

Post-Model Fusion of Speech, Drawing, and Movement Data to Classify Parkinson's Disease

Attila Zoltán Jenei

Department of Telecommunications and Artificial Intelligence
Budapest University of Technology and Economics
Budapest, Hungary
jenei.attila.zoltan@vik.bme.hu

Dávid Sztahó

Department of Telecommunications and Artificial Intelligence
Budapest University of Technology and Economics
Budapest, Hungary
sztaho.david@vik.bme.hu

Abstract—Parkinson's disease, a common movement disorder, remains incurable but benefits from early, accurate diagnosis to maintain quality of life. Since the symptoms at the early stage are heterogeneous, the detection is challenging, which highlights the investigation of several modalities. Speech, drawings, and specific movements are studied, and promising performances are shown to assist the doctors. This study investigates combining speech, drawing, and movement data to improve classification. Acceleration data from six movements and X-Y coordinates from Archimedean spiral drawings were processed in image and time-series representations while the speech was processed with x-vector technology. Support Vector Machine, Random Forest, and k-nearest Neighbors algorithms were trained and tested to classify 33 Parkinson's disease patients and 47 healthy controls. Next to the single modalities, post-model fusions were examined with all combinations of the modalities. Mann-Whitney U test was used to compare the performances of the models next to a 0.05 significance level. The speech significantly outperformed the drawing and movement activities. Furthermore, any combination of the modalities resulted in significantly better balanced accuracy than using movements and drawings alone. Lastly, the speech performed comparably with any combination of the modalities. In conclusion, speech data is a strong standalone predictor, while drawing and movement data enhance detection together when speech data is unavailable.

Keywords—acceleration, image, time series, multimodality, deep learning, machine learning, speech, Parkinson's disease

I. INTRODUCTION

Parkinson's disease (PD) is one of the most common neurodegenerative disorders, which affects the population mainly after the age of 60 [1]. The origin is unknown according to recent clinical knowledge, and the diagnosis is also problematic at the early stage. Therefore, solutions that help the diagnostic procedure are desirable in clinical practice.

The disease affects around 1% of people older than 60 years in the United States and 0.3% globally [1], [2]. The incidence of PD is low before 50 but rapidly increases with age. Furthermore, men are more likely to have the disease than women. The incidence ratio between men and women is 1.46 [2]. Other risk factors are genes, environment, and toxic chemicals [3].

PD is defined by the loss of dopamine-producing neurons in the substantia nigra pars compacta region of the brain [4]. The disease is also associated with Lewy-body aggregation. It was discovered that 50% of the dopaminergic neurons die already when the first motor symptoms appear [5].

Non-motor symptoms may appear years before the motor ones but are less homogenous among the population with PD. These can be constipation, urinary dysfunction, problems with

sleep, cognitive function, and mood. According to [6], sleep and mood disorders are the most frequent ones.

The disease is associated with bradykinesia, tremor, rigidity, and postural instability, which are primary motor symptoms [7]. Additionally, micrographia, gait disturbance, and speech problems also can appear (these are called secondary motor symptoms).

The diagnostic approach relies on the patient's history, observing motor and non-motor symptoms, excluding alternative causes, and doing drug tests. Even though it is a routine clinical practice, misclassification happens, ranging from 15% to 24% in cases. According to [8], the diagnostic accuracy was around 80% performed by movement disorder experts. Common misdiagnoses are between PD, essential tremor, and secondary parkinsonism. The diagnosis of the disease largely depends on the proper categorization of the symptoms, which requires the deep expertise of the examining physician. This recognition is problematic in the early stages and for professionals with less experience.

Since there is no direct diagnostic procedure for PD, supplementary tools may be needed to assist. Non-invasive biomarkers such as speech, handwriting, and movements are recorded and analyzed to capture motor disturbances [9]. These modalities have already reached high-performance with the help of machine learning algorithms. These are the most common, distinct modalities from the machine learning perspective while overlap is possible in the rating exercises at the neurologist (like drawing task in the movement modality). Recently, the joint applicability of these signals has been highlighted for several reasons, such as 1) it improves discrimination or 2) it helps to detect early signs [10].

Therefore, we recorded speech, drawing, and movement signals from PD patient to study their joint use of the classification. The drawing was chosen because it is less influenced by the writing style of the person such as the handwriting. Our contribution is a state-of-the-art framework to examine such signals and fuse modalities together with post-model prediction fusion.

After the Introduction, the Literature study summarizes the findings regarding PD classification using different modalities. Then, Methods will describe the dataset, signal processing and the model evaluation methods. The Results section details the outcomes after the model evaluation. The Discussion and Conclusion section will discuss and conclude the key findings.

