ISSN: 0974-8571 Vol.13 No. 1 August, 2021

International Journal of Computational Intelligence in Control

Machine Learning Based Parkinson's Disease Diagnosis using Hand Writing Related Variables

Atiqur Rahman[1], Sardar Ali[2,3], Amjad Iqbal[4], Shafqat Ullah Khan[5], Aurangzeb Khan[1], Yakubu Imrana[6], Arif Ali[1]
[1]Department of Computer Science, University of Science and Technology Bann, Pakistan, atiqrehman37@gmail.com, aurangzeb.ustb@gmail.com dr.arifali92@gmail.com
[2]Department of Electronics, University of Buner, Pakistan, sardar33@gmail.com
[3]School of Engineering and Applied Sciences, ISRA University Islamabad, Pakistan
[4]Department of Neurology, Lady Reading Hospital, Peshawar, Pakistan , dramjadiqbal2013@gmail.com
[5]Department of Electronics, University of Buner, Pakistan, shafqatphy@yahoo.com
[6]University of Electronic Science and Technology of China (UESTC), China, 201914080126@std.uestc.edu.cn

Date of Submission: 15th March 2021 Revised: 27th May 2021 Accepted: 8th August 2021

***Abstract* - Parkinson's disease (PD) is a well-known neurodegenerative condition that affects the central nervous system (CNS). PD patients cannot be diagnosed through clinical examination. However, it has been noted that such patients have difficulty in writing. This characteristic of PD patients has been utilized by researchers to present various computer vision and machine learning (ML) based approaches like micrography etc. for its non-invasive detection. However, the presented approaches are associated with two major problems- biasedness in the models and low accuracy rate of classification. The former is triggered by imbalanced data, which means that the performance of ML models is dependent on the classes of the data; such that poor performance on minority class and good performance on majority class. To deal with it, in this paper we propose to use under-sampling method. To deal with the second problem, we propose integration of Chi2 model with deep neural network (DNN). Obtained results are good when compared with previous methods.**

Index Terms - Machine learning, feature selection, deep neural network, handwritten variables, Parkinson's disease.

1. INTRODUCTION

Parkinson's disease (PD) is a common neurological disease that affects central nervous system [1]. It is found mostly in elderly persons above the ages of sixty [2]. Among the many symptoms, PD patients suffer from are tiredness in movement, vocal impairments, rigid behavior, tremor in actions, and weak balance [3]. Despite PD being so common, there is no definitive diagnostic process for it through clinical measures [4]. However, there are indications that PD patients have problems in handwriting. Therefore, various methods based on computer vision and machine learning (ML) using micrography and visioning procedures have been utilized by researchers for PD detection through writing examinations.

Recent research has shown state-of-the-art performance in many disease detection and mortality prediction problems using automated methods. These automated methods were developed using machine learning and deep learning methods. For example, expert systems were developed for hepatitis disease detection [5], mortality prediction [6,7], heart disease detection [8-11], hepatocellular carcinoma detection [12], lung cancer [13] etc. Ali et al. delineated moving hands in the air while writing has noticeable effect on the accuracy of PD detection [14]. Rosenblum et al. demonstrated in [15] that by undertaking handwriting examinations, patients with PD can be identified among stable (healthy) people. To this end, they carried out a study of 20 persons with PD and 20 persons with no PD symptoms. Using a digital table, each person was required to put down his or her name together with an address in a written form on paper. They utilized expected values of velocity and pressure related metrics and obtained detection accuracy of up to 97.5%, 100% specificity, and 95% sensitivity. However, the size of the dataset used in the experimentation was limited which mitigated significance of these studies. To address this issue, Ali et al. used data collected from 37 persons with PD symptoms and from 38 control subjects each one performing eight various handwriting activity [14]. The authors developed well-known ML models which included the support vector machine (SVM), Adaboost (AD) ensemble model, and the k-nearest neighbors (KNN), and obtained 81.3% classification accuracy. Pereira et al. proposed ML and computer vision based methodologies to

