	0.3438 ± 0.1161	0.4132 ± 0.1101
Random Forest	HDRS	**0.5488 ± 0.1217**	**0.4262 ± 0.3777**	0.1210 ± 0.1015	0.1722 ± 0.1327
	QIDS	**0.6245 ± 0.1270**	0.5908 ± 0.1594	0.5022 ± 0.1601	0.5289 ± 0.1324
Gradient Boosting	HDRS	0.5234 ± 0.1284	0.3656 ± 0.2324	**0.2154 ± 0.1589**	**0.2633 ± 0.1785**
	QIDS	0.6052 ± 0.1203	0.5459 ± 0.1154	**0.5585 ± 0.1704**	**0.5430 ± 0.1268**
AdaBoost	HDRS	0.5393 ± 0.1294	0.1400 ± 0.3273	0.0389 ± 0.0830	0.0599 ± 0.1298
	QIDS	0.5929 ± 0.1117	**0.6985 ± 0.3385**	0.2682 ± 0.1458	0.3714 ± 0.1858
MLP	HDRS	0.5249 ± 0.1120	0.1199 ± 0.1973	0.2000 ± 0.4010	0.1166 ± 0.2129
	QIDS	0.5427 ± 0.1334	0.3837 ± 0.3178	0.4231 ± 0.4745	0.3167 ± 0.2987

