Computing, Telecommunication and Control, 2023, Vol. 16, No. 1, Pp. 69–##.
Информатика, телекоммуникации и управление. 2023. Том 16, № 1. С. 69–##.

Intellectual Systems and Technologies
Интеллектуальные системы и технологии

Research article
DOI: https://doi.org/10.18721/JCSTCS.16106
UDC 004.85

IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS FOR PARKINSONIAN GAIT DATA

O. Unal[1] ✉ , *V.V. Potekhin*[2]

[1,2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ ogulunal@gmail.com

Abstract. In this study, we used the Physionet gait database and extracted gait features such as step/stride regularities and symmetries to build a classifier for Parkinson's disease (PD) subjects and healthy controls. We also improved the number of features using the mean and standard deviation of step times during their usual, self-selected pace for approximately 2 minutes on level ground. Extracted features were used in three different machine learning algorithms.

PD is a neurodegenerative disorder caused by the neurodegeneration of regions of the basal ganglia. Gait abnormality is one of the main symptoms of PD. Motor symptoms in Parkinson's disease cause a lack of control over movements and difficulty initiating muscle movements such as shuffling steps, quicker strides, or moving slower than expected for the corresponding age. The proposed approach can be used for the diagnosis of PD that can be automated or performed remotely.

Keywords: machine learning, supervised learning, Parkinson's disease, gait, feature analysis

Citation: Unal O., Potekhin V.V. Implementation of machine learning algorithms for Parkinsonian gait data. Computing, Telecommunications and Control, 2023, Vol. 16, No. 1, Pp. 69–78. DOI: 10.18721/JCSTCS.16106

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.16106
УДК 004.85

РЕАЛИЗАЦИЯ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДАННЫХ О ПОХОДКЕ ПРИ БОЛЕЗНИ ПАРКИНСОНА

О. Юнал[1] ✉ , *В.В. Потехин*[2]

[1,2] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

✉ ogulunal@gmail.com

Аннотация. Изучена база данных походки Physionet и рассмотрены такие характеристики походки, как регулярность шага и симметрии, для построения классификатора для пациентов с болезнью Паркинсона (БП) и здоровых людей контрольной группы. Увеличено количество функций, используя среднее значение и стандартное отклонение времени выполнения шагов в их обычном, самостоятельно выбранном темпе, в течение примерно двух минут на ровном месте. Извлеченные функции использованы в трех различных алгоритмах машинного обучения.

БП – нейродегенеративное заболевание, вызванное нейродегенерацией областей базальных ганглиев. Нарушение походки является одним из основных симптомов БП. Двигательные симптомы при болезни Паркинсона вызывают отсутствие контроля над движениями и трудности с инициированием мышечных движений, таких как шаркающие шаги, более быстрые шаги или движение медленнее, чем ожидалось для соответствующего возраста. Предложенный подход рекомендуется для диагностики БП, которая может быть автоматизирована или выполнена удаленно.

Ключевые слова: машинное обучение, контролируемое обучение, болезнь Паркинсона, походка, анализ признаков

Для цитирования: Unal O., Potekhin V.V. Implementation of machine learning algorithms for Parkinsonian gait data // Computing, Telecommunications and Control. 2023. Т. 16, № 1. С. 69–78. DOI: 10.18721/JCSTCS.16106

Introduction

Almost all neurodegenerative diseases of the brain begin insidiously and progress over the span of years. Parkinson's disease (PD) is a chronic and progressive neurological disorder that results in rigidity, tremor, postural instability, and slowness. The main symptoms appear gradually and worsen over time. As the disease progresses, people may have difficulty walking, mental and behavioral changes, depression, memory problems, sleep problems, and fatigue may also occur. PD is caused by selective cell death of dopamine-producing neurons in the brain. There are a variety of theories about what causes the neurons to die. Parkinson's disease has no cure, per se, and current treatments consist of externally supplying the dopamine that the dying neurons stop producing naturally [18].

