Hypomimia and Facial Action Units for Distinguishing Parkinson's Disease from Depression: A Foundation Model Approach

Justyna Skibińska
Faculty of Computer and Information Science, University of Ljubljana
Ljubljana, Slovenia
justyna.skibinska@fri.uni-lj.si

Teodora Matić
Faculty of Computer and Information Science, University of Ljubljana
Ljubljana, Slovenia
teodora.matic@fri.uni-lj.si

Aleksander Sadikov
Faculty of Computer and Information Science, University of Ljubljana
Ljubljana, Slovenia
aleksander.sadikov@fri.uni-lj.si

Vida Groznik
Faculty of Computer and Information Science, University of Ljubljana
Ljubljana, Slovenia
vida.groznik@fri.uni-lj.si

Dejan Georgiev
Faculty of Computer and Information Science, University of Ljubljana
Department of Neurology, University Medical Centre Ljubljana
Ljubljana, Slovenia
dejan.georgiev@kclj.si

Abstract

Parkinson's disease (PD) and depression share overlapping facial motor symptoms, particularly hypomimia, which complicates differential diagnosis when relying on subjective clinical rating scales. Quantitative, video-based analysis of facial Action Units (AUs) offers a promising, objective alternative, yet the only prior study to address this specific distinction is that of Matic [1]. This work extends that study by combining AUs extracted with the open-source, transparent tool OpenFace 2.0 with tabular foundation models (TabPFN and TabICL) to differentiate PD patients, depressed individuals, and healthy controls. The reading task combined with the TabPFN classifier achieved the highest binary classification accuracy of 0.78, slightly below that reported in the original study using the same task and AU features (0.83). The multiclass classification using TabPFN on the reading task reached an accuracy of 0.56, outperforming the 0.52 achieved previously. These results demonstrate that foundation models paired with open-source, interpretable feature extraction can match or exceed prior benchmarks while improving reproducibility and transparency.

Keywords

action units, artificial intelligence, depression, foundation models, Parkinson's disease, video analysis

1 Introduction

Parkinson's disease (PD) stands as the second most prevalent neurodegenerative condition worldwide [2]. Its frequency climbs steadily with advancing age, the disease is set to place growing strain on healthcare systems in the decades ahead [3]. Its clinical presentation is dual: a motor syndrome comprising tremor, bradykinesia, gait freezing, micrographia, and dysphagia [4], and a broad non-motor syndrome encompassing sleep disruption, mood disturbance, hallucinations, and cognitive impairment [5].

Reduced facial mobility, or hypomimia, is a particularly salient and clinically informative motor sign of PD [6]. It presents as a fixed, mask-like expression, infrequent blinking, flattened nasolabial folds, and visible asymmetry between the two sides of the face [7, 8]. Patients also show difficulty both producing and recognising emotional expressions in others [9]. Although blunted facial expression is a feature of depression [10], current findings suggest that facial hypomimia in PD is largely independent of depressive status and reflects a distinct mechanism [6], underscoring the need to disentangle the two conditions. Moreover, voluntary and spontaneous expressions further appear to draw on separate neural pathways, which argues for evaluating both [11].

A common strategy for quantifying such changes relies on the Facial Action Coding System, which decomposes expressions into discrete, anatomically grounded action units (AUs) tied to individual muscle movements [12]. Tracking the probability of them within time offers a finer-grained, muscle-level lens on expressivity and is well suited to capturing the subtle muscle activity deficits characteristic of hypomimia in PD [13].

Clinical artificial intelligence (AI) systems are being adopted across medicine to support diagnostic reasoning and guide individualised care decisions [14]. This matters particularly here because widely used clinical instruments—including the Movement Disorder Society-sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and standard depression scales—rely on subjective rater judgement [15], [16], whereas quantitative digital measures promise a more consistent way of teasing apart PD-related and depressive symptomatology [17].

The main contribution of this work is extending the only existing study on distinguishing PD from depression via hypomimia (Matic [1]) by applying foundation models combined with AUs extracted using the open-source, transparent tool OpenFace 2.0. The work achieves comparable binary classification performance (PD vs. depression) and improves multiclass classification accuracy (PD vs. depression vs. healthy controls (HC)) over the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
http://doi.org/10.70314/is.2026.skui.135

original study, while offering a more interpretable, and reproducible pipeline due to the use of open-source feature extraction.

