UCD 616.858-008.6:004.8:004.94

COMPREHENSIVE INTEGRATION OF MACHINE LEARNING AND CLINICAL DATA FOR CLASSIFICATION OF EARLY AND LATE STAGES OF PARKINSON'S DISEASE

I. M. Liakh[1], R. Ya. Zhovtani[2], M. Yu. Kashka[3], A. Yu. Tsipino[4]

[1] Uzhhorod National University, 3 Narodna Sq., Uzhhorod, 88000, Ukraine, https://orcid.org/0000-0001-5417-9403, e-mail: igor.lyah@uzhnu.edu.ua
[2] Uzhhorod National University, 3 Narodna Sq., Uzhhorod, 88000, Ukraine https://orcid.org/0000-0002-7421-148X, e-mail: ruslana.zhovtani@uzhnu.edu.ua
[3] Uzhhorod National University, 3 Narodna Sq., Uzhhorod, 88000, Ukraine, https://orcid.org/0000-0001-7437-6156, e-mail: mariia.kashka@uzhnu.edu.ua
[4] Uzhhorod National University, 3 Narodna Sq., Uzhhorod, 88000, Ukraine https://orcid.org/0009-0007-1704-2732, e-mail: tsipino.artemii@student.uzhnu.edu.ua

The purpose of this study is to evaluate the effectiveness of public health information campaigns, determine their impact on public awareness, and formulate recommendations to optimise communication strategies. The mPower Public Researcher Portal open dataset (BrianMBot, 2015), which contains information on participants' behaviour and activity in a mobile application study for Parkinson's disease patients, was used for the analysis. The data was processed in a Python environment using statistical analysis and machine learning libraries.

Descriptive statistics, correlation and regression analysis, and clustering were used to assess the effectiveness of information campaigns to identify groups of participants with different levels of engagement and information perception. The study showed that campaigns with interactive online tools and multimedia content have the highest level of reach and impact on population behavioural changes. Traditional methods, such as print and radio, demonstrate lower effectiveness, especially among younger audiences. In addition, a significant difference in information perception was found across regions and respondents' socio-demographic characteristics. The effectiveness of public health information campaigns largely depends on interactivity and adapting content to the specifics of target audiences. It is recommended to apply a combination of digital and traditional communication channels, tailored to regional specifics and population needs.

Keywords: *Parkinson's disease, stage prediction, machine learning, Python, Random Forest, XGBoost, LightGBM, clinical data, automated inference, binary classification.*

Problem Statement. Parkinson's disease (PD) remains one of the leading neurodegenerative disorders with a rising global prevalence, including Ukraine. Modern research covers a wide range of PD aspects – from epidemiology and the effectiveness

of surgical interventions to non-motor manifestations and cognitive impairments – while machine learning is increasingly used to predict disease progression and support clinical decisions.

Given the complexity and heterogeneity of clinical and molecular data, modern approaches combine traditional diagnostic methods with machine learning algorithms and hybrid models, which allows not only to increase the accuracy of disease stage classification but also to identify patients at high risk of complications, predict response to therapy, and improve personalised PD management. There is an urgent need to develop optimised machine learning models (particularly Random Forest-based) that can effectively process multidimensional data to support clinical decision-making and personalise therapy, an area that is currently underdeveloped.

Analysis of Recent Research and Publications. In modern Parkinson's disease research, machine learning methods are increasingly used to predict disease stages and support clinical decisions. The use of programming environments, such as Python, in conjunction with ML libraries enables processing large, heterogeneous clinical data, deriving informative features, and improving the accuracy of models for identifying patients at early and late stages.

Particular attention is paid to automated inference for new patients, ensuring the integration of data preparation, scaling, and the encoding of categorical variables into a single pipeline. This approach not only improves models' generalisation but also makes the system more practical for clinical use, reducing decision-making time and minimising the risk of errors. These aspects lay the groundwork for a detailed review of current literature on methods for predicting Parkinson's disease progression.

Article [1] presents data on the prevalence of Parkinson's disease in Ukraine for the period 20102020, including information during the COVID-19 pandemic. The authors show that the overall incidence increased from 59.6 to 67.5 cases per 100,000 population between 2010 and 2017, and data for the Kyiv region for 2019-2020 indicate fluctuating rates, reflecting the impact of healthcare accessibility and diagnostic quality.

Article [2] evaluates the results of surgical treatment for Parkinson's disease in Ukraine, specifically radiofrequency destruction of subcortical nuclei and implantation of deep-brain stimulation (DBS) systems. The authors show high effectiveness of interventions in reducing tremor and rigidity, with DBS demonstrating the best results among all methods, while combined and bilateral interventions require careful patient selection due to an increased risk of complications.

In article [3], a machine learning-based clinical tool is presented to support the diagnosis of mild cognitive impairment (MCI) in Parkinson's disease patients. The model, based on long-term observational data, demonstrated noninferior performance compared with the standard clinical test (MDS PD-MCI Level II) and detected a subgroup of MCI patients not identified by the clinical test, highlighting the potential of ML to complement traditional diagnostic approaches.

In article [4], the impact of non-motor symptoms on the quality of life of Parkinson's disease patients was investigated using machine learning. The authors showed that non-motor symptoms are key predictors of PDQ-39 scores, and ML models can effectively

predict overall HRQoL and cognitive aspects, emphasising the importance of a comprehensive approach to PD treatment.

The study [5] proposed a hybrid cancer diagnosis model based on gene expression data, combining spectral inductive clustering, Random Forest, CNN, and alternative voting for the final decision. The model's methods are effective in identifying interconnected clusters and multi-class classification, making them potentially applicable for predicting Parkinson's disease progression and analysing complex clinical and molecular data.

In article [6], a systematic analysis of 178 clinical studies aimed at evaluating cognitive impairments in Parkinson's disease was conducted, with an emphasis on trial design, observation duration, and metrics used. The authors found that only a small fraction of completed studies demonstrated short-term improvement in cognitive functions, indicating a need to improve methodology, sample selection, and assessment tools in future clinical trials.

In article [7], the incidence rates of Parkinson's disease in five large North American epidemiological cohorts in 2012 were analysed, encompassing over 6.7 million person-years of observation. The results showed significant variability in incidence rates by age, sex, and geographical location; rates increased with age and were higher among men, underscoring the need to elucidate risk factors and case registration methods.

Article [8] investigated, for the first time, the ability of patients with Parkinson's disease to perform affective forecasting – predicting their emotional reactions to future positive and negative situations. The results revealed no statistically significant differences between the PD group and neurotypical controls, indicating that emotional self-forecasting mechanisms remain intact despite expected impairments in future self-projection.

