International Journal
Information Technology Applications (ITA)
Volume 11, Number 2, December 2022

DETECTION OF PARKINSON'S DISEASE WITH MACHINE LEARNING SUPPORT

**Zuzana Képešiová, Štefan Kozák, Eugen Ružický,
Alfréd Zimmermann, Richard Malaschitz**

Abstract:

Parkinson's disease (PD) is a neurodegenerative disorder causing partial or complete loss of motor reflexes and speech. It affects the patients significantly in their daily life. The accurate diagnosis of PD is quite complex, requires a lot of resources, and equipment and is time-consuming for diagnosis, especially in its initial stage's occurrence of the disease. To help the medical experts and researchers diagnose PD, we developed a machine learning approach based on the simple descriptive tasks captured in audio on the smartphone. The dataset for demonstrations consists of over 1500 patients with approximately 9% of patients diagnosed with PD. Hence the imbalance of the dataset the evaluation metrics such as Mathews correlation coefficient (MCC), sensitivity - specificity, and ROC curve were picked to describe the selected machine learning algorithms performance. The analysis of tested results in a single approach resulting with 73.49% (52.53% - 91.41%) MCC.

Keywords:

Machine learning, medical voice recordings, Parkinson disease, early detection, classification.

ACM Computing Classification System:

Machine learning.

1 Introduction

Parkison's disease (PD) is the second most common neurological disease, right after Alzheimer disease. At present, there is no cure nor a known root for PD, only a treatment, that can ease the symptoms and increase the quality of life of those affected. The symptoms of PD might demonstrate as problems with movement when the person is slow or even sometimes seems to be still. A well-known sign is rigidity (especially wrist, shoulder, and neck), and imbalance of neurotransmitters, dopamine, and acetylcholine. PD is mostly a movement disorder, but other weakenings also commonly develop, including psychiatric problems such as depression and dementia. Other noticeable sign of PD is the quality of speech.

The association of speech disorders with Parkinson's disease has been verified in various studies [1] [2]. Other studies have shown the progression of the disease is linked to a decrease in voice performance over time [3]. Therefore, the speech samples are ideal for the detection of the PD, while the data can be easily collected. Studies on the PD are usually focused on voice problems. Such practices also assist in the early diagnosis of the disease. People with Parkinsonism have vocal disorders affecting their speech in areas like volume level, the correct pronunciation of syllables, etc [4].

A Parkinson disease detection in machine learning related studies is being detected from various data sources such as handwriting database [5], dataset consisting of local field potential (LFP) recordings [6], gait features [7], single-photon emission computed tomography (SPECT) images [8], vocal recordings converted into images [9], and speech recordings converted into features [10]. In the previous study [11] we have shown a first comparative analysis of the machine learning algorithms for the early detection of PD, while in this paper we focus on a more specific treats between PD and speech symptoms with enlarged dataset focused on more specific tasks.

The paper is organized as follows. Section 2 describes the obtained dataset and its features, and introduces machine learning methods for classification, scaling methods of data, sampling methods for imbalanced datasets, and feature selection methods. Section 3 shows experimental results for the comparative analysis of proposed machine learning algorithms in a combination of sampling methods, scaling methods, and selection of features.

2 Machine Learning Methods

This section navigates through technical information about the analysis and the developed mathematical model. It starts with the dataset and its description, continues with data scaling methods, hence the imbalance of the dataset it continues with the sampling methods and finishes with the proposed machine learning algorithms to be analyzed.

A Dataset and Features

The collected dataset is containing over 1500 recorded samples. The data is consisting of 2 main data sources: basic information about the patient such as age, sex, education, etc; speech recordings statistics such as number of gaps, length of the gaps between words, number of used words, phonetism, etc; resulting in over 66 000 various features. The features are being computed from vocal recordings of subjects describing the seen images shown by the examiner. There are 2 types of images: small (62 images) and big (5 images). While small images depict one object or action, the big images consist of sceneries, which require more attention to detail. The dataset containes subjects falling into one or more of following groups:

- No neurodegenerative disorder
- Mild Cognitive Impairment (MCI)
- Alzheimer disease
- Parkinson disease
- Other neurodegenerative disorder

The dataset was cleaned up of low-quality recordings to ensure the transformation algorithms to extract the features correctly. Later, the subjects with diagnosed MCI and / or Alzheimer disease were filtered out of the dataset for early detection of Parkinson disease experiments. The dataset was filtered not only horizontally, but also vertically by selecting only 40 small images and 2 big images for the experiments.