enhance PD detection accuracy. To this end, Pereira et al. in [16] used data of spiral drawings collected from 55 subjects, pertaining 37 persons with PD symptoms and 18 persons with no PD symptoms. The used SVM, Naïve Bayes (NB), and optimum path forest (OPF) supervised models for distinguishing between writing of PD patients and that of persons with no PD symptoms. By use of this method, nearly 78.9% accuracy with NB model is achieved. The authors further obtained another dataset, named HandPD, collected from 18 persons with no PD symptoms and 74 persons with PD symptoms. HandPD is containing spiral and meander drawings and is perceived to be the largest publicly available handwritten dataset developed so far in terms of sample space of up to 736 samples. HandPD is developed by extracting features from handwritten drawings using computer vision based methodologies. Moreover, classification is carried out using OPF, SVM, and NB. Following the above framework and dataset, the authors achieved 67% classification accuracy. However, there are certain limitations associated with HandPD dataset, as it leads towards biasedness in the modeling and mitigation in the PD detection accuracy.

To address the problems of previously developed models been bias and having low accuracy rate in detecting PD, we develop our methodology in this paper. To show the impact of biasedness in the models, we carry out development of four diverse ML based models, namely, KNN, linear discriminant analysis (LDA), Decision Tree (DT) and the Gaussian Naïve Bayes (GNB). To counter the effect of this issue, we use the method of random under-sampling. Similarly, we combat the issue of low accuracy rate of detecting PD, by proposing a deep learning-based method. Through experimental findings, we demonstrate that our suggested approach will minimize the impact of both concerns considerably.

The following is the organization of the remaining parts of this paper: Section 2 elaborates the dataset used in our model. Section 3 elaborates the problems and the proposed solutions. Section 4 validates and evaluates the proposed solution. Section 5 presents the experiments and discussion. Section 6 concludes the paper.

2. DATASET

In this study, we collected a dataset having Urdu language sentences from Healthy Group (HG) of people and from people having Parkinson 's disease (PDG). The database was collected after the approval of ethical review board of Lady Reading Hospital (Medical Teaching Institution), Pakistan. The main database included four modalities i.e. voice data recorded using SONY digital voice recorder, voice data collected using smart phones, Urdu language hand written sentences and hand written drawings dataset. However, in this study we only analyzed hand written Urdu sentences dataset. The database was collected from 89 subjects (29 PD patients and 60 age matched healthy subjects). Figure 1 shows data form filled by a healthy subject and Figure 2 shows data form filled by a PD patient. The sentence in Urdu language means "knowledge is power" and can be read as "Ilm barhi dowlat hai". Moreover, the database can be obtained from the corresponding author upon request.

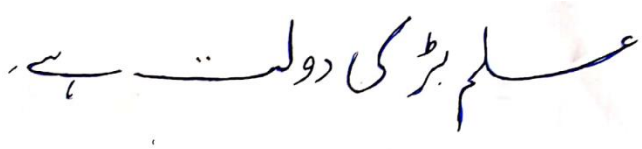

Figure 1: Data form filled by a healthy subject

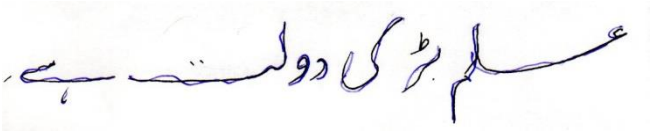

Figure 2: Data form filled by a PD patient.

3. PROBLEMS AND PROPOSED SOLUTIONS

As discussed above, number of automated methods have been proposed by different researchers to present various computer vision and machine learning (ML) based approaches like micrography etc. for its non-invasive detection. However, the presented approaches are associated with two major concerns which include biasedness and low accuracy rate of classification. The former is triggered by imbalanced data, which means that the performance of ML models is dependent on the classes of the data; such that poor performance on minority class and good performance on majority class. To deal with it, in this paper we suggest using the under-sampling approach. To deal with the second concern, we propose the integration of Chi2 model with artificial neural network (ANN). Initially, we use pixel level feature by vectoring them and then extract meaningful feature using autoencoders. Then Chi2 model is used for selecting highly relevant features. Chi2 is a statistical approach for feature selection that uses univariate analysis of each feature and the features are based on their importance. For more details about Chi2 statistics, readers are referred to [3, 14]. Here we select the features using best first strategy. The selected features are given to ANN for classification. The proposed approach and its flow diagram is given in Figure 3. The formulation of the ANN is given as follows:

We denoted the dataset used in the study with D. Moreover, we use the terms train, validation and test sets for D_train, D_valid, and D_test, respectively, whenever we intend to make distinction among the sets. We perform model and hyper parameter selection through validation set, evaluate the final generalization error through test set. This way, we compare various algorithms impartially.