Work [9] proposes an approach to predicting Parkinson's disease progression based on temporal MDS-UPDRS data and machine learning, specifically by reformatting time-series data for supervised learning. The Multi Layer Perceptron model demonstrated the best results, outperforming Random Forest on MAE and MSE, which confirms the effectiveness of neural network approaches for modelling the dynamics of PD progression.

Article [10] explores the use of machine learning to predict Parkinson's disease progression by leveraging clinical and molecular data from the AMPPD database, which encompasses over 10,000 patients. The authors emphasise the potential of models to identify high-risk patients and predict treatment response, paving the way for personalised, more effective PD treatment.

Article [11] presents a global modelling of Parkinson's disease prevalence through 2050, accounting for age, gender, and the socio-demographic index. The authors predict a doubling of the number of patients to 25.2 million, with population ageing as the main growth factor, and the largest growth observed in countries with a medium socio-demographic index and among men, underscoring the urgent need for strategic healthcare resource planning and policy development.

Work [12] investigates the prediction of the «wearing-off» phenomenon in Parkinson's disease patients using wearable devices and smartphone data. The authors

employed six deep learning architectures to predict symptom fluctuations over the subsequent hour, demonstrating the potential of such methods for personalised PD monitoring and management, consistent with earlier findings in the field.

Analysis of current research on Parkinson's disease highlights several systemic problems, the primary being the high heterogeneity and complexity of clinical and molecular data. Existing methodological approaches often show significant variability in results depending on the demographic characteristics of the sample and require substantial improvement of assessment tools to achieve a stable clinical effect. In addition, considerable overlap of motor and non-motor symptoms at different stages of disease progression complicates the construction of a clear separating hyperplane between stages. Despite the existence of some successful models, clinical practice still lacks integrated automated systems capable of providing end-to-end data processing and reliable decision support for new patients.

Purpose of the Article. The purpose of the study is to develop and validate an integrated automated system for the classification of early and late stages of Parkinson's disease based on a comprehensive integration of clinical data and machine learning methods. The work seeks to improve the accuracy, generalisability, and clinical interpretability of models by building a reproducible data processing pipeline and conducting a comparative analysis of composite machine learning models, taking into account class imbalance.

Presentation of the Main Research Material. In the initial stage, the clinical dataset of patients with Parkinson's disease was loaded into the Python environment. The sample contained 12,608 records for unique patients, indicating the presence of repeated clinical observations [13].

Initial cleaning included checking the table structure, standardising variable types, and eliminating technical inaccuracies. Records with contradictory or clinically implausible indicator values were logically filtered out.

Missing values were handled according to variable type: statistical imputation was used for numerical features, and modal or special labelling, followed by subsequent pipeline processing, was applied for categorical features. Records lacking key date information (specifically the day) were removed separately, as this prevented the accurate formation of temporal characteristics and further analysis. Implementing this stage ensured data integrity and consistency before the next stage of feature engineering and model building.

During the feature engineering stage, a structural transformation of primary clinical indicators was performed to increase their informativeness for machine learning models. Specifically, derivative variables were created from existing numerical parameters (aggregated characteristics, relative ratios, temporal indicators of disease progression) to capture the dynamic aspects of Parkinson's disease progression.

Categorical features underwent logical transformation, including the standardisation of Boolean variables and the conversion of disparate formats into a uniform representation. and converting different recording formats to a single representation. This ensured correct subsequent processing within the pipeline using categorical data encoding mechanisms.

Special emphasis was placed on aligning features with the clinical interpretation of disease stages. At this stage, the data structure was prepared for target variable definition; specifically, we verified that disease duration and other relevant indicators met the requirements of the binary classification task. These transformations enhanced the representativeness of the feature space and minimised the risk of logical inconsistencies prior to forming the target variable.

The original dataset featured a multi-class structure for Parkinson's disease stages, including *Early, Mid,* and *Late* categories.

To reduce label noise and decrease intra-class variability, an uncertainty margin was introduced, within which observations were excluded from further analysis. Formally, the target variable is defined based on the disease duration :

$$Y_{class} = \begin{cases} early & if T_{duration} < 5 \\ late & if T_{duration} > 8 \\ excluded & if\ 5 \le T_{duration} \le 8 \end{cases} \quad (1)$$

where $T_{duration}$ is the disease duration in years (with a fractional part).

Consequently, patients with a disease duration between 5 and 8 years were excluded from the sample to enhance class separability. The transition zone is clinically characterised by a gradual change in motor and non-motor manifestations, which can result in overlapping features and complicate the delineation of a clear-cut hypersurface.

At the stage of final dataset formation, additional statistical data cleaning was performed by removing numerical feature outliers. For this, a modified interquartile range (IQR) method was applied, which uses robust estimates of extreme values.

For each numerical feature , robust quartiles were calculated:

$$\begin{aligned} Q_1 &= P_5(x) \\ Q_3 &= P_{95}(x) \end{aligned} \quad (2)$$

where P_5 is the 5th percentile of the feature distribution, P_{95} is the 95th percentile.

The interquartile range was defined as:

$$IQR = Q_3 - Q_1. \quad (3)$$

The boundary values of the permissible interval were calculated using the formulas:

$$\begin{aligned} lower &= Q_1 - 1.5 \cdot IQR \\ upper &= Q_3 + 1.5 \cdot IQR \end{aligned} \quad (4)$$

An observation was considered an outlier if the condition was met:

$$x < lower \vee x > upper. \quad (5)$$

The procedure was applied sequentially to each automatically pre-identified numerical feature, removing records that exceeded these limits. The resulting cleaned sample was then retained to form the final feature set.

Upon completion of the cleaning process, the resulting dataset, comprising both numerical and categorical variables, was prepared for further scaling and encoding. At this stage, the dataset structure ensured statistical stability and logical consistency of the feature space prior to visualisation and the construction of classification models.

Before applying dimensionality reduction methods, all numerical features were scaled to eliminate the influence of different orders of magnitude on distance and

maximum-variance direction calculations. This approach ensures the correct operation of algorithms sensitive to the scale of variables.

For the initial analysis of the data structure, *the Principal Component Analysis (PCA)* method was used, which allows projecting a multidimensional feature space into a lower-dimensional space while preserving the maximum proportion of variance. Visualising the first components enabled assessment of the degree of class separation and the nature of their overlap in the latent space.

