IRSTI 28.23.01 https://doi.org/10.26577/jpcsit2025337

Aizhan Nurzhanova[1], Miras Mussabek[2], Gokhan Ince[3], Mas Rina Mustaffa[4], Ainur Zhumadillayeva[1*]

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Astana IT University, Astana, Kazakhstan
[3]Istanbul Technical University, Istanbul, Turkey
[4]Universiti Putra Malaysia, Serdang, Malaysia
*e-mail: zhumadillayeva_ak@enu.kz

DETECTION OF MENTAL DISORDERS BASED ON THE ANALYSIS OF EMOTION, FACIAL EXPRESSIONS AND FACIAL MOVEMENTS IN A VIDEO STREAM

Abstract. Traditional emotion recognition systems typically use publicly available models without regard to differences in future emotions. In this paper, we find the possibility of violations based on the analysis of facial expressions and movements in a video stream taking into account features. A personalized approach to emotion recognition is proposed, implemented using machine and deep learning algorithms. The system uses the MediaPipe FaceMesh to extract 468 facial keypoints and analyze expressions associated with conditions such as anxiety, depression, post-traumatic stress disorder(PTSD), mania, or fatigue. Experiments have confirmed that personalized models provide higher accuracy compared to robustness. We propose a video-based framework for predicting mental disorders by analyzing temporal facial dynamics, micro-expression volatility, and asymmetry using RGB streams. Our system identifies disorder-specific biomarkers like delayed emotional reactivity and slow blinks, enabling real-time mental health detection.

Keywords: emotion recognition, video models, individual differences, personalized models, deep learning, affective computing.

1. Introduction

Emotions and facial expressions play an important role in human communication, influencing interpersonal interactions, cognitive processes, and decision-making. Automated emotion recognition systems have gained great importance in affective computing and human-computer interaction systems. However, traditional models generalize emotional expressions without taking into account individual differences. The variability of facial expressions, head movements, and micro-emotions requires the development of personalized recognition models that are sensitive to the characteristics of a particular person.

In light of the growing importance of human-computer interaction, research in the field of psycho-emotional state recognition is becoming especially relevant. Automatic emotion recognition attracts the attention of both researchers and developers, which makes this task significant for fundamental research, applied developments in the field of information technology and neural interfaces.

Mental health is an integral part of a person's overall condition, determining their emotional well-being, cognitive abilities, and quality of life. In recent decades, there has been a steady increase in the number of people suffering from mental disorders, such as depression, anxiety disorders, bipolar disorder, emotional burnout, and chronic fatigue. These conditions often develop gradually and may remain unnoticed until a critical phase. Therefore, early detection of signs of psychoemotional disorders is of particular importance.

One of the key indicators of a person's internal state is emotion. Emotions reflect a person's reaction to external and internal stimuli, and form the basis of motivation and behavior. They are accompanied by physiological changes, behavioral reactions, and, above all, are expressed through facial expressions – movements of the facial muscles that form facial expressions. Facial expressions, in turn, are a universal and intuitive communication channel: they allow you to convey emotional states even without words.

Of particular importance are microexpressions – short-term, involuntary facial expressions that last less than 0.5 seconds. These expressions arise in response to real emotions, even if a person tries to hide them. Scientific research has proven that microexpressions are a highly informative feature that can indicate internal conflicts, stress, anxiety or suppressed emotions. Changes in the nature of emotional manifestations – their frequency, intensity, symmetry and diversity – can be a sign of disorders in the psycho-emotional sphere. In recent years, the rapid development of computer vision, machine and deep learning methods has made it possible to automate the process of analyzing emotions. Modern systems are able to accurately detect a face on video, track key points (for example, using technologies such as MediaPipe FaceMesh), extract facial features and classify emotional states. This has created the prerequisites for building intelligent systems for detecting mental disorders that can assess a person's emotional reactions in real time. In modern society, there is an increase in the number of psycho-emotional disorders, such as depression, anxiety, bipolar disorder, chronic fatigue and other forms of mental disorders. These conditions often go unnoticed in the early stages due to the subjectivity of self-diagnosis, social stigma, and limited access to mental health professionals. In this regard, there is an increasing need to develop non-invasive, automated, and objective methods for the early detection of signs of mental disorders.

One of the promising areas is the analysis of a human face video stream to assess facial expressions, microexpressions, and emotional reactions. The face reflects a wide range of psycho-emotional states, and its movements are a powerful indicator of an individual's internal state. Modern computer vision and machine learning technologies make it possible to track the smallest changes in facial expressions and analyze them in real time, which opens up opportunities for continuous monitoring of the emotional state without the need for direct intervention.