The literature largely discuss classification problems in which the data set D denotes an indexed set containing pairs (x(i),y(i)). The superscripts are used to differentiate the training set examples such that x(i) ∈ R^D represent the i_th training example having D as dimension. In the same manner, y(i)∈ {0,… L} denote the i_th label originally assigned to input x(i).

Math Conventions

i. W: denotes Matrix.
ii. W_{ij}: an element in i_{th} row and j_{th} column of matrix W
iii. $W_{i\cdot}, W_i$: vector, i_{th} row of matrix W
iv. $W_{\cdot j}$: vector, j_{th}columnof matrix W
v. b: the symbols written in lower case denote a vector unless we specifically state otherwise
vi. b_i: i_{th}member of vector b

List of Symbols and acronyms

i. D: the number of input dimensions
ii. $D_h^{(i)}$: the number of hidden units in the i_{th} layer
iii. $f_\theta(x), f(x)$: the classification function used for a model $P(Y|x,\theta)$. This is defined as $argmax_k\, P(Y=k|x,\theta)$. It is worth noting that we will frequently disuse the θ subscript
iv. L: denotes the number of labels
v. $L(\theta, D)$: Log likelihood D of the model defined θ
vi. $l(\theta, D)$: empirical loss of the prediction function f parameterized by θ on data set D.
vii. NLL: negative log likelihood.
viii. θ: set of all parameters for a given model

Deep learning is exciting largely due to unsupervised learning of deep networks, in addition to a huge role of supervised learning [17]. In order to evaluate the of unsupervised pre training, it is often dependent on the fact that how much performance gain can be achieved after supervised fine tuning. In this section, we review the fundamentals of supervised learning for classification models, and study the minibatch stochastic gradient descent algorithm, which is utilized for fine-tuning multiple models in deep learning [18-20].

The main objective to train a DNN model is minimizing the number of errors, also called zero-one loss, on unnoticed examples. Given that f:R^D → {0,….L} denote the prediction function, then the loss can be represented by :

$$l_{0,1} = \sum_{i=0}^{|D|} I_{f(x^{(i)})} \neq y^{(i)} \qquad (1)$$

it is worth noting that the notation D either represent the training set or it is an intersection with D_train such that D ∩ D_train=0 (in order for skipping biasedness in the evaluation of test or validation error, where I being the indicator function and is represented as

$$I_x = \{1\ if\ x\ is\ true\ 0\ otherwise \qquad (2)$$

Similarly, f is defined as

$$f(x) = argmax_k\, P(x,\theta) \qquad (3)$$

As the zero-one loss cannot be differentiate, its optimization for large modes, where millions of parameters are involved, is too much computationally expensive. Therefore, we maximize the log-likelihood of our classifier (DNN) provided all the labels in a training set.

$$L(\theta, D) = \sum_{i=0}^{|D|} log\ log\ P\left(x^{(i)}, \theta\right) \qquad (4)$$

The likelihood the correct class carry is different from the number of correct predictions. However, if seen from the standpoint of a randomly initiated DNN, they are largely identical. It is worth noting that both likelihood and zero-one loss are dissimilar objectives, and one can see that they both are most often correlated on the validation set; however, they sometimes vary and rise or fall in accordance to each other. In this way, learning will minimize the negative log-likelihood expressed as

$$NLL(\theta, D) = -\sum_{i=0}^{|D|} log\ log\ P\left(x^{(i)}, \theta\right) \qquad (5)$$

It is worth noting that the NLL used in our classifier is differentiable for the zero one loss. Therefore, as a classifier, this function's gradient over our training data can be utilized for deep learning as a signal of supervised learning. To solve Eq 5, we use stochastic gradient descent in this paper.

Stochastic gradient descent assumes the same principle of working as that of standard gradient descent but advances more rapidly by approximating the gradient value from a few instances only at any point in time rather than to examine the whole training set. The recommended variant for deep learning is an additional twist on stochastic gradient descent with the help of minibatches. Except that multiple training examples are used in minibatches to approximate the gradient, with the consequence of reducing variance in calculating the gradient and allowing greater utilization of memory management hierarchically in current computers, the concept of minibatch operation is analogous to stochastic gradient descent.