Table 4: Binary classification experiments with audio features

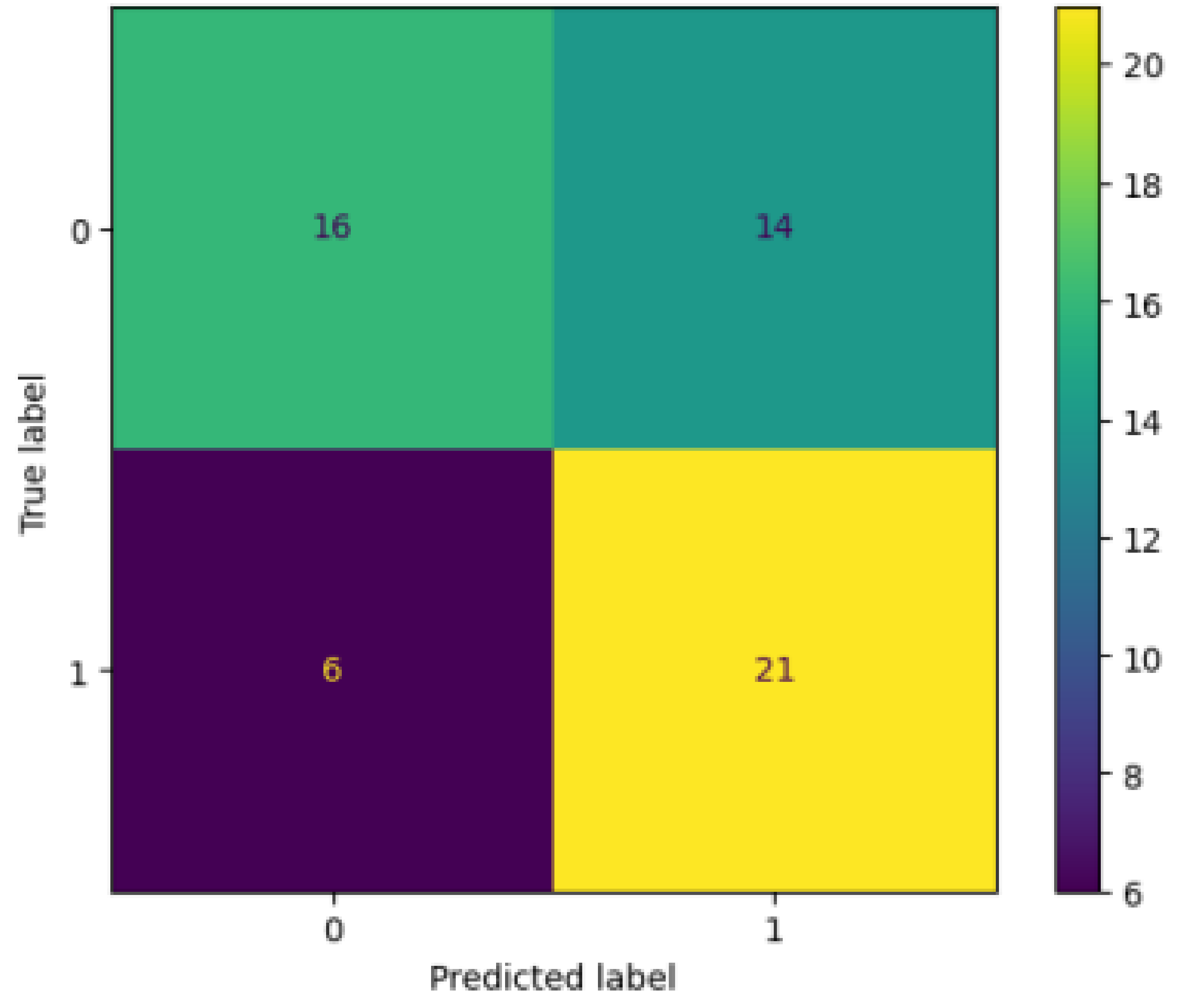


Figure 8: Confusion matrix for Random Forest (QIDS, eGeMAPS features, binary classifier) for one of the data splits, figures represent the number of audio files

Experimental results

DL models results

Classification methods	Data representation	Scale	ROC-AUC	Precision	Recall	F1-Score	Number of epochs
Random prediction		HDRS	0.5156 ± 0.1185	0.3146 ± 0.1080	0.5119 ± 0.1893	0.3880 ± 0.1341	
		QIDS	0.5378 ± 0.0686	0.4948 ± 0.1142	0.4892 ± 0.0845	0.4890 ± 0.0920	
Basic CNN	№1	HDRS	**0.6237 ± 0.1737**	**0.4667 ± 0.5018**	**0.1253 ± 0.1404**	**0.1952 ± 0.2150**	30
		QIDS	0.5173 ± 0.0935	0.1896 ± 0.2448	0.3452 ± 0.4533	0.2436 ± 0.3155	30
	№2	HDRS	0.5681 ± 0.1083	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
		QIDS	0.5032 ± 0.0573	0.3379 ± 0.2400	0.2575 ± 0.3194	0.2537 ± 0.2366	30
Deeper CNN	№1	HDRS	0.4678 ± 0.1459	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
		QIDS	0.4655 ± 0.0994	0.2985 ± 0.2735	0.5054 ± 0.5144	0.3241 ± 0.3117	30
	№2	HDRS	0.5421 ± 0.1400	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
		QIDS	0.4737 ± 0.0925	0.3810 ± 0.3120	0.5292 ± 0.4588	0.3703 ± 0.3062	30
ViT	№1	HDRS	0.6189 ± 0.1763	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
		QIDS	**0.5200 ± 0.0575**	**0.4339 ± 0.1677**	**0.6872 ± 0.3986**	**0.5075 ± 0.2420**	30
	№2	HDRS	0.5749 ± 0.1154	0.1000 ± 0.3162	0.0111 ± 0.0351	0.0200 ± 0.0632	30
		QIDS	0.4913 ± 0.1028	0.2011 ± 0.2608	0.2376 ± 0.4099	0.1901 ± 0.1901	30

Table 5: Binary classification results

- Zero values of precision, recall, and F1-score relate to the situations when probabilities are predicted in the range nearly from 0.1 to 0.4.
- No decent model compared to both acoustic features benchmark and random prediction baseline.