Analysis of non-verbal signs, including facial expressions, microexpressions, and motor patterns, opens up new opportunities in the early detection of psychoneurological and mental disorders, such as depression, anxiety, autism, post-traumatic stress disorder, and schizophrenia.

The use of algorithms based on facial recognition features such as expression asymmetry, blink rate, muscle tension, lip and eye movement dynamics, in combination with emotion classification, provides a deeper understanding of a person's internal psycho-emotional state. Such systems can be effectively used in telemedicine, psychotherapy, education, as well as for monitoring the condition of employees in responsible areas of activity.

Today's algorithms provide high accuracy and can serve as the basis for scalable, non-invasive, and personalized mental health support systems. Among all approaches, deep neural networks can be noted, where deep learning algorithms, in particular convolutional neural networks (CNN), demonstrate high efficiency in the task of facial expression recognition (FER), significantly surpassing traditional methods in the accuracy and stability of results [1],[2].

The paper discusses the application of video analysis and deep learning methods to assess a person's mental state in order to detect signs of mental disorders. The approaches are based on the synthesis of computer vision, affective computing, and machine learning methods.

A person's mental health can manifest itself in various behavioral features or physiological changes. For example, micro-movements of the eyes and pupil dilation can be markers of internal stress. Deep gaze features extracted by autoencoders are effectively classified using Random Forest, achieving 100% accuracy [3]. Pupil diameter is tracked in the video stream and complements facial expression analysis to improve accuracy [4].

Body posture and gestures also reflect the psycho-emotional state. Using Pose Net, it is possible to analyze posture and body movements associated with anxiety (e.g. fidgeting, closed postures). Accuracy is F1: 0.98–0.99 [5]. Head movements (nodding, tilting) and body movements are analyzed together with facial expressions and gaze to obtain a comprehensive picture [6].

Video data allows non-invasive extraction of physiological signals: remote photoplethysmography (rPPG) is used to estimate HR and HRV. Random Forest classifiers achieve 99% accuracy [1,7]. Hemoglobin (HC) changes are estimated by skin color. The analysis is based on bit planes and ML models [8].

Facial expressions are one of the most informative channels for identifying emotional states, including anxiety. Active shape models (ASM) capture key facial landmarks (eyes, lips, eyebrows) and generate feature vectors classified by

SVM to determine anxiety and stress [9]. Deep neural networks trained on specialized datasets allow accurate recognition of anxious expressions and are integrated into web applications for real-time analysis [10]. Multi-task learning with attention mechanism combines facial expression with physiological characteristics (HR, HHR), achieving an accuracy of up to 94.33% [11].

We propose a video-centric framework to predict mental disorders by analyzing temporal dynamics, micro-expression volatility, and facial asymmetry metrics derived exclusively from RGB video streams. Leveraging the MediaPipe FaceMesh tool, our system quantifies disorder-specific biomarkers–such as delayed emotional reactivity in depression and slow blinks with yawning in fatigue. Our framework enables real-time mental health identification through optimized edge deployment, processing video streams directly on consumer-grade devices.

2. Materials and methods

In this study, we utilized a real-time emotion detection system with the help of the MediaPipe Face Mesh model developed by Google. We analyzed individual differences in emotional expression and recognition using 468 facial landmarks of the human face. The eyes, eyebrows, nose, mouth, and jawline were among the important face traits that the system tracked. Geometric features such as ratio of the human face, and distances between each keypoints were extracted using these landmarks.

In face detection side MediaPipe Face Mesh's utilize BlazeFace, a neural network architecture tailored for mobile GPU inference [21]. BlazeFace uses a simplified feature extractor made especially for faces, but it is based on the Single Shot MultiBox Detector (SSD). It uses a unique backbone that is similar to but different from MobileNetV1/V2 with custom residual structures known as BlazeBlocks and double BlazeBlock, which contains 5×5 depthwise separable convolutions instead of the typical 3×3 kernels. These design components aid in expanding the receptive field with little overhead and are computationally efficient. The model processes inputs at 128×128 resolution and uses an anchor-based detection scheme that stops at an 8×8 spatial resolution to minimize latency, replacing non-maximum suppression with a regression-weighted blending strategy to enhance stability over time.

The face landmark side generates a dense 3D mesh with 468 vertices, each of which is treated as an independent landmark [22]. These points are placed to capture perceptually significant facial regions. The model is trained using a combination of synthetic 3D renderings and annotated 2D landmarks from real-world mobile images, as well as special noise modeling and lighting variability techniques.

For our experiments we implemented the rule-based approach to identify different human mental disorders, including anxiety, depression, panic disorder, bipolar disorder and fatigue based on the relative positions and movements of facial features (see Table 1). It takes position coordinates of face landmarks and calculates the differences and distances between them. All positions are relative, so the distance between face and recording device will not affect the results.