The main problems of Parkinsonian gait are step length reduction with speed reduction impaired coordination, a decrease in the length of the step with an increase in the frequency of steps, the inability to produce effective steps at the beginning of walking, or complete cessation of steps during gait, and problems with dual tasking while gait. In this study, we extracted 12 different features using the Physionet database. Extracted features are correlated to Parkinsonian gait abnormalities therefore, we applied machine learning algorithms to make classification using a new dataset. Machine learning is a powerful technique for effectively analyzing data like gait signals. In this study, machine learning is used to analyze the gait data of the PD patients to classify them into "healthy" and "not healthy" classes based on extracted gait features.

Literature review

The human brain & neurotransmitters. Approximately one and a half kilogram organ that organizes every function to manage the body itself. The brain is also responsible to analyse the information as well as controlling emotions, intelligence, memory and movement. The main parts of the brain are Cerebrum, Cerebellum, and Brainstem [1].

Neurons are communicating through the body and communicate with one another to transmit signals. Although, neurons are not simply connected physically. Each neuron's end is a small gap that is a synapse and to communicate with the next cell, the signal has to be able to cross the space [5]. This process is known as neurotransmission. There are excitatory and inhibitory neurotransmitters. Excitatory will increase the likelihood that the neuron will fire an action potential while inhibitory will decrease it. Neurotransmitters and their functions are explained as [2]:

- Adrenaline (excitatory): also called epinephrine, it increases blood flow and heart rate.
- Norepinephrine (NE): also known as noradrenaline (excitatory) increases blood flow and attention.
- Dopamine (both excitatory and inhibitory): is mostly responsible for feelings of reward and pleasure. Low dopamine level is related to a specific disorders such as Parkinson's disease [3].
- Serotonin (inhibitory): feelings of happiness and well-being, stable mood, sleep cycle, and digestive system.
- Gamma-aminobutyric acid (GABA) (inhibitory): inhibits neuron firing in the Central Nervous System, high levels improve sleep quality and provide calming effect.
- Acetylcholine (excitatory): learning, memory, muscle contraction, awakening.
- Glutamate (excitatory): learning, memory, and creating new nerve pathways.
- Endorphins: natural pain killer, excitement, and exercise.

Neurodegeneration. Different brain diseases attack different brain regions such as: Cerebrovascular diseases (primarily arteries), Infectious diseases (various substrates), Demyelinating diseases (primarily myelin) and Neurodegenerative diseases (primarily neurons). Neurodegeneration is progressive loss of structure/function of neurons and the death of neurons. Each neurodegenerative disease of the brain is caused by the progressive accumulation of a specific protein inclusion (proteinopathy) [6]. Over time this accumulation becomes toxic to the brain leading to irreversible degeneration (death) of neurons and atrophy. There are many neurodegenerative diseases, some of them are; Alzheimer's, Huntington's and Parkinson's disease (see Fig. 1).

Parkinson's disease. Parkinson's disease, also called primary parkinsonism, paralysis agitans, or idiopathic parkinsonism is a degenerative neurological disorder that is characterized by the onset of tremor, muscle rigidity, slowness in movement (bradykinesia), and stooped posture [7]. The disease was first described in 1817 by British physician James Parkinson in his Essay on the Shaking Palsy [7]. In PD, neurons in the part of the brain called the basal ganglia to begin to die off and produce less dopamine causing dopamine levels to fall. As dopamine starts to decrease, signs and symptoms of Parkinson's disease begin to appear [4] (see Fig. 2).

There are various kinds of research aiming at early detection of PD. In [19], focuses on vocal detection of Parkinson's disease while in [20] their research focuses on finger alterations during keyboard usage. In [21], gait data were used however their research was based on stride time variability and swing time variability. In our study, we use mean, and variance to detect the variabilities and we applied the coefficient of variation that shows the degree of variability relative to the mean value of the data.

Our study provides another possible usage of machine learning classifiers for PD patient diagnosis that is cost-effective, and provides more information for PD.