II. LITERATURE STUDY

Generally, the acquired signals are cleaned/filtered before extracting features. Then, the training part of these features is

fed into the machine learning algorithm to obtain outcomes like labels or scores on the test data. Modalities like speech, handwriting, and movements support PD diagnostic procedures. The first subsection summarizes findings with single modalities, while the second summarizes findings with modality fusion.

A. Findings with single modalities

Modalities like speech, handwriting, and movements support PD diagnostic procedures. Multiple speech tasks are available to examine the PD state, such as sustained vowels, syllables, scripted texts, and monologues. Even though speech impairment is not the primary symptom, the high sensitivity of speech analysis makes it an effective early biomarker.

188 PD patients were used in the study [11] with 64 healthy controls (HC). Sustained /a/ vowels were recorded three times, and features like Mel Frequency Cepstral Coefficient (MFCCs), Vocal Fold Feature, and Wavelet Transform-based features were extracted. Naïve Bayes, k-Nearest Neighbor (k-NN), and XGBOOST algorithms were trained and tested. In the cross-validation setup, 98.75% accuracy was achieved.

Deep Neural Network (DNN) based classification was done on spectrograms using the PC-GITA and ItalianPVS datasets [12]. The two datasets together included 78 PD and 72 HC. On the PC-GITA dataset, modulated vowels outperformed all other speech exercises, resulting in a maximum of 92% accuracy. 96% accuracy was achieved with vowel /a/ and vowel /o/ in the ItalianPVS dataset.

In the study [13], 39 PD and 39 HC were used with multiple speech tasks from sustained vowels to monologue. The authors extracted features with e-capac and x-vector deep learning models and classified these features with leave-one-out cross-validation using a Support Vector Machine (SVM). They found that the x-vector outperforms the e-capac technology. Furthermore, the scripted text and the monologue achieve the best performances in classifying PD and HC labels.

Drawing different shapes like spirals, waves, or lines is common to detect tremors, bradykinesia, and muscle rigidity. The authors in the study [14] used the publicly available Parkinson's Disease Spiral Drawings dataset with 62 PD and 15 HC. They compared four signal processing methods: signal, visual, hand-crafted, and fusion. Random Forest (RF) was used as a classifier. The best result was a 93.0% F1 score on the signal-processed input. This method represented the signals as one-dimensional vectors and they were fed into a convolutional neural network (CNN) architecture.

In the study [15], spiral and wave drawings were used from 175 PD and 192 HC, merging two datasets together. They proposed the VGG19-INC hybrid transfer learning model to predict PD labels. They concluded that the results (98.5% accuracy) outperformed the state-of-the-art solutions.

DraWritePD and PaHaW datasets were used to classify PD in the study [16]. Both datasets included multiple drawing tasks. The authors calculated dynamic features with different CNN models. On the PaHaW dataset, they reported 84.7% accuracy as a maximum performance. They compared the results to other works like [17] with 62.8% accuracy, [18] with 93.8% accuracy, and [19] with 75.0% accuracy.

Acceleration data from different movement tasks are widespread in the classification of PD. Exercises based on clinical evaluation are chosen for machine learning examination since they were created to capture PD symptoms.

Uchitomi and his colleagues examined the walking movement of 46 PD and 44 HC [20]. They explore several augmentation techniques, retaining the raw signals as well. Without augmentation, 81.9% accuracy was achieved, which increased to 86.4% with rotation-based augmentation. The authors used CNN-based classification on short-time Fourier transform images.

8 PD and 18 HC were examined with mobile phone inner sensors [21]. They performed walking, standing, sitting, holding, and not wearing the phone. They applied SVM and regularized logistic regression on standard features from time series. The aim was to identify the different activities. The result showed that the activities were classified correctly, with 96.1% accuracy for HC and 92.2% for PD.

Finger-tapping tests were used with 55 PD and 65 HC in the study [22]. Performing either test, the best result was 0.9 AUC (of the area under the receiver operating characteristic curve). Combining the two tests, the result improved to 0.95.

B. Approaches with multimodality

Neuroimages (MRI and SPECT) and biological (CSF) features were combined in the study [23] with 72 PD and 59 HC. The authors performed feature-level and model-level fusion. After the experiments, 93.3% accuracy resulted for the feature-level and 81.4% for the model-level fusion.

42 HC and 40 PD were examined with MRI images and clinical data [24]. On the input data, feature extraction was performed with deep-learning models separately for the two modalities. They concluded that by supplementing MRI images with clinical data, improved classification can be achieved (up to 96.5% F1-score).