Additionally, *the Multidimensional Scaling (MDS)* method was applied, which reproduces distances between objects in a two-dimensional representation while preserving the similarity structure of the observations. Given the computational complexity of MDS for large samples, a sampling procedure was used, which allowed to reduce the number of observations without significant loss of representativeness.

After distribution into the multiclass structure of the *Late* limit, *PCA* and *MDS* visualisation is performed. Based on the visualisation and considering the research goal with a clear distinction between the early and late phases of prediction, as well as due to the absence of clear demarcation and class overlap, it was decided to use binary classification, which led to the removal of the *Mid* classifier, for a clear definition of the *Early* and *Late* stages.

The obtained visualisations enabled assessment of the geometric structure of the feature space and the complexity of the classification task, in particular, the presence of partial class overlap in a multidimensional environment.

Given the presence of repeated observations for each patient, a *grouped train/test split* strategy was applied, grouping by patient ID. This was achieved by distributing patients across the training and test subsets, ensuring that each patient's records fell into only one subset, preventing data leakage during model training and evaluation.

This approach prevents data leakage and eliminates artificial metric inflation caused by observation correlation. As a result, training and test samples were formed, ensuring a correct assessment of the models' generalisation capability.

During data preprocessing, numerical features were standardised to a uniform scale, preventing some variables from dominating others during model training. Categorical features were transformed using *OneHotEncoding*, ensuring correct representation of nominal variables in a format compatible with most machine learning algorithms.

To organise the processing of different feature types, ColumnTransformer was used, which allows simultaneous application of scaling to numerical columns and encoding to categorical ones. The handle_unknown='ignore' parameter ensures robustness to new or missing categories in test data, preventing inference errors. This approach guarantees unified, reproducible data processing and lays the foundation for reliable model training.

For classifying patients with Parkinson's disease, three modern machine learning algorithms were selected: *RandomForest, XGBoost,* and *LightGBM*. The selection of models was based on their ability to effectively work with heterogeneous features and handle class imbalance. For each algorithm, a pipeline was built, integrating preliminary preprocessing (scaling numerical features, encoding categorical ones) with the classifier.

This organisation ensures automated and reproducible data processing at all stages of training and prediction, reducing the risk of errors when working with new patient data.

The *Random Forest* algorithm implements bagging (Bootstrap Aggregating), which aims to reduce model variance without significantly increasing bias. Let D be the training sample. The algorithm forms B bootstrap samples (samples with replacement), on each of which a separate decision tree $h_b(x)$ is built.

A key feature of the method is the random selection of a subset of $m < M$ features (where M is the total number of features) at each node split. Gini Impurity is used as the splitting quality criterion:

$$G(t) = 1 - \sum_{k=1}^{K} p_{tk}^2, \tag{6}$$

where p_k is the proportion of objects of class k in node t, and K is the number of classes.

When splitting node t into child nodes t_L and t_R, the information gain is maximised, which is equivalent to minimising the weighted impurity:

$$\Delta G = G(t) - \left(\frac{N_{tL}}{N_t} G(t_L) + \frac{N_{tR}}{N_t} G(t_R) \right), \tag{7}$$

where N is the number of objects in node t, N_L and N_R, are the number of objects in the corresponding child nodes.

The final prediction for a new object x' is formed by majority voting:

$$\hat{y} = ar\ gmax \sum_{b=1}^{B} I\left(h_b(x') = k\right), \tag{8}$$

where $I(\cdot)$ is an indicator function that equals 1 if the tree's prediction belongs to class k, and 0 otherwise. This aggregation mechanism neutralises the influence of individual overfitted trees and ensures the robustness of the final decision.

Model training was performed using *RandomizedSearchCV* to find optimal hyperparameters via cross-validation. To account for class imbalance in the training sample, *SMOTE oversampling* was applied, increasing the representation of the minority class and enhancing the models' sensitivity to late-stage disease.

Unlike *Random Forest*, the *XGBoost* algorithm implements a gradient boosting strategy, where base models (decision trees) are built sequentially. Each new tree is trained to minimise the errors of the previous model, thereby correcting residuals and gradually improving the approximation quality.

At step t, the regularised objective function is minimised:

$$\mathcal{L}^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \tag{9}$$

where $l(\cdot)$ is a differentiable convex loss function (in this study, *LogLoss* was used for binary classification), y_i is the true class label, $\hat{y}_i^{(t-1)}$ is the ensemble prediction at the previous step, $f_t(x_i)$ is the new tree added to the model, and, $\Omega(f_t)$ is the regularisation term.

LogLoss has the following form:

$$l(y, p) = -log(p) + (1 - y) \cdot log(1 - p). \tag{10}$$

The regularisation term is given by:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{11}$$

where T is the number of leaves in the constructed tree, reflecting its structural complexity, w is the vector of leaf weights, and $\|w\|^2$ is the square of the L2-norm of this vector (the sum of squares of weights in all leaves), γ is the penalty coefficient for the number of sheets (structural regularisation), and λ is the L2-regularisation coefficient for weights.

The first term γT limits the excessive growth of the tree, while the second term $\frac{1}{2}\lambda \|w\|^2$ controls the magnitude of the weights in the leaves. Collectively, this reduces the risk of overfitting and ensures better generalisation ability of the model.

Parameter optimisation is carried out via a second-order Taylor expansion of the objective function, which allows consideration of both the gradient (first derivative) and the Hessian (second derivative) of the loss function and ensures efficient convergence of the algorithm.

LightGBM is an optimised implementation of gradient boosting, focused on high computational efficiency, training speed, and scalability when working with large datasets. The algorithm retains the general idea of building an ensemble of trees additively, but uses a number of engineering and mathematical optimisations.

Unlike the level-wise approach, *LightGBM* uses a leaf-wise strategy, meaning that at each step, the leaf that provides the maximum reduction in the loss function L is split. Formally, a split is chosen that maximises the gain:

$$G_{ain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{G^2}{H + \lambda} \right) - \gamma, \tag{12}$$

where G and H are the sums of gradients and Hessians in the node; G_L, H_L and G_R, H_R are the corresponding values for the left and right child nodes, λ is the L2-regularization coefficient, γ is the penalty for creating a new leaf.

This approach allows the formation of deeper and more adaptive trees, which increases accuracy but requires depth control to prevent overfitting.

The method is based on the selective sampling of objects by gradient magnitude. Instances with large gradients (i.e., with large errors) are fully retained, while a portion of objects with small gradients are randomly sampled. This allows a reduction in computational volume without a significant loss of information, as the most informative observations remain in the training set.