Gait in Parkinson's disease. Gait is one of the neurological examinations for neurodegenerative diseases as well as for Parkinson's disease. Motor symptoms in PD come from a lack of control over movements and difficulty initiating muscle movements [8]. Some features of Parkinsonian gait are:

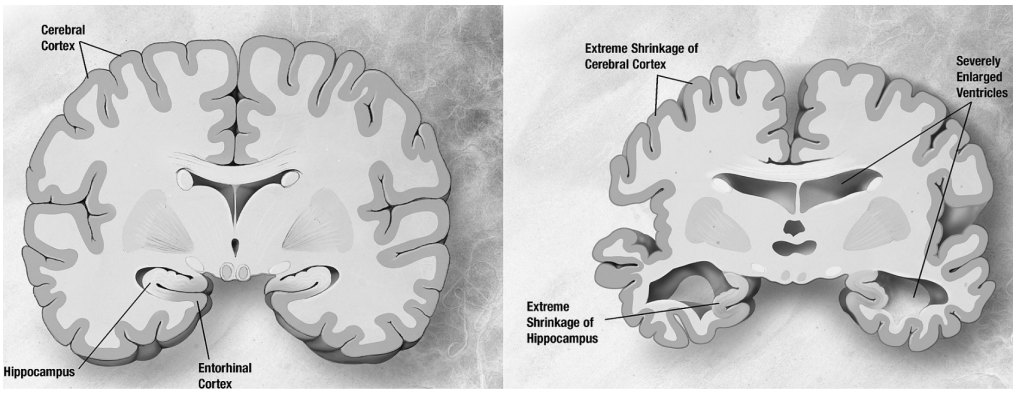

Fig. 1. Example diagram of normal and neurodegenerated brain

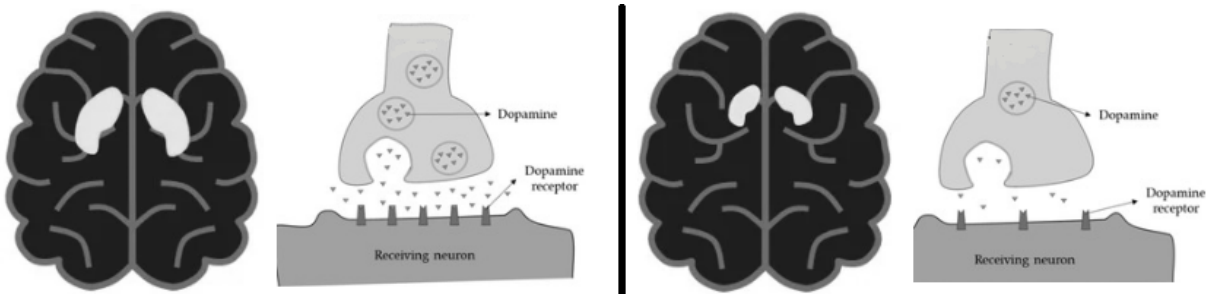

Fig. 2. SPECT scan and synaptic terminal of Healthy controls (left) and Parkinson patients on the right

- Taking small and shuffling steps.
- Move more slowly than expected for the corresponding age.
- Strides becoming quicker.
- Freezing of gait.

In this research, Gait data was used to analyse Parkinsonian gait characteristics, Afterward, Features related to gait data were extracted and a new dataset was created to test with different machine learning algorithms for classification. Extracted features were explained in detail in the Preprocessing section below.

Methods & materials

Physionet gait dataset. The database contains measures of gait from 93 patients with idiopathic PD (mean of age: 66.3 years; 63 % men), and 73 healthy controls (mean age: 66.3 years and 55 % men). The database includes the vertical ground reaction force (VGRF) records of subjects as they walked at their usual pace for 2 minutes on the level ground [9]. Underneath each foot were 8 sensors (Ultraflex Computer Dyno Graphy) that measure the force (in Newtons) as a function of time [9]. The output of each of these 16 sensors has been digitized and sampling was 100 per second [9]. In the database, two signals indicate the sum of the 8 sensor outputs for each foot. Among 8 sensors of each foot, sensors close to the toe were chosen for further preprocessing (1 sensor for the left and 1 for the right foot).

Column 1: Time (in sec).

Columns 2-9: Vertical ground reaction force on each of 8 sensors located under the left foot.

Columns 10-17: VGRF on each of the 8 sensors located under the right foot.

Columns 18,19: Total force under the left and right foot respectively. Ga, Ju, or Si in the database indicate the study from which the data originated:

Ga: Galit Yogev et al (dual-tasking in PD; Eur. J. Neuro, 2005);

Ju: Hausdorff et al. (RAS in PD; Eur. J. Neuro, 2007);

Si: Silvi Frenkel-Toledo et al. (Treadmill walking in PD; Mov. Disorders, 2005).