B Scaling Methods

Many machine learning algorithms require scaling the input data to converge to the optimum as fast and smooth as possible. During the experimentation we used several different scaling methods.

Standard scaler, also known as z-score, standardize features by removing the mean and scaling to unit variance.

Min-max scaler transforms given values to fit into a given range set up to <0,1>.

Max-abs scaler changes each feature by its maximum absolute value resulting in values set up in a maximum range of <-1,1>.

Robust scaler removes the median and scales the data according to the quantile range, a measure of statistical dispersion

A Quantile transformer maps a variable's probability distribution to another probability distribution. The quantile function ranks or smooths out the relationship between observations and can be mapped onto other distributions, such as the uniform or normal distribution [12].

Yeo–Johnson transformer applies a transformation to stabilize variance and to make data more normal distribution-like, more Gaussian-like. Yeo–Johnson transformation as a power transformer can optimize the negative values in contradiction to the Box-Cox transformation method [13].

L2 normalizer, also known as Euclidean normalization of the group of L^p function spaces, makes the sum of the squares always be up to 1 normalizing the values.

C Sampling Techniques

The dataset consists of only 9% of positive cases for PD, which makes it unbalanced and makes it harder for the machine learning algorithms to learn efficiently, as there are not many samples with the class describing the PD. Some of the available methods dealing with such problem is to artificially create the underclassed samples or in contradiction, cut numbers of overclassed samples. In the paper, we propose four techniques, which may help the algorithms to find the links needed between the features [14].

Random oversampling is a method, where the minority class is evened out by enlarging the dataset with random copies of the minority class to equal the number of minority class compared to the majority class.

Synthetic Minority Over-sampling Technique (SMOTE) as the over-sampling method is a kind of data augmentation, where the minority class samples are not only randomly copied but are synthetically created. SMOTE picks samples close in the feature space from the minority class based on K-Neighbors and artificially creates a new sample with the feature values lying between the neighbors [15].

Adaptive Synthetic algorithm (ADASYN) oversample the dataset in a similar way as SMOTE as it is its extension. In contradiction to SMOTE, ADASYN generates synthetic samples inversely proportional to the density of the samples in the underclass area [16].

TomekLinks is in contradiction to an undersampling technique, where the majority class lowers the number of samples by deletion. TomekLinks uses the modified Condensed Nearest Neighbor method to choose which samples to delete. This modification creates so-called Tomek-link pair of the samples with different output classes and together they have the smallest Euclidean distance to each other in the feature space. This small distance creates close neighbors which implies high noise samples or very close samples to the minority class and therefore they are therefore removed [17].

D Machine Learning Algorithms

For the experiments we propose the comparison of the following machine learning algorithms: Extremely randomized trees, Random Forest, Linear regression, Logistic regression, Ridge regression, Passive Aggressive, Support vector classifier (SVC), Stochastic gradient descent (SGD), Light gradient-boosting machine (LGBM), Perceptron, Linear discriminant analysis (LDA), and k-nearest neighbors (KNN).

Extremely randomized trees propose an ensemble method of randomizing strongly both attribute and cut-point choice while splitting a tree node. The algorithm fits randomized decision trees on various batches of the dataset and utilizes averaging to improve the outcome accuracy [18].

The next tree-based method is called Random Forest. It utilizes the ensembling of decision trees method. In a classification task, the result of a random forest depends on most of the predicted class from all the trees of the forest [19].

The second class of machine learning algorithms is well known as linear algorithms. Linear algorithms predict the outcome of the data by a linear combination of the features and compute the weight of each feature to what extent it affects the predicted outcome. The best known is Linear regression, which fits the linear model with weights to minimize the residual sum of squares between the samples from the dataset and the predicted outcomes. Linear regression predicts the outcome value, which may lay outside of range <0,1> and therefore the input must be scaled properly.

Logistic regression is estimating a probability of the outcome and therefore is used for the classification problem more frequently. The logistic regression sums the weighted input and passes it through the sigmoid activation function resulting in a value in the range of <0,1>.

A Ridge regression extends the typical linear regression augmented by a penalty on the size of the coefficients resulting in a machine learning algorithm used for solving a problem of highly correlated independent variables.

For large-scale learning, there is a subclass of linear algorithms called Passive Aggressive. As the name of the subclass denotes, it is composed of two main parts: passive – if the prediction is correct, keep the model as it is; aggressive – on the other hand, if the prediction is not correct, then update the model. Passive Aggressive models use a regularization parameter to penalize the model in a case of an incorrect prediction [20]. In this comparative analysis, we chose a Passive Aggressive model utilizing a hinge loss function.