Vásquez-Correa et al. combined speech, handwriting, and movement data in a study involving 44 PD and 40 HC [25]. The dataset included six diadochokinetic (DKK) words, read sentences, a story, a monologue, 14 drawings and writings (captured using a Wacom Cintiq 13-HD), and walking tasks (recorded with the eGait system). Each recording session lasted approximately one hour per participant. A short-time Fourier transform was applied at transient points in the recordings and analyzed with CNN models. The test set yielded top accuracies of 92.3% for speech, 80.3% for walking, and 67.1% for handwriting. They examined fusion on feature maps, embeddings, and feature vectors of the three bio-signals. The fusion model outperformed all individual modalities, achieving 97.6% accuracy. The authors noted that most misclassifications occurred in patients at the early stages of PD. Additionally, they developed an Android application called Apkinson, which uses the phone's built-in microphone and inertial sensors to monitor PD symptoms [26]. The fusion of speech and movement data achieved over 95% accuracy.

Based on the literature, the individual modalities can result in high accuracy. However, they decline with early-stage PD. Multimodal fusion helps with this limitation. Yet, it still requires experiments since the multimodal examinations can be time- and resource-consuming. Our aim is to examine post-model fusion with rapid speech, drawing, and movement data.

III. METHODS

Fig. 1 summarizes the examination process. The different modalities are processed accordingly, then classification is performed on the extracted features. With the model predictions, post-model fusion was created to give the final prediction.

A. The Hungarian MultiPark dataset

The dataset includes three modalities: speech, drawing, and movements. 33 PD and 47 HC were recorded with a Lenovo Tab M10 tablet, a commercial passive tablet pen, and wrist-worn Meta-Sensors from MbiEntLab. All participants were informed in advance and gave informed consent. None of the parts of the recording sessions included invasive intervention.

Speech: Participants read the North Wind and the Sun scripted text displayed at once on the tablet in Hungarian language. The test took about 1 minute to complete. The text size was 34 scaleable pixels. The recordings were stored in 16 kHz sampling frequency, 16-bit quantization with Pulse-Code Modulation (PCM) format.

Drawing: Standard Archimedean spiral was drawn from the inside out, moving between the template lines. The template included four rounds to reach the outside. The width of the templates was a bit smaller than the width of the tablet. The sampling frequency was 110 Hz. X and Y coordinates with timestamps were stored. The test was performed with a passive stylus pen. The subjects performed the tests with their right hand.

Movements: Movements were selected from the Unified Parkinson's Disease Rating Scale (UPDRS) [27]. Gait (3.10), kinetic tremor of the hands (3.16), postural tremor of the hands (3.15), rest tremor amplitude (3.17), pronation-supination (3.6) movements of the hands were instructed to the subjects. These tasks were selected based on the literature and because they involve the hand; hence, they can be measured by wrist-worn sensors. The numbers in parentheses mark the task's number in the scale. The repetitions and the walking distance were halved due to the limited time and space (compared to what is in the rating scale). The instruction was the same as in the rating scale. The wrist-worn sensor captured acceleration data (X, Y, Z) with a 50 Hz sampling frequency and transferred data with Bluetooth. The subjects performed the tests with their right hand.

30 males and three females were in the PD class, with an average age of 63.4 and a 14.9 standard deviation. The average severity was 1.0 based on subtests of the UPDRS. These subtests were selected and recorded by the neurologist. The

subtests included the resting tremor (3.17), postural tremor (3.15), rigidity (3.3), and finger tapping (3.4) tasks. The numbers in parentheses mark the task's number in the scale. The medical diagnosis was given by the neurologist. Patients reported no speech problems and were allowed to use glasses/contact lenses. Patients were retained in the database who had all the tasks recorded and performed the tasks without any aid.

20 males and 27 females were recorded in the HC class. The average age was 56.7, with a standard deviation of 15.2. The HC participants were reported to have no neurological disease diagnosed.

B. Data pre-processing and feature extraction

Speech data was normalized to peak, and features were extracted with an x-vector time-delay feed-forward deep neural network [28]. The model was pre-trained on the Voxceleb dataset and was used (with the SpeechBrain toolkit (v.0.5.13) from the Huggingface website) for feature extraction without fine-tuning.

In the case of motion data, magnitude vectors were calculated from the X, Y, and Z signals after zero offset, and a fourth-order Butterworth filter was applied. The X and Y coordinates of the drawing data were similarly processed. Time Series Feature Extraction Library (TSFEL v0.1.4) was used to calculate 60 time-series-based features in statistical, temporal, and spectral domains. Next to the time-series representation, the Recurrence Plot [29] was also used to create an image representation from the magnitude vectors of the motion data. For this, `pyts.image.RecurrencePlot` class was used with default parameters.