Feature extraction & preprocessing. A step is the movement made from one foot to the other while stride is a long step. The coefficient of variation (CV) is the ratio of the standard deviation to the mean value and shows the degree of variability relative to the mean value of the data. These features were generated using Physionet gait data in order to use in ML models. The higher the CV indicates greater variance. mean, standart deviation and CV formulas are as follows:

$$\mu = \frac{\sum_{i=1}^{n} xi}{n};$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (xi - \mu)^2}{n}};$$

$$Cv = \frac{\sigma}{\mu}.$$

Like many other motor symptoms that can occur in patients with PD, gait changes and freezing are caused by a loss of control in areas of the brain responsible for producing intentional movements. There are mainly 2 different gait disturbances that are episodic and continuous [10]. The episodic gait disturbances are freezing of gait and hesitation, the continuous changes indicates alterations in the walking pattern that appear, at least at first glance, to be more or less consistent from one step to the next, i.e., they persist and are apparent all the time [10]. Step, stride regularity and step symmetry features were generated to analyse and compare differences between PD patients and healthy controls.

Coordination of steps has also been shown to be dysfunctional in those with PD during gait therefore step times in PD for subjects in the database were used [11]. In the Fig. 3 represents peak detection using gait data. Peak times and peak values were also used to derive other features (step/stride regularity and step symmetry).

Gait data can be analyzed by unbiased autocorrelation procedures to give cadence, step length and measures of gait regularity and symmetry [12]. Step timing information can be calculated using the peak

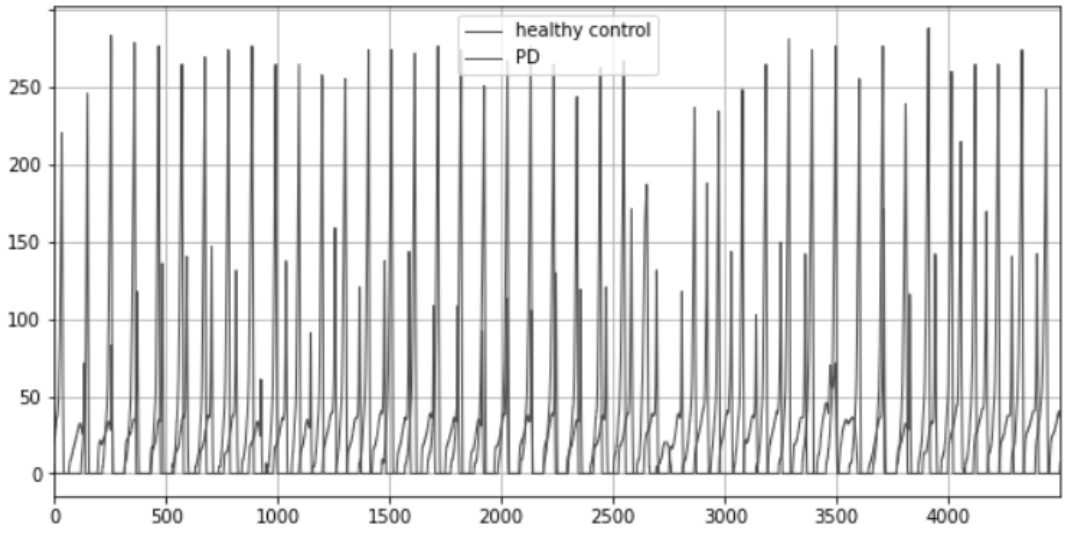

Fig. 3. Comparison of heathy and PD signals

times in raw gait data. Therefore, three features were generated that are: mean time between steps, standard deviation of step time, and the coefficient of variation (see Fig. 4).

As a result of feature extraction, there were 12 new columns representing step regularity, stride regularity, step symmetry, step time mean, step time standard deviation, and step time coefficient of variation for left and right leg electrodes. Ga, Ju, and Si studies were tested individually. This means the new dataset for Ga (dual tasking) has 113, Ju (Rhythmic Auditory Stimulation) has 129 and Si has 64 (Treadmill walking) rows that are extracted from each trial in the Physionet dataset. Thus, the dataset has 214 PD and 92 healthy controls. Below generated features are shown using the seaborn library in python programming language (see Fig. 5).