Table 1 – Emotion detection criteria based on facial landmark features.

Condition	Facial Signs	Detection Logic
Anxiety	Rapid blinking, lip compression [12],[13]	Blink rate: EAR < 0.2 for 3+ frames = blink. >20/min = anxiety.
		Lip compression: Vertical distance between [61] and [291] < 0.05 (normalized).
Depression	Reduced smiling, flat affect [14],[15]	Smile absence: Mouth corner distance (61-291) < 0.2 (neutral) for >80% of frames.
		Slow blinks: EAR < 0.2 for >0.5s.
Hypervigilance (Post-traumatic stress disorder)	Sudden eye widening, mouth slackening [16],[17]	Eye widening: EAR > 0.3 (baseline: 0.25) + mouth openness (13-14 distance > 0.15).
Mania (Bipolar disorder)	Excessive smiling, eyebrow raises [18],[19]	Smile intensity: Mouth corner distance > 0.4 for >50% of video.
		Eyebrow volatility: Rapid AU2 (brow raise) spikes.
Fatigue (Sleep Deprivation)	Slow blinks, yawning [20]	Slow blinks: EAR < 0.2 for >0.5s.
		Yawning: Jaw (17-181 distance) > 0.3 for 2+ sec.

We manually annotated 100 video samples from human subjects, labeling each with one of five mental states. Each sample consisted of a short video sequence in which facial behavior was observed and classified using established psychological markers. The system's predictions, based on facial landmark thresholds, were then compared to these ground truth labels to evaluate classification performance. Experiments conducted on the machine with these specifications: NVIDIA GeForce RTX 3080 Laptop GPU and 16GB 3200MHz RAM.

For the real-time our system processed video frames at 30 frames per second using a standard webcam(see Figure 1). Each frame was converted to RGB format and passed through the MediaPipe Face Mesh pipeline to extract facial landmarks. They helped us to build prediction statements for human mental disorders. Facial signs conditions are written in python code as well as implementation of MediaPipe Face Mesh.

In comparison with usual similar systems, our framework employs a multi-stage adaptive architecture that integrates personalized baseline profiling, dynamic threshold adjustment (see Figure 2). By that we can address critical gaps in handling individual variability for each person. Also the system uses its lightweight CNNs for real-time detection for possibility to implement it in edge devices, where other models can use multiple models, separate for detection and disorder classification (see Figure 3).

Figure 1 – Emotion recognition examples based on facial expression analysis with one member of our team.

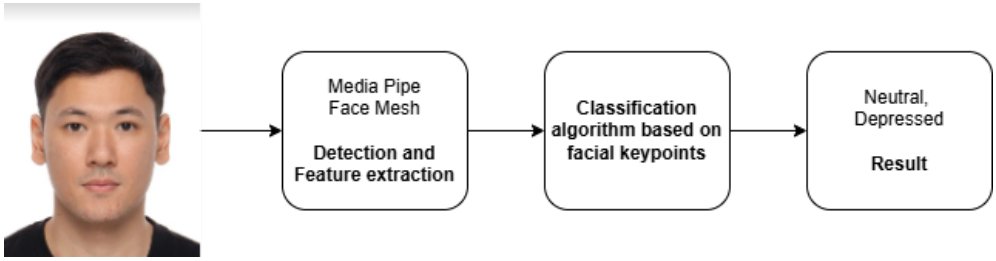

Figure 2 – Proposed mental disorder detection system architecture.

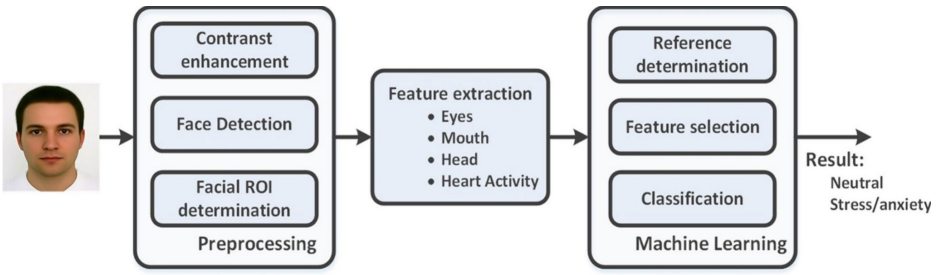

Figure 3 – Emotion recognition system example in [12].

To assess the performance of our rule-based system, we used standard classification metrics like precision, recall, F1-score and accuracy. These metrics enable us to quantify how well the system identifies each mental state, particularly in the presence of class overlap or imbalanced predictions.