The drawing data was also processed as a two-dimensional spiral image. Both the drawings and image representation of the movements were resized to 224x224 with 24-bit depth. MobileNet pre-trained CNN model [30] was used to extract features from the images. Keras API was used with the Tensorflow (v2.0.0) machine-learning platform to implement the model. The classification layer (Dense layer) was replaced with a GlobalAveragePooling2D layer to get a one-dimensional feature vector.

C. Classification and model evaluation

SVM, RF, and k-NN classification were applied to the extracted features in a 10-fold nested cross-validation setup. The outer cross-validation separated the test set, while the inner cross-validation loop split the remaining set into training and validation sets. The validation set was used to optimize the parameters of the classifier as follows: *kernel* (linear, rbf), *C* value (0.001, 0.01, 0.1, 1, 10, 100), and *gamma* (10, 1, 0.1,

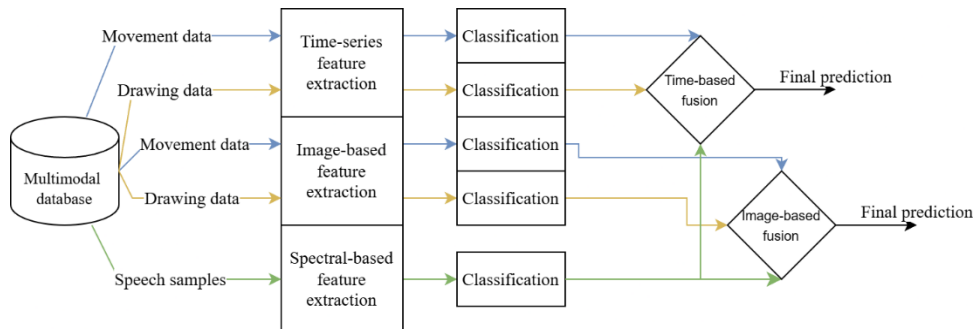


Figure 1. Examination setup: The signals were processed, then classification was performed. With the predictions, post-model fusion was created to get the final predictions.

0.01, 0.001, 0.0001) for SVM; the *max_depth* (10, 30, 50, 70, 90, None), *max_features* (auto, sqrt), and *min_samples_leaf* (1, 2, 4) for RF; the *number of neighbors* (2, 3, 5, 10, 20) for k-NN. Lastly, feature selection was done with Analysis of Variance (ANOVA) F-value using the best (highest feature importance) 100 features from the feature vectors.

After the classification, predictions between 0 and 1 were made for all participants per activity. This was converted into a label as if the prediction is above 0.5, then the subject is classified as PD, otherwise as HC. On these labels, sensitivity, specificity, and balanced accuracy were calculated to measure the model performances.

Using the decimal prediction values, post-model fusion was applied with the training and testing of SVM, RF, and k-NN in the same nested cross-validation setup. The input variables were the predictions from the different modalities, the output was the final label for the participants. This fusion technique was used with the Time-based and Image-based approaches separately for motion and drawing data supplemented with predictions with speech.

The following experiments were conducted in the classification of PD and HC: 1) examine speech, drawing, and all movements separately, 2) perform post-model fusion on movements, 3) perform post-model fusion pairwise on modalities, 4) perform post-model fusion on all the three modalities.

The performance of the classifiers is measured primarily on the balanced accuracy and compared with two sample Mann-Whitney non-parametric tests. The significance level was set to 0.05 based on the literature. The t-distributed stochastic neighbor embedding (t-SNE) was plotted with the model predictions of the modalities and colored by different properties.

IV. RESULTS

Table 1 summarizes the classification results achieved with the image representation approach. The markings *drawing*, *move*, and *speech* stand for the three modalities (experiment 1). The *m* denotes the post-model fusion of all movements (experiment 2). The signs *dm*, *sd*, and *sm* are the classification of pairwise modalities with post-model fusion (experiment 3). The abbreviations are the following: *dm* – drawing + move, *sd* – speech + drawing, *sm* – speech + move. The *sdm* represents the joint usage of all three modalities (experiment 4). Sensitivity (*sens*), specificity (*spec*), and balanced accuracy (*bacc*) are presented with average values and standard deviation.