A support vector machine (SVM) is a machine learning method used for a large variety of tasks, which includes a classification problem ending in the term Support vector classifier (SVC). SVM maps the training samples to points in the space in order to maximize the gaps between given categories. The samples to be predicted are later mapped into the same space and depending on which side of the gap they fall the gap they belong to.

Gradient descent is an optimization method, where the main technique is a performance of minimizing cost function $J(\theta)$ with the parameter θ updating parameters in the reverse direction of the gradient of the cost function $\nabla\_\theta \; J(\theta)$ with respect to the parameters [21]. In the study, we used an SVM with Stochastic gradient descent (SGD).

Light gradient-boosting machine (LGBM) is based on decision tree algorithms. LGBM implements a highly optimized histogram-based decision tree learning algorithm. The LightGBM algorithm utilizes two novel techniques called Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) which allow the algorithm to run faster while maintaining a high level of accuracy [22].

Perceptron is an algorithm used as a classifier. In a binary classification problem, it maps the input vector onto a single binary output value. The output value depends on the sum of weights multiplied with the input in a dot product manner regularized with the bias factor [21].

A Linear discriminant analysis (LDA) tries to find a projection vector that can enlarge the distance of samples from different classes and reduce the distance of samples from the same class [23]. It may be used not only for a classification task but also for a dimensional reduction with the use of covariance parameters.

A possibility how classifying the samples is to place them into one of the defined groups. The groups might be created by a clustering algorithm known as the k-nearest neighbors (KNN) algorithm. In a contradiction to the previous algorithms, a KNN is a non-parametric supervised learning method, wherein in a classification task the goal is to assign the provided sample to the class based on the votes of its n closest neighbors and their class reference [24].

3 Experiments and Results

A Evaluation Metrics

Due to the high imbalance of the used dataset conventional evaluation metrics would not describe the efficiency of the trained model accurately. Therefore, we utilized the following evaluation metrics: Matthew's correlation coefficient (MCC), precision, and recall. An MCC is based on a confusion matrix and can be described as accuracy for the imbalanced dataset. The specificity metric defines how many negative samples were predicted correctly, while the sensitivity metric describes how many positive samples have been predicted correctly. The evaluation metrics are defined by the following formulas:

$$\text{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}, \quad (1)$$

$$\text{Specificity} = \frac{tn}{tn+fp}, \quad (2)$$

$$\text{Sensitivity} = \frac{tp}{tp+fn}, \quad (3)$$

tn denotes true negatives, tp are true positives, fn are false negatives, fp are false positives [25]

B Experimental Results

The provided dataset of medical voice recordings transformed into the statistics combined with the basic information about patient with imbalanced data distribution over the Parkinson's disease categorization with 9% of positive and 91% of negative cases was divided into training and validation subsets in ratio of 80% training - 20% validation data 5 times in total utilizing stratified cross validation. The data split via stratified cross-validation ensures all samples to be included in validation split and provides better understanding of the absolute results of the algorithms. The data were divided with age and neurodegenerative disorder in mind.

(Tab.1) is providing a table of the 20 best results for the combination of the scaling method, the machine learning algorithm, and the sampling method. The results are sorted by the total MCC from the highest to lowest. The total values of the provided metrics are computed based on the prediction of the samples, while the samples were in the validation fold. The ranges of the metrics are computed across the cross-validation folds.

Regardless of the sampling method, the best results yields Linear Regression scaled by Quantile Transformer. For this Algorithm, the best sampling method is ADASYN. Other valuable algorithms used for successful early detection of Parkison's disease are Ridge, and LGBM regardless of the sampling method, and SVM, LDA and Logistic Regression in combination with None or TomekLinks sampling method and in use of Quantile Transformer. Overall, Quantile Transformer has proven to be the best scaling method for this task, especially in a combination with Linear Regression.

Table 1. comparative analysis results of selected best 20 settings.