Precision measures the reliability of the system when it predicts a specific class – in other words, how many of those predictions were actually correct. For example, when the system predicts that someone is experiencing mania, precision tells us how often that prediction is accurate (1):

$$\text{Precision} = \frac{\text{True Positives+False Positives}}{\text{True Positives}} \quad (1)$$

Recall quantifies the system's ability to detect all actual instances of a given class – for example, how many true cases of depression the system was able to identify (2):

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}} \quad (2)$$

The F1-score strikes a balance between precision and recall, making it an especially useful metric when dealing with uneven class distributions or overlapping symptoms. It is the harmonic mean of precision and recall, yielding a single score that reflects both the system's accuracy and sensitivity (3).

$$\text{F1 score} = 2 * \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (3)$$

Accuracy represents the overall effectiveness of the rule-based classification system (4):

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad (4)$$

3. Results

Anxiety is identified through heightened blink rates and lip compression [12,13]. The algorithm calculates blink frequency using the Eye Aspect Ratio (EAR), where an EAR below 0.2 for three consecutive frames registers as a blink. A sustained rate exceeding 20 blinks per minute signals potential anxiety. At the same time, lip compression measured by the vertical distance between landmarks 61, which is upper lip, and 291, which is lower lip, is flagged when normalized to less than 0.05. MediaPipe Face Mesh tracks these metrics via eyelid landmarks, for example indices 362, 385 are for the left eye and 33, 160 for the right eye, and lip landmarks.

According to [15], depression is associated with reduced smiling and prolonged blinks. Our algorithm monitors the distance between mouth corners, the landmarks 61 and 291, where a value below 0.2 for over 80% of frames suggests diminished positive affect. Slow blinks, defined as

eyelid closure, with EAR smaller than 0.2 lasting longer than 0.5 seconds, are tracked using the same eye landmarks as anxiety. MediaPipe's precision in capturing subtle lip and eyelid movements allows for continuous assessment of emotional withdrawal.

Hypervigilance in Post-traumatic stress disorder(PTSD), is detected through sudden eye widening and mouth slackening [16,17]. We fully used MediaPipe's capability to track subtle changes in eye and inner lip landmarks for real-time detection. An EAR exceeding 0.3, signals exaggerated eye openness. Meanwhile inner lip distance, which are the landmarks 13 and 14, surpassing 0.15 reflects mouth tension release.

Mania is characterized by excessive smiling and erratic eyebrow movements and most of the time is a Bipolar disorder [18,19]. In our algorithm a mouth corner distance, which are landmarks from 61 to 291, exceeding 0.4 for more than half of observed frames indicates intense, sustained smiling. Also eyebrow volatility, which is measured by rapid displacement of brow landmarks 336 and 296 for the left, 66 and 105 for the right brow, reflects heightened arousal. We map these dynamics through upper lip curvature, 37 and 267, and brow motion.

Figure 4 illustrates temporal trajectories of selected facial landmarks**.** For example, eyebrow landmark movements (66, 105, 296, 336) display pronounced volatility in mania compared to stable patterns in neutral or depressive states. Such temporal dynamics highlight the relevance of personalized thresholding.