2 Related Work

Hypomimia can be captured through a range of complementary modalities—static images, video sequences, electromyographic recordings, and optoelectronic motion capture—each offering a different resolution on facial movement deficits. Existing studies analysing hypomimia can broadly be categorised into two main categories: emotional and non-emotional assessments [18].

Facial features form the backbone of most AI-driven diagnostic support systems, extracted through several complementary strategies: geometric features derived from facial landmark positions, AUs capturing discrete muscle-level movements, and textural descriptors encoding surface appearance. Beyond these hand-crafted approaches, features can also be learned end-to-end through deep learning methods, which extract representations directly from raw image or video data [18].

To disentangle spontaneous, voluntary, and mimicked facial expressions, 80 participants were recruited: 30 with PD, 25 with depression, and 25 HC by Matic [1]. Participants performed five tasks: neutral still expression, spontaneous emotion, voluntary emotion, mimicked emotion, and reading emotionally-neutral text. iMotions software yielded seven emotion features alongside a complementary set of AUs features. A Random Forest classifier with leave-one-out cross-validation reached 0.68 accuracy on the three-way classification using spontaneous-expression emotion features, and 0.89 accuracy for distinguishing PD from depressed patients specifically, using emotion features from the reading task. Results using AUs were more modest, reaching 0.83 accuracy for the reading task when differentiating PD from depressed patients, while multiclass classification using the same task achieved 0.52 accuracy.

One study drew on video recordings from 109 PD subjects and 45 HC. Two diadochokinesia tasks—/bagada/ and /pataka/— were used in this study. From these recordings, the authors derived global features capturing both AUs intensities and their temporal derivatives, then fed these into a Random Forest classifier to distinguish PD from HC. Using 5-fold nested cross-validation, the approach achieved subject-level classification performance of up to 79.1 % balanced accuracy and an AUC of 83.7 % [19].

Tabular foundation models, including the Tabular Prior-data Fitted Network (TabPFN) and Tabular In-Context Learning (TabICL), have recently demonstrated strong predictive performance on small- and medium-sized tabular datasets, making them particularly well suited to medical and clinical applications where cohort sizes are often limited [20, 21]. Crucially, these methods dispense with the need for hyperparameter tuning. TabPFN is a transformer pretrained via in-context learning on millions of synthetic datasets [20]. TabICL combines a two-stage approach: a column-then-row attention architecture with a transformer to deliver efficient in-context learning on substantially larger datasets [21]. Applying such models to tabular features derived from facial video offers a promising approach for building robust classifiers from the modest cohort sizes typical of studies distinguishing PD from depression.

In this paper, we combine video-derived facial expressivity features with a tabular foundation models to distinguish individuals with PD from those with depression, aiming to provide an objective complement to clinical assessment. This work is an extension of the original research carried out by Matic [1], proposing the use of fully explainable hand-crafted features together with promising tabular foundation models. Such an approach could be explored further to provide full interpretability of the machine learning pipeline through the application of explainable AI.

3 Methodology

3.1 Experiment Description

The primary aim of this experiment was to assess the accuracy with which individuals with PD could be distinguished from depressed patients. Beyond this binary comparison, we further sought to differentiate among PD, HC, and depressed cases in a three-way classification. Classification was subsequently carried out using Leave-One-Out Cross-Validation: low-variance features were first eliminated, followed by selection of 500 features based on mutual information. TabPFN and TabICL served as the classifiers for this task.

This study used the data previously collected by Matic [1]. The dataset consisted of records from 30 PD patients, 25 individuals with depression, and 25 HC. Participants completed five tasks: neutral still-expression, spontaneous emotional expression, voluntary (posed) emotional expression, mimicked emotional expression, and reading. In the spontaneous task, participants viewed nine video fragments; in the voluntary task, they imitated happiness, anger, or surprise per on-screen instructions; and in the mimicked task, they replicated the emotion (anger, surprise, happiness, or neutral) shown in a displayed image of a man or woman. The videos were selected based on their ability to elicit an emotional response. A variety of videos was included to cover a range of emotions, both positive and negative in valence. Matic conducted a small pilot study to identify the stimuli that elicited the strongest emotional responses in participants. However, she did not focus on any particular emotion; rather, her goal was to elicit an emotional response that was as strong as possible, regardless of its valence or the specific emotion elicited.