The drawing and move modalities resulted in higher specificity than sensitivity, with an average difference of 21.1%. This difference is smaller but still exists with the speech and the post-model fusion of movements. With the post-model fusion of modalities, the difference between the sensitivity and specificity seems to be diminished. The highest balanced accuracy was reached with *sd* and *sdm*; however, *dm*, *sm*, and *speech* also approached this performance.

TABLE I. CLASSIFICATION RESULTS OF THE IMAGE REPRESENTATION APPROACH. ROWS MARK THE MODALITIES AND FUSIONS; COLUMNS SHOW THE SENSITIVITY, SPECIFICITY, AND BALANCED ACCURACY METRICS WITH STANDARD DEVIATION.

	sens		spec		bacc	
	mean	std	mean	std	mean	std
drawing	46.6%	27.7%	70.2%	17.9%	61.7%	13.7%
move	46.8%	14.7%	74.8%	8.5%	62.6%	9.3%
speech	88.9%	16.4%	94.8%	8.7%	91.7%	9.3%
m	71.8%	28.7%	83.4%	14.7%	76.7%	15.4%
dm	93.8%	11.6%	93.9%	9.3%	93.3%	8.4%
sd	93.9%	12.9%	96.0%	8.0%	94.6%	7.7%
sdm	95.2%	9.8%	95.3%	8.5%	94.6%	7.0%
sm	93.0%	12.0%	94.2%	10.3%	92.9%	9.5%

The joint use of at least two modalities appeared to be significantly better from the *drawing*, *move*, and *m* cases ($p < 0.000$). Moreover, the speech also appeared to be significantly better ($p < 0.000$) than the above-mentioned cases and be significantly not different ($p > 0.193$) from the fusion cases (*dm*, *sd*, *sm*, *sdm*). Case *m* is significantly better ($p < 0.000$) than the drawing and moving, but it is still significantly worse ($p < 0.000$) than the other cases (speech and joint modalities).

Table 2, similar to Table 1, summarizes the classification results achieved with time-based features (time representation). The notions and abbreviations are the same as in Table 1.

The average difference between the specificity and sensitivity is 24.9% in the drawing and move cases. With the speech modality, this difference is 5.8%, showing a more balanced trade-off. The post-model fusion brought a better balance between these two metrics. However, this difference is higher than in the image representation results. Here, the average difference is 7.9%, while it is 0.9% with image-represented signals.

The average balanced accuracy is 91.2% with time representation using post-model fusion on at least two modalities. This metric is 75.6% on average using the modalities separately. Yet, the speech modality reaches up to 91.7% balanced accuracy to a similar level as the joint

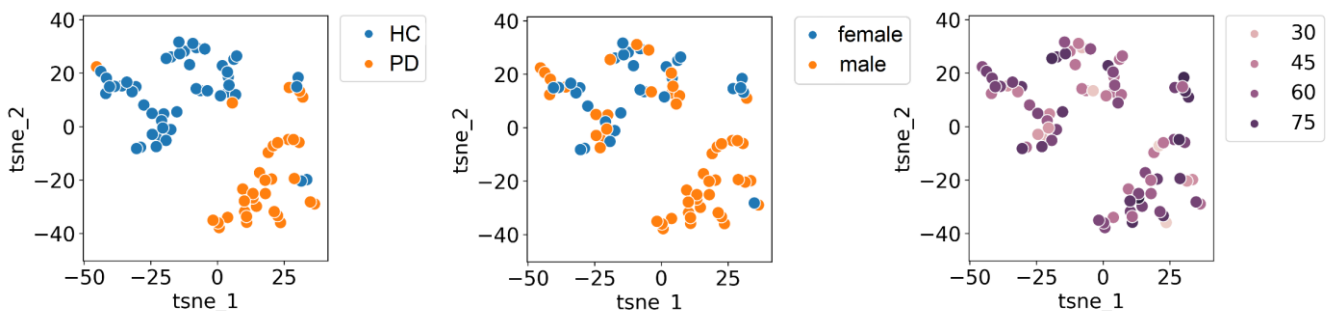


Figure 2. The t-SNE plots on the predictions from the modalities. The dots are colored by the original label (left), gender (middle) and age (right) of the participants.

modalities. The image representation had an average balanced accuracy of 93.9% on joint modalities and 73.1% on separate modalities.