There is, however, a tradeoff in selecting a minibatch with the size of B. With reduced variance and with usage of SIMD instructions assist in a better way when B is increased

from 1 to 2, yet this fraction of improvement rapidly drops to zero improvement. Moreover, when the size of B is kept larger, the time taken to reduce the variance of the gradient estimator grows high and thus get wasted. This large amount of time can be efficiently spent on utilizing more gradient steps. In this study, we used B value to be 20.

One usual problem with DNN models is overfitting. Training our model from data means that the model is being prepared for performing well in new examples different from those it has already undergone. The training loop for MSGP does not consider this factor and leads to be overfit the training examples. In this paper, to solve overfitting problem we use regularization (i.e, L1 and L2).

The regularization L1 and L2 , add to the loss function an additional term. This addition compromises configuarion of some parameters. Let our loss function is expressed as

$$NLL(\theta, D) = -\sum_{i=0}^{|D|} log\ log\ P\left(x^{(i)}, \theta\right) \quad (6)$$

The loss regularized can then be express as

$$E(\theta, D) = NLL(\theta, D) + \lambda\, R(\theta) \quad (7)$$

In our case it can be expressed as

$$E(\theta, D) = NLL(\theta, D) + \lambda \left|\left|\theta\right|\right|_P^p \quad (8)$$

where

$$\left|\left|\theta\right|\right|_P = \left(\sum_{j=0}^{|\theta|} |\theta_j|^P\right)^{\frac{1}{P}} \quad (9)$$

Which denotes the L_P norm of θ. The notation λ represent a hybper parameter which manage the relative significance of the regularization parameter. The most common values for p fixed are 1 and 2, which is why L1/L2 names are used. Moreover, if the value of p is set to be 2, then term weight decay is used for the regularizer.

Normally, in neural networks, the terms (loss and regularization) are added to produce an unobstructed mapping of the network. However, this comes with compromising high parameter values, which mitigates the non-linearity level modelled by a network. Stated in a simplified way, NLL and R(θ) results in a well data modelling (NLL) and smooth solutions (R(θ)). Therefore, if the sum of both terms is reduced, this means, in principle, that there is a favorable trade-off between the generalization of an eventual solution and fitting of the training data. According to Occam's razor principle, the minimization thus obtained should lead us to find the simplest solution, measured with respect to simple criterion, for the solution to be fitting the training data.

It is worth noting that simple solution does not imply well generalization. It has been found through empirical measurements that conducting regularization in this way in the context of network assist in generalization particularly in small sized datasets. This factor motivated us to use L1 and L2 regularization to avoid overfitting of the DNN model. The working of the automated PD detection system is shown in Figure 3.

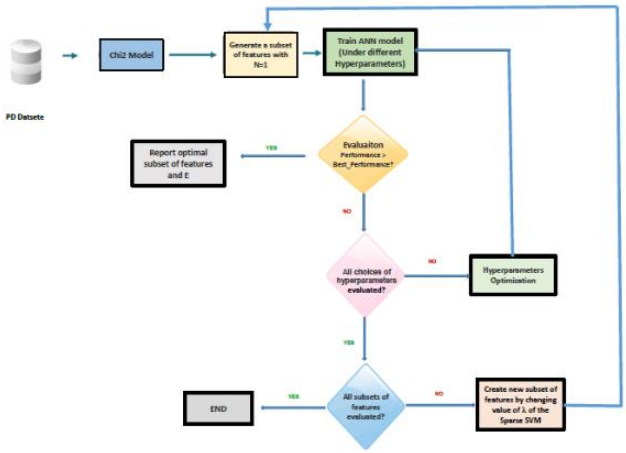

Figure 3: Block diagram of the proposed approach

4. Validation and Evaluation

For validating the proposed cascaded learning system (CLS), we use stratified k-fold validation scheme with values of k is fixed 5 and 4. Also, for evaluation purposes, we used specificity, sensitivity, accuracy, F-score or the F-measure and the Mathews correlation coefficient (MCC). The dataset used in this paper has imbalance classes which makes accuracy not a good measure for such data [14]. Thus, we also used ACCbal which stands for balanced accuracy.