3.2 Feature Extraction

Feature extraction proceeded in two stages. In the first stage, facial behavior was analysed frame by frame from the video recordings using the OpenFace 2.0 toolkit [22]. For each recording, this yielded multivariate time series of the intensities of seventeen facial AUs [23], bilateral eye gaze vectors and gaze angles [24], and head orientation expressed as Euler angles. In the second stage, each of these time series was treated as a 1-D signal. For this purpose, we used two feature extraction libraries: the Time Series Feature Extraction Library (TSFEL), which provides features from the statistical, frequency, spectral, and fractal domains, and Catch22, a set of 22 canonical time series features selected for their discriminative power and low redundancy. To further capture signal complexity and predictability, we additionally computed approximate entropy, sample entropy, and relative standard deviation for each signal.

Fractal analysis was included as part of the feature set because facial behavior is a non-stationary signal whose dynamics could not be fully captured by standard statistical or spectral descriptors. Fractal measures quantify self-similarity and structural complexity across time scales, which may reflect differences in how smoothly or irregularly facial movements unfold. These measures were computed directly from each time series (AU intensities, gaze, and head pose). The fractal measures extracted included

intr alia: Petrosian Fractal Dimension, Detrended Fluctuation Analysis, Hurst Exponent, Katz's Fractal Dimension, Normalised Length Density Fractal Dimension. Lyapunov exponents were additionally computed to characterise the divergence of neighbouring trajectories and thereby assess the chaotic behaviour of the signal.

Finally, demographic variables — namely gender and age — were incorporated into the analysis to control for potential confounding effects.

4 Results

The outcomes of classifying PD cases and depressed patients are presented in Table 1. The best results were obtained for the reading task using the TabPFN algorithm, with balanced accuracy and accuracy reaching 0.78, while sensitivity, specificity, and F1-score were 0.77, 0.80, and 0.79, respectively. The next most powerful exercises were mimicking neutral male emotion, maintaining a still face, and watching all movies. The multiclass classification was performed for the reading exercise, since it was found to be the most significant exercise based on the binary classification of PD and depressed cases. The obtained results were slightly better for the TabPFN than TabICL classifier and reached an accuracy of 0.56, a balanced accuracy of 0.56, a sensitivity macro of 0.56, a specificity macro of 0.78, and an F1-score macro of 0.56. The normalised confusion matrix for this classification is depicted in Fig. 1. The easiest class to be correctly recognised appeared to be PD, and it was successfully identified in 66.6 % of cases. Next, 56 % of HC cases were detected correctly, while patients suffering from depression were recognised correctly in 44 %.

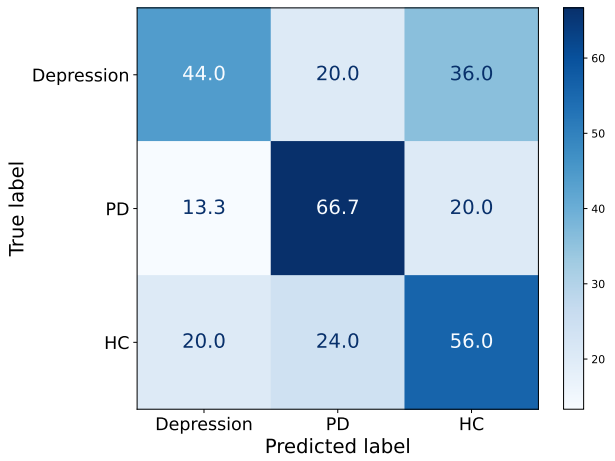

Figure 1: The Normalised Confusion Matrix for the Multiclass Classification.

