Computer Science

Machine Learning and High-load Systems

Moscow
2024

DIAGNOSIS OF DEPRESSION USING AUDIO DATA AND ITS DERIVATIVES

Author: Aleksandra Kovaleva
Supervisor: Soroosh Shalileh

Depression statistics

According to [1], National Institute of Health (NIH):

25,6 %

of people in Russia suffer from depression or high levels of anxiety.

18,1 %

of people in Russia experienced clinical cases in 2023.

50,2 %

of people in Russia receive proper treatment in psychological hospitals.

[1] Maksimov, S., M.B., K., Gomanova, L., Balanova, Y., Evstifeeva, S., and Drapkina, O. Mental health of the russian federation population versus regional living conditions and individual income. International Journal of Environmental Research and Public Health 20 (05 2023), 5973.

Introduction

As it was reported in [2], a depressed individual's:

- range of pitch and volume drop, so they tend to speak lower, flatter and softer.
- speech also sounds labored, with more pauses, starts and stops.
- vocal cords experience tension or relaxation, which can make speech sound strained or breathy.

Goal: by extracting acoustic features from audio files, such as tone, fluency and pitch, use this data as an input for classification of individuals into depressed and non-depressed.

Novelty: focusing solely on raw acoustic data for detecting depression.

[2] S. Scherer, G. M. Lucas, J. Gratch, A. Skip Rizzo and L. -P. Morency, "Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews," in IEEE Transactions on Affective Computing, vol. 7, no. 1, pp. 59-73, 1 Jan.-March 2016, doi: 10.1109/TAFFC.2015.2440264.

Questions

- Which of the seven AI methods under consideration will perform most efficiently in classifying depressed and non-depressed classes?
- Which of the two depression assessment scales is more effective, by leading to more accurate classification predictions?
- Which of the three elicitation tasks (stimuli) is more effective for classifying patients?

Literature review

According to [3], the latest and most comprehensive review on the applications of AI to identify depression:

1. K-nearest neighbor, multi-layer perceptron, and gradient boosting, in descending order, are the three most commonly applied AI classification algorithms.
2. Majority of the previous research obtained approximately, in the best case, scenario, 77-84% accuracy (ROC AUC)
3. Considering the number of unique data sets: Distress Analysis Interview Corpus/ Wizard-of-Oz set (DAIC-WOZ) is the top frequently used open dataset.

[3] Mamidisetti, S., and Reddy, A. M. A stacking-based ensemble framework for automatic de- pression detection using audio signals. International Journal of Advanced Computer Science and Applications 14, 7 (2023).

Data set

Our data set:
200 control group
146 participants with depression symptoms

Assessment techniques:
45 assessed by Hamilton Depression Rating Scale (HDRS)
141 assessed by Quick Inventory of Depressive Symptomatology (QIDS)

Assessment criteria: raw scores re-scaled between 0 and 3, where 0 represents no symptoms of depression, i.e., the control group, and 3 represents the existence of severe depression symptoms

	Depression symptoms							
Depression scale	Control group (0)	1	2	3	Control group (0)	1	2	3
HDRS	91	41	3	1	26%	12%	1%	0%
QIDS	109	52	33	16	32%	15%	10%	5%
Total	200	93	36	17	58%	27%	11%	5%

Data set

QIDS scoring criteria	Score	HDRS scoring criteria	Score
Normal	0-7	Normal	0-7
Mild	8-12	Mild	8-16
Moderate	13-16	Moderate	17-23
Moderate to severe	17-20		
Severe	21+	Severe	24+

	Depression symptoms			
Depression scale	Control group (0)	1	Control group (0)	1
HDRS	91	45	26%	13%
QIDS	109	101	32%	29%
Total	200	146	58%	42%