Machine learning algorithms

Random forest (RF). RF is an ensemble learning method, namely a random forest model consisting of a large number of small decision trees (estimators) that creates their predictions. The RF combines the result to obtain accuracy [13]. A random forest builds trees in parallel (bagging in ensemble learning) while boosting builts trees sequentially. Hyperparameters chosen for RF algorithm: max_depth (longest path between the root and the leaf node) = 3, max_features (RF is allowed to try in an individual tree) = 1, min_samples_leaf (minimum number of samples required to split an internal node) = 3.

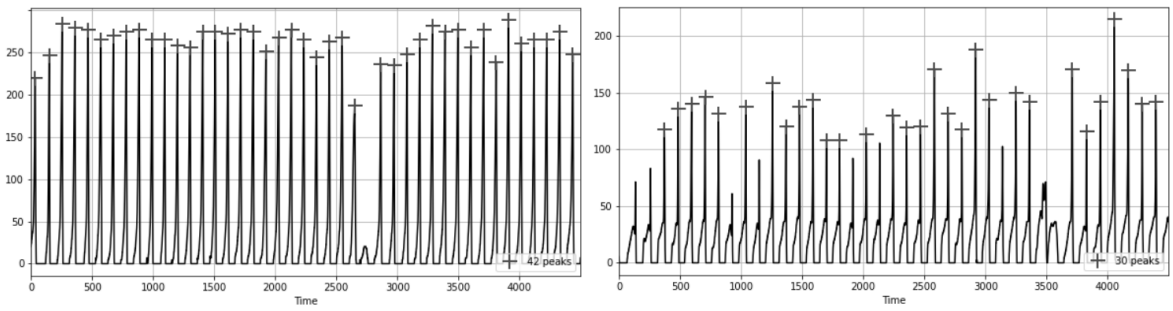

Fig. 4. Detected peaks of healthy controls and PD patients during 45 seconds interval

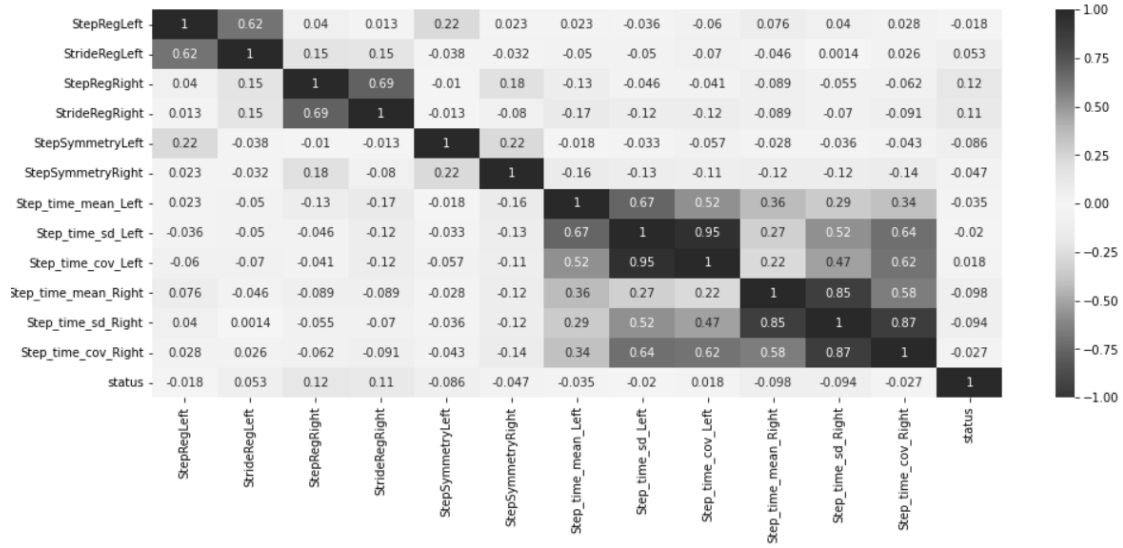

Fig. 5. Correlation Heat map of features

The model gave best performance using RF that construct many individual decision trees at training. Predictions from all trees are pooled to make the final prediction, to calculate Gini importance:

$$ni_j = w_j c_j - w_{left(j)} c_{left(j)} - w_{right(j)} c_{right(j)}.$$

Afterwards, importance for each feature on a decision tree is then calculated:

$$fi_j = \frac{\sum j \; Node \; j \; splits \; on \; feature \; i \; Nij}{\sum k \in all \; nodes \; ni_k}.$$