Scaling Method	ML Algorithm	Sampling Method	Accuracy	MCC	Sensitivity	Specificity
Quantile Transformer	Linear Regression	ADASYN	0.9434 (0.913-0.9826)	0.7349 (0.5253-0.9141)	0.7753 (0.4285-0.8928)	0.9664 (0.9257-0.995)
Quantile Transformer	Linear Regression	None	0.9452 (0.913-0.9782)	0.7336 (0.5253-0.8969)	0.7391 (0.4285-0.9259)	0.9733 (0.9257-0.995)
Quantile Transformer	Linear Regression	Random Over Sampler	0.9434 (0.913-0.9739)	0.7298 (0.5253-0.8741)	0.7536 (0.4285-0.8928)	0.9693 (0.9356-0.9852)
Quantile Transformer	Linear Regression	TomekLinks	0.9426 (0.9043-0.9782)	0.7265 (0.5161-0.8969)	0.7536 (0.4642-0.9259)	0.9683 (0.9207-1.0)
Quantile Transformer	Linear Regression	SMOTE	0.9408 (0.9043-0.9695)	0.7253 (0.5253-0.8614)	0.7753 (0.4285-0.9285)	0.9634 (0.9158-0.9852)
Quantile Transformer	Ridge	TomekLinks	0.9391 (0.9043-0.9608)	0.6909 (0.4856-0.7989)	0.6521 (0.4285-0.8571)	0.9782 (0.9603-1.0)
Quantile Transformer	LGBM Regressor	ADASYN	0.94 (0.9173-0.9739)	0.6841 (0.5478-0.8691)	0.5724 (0.4285-0.7777)	0.9901 (0.9801-1.0)
Quantile Transformer	SVM	TomekLinks	0.9373 (0.8956-0.9608)	0.6833 (0.4208-0.8047)	0.6521 (0.3571-0.9285)	0.9762 (0.9554-0.995)
Quantile Transformer	Linear Discriminant Analysis	None	0.9382 (0.9-0.9652)	0.6806 (0.4538-0.8228)	0.6159 (0.3928-0.8571)	0.9822 (0.9653-1.0)
MaxAbs Scaler	LGBM Regressor	SMOTE	0.9382 (0.9173-0.9565)	0.6795 (0.5744-0.7798)	0.6086 (0.4814-0.8214)	0.9832 (0.9702-1.0)
Quantile Transformer	LGBM Regressor	TomekLinks	0.9382 (0.9043-0.9608)	0.6748 (0.4395-0.8066)	0.5724 (0.2142-0.75)	0.9881 (0.9653-1.0)
MaxAbs Scaler	LGBM Classifier	TomekLinks	0.9382 (0.9086-0.9565)	0.6725 (0.4846-0.7746)	0.5434 (0.3571-0.6428)	0.992 (0.9801-1.0)
MaxAbs Scaler	LGBM Regressor	ADASYN	0.9365 (0.9043-0.9521)	0.6725 (0.4613-0.7798)	0.6159 (0.3571-0.8214)	0.9802 (0.9702-0.9901)
Standard Scaler	LGBM Regressor	ADASYN	0.9365 (0.9086-0.9608)	0.6725 (0.5046-0.8066)	0.6159 (0.4285-0.7777)	0.9802 (0.9554-1.0)
Quantile Transformer	Linear Discriminant Analysis	TomekLinks	0.9356 (0.9-0.9608)	0.6714 (0.4538-0.7982)	0.6304 (0.3928-0.8571)	0.9772 (0.9554-0.995)
Quantile Transformer	Logistic Regression	TomekLinks	0.9356 (0.8956-0.9608)	0.6686 (0.4208-0.7989)	0.6159 (0.3571-0.8571)	0.9792 (0.9653-1.0)
Quantile Transformer	Ridge	ADASYN	0.9365 (0.9-0.9608)	0.6678 (0.4401-0.7989)	0.5869 (0.3571-0.7857)	0.9841 (0.9702-1.0)
Quantile Transformer	Ridge	None	0.9365 (0.9-0.9608)	0.6678 (0.4401-0.7989)	0.5869 (0.3571-0.7857)	0.9841 (0.9702-1.0)
Quantile Transformer	Ridge	Random Over Sampler	0.9365 (0.9-0.9608)	0.6678 (0.4401-0.7989)	0.5869 (0.3571-0.7857)	0.9841 (0.9702-1.0)
Quantile Transformer	Ridge	SMOTE	0.9365 (0.9-0.9608)	0.6678 (0.4401-0.7989)	0.5869 (0.3571-0.7857)	0.9841 (0.9702-1.0)

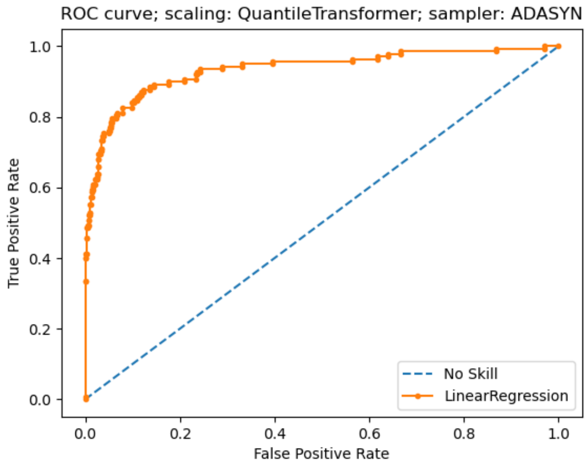


Fig.1. ROC curve of a combination of Quantile Transformer sampling method, Linear Regression machine learning. algorithm and ADASYN sampling method.

