Proceedings of the Federated Conference on Computer Science and Information Systems pp. 371–380
DOI: 10.15439/2020F188
ISSN 2300-5963 ACSIS, Vol. 21

BiLSTM with Data Augmentation using Interpolation Methods to Improve Early Detection of Parkinson Disease

Olusola O. Abayomi-Alli, Robertas Damaševičius
Department of Software Engineering, Kaunas University of Technology, Kaunas, Lithuania
olusola.abayomi-alli@ktu.edu, robertas.damasevicius@ktu.lt

Rytis Maskeliūnas
Department of Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania
rytis.maskeliunas@vdu.lt

Adebayo Abayomi-Alli
Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria
abayomiallia@funaab.edu.ng

***Abstract*—The lack of dopamine in the human brain is the cause of Parkinson disease (PD) which is a degenerative disorder common globally to older citizens. However, late detection of this disease before the first clinical diagnosis has led to increased mortality rate. Research effort towards the early detection of PD has encountered challenges such as: small dataset size, class imbalance, overfitting, high false detection rate, model complexity, etc. This paper aims to improve early detection of PD using machine learning through data augmentation for very small datasets. We propose using Spline interpolation and Piecewise Cubic Hermite Interpolating Polynomial (Pchip) interpolation methods to generate synthetic data instances. We further investigate on reducing dimensionality of features for effective and real-time classification while considering computational complexity of implementation on real-life mobile phones. For classification we use Bidirectional LSTM (BiLSTM) deep learning network and compare the results with traditional machine learning algorithms like Support Vector Machine (SVM), Decision Tree, Logistic regression, KNN and Ensemble bagged tree. For experimental validation we use the Oxford Parkinson disease dataset with 195 data samples, which we have augmented with 571 synthetic data samples. The results for BiLSTM shows that even with a holdout of 90%, the model was still able to effectively recognize PD with an average accuracy for ten rounds experiment using 22 features as 82.86%, 97.1%, and 96.37% for original, augmented (Spline) and augmented (Pchip) datasets, respectively. Our results show that proposed data augmentation schemes have significantly ($p < 0.001$) improved the accuracy of PD recognition on a small dataset using both classical machine learning models and BiLSTM.**

***Index Terms*—speech impairment, Parkinson's, voice analysis; deep learning, data augmentation, interpolation, small data.**

I Introduction

PARKINSON Disease (PD) is a degenerative disorder of the central nervous system with major damage affecting the motor system in the brain cells [1]. This disease is among the most common and fastest growing neurodegenerative disorders affecting close to 7 to 10 million people globally [2-3]. It is majorly caused by the lack of dopamine (neurotransmitter) in the human brain [4] and its effect can be categorized into motor and non-motor symptoms such as voice/speech impairment, dementia, depression, slow thinking, rigidity, tremor, bradykinesia, and other cognitive disabilities [4-5]. From 60% to 90% of PD patients suffer from speech impairment such as slurred, mumbled or slow speech [6-7], among other symptoms.

Quite a number of research progress have been accounted in previous studies but the need to further explore more sophisticated algorithms of artificial intelligence (AI) methods is still ongoing with the aim of improving the health of the aged citizen through early detection of the PD disease and other diseases with similar symptoms. Several databases have been created for easing research output in the detection of neurodegenerative disorder and these databases presented in existing literature for detection of PD include dataset for detecting speech impairment (dysphonia) [1], drawing movement [8], Volatile Organic Compounds (VOCs) in blood [9], cognitive impairment [10], electroencephalohraphy (EEG) and electromyography (EMG) bio-signals [11], images such as magnetic resonance imaging (MRI), functional MRI (fMRI), positron emission tomography (PET) [12], etc. In majority of cases, once the symptoms of the neurodegenerative disorder such as PD have been validated by a medical expert, the chances of disease progress in patients becomes higher due to late detection [13]. Therefore, further research endeavors in early diagnosis of PD before it progresses any further making any medical assistance and treatment ineffective are very important [14].

The traditional methods require a lot of monitoring of living activities, motor skills, and other neurological parameters to determine the PD progress in a patient [5]. Recent advancement in AI methods have increased research focus towards adopting algorithms to enhance diagnostics of PD among patients. Existing research contributions include the implementation of mobile applications for PD diagnosis and monitoring [14-18]. The contribution of this paper is:

- Effective interpolation-based data augmentation techniques to generate synthetic data samples for training of machine learning models.
- To explore dimensionality reduction with the aim of identifying the best set of features for classification.
- Finally, to investigate and compare the performance of BiLSTM deep learning models and traditional machine learning algorithms in early detection of PD using the original (Oxford Parkinson) and augmented datasets.

The rest of the paper is organized as follows: Section II discusses in details the related works with highlights on the shortcomings of existing solutions. Section III presents the methods used in this study with an emphasis on the proposed methods and data used with introductory explanation of neural network models. Section IV describes the implementation details and the results achieved from our proposed models and presents a comparison of our results with existing studies using the same dataset. The paper concludes in Section V with future research recommendations.

II Literature Review

This section discuss the various studies tailored towards detecting and classifying PD with a focus on previous work based on speech impairment. A typical wave form variance of a healthy person and an individual suffering from speech impairment is depicted in Fig. 1.

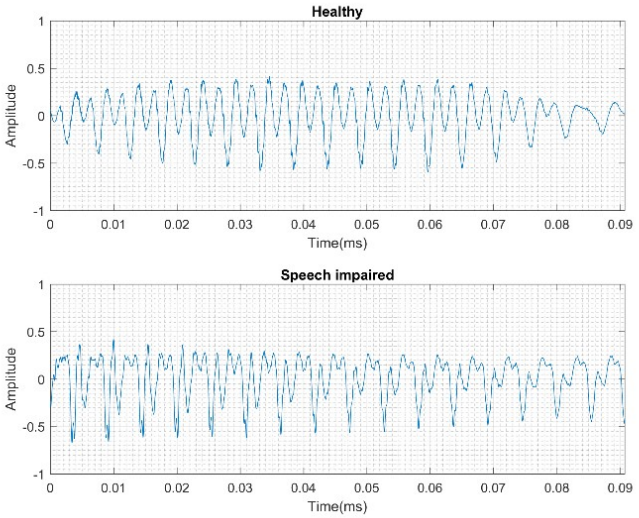

Fig. 1. A typical wave form variance of a healthy person and an individual suffering from speech impairment (data taken from the dataset described in [20,21])

A Related Studies on Speech Impairment

Previous study on early diagnosis of PD include [19], which presented an ensemble classifier based on Deep Belief Network (DBN) and Self-Organizing Map (SOM) for remote tracking of PD progress. Recent studies [20, 21] proposed a hybrid model based on bidirectional LSTM (Bi-LSTM) neural network and wavelet scattering transform (WST) and SVM classifier to detect speech impairments. Authors experimented on 15 subjects and 7 diseased subjects making up for 339 voice samples. The results showed that the proposed based on WST and SVM outperform Bi-LSTM and is expected to improve the decision systems for speech impairment detection with accuracy of 96.3%. Similar study was conducted in [22] using online handwriting dynamic signals for detection of PD. The authors investigated the impact of transfer learning and data augmentation methods, and evaluated the classification approach based on CNN-BLSTM and SVM. The study concluded that integrating data augmentation helps to achieve favourable results.