The final feature importance, is the average over all the trees:

$$RFfi_j = \frac{\sum j \in all \; trees^{normfi_{ij}}}{T}.$$

KNN. K-Nearest Neighbour is a supervised learning algorithm. KNN measures the similarity between the new case/data and training cases therefore the algorithm can classify the new case into the categories. KNN can be used for regression tasks or classification tasks. The default number of neighbors (five) was chosen.

AdaBoost. AdaBoost is an ensemble learning method, namely, an AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases [14].

Results

Selected algorithms (RF, KNN, AdaBoost) were applied to three different data studies (Ga, Ju and Si). Each of them gave a different result. After different combinations, among 12 extracted features, only six of them were used in algorithms. Those are: StrideRegLeft, StrideRegRight, StepSymmetryLeft, StepSymmetryRight, Step_time_mean_Left Step_time_sd_Left gave highest accuracies for each data study. Random forest gave the highest accuracy for each task of binary classification. Note that this result indicates 100 % accuracy and Healthy vs PD which are presented in the matrix confusion, along with evaluation formulas below using test data (see Fig. 6).

Accuracy (TP + TN/(TP + TN + FP + FN)).

Precision = TP/(TP+FP).

Recall = TP/(TP+FN).

F1 Score = 2*(Recall * Precision) / (Recall + Precision).

Ga is dual-tasking for healthy controls and PD patients. Ju's study contains PD records using Rhythmic Auditory Stimulation (RAS) which is a treatment technique for patients. RAS is designed to improve gait by providing the patient with auditory cues throughout the gait. Treadmill walking can improve gait stability in patients who have Parkinson's disease [15]. Combining Ga, Ju, and Si studies provided the worst result with high false positives for each algorithm. Different ML models were created and results were compared for each different Physionet database study (Ga, Ju, Si).

Patients having PD are supposed to have reduced step length people with Parkinsonian gait usually take small and shuffling steps. It may be difficult for them to lift their legs [16]. However, Parkinsonian gait is not the only symptom. There are other symptoms such as tremor, which usually begins in the hand and is more likely to occur when the limb is relaxed. Other symptoms are slowness of movement, Bradykinesia, and rigidity. It is also difficult to understand the common characteristics of gait because it might be differ-