Experimental results

Transfer learning results

Classification methods	Data representation	Scale	ROC-AUC	Precision	Recall	F1-Score	Number of epochs
Random Prediction		HDRS	0.5156 ± 0.1185	0.3146 ± 0.1080	0.5119 ± 0.1893	0.3880 ± 0.1341	
		QIDS	0.5378 ± 0.0686	0.4948 ± 0.1142	0.4892 ± 0.0845	0.4890 ± 0.0920	
InceptionV3	№1	HDRS	0.6464 ± 0.1312	**0.5317 ± 0.1735**	**0.3553 ± 0.1945**	**0.4023 ± 0.1835**	30
		QIDS	0.5990 ± 0.1809	0.5558 ± 0.1817	0.5403 ± 0.1906	0.5387 ± 0.1668	30
	№2	HDRS	0.5692 ± 0.1372	0.2250 ± 0.1715	0.1035 ± 0.1055	0.1293 ± 0.1042	30
		QIDS	**0.6099 ± 0.0982**	0.5339 ± 0.0755	**0.6836 ± 0.1078**	**0.5969 ± 0.0798**	30
ResNet50	№1	HDRS	0.6770 ± 0.1227	0.4983 ± 0.2566	0.3109 ± 0.1995	0.3697 ± 0.1965	30
		QIDS	0.5901 ± 0.1021	0.5410 ± 0.1103	0.5570 ± 0.1594	0.5414 ± 0.1138	30
	№2	HDRS	0.4918 ± 0.1061	0.1500 ± 0.3375	0.0306 ± 0.0723	0.0462 ± 0.1038	30
		QIDS	0.6093 ± 0.0906	**0.5691 ± 0.1004**	0.5597 ± 0.1036	0.5598 ± 0.0859	30
InceptionResNet	№1	HDRS	0.6046 ± 0.1267	0.5077 ± 0.2789	0.2505 ± 0.1501	0.3198 ± 0.1691	30
		QIDS	0.4973 ± 0.0668	0.4079 ± 0.1581	0.4291 ± 0.2389	0.4077 ± 0.1896	30
	№2	HDRS	0.5314 ± 0.1096	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
		QIDS	0.5693 ± 0.0736	0.5155 ± 0.0887	0.5116 ± 0.1037	0.5072 ± 0.0716	30
ViT	№1	HDRS	**0.7050 ± 0.0965**	0.5250 ± 0.4158	0.1732 ± 0.1491	0.2544 ± 0.2120	10
		QIDS	0.5597 ± 0.1327	0.5441 ± 0.1208	0.5655 ± 0.1326	0.5424 ± 0.1002	10
	№2	HDRS	0.5235 ± 0.1011	0.2000 ± 0.2297	0.0487 ± 0.0521	0.0768 ± 0.0821	10
		QIDS	0.5787 ± 0.1197	0.5337 ± 0.0979	0.6069 ± 0.0825	0.5611 ± 0.0650	10

Table 6: Binary classification experiments with transfer learning, feature extraction sub-technique

Problem of constant prediction was almost solved. HDRS and data representation №1 were predicted more accurately.