Authors in [23] introduced a bio-inspired algorithm for decision support system to evaluate voice challenges. The study compared the performance of different mathematical solution such as Fourier and Gabor transformation with bio-inspired algorithms. The result of the proposed voice analysis system using heuristic and spiking neural network gave a promising results when compared with the state-of–the–art methods with the average effectiveness as 87%. Author in [13] presented four machine learning (ML) methods for detection of PD from sustained phonation and speech signals. Authors applied eighteen feature extraction approach obtained from acoustic cardioid and smartphone recording on four ML algorithms KNN, MLP, optimum-path forest (OPF) and SVM. Authors in [1] proposed an extreme learning machine (ELM) for predicting PD. Authors in [5] presented a Principal Component Analysis (PCA) algorithm on original feature sets and other non-linear classifiers. The study gave an impressing performance of random forest accuracy as 96.87%. Authors claimed that reducing dimensionality plays an important role in improving overall classification of PD. Authors in [7] introduced the combination of Gaussian processes and automatic relevance determination for detecting PD. The study was conducted on two PD dataset and the focus was based on the using small amount of relevant acoustic features for detection. Authors in [24] presented an automatic analyses of PD using Mel-Frequency Cepstral Coefficients (MFCC), combined with Gaussian Mixture Models (GMM). In addition, authors in [25] incorporated MFCC and Intrinsic Mode Functions (IMF) for detection of PD. Authors in [26] also proposed using MFCC and glottal pulse for early detection of PD.

A number of ML algorithms have been implemented by researchers in detection of PD such as the application of supervised classification algorithms was presented in [27]. The classification result gave a peak accuracy of 85% while it is promising when compared to diagnosis accuracy of non-experts and specialists. Multiple learner for PD detection was investigated by authors in [25, 2]. Authors in [28] presented an ensemble classification methods based on random subspace classifier using kNN. While the latter [2] utilized ensemble bagging with genetic algorithm for detection of PD. Similarly, the study [29] presented a hybrid approach based on Synthetic Minority Over-Sampling Techniques (SMOTE) and Random Forest (RF) classifier for classifying PD. The overall classification result showed significant improvement with accuracy of 94.89% with the 10-fold validation test.

Furthermore, deep learning methods were applied in [30] for the diagnosis of PD. The study applied Multilayer Feedforward Neural Network (MLFNN) with Back-propagation (BP) algorithm for early detection of PD. Their experimental result gave a low specificity of 63.6% and a fair accuracy of 80% compared with other studies. Despite, the application of various techniques and methods on detecting PD, it is concluded these methods are still far from obtaining desired result in terms of accurate identification of PD [1].

TABLE I. SUMMARY OF RELATED WORKS

References	Methods		Contributions	Limitations	Type of Data
	Classification	Data Augmentation			
[31]	Convolutional Neural Network (CNN)	Jittering, scaling, rotating, permutating, magnitude warping, time-warping methods	Application of data augmentation improved generalization performance	Fluctuations in misclassification due to noisy labels.	Wearable Sensor Data (Motor State)
[32]	long-short-term memory recurrent neural network (LSTM-RNN)	doubling the number of datapoints for non-zero; converted all non-zero tremor scores to a single value (positive)	Balancing of training data	No significant improvement in accuracy; Issues with overfitting	Motion data (Tremor)
[33]	CNN	Image interpolation	Best detection rate was achieved based on sentence segments	Authors did not compare the performance with existing studies	Speech Data
[34]	Different regression models	Random resize and crop, random horizontal flip, and color jitter	Pitch-related features perform better than alternative features	The accuracy of interference need to be improved considerable.	Speech Data
[18]	CNN	magnitude perturbation, temporal perturbation, and random rotation	achieved satisfactory performance	overenrolled tremor-dominant PD subjects	wearable sensor devices

The summary of some related work that applied data augmentation techniques for PD detection is presented in Table I. Some of the challenges affecting research efforts and the performance of learning are still centered on insufficient data, noisy labels, and large intra-class variability [31]. However, there is still a need for more efficient and reliable data augmentation techniques to improve accuracy and the reliability of the detection thus reducing error rate [5].

B Data Augmentation

Data Augmentation have been successful applied in many classification application due to the fact that it leverages on small data by transforming existing samples and generating new ones [31,35]. The application of data augmenting have improved generalization of deep learning models and prevented overfitting of trained data. Some of its application areas include in face recognition system [36], motion detection system [37], etc. In addition, it also enhance deep learning model performance and overall stability of training results. The latter is especially relevant for the so called "small data problem" [38], when only little data is available for training of machine learning models.

Some application of the commonly used data augmentation techniques in image processing or signal processing, are geometric transformation which include scaling, shifting, rotation/ reflection, time wrapping and addition of noise. Recent studies in detection of PD have applied variety of data augmentation methods to accelerometer and gyroscope recordings such as magnitude scaling, rotation and magnitude scaling [18], cropping methods, window slicing, jittering, etc. [31,35]. Some of the drawbacks affecting the application of data augmentation include the need to maintain correct annotations/ labels which mostly requires expert knowledge.

Based on these, we present an effective data augmentation technique based on interpolation methods (spline and pchip) for generating synthetic values for further classification analysis. We further investigate the impact of data augmentation and feature reduction of the performance of the neural network model for detection of PD from speech data.

III METHODS AND MATERIALS

We discus the various steps involved in our proposed model as depicted in the functional block diagram in Fig. 2.

A Data Source

For this study, we used the Oxford Parkinson Disease dataset [39], which comprises of biomedical voice measurement from 31 individuals with 23 individuals suffering from PD. The dataset description is summarized in Table II and it consists of 195 voice recordings (147 PD and 48 healthy voice recordings), 22 real-value features.