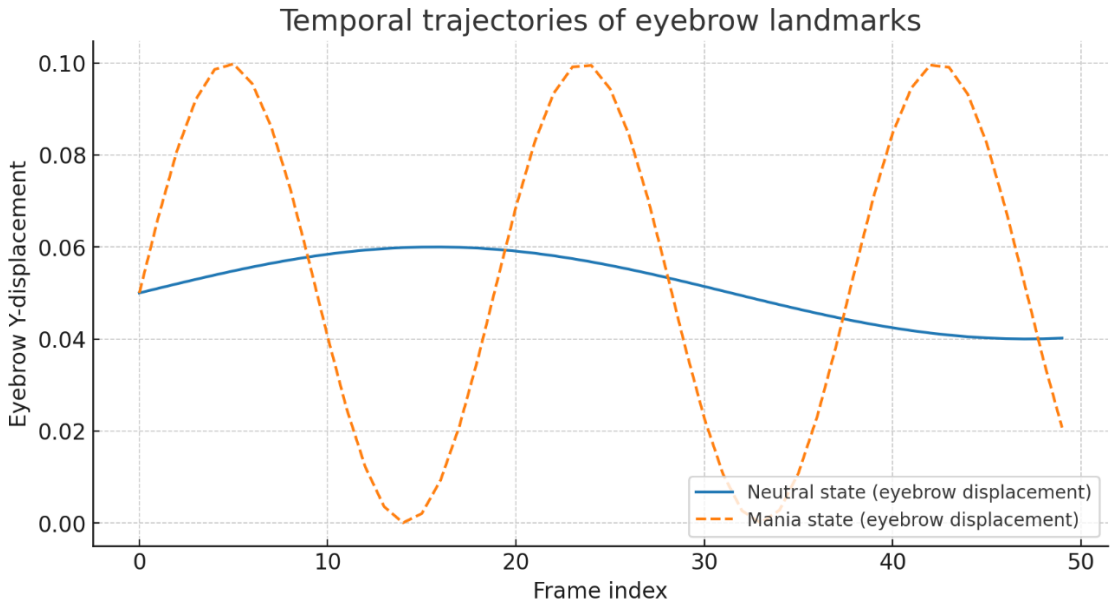

Figure 4 – Temporal trajectories of selected facial landmarks

Fatigue measured in our algorithm through prolonged blinks and yawning [20]. Slow blinks, EAR smaller than 0.2 for more than 0.5 seconds, are paired with jaw dropping. It is measured by the distance between chin landmark 17 and jawline point 181 exceeding 0.3 for two seconds. We used MediaPipe's jaw and eyelid tracking to ensure robust detection across varied scenarios like stretching or screen interaction.

In Table 2 we show the confusion matrix, which summarizes the classification results across the five mental states. Each row represents the actual label, and each column represents the predicted label assigned by the system. Diagonal entries represent correctly classified samples, whereas off-diagonal entries indicate misclassification. The model was highly accurate in detecting mania and depression, with few misclassifications. However, there was a significant overlap between fatigue, anxiety, and PTSD, owing to the visual similarity of certain facial cues such as droopy eyelids and decreased facial tension across these states.

Table 2 – Confusion matrix illustrating the classification performance of the rule-based system across five mental state categories.

Actual \ Predicted	Anxiety	Depression	PTSD	Mania	Fatigue
Anxiety	14	2	1	0	3
Depression	1	16	0	0	3
PTSD	2	1	13	1	3
Mania	0	0	1	18	1
Fatigue	3	2	2	1	12

We calculated the precision, recall, and F1-scores for each of the five mental state categories using the confusion matrix(see Table 3). Mania performed the best in all metrics, with precision and recall both at 0.90, followed by depression and PTSD. Fatigue had the lowest scores, owing to visual similarities with other classes, which resulted in a higher false positive rate. The system's overall accuracy was 73%, which demonstrates the reliability of the threshold-based approach.

Table 3 – Performance metrics of the system across five mental state categories.

Metric	Anxiety	Depression	PTSD	Mania	Fatigue
Precision	0.700	0.762	0.765	0.900	0.545
Recall	0.700	0.800	0.650	0.900	0.600
F1-score	0.700	0.780	0.703	0.900	0.571
Accuracy	**0.730**				

All information taken from a person's facial behaviour can predict some mental disorders. Nonetheless, similar patterns in each case may arise during different human moods. For example for Hypervigilance conditions, the same patterns can happen in the moments of surprise or concentration, underscoring the need for corroborating behavioral data to distinguish pathology from transient reactions. It risks misinterpreting naturally expressive individuals, emphasizing the importance of longitudinal tracking to differentiate episodic mental disorder actions from temperamental exuberance.

The experimental results demonstrate that personalized video-based emotion recognition models can be used in real-world applications, with good camera conditions, where a person can be detected with the Mediapipe library. This result stems from the models' ability to account for individual variations in facial dynamics, such as unique microexpression patterns and temporal differences in emotional expression transitions (Figure 1). However, the analysis revealed challenges in distinguishing between visually similar states, such as fatigue and depression, because of overlapping facial conditions like drooping eyelids and reduced smile intensity.

MediaPipe's facial landmark tracking proved effective in capturing facial features like jaw tension, eyelid closure duration, and inner lip compression. They are all associated with mental states. For instance, we can differentiate landmark-based metrics with intentional smiles from spontaneous ones. To address limitations for individual persons, we incorporated adaptive thresholds that adjusted classification criteria based on user-specific baselines. For example, resting lip distance and blink frequency were calibrated during initial sessions, reducing false positives in natural settings.

4. Discussion

In this study, we conducted research aimed at detecting mental disorders by analyzing videos that assess emotions, facial expressions, and facial movements. The proposed approach, based on computer vision technologies, demonstrates strong potential for the early diagnosis of mental disorders. The developed technology can be useful in healthcare practice for pre-identifying signs of mental distress before a patient visits a physician, thereby accelerating diagnosis and improving treatment outcomes. In this study, only possible

mental disorders were considered, but in the future, this method may be extended to diagnose neurological conditions, such as Bell's palsy, which is characterized by facial asymmetry.

A promising direction for the further development of this technology is its application in the diagnosis of neurological disorders. In particular, video stream analysis may be beneficial in detecting Parkinson's disease, Huntington's disease, bulbar and pseudobulbar syndromes, multiple sclerosis, facial neuritis (including Bell's palsy), and Tourette's syndrome. Including such conditions would significantly increase the diagnostic value of the proposed approach and expand the possibilities for early detection of diseases manifesting through facial and motor patterns.