Data set

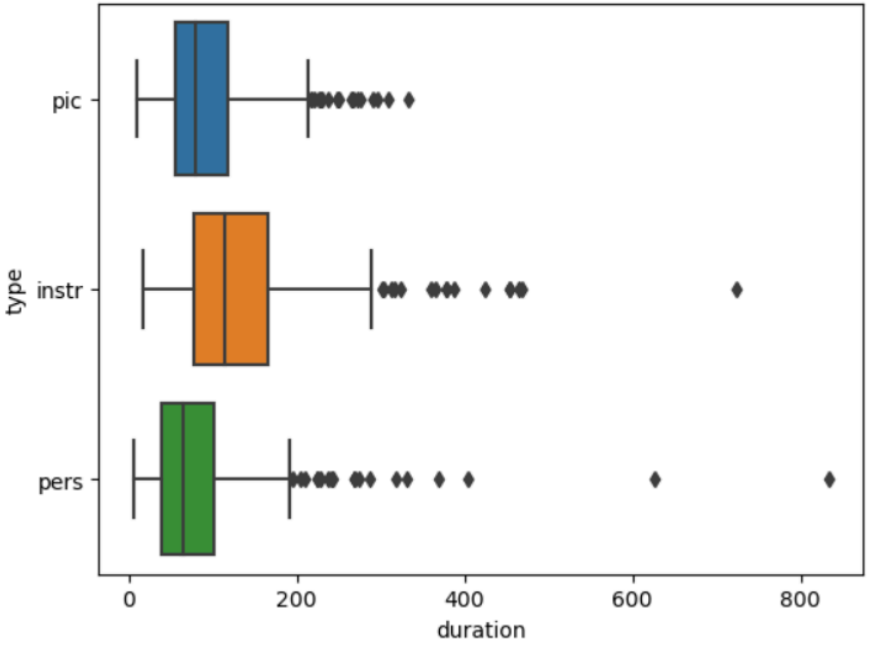
Elicitation tasks for acoustic data collection:

(a) picture-elicited narratives (PICS)

(b) personal stories (PERS)

(c) picture-based instructions, with IKEA's self-assembly furniture manuals for picture-elicited instructions (INSTR)

Stimulus	Min	Mean	Max
INSTR	18	149	724
PERS	6	126	833
PIC	10	117	332



Data preprocessing

Cutting audio files to 60 seconds and resampling to 48 000 kHz

↓

Extracting 88 acoustic features from each audio with with openSMILE library using eGeMAPS