Each recording was subjected to different measurements, consisting of vocal fundamental frequency (average, maximum and minimum) measured in Hertz, Multi-Dimensional Voice Program (MDVP) for percentage measurement of variations of frequency (Jitter) and amplitude (Shimmer), harmonicity measurement (HNR and NHR) and records of non-linear dynamics (NLD) features namely: correlation dimension (D2), Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA) and frequency variation measurement which include spread1, spread2 and Pitch Period Entropy (PPE). The data is divided randomly using the ratio of training and testing as 70:30, respectively.

The training dataset comprises of 103 PD and 34 Healthy which we further used in the generation of synthetic dataset. A total number of 571 synthetic data samples was generated (consisting of 320 PD and 251 Healthy) and the overall data used for training our the deep network model is 708. The testing data consist of 58 instances from the original data.

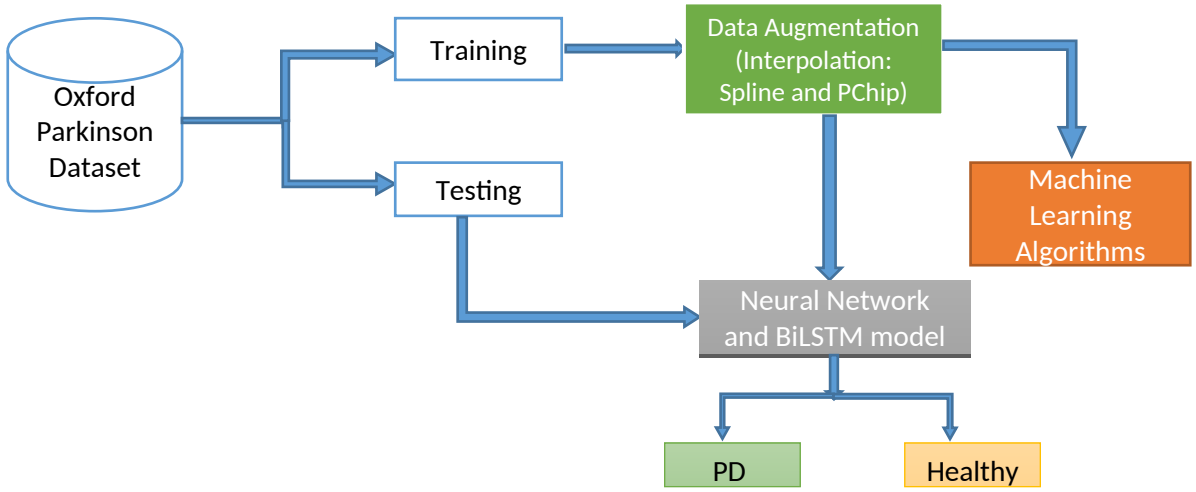

Fig. 2. Block diagram of our proposed model

B *Data augmerntation using interpolatiom*

Interpolation can be described as the method for calculating unknown values from a specified values or input with the goal of identifying analytic functions that moves through a given points to interpolate for any arbitrary point. Some of the most commonly used interpolation techniques in literature include, but are not limited to linear, polynomial, spline, pchip, nearest neighbor, multi-dimensional etc. Take an unknown function $f(x)$ assuming we are given exact values at $(n+1)$ distinct points $x_0<x_1<\ldots<x_n$ such that the values of $f(x_0), f(x_1), \ldots, f(x_n)$ are already known.

Interpolation generates a function $Q(x)$ that moves through the known points thus identifying the function with the aim of satisfying interpolation requirements (see Fig. 3) given in Eq. (1).

$$Q(x_j)=f(x_j), 0\le j\le n, \qquad (1)$$

For spline interpolation, we use a common cubic spline function. The cubic spline is a third degree derivative polynomial using a continuity conditions of spline interpolation. Therefore, spline interpolation is referred to as finding a polynomial on subintervals that are connected in a smooth manner. A spline of degree k is said to have a knots assuming we pick points of $(n+1)$ at $t_0<t_1<\ldots<t_n$. Therefore, a spline of degree k having $t_0, t_1, \ldots, t_n$ is a function of $s(x)$ which satisfies the two major properties:

- On $((t_{i-1}, t_i), s(x)$ is a polynomial of degree $\le k$, where $s(x)$ is a polynomial on every subinterval defined by the knots.
- Smoothness: $s(x)$ has a continuous $(k-1)$-th derivative on the interval $[t_0, t_n]$.

TABLE II. SUMMARY OF FEATURES (OXFORD PD DATASET [37])

Categories	Features
Vocal Fundamental Frequency	Average: MDVP:Fo(Hz), Maximum: MDVP:Fhi(Hz), Minimum:MDVP:Flo(Hz),
Frequency Parameters	MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP
Amplitude Parameters	MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA
Harmonicity Parameters	Noise-to-Harmonic (NHR), Harmonic-to-Noise (HNR)
Other Parameters	RPDE, D2: (Non-linear dynamical complexity measures), DFA: (Signal fractal scaling exponent), spread1, spread2, PPE: (Three nonlinear measures of fundamental frequency variation)

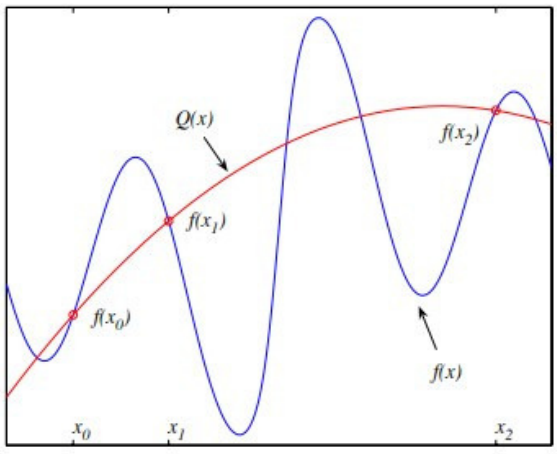

Fig. 3. A typical function $f(x)$ showing the interpolation points x_0, x_1, x_2 and the interpolating polynomial $Q(x)$ (adopted from [40])

C Neural Network Model

Neural network models are biological inspired method which defines a function as an input (set of observations) to produces an output or decision. The elements of a neural network include the input layer (X_t) with each input layer having a neuron and the weight (Ʊ), a hidden layer (H_t) and an output (Y_t). The input layer accepts signals of examination measurements which varies from ($X_{n=iton}$) while the hidden layer processes the input signals and passes them forward to the output layer for classification.

The deep learning model used in this study is a variant of recurrent neural network (RNN) known as bi-directional LSTM (BiLSTM) model (see Fig. 4). The BiLSTM model was used with the training options: Adam optimizer, maxepoch size of "250" and gradient threshold "1", initial learning rate of 0.005. We also used a verbose of 0, piecewise learning rate schedule, and the learning rate drop period, and drop factor values are 125 and 0.2, respectively.