To reduce dimensionality, *LightGBM* merges mutually exclusive sparse features (e.g., those obtained after One-Hot encoding) into a single dense feature. Since such features do not take on non-zero values simultaneously, they can be combined without loss of information, thereby significantly reducing memory and computational costs.

Given the medical specificity of the task and the presence of imbalance, using only the Accuracy metric is insufficient. Therefore, model evaluation was carried out using a set of metrics that allow for a more complete characterisation of classification quality. Thanks to these mechanisms, *LightGBM* achieves a balance between high accuracy

and efficiency, which explains its best results for *Recall* (*macro*) and *F1 (macro*) in this study.

Macro-averaged F1-Score: the harmonic mean of *Precision* and *Recall* is averaged across classes without weighting by class.

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C}\frac{2P_i \cdot R_i}{P_i + R_i}, \tag{13}$$

where P_i is the accuracy, R_i is the recall for class i, and $C = 2$ is the number of classes. This metric is critically important in imbalanced datasets because it equally accounts for the classification quality of both classes and «penalizes» the model for neglecting the underrepresented class. $1/C$ is the uniform averaging coefficient, which effectively acts as a weighting factor for each class.

In macro-averaging, each class is assigned equal weight:

$$w_i = \frac{1}{C}, \tag{14}$$

regardless of the number of its representatives in the sample. Thus, the influence of each class on the overall assessment is symmetrical, which prevents a more numerous class from dominating.

By comparison, weighted averaging involves the use of weighting factors proportional to the number of objects in the respective class:

$$w_i = \frac{n_i}{N}, \tag{15}$$

where n_i is the number of objects of class i in the sample; N is the total number of objects; w_i is the weighting coefficient for class i.

In this case, the metric takes the form:

$$F1_{weighted} = \sum_{i=1}^{C} w_i \cdot F1_i. \tag{16}$$

Macro-averaging is critically important in imbalanced settings, as it equally accounts for the classification accuracy of all classes and reduces the risk of obtaining inflated quality estimates due to the correct classification of only the dominant class.

The completeness for class k is defined as:

$$R_k = \frac{TP_k}{TP_k + FN_k}, \tag{17}$$

where TP_k is the number of true positives in class k, FN_k is the number of objects in class k, which the model misclassified.

Similarly, the accuracy for class k is defined as:

$$P_k = \frac{TP_k}{TP_k + FP_k}, \tag{18}$$

where FP_k is the number of false positives, i.e., objects from other classes that are incorrectly classified into the class k.

In a clinical setting, a high *Recall* rate is of fundamental importance because it minimizes Type II errors (*False Negative*), i.e., it reduces the risk of missing a patient with advanced-stage disease and prescribing inadequate treatment.

Along *Recall,* a critically important reliability indicator of a diagnostic system is *Precision,* which in biostatistics is also referred to as the positive predictive value.

Precision is defined as the ratio of the number of true positives to the total number of objects which the algorithm classified as positive. Precision for class *k* is defined as:

$$P_k = \frac{TP_k}{TP_k + FP_k}, \tag{19}$$

where TP_k is the number of patients correctly classified as belonging to the class *k*, FP_k is the number of patients incorrectly classified by the model as belonging to the class *k*.

A confusion matrix allows us to visualize the structure of classification errors *(TP, TN, FP, FN),* and analyse any potential systematic bias in the model towards one of the classes. Analysing it is essential for the clinical interpretation of results.

To predict the stage of Parkinson's disease in a new patient, the *predict_new_patient* function was implemented, which automatically processes the input data and incorporates all necessary transformations from the training phase.

A key component is the *_coerce_categoricals_to_train* function, which converts categorical features into the format that the classifier saw during training. This allows the correct processing of *bool* values, text representations of logical variables, and unknown categories, avoiding errors when applying OneHotEncoding in the pipeline.

The *predict_new_patient* function generates *DataFrame* from the provided values, expands it to the full set of features, and applies scaling and encoding using the trained pipeline. Both *predict* (patient class) and `*predict_proba*` (class probabilities) are available for prediction, allowing you to assess the model's confidence in its prediction.

For better understanding, a diagram of the class balancing process has been developed, as shown in Figure 1.

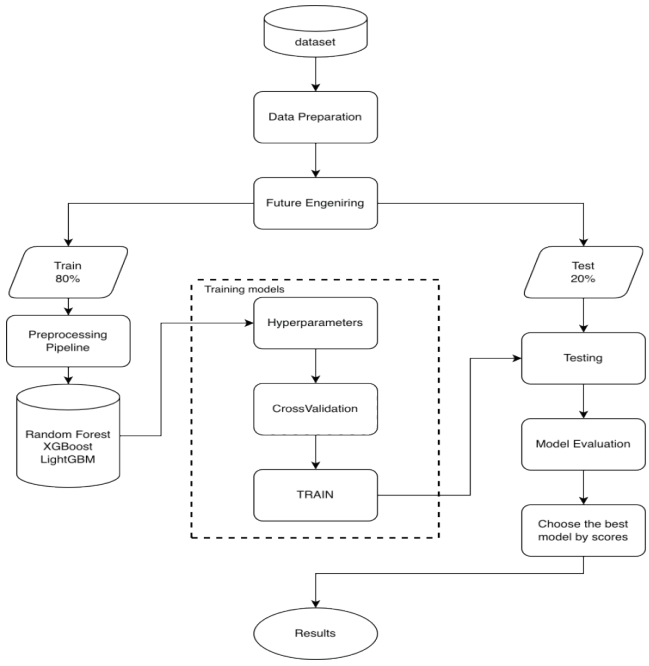

Figure 1. Data projection onto the first two principal components (PCA)

The dataset under study contained 12,608 records corresponding to 100 unique patients, with an average of approximately 126 observations per patient. This structure indicates that the data is panel-based, with each patient represented by a series of repeated clinical measurements.

To improve data quality and model stability, the interquartile range (IQR) method was applied to numerical features. As a result, 281 records were removed, representing approximately 2.23% of the total sample.

Such cleaning allows reducing the impact of abnormal clinical indicators that potentially distort model training, and increasing the stability and generalization ability of predictions. After outliers removal, a binary target variable with two classes was formed:

– 0 (Early Stage) – 7 435 records.
– 1 (Late Stage) – 2 650 records.

A moderate class imbalance is observed (~74% vs. ~26%), which may affect model training. To balance class representation in the training sample, *SMOTE oversampling* was applied, which allows increasing the sensitivity of models to the minority class and ensuring a balanced assessment of classification quality.