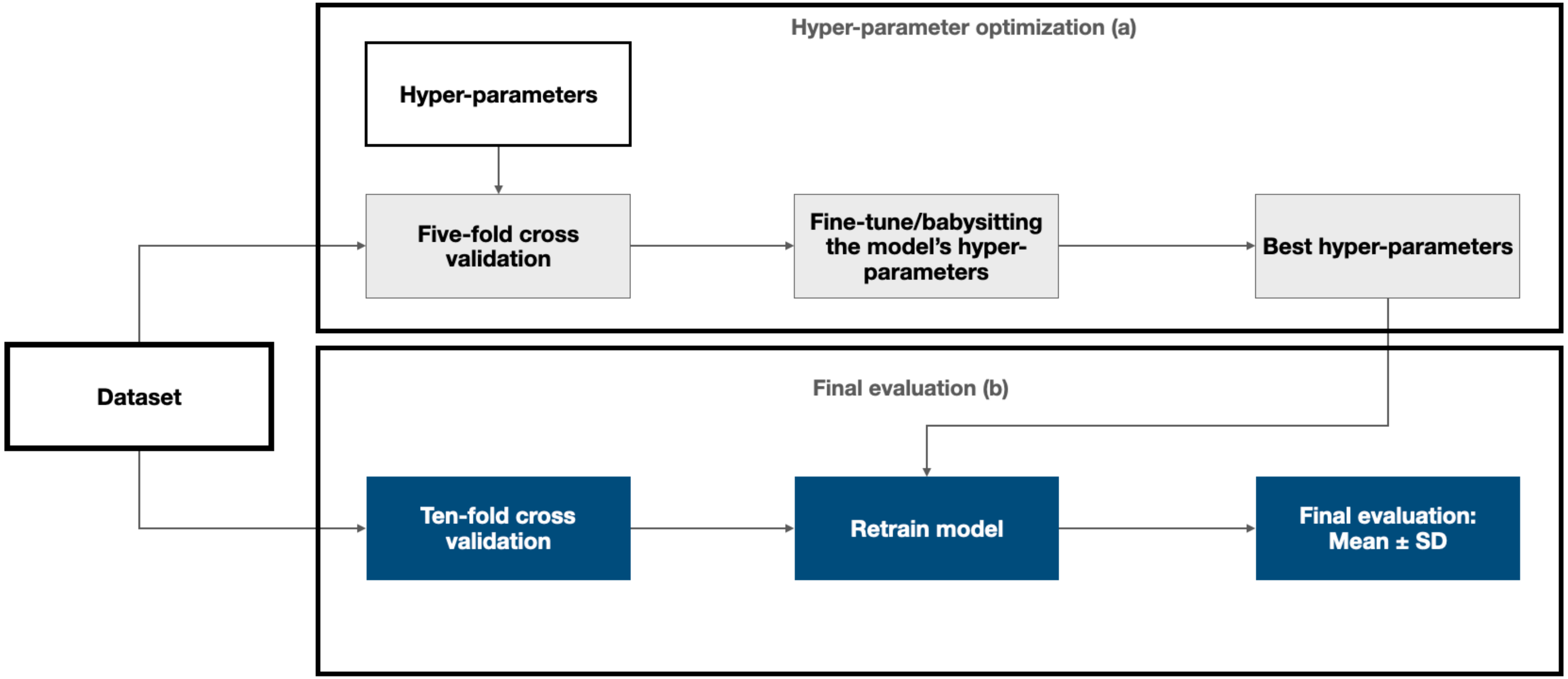
↓

Min-max scaling acoustic features

↓

Merging labels (binary output)

Computational settings



Evaluation metrics

- Precision: $\frac{TP}{TP + FP}$
- Recall: $\frac{TP}{TP + FN}$
- F1-Score: $\frac{2 \times Precision \times Recall}{(Precision + Recall)}$
- ROC-AUC score: $\int_0^1 TPR(FPR)\, dFPR$, where $TPR = \frac{TP}{TP + FN}$ $FPR = \frac{FP}{FP + TN}$

where true positives (TP) are all truly predicted depressed individuals, false positives (FP) are non-depressed individuals that algorithm predicts as depressed, and false negatives (FN) are depressed patients that algorithm attributes to control group.

Methods overview

Machine learning methods:

- Logistic regression: is a machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome
- Random forest: is an ensemble learning method that builds multiple decision trees for classification or regression tasks, and outputs the most common class or the average prediction
- Gradient boosting: is a machine learning method that incrementally improves its predictions by correcting its own mistakes in a step-by-step manner, enhancing the accuracy of the model as it progresses
- K-nearest neighbor: is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point

Methods overview

Deep learning methods:

- Multi-layer perceptron: a type of feedforward neural network consisting of fully connected neurons with a nonlinear kind of activation function

- Attentive Interpretable Tabular Learning (TabNet): a deep tabular data learning architecture that uses sequential attention to choose which features to reason from at each decision step.

- Wide and Deep Learning architecture (W&DL): an architecture that jointly trains wide linear models and deep neural networks to combine the benefits of memorization and generalization

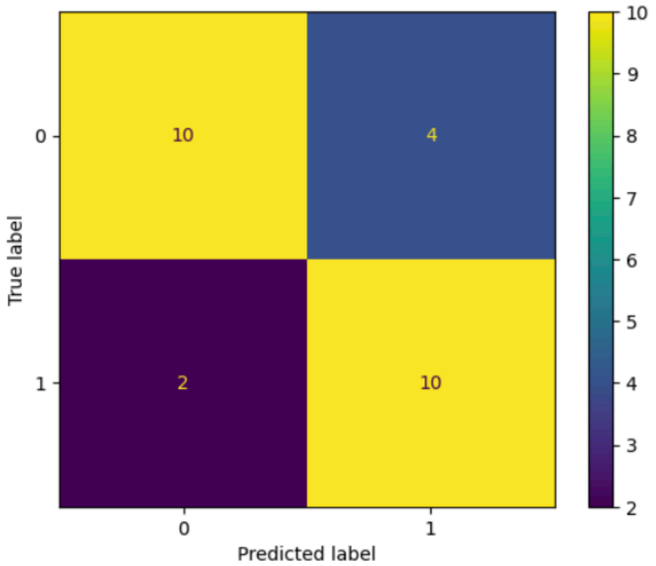
All data (no depression scale split)

Classifcation	Metrics			
	Precision	Recall	F1-Score	ROC AUC
Random prediction	0.51 ± 0.01	0.50 ± 0.02	0.50 ± 0.02	0.50 ± 0.02
Logistic regression	0.60 ± 0.07	0.60 ± 0.07	**0.60 ± 0.06**	**0.59 ± 0.06**
Random forest	0.54 ± 0.09	0.54 ± 0.07	0.52 ± 0.07	0.52 ± 0.07
Gradient Boosting	**0.62 ± 0.07**	**0.61 ± 0.04**	0.55 ± 0.05	0.57 ± 0.04
K-Nearest Neighbor	0.49 ± 0.07	0.51 ± 0.06	0.49 ± 0.07	0.49 ± 0.07
MLP	0.53 ± 0.06	0.52 ± 0.06	0.52 ± 0.06	0.52 ± 0.06
TabNet	0.57 ± 0.08	0.57 ± 0.03	0.48 ± 0.05	0.52 ± 0.03
Wide and Deep Learning	0.53 ± 0.08	0.52 ± 0.08	0.52 ± 0.09	0.52 ± 0.09

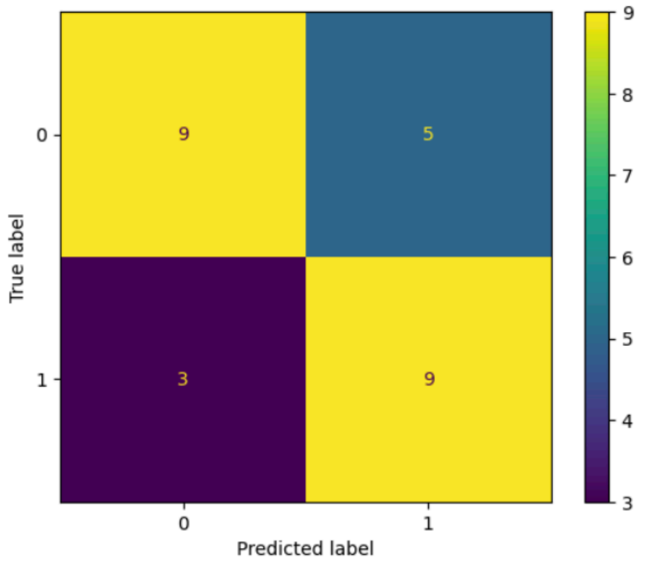
	Classifier performance				
Stimulus	Top-3 models	Precision	Recall	F1-Score	ROC AUC
PICS	MLP	**0.62 ± 0.07**	**0.62 ± 0.07**	**0.61 ± 0.07**	**0.62 ± 0.07**
	Logistic regression	0.60 ± 0.10	0.60 ± 0.09	0.60 ± 0.09	0.60 ± 0.09
	Random forest	0.60 ± 0.09	0.60 ± 0.09	0.59 ± 0.09	0.59 ± 0.09
INSTR	Logistic regression	**0.62 ± 0.11**	**0.62 ± 0.10**	**0.61 ± 0.10**	**0.61 ± 0.10**
	TabNet	0.58 ± 0.10	0.58 ± 0.10	0.56 ± 0.10	0.56 ± 0.09
	Gradient Boosting	0.42 ± 0.20	0.55 ± 0.07	0.44 ± 0.10	0.52 ± 0.08
PERS	MLP	**0.56 ± 0.17**	**0.56 ± 0.17**	**0.56 ± 0.17**	**0.55 ± 0.17**
	K-Nearest Neighbor	0.54 ± 0.09	0.54 ± 0.09	0.53 ± 0.08	0.53 ± 0.08
	Random forest	0.48 ± 0.08	0.47 ± 0.08	0.47 ± 0.06	0.45 ± 0.06

- Logistic regression model obtained the highest F1 and ROC-AUC scores, yet far from being acceptable.
- Gradient boosting obtained slightly better results with precision equal to 0.62 and recall equal to 0.61.
- Models with PICS and INSTR stimuli produced results with ROC-AUC score of around 0.61

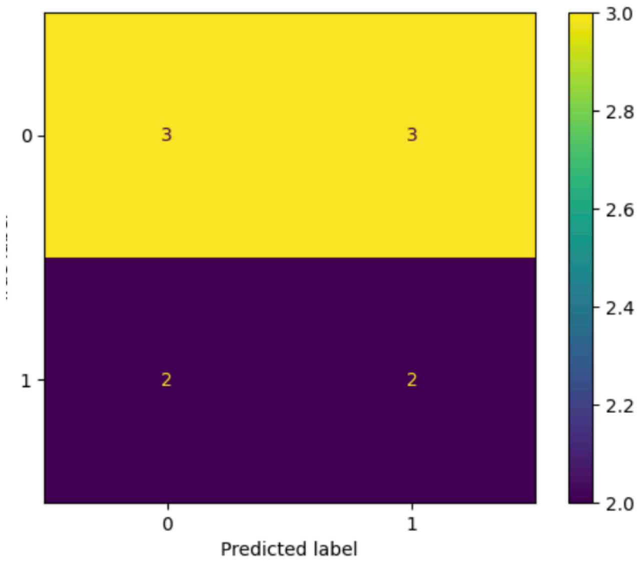
All data (no depression scale split)



Confusion matrix with INSTR stimulus of Logistic regression



Confusion matrix with PIC stimulus of MLP



Confusion matrix with PERS stimulus of MLP

- Both MLP and Logistic regression show high performance in classifying individuals with ROC-AUC score of the best models at approximately 0.7
- Overall, there is no drastic performance improvement after splitting the whole dataset into various stimuli

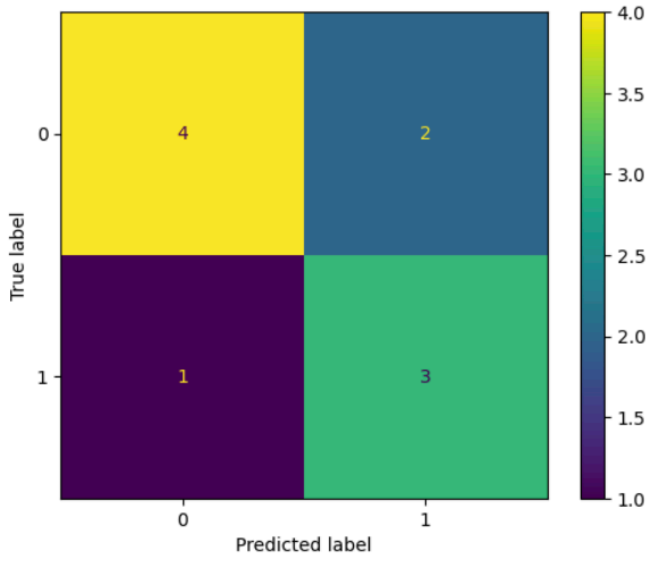
HDRS

Classifcation	Metrics			
	Precision	Recall	F1-Score	ROC AUC
Random prediction	0.56 ± 0.03	0.50 ± 0.03	0.52 ± 0.03	0.50 ± 0.03
Logistic regression	0.39 ± 0.00	0.62 ± 0.00	0.48 ± 0.00	0.50 ± 0.0
Random forest	0.39 ± 0.00	0.62 ± 0.00	0.48 ± 0.00	0.50 ± 0.0
Gradient Boosting	0.39 ± 0.00	0.62 ± 0.00	0.48 ± 0.00	0.50 ± 0.0
K-Nearest Neighbor	0.42 ± 0.07	0.57 ± 0.03	0.47 ± 0.04	0.46 ± 0.03
MLP	0.39 ± 0.00	0.62 ± 0.00	0.48 ± 0.00	0.50 ± 0.0
TabNet	**0.60 ± 0.22**	**0.67 ± 0.05**	**0.56 ± 0.09**	**0.56 ± 0.06**
Wide and Deep Learning	0.39 ± 0.00	0.62 ± 0.00	0.48 ± 0.00	0.50 ± 0.0

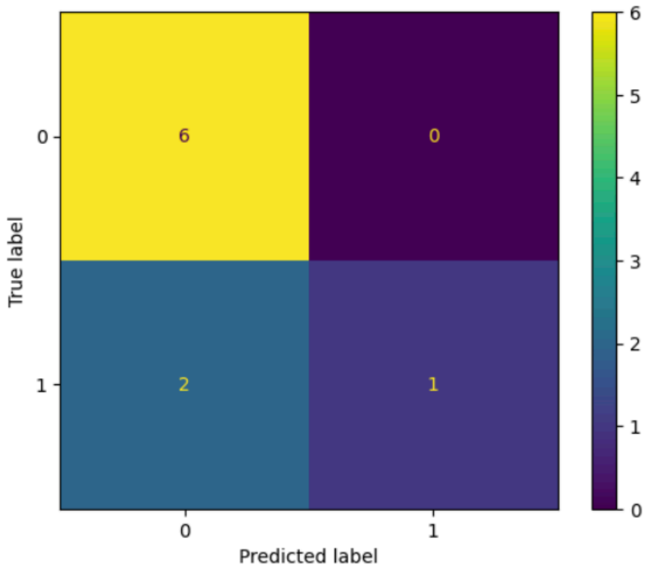
	Classifier performance				
Stimulus	Top-3 models	Precision	Recall	F1-Score	ROC AUC
PICS	TabNet	**0.69 ± 0.12**	**0.63 ± 0.09**	**0.62 ± 0.10**	**0.65 ± 0.10**
	Logistic regression	0.52 ± 0.09	0.53 ± 0.07	0.51 ± 0.08	0.51 ± 0.07
	Random forest	0.50 ± 0.22	0.59 ± 0.10	0.50 ± 0.11	0.51 ± 0.10
INSTR	TabNet	**0.66 ± 0.23**	**0.72 ± 0.16**	**0.67 ± 0.19**	**0.62 ± 0.20**
	K-Nearest Neighbor	0.62 ± 0.20	0.68 ± 0.13	0.64 ± 0.16	0.58 ± 0.17
	Random forest	0.60 ± 0.18	0.70 ± 0.05	0.62 ± 0.09	0.57 ± 0.08
PERS	TabNet	**0.50 ± 0.24**	**0.64 ± 0.13**	**0.54 ± 0.17**	**0.56 ± 0.14**
	Wide and Deep Learning	0.41 ± 0.03	0.62 ± 0.03	0.48 ± 0.04	0.52 ± 0.04
	Gradient Boosting	0.44 ± 0.20	0.56 ± 0.16	0.48 ± 0.16	0.51 ± 0.17

- TabNet model achieved the most accurate results, both F1-Score and ROC-AUC scored 0.56.
- After splitting the data set TabNet neural network outperformed all other methods: ROC-AUC score is on average higher than other top-2 models by at least 4 b.p, and F1-score is higher by at least 3 b.p

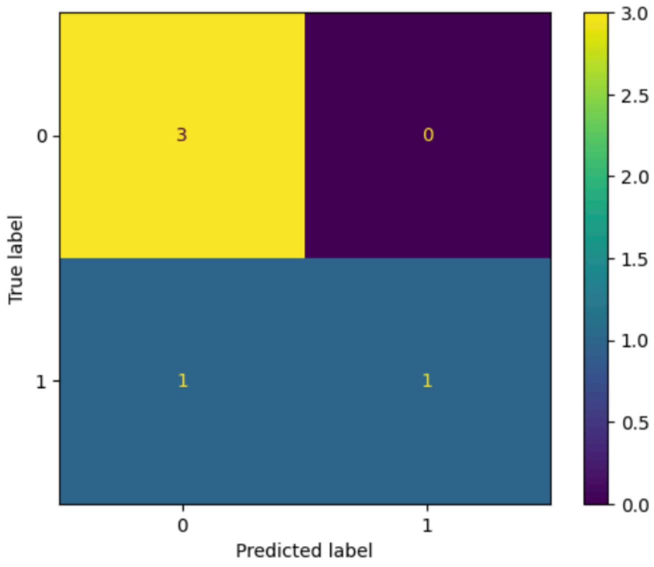
HDRS



Confusion matrix with PIC stimulus of TabNet



Confusion matrix with INSTR stimulus of TabNet



Confusion matrix with PERS stimulus of TabNet

- Overall, the metrics yielded higher scores after stimuli split only for PICS and INSTR stimulus.
- TabNet with PERS stimulus produced slightly lower results compared to the TabNet without stimuli split.

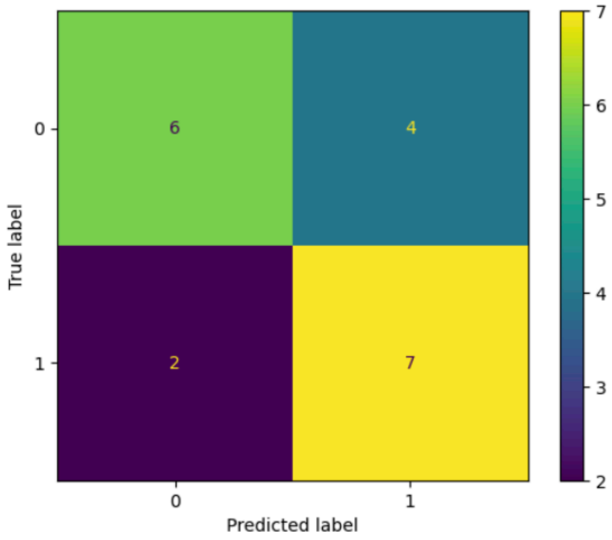
QIDS

Classifcation	Metrics			
	Precision	Recall	F1-Score	ROC AUC
Random prediction	0.50 ± 0.02	0.50 ± 0.02	0.50 ± 0.02	0.50 ± 0.02
Logistic regression	0.52 ± 0.09	0.50 ± 0.08	0.48 ± 0.09	0.51 ± 0.08
Random forest	0.54 ± 0.10	0.51 ± 0.10	0.49 ± 0.09	0.52 ± 0.08
Gradient Boosting	0.60 ± 0.10	0.52 ± 0.06	0.47 ± 0.08	0.55 ± 0.06
K-Nearest Neighbor	0.55 ± 0.08	$\mathbf{0.54 \pm 0.06}$	$\mathbf{0.53 \pm 0.07}$	0.54 ± 0.07
MLP	0.54 ± 0.07	0.53 ± 0.07	0.53 ± 0.08	0.54 ± 0.07
TabNet	$\mathbf{0.60 \pm 0.09}$	0.53 ± 0.06	0.50 ± 0.06	$\mathbf{0.56 \pm 0.06}$
Wide and Deep Learning	0.54 ± 0.07	0.53 ± 0.08	0.53 ± 0.07	0.54 ± 0.07

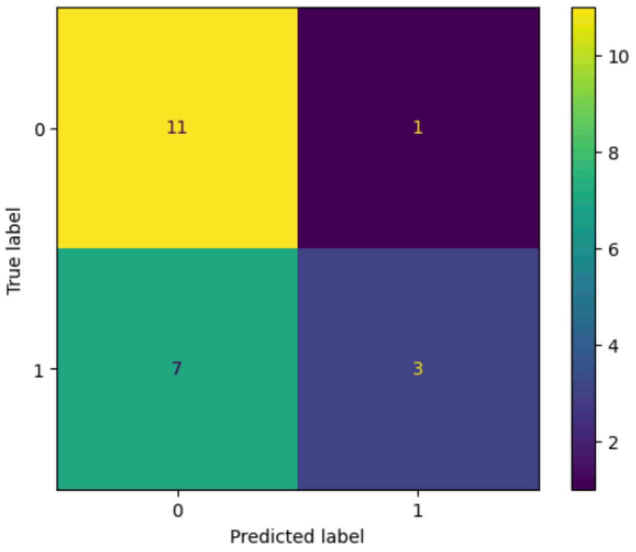
		Classifier performance			
Stimulus	Top-3 models	Precision	Recall	F1-Score	ROC AUC
PICS	MLP	$\mathbf{0.71 \pm 0.10}$	$\mathbf{0.70 \pm 0.10}$	$\mathbf{0.70 \pm 0.10}$	$\mathbf{0.70 \pm 0.09}$
	Logistic regression	0.68 ± 0.10	0.67 ± 0.10	0.68 ± 0.10	0.67 ± 0.10
	Gradient Boosting	0.59 ± 0.12	0.57 ± 0.11	0.56 ± 0.11	0.57 ± 0.11
INSTR	TabNet	$\mathbf{0.57 \pm 0.08}$	$\mathbf{0.55 \pm 0.07}$	$\mathbf{0.54 \pm 0.08}$	$\mathbf{0.56 \pm 0.07}$
	Gradient Boosting	0.52 ± 0.12	0.52 ± 0.12	0.52 ± 0.12	0.52 ± 0.12
	Logistic regression	0.51 ± 0.09	0.50 ± 0.08	0.50 ± 0.08	0.51 ± 0.08
PERS	Logistic regression	$\mathbf{0.58 \pm 0.12}$	$\mathbf{0.58 \pm 0.11}$	$\mathbf{0.57 \pm 0.12}$	$\mathbf{0.57 \pm 0.12}$
	TabNet	0.53 ± 0.10	0.52 ± 0.09	0.51 ± 0.09	0.52 ± 0.09
	MLP	0.47 ± 0.22	0.49 ± 0.11	0.43 ± 0.15	0.51 ± 0.10

- TabNet scored the highest ROC-AUC with mean equal to 0.56
- KNN achieved the highest F1-Score, 0.53
- TabNet yielded the highest precision, 0.56, while KNN score the highest recall, 0.54.
- MLP with PICS stimulus showed the highest ROC-AUC and F1-Score score out of other stimuli in QIDS dataset and out of all models, 0.7

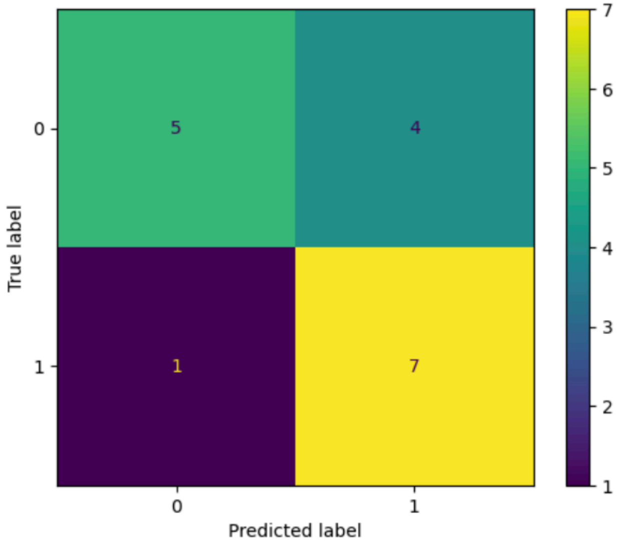
QIDS



Confusion matrix with PIC stimulus of MLP



Confusion matrix with PIC stimulus of TabNet



Confusion matrix with PERS stimulus of Logistic regression

- The Type I error presence: MLP and logistic regression classifiers tends to label patients from control group as depressed
- MLP metrics improved by 14 b.p with PICS stimulus with MLP model
- Other splits did not show any remarkable improvement.

Conclusion

	Classifier performance				
Dataset configuration	Model	Precision	Recall	F1-Score	ROC AUC
QIDS-PICS	MLP	**0.71 ± 0.10**	**0.70 ± 0.10**	**0.70 ± 0.10**	**0.70 ± 0.09**
HDRS-PICS	TabNet	**0.69 ± 0.12**	**0.63 ± 0.09**	**0.62 ± 0.10**	**0.65 ± 0.10**
ALL DATA-PICS	MLP	**0.62 ± 0.07**	**0.62 ± 0.07**	**0.61 ± 0.07**	**0.62 ± 0.07**

- MLP and TabNet neural networks demonstrated the highest ROC-AUC scores, achieving 0.70 and 0.65, respectively
- QIDS depression scale was the most effective for depression detection
- PIC-stimulus yielded the highest scores among all the metrics with different models

Future work

1. Improving quality of metrics with more advanced models.
2. Using more data for future research from Mental Health Research Center in Moscow, RF
3. Applying transformers developed for working with audio data.
4. Running clinical trials.
5. Interpreting the performance of our best model on depression detection.
6. Publication in Q2/Q3 journals

Thank you!