TABLE II. CLASSIFICATION RESULTS OF THE TIME REPRESENTATION APPROACH. ROWS MARK THE MODALITIES AND FUSIONS; COLUMNS SHOW THE SENSITIVITY, SPECIFICITY, AND BALANCED ACCURACY METRICS WITH STANDARD DEVIATION.

	sens		spec		bacc	
	mean	std	mean	std	mean	std
drawing	49.4%	30.9%	72.7%	24.2%	62.5%	14.8%
move	54.4%	14.8%	80.9%	9.0%	69.5%	9.4%
speech	88.9%	16.4%	94.8%	8.7%	91.7%	9.3%
m	73.7%	24.3%	82.2%	20.6%	78.8%	18.6%
dm	85.3%	21.7%	95.3%	8.5%	91.7%	8.7%
sd	88.2%	19.8%	94.3%	11.0%	91.7%	11.3%
sdm	86.1%	20.8%	94.0%	10.5%	90.8%	10.7%
sm	86.4%	19.8%	94.0%	10.5%	90.8%	10.7%

The significance test gave results similar to those of the time representation compared to the image representation. Case *m* appeared to be significantly better ($p < 0.019$) from *drawing* and *move* but significantly worse ($p < 0.013$) than *speech* and the joint modalities. The post-model fusion of at least two modalities appeared to be better ($p < 0.000$) than the *drawing* and *move* cases but significantly not different ($p > 0.722$) from the *speech*.

Lastly, t-SNE plots were examined to determine any biases that occurred during the classification. Using the predictions from the modalities (*speech*, *drawing*, and *move*) as input features, two vectors were resulted and plotted in Figure 2. The plots are colored by the original labels, by the gender, and by the age of the participants. The two classes by the original labels can be seen on the left plot, which is linked to the high balanced accuracy of the post-model results. In the middle plot, a similar gender distribution can be seen as it is in the dataset. In the case of gender bias, there would be no heterogeneous aggregation. At last, the right plot shows the age distribution of the two classes, which seems to be homogenous in the figure. While this technique gives an insight into the data, may not reflect symmetrical bias and lack of quantitative measure.

V. DISCUSSION AND CONCLUSION

PD is one of the most common neurodegenerative disorders that cannot be cured, according to recent clinical knowledge. However, early detection gives the chance to sustain the quality of the patient by addressing appropriate treatment. Due to advances in technology, an improved classification of PD is realized in the literature. Modalities such as speech, drawings, and movements are promising to capture the motor symptoms since their examination may also be part of the patient’s visit.

The joint applicability of these modalities proved to be more selective and robust to detect PD, especially if the patient has mild symptoms that would be hard to notice by visual inspection. Since the symptoms at an early stage are diverse, find the optimal set of activities and recordings cumbersome.

In this study, three modalities and their post-model fusion were examined to classify PD next to HC. The recording setup was planned to be carried out in 15 minutes with minimal tools. The recordings can be acquired with a tablet and a wrist-worn sensor. During this, read scripted text, drawing of an

Archimedean spiral, and acceleration data from five different movements were recorded. They were preprocessed and described with features (with pre-trained models and descriptive features). The pre-trained features were not specific for PD since fine-tuning was not performed but was chosen because of the limited dataset. With these features, machine learning algorithms were trained and tested to classify PD and HC.

Modalities were examined individually and jointly with post-model prediction fusion. Two approaches were examined: image and time-series-based representation. At the former, RPs were created from the input signals, while time series features were extracted from the latter.

From the three modalities, the speech appeared to be the top, overtaking the drawing and movements. This difference was significant for both processing approaches. This can be explained as the speech is the most complex signal (and also the longest recording) among these modalities which can capture the symptoms better. Among the drawing and move cases, the move tended to be better regarding balanced accuracy; however, the results were not significant. Applying post-model fusion on movements resulted in a significant improvement compared to using the movement tasks separately. In this way, the model can handle that the early symptoms may not appear in one or two movement tasks (but in a collection of movements).

Lastly, more modalities with post-model fusion appeared to be better than using drawing or movement. This difference was significant for any combination (even when the comparison was made to the post-model fused movements). This supports the idea that multiple activities capture symptoms better than using one kind of modality. Nevertheless, the speech did not reach significantly different results in the joint modalities. This calls attention to that the speech as a single modality can reach the same performance as the combination of other modalities. Next to the signal complexity, the feature extractor algorithm can also play a vital role in the final results.

Comparing the result with Vásquez-Correa et al. work mentioned in the literature study, a comparable performance can be seen. They achieved an accuracy of 92.3% for speech, 80.3% for walking, and 67.1% for handwriting. The results in this study are similar since the highest performance was reached with speech (91.7% accuracy), then with movements and drawings (61.7%-69.5% accuracy). Their final result after the fusion was 97.6% accuracy, whereas in this study, it was 94.6% accuracy.