D Performance Metrics

The experimental result was evaluated using Accuracy (percentage of true correctness of both PD patient and healthy patients), Sensitivity (percentage of PD test for PD patients), and Specificity (percentage of healthy test for healhy patients).

IV Results and Discussion

The proposed model was implemented on Matlab R2019 (MathWorks Inc., USA) using some specific toolboxes such as classification learner for analysis on supervised ML algorithms. The result of our findings is divided into two subsection and also the comparison of our work with existing study using the same dataset is presented as well.

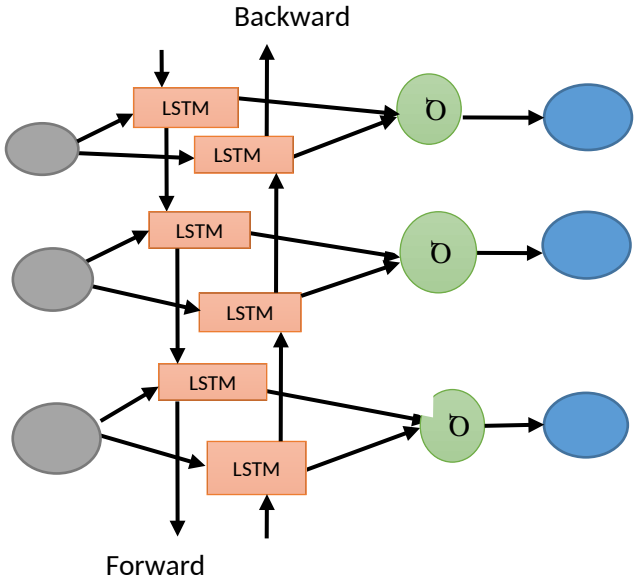

Fig. 4. The BiLSTM model used for classification

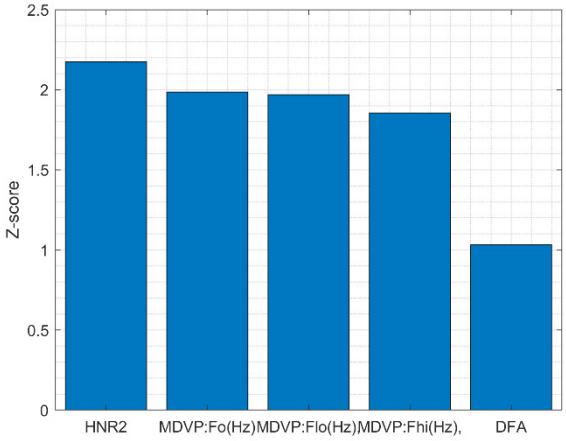

Fig. 5. The most informative features from Oxford Parkinson dataset identified using feature ranking based on non-parameteric (Wilcoxon) criterion

A Result Based on Machine Learning Algorithms

In this study, we evaluated the performance of different supervised ML algorithms such as Decision Tree, Linear Discriminant, logistic regression, SVM, KNN, and other ensemble algorithms to identify the best classifier. We investigated the performance using 5-fold cross validation. The experimental results based on the 22 features for original, augmented (spline) and augmented (pchip) datasets respectively is summarized in Table II.

Furthermore, we explored reducing feature dimensionality reduction using feature ranking based on non-parameteric (Wilcoxon) criterion and removing the highly correlated features to obtain the most relevant features. The best 5 features captured in our experiments for dimensionality reduction are: HNR2, MDVP:Fo(Hz), MDVP:Flo(Hz), MDVP:Fhi(Hz), and DFA. The results of feature ranking are presented in Fig. 5. The performance of these 5 features on traditional machine learning algorithms and BiLSTM network is presented in Table III.

Our results show that both spline and pchip augmentation was effective in allowing increasing the accuracy of classification by 4.45% ($p < 0.01$) and 4.11% ($p < 0.05$) for the 22 feature dataset, and by 4.80% ($p < 0.01$) and 1.93% (not significant) for the 5 feature dataset, when using spline and pchip augmentation, respectively (an average increase of accuracy calculated over 8 different machine learning methods). For evaluation of statistical significance, Student's two-sample t-test was used, which assumes that data are independent random samples from normal distributions.

B Results Based on BiLSTM Model

This subsection discussed the result obtained from our proposed BiLSTM model as depicted in Fig. 4. For our BiLSTM model, we used 10 % of training samples and 90% for testing for the three datasets. We selected such a small

number of training samples, which is not commonly used, in order to validate our approach for solving the small dataset problem. We also analyzed with 20 hidden units and our best performance was achiedved on 250 epoch. The summary of experiment on 20 hidden neurons and 250 epochs for holdout of 90% is presented in Fig. 6. To estimate statistical confidence limits, all experiments were repeated 10 times. Our experimental results on the three data sets show a mean accuracy for 22 features as 83.49±2.33%, 96.59±1.18% and 96.2±0.75% for original, augmented with spline interpolation (Spline) and augmented with Pchip interpolation (Pchip) datasets, respectively (assuming 95% confidence level). The performance of the BiLSTM model on spline interpolated dataset gave better results when compared with overall performance on the original and augmented (Spline) dataset.

When considering different levels of holdout, the proposed data augmentation techniques allowed to improve accuracy both in case of high holdout values when little data is available for training, and in case or small holdout, when accuracy is reduced by overfitting. The results, presented in Fig. 7 show the benefit of using data augmentation techniques for both high and low holdout values.

Our results for all machine learning models trained using an augmented (spline) dataset are summarized in Fig. 8. The results show that the best results were achieved using Weighted KNN at 98%, however, the BiLSTM also achieved comparatively good results at 97.1%. In addition, the confusion matrix the best BiLSTM model with spline augmentation using 22 features) is depicted on Fig. 9.