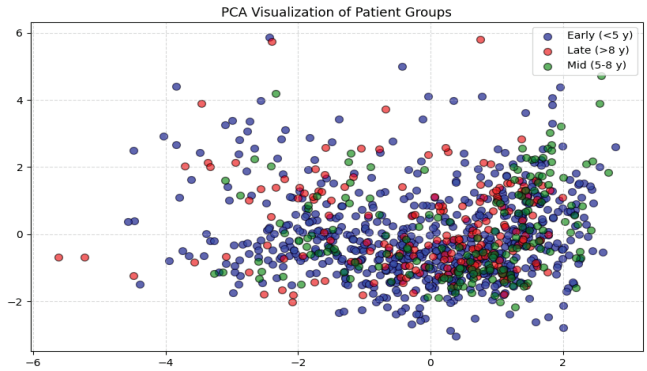

Figure 2. Data projection onto the first two principal components (PCA)

Before applying PCA, the numerical variables were standardized to ensure accurate calculation of the variance and the directions of the principal components.

Analysis of the first two components showed that they explain a significant portion of the total variance. However, the patient classes exhibit only partial separation and substantial overlap, indicating a lack of linear separability. Therefore, the use of nonlinear models is appropriate for accurate disease stage classification.

To ensure computational efficiency, the data were sampled prior to applying the MDS. This method preserves the distances between objects in multidimensional space, reflecting the similarity structure of patients in a two-dimensional representation.

The visualization shows the formation of clusters, though with some class overlap, confirming the complexity of the classification task and the need to use more flexible algorithms.

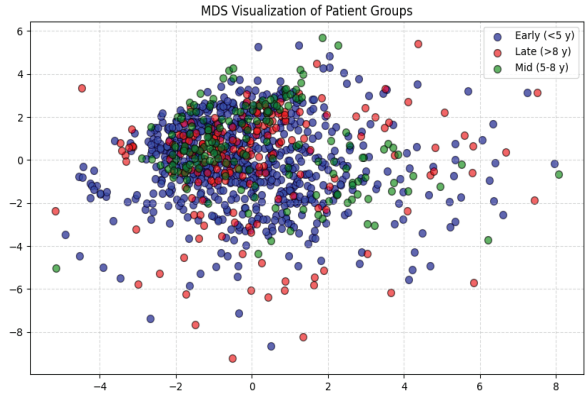

Figure 3. Two-dimensional data projection using multidimensional scaling (MDS)

A *grouped split* by patient ID was used to split the data. As a result, the following was formed:

– *Train*: 8 118 records.
– *Test*: 1 967 records.

This approach ensures that there is no data leakage between patients, since all observations for a single patient are included in only one subsample. This allows for more accurate clinical validation and enables the assessment of the models' generalization ability on new, previously unknown patients.

To address the class imbalance problem, *SMOTE oversampling* is applied, which increases the representation of the minority class in the training set.

Given the medical specificity of the task and the presence of class imbalance, using only the *Accuracy* metric does not provide an objective assessment of the model's quality. Therefore, the evaluation was performed based on a complex set of metrics. The primary metric was the macro-averaged *F1-score* – a generalized indicator that combines *Precision* and *Recall,* and is calculated separately for each class, followed by averaging without considering their size. This approach allows for equally considering the recognition quality of both larger and less represented classes and prevents situations where the model demonstrates high overall accuracy by ignoring a clinically significant group of patients.

Special attention is paid to *Recall,* which reflects the model's ability to identify all instances of a specific class. In a clinical context, this is critically important, as high *Recall* reduces the risk of false-negative results, i.e., missing patients with late-stage disease. At the same time, *Precision* characterises the reliability of a positive prediction and shows how justified it is to assign a patient to a certain stage. For a detailed analysis of the error structure, a confusion matrix was also used, which allows evaluating the ratio of truly and falsely classified cases and identifying possible systematic bias of the model towards one of the classes.

The optimal hyperparameters were searched through *RandomizedSearchCV* with cross-validation (CV). The *F1-score (macro)* was chosen as the optimization criterion

because it provides a balanced assessment of model quality for both classes, taking into account the uneven distribution of records.

To solve the problem of Parkinson's Disease stage prediction, an approach based on Ensemble Learning methods was chosen. This strategy is due to the ability of ensembles to effectively model complex non-linear dependencies in a high-dimensional feature space, as well as their robustness to noise and variability, which are typical for biomedical mHealth data. Within the study, a comparison of three algorithms was conducted: *Random Forest, XGBoost,* and *LightGBM.*

Random Forest is based on the bagging approach, which involves building an ensemble of decision trees on different random subsamples of training data to reduce model variance without significantly increasing bias. Each tree is trained on its own bootstrap sample, and during branching, a random subset of features is used, which further increases the diversity of the ensemble. The quality of node splitting is evaluated by a criterion that reflects the degree of class heterogeneity in the node, and the optimal split is chosen to maximally reduce this heterogeneity. The final prediction for a new object is determined by aggregating the predictions of all trees based on the majority principle, which reduces the influence of individual overfitted models and ensures the stability and reliability of the final solution.

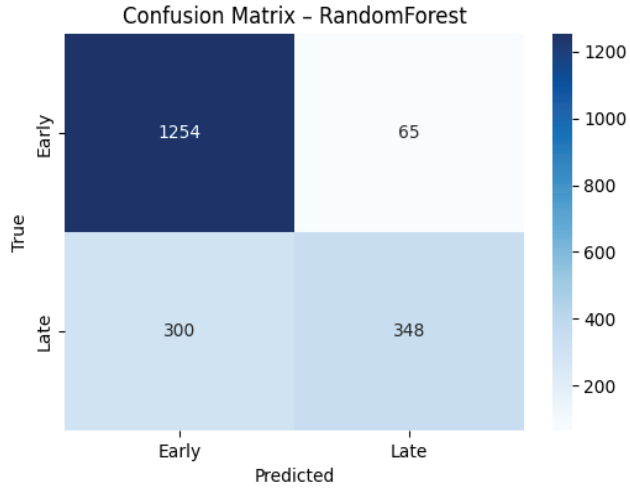

Figure 4. Confusion Matrix for RandomForest

The *RandomForest* confusion matrix showed the following results on the test set:

- *Accuracy*: 0.8144.
- *Precision* (*macro*): 0.8248.
- *Recall* (*macro*): 0.7439.
- *F1* (*macro*): 0.7645.

The advantages of this algorithm include operational stability and reduced sensitivity to data noise, making it a reliable choice for classifying the stages of Parkinson's disease in clinical samples with repeated measurements.