5 Discussion

The most significant exercise for distinguishing between PD cases and depression with our settings was found to be the reading task. This exercise does not trigger as many emotions as, for instance, watching videos. This could be explained by the fact that AUs objectify the movements of muscle units. Moreover, a similar observation was reported in [1]. The accuracy obtained in this work for the reading task based on AUs was lower than in the original work [1] (0.78 vs. 0.83). This difference may be partly attributable to methodological differences between the two studies: AUs were extracted with OpenFace 2.0 in the present work and with iMotions [25] in the original study, and classification was performed with tabular foundation models here versus Random Forest in the original study. Because these factors changed simultaneously, and we did not evaluate them separately, we cannot determine the contribution of each. Nevertheless, for the multiclass classification, this paper presents higher accuracy for reading and AUs (0.56 vs. 0.52). The results achieved for the keeping still face and mimicking neutral man face exercises indicate that the facial expressions of persons with depression vary from those of PD patients. Finally, combining features from all exercises into a single feature set did not improve classification performance (accuracy = 0.53) compared with the best individual exercise (reading, accuracy = 0.78). This fact could suggest that the feature selection method used is not the best, or that the selected features are not appropriate for the TabPFN or TabICL classifiers. Moreover, the features were extracted here with OpenFace 2.0, whereas the features in [1] were generated by the iMotions software. The advantage of this presented approach is that OpenFace 2.0 provides explanations of how the algorithms work, whereas the iMotions software does not.

6 Conclusion

This study investigates the usability of the AUs and foundation models to differentiate PD patients, depressed individuals, and HC. Moreover, the open-source software OpenFace 2.0 was used to generate the features, which provides transparency to the approach. This work is an extension of the only prior work by Matic [1] in this field to distinguish PD cases from depression based on hypomimia. The differentiation of PD patients from depression reached its highest accuracy for the reading task and the TabPFN classifier, equal to 0.78. This approach did not outperform the results from the original work for the reading task and AUs (accuracy of 0.83) [1]. However, the accuracy for the multiclass classification obtained for the foundation model TabPFN, reading task, and AUs was higher than in [1] (0.56 vs. 0.52).

Possible future work includes investigating different feature selection methods together with the foundation models, computing emotion features using open-source software, likewise combining emotion and AU features to distinguish between PD patients and depressed people. Another direction of research is to check the potential of the support system methodology to distinguish PD cases from HC based on the same dataset.

7 Acknowledgments

This publication is supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-founded under the grant agreement No. 101081355. The operation (SMASH project) is co-founded by the Republic of Slovenia and the European Union from the European Regional Development Fund. Co-founded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the REA can be held responsible for them. This research was supported by the Slovenian Research and Innovation Agency (ARIS) under the Artificial Intelligence and Intelligent Systems Programme and AI for Science (Grant No. P2-0209 and GC-0001).

Table 1: The Results of the Distinction Between PD Patients and Depressed Individuals.

Task	Classifier	Baseline Accuracy	Accuracy	Balanced Accuracy	Sensitivity	Specificity	F1 score
All	TabPFN	-	0.49	0.48	0.57	0.40	0.55
All	TabICL		0.53	0.51	0.67	0.36	0.61
Video all	TabPFN	0.75	0.67	0.68	0.63	0.72	0.68
Video all	TabICL		0.69	0.69	0.70	0.68	0.71
Mimicked all	TabPFN	0.80	0.45	0.45	0.50	0.40	0.50
Mimicked all	TabICL		0.51	0.50	0.63	0.36	0.58
Neutral man	TabPFN	-	0.69	0.70	0.63	0.76	0.69
Neutral man	TabICL		0.56	0.55	0.67	0.44	0.63
Voluntary all	TabPFN	0.78	0.42	0.42	0.37	0.48	0.41
Voluntary all	TabICL		0.55	0.54	0.63	0.44	0.60
Reading	TabPFN	0.83	**0.78**	**0.78**	**0.77**	0.80	**0.79**
Reading	TabICL		0.76	0.77	0.70	**0.84**	0.76
Still face	TabPFN	-	0.69	0.69	0.67	0.72	0.70
Still face	TabICL		0.55	0.53	0.67	0.40	0.62

References

[1] Teodora Matić. 2026. *Objectivisation of Facial Expression Assessment in Parkinson's Disease and Depression.* PhD thesis. Univerza v Ljubljani, Filozofska fakulteta.