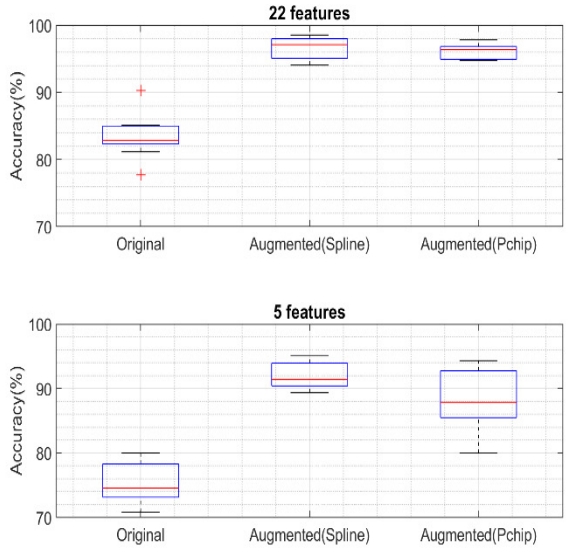

Fig. 6. Classification accuracy using original and augmented datasets and BiLSTM model with 22 and 5 dataset features

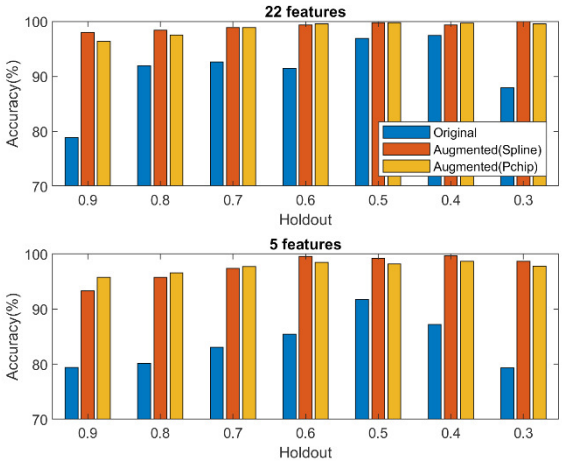

Fig. 7. Accuracy of the BiLSTM models trained with different holdout values for original and augmented datasets

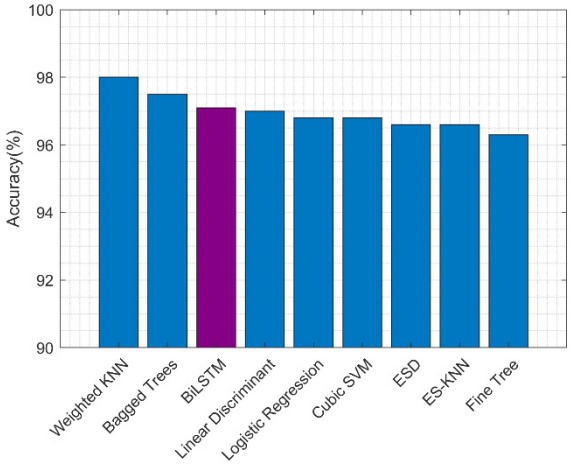

Fig. 8. Comparison of results of machine learning methods and BiLSTM model on augmented (spline) dataset

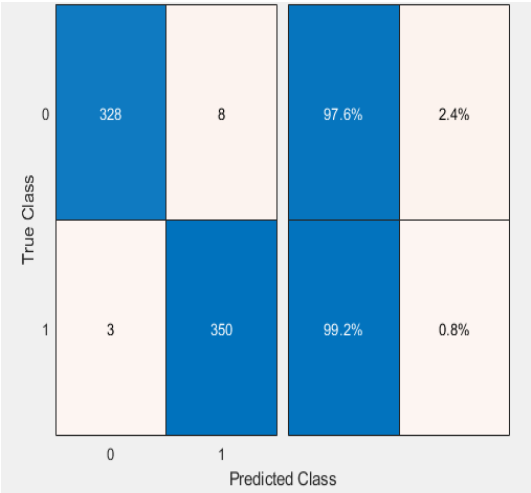

Fig. 9. Confusion matrix for best classification model using BiLSTM with 22-feature dataset augmented by spline interpolation

TABLE II.

CLASSIFICATION RESULTS ON OXFORD PARKINSON DATASET (22 FEATURES) . BEST RESULTS ARE SHOWN IN BOLD.

Algorithms	Original Data			Spline			Pchip		
	Acc (%)	Sp (%)	Sen (%)	Acc (%)	Sp (%)	Sen (%)	Acc (%)	Sp (%)	Sen (%)
Fine Tree	89.7	92.3	94.4	96.3	97.40	95.20	97.6	**97.73**	97.46
Linear Discriminant	87.2	96.3	86.5	97.0	96.37	97.74	97.0	96.37	97.74
Logistic Regression	86.2	94.4	97.7	96.8	96.36	97.18	95.3	91.47	**100**
Cubic SVM	96.6	97.8	97.8	96.8	95.85	97.74	95.1	94.43	95.76
Weighted KNN	96.6	98.8	95.5	**98.0**	**97.49**	98.59	97.3	96.92	97.74
Bagged Trees	94.0	95.6	96.6	97.5	97.46	97.46	**97.6**	97.2	98.0
Ensembled Subspace Discriminant (ESD)	92.3	93.5	96.6	96.6	94.60	**98.87**	96.5	95.57	97.46
Ensemble Subspace KNN (ES-KNN)	**97.4**	**98.9**	**97.8**	96.6	95.58	97.74	96.5	95.32	97.74

TABLE III.

CLASSIFICATION RESULTS WITH A REDUCED SET OF 5 FEATURES. BEST RESULTS ARE SHOWN IN BOLD.

Algorithms	Original Data			Spline			Pchip		
	Acc (%)	Sp (%)	Sen (%)	Acc (%)	Sp (%)	Sen (%)	Acc (%)	Sp (%)	Sen (%)
Fine Tree	94.9	96.6	96.6	97	96.9	97.12	97.6	**98.28**	96.89
Linear Discriminant	86.3	94.4	88.4	95.2	91.23	**100**	95.2	91.23	**100**
Logistic Regression	87.2	95.5	88.5	94.2	91.95	96.89	87.6	94.03	80.22
Cubic SVM	91.5	94.4	94.4	95.3	91.68	99.72	80	98.18	61.07
Weighted KNN	97.4	98.9	97.8	95.6	94.49	96.89	95.9	94.04	98.02
Bagged Trees	**94.9**	**96.6**	**96.6**	**97.5**	**97.20**	97.74	94.2	93.83	94.63
Ensembled Subspace Discriminant (ESD)	88.0	88.7	96.6	95.2	91.23	**100**	**98**	98.02	98.02
Ensemble Subspace KNN (ES-KNN)	88	87.6	96.3	96.6	95.33	98.02	95.2	91.23	**100**

TABLE IV.