Unlike *Random Forest*, the *XGBoost* algorithm implements a gradient boosting approach, where decision trees are built sequentially, and each subsequent tree aims to correct the errors of the previous model composition. At each step, a regularised loss function is optimised, within which the difference between the true labels and the current ensemble predictions is evaluated; a logarithmic loss function is used for binary classification tasks. A regularisation component is added to the target function that simultaneously limits the structural complexity of the tree and controls the magnitude of weights in its leaves, reducing the risk of overfitting and improving the model's generalization ability. For efficient optimisation, a second-order approximation is used, which considers both the gradient and the curvature of the loss function, ensuring stable and fast convergence of the algorithm.

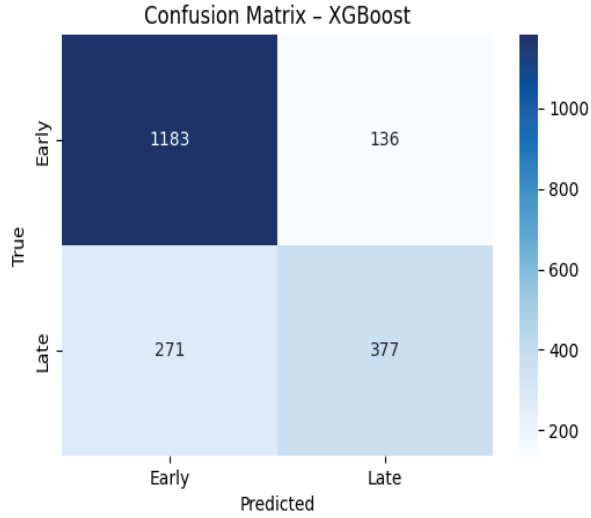

Figure 5. Confusion Matrix for XGBoost

Confusion matrix *XGBoost* showed the following results on the test sample:

– *Accuracy*: 0.7931.
– *Precision* (*macro*): 0.7743.
– *Recall* (*macro*): 0.7393.
– *F1* (*macro*): 0.7513.

A key feature of *XGBoost* is its improved probability calibration, which makes it suitable for testing the new patient prediction function and assessing the model's confidence in predicting disease stage.

LightGBM is an optimised implementation of gradient boosting, designed for high training speed and efficient handling of large datasets. While retaining the additive principle of ensemble tree construction, the algorithm applies a series of engineering enhancements, including a leaf-wise growth strategy, where at each step, the leaf that provides the greatest improvement in model quality is expanded. This allows the formation of deeper, more adaptive structures while requiring complexity control to prevent

overfitting. To increase computational efficiency, selective sampling of observations with the largest errors is used, which reduces the number of computations without significant loss of information, as well as a mechanism for merging mutually exclusive sparse features to reduce dimensionality. The combination of these approaches ensures a balance between high accuracy and speed, resulting in the best model performing across key quality metrics.

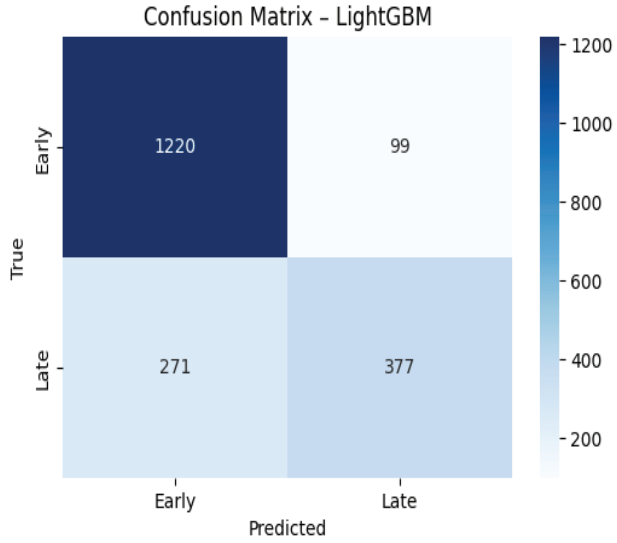

Figure 6. Confusion Matrix for LightGBM

Confusion Matrix *LightGBM* showed the following results on the test sample:

– *Accuracy*: 0.8119.
– Precision (macro): 0.8051.
– *Recall* (*macro*): 0.7534.
– *F1* (*macro*): 0.7696.

This model demonstrated the highest *Recall (macro)* and *F1 (macro)* scores among all tested algorithms, indicating its effectiveness in recognising both early and late stages of Parkinson's disease.

A summary comparison of key quality metrics for all tested models on the test sample, allowing for an assessment of their effectiveness in classifying patients' disease stages, is provided below.

Table 1. Final results table

Model	Accuracy	Precision	Recall (macro)	F1 (macro)
RandomForest	0.8144	0.8247	0.7439	0.7645
XGBoost	0.7931	0.7742	0.7393	0.7513
LightGBM	0.8119	0.8051	0.7534	**0.7696**

The primary criterion for the final model selection was the maximum *F1 (macro) score*. ***LightGBM*** proved to be the best solution, delivering the highest balanced classification quality, while *RandomForest* showed similar results, and *XGBoost* had a slightly lower *F1 score*.

The new patient prediction model was tested on the *LightGBM* model. It demonstrated correct handling of categorical variables, automatic scaling of numerical features through the pipeline, and the ability to obtain *'predict_proba'* for assessing prediction confidence. In the example provided, the patient is a 72-year-old man with 134 tapping measurements, without deep brain stimulation (DBS), with a «smoking» status, and moderate indicators of tapping interval variability. Based on these characteristics, the model generated a clinically interpretable conclusion: the prognosis is *Early Stage* (label=0) with a probability of *82.39%*, which can be used as an auxiliary tool to support clinical decision-making.

The *LightGBM* model proved to be the most effective due to its ability to work with heterogeneous features, including numerical and categorical variables, and to effectively represent non-linear dependencies between them. In addition, its algorithmic architecture ensures less susceptibility to overfitting, which allows getting stable results on the test sample and increases the overall balance of classification.

High Recall (macro) is critical in a clinical context, as it helps minimise false-negative cases of patients with late-stage disease that might otherwise go without timely intervention. A balanced recall score ensures uniform sensitivity of models across both classes and is essential for making informed decisions in medical practice.

The use of *SMOTE oversampling* helps reduce model bias towards the more prevalent Early class, increasing sensitivity to the underrepresented *Late* class. This contributes to improving the *F1-score* and *recall* for the critically important late-stage patient class, thereby enhancing the practical value of the models.