NATIONAL RESEARCH UNIVERSITY

AICS

Laboratory of Artificial Intelligence for Cognitive Sciences

Experimental results

Transfer learning results

Classification methods	Data representation	Scale	ROC-AUC	Precision	Recall	F1-Score	Number of epochs
Random prediction		HDRS	0.5156 ± 0.1185	0.3146 ± 0.1080	0.5119 ± 0.1893	0.3880 ± 0.1341	
		QIDS	0.5378 ± 0.0686	0.4948 ± 0.1142	0.4892 ± 0.0845	0.4890 ± 0.0920	
Inception V3	№1	HDRS	0.6946 ± 0.1327	0.5202 ± 0.3110	**0.4795 ± 0.3165**	**0.4505 ± 0.2355**	30
		QIDS	0.6046 ± 0.1461	0.3264 ± 0.3066	0.3227 ± 0.3609	0.2881 ± 0.2707	30
	№2	HDRS	0.4623 ± 0.1344	0.2617 ± 0.3304	0.2167 ± 0.2786	0.1961 ± 0.2234	30
		QIDS	0.6010 ± 0.0569	**0.6280 ± 0.1572**	0.4770 ± 0.3265	0.4578 ± 0.2224	60
ResNet50	№1	HDRS	0.6388 ± 0.1443	0.3933 ± 0.3654	0.2535 ± 0.2872	0.2824 ± 0.2736	30
		QIDS	0.6174 ± 0.0978	0.5467 ± 0.1078	**0.5770 ± 0.2385**	**0.5388 ± 0.1491**	30
	№2	HDRS	0.4288 ± 0.1535	0.2956 ± 0.4003	0.1515 ± 0.2245	0.1388 ± 0.1673	30
		QIDS	0.5547 ± 0.0953	0.5328 ± 0.0741	0.5612 ± 0.1825	0.5320 ± 0.0996	30
InceptionResNet	№1	HDRS	0.5 ± 0.0	0.3292 ± 0.0348	1.0 ± 0.0	0.4944 ± 0.0	30
		QIDS	**0.6542 ± 0.0918**	0.6016 ± 0.1147	0.5695 ± 0.2936	0.5379 ± 0.1884	30
	№2	HDRS	0.4970 ± 0.0870	0.3003 ± 0.1487	0.2439 ± 0.1297	0.2630 ± 0.1302	30
		QIDS	0.5106 ± 0.0855	0.4717 ± 0.3366	0.3816 ± 0.3845	0.3254 ± 0.2464	30
ViT	№1	HDRS	**0.7082 ± 0.1115**	**0.5649 ± 0.3162**	0.3174 ± 0.1647	0.3743 ± 0.1608	10
		QIDS	0.5839 ± 0.1356	0.5180 ± 0.2350	0.5323 ± 0.2958	0.4793 ± 0.1995	10
	№2	HDRS	0.5800 ± 0.1272	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	10
		QIDS	0.5454 ± 0.0855	0.4217 ± 0.2377	0.3934 ± 0.2953	0.3856 ± 0.2231	10

Table 7: Binary classification experiments with transfer learning, fine-tuning sub-technique