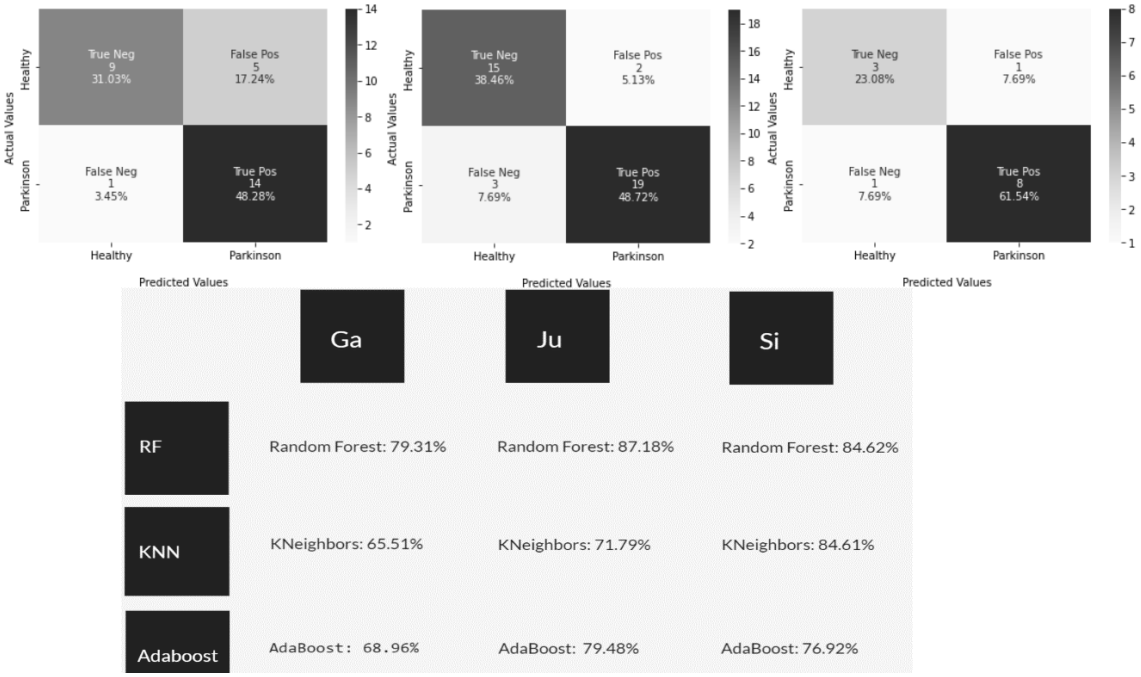

Fig. 6. Test accuracy results of algorithms and confusion matrix of RF using test data of Ga, Ju and Si respectively

ent for each person. As a result, using gait data alone might not be enough to classify or analyze PD using machine learning.

Because of having small datasets, we did not use neural networks in this study. Neural networks require a bigger size of data to achieve better results. The random forest algorithm provides a higher level of accuracy in predicting outcomes than the decision tree algorithm. RF performs well when there are imbalanced datasets. RF has methods for balancing errors in class population unbalanced data sets. Random forest minimizes the overall error rate when there is a data set. Thus, the larger class will get a low error rate while the smaller class will have a higher error rate. Hyperparameters of RF were tuned (explained in the previous section) KNN is mainly based on feature similarity. It is sensitive to outliers and noise. The performance of the KNN algorithm gets worse as the number of features increases. Another algorithm, AdaBoost, is another ensemble learning algorithm that is also sensitive to outliers and noise [17]. As a result, among chosen algorithms, RF gave the highest accuracies.

Conclusions

In this study, we presented a method to classify Parkinsonian gait data using extracted features from the Physionet gait database. Gait disorder is one of the main symptoms of Parkinson's disease. Features based on gait characteristics were extracted and we used supervised learning algorithms to implement three machine learning models based on three datasets. The proposed models gave the highest accuracies using Random forest compared to KNN and AdaBoost algorithm. The results of our approach can be further improved by implementing it on a larger dataset. The classification results based on gait data can be improved by combining patients' EEG and EMG data. As future work, EEG data of PD patients and healthy controls will be used for preprocessing and ML algorithms. Parkinson's disease is a neurodegenerative disease thus, it can be helpful to observe the effects of dopamine alterations in the brain using EEG signals. Another future work will be using EMG data to analyze muscle signals during the tremor stages. The gait data analysis and implemented ML models for the classification of Parkinson's disease can provide support for patients to improve their quality of life.

Acknowledgments

Firstly, I would like to thank my father Dr. Turgay Ünal for his help in Parkinson studies also would like to thank Associate professor Vyacheslav Potekhin for all his help during my studies. Finally, I thank my mom Dr. Seher Gülçin Ünal for her assistance in giving all the support in my life.

REFERENCES

1. **Marieb H., Hoehn K.** *Brain anatomy adapted from human anatomy & physiology*. 9th ed., Pearson, 2009, Vol. 9.

2. **Marianne K., Daniella S., Ariel R., David N., Jackson C., Ricardo G.** Dopamine: Functions, signaling, and association with neurological diseases. *Cellular and Molecular Neurobiology*, 2019, Pp. 1–30.

3. **Meder D., Herz D.M., Rowe J.B., Lehéricy S., Siebner H.R.** The role of dopamine in the brain – lessons learned from Parkinson's disease. *NeuroImage*, 2019, Pp. 79–93.

4. **Sergei G.** Dopamine function and the efficiency of human movement. *Journal of Cognitive Neuroscience*, 2016, Pp. 1–10.