The challenge is the variety of the symptoms at the early stage. Namely, it is possible that the patient may not manifest symptoms in every modality or in every task per modality. This is reflected by the average performance of the individual movement tasks. However, the combination of multiple tasks or modalities leaves room for the algorithm to pick up the information needed to predict the right label. A further challenge is the high standard deviation, especially with drawing and the movements. This implies instability which may come from the early symptoms and the limited size of the dataset. Furthermore, the drawing and movement signals are simpler and shorter than the speech.

The limitation of the study is the imbalanced dataset, which is not available yet publicly and is still under development. Analyzing the models’ performances, we found

no considerable bias in the predictions. However, this technique is not a quantitative measure of the biases and may not reveal all of it. Therefore, repeating the study on other datasets with similar modalities can increase the reliability of the methods. As a possible future work, experimenting with other fusion techniques could extend these results by including state-of-the-art technologies and other modalities. Furthermore, analyzing the importance of different tasks would be interesting to understand early symptoms and control the weight in the final prediction. Lastly, experimenting with data augmentation would also be welcomed next to balancing the dataset.

ACKNOWLEDGMENT

The work was funded by the National Research, Development and Innovation Office – NKFIH, project K143075.

REFERENCES

- [1] R. Balestrino and A. H. V. Schapira, 'Parkinson disease', *Euro J of Neurology*, vol. 27, no. 1, pp. 27–42, Jan. 2020, doi: 10.1111/ene.14108.
- [2] T. A. Zesiewicz, 'Parkinson Disease', *CONTINUUM: Lifelong Learning in Neurology*, vol. 25, no. 4, pp. 896–918, Aug. 2019, doi: 10.1212/CON.0000000000000764.
- [3] D. Belvisi *et al.*, 'Risk factors of Parkinson disease: Simultaneous assessment, interactions, and etiologic subtypes', *Neurology*, vol. 95, no. 18, Nov. 2020, doi: 10.1212/WNL.0000000000010813.
- [4] D. K. Simon, C. M. Tanner, and P. Brundin, 'Parkinson Disease Epidemiology, Pathology, Genetics, and Pathophysiology', *Clinics in Geriatric Medicine*, vol. 36, no. 1, pp. 1–12, Feb. 2020, doi: 10.1016/j.cger.2019.08.002.
- [5] J. M. Fearnley and A. J. Lees, 'AGEING AND PARKINSON'S DISEASE: SUBSTANTIA NIGRA REGIONAL SELECTIVITY', *Brain*, vol. 114, no. 5, pp. 2283–2301, 1991, doi: 10.1093/brain/114.5.2283.
- [6] K. Berganzo *et al.*, 'Motor and non-motor symptoms of Parkinson's disease and their impact on quality of life and on different clinical subgroups', *Neurología (English Edition)*, vol. 31, no. 9, pp. 585–591, Nov. 2016, doi: 10.1016/j.nrleng.2014.10.016.
- [7] A. A. Moustafa *et al.*, 'Motor symptoms in Parkinson's disease: A unified framework', *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 727–740, Sep. 2016, doi: 10.1016/j.neubiorev.2016.07.010.
- [8] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, 'Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis', *Neurology*, vol. 86, no. 6, pp. 566–576, Feb. 2016, doi: 10.1212/WNL.0000000000002350.
- [9] G. Pahuja and T. N. Nagabhushan, 'A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection', *IETE Journal of Research*, vol. 67, no. 1, pp. 4–14, Jan. 2021, doi: 10.1080/03772063.2018.1531730.
- [10] L. M. Chahine and M. B. Stern, 'Parkinson's Disease Biomarkers: Where Are We and Where Do We Go Next?', *Movement Disord Clin Pract*, vol. 4, no. 6, pp. 796–805, Nov. 2017, doi: 10.1002/mdc3.12545.
- [11] A. Shrivastava, M. Chakkaravarthy, and M. Asif Shah, 'A Novel Approach Using Learning Algorithm for Parkinson's Disease Detection with Handwritten Sketches', *Cybernetics and Systems*, vol. 55, no. 8, pp. 2388–2404, Nov. 2024, doi: 10.1080/01969722.2022.2157599.
- [12] K. Bhatt, N. Jayanthi, and M. Kumar, 'High-resolution superlet transform based techniques for Parkinson's disease detection using speech signal', *Applied Acoustics*, vol. 214, p. 109657, Nov. 2023, doi: 10.1016/j.apacoust.2023.109657.