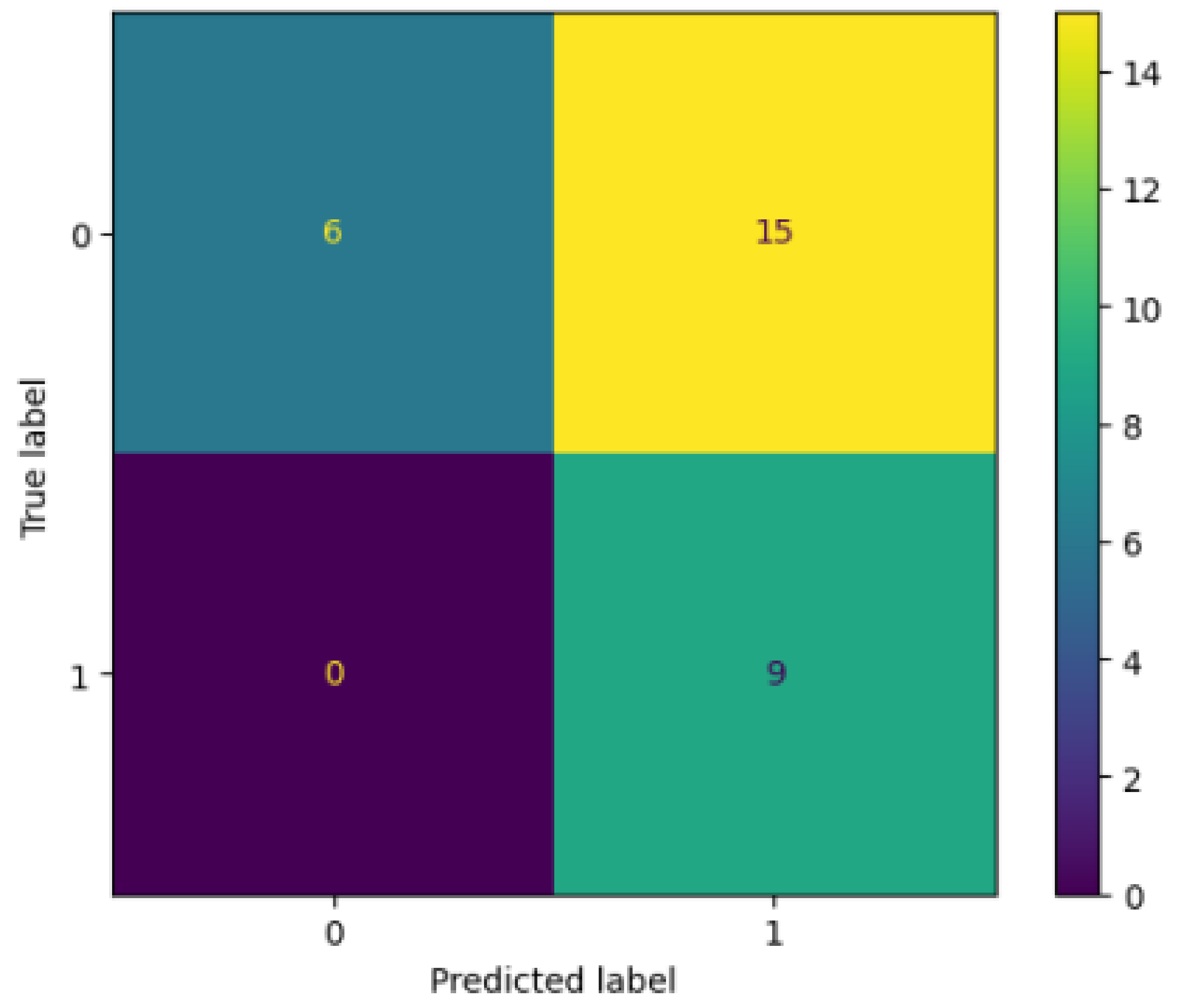


Figure 9: Confusion matrix for fine-tuned InceptionV3 (HDRS, data representation №1, binary classifier) for one of the data splits, figures represent the number of audio files

Scores are mostly better than feature extraction results. Inception and ViT provided are the most accurate models.

Experimental results

One-plus-epsilon classification results

Classification methods	One-class classification method	Data representation	Scale	ROC-AUC	Precision	Recall	F1-Score	Number of epochs
Random prediction			HDRS	0.5156 ± 0.1185	0.3146 ± 0.1080	0.5119 ± 0.1893	0.3880 ± 0.1341	
			QIDS	0.5378 0.0686	0.4948 ± 0.1142	0.4892 ± 0.0845	0.4890 ± 0.0920	
InceptionV3 (fine-tuned)	Brute-Force OPE	№1	HDRS	**0.7183 ± 0.1160**	**0.1000 ± 0.3162**	**0.0083 ± 0.0264**	**0.0154 ± 0.0487**	30
			QIDS	**0.6127 ± 0.1510**	**0.6600 ± 0.2462**	**0.2690 ± 0.1863**	**0.3624 ± 0.2126**	30
		№2	HDRS	0.5021 ± 0.1310	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
			QIDS	0.5917 ± 0.0982	0.5279 ± 0.2083	0.3841 ± 0.2510	0.4202 ± 0.2295	30
ViT (fine-tuned)	Brute-Force OPE	№1	HDRS	0.6608 ± 0.1584	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
			QIDS	0.4786 ± 0.1231	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
		№2	HDRS	0.5604 ± 0.1171	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30
			QIDS	0.5585 ± 0.1000	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	30

Table 8: One-plus-epsilon computations results

Problem formulation as the one-class classification may improve the model accuracy in some cases, but it did not demonstrate any drastic and sustainable effect.