The difficulty in distinguishing between fatigue and depression highlights one of the key limitations of unimodal systems. Although both conditions are associated with reduced ocular activity and slowed facial expressions, depression is often characterized by prolonged microexpressions of sadness. The integration of speech analysis or a combination of head pose tracking with galvanic skin response (GSR) data may allow for a more accurate differentiation between cognitive fatigue and emotional withdrawal.

Thresholds for indicators such as lip compression, gaze deviation, or smile intensity may vary significantly across ethnicities, ages, and cultural backgrounds. For instance, norms for lip distance are influenced by facial structure, while eye contact conventions are determined by cultural norms. It is well known that individuals with darker skin tones often experience higher false-positive stress detections due to improper calibration of infrared cameras. This can lead to entrenched diagnostic bias. To reduce such risks, participatory design principles should be applied: involving diverse user groups in threshold calibration, using synthetic datasets to supplement underrepresented facial structures, and including cultural consultants to adapt models to regional norms. For example, increasing the acceptable threshold for gaze deviation in cultures where avoiding eye contact is perceived as a sign of respect rather than an attempt at deception.

While the current version of the system applies rule-based thresholds, including metrics like eye aspect ratio, lip distance, and blink rate, based on existing empirical studies, we acknowledge the need for a more rigorous scientific foundation. In future studies, we will perform systematic threshold validation across larger and more diverse samples using statistical analyses, expert clinical input, and data-driven optimization techniques. This will ensure that the selected facial parameters have robust diagnostic relevance rather than relying solely on heuristic definitions.

Another important aspect not yet addressed in this study is the comorbidity of mental disorders. Many individuals exhibit overlapping symptoms from multiple conditions–such as depression co-occurring with anxiety, or PTSD alongside fatigue–which complicates classification based on isolated features. Since our current framework is designed to detect single-condition markers, it may fail to capture the interaction between multiple symptom domains.

To address this, we plan to introduce multilabel classification models capable of identifying multiple co-occurring conditions simultaneously. Additionally, we aim to adopt temporal modeling techniques (e.g., LSTMs, Transformers) that account for behavioral patterns over time–patterns which often reflect comorbid states more reliably than static snapshots. These enhancements will be supported by expanding the dataset to include clinically verified cases with annotated comorbid symptoms.

Improving the accuracy and reliability of the system requires refinement of the algorithms and expansion of the training data. Introducing more complex analysis conditions and training the model on larger datasets with annotated mental disorders will allow much of the diagnostic process to be delegated to artificial intelligence. This will not only enhance recognition accuracy but also improve model robustness to environmental factors such as lighting, camera angle, background, and more.

Nevertheless, this study has a number of significant limitations related to the dataset used. First, the dataset is proprietary and private, created solely for experimental validation. It contains video recordings of only one subject, captured under controlled conditions. As such, the dataset lacks diversity in facial expressions, facial structure, and behavioral responses. This limited scope restricts the generalizability of results and prevents broader validation of model robustness.

In future research, we plan to address these limitations. First, we will develop a larger dataset involving a diverse group of participants, including

both clinical cases (e.g., depression, anxiety, bipolar disorder) and healthy individuals. This will allow us to balance class representation and improve the generalizability of the model.

Second, we plan to ensure demographic diversity across factors such as gender, age, ethnicity, and cultural background. This is essential for creating a robust and ethically sound model that does not discriminate based on appearance or cultural behavior.

Third, in future work, we intend to shift from a unimodal approach to a multimodal system, integrating visual cues with speech characteristics, physiological data (e.g., heart rate, EEG, GSR), and audio signals. This approach will enhance the diagnostic accuracy, particularly in complex cases where behavioral cues may be absent or masked.

Furthermore, we aim to implement an open and reproducible methodology, including the publication of anonymized data (with participant consent), label definitions, feature descriptions, and transparent evaluation protocols. This will support scientific verification and comparability with other research efforts.

Thus, the future development of our system is focused on creating a scalable, ethically responsible, and personalized tool for the early diagnosis of mental and neurological disorders in real-world conditions.

5. Conclusions

This research demonstrates the feasibility of using video-centric affective computing as a non-invasive, scalable tool for mental health assessment. By analyzing temporal facial dynamics and microexpressions extracted from RGB video using MediaPipe FaceMesh, the system identifies disorder-specific facial markers such as delayed emotional reactivity (in depression) and microexpression volatility (in anxiety). The real-time capability of the framework, coupled with personalized adaptation that adjusts thresholds to individual facial baselines, highlights its potential for practical deployment across neurodiverse and demographically varied populations.