COMPARISON OF CLASSIFICATION RESULTS WITH KNOWN STUDIES

Reference	Methodology	Validation Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
[5]	PCA with Random Forest (RF)	-	96.87	99.85	99.75
[7]	Gaussian Process+ 5 features	10-fold CV	96.92	99.29	90
[2]	Ensemble bagging +Genetic Algorithm (GA)	-	98.28	-	-
[27]	Multilayer Feedforward Neural Network (MLFNN) with Back-propagation (BP)	10-fold CV	80.0	63.6	83.3
[39]	Kernel support vector machine	bootstrap with 50 replicates	91.4	-	-
[41]	Rough set theory	Split validation	95.0	94.0	95.0
[42]	Hybrid Relief prior and Bacterial Foraging Optimization SVM (RF-BFO-SVM)	5-fold CV	97.42	91.50	99.29
[43]	Artificial neural networks (ANN)	10-fold CV	96.88	100	95.74
[44]	Linear kernel SVM	10-fold CV	65.12	-	-
[45]	Hybrid kernel extreme learning machine approach	average 10-fold CV	95.97	91.11	97.27
[46]	Deep Autoencoder Neural Network	-	96.11	89.78	98.15
[47]	k-NN and PCA using the created ParkDet 2.0	10-fold CV	99.1	-	-
[48]	Stability Selection method using Random Forest and Logistic Regression algorithms	5 fold CV	94.36	-	-
[49]	Complex-Valued Neural Networks and mRMR Feature Selection Algorithm	10-fold CV	98.12	98.96	99.24
Our Model	BiLSTM with Original Data	Holdout	82.86	90.5	87.97
	BiLSTM with Augmentation (Spline)	Holdout	97.1	98.78	95.57
	BiLSTM with Augmentation (Pchip)	Holdout	96.37	97.94	93.14

C Statistical analysis

To compare the performance of proposed data augmentation schemes and to assess the statistical significance of the results, we have adopted the non-parametric Friedman test and post-hoc Nemenyi test, which compare the mean ranks of the methods across multiple classification runs. The results of the Nemenyi test regarding original, augmented (Spline) and Augmented (Pchip) datasets (see Fig. 10) show that the differences between mean ranks of the methods are statistically significant (Friedman's $p < 0.001$). The Critical Difference (CD) shows the smallest difference in mean ranks, where the difference is not statistically significant. Note that for the 22-feature dataset both Spline and Pchip augmentation schemes work equally well, but significantly better than using the original dataset without augmentations. The same observation is confirmed for the 5-feature dataset: both Spline and Pchip augmentation schemes allow to achieve significantly better results as compared to the results without augmentation. In the latter case, the Spline-based augmentation works better than the Pchip augmentation, but the difference is not significant (the difference between mean ranks is smaller than the CD value).

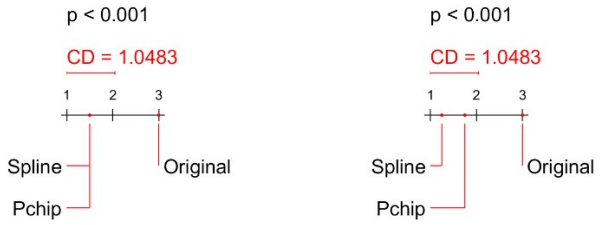

Fig. 10. Critical distance diagrams for full dataset (22 features) (left) and reduced feature dataset (5 features) (right). CD – critical distance.

D Comparison with Existing Work

For the purpose of validating our proposed method, we compared our results with previous work as shown in Table IV. Considering that the PD dataset utilized in this study has 22 features, we summarized and compared the result based on the original dataset as well as with the features obtained using feature selection. The comparison table shows the various algorithm proposed in literature with our proposed model. Our experimental results using BiLSTM with Augmentation (Spline) achieved a significant improvement in accuracy, specificity and sensitivity. However, some existing methods such as proposed in [2],[42],[47],[49] achieved accuracy between 97.42-99.1% and thereby outperformed our method. Some of limitations of this state-of-the-art methods is increasing computational complexity. Therefore, we can argue that our proposed model reveals a simple and effective method for detecting PD.

One key limitations of our data augmentation method (Interpolation) is generating noisy (out of range) values. This play a major role in affecting the performance of our model. Thus, further study is to consider more diversity among data augmentation techniques with the aim of reducing noise and error rates, and improving performances.

E Evaluation and discussion

Data augmentation using the interpolation methods allowed us to increase the accuracy of PD recognition using voice data. The application of interpolation effectively increases the resolution of captured signal, which allows to recover some of the information lost due to the microphone sampling rate that is lower than needed to solve this task. In the datased we used (Oxford Parkinson [39]), voice data was captured using a microphone with a sampling rate of 44.1 kHz. However, the study [50] concluded that a sampling rate of 96 kHz is preferred for effective PD recognition, which makes the interpolation techniques an attractive method for dealing with low resolution voice data.

Another advantage of data augmentation is the ability to increase data volume for model training. Effective training of neural networks, especially deep learning models, usually require having large amounts of data. However, in case of niche applications, such as diagnosing rare diseases, the datasets are usually small. Generation of the synthetic (surrogate) data for training allows to obtain better models, thus increasing the accuracy of classification, as also was demonstrated in this paper.

V Conclusion

The need to increase available data for classification when using small datasets with the aim of improving recognition of Parkinson disease (PD) cannot be over-emphasized. This paper effectively applied the interpolation (spline and pchip) methods for the generation of synthetic data instances thus increasing the learning samples available for training of machine learning models and improving the classification performance. This study was able to effectively address the problem of class imbalance by augmenting the original data samples using the interpolation method. Two interpolation techniques (spline and pchip) were used to generate synthetic data. A total number of 571samples was generated by each technique consisting of 320 Healthy and 251 Parkinson disease samples.

This paper investigated the performance of traditional machine learning algorithms and BiLSTM model in classifying the three categories of data samples. Our results showed that for an efficient and simple data augmentation technique based on spline and pchip interpolation have proven to be effective in the detection of PD. The analysis results for BiLSTM shows that even with a holdout of 90% for testing, the model was still able to effectively classify PD on three datasets (original Oxford Parkinson dataset, original dataset augmented using spline interpolation and original dataset using pchip intermolation) with an average accuracy

of 82.86%, 97.1%, and 96.37% for the original, spline and pchip datasets, respectively (all 22 features were used). Further experiments was carried out for feature dimensionality reduction and the best results were obtained on 5 features and the the average accuracy on the 90% holdout was 74.14%, 91.44%, and 87.88% for for the original, spline and pchip datasets, respectively. The experimental results using spline augmentation have shown statistically significant ($p < 0.001$ using Friedman's test) consistency in impoving the accuracy for both 22 feature and 5 feature datasets.

The comparison of our results with the existing studies shows that the application data augmentation did not only improved accuracy, but was also able to reduce overfitting and improve the overall performance. This study was able to apply a simple BiLSTM model to effective classify speech impairment which will efficiently enhance the early detection of PD. Our proposed model based on using data augmentation techniques for small datasets showed a significant improvement in accuracy, when only a small amount of data is available for training. Note that we simulated a small dataset using an extreme value of 90% holdout for training data, which has not been used by other authors before.