5. Results and Discussion

Python programming language has been used to simulate the proposed methods. Two types of cross validation experiments were simulated, the first one used 4-fold cross validation and the second one used 5-fold cross validation. The details of these experiments are given in the below subsections.

5.1 4-fold cross validation

During the 4-fold cross validation, we split the data into four parts. During the first loop of cross validation, the first part of the splitted data is used for test purpose while the reaming three parts of the data are used for training

purposes. The loop is repeated for four times thus giving chance to all the four part to be tested while training the model on the remaining parts of the data during each loop. Also, we used Chi2 for feature selection and seven different ML classifiers for classification purposes including the proposed DNN. The other ML models include Decision Tree, Naïve Bayes, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and the K-nearest neighbours (KNN) having linear function and RBF function as kernel. The results of the 4-fold cross validation are given in the Table given below.

TABLE 1. Results of the 4-fold cross validation method

Method	**S**	**ACC_{bal}**	**Sen(%)**	**Spec(%)**	**F**	**MCC**
Navie Bayes	1	64	96	33	0.9	0.4
Decision tree	2	56	97	15	0.8	0.2
LDA	1	60	82	38	0.8	0.2
KNN	2	63	60	65	0.7	0.2
SVM(Lin)	5	74	76	72	0.8	0.4
SVM(RBF)	2	60	67	54	0.7	0.1
DNN	**2**	**62**	**61**	**62**	**0.7**	**0.1**

Table 1 shows that KNN model resulted very poor performance. On the other hand, the proposed DNN results with highest performance. This clearly shows the importance of the DNN model over the other ML classifiers.

5.2 5-fold cross validation

During this experiment, we split the data into five parts. During the first loop of cross validation, the first part of the splitted data is used for test purpose while the reaming three parts of the data are used for training purposes. The loop is iterated for five times. Finally, the average results obtained on all the five testing datasets is reported. Results of the five-fold cross validation is given in the Table given below.

TABLE 2. Results of 5-fold cross validation

Method	**S**	**ACC_{bal}**	**Sen(%)**	**Spec(%)**	**F**	**MCC**
Naïve Bayes	5	64	96	33	0.9	0.4
Decision Tree	5	56	97	15	0.8	0.2
LDA	8	60	82	38	0.8	0.2
KNN	3	63	60	65	0.7	0.2
SVM(Lin)	4	74	76	72	0.8	0.4
SVM(RBF)	4	60	67	54	0.7	0.1
DNN	**5**	**62**	**61**	**62**	**0.7**	**0.1**

6. Conclusion and future direction

Hand written sentences in Urdu language for diagnosing Parkinson's disease have been used in this research work. It is a common problem in health informatics that classes are imbalanced because of the lower samples of patient class compared to the healthy class. This kind of imbalance causes biased performance. Hence, random under-sampling method was used. Also, Chi2 model was used with DNN model to improve PD detection accuracy. Results validate the use of hand-written Urdu language sentences for the diagnosis of PD. In future, other hand written modalities like spiral and meander drawings and different types of sentences can be evaluated to find out better modality for PD detection. Moreover, for optimizing the DNN model, in future some robust, fast and automated algorithms like evolutionary [21, 22] algorithms should be adopted.

REFERENCES

[1] L. Ali, S. U. Khan, M. Arshad, S. Ali, and M. Anwar, "A multimodel framework for evaluating type of speech samples having complementary information about parkinson's disease," in *2019 International Conference on Electrical, Communication, and Computer Engineering(ICECCE)*. IEEE, 2019, pp. 1–5.

[2] L. Ali, Z. He, W. Cao, H. T. Rauf, Y. Imrana, and M. B. B. Heyat, "Mmdd-ensemble: A multimodal data–driven ensemble approach for parkinson's disease detection," *Frontiers in Neuroscience*, vol. 15, 2021.

[3] L. Ali, C. Zhu, M. Zhou, and Y. Liu, "Early diagnosis of parkinsons disease from multiple voice recordings by simultaneous sample and feature selection," *Expert Systems with Applications*, vol. 137, pp. 22– 28, 2019.

[4] L. Ali, C. Zhu, Z. Zhang, and Y. Liu, "Automated detection of parkinsons disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network," *IEEE journal of translational engineering in health and medicine*, vol. 7, pp. 1–10, 2019.