NATIONAL RESEARCH UNIVERSITY

Experimental results

Number of epochs

AICS

Laboratory of Artificial Intelligence for Cognitive Sciences

Number of epochs was limited to avoid overfitting issues. Selective experiments with increasing the number of epochs demonstrated lower test accuracies.

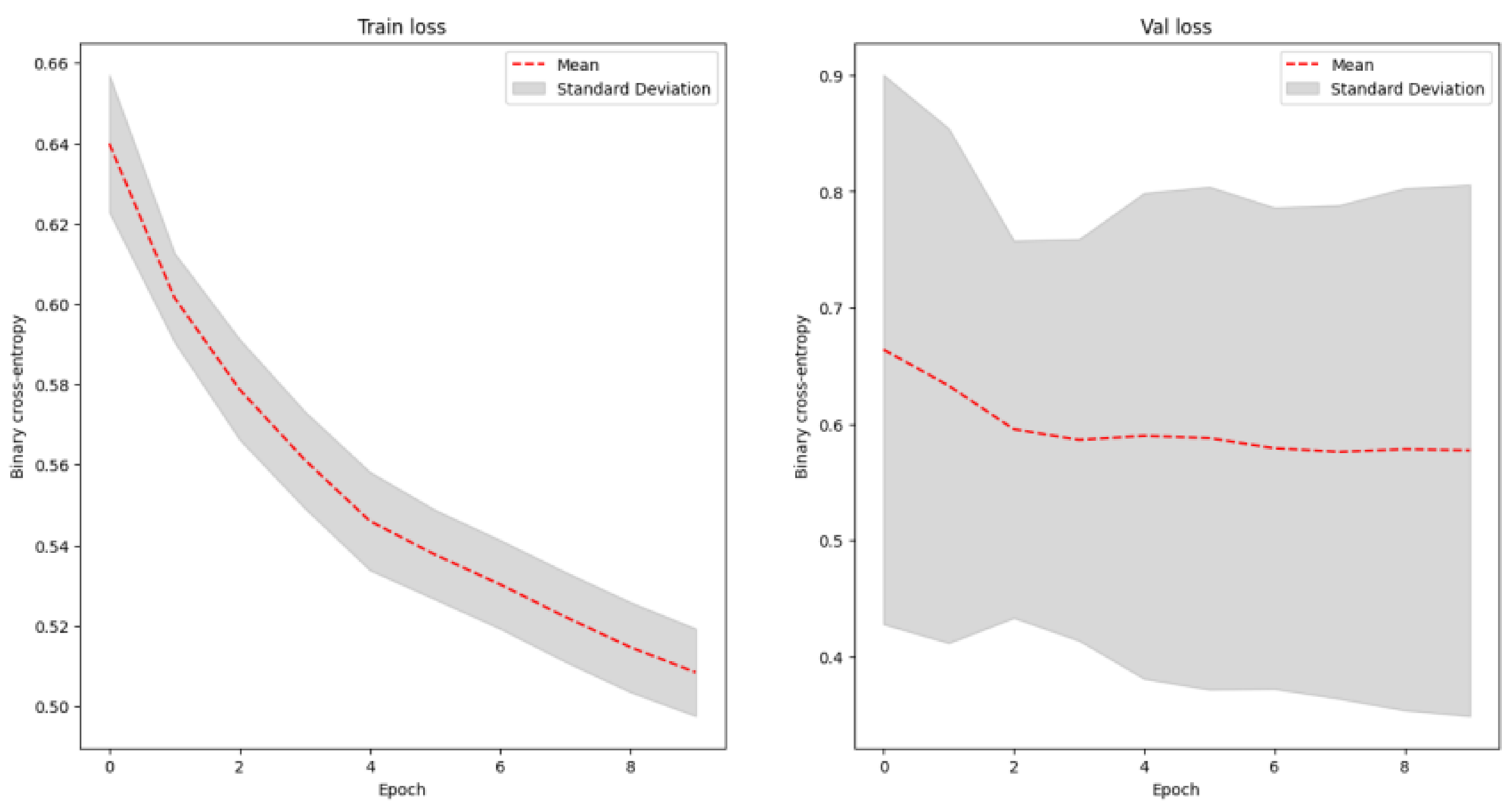


Figure 10: Losses history feature-extraction with ViT (HDRS, data representation №1, binary classifier)