However, the study also reveals important limitations. The model was developed and tested on a private, single-subject dataset without class balance or demographic variation. Rule-based labeling was applied without clinical verification, which limits generalizability and diagnostic reliability. In addition, the current unimodal design restricts the system's ability to resolve overlapping emotional states–such as distinguishing between fatigue and depression–both of which may exhibit similar facial attenuation.

To overcome these challenges, future work will focus on developing a large, demographically diverse dataset with clinically annotated cases, including both isolated and comorbid mental health conditions. The system will evolve toward a multimodal architecture by integrating facial analysis with speech prosody, physiological signals (e.g., heart rate variability, EEG, GSR), and temporal modeling techniques (e.g., LSTMs, Transformers) to better capture dynamic and overlapping behavioral cues.

Ethical deployment is central to this work. As systems like this infer highly sensitive mental states, safeguards must include transparency at multiple levels: how decisions are made (e.g., which facial features contribute), how data is stored and used, and how predictions are interpreted. In high-stakes applications such as workplace monitoring or insurance assessment, dynamic consent mechanisms and user-facing explainability are critical to ensuring trust and preventing misuse.

Cultural and neurodivergent inclusion must remain a design priority. Participatory design with underrepresented groups, systematic bias audits, federated learning methods, and region-specific model calibration will help reduce risks of stigmatization and increase fairness and model robustness across diverse real-world contexts.

In conclusion, while this work serves as a foundational proof of concept, its continued development–through theoretical grounding, multimodal expansion, and ethical reinforcement–will be essential for building a clinically reliable and socially responsible system for early mental health and neurological disorder detection.

Funding

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Grant No. AP26100559 "Development of computational methods for determining and generating positive and negative emotions from photo and video data for monitoring of the educational process".

Author Contributions

Conceptualization: A.N., M.M. and A.Z.; Methodology: A.N., M.M.; Software: M.M.; Validation: G.I., M.R. M. and A.Z.; Formal analysis: A.N., M.M.; Investigation: A.N., M.M.; Resources: A.N., M.M.; Data Curation: M. M.; Writing – Original Draft Preparation: A.N., M.M.; Writing – Review & Editing: A.N., M.M.; Visualization: M.M.; Supervision: G.I., M.R.M. and A.Z.; Project administration: A.Z.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. A. Devarapalli and J. Gonda, "Investigation into facial expression recognition methods: a review," *Indonesian Journal of Electrical Engineering and Computer Science*, 2023. doi: 10.11591/ijeecs.v31.i3.pp1754-1762.
2. D. Uneza and S. Gupta, "Facial Expression Analysis: Unveiling the Emotions Through Computer Vision," in *Proc. Int. Conf. Recent Innovations in Technology and Optimization (ICRITO)*, 2024, pp. 1–5. doi: 10.1109/icrito61523.2024.10522418.
3. N.S. Harshit, N. K. Sahu, and H. R. Lone, "Eyes Speak Louder: Harnessing Deep Features From Low-Cost Camera Video for Anxiety Detection," in *Proc. ACM Conf.*, 2024, pp. 23–28. doi: 10.1145/3662009.3662021.
4. A. Tiwari, B. Matejek, and D. Haehn, "Non-Invasive Stress Monitoring From Video," in *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, 2024, pp. 1–5. doi: 10.1109/isbi56570.2024.10635725.
5. Y. Amirgaliyev, I. Krak, I. Bukenova, B. Kazangapova, and G. Bukenov, "Determining the psycho-emotional state of the observed based on the analysis of video observations," *Eastern-European Journal of Enterprise Technologies*, 2024. doi: 10.15587/1729-4061.2024.296500.
6. A. Kargarandehkordi and P.S.W. Vecilla, "Computer Vision Estimation of Stress and Anxiety Using a Gamified Mobile-based Ecological Momentary Assessment and Deep Learning: Research Protocol," *medRxiv*, 2023. doi: 10.1101/2023.04.28.23289168.
7. M. Khomidov, D. Lee, C. Kim, and J.-H. Lee, "The Real-Time Image Sequences-Based Stress Assessment Vision System for Mental Health," *Electronics*, vol. 13, no. 11, p. 2180, 2024. doi: 10.3390/electronics13112180.
8. K. Lee, P. Zhang, and S. Wu, "System and method for camera-based stress determination," 2019.
9. M. Penev, A. Manolova, and O. L. Boumbarov, "Active Shape Models with 2D profiles for Stress/Anxiety recognition from face images," in *Proc. Int. Conf. Communications*, 2014, vol. 1, pp. 108–112. [Online]. Available: http://e-university.tu-sofia.bg/e-publ/files/1697_CEMA14_Martin_Agata.pdf
10. H. Chandika, B. Soumya, B.N.E. Reddy, and B.M.S. SaiManideep, "Real-Time Stress Detection and Analysis using Facial Emotion Recognition," *Int. J. Adv. Res. Comput. Commun. Eng.*, 2024. doi: 10.17148/ijarcce.2024.13324.
11. J. Xu, C. Song, Z. Yue, and S. Ding, "Facial Video-Based Non-Contact Stress Recognition Utilizing Multi-Task Learning With Peak Attention," *IEEE J. Biomed. Health Inform.*, pp. 1–12, 2024. doi: 10.1109/jbhi.2024.3412103.
12. G. Giannakakis et al., "Stress and anxiety detection using facial cues from videos," *Biomed. Signal Process. Control*, vol. 31, pp. 89–101, 2017. doi: 10.1016/j.bspc.2016.06.020.
13. H. Chandika, B. Soumya, B. N. E. Reddy, and B. M. S. SaiManideep, "Real-Time Stress Detection and Analysis using Facial Emotion Recognition," *Int. J. Adv. Res. Comput. Commun. Eng.*, 2024. doi: 10.17148/ijarcce.2024.13324.
14. G. Orrù et al., "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neurosci. Biobehav. Rev.*, vol. 36, no. 4, pp. 1140–1152, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22305994
15. J. Singh and G. Goyal, "Decoding depressive disorder using computer vision," *Multimedia Tools Appl.*, vol. 80, pp. 8189–8212, 2021. doi: 10.1007/s11042-020-10128-9.
16. W.R. Ringwald et al., "Day-to-day dynamics of facial emotion expressions in posttraumatic stress disorder," 2024. doi: 10.31234/osf.io/6fqg9.
17. L.R. Enders, H. Roy, T. Rohaly, A. Jeter, and J. Villarreal, "Impacts of Posttraumatic Stress Disorder on Eye-Movement during Visual Search in an Open Virtual Environment under High and Low Stress Conditions," *J. Vis.*, vol. 23, no. 9, p. 5691, 2023. doi: 10.1167/jov.23.9.5691.
18. J. Gruber et al., "Associations between hypomania proneness and attentional bias to happy, but not angry or fearful, faces in emerging adults," *Cognition & Emotion*, vol. 35, no. 1, pp. 207–213, 2021. doi: 10.1080/02699931.2020.1810638.
19. J. Wang, Y. Song, H. Li, Z. Leng, M. Li, and H. Chen, "Impaired Facial Emotion Recognition in Individuals with Bipolar Disorder," *Asian J. Psychiatry*, vol. 102, p. 104250, 2024. doi: 10.1016/j.ajp.2024.104250.
20. R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?," *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008. doi: 10.1080/00140130701817062.
21. V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," *arXiv preprint*, arXiv:1907.05047, 2019. doi: 10.48550/arXiv.1907.05047.
22. Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," *arXiv preprint*, arXiv:1907.06724, 2019. doi: 10.48550/arXiv.1907.06724.

