Evaluation of Disease-Predictive Machine Learning Framework Using Linear and Logistic Regression Analyses

A.M. Adeshina, Siti Fatimah Abdul Razak, Sumendra Yogarayan, Md Shohel Sayeed
Faculty of Information Science and Technology, Multimedia University, Malaysia
E-mail: am.adeshina@mmu.edu.my, codedengineer@yahoo.com, fatimah.razak@mmu.edu.my, sumendra@mmu.edu.my, shohel.sayeed@mmu.edu.my

Keywords: cancer, disease prediction, linear regression, logistic regression, Parkinson's disease

Received: December 24, 2023

Predicting diseases with revolutionary tools has recently been so significant in the domain of computer-assisted software for disease diagnosis, treatment and therapy management. Some of these life-threatening diseases could be challenged with early prediction and diagnoses thereby saving lives as well as healthcare cost. Unfortunately, among these diseases are those with more extreme challenges when it comes to diagnosing, as there are no specific tests to confirm the disease and many symptoms overlap with those of other neurodegenerative disorders such Parkinson's diseases, essential tremor, a condition that causes involuntary and rhythmic shaking seen similar to Parkinson's, such could also be traced to most forms of cancer diseases. Apparently, harnessing Artificial Intelligence to predicting diseases has been the recent focus in disease diagnosis and therapy management. Considering the fact that medics confirmed Artificial Intelligence to be a non-invasive, predictive, preventive and personalized medical technique, it has therefore been cited as an alternate approach to helping physicians making accurate diagnosis, more quickly. This study proposed a machine learning framework for predicting diseases. The study was evaluated using linear and logistic regression analyses. The framework was designed and implemented to function in multimodal capacities hence datasets of one of the most prevalent cancers, breast cancer, seen as the leading causes of mortality among women, were also used in the evaluation of the framework. Interestingly, logistic regression analysis recorded the best accuracy in all the experimentations conducted with seed 6 and approximately 92.021% of accuracy. Linear Regression algorithm was rated poor with low prediction of 67.293% with the seed of 7 whereas logistic regression recorded approximately 88.298% with the same number of seed. While more efforts would be required to ensuring a firm prediction of Parkinson's Diseases through the proposed framework, with the few datasets of Parkinson's diseases used in the study, the framework was able to detect the presence of Parkinson through status indication of '1' or its absence with the status '0' using logistic regression analysis. Significantly, the evaluation proves the framework resourceful, not only in multimodal capacities but also to an appreciable extent in predicting diseases.

Povzetek: Strojno učenje je uporabljeno za napovedovanje bolezni raka in Parkinsove bolezni, pri čemer je linearna regresija dosegla 67,3 % točnosti, logistična regresija pa 92,0 %.

1 Introduction

Predicting diseases to save lives with revolutionary tools has been the focus of researchers. Apparently, early prediction and diagnoses of diseases remain the most beneficial in terms of healthcare costs and possible patient outcomes. For instance, cancer disease has been a major issue in medical sciences challenging the pathologists in the disease, diagnosis and therapy management. Apparently, there are quite a number of possible cancer therapies for patients diagnosed of cancer such as chemo, radio, and hormone therapy. Unfortunately, some of these remedies are usually associated with one challenge or another, such high load of reversible and irreversible adverse effects, leading to limited therapeutic efficacy, and low chances of quality survival. The mentioned methods also may have some side effects and false positives, in the similar circumstances we have with tumors [1]. Similar situations are applicable to other life-threatening diseases. Furthermore, with the current rate at which volume of data generated extremely fast in the field of disease diagnosis and therapy management, machine learning techniques have been found to offer quite promising results. With the use of the machine learning techniques, required information could be extracted quickly on the basis of past experiences while also facilitating accurate detection of the patterns within a short period. This study proposes a machine learning framework for predicting diseases and evaluates the accuracy, sensitivity, specificity, and the precision of the framework with cancer datasets using Regression Analysis Linear and Logistic Regression Analysis.

2 Related works

A number of efforts have been contributed to the development of resourceful computer-aided framework for cancers and Parkinson's diseases. Cancer of the blood, Leukemia, the situations whereby cancer starts in blood-forming tissue, such as the bone marrow, causes large numbers of abnormal blood cells to be produced and enter the bloodstream. Machine learning algorithms have been widely utilized in the treatment of leukemia. Techniques like image processing and pattern recognition can be utilized to help in many capacities [2]. Leukemia can be automatically detected using computer-aided diagnostic (CAD) models, which can help doctors greatly in their procedures and thus becoming useful for leukemia early identification [3]. Towards the development of application of machine learning framework for prostate cancer, some studies focused on feature selection using Random Forest Classifier to detect prostate cancer [4].

Shetty and Rao in 2016 [5] concentrated on gait signals in Parkinson's disease diagnosis and other neurological disorders using Support Vector Machine (SVM). Similarly, the study of Prashanth and the team [6] utilized non motor signals in the early diagnosis of Parkinson's disease by deploying Naïve Bayes, Support Vector Machine (SVM), Boosted Trees, and Random Forest. The authors documented Support Vector Machine as an approach with the highest accuracy value of 96.40%. Submission of Bloem in 2021 [7] added values to the efforts in Parkinson's disease prediction studies. The authors implemented wavelet analysis Support Vector Machine as paired approach for efficient classification accuracy of 90.32%. Abiyev and Abizade [8] presented Parkinson's Disease diagnosis methodology using Fuzzy Neural System (FNS) and Neural Network (NN). The findings indicated that FNS is better off compared to NN. Singh et al. [9] presented Parkinson's Disease detection using Support Vector Machine with overall accuracy of 100%.

Çimen and Bolat [10] documented Generalized Regression Neural Network (GRINN) as the best performed approach among Artificial Neural Network (ANN), Multilayer Perceptron (MLP), and Generalized Regression Neural Networks (GRNN) in the Parkinson's Disease diagnosis studies. Similarly, Gao et. al. [11] implemented several model-based and model-free Machine Learning techniques including the Random Forests (RF), the Logistic Regression (LR), the XGboost and the Support Vector Machines (SVM). The study recorded a machine learning model with 80% accuracy. Huang et. al. [12] investigated the deep Convolutional Neural Network (CNN) framework with pre-trained model which was found to be able to successfully differentiate between patients with Parkinson's disease and normal subject. The general results presented by Rehman et. al. [13] on evaluation of Support Vector Machine and Random Forests (RF) for Parkinson's disease predictions showed that Support Vector Machine performed better than Random Forest. The study trained Random Forest and Support Vector Machine models with normalized data and were being evaluated using cross validation with area under the curve. Table 1 summarizes related contributions to this study.

Table 