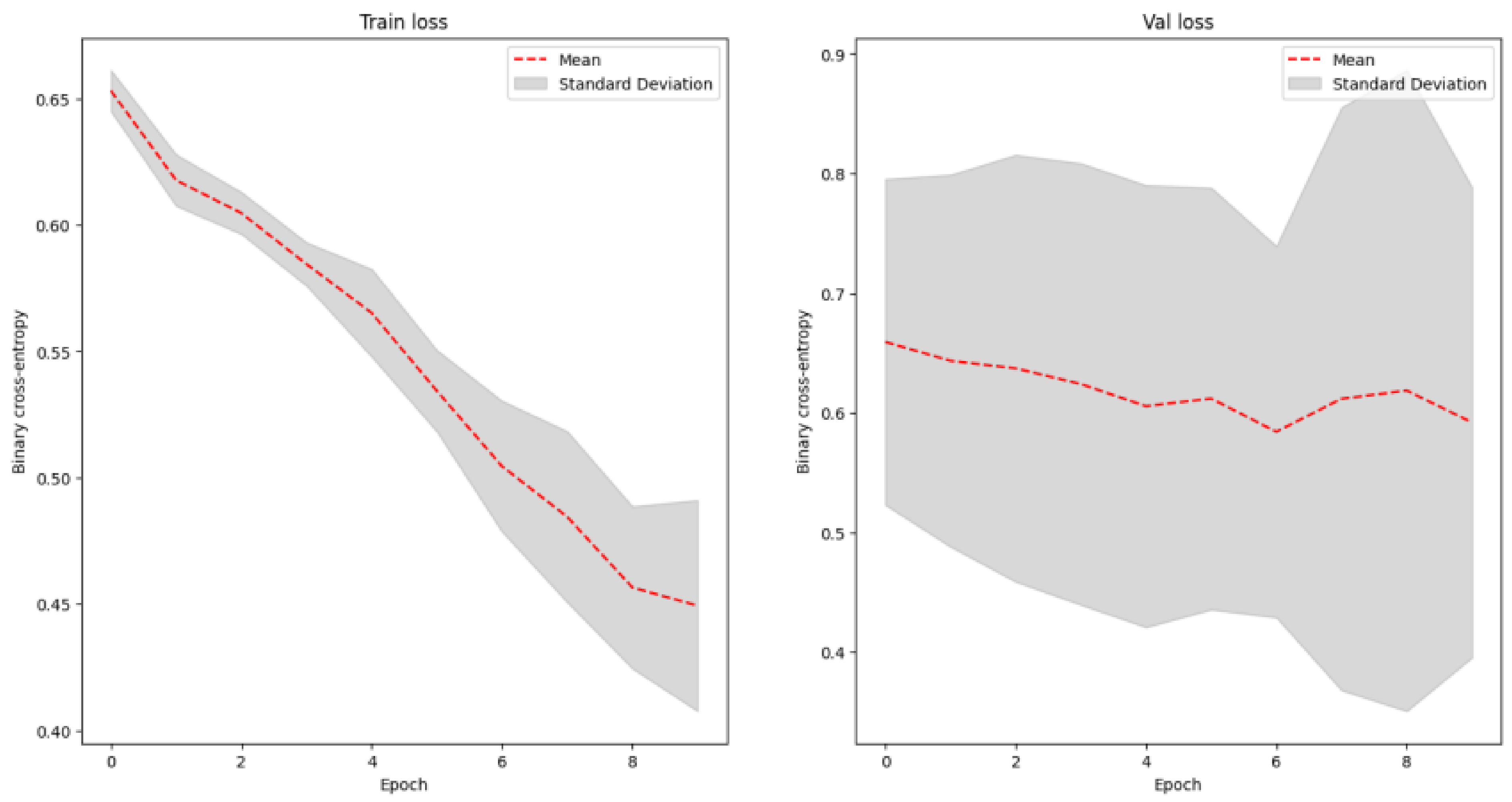


Figure 15: Losses history for fine-tuned ViT (HDRS, data representation №1, binary classifier)

NATIONAL RESEARCH UNIVERSITY

Experimental results

Key results

- Binary classifiers demonstrated relatively acceptable accuracy in terms of ROC-AUC.
- **Transfer learning** boosted the performance, especially **fine-tuning** technique.
- **ViT** and **Inception** architectures demonstrated the highest accuarcies.
- **HDRS** scale was predicted better.
- Data representation №1 outperformed both acoustic features and other spectrograms.

Classification methods	**Data represen-tation**	**Scale**	**ROC-AUC**	**Precision**	**Recall**	**F1-Score**
ViT (fine-tuned)	№1	HDRS	0.7082 ± 0.1115	0.5649 ± 0.3162	0.3174 ± 0.1647	0.3743 ± 0.1608
ViT (feature extraction)			0.7050 ± 0.0965	0.5250 ± 0.4158	0.1732 ± 0.1491	0.2544 ± 0.2120
Inception (fine-tuned)			0.6946 ± 0.1327	0.5202 ± 0.3110	0.4795 ± 0.3165	0.4505 ± 0.2355
InceptionResNet (fine-tuned)		QIDS	0.6542 ± 0.0918	0.6016 ± 0.1147	0.5695 ± 0.2936	0.5379 ± 0.1884

Table 9: Pivotal results of binary classification

Experimental results

Interpretation

- **SHAP values** for fine-tuned Inception in terms of BOPE algorithm. The higher the SHAP values, the less probability of depression. Left columns: spectrograms and ground-truth binary labels. Right columns: SHAP values and predicted values on test.

- It can be suggested, that a large spread of frequencies decreases depression probability, while a more uniform spectrum is recognized as depression.

Conclusion and future work