[5] W. Akbar, W.-p. Wu, S. Saleem, M. Farhan, M. A. Saleem, A. Javeed, and L. Ali, "Development of hepatitis disease detection system by exploiting sparsity in linear support vector machine to improve strength of adaboost ensemble model," *Mobile Information Systems*, vol. 2020, 2020.

[6] F. S. Ahmad, L. Ali, H. A. Khattak, T. Hameed, I. Wajahat, S. Kadry, S. A. C. Bukhari *et al.*, "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (ehrs)," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3283–3293, 2021.

[7] F. S. Ahmed, L. Ali, B. A. Joseph, A. Ikram, R. U. Mustafa, and S. A. C. Bukhari, "A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit," *Journal of Trauma and Acute Care Surgery*, vol. 89, no. 4, pp. 736–742, 2020.

[8] L. Ali and S. Bukhari, "An approach based on mutually informed neural networks to optimize the generalization

capabilities of decision support systems developed for heart failure prediction," *Irbm*, vol. 42, no. 5, pp. 345–352, 2021.

[9] A. Rehman, A. Khan, M. A. Ali, M. U. Khan, S. U. Khan, and L. Ali, "Performance analysis of pca, sparse pca, kernel pca and incremental pca algorithms for heart failure prediction," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–5.

[10] L. Ali, S. U. Khan, N. A. Golilarz, I. Yakubu, I. Qasim, A. Noor, and R. Nour, "A feature-driven decision support system for heart failure prediction based on statistical model and gaussian naive bayes," *Computational and Mathematical Methods in Medicine*, vol. 2019, 2019

[11] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54 007–54 014, 2019.

[12] L. Ali, I. Wajahat, N. A. Golilarz, F. Keshtkar, and S. A. C. Bukhari, "Lda–ga–svm: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," *Neural Computing and Applications*, vol. 33, no. 7, pp. 2783–2792, 2021.

[13] T. Meraj, H. T. Rauf, S. Zahoor, A. Hassan, M. I. Lali, L. Ali, S. A. C. Bukhari, and U. Shoaib, "Lung nodules detection using semantic segmentation and classification with optimal features," *Neural Computing and Applications*, vol. 33, no. 17, pp. 10 737–10 750, 2021

[14] L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou, and Y. Liu, "Reliable parkinsons disease detection by analyzing handwritten drawings: construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model," *Ieee Access*, vol. 7, pp. 116 480–116 489, 2019.

[15] S. Rosenblum, M. Samuel, S. Zlotnik, I. Erikh, and I. Schlesinger, "Handwriting as an objective tool for parkinsons disease diagnosis," Journal ´ of neurology, vol. 260, no. 9, pp. 2357–2361, 2013.

[16] C. R. Pereira, D. R. Pereira, F. A. da Silva, C. Hook, S. A. Weber, L. A. Pereira, and J. P. Papa, "A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements," in 2015 IEEE 28th international symposium on computer-based medical systems. IEEE, 2015, pp. 171–176.

[17] Y. Imrana, Y. Xiang, L. Ali, and Z. Abdul-Rauf, "A bidirectional lstm deep learning approach for intrusion detection," *Expert Systems with Applications*, vol. 185, p. 115524, 2021.

[18] Y. Zhang, "Can a smartphone diagnose parkinson disease? a deep neural network method and telediagnosis system implementation," *Parkinsons Disease*, vol. 2017, 2017

[19] M. Vadovsky and J. Parali ` c, "Data collection methods for the diagnosis ˇ of parkinsons disease," *International Journal on Biomedicine and Healthcare*, vol. 5, no. 1, pp. 28–32, 2017.

[20] M. Su and K.-S. Chuang, "Dynamic feature selection for detecting parkinson's disease through voice signal," in *2015 IEEE MTT-S 2015 International Microwave Workshop Series on RF and Wireless Technologies for Biomedical and Healthcare Applications (IMWS-BIO)*. IEEE, 2015, pp. 148–149.

[21] N. A. Golilarz, H. Gao, R. Kumar, L. Ali, Y. Fu, and C. Li, "Adaptive wavelet based mri brain image de-noising," *Frontiers in neuroscience*, vol. 14, 2020.

[22] S. U. Khan, M. Rahim, and L. Ali, "Correction of array failure using grey wolf optimizer hybridized with an interior point algorithm,"*Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 9, pp. 1191–1202, 2018.