5. **Potekhin V.V., Ünal O.** Analysis of emotions using EEG data and machine learning. *Proceedings of the 32nd DAAAM International Symposium*, Publ. by DAAAM International, Vienna, Austria, 2021, Pp. 158–167.

6. **Li D., Liu C.** Conformational strains of pathogenic amyloid proteins in neurodegenerative diseases. *Nat. Rev. Neuroscience*, 2022.

7. Parkinson disease. *Encyclopedia Britannica*, 21 Jan. 2021, *https://www.britannica.com/science/Parkinson-disease* (Accessed 2 June 2022).

8. **Lazzaro D.B.** Gait analysis in Parkinson's disease. *Overview of the Most Accurate Markers for Diagnosis and Symptoms Monitoring, Sensors*. Basel, Switzerland, 2020, Vol. 20.

9. **Goldberger A., Amaral L., Glass L., Hausdorff J., Ivanov P., Mark R., Stanley H.E.** *PhysioBank, PhysioToolkit and PhysioNet: Components of a new research resource for complex physiologic signals*, 2000, Pp. 215–220.

10. **Hausdorff J.M.** Gait dynamics in Parkinson's disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling. *Chaos*. Woodbury, N.Y., 2009. DOI: 10.1063/1.3147408

11. **Plotnik M., Giladi N., Hausdorff J.M.** Bilateral coordination of walking and freezing of gait in Parkinson's disease. *Eur. J. Neurosci*, 2008. DOI: 10.1111/j.1460-9568.2008.06167.x

12. **Nilssen R., Helbostad L.J.** Estimation of gait cycle characteristics by trunk accelerometry. *Journal of Biomechanics*, 2004, Vol. 37.

13. **Ali J., Khan R., Ahmad N., Maqsood I.** Random forests and decision trees. *International Journal of Computer Science Issues*, 2012.

14. **Freund Y., Schapire R.** *A decision-theoretic generalization of on-line learning and an application to boosting*, 1995.

15. **Earhart G.M., Williams J.** Treadmill training for individuals with Parkinson disease. *Physical Therapy*, 2012, Pp. 893–897. DOI: 10.2522/ptj.20110471

16. **Noh B., Youm C., Lee M., Cheon S.M.** *Gait characteristics in individuals with Parkinson's disease during 1-minute treadmill walking*, 2020. DOI: 10.7717/peerj.9463

17. **Chengsheng T., Liu H., Bing X.** AdaBoost typical algorithm and its application research. *MATEC Web of Conferences*, 2017.

18. **López-Grueso M.J., Padilla C.A., Bárcena J.A.** Deficiency of Parkinson's related protein DJ-1 alters Cdk5 signalling and induces neuronal death by aberrant cell cycle re-entry. *Cell Mol. Neurobiology*, 2022. DOI: 10.1007/s10571-022-01206-7

19. **Parisi L., RaviChandran N., Manaog M.L.** Feature-driven machine learning to improve early diagnosis of Parkinson's disease. *Expert Systems with Applications*, 2018, Vol. 110, Pp. 182–190. DOI: 10.1016/j.eswa.2018.06.003

20. **Giancardo L., Sánchez-Ferro A., Arroyo-Gallego T., Butterworth I., Mendoza C.S., et al.** *Computer keyboard interaction as an indicator of early Parkinson's disease*, 2016. DOI: 10.1038/s41598-018-32121-x

21. **Herman T., Gruendlinger L., Peretz C., Frenkel-Toledo S., Giladi N., Hausdorff J.M.** *Effect of gait speed on gait rhythmicity in Parkinson's disease: Variability of stride time and swing time respond differently*, 2005. DOI: 10.1186/1743-0003-2-23

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Юнал Огюл
Ogul Unal
E-mail: ogulunal@gmail.com

Потехин Вячеслав Витальевич
Vyacheslav V. Potekhin
E-mail: Slava.Potekhin@spbstu.ru
ORCID: https://orcid.org/0000-0001-9850-9558

Submitted: 24.11.2022; Approved: 10.05.2023; Accepted: 17.05.2023.
Поступила: 24.11.2022; Одобрена: 10.05.2023; Принята: 17.05.2023.