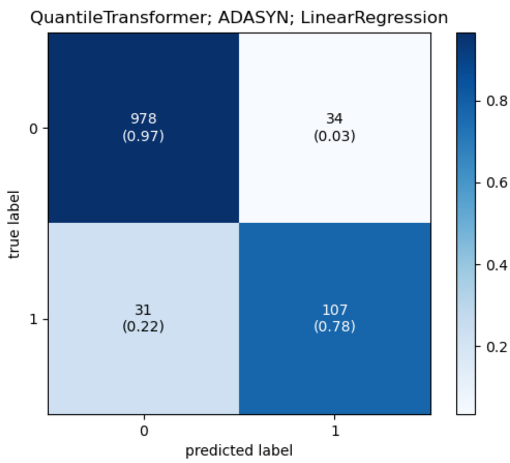


Fig.2. Confusion matrix of the selected model, linear regression with Quantile transformed data with ADASYN sampling method used.

(Fig.1) shows ROC curve of a combination of Quantile Transformer sampling method, Linear Regression machine learning algorithm and ADASYN sampling method. ROC curve shows a great decision skill of the selected algorithm when keeping a high true positive rate while keeping low false positive rate.

(Fig.2) shows a confusion matrix of the selected combination of Quantile Transformer sampling method, Linear Regression machine learning algorithm and ADASYN sampling method. As it is depicted on the cinfusion matrix, we can see the algorithm is very succcesfull determine the healthy patient, while keeping the high performance detecting a patient with PD.

4 Conclusion

In this paper, performance evaluation of machine learning algorithms in a combination with sampling and scaling methods was proposed and verified done. The aim of analysis the performance evaluation is to find the best modified algorithm available to solve the complex problem, and in this study for early detection of Parkinson's disease based on voice recordings of patients describing the images. These patients were describing the provided images divided into two groups: small and big images, differentiating on a scale of the descriptive manner. These voice recordings showing the light mental exercises were preprocessed and statistically described in a combination with basic information about a patient available. Out of all provided machine learning algorithms, and scaling and sampling methods the best performing model was chosen as a linear regression with a threshold with a Quantile transformed input data and with ADASYN sampling method resulting in 94.34% (91.30%-98.26%) accuracy, 73.49% (52.53%-91.41%) MCC, 77.53% (42.85%-89.28%) sensitivity, and 96.64% (92.57%-99.50%) specificity being able to successfully predict the Parkinson's disease in most of the positive subjects and with suspicion for healthy subjects to be monitored and followed up with a doctor. The proposed approach and efficient algorithmic processing can be a suitable and effective means for early diagnosis of neurodegenerative diseases.

Acknowledgement

This paper was supported by Early Warning of Alzheimer and other neurodegenerative diseases from Operational Programme Integrated Infrastructure (EU Cohesion Fund) - Code n. ITMS2014+/313022V631 and by Research of Advanced Algorithms and Modeling of Processes in Domains of Applied Informatics from Grant Agency Academia Aurea (GAAA) - Code n. GAAA/2022/1.

References

[1] J. Gamboa, F. J. Jiménez-Jiménez, A. Nieto, J. Montojo, M. Ortí-Pareja, J. A. Molina, E. García-Albea and I. Cobeta, "Acoustic voice analysis in patients with Parkinson's disease treated with dopaminergic drugs," *J Voice,* pp. 314-320, September 1997.

[2] A. K. Ho, J. L. Bradshaw and R. Iansek, "For better or worse: The effect of levodopa on speech in Parkinson's disease," *Mov Disord.,* vol. 23, no. 4, pp. 574-580, 15 March 2008.

[3] B. Harel, M. Cannizzaro and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study," *Brain Cogn.,* vol. 56, no. 1, pp. 24-29, October 2004.

[4] B. Sakar , M. Isenkul, C. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin and O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings" *IEEE J Biomed Health Inform,* vol. 17, no. 4, pp. 828-834, 2013.