- [13] J. Attila Zoltán, V. Zalán, and S. Dávid, 'Comparative analysis of multiple speech tasks to recognise Parkinson's disease using pre-trained feature extractor embeddings', in *XX. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary: Szegedi Tudományegyetem, Jan. 2024, pp. 179–186. [Online]. Available: <https://rgai.inf.u-szeged.hu/sites/rgai.inf.u-szeged.hu/files/mszny2024%20%281%29.pdf#page=178.11>
- [14] M. Gil-Martín, C. Luna-Jiménez, F. Fernández-Martínez, and R. San-Segundo, 'Signal and Visual Approaches for Parkinson's Disease Detection from Spiral Drawings', *nldl*, vol. 4, Jan. 2023, doi: 10.7557/18.6809.
- [15] S. Saravanan, K. Ramkumar, K. Narasimhan, S. Vairavasundaram, K. Kotecha, and A. Abraham, 'Explainable Artificial Intelligence (EXAI) Models for Early Prediction of Parkinson's Disease Based on Spiral and Wave Drawings', *IEEE Access*, vol. 11, pp. 68366–68378, 2023, doi: 10.1109/ACCESS.2023.3291406.
- [16] X. Wang *et al.*, 'Comparison of one- two- and three-dimensional CNN models for drawing-test-based diagnostics of the Parkinson's disease', *Biomedical Signal Processing and Control*, vol. 87, p. 105436, Jan. 2024, doi: 10.1016/j.bspc.2023.105436.
- [17] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, 'Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease', *Artificial Intelligence in Medicine*, vol. 67, pp. 39–46, Feb. 2016, doi: 10.1016/j.artmed.2016.01.004.
- [18] M. Diaz, M. Moetesum, I. Siddiqi, and G. Vessio, 'Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and BiGRUs', *Expert Systems with Applications*, vol. 168, p. 114405, Apr. 2021, doi: 10.1016/j.eswa.2020.114405.
- [19] M. Diaz, M. A. Ferrer, D. Impedovo, G. Pirlo, and G. Vessio, 'Dynamically enhanced static handwriting representation for Parkinson's disease detection', *Pattern Recognition Letters*, vol. 128, pp. 204–210, Dec. 2019, doi: 10.1016/j.patrec.2019.08.018.
- [20] H. Uchitomi, X. Ming, C. Zhao, T. Ogata, and Y. Miyake, 'Classification of mild Parkinson's disease: data augmentation of time-series gait data obtained via inertial measurement units', *Sci Rep*, vol. 13, no. 1, p. 12638, Aug. 2023, doi: 10.1038/s41598-023-39862-4.
- [21] M. V. Albert, S. Toledo, M. Shapiro, and K. Kording, 'Using Mobile Phones for Activity Recognition in Parkinson's Patients', *Front. Neur.*, vol. 3, 2012, doi: 10.3389/fneur.2012.00158.
- [22] N. Akram *et al.*, 'Developing and assessing a new web-based tapping test for measuring distal movement in Parkinson's disease: a Distal Finger Tapping test', *Sci Rep*, vol. 12, no. 1, p. 386, Jan. 2022, doi: 10.1038/s41598-021-03563-7.
- [23] G. Pahuja and B. Prasad, 'Deep learning architectures for Parkinson's disease detection by using multi-modal features', *Computers in Biology and Medicine*, vol. 146, p. 105610, Jul. 2022, doi: 10.1016/j.combiomed.2022.105610.
- [24] V. Dentamaro, D. Impedovo, L. Musti, G. Pirlo, and P. Taurisano, 'Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database', *Sci Rep*, vol. 14, no. 1, p. 20941, Sep. 2024, doi: 10.1038/s41598-024-70165-4.
- [25] J. C. Vasquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Noth, 'Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach', *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019, doi: 10.1109/JBHI.2018.2866873.
- [26] J. R. Orozco-Arroyave *et al.*, 'Apkinson: the Smartphone Application for Telemonitoring Parkinson's Patients Through speech, Gait and Hands Movement', *Neurodegener. Dis. Manag.*, vol. 10, no. 3, pp. 137–157, Jun. 2020, doi: 10.2217/nmt-2019-0037.
- [27] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, 'The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations', *Movement Disorders*, vol. 18, no. 7, pp. 738–750, Jul. 2003, doi: 10.1002/mds.10473.
- [28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, 'X-Vectors: Robust DNN Embeddings for Speaker Recognition', in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.
- [29] L. C. S. Afonso *et al.*, 'A recurrence plot-based approach for Parkinson's disease identification', *Future Generation Computer Systems*, vol. 94, pp. 282–292, May 2019, doi: 10.1016/j.future.2018.11.054.
- [30] A. G. Howard *et al.*, 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', 2017, *arXiv*. doi: 10.48550/ARXIV.1704.04861.