Relying on the conducted experiments on the 3D dataset, we answered the research questions:

Q1: The best achieved ROC-AUC for binary classification was 0.72, which is relatively acceptable. Revealing severity of depression, i.e., employing regression or multi-class classification formulations, remains unresolved.

Q2: DL methods for spectrograms outperform simpler algorithms for acoustic features; spectrograms without implementation of normalizing and pseudo-coloring operations provided higher scores.

Q3: Inception and ViT were the most promising architectures.

Q4: Transfer learning significantly boosted performance, especially fine-tuning technique.

Q5: One-class classification is also an acceptable method, however, it does not provide significant and consistent improvement.

Q6: HDRS scale is definitely better predicted.

NATIONAL RESEARCH UNIVERSITY

Conclusion and future work

AICS
Laboratory of Artificial Intelligence for Cognitive Sciences

Future work:

- Contemplate improvement of recall and F1-score (another architecture, another probability threshold);
- Experiments with other audio preprocessing (noise reduction techniques, Mel-scale);
- More extensive study of models pre-trained on speech data and architectural modification of the already employed models;
- Experiments with a combined approach of employing spectrograms and acoustic features;
- Experiments with including personal attributes (gender, age, or education).

Thank you!

Anna Kazachkova

Soroosh Shalileh

@SSHALILEH

@ANYAKAZACHKOVA

sr.shalileh@gmail.com

anya.kazachkova98@gmail.com