The results of *PCA* and *MDS* show only partial separation of classes in the feature space, which explains why the *Accuracy* value does not exceed 0.85 even for the best models. Such visualisations confirm the complexity of the biomedical problem and the need to apply more flexible, non-linear algorithms to achieve high classification quality.

Conclusions. As a result of the research, a proprietary solution to the pressing problem of Parkinson's disease stage classification was implemented by creating a comprehensive machine learning pipeline, ensuring automated inference and reproducible data processing. Initial data preparation, including outlier removal using the *IQR* method and feature standardisation, ensured the consistency and representativeness of the feature space. A key innovation of the study was the introduction of a margin-based exclusion strategy for disease duration, which artificially increased class separability and reduced labelling noise. Data visualisation using *PCA* and *MDS* confirmed the presence of non-linear separability in biomedical data, justifying the transition to binary classification of *Early* and *Late* stages.

The constructed ensemble models, *RandomForest*, *XGBoost*, and *LightGBM*, demonstrated high predictive capability, which is especially important in conditions of class imbalance. The *LightGBM* algorithm, combined with *SMOTE* balancing, yielded

the best results, achieving the highest *Recall* and *F1-score*. This allowed for mitigating the risk of Type II errors and increasing the reliability of detecting patients in the late stage. The developed automated inference function integrates all preprocessing steps for new patients, enabling reliable predictions with probability estimates for each class. The proposed approach is a promising practical tool for personalised management of patients with Parkinson's disease, creating a solid foundation for the further implementation of clinical decision support systems.

Further research may include SHAP analysis of feature importance, enabling a detailed assessment of the contribution of each clinical and demographic characteristic to disease-stage prognosis. It is also advisable to conduct ROC-AUC analysis to evaluate the models' ability to distinguish between classes at different thresholds, thereby improving the interpretation of their sensitivity and specificity.

In the future, it may be possible to restore a multi-class model that includes an intermediate Mid class for more accurate modelling of disease progression. In addition, an important step is to conduct external validation in an independent cohort of patients, which will allow assessment of the models' generalisation ability in real clinical settings.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Trufanov Y., Machado de Oliveira L., Svyrydova N., Suchowersky O. The prevalence of Parkinson disease in Ukraine. *Movement Disorders Clinical Practice*, 2023, 10(3), P. 524-525. DOI: 10.1002/mdc3.13668.
2. Kostiuk K., Lisianyi A., Medvediev Y., Popov A., Cheburakhin V., Buniakin V., Tevzadze D. Prospects for the use of deep brain stimulation in the treatment of Parkinson's disease in Ukraine. *Medicni Perspektivi*, 2024, 29(4), P. 193-207. DOI: 10.26641/2307-0404.2024.4.319368.
3. Martínez Tirado G., Martins Conde P., Sapienza S., Fröhlich H., Pauly C., Schröder V. E., Klucken J. Data-driven clinical decision support tool for diagnosing mild cognitive impairment in Parkinson's disease. *npj Parkinson's Disease*, 2026. DOI: 10.1038/s41531-025-01222-6.
4. Magano D., Barros A. S., Massano J., Alsuwaidi L., Taveira-Gomes T. Non-motor symptoms as critical predictors of quality of life in Parkinson's disease: a machine learning approach. *Health and Quality of Life Outcomes*, 2026, 24(1), P. 7. DOI: 10.1186/s12955-025-02451-2.
5. Babichev S., Yasinska-Damri L., Liakh I. A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques. *Applied Sciences*, 2023, 13(10), P. 6022. DOI: 10.3390/app13106022.
6. Bayram E., Batzu L., Tilley B., Gandhi R., Jagota P., Biundo R., Garon M., Prasertpan T., Lazcano-Ocampo C., Chaudhuri K. R., Weil R. S. Clinical trials for cognition in Parkinson's disease: Where are we and how can we do better? *Parkinsonism and Related Disorders*, 2023, 112, 105385. DOI: 10.1016/j.parkreldis.2023.105385.
7. Willis A. W., Roberts E., Beck J. C., Fiske B., Ross W., Savica R. Incidence of Parkinson disease in North America. *npj Parkinson's Disease*, 2022, 8(1), P. 170. DOI: 10.1038/s41531-022-00410-y.
8. Coundouris S. P., Henry J. D., Suddendorf T., Lehn A. C. Affective forecasting in Parkinson's disease. *Journal of the International Neuropsychological Society*, 2023, 29(4), P. 406-409. DOI: 10.1017/S1355617722000388.

9. Chowdhury A. R., Ahuja R., Manroy A. A Machine learning driven approach for forecasting Parkinson's Disease progression using temporal data. In I*nternational Conference on Distributed Computing and Intelligent Technology*, 2024, P. 266-281. *Cham: Springer Nature Switzerland*. DOI: 10.1007/978-3-031-50583-6_18.
10. Dileep, Jaswath, Agrawal K., Prathima T. Forecasting Parkinson's disease progression: Leveraging machine learning techniques. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 2024, P. 1-7. *IEEE*. DOI: 10.1109/APCIT62007.2024.10673685.
11. Su D., Cui Y., He C., Yin P., Bai R., Zhu J., Feng T. Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. *BMJ*, 2025, 388. DOI: 10.1136/bmj-2024-080952.
12. Victorino J. N., Shibata Y., Inoue S., Shibata T. Forecasting Parkinson's disease patients' wearing-off using wrist-worn fitness tracker and smartphone dataset. In *Human Activity and Behavior Analysis*, 2024, P. 3-22. *CRC Press.*
13. BrianMBot. mPower Public Researcher Portal [Dataset]. *Synapse*, 2015. DOI: 10.7303/SYN4993293.