[5] 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, 2016.

[6] A. T. Connolly, W. F. Kaemmerer, S. Dani, S. R. Stanslaski, E. Panken, M. D. Johnson and T. Denison, "Guiding deep brain stimulation contact selection using local field potentials sensed by a chronically implanted device in Parkinson's disease patients" in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2015.

[7] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge and D. C. Ackland, "Classification of Parkinson's Disease Gait Using Spatial-Temporal Gait Features" *IEEE Journal of Biomedical and Health Informatics,* vol. 19, no. 6, pp. 1794-1802, 2015.

[8] T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul and T. Wilaiprasitporn, "Parkinson's Disease Recognition Using SPECT Image and Interpretable AI: A Tutorial" *IEEE Sensors Journal,* vol. 21, no. 20, pp. 22304 - 22316, 2021.

[9] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave and E. Nöth, "Deep Learning Approach to Parkinson’s Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 2019.

[10] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas and V. H. C. de Albuquerque, "Detecting Parkinson’s disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters,* vol. 125, pp. 55-62, 2019.

[11] Z. Képešiová, Š. Kozák, E. Ružický, A. Zimmermann and R. Malaschitz, "Comparative Analysis of Advanced Machine Learning Algorithms for Early Detection of Parkinson Disease," in *2022 Cybernetics & Informatics (K&I)*, Visegrád, Hungary, 2022.

[12] B. M. Bolstad, R. A. Irizarry, M. Åstrand and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics,* vol. 19, no. 2, pp. 185-193, January 2003.

[13] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry" *Biometrika,* vol. 87, no. 4, pp. 954-959, December 2000.

[14] "Training and assessing classification rules with imbalanced data" *Data Mining and Knowledge Discovery,* vol. 28, pp. 92-122, 2014.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[16] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE International Joint Conference on Neural Networks*, 2008.

[17] D. Devi, S. k. Biswas and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance," *Pattern Recognition Letters,* vol. 93, pp. 3-12, 2017.

[18] P. Geurts, D. Ernst and L. Wehenkel, "Extremely randomized trees" *Machine Learning,* vol. 63, pp. 3-42, 2006.

[19] K. Polat, "A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests," in *Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT),* 2019.

[20] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz and Y. Singer, "Online Passive-Aggressive Algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 551-585, 2006.

[21] Z. Képešiová, J. Cigánek and Š. Kozák, "Driver Drowsiness Detection Using Convolutional Neural Networks," in *2020 Cybernetics & Informatics (K&I),* 2020.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA, USA, 2017.

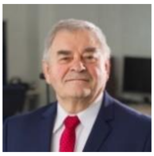
[23] J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen and Y. Xu, "Robust sparse linear discriminant analysis," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 29, no. 2, pp. 390-403, 2018.

[24] A. R. Lubis, M. Lubis and Al-Khowarizmi, "Optimization of distance formula in K-Nearest Neighbor method," *Bulletin of Electrical Engineering and Informatics,* vol. 9, no. 1, pp. 326-338, 2020.

[25] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics,* vol. 21, 2020.

Authors

Ing. Zuzana Képešiová, PhD.
Sygic, Ltd.
zuzana.kepesiova@gmail.com
research and applications worker Sygic, Ltd. as machine learning specialist. Her professional focus is mostly research and programming in intelligent systems focusing on pattern recognition, optimization, big data analysis and prediction.



prof. Ing. Štefan Kozák, PhD.
Faculty of Informatics
Pan-European University in Bratislava, Slovakia
stefan.kozak@paneurouni.com
His research interests include system theory, linear and nonlinear control methods, numerical methods and software for modeling, control, signal processing, IoT, IIoT and embedded intelligent systems for digital factory in industry and medicine.

Assoc. Prof. RNDr. Eugen Ružický, PhD.
Faculty of Informatics
Pan-European University in Bratislava, Slovakia
eugen.ruzicky@paneurouni.com
His research interests include applied informatics, system analysis, modelling, visualisation and applications in medicine.

RNDr. Alfréd Zimmermann
AXON PRO Ltd. Bratislava, Slovakia
zimmermann@axonpro.sk
He is owner and CEO of the company. Professional and scientific interests include artificial intelligence in general and natural language processing.

Richard Malaschitz
AXON PRO Ltd. Bratislava, Slovakia
richard.malaschitz@axonpro.sk
Research worker at AXON PRO, data preprocessing for research projects.