[2] Eduardo Tolosa, Alicia Garrido, Sonja W Scholz, and Werner Poewe. 2021. Challenges in the diagnosis of parkinson's disease. *The Lancet Neurology*, 20, 5, 385–397.

[3] E Ray Dorsey and Bastiaan R Bloem. 2018. The parkinson pandemic—a call to action. *JAMA neurology*, 75, 1, 9–10.

[4] Joseph Jankovic. 2008. Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry*, 79, 4, 368–376.

[5] K Ray Chaudhuri and Anthony HV Schapira. 2009. Non-motor symptoms of parkinson's disease: dopaminergic pathophysiology and treatment. *The Lancet Neurology*, 8, 5, 464–474.

[6] Edoardo Bianchini, Domiziana Rinaldi, Marika Alborghetti, Marta Simonelli, Flavia D'Audino, Camilla Onelli, Elena Pegolo, and Francesco E Pontieri. 2024. The story behind the mask: a narrative review on hypomimia in parkinson's disease. *Brain sciences*, 14, 1, 109.

[7] YANG Liqiong, CHEN Xiangling, GUO Quanhao, LUO Man, CHEN Xiaqing, WEN Yanxia, ZOU Xianwei, XU Fan, et al. 2022. Changes in facial expressions in patients with parkinson's disease during the phonation test and their correlation with disease severity. *Computer Speech & Language*, 72, 101286.

[8] Juan Diego Guerra-Hiraldo et al. 2023. Hemihypomimia in parkinson's disease: an under-recognized clinical sign? *Journal of Neurology*, 270, 1, 548–551.

[9] Margaret TM Prenger and Penny A MacDonald. 2018. Problems with facial mimicry might contribute to emotion recognition impairment in parkinson's disease. *Parkinson's Disease*, 2018, 1, 5741941.

[10] Jean-Guido Gehricke and David Shapiro. 2000. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research*, 95, 2, 157–167.

[11] William E Rinn. 1984. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 95, 1, 52.

[12] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental psychology & nonverbal behavior*.

[13] Anas Filali Razzouki et al. 2025. Explaining facial action units' correlation with hypomimia and clinical scores in parkinson's disease. *npj Parkinson's Disease*, 11, 1, 53.

[14] Eren Ogut. 2025. Artificial intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. *Clinics and practice*, 15, 9, 169.

[15] Christopher G Goetz et al. 2008. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23, 15, 2129–2170.

[16] Yuvaraj Krishnamoorthy, Sathish Rajaa, and Tanveer Rehman. 2020. Diagnostic accuracy of various forms of geriatric depression scale for screening of depression among older adults: systematic review and meta-analysis. *Archives of gerontology and geriatrics*, 87, 104002.

[17] Ana Karen Macias Alonso, Julian Hirt, Tim Woelfle, Perrine Janiaud, and Lars G Hemkens. 2024. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. *BMJ health & care informatics*, 31, 1, e100914.

[18] Anas Filali Razzouki, Laetitia Jeancolas, Dijana Petrovska-Delacrétaz, and Mounim A El-Yacoubi. 2025. Facial digital markers for hypomimia detection in parkinson's disease: a systematic review. *Pattern Recognition*, 112573.

[19] Anas Filali Razzouki et al. 2025. Leveraging action unit derivatives for early-stage parkinson's disease detection. *IRBM*, 46, 1, 100874.

[20] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. Tabpfn: a transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

[21] Jingang Qu, David HolzmÃžller, GaÃl Varoquaux, and Marine Le Morvan. 2025. Tabicl: a tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*.

[22] Baltrusaitis Tadas, Zadeh Amir, Lim Yao Chong, and Louis-Morency Philippe. 2018. Openface 2.0: facial behavior analysis toolkit. In *13th IEEE international conference on automatic face & gesture recognition*. IEEE NY, USA.

[23] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 6. IEEE, 1–6.

[24] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE international conference on computer vision*, 3756–3764.

[25] 2025. Imotions software. https://imotions.com/. Accessed: 2026-09-20. (2025).