REFERENCES

1. Trufanov, Y., Machado de Oliveira, L., Svyrydova, N., & Suchowersky, O. (2023). The prevalence of Parkinson disease in Ukraine. *Movement Disorders Clinical Practice*, 10(3), 524-525. https://doi.org/10.1002/mdc3.13668.
2. Kostiuk, K., Lisianyi, A., Medvediev, Y., Popov, A., Cheburakhin, V., Buniakin, V., & Tevzadze, D. (2024). Prospects for the use of deep brain stimulation in the treatment of Parkinson's disease in Ukraine. *Medicni Perspektivi*, 29(4), 193-207. https://doi.org/10.26641/2307-0404.2024.4.319368.
3. Martínez Tirado, G., Martins Conde, P., Sapienza, S., Fröhlich, H., Pauly, C., Schröder, V. E., ... & Klucken, J. (2026). Data-driven clinical decision support tool for diagnosing mild cognitive impairment in Parkinson's disease. npj Parkinson's Disease. https://doi.org/10.1038/s41531-025-01222-6.
4. Magano, D., Barros, A. S., Massano, J., Alsuwaidi, L., & Taveira-Gomes, T. (2026). Non-motor symptoms as critical predictors of quality of life in Parkinson's disease: a machine learning approach. *Health and Quality of Life Outcomes*, 24(1), 7. https://doi.org/10.1186/s12955-025-02451-2.
5. Babichev, S., Yasinska-Damri, L., & Liakh, I. (2023). A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques. *Applied Sciences*, 13(10), 6022. https://doi.org/10.3390/app13106022.
6. Bayram, E., Batzu, L., Tilley, B., Gandhi, R., Jagota, P., Biundo, R., Garon, M., Prasertpan, T., Lazcano-Ocampo, C., Chaudhuri, K. R., & Weil, R. S. (2023). Clinical trials for cognition in Parkinson's disease: Where are we and how can we do better? *Parkinsonism & Related Disorders*, 112, 105385. https://doi.org/10.1016/j.parkreldis.2023.105385.
7. Willis, A. W., Roberts, E., Beck, J. C., Fiske, B., Ross, W., Savica, R., ... & Parkinson's Foundation P4 Group Alcalay Roy Schwarzschild Michael Racette Brad Chen Honglei Church Tim Wilson Bill Doria James M. (2022). *Incidence of parkinson disease in North America. npj Parkinson's Disease*, 8(1), 170. https://doi.org/10.1038/s41531-022-00410-y.

8. Coundouris, S. P., Henry, J. D., Suddendorf, T., & Lehn, A. C. (2023). Affective forecasting in Parkinson's disease. *Journal of the International Neuropsychological Society*, 29(4), 406-409. doi:10.1017/S1355617722000388.
9. Chowdhury, A. R., Ahuja, R., & Manroy, A. (2024, January). A Machine learning driven approach for forecasting Parkinson's Disease progression using temporal data. In *International Conference on Distributed Computing and Intelligent Technology* (pp. 266-281). *Cham: Springer Nature Switzerland*. https://doi.org/10.1007/978-3-031-50583-6_18.
10. Dileep, Jaswath, Agrawal, K., & Prathima, T. (2024). Forecasting Parkinson's disease progression: Leveraging machine learning techniques. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* (pp. 1-7). *IEEE*. https://doi.org/10.1109/APCIT62007.2024.10673685.
11. Su, D., Cui, Y., He, C., Yin, P., Bai, R., Zhu, J., ... & Feng, T. (2025). Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. *bmj*, 388. https://doi.org/10.1136/bmj-2024-080952.
12. Victorino, J. N., Shibata, Y., Inoue, S., & Shibata, T. (2024). Forecasting Parkinson's disease patients' wearing-off using wrist-worn fitness tracker and smartphone dataset. In *Human Activity and Behavior Analysis* (pp. 3-22). *CRC Press.*
13. BrianMBot. (2015). mPower Public Researcher Portal [Dataset]. *Synapse*. https://doi.org/10.7303/SYN4993293.

doi: 10.32403/1998-6912-2026-1-72-139-158

КОМПЛЕКСНА ІНТЕГРАЦІЯ МАШИННОГО НАВЧАННЯ ТА КЛІНІЧНИХ ДАНИХ ДЛЯ КЛАСИФІКАЦІЇ РАННІХ І ПІЗНІХ СТАДІЙ ХВОРОБИ ПАРКІНСОНА

І. М. Лях[1], Р. Я. Жовтані[2], М. Ю. Кашка[3], А. Ю. Ціпіньо[4]

[1] Державний вищий навчальний заклад «Ужгородський національний університет», пл. Народна, 3, м. Ужгород, 88000, Україна, https://orcid.org/0000-0001-5417-9403, e-mail: igor.lyah@uzhnu.edu.ua
[2] Державний вищий навчальний заклад «Ужгородський національний університет», пл. Народна, 3, м. Ужгород, 88000, Україна, https://orcid.org/0000-0002-7421-148X, e-mail: ruslana.zhovtani@uzhnu.edu.ua
[3] Державний вищий навчальний заклад «Ужгородський національний університет», пл. Народна, 3, м. Ужгород, 88000, Україна, https://orcid.org/0000-0001-7437-6156, e-mail: mariia.kashka@uzhnu.edu.ua
[4] Державний вищий навчальний заклад «Ужгородський національний університет», пл. Народна, 3, м. Ужгород, 88000, Україна, https://orcid.org/0009-0007-1704-2732, e-mail: tsipino.artemii@student.uzhnu.edu.ua

Метою даного дослідження є оцінка ефективності інформаційних кампаній у сфері громадського здоров›я, визначення їх впливу на рівень обізнаності населення та формування рекомендацій щодо оптимізації комунікаційних стратегій. Для аналізу використано відкритий датасет mPower Public Researcher Portal (BrianMBot, 2015), який містить інформацію про поведінку та активність учасників дослідження мобільного додатку для пацієнтів із хворобою Паркінсона. Дані оброблено у середовищі Python із застосуванням бібліотек для статистичного аналізу та машинного навчання.

Для оцінки ефективності інформаційних кампаній використано описову статистику, кореляційний та регресійний аналіз, а також кластеризацію для виявлення груп учасників із різним рівнем залученості та сприйняття інформації. Дослідження показало, що кампанії з інтерактивними онлайн-інструментами та мультимедійним контентом мають найвищий рівень охоплення та впливу на поведінкові зміни населення.Традиційні методи, такі як друковані матеріали та радіо, демонструють меншу ефективність, особливо серед молодих аудиторій. Крім того, було виявлено значущу різницю у сприйнятті інформації залежно від регіону та соціально-демографічних характеристик респондентів. Ефективність інформаційних кампаній у сфері громадського здоров›я значною мірою залежить від інтерактивності та адаптації контенту до специфіки цільових аудиторій. Рекомендовано застосовувати комбіновані стратегії, що поєднують цифрові та традиційні канали комунікації, з урахуванням локальних особливостей та потреб населення.

Ключові слова: *хвороба Паркінсона, прогнозування стадії, машинне навчання, Python, Random Forest, XGBoost, LightGBM, клінічні дані, автоматизована інференція, бінарна класифікація.*

Стаття надійшла до редакції 10.05.2026.
Submitted: 10.05.2026.
Прийнято до друку: 14.05.2026.
Accepted: 14.05.2026.
Опубліковано: 30.05.2026.
Published: 30.05.2026.