Future recommendation is to explore other data augmentation methods based on different AI methods and architectural frameworks with the aim of developing an intelligent model for speech recognition for small datasets.

REFERENCES

[1] Agarwal, A., Chandrayan, S. and Sahu, S.S., 2016. Prediction of Parkinson's disease using speech signal with Extreme Learning Machine. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 3776-3779). IEEE. Doi: 10.1109/ICEEOT.2016.7755419

[2] Fayyazifar, N. and Samadiani, N., 2017. Parkinson's disease detection using ensemble techniques and genetic algorithm. In 2017 Artificial Intelligence and Signal Processing Conference (AISP) (pp. 162-165). IEEE. Doi: 10.1109/AISP.2017.8324074

[3] Dorsey, E.R., Elbaz, A., Nichols, E., Abd-Allah, F., Abdelalim, A., Adsuar, J.C., Ansha, M.G., Brayne, C., Choi, J.Y.J., Collado-Mateo, D. and Dahodwala, N., 2018. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology, 17(11), 939-953. doi:10.1016/S1474-4422(18)30295-3

[4] Saikia, A., Majhi, V., Hussain, M. and Paul, S., 2019. A Systematic review on Application based Parkinson's disease Detection Systems. International Journal on Emerging Technologies 10(3): 166-173.

[5] Aich, S., Younga, K., Hui, K.L., Al-Absi, A.A. and Sain, M., 2018, February. A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data. In 2018 20th International Conference on Advanced Communication Technology (ICACT) (pp. 638-642). IEEE. Doi: 10.23919/ICACT.2018.8323864

[6] Chan, M.Y., Chu, S.Y., Ahmad, K. and Ibrahim, N.M., 2019. Voice therapy for Parkinson's disease via smartphone videoconference in Malaysia: A preliminary study. Journal of telemedicine and telecare, doi:10.1177/1357633X19870913

[7] Despotovic, V., Skovranek, T. and Schommer, C., 2020. Speech Based Estimation of Parkinson's Disease Using Gaussian Processes and Automatic Relevance Determination. Neurocomputing, , 401, 173–181. doi:10.1016/j.neucom.2020.03.058

[8] Gil-Martín, M., Montero, J.M. and San-Segundo, R., 2019. Parkinson's disease detection from drawing movements using convolutional neural networks. Electronics, 8(8), p.907. doi:10.3390/electronics8080907

[9] Lavner, Y., Khatib, S., Artoul, F. and Vaya, J., 2014, December. An algorithm for processing and analysis of Gas Chromatography-Mass Spectrometry (GC-MS) signals for early detection of Parkinson's disease. In 2014 IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI) (pp. 1-5). IEEE. Doi: 10.1109/EEEI.2014.7005772

[10] Liu, H.J., Li, X.Y., Chen, H., Yu, H.L., Tao, Q.Q. and Wu, Z.Y., 2020. Identification of susceptibility loci for cognitive impairment in a cohort of Han Chinese patients with Parkinson's disease. Neuroscience Letters, 135034. Doi: doi:10.1016/j.neulet.2020.135034

[11] Saikia, A., Hussain, M., Barua, A.R. and Paul, S., 2019. EEG-EMG correlation for parkinson's disease. International Journal of Engineering and Advanced Technology, 8(6), pp.1179-85. Doi: 10.35940/ijeat.F8360.088619

[12] Rumman, M., Tasneem, A.N., Farzana, S., Pavel, M.I. and Alam, M.A., 2018. Early detection of Parkinson's disease using image processing and artificial neural network. In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 256-261). IEEE. Doi: 10.1109/ICIEV.2018.8641081

[13] Almeida, J.S., Rebouças Filho, P.P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R. and de Albuquerque, V.H.C., 2019. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters, 125, pp. 55-62. doi:10.1016/j.patrec.2019.04.005

[14] Lauraitis, A., Maskeliūnas, R., Damaševičius, R., Połap, D. and Woźniak, M., 2019. A smartphone application for automated decision support in cognitive task based evaluation of central nervous system motor disorders. IEEE journal of biomedical and health informatics, 23(5), pp. 1865-1876. Doi: 10.1109/JBHI.2019.2891729

[15] Gatsios, D., Antonini, A., Gentile, G., Marcante, A., Pellicano, C., Macchiusi, L., Assogna, F., Spalletta, G., Gage, H., Touray, M. and Timotijevic, L., 2020. Mhealth for remote monitoring and management of Parkinson's disease: determinants of compliance and validation of a tremor evaluation method. JMIR mHealth and uHealth.

[16] Linares-Del Rey, M., Vela-Desojo, L. and Cano-de la Cuerda, R., 2019. Mobile phone applications in Parkinson's disease: a systematic review. Neurología (English Edition), 34(1), pp. 38-54. doi:10.1016/j.nrleng.2018.12.002

[17] Zhang, H., Song, C., Rathore, A.S., Huang, M., Zhang, Y. and Xu, W., 2020. mHealth Technologies towards Parkinson's Disease Detection and Monitoring in Daily Life: A Comprehensive Review. IEEE Reviews in Biomedical Engineering. DOI: 10.1109/RBME.2020.2991813

[18] Zhang, H., Deng, K., Li, H., Albin, R.L. and Guan, Y., 2020. Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson's Disease. Patterns, 100042. doi:10.1016/j.patter.2020.100042

[19] Nilashi, M., Ahmadi, H., Sheikhtaheri, A., Naemi, R., Alotaibi, R., Alarood, A.A., Munshi, A., Rashid, T.A. and Zhao, J., 2020. Remote Tracking of Parkinson's Disease Progression Using Ensembles of Deep Belief Network and Self-Organizing Map. Expert Systems with Applications, 113562. doi:10.1016/j.eswa.2020.113562

[20] Lauraitis, A., Maskeliūnas, R., Damaševičius, R. and Krilavičius, T., 2020. Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features. IEEE Access, 8, 96162 – 96172. Doi: 10.1109/ACCESS.2020.2995737

[21] Lauraitis, A., Maskeliūnas, R., Damaševičius, R. and Krilavičius, T., 2020. A Mobile Application for Smart Computer-Aided Self-Administered Testing of Cognition, Speech, and Motor Impairment. Sensors, 20, 3236. doi:10.3390/s20113236

[22] Taleb, C., Likforman-Sulem, L. and Mokbel, C., 2019. Improving Deep Learning Parkinson's Disease Detection Through Data Augmentation Training. In Mediterranean Conference on Pattern Recognition and Artificial Intelligence (pp. 79-93). Springer, Cham. Doi:

[23] Połap, D., Woźniak, M., Damaševičius, R. and Maskeliūnas, R., 2019. Bio-inspired voice evaluation mechanism. Applied Soft Computing, 80, pp. 342-357. doi:10.1016/j.asoc.2019.04.006

[24] Jeancolas, L., Benali, H., Benkelfat, B.E., Mangone, G., Corvol, J.C., Vidailhet, M., Lehericy, S. and Petrovska-Delacrétaz, D., 2017. Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients. In

2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-6). IEEE. Doi: 10.1109/ATSIP.2017.8075567

[25] Rueda, A. and Krishnan, S., 2017. Feature analysis of dysphonia speech for monitoring Parkinson's disease. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2308-2311). IEEE. Doi: 10.1109/EMBC.2017.8037317

[26] Vikas, and Sharma, R.K. 2014, May. Early detection of Parkinson's disease through Voice. In 2014 International Conference on Advances in Engineering and Technology (ICAET) (pp. 1-5). IEEE. Doi: 10.1109/ICAET.2014.7105237

[27] Wroge, T.J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D.C. and Ghomi, R.H., 2018, December. Parkinson's disease diagnosis using machine learning and voice. In 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1-7). IEEE. Doi: 10.1109/SPMB.2018.8615607

[28] Eskıdere, Ö., Karatutlu, A. and Ünal, C., 2015, September. Detection of Parkinson's disease from vocal features using random subspace classifier ensemble. In 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-4). IEEE. Doi: 10.1109/ICECCO.2015.7416886

[29] Polat, K., 2019. A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests. In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (pp. 1-3). IEEE. Doi: 10.1109/EBBT.2019.8741725

[30] Olanrewaju, R.F., Sahari, N.S., Musa, A.A. and Hakiem, N., 2014. Application of neural networks in early detection and diagnosis of Parkinson's disease. In 2014 International Conference on Cyber and IT Service Management (CITSM) (pp. 78-82). IEEE. Doi: 10.1109/CITSM.2014.7042180

[31] Um, T.T., Pfister, F.M., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U. and Kulić, D., 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 216-220). doi:10.1145/3136755.3136817

[32] Bourdillon, A., Sawhney, K., Mehra, R., O'Grady, P. and Liu, T., Extracting kinetic features from wearable tech for clinical symptoms of Parkinsons Disease.

[33] Vaiciukynas, E., Gelzinis, A., Verikas, A. and Bacauskiene, M., 2017. Parkinson's disease detection from speech using convolutional neural networks. In International Conference on Smart Objects and Technologies for Social Good (pp. 206-215). Springer, Cham.

[34] Bayestehtashk, A., Asgari, M., Shafran, I. and McNames, J., 2015. Fully automated assessment of the severity of Parkinson's disease from speech. Computer speech & language, 29(1), pp.172-185. Doi: 10.1016/j.csl.2013.12.001

[35] Pan Q., Li, X., and Fang L. 2020. Data Augmentation of Deep learning-based on ECG Analysis. Feature Engineering and Computational Intelligence in ECG Monitoring, 91-111. Springer Nature Singapore Pte. Doi: 10.1007/978-981-15-3824-7_6

[36] Kutlugün, M.A., Sirin, Y. and Karakaya, M., 2019. The Effects of Augmented Training Dataset on Performance of Convolutional Neural Networks in Face Recognition System. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 929-932). IEEE. Doi: 10.15439/2019F181

[37] Lee, J.W., Nam, D.W., Yoo, W.Y., Kim, Y., Jeong, M. and Kim, C., 2018. Soccer object motion recognition based on 3D convolutional neural networks. In FedCSIS (Communication Papers) (pp. 129-134). Doi: 10.15439/2018F48

[38] Li, Z., Yao, H., and Ma, F. (2020). Learning with Small Data. Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20. doi:10.1145/3336191.3371874

[39] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. 2009. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on bio-medical engineering, 56(4), 1015. doi:10.1038/npre.2008.2298.1

[40] Levy, D., 2010. Introduction to numerical analysis. Department of Mathematics and Center for Scientific Computation and Mathematical Modeling (CSCAMM) University of Maryland, pp.2-2.

[41] Revett, Kenneth, Florin Gorunescu, and Abdel-Badeeh Mohamed Salem. "Feature selection in Parkinson's disease: A rough sets approach." In 2009 International Multiconference on Computer Science and Information Technology, pp. 425-428. IEEE, 2009. Doi: 10.1109/IMCSIT.2009.5352688

[42] Cai, Z., Gu, J. and Chen, H.L., 2017. A new hybrid intelligent framework for predicting Parkinson's disease. IEEE Access, 5, pp.17188-17200. Doi: 10.1109/ACCESS.2017.2741521.

[43] Wang, X., 2014. Data Mining Analysis of the Parkinson's Disease. Masters thesis Submitted to the College of Arts and Sciences, Georgia State University.

[44] Bhattacharya, I. and Bhatia, M.P.S., 2010. SVM classification to distinguish Parkinson disease patients. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (pp. 1-6). Doi: 10.1145/1858378.1858392

[45] Chen, H.L., Wang, G., Ma, C., Cai, Z.N., Liu, W.B. and Wang, S.J., 2016. An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson′ s disease. Neurocomputing, 184, pp.131-144. doi:10.1016/j.neucom.2015.07.138

[46] Kose, U., Deperlioglu, O., Alzubi, J. and Patrut, B., Diagnosing Parkinson by Using Deep Autoencoder Neural Network. In Deep Learning for Medical Decision Support Systems (pp. 73-93). Springer, Singapore. doi:10.1007/978-981-15-6325-6_5

[47] Ozkan, H., 2016. A comparison of classification methods for telediagnosis of Parkinson's disease. Entropy, 18(4), p.115. Doi: 10.3390/e18040115

[48] Akyol, K., Bayir, Ş. and Baha, Ş.E.N., Importance of Attribute Selection for Parkinson Disease. Akademik Platform Mühendislik ve Fen Bilimleri Dergisi, 8(1), pp.175-180. Doi: 10.21541/apjes.541637

[49] Peker, M., Sen, B. and Delen, D., 2015. Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm. Journal of healthcare engineering, 6. Doi: 10.1260/2040-2295.6.3.281

[50] Wu, K., Zhang, D., Lu, G., and Guo, Z. 2018. Influence of sampling rate on voice analysis for assessment of Parkinson's disease. The Journal of the Acoustical Society of America, 144(3), 1416–1423. doi:10.1121/1.5053681