Information about authors

Aizhan Nurzhanova is a 1st year doctoral student in the Department of Computer and Software Engineering at L.N. Gumilyov Eurasian National University (Astana, Kazakhstan, nuraizhan87@mail.ru, +77028307620). Her research interests include video-based emotion recognition, facial expression analysis, and machine learning applications in mental health. ORCID iD: 0009-0006-9871-9823.

Miras Mussabek is a 1st year doctoral student in the Department of Computer Engineering at Astana IT University (Astana, Kazakhstan, miras.k@astanait.edu.kz, +77071771011). His research interests include video-based detection, recognition. ORCID iD: 0009-0009-2353-3524.

Dr. Gokhan Ince is an Associate Professor in the Computer Engineering Department, Faculty of Computer and Informatics Engineering at Istanbul Technical University (Istanbul, Turkey, gokhan.ince@itu.edu.tr, +90 (212) 285 69 86 ext: 6986). He has extensive experience in signal processing, affective computing, and human–robot interaction. ORCID iD: 0000-0002-0034-030X.

Dr. Mas Rina Mustaffa is an Associate Professor at the Faculty of Computer Science, University Putra Malaysia (Serdang, Malaysia, MasRina@ump.edu.my). Her research interests include pattern recognition, emotion detection, and deep learning for intelligent systems. ORCID iD: 0000-0001-5088-2871.

Dr. Ainur Zhumadillayeva is an Associate Professor in the Faculty of Information Systems, Department of Computer and Software Engineering at L.N. Gumilyov Eurasian National University (Astana, Kazakhstan, zhumadillayeva_ak@enu.kz, +77025295999). Her research focuses on machine learning, data mining, and educational technologies. ORCID iD: 0000-0003-1042-0415.

Submission received: 16 April, 2025.
Revised: 30 August, 2025.
Accepted: 30 August, 2025.