

# End-to-end multimodal system for depression detection from online recordings

Mateusz Kowalewski<sup>1</sup>, Maciej Stroinski<sup>1,2</sup>, Kamil Kwarciak<sup>1,2</sup>,  
 , Volodymyr Laptiev<sup>1</sup> and Daria Hemmerling<sup>1,2</sup>

**Abstract**—Depression is one of the most occurring civilizational diseases. In this paper, we propose a new approach for detecting depression through the analysis of social media content using face analysis, emotion recognition neural networks, and speech processing. We utilized audio-visual analysis and acquired more than 605 features in the time domain. Those are fed to machine learning and deep learning models for depression classification. Our approach outperforms the other state-of-the-art models, achieving the F1-score 0.77. The results have the potential to provide valuable insights for mental health professionals, offer early detection and intervention, and serve as a resource for individuals seeking help with their mental health. This study enables real-time analysis and represents a significant advancement in mental health and technology and has the potential to impact society.

**Clinical relevance**—The system aims to provide a fast and accurate way to detect depression in individuals through online recordings. The use of multimodal information (e.g. audio, image) enhances the performance of the non-verbal behavioral analysis. The end-to-end system reduces the need for manual analysis by mental health professionals and increases the efficiency of depression screening. The system can potentially help identify individuals who are at risk for depression, enabling early intervention and treatment. The results from the system can complement traditional assessments and support mental health professionals in making a diagnosis. The system can be used in real-time processing, f.e. during online calls, and provide objective measurements summarizing the overall behavior based on computer vision and audio analysis.

**Index Terms**—speech processing, computer vision, feature extraction, biomarker

## I. INTRODUCTION

Depression is one of the most common mental illnesses. It belongs to the diseases of civilization. Experts at the World Health Organization (WHO) estimate that about 280 million (3.8 percent) people worldwide suffer from it [1]. Depression is an illness that manifests itself as a lowering of mood (apathy, chronic feelings of sadness, irritability or emptiness) and/or an inability to enjoy life and things that were previously a source of pleasure (this is known as anhedonia). Patients also often have a negative evaluation of themselves, feelings of guilt, lack of faith in their ability to improve, and thoughts of death or suicide.

In addition, a person suffering from depression is characterized by a slowdown in motor and intellectual activity,

\*This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no 101017746

<sup>1</sup>Mateusz Kowalewski, Daria Hemmerling, Kamil Kwarciak, Maciej Stroinski and Volodymyr Laptiev are with SoftServe, Wroclaw, Poland, dhemm@softserveinc.com

<sup>2</sup>Daria Hemmerling, Kamil Kwarciak, and Maciej Stroinski are with the AGH University of Science and Technology, Krakow, Poland

which is expertly described as a reduction in psychomotor drive. Patients complain of physical symptoms - lack of strength, energy, and constant fatigue. They also have problems concentrating attention, remembering, or making decisions. Depression is also characterized by a disruption of biological rhythms. Patients experience a decrease or increase in appetite and body weight. In a standard diagnosis of depression, the doctor conducts a medical history, observation, and physical examination. In the process, he also diagnoses other possible mental disorders (such as anxiety or substance abuse) and rules out the presence of other symptoms of bipolar disorder. Generally, he also orders laboratory tests to make sure that depression is not caused by, or does not co-occur with, another disease (e.g., thyroid). In the diagnostic process, the doctor may also use special tests or questionnaires, filled out by himself or the patient. These include, for example, the Beck Depression Inventory, Hamilton Depression Scale, CES-D (Center for Epidemiological Studies Depression Scale). These scales are not sufficient to diagnose or rule out depression. They are used auxiliary to assess its severity. They can also be used to periodically assess the effects of therapy.

Recently, we observe a growing interest in the acquisition and processing of observable cues such as speech, and facial expressions in the screening and prediction of depression [2]. It is proven, that the observable cues are one of the visible symptoms of depression [3]. Among such cues are facial muscle movements, slower pupil dilation, and changes in vocal tract dynamics [2], [4]. To perform successful classification, digital signal processing, machine learning, and deep learning approaches can contribute to analyzing multimodal information and monitoring the signs of depression as well as other psychiatric disorders.

**Related work:** In literature, different approaches are presented to detect depression [5]. For complementary human behavioral analysis, usually, the data are accompanied by voice, video, and text. These aim to provide the diagnosis and increase the efficiency of depression screening. In multimodal classification, these are connected in different configurations. The paper [6] shows the application of transformer-based multimodal deep learning model that uses a cross-attention mechanism to generate multimodal representations to detect depression. The data used for the experiment purposes was acquired from social media consisting of 961 vlogs. The audio and visual cues were analysed. The F1-score is 0.63, precision 0.65 and a recall 0.66. The same dataset was also used by the authors of [7] for depression

detection based on non-verbal (acoustic and visual) behaviors. The authors proposed a time-aware attention-based multimodal fusion depression detection network (TAMFN) to mine and fuse the multimodal features. The F1-score is 0.65, precision is 0.66, and recall is equal to 0.66. Another approach to depression detection is the correlation-based anomaly detection framework and a measure of similarity to depression where depression relapse is detected when the deep audiovisual patterns of a depression-free subject become close to the deep audiovisual patterns of depressed subjects [8]. The correlation between the audiovisual encoding of a test subject and a deep audiovisual representation of depression is computed and is used for monitoring depressed subjects and for predicting relapse after depression. The detection accuracy was 80.99-82.55% on DAIC-Woz dataset. In the paper [9] authors applied a probabilistic model analyzing users' social activities, emotions, and language signals. As the result, the social media depression index was proposed to evaluate the user's depression level.

**Contribution:** In this work, we propose a multimodal platform for objective human behavior analysis and provide end-to-end system for depression detection. We adopted computer vision techniques to extract information about emotions, attention level, gaze analysis, movement energy and audio processing to analyze voicing and spectral information like phonation, articulation, and prosody. We proposed machine learning classification algorithms, which enable the highest performance equal to F1-score 0.77 in comparison to the state-of-the-art results. We show, that the multimodal approach enhances the performance of the non-verbal behavioral analysis.

## II. MATERIALS AND METHODS

### A. Overview

The processing pipeline consists of uploading the video, performing image and audio analysis. Image processing includes (i) frames and resolution normalization, (ii) face detection, (iii) multi-dimensional feature extraction in the time domain, (iv) providing descriptive statistics. The audio processing is conducted with (i) silence removal with voice activity detection, (ii) feature extraction, (iii) descriptive statistics. The results from the audio-visual analysis consist of the input for the classification model. The experiment pipeline is shown in Figure 1.

### B. Database

For the purpose of this research, we used D-Vlog database [6]. D-Vlog consists of 961 vlogs ( 160 hours in length), 555 are annotated with depression and 406 as non-depressed. In our study, we used 736 recordings, 423 with depression and 313 without, respectively. The descriptive statistics of the dataset used for the classification model are presented in table I. The videos were processed with their original resolution and frame rate. The audio processing was analyzed with a sampling frequency 44.1 kHz and 16-bit resolution. The dataset with a large amount of data and a variety of people can be utilized in developing depression

classification models based on computer vision and audio processing models in real-world scenarios.

TABLE I  
DESCRIPTIVE STATISTICS OF THE DATASET.

	Recordings	Avg. Duration[s]
Depression	423	616.85
Non-depression	313	522.55

### C. Multimodal approach

In our work, we applied a multimodal approach consisting of vision and audio. For the video, we first normalize all recordings. Depending on the video frame rate varies from 8 to 30 per second. Each video was later cut into frames and was prepossessed using our emotion recognition platform, which is described below. In that process, several features were obtained for each frame: blink rate, valence, arousal, number of eye saccadic movement, dominant emotion, and heart rate.

### D. Emotion recognition platform

Emotion recognition platform was developed for purpose of extracting various biosignals, such as: dominant emotion, valence, arousal, movement energy, heart rate, saccades, and blinks. Initially, the video processing first step is face detection and facial landmarks analysis. We extracted 68 landmarks and applied them to calculate blink rate, saccadic eye movement and movement energy. To extract the heart rate from the video, we applied the BioFeedback application [10]. Initially, it calculates the remote photoplethysmography (rPPG), which is a variance of red, green, and blue light reflected from the skin. These differences are caused by the changes in the blood volume in the arteries. Later the algorithm uses the acquired rPPG values to calculate the heart rate. The BioFeedback SDK applies remote photoplethysmography (rPPG) approaches to extract rPPG signals from specific areas on the human face and then signal processing methods to estimate heart rate. The pipeline performs in real-time because of multithreading architecture. The input of the SDK is a sequence of RGB video frames and the output is heart rate measurements with timestamps. Dominant emotion, valence, and arousal are implemented using EmoNet[11]. The network was trained using images of people in various emotional states (8 classes) with assigned emotion, valance, and arousal states. It uses heatmaps from the face alignment network as features for its fully connected classification head. All of those attributes are used in one of the approaches described in this work. Authors report up to 75% accuracy. Below you can see in Figure 2 the demo application that uses a web camera to predict user's biosignals. The application can use input from a web camera and predict biosignals online or load prerecorded video and work offline.

When it comes to acoustic analysis, prior to processing, a voice activity detector was applied to extract and remove silent segments from the speech [12]. The extracted speech

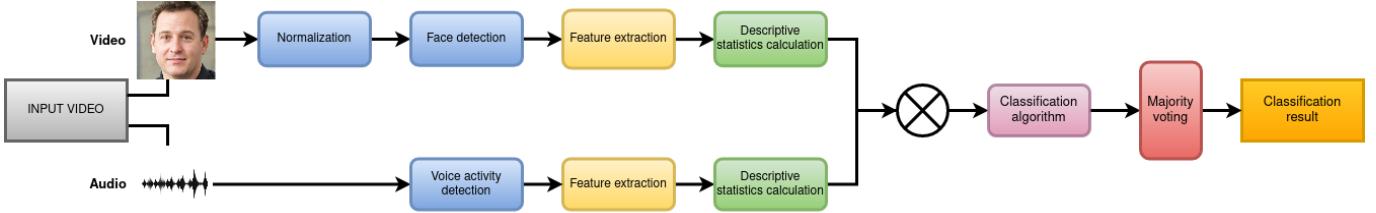


Fig. 1. Experiment pipeline used in our approach.

Energy:	0.001
Attention prob	1.0
Attention	2
Emotion	neutral
Valence	0.47
Arousal	-0.08
Direction	center
Saccades	14
Total number of blinks	20
Blink rate per minute	20
Heart Rate	45.86



Fig. 2. Demo application for biosignal sensing.

was further analyzed with OpenSMILE v2.0.0 [13] using 'ComParE 2016' being extracted at Low-level descriptors (LLD) level. This configuration enabled to extract 594 features, each was extracted with a frame size of 25 ms and a frame step of 10 ms. Exemplary features are fundamental frequency, mel-frequency cepstral coefficients (MFCC) and their deltas, Jitter, Shimmer, energy slope, psychoacoustic sharpness, zero crossing rate.

#### E. Depression classifier

We decided to use a few different approaches to binary classify depression. The output of all the methods is a binary score that indicates

1) *EfficientNetB0*: As our baseline we used EfficientNetB0 model. This architecture uses mobile inverted bottleneck convolution. As an input we use normalized frames from videos. We use every 15th frame to make prediction. After every frame has its label assigned majority voting is applied. It is a baseline model so no audio features are used in this approach.

2) *Numerical feature approach*: Further analysis included statistical analysis of acquired features over time. To do this we calculated mean, quantiles, interquartile range, skewness and kurtosis from each feature over 30 second windows with 50% overlap. Every video has a different frame rate so windows might vary in frame number but since we are calculated normalized features it does no matter. No means of frame rate equalization was not used. We did same thing for acoustic features. Since our features are purely numerical we decided to use XGBoost and fully connected neural network as classifiers. At the end prediction from each frame are taken and majority voting is applied.

### III. RESULTS

Table II shows results comparison between used methods. For each method 5-fold stratified validation was used because the dataset doesn't contain a separate test set. Each fold class distribution resembles the distribution of the whole dataset.

TABLE II  
METHODS EVALUATION RESULTS.

Architecture	Modalities	F1-Score	Recall	Precision
Cross attention [5]	audio + video	63.50	65.57	65.40
EfficientNet	video	0.55	0.46	0.69
NN	audio	0.72	0.67	0.77
NN	video	0.72	0.67	0.77
NN	audio+video	0.72	0.67	0.77
XGBoost	audio	0.76	0.83	0.70
XGBoost	video	0.73	0.87	0.63
XGBoost	audio+video	<b>0.77</b>	0.84	0.71

In Figure 3 we can see precision-recall curve over 5-folds from our best approach.

Thanks to its nature XGBoost classifier supports Gini Coefficient based feature importance that we can see in Figure 4.

### IV. CONCLUSION

The Table II presents that the XGBoost outperforms other methods in this task. It is expected behavior compared to the baseline model. But it might be quite surprising that it outperforms neural network on the same features.

Differences between XGBoost and neural network performance can be explained when we investigate the nature of the feature fed to those methods. In this study, we used features

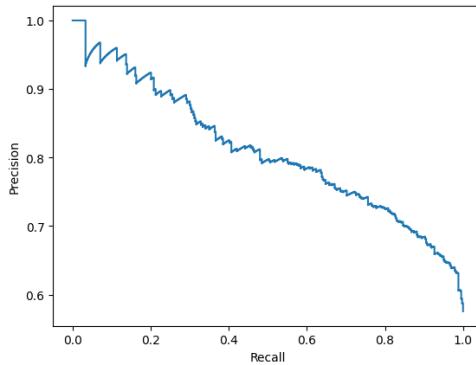


Fig. 3. Precision-recall curve over 5-fold from XGBoost classifier.

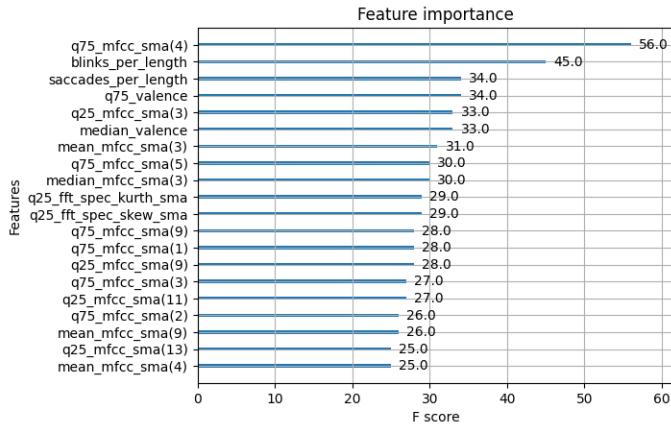


Fig. 4. Feature importance from XGBoost classifier. Kurth - kurtosis, skew - skewness, sma - simple moving average, mfcc - mel-frequency cepstral coefficients, fft-spec - Fourier spectrum.

that are heterogeneous in its nature. It means that they don't have the same scale and unit of measure. Heart rate is in beats per minute, emotions have discrete values and valence is capped between -1 and 1. Since tree-based methods treat each feature independently above problem doesn't affect them. This is confirmed by obtained results.

We can clearly see that adding more modalities improves results with XGBoost classifier. Since neural network does not show improvements upon adding more features we might suspect that it gets stuck in a local minimum.

Compared to previous works our solution achieves state-of-the-art results in terms of F1 score.

Figure 4 shows future importance extracted from XGBoost model. We can see that model learns a lot from MFCC coefficients. That might be due to the rather sad tone of people's voices. It also focuses on valence and arousal that are directly connected with a person's emotional state during the video recording. The model is actually especially dependent on valence which might indicate the negative mood of the participant. The last very important features are blink rate and a number of saccadic eye movements. It would validate the statement from this paper [14] that says "blink rate is increased in depression and falls to normal levels during treatment".

In future work, we plan to extend the research by adding sentiment analyses of the recording. It can be done by running an automatic speech recognition tool on the recording to extract text. From the text using pre-trained models, we can extract the sentiment of the recording. That would bring additional information about participant mood and potentially increase the performance of our classifier.

To conclude, we proposed an XGBoost-based depression classification model. Introduced emotion recognition tool to extract important features and proved that depression can be detected using classical machine learning algorithms.

## REFERENCES

- [1] World Health Organization et al., "Depression and other common mental disorders: global health estimates," Tech. Rep., World Health Organization, 2017.
- [2] Umut Arioiz, Urška Smrke, Nejc Plohl, and Izidor Mlakar, "Scoping review on the multimodal classification of depression and experimental study on existing multimodal models," *Diagnostics*, vol. 12, no. 11, pp. 2683, 2022.
- [3] Peter H Waxer, "Therapist training in nonverbal communication: I. nonverbal cues for depression.," *Journal of clinical psychology*, 1974.
- [4] Anastasia Pampouchidou, Panagiotis G Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 445–470, 2017.
- [5] 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*, vol. 80, pp. 56–86, 2022.
- [6] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 12226–12234.
- [7] Li Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu, "Tamfn: Time-aware attention multimodal fusion network for depression detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.
- [8] Alice Othmani and Assaad Oussama Zeghina, "A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept," *Healthcare Analytics*, vol. 2, pp. 100090, 2022.
- [9] Munmun De Choudhury, Scott Counts, and Eric Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th annual ACM web science conference*, 2013, pp. 47–56.
- [10] SoftServe, "BioFeedback," <https://demo.softserveinc.com/biofeedback-sdk/>, 2023, [Online; accessed 19-January-2023].
- [11] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, pp. 42—50, 2021.
- [12] Johannes Wagner, Dominik Schiller, Andreas Seiderer, and Elisabeth André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?," in *Proceedings of Interspeech*, Hyderabad, India, 2018, pp. 147–151.
- [13] Audeering, "OpenSMILE," <https://www.audeering.com/research/opensmile/>, 2023, [Online; accessed 19-January-2023].
- [14] Kitamura T. Mackintosh JH, Kumar R, "Blink rate in psychiatric illness.," *The British journal of psychiatry : the journal of mental science* vol. 143, pp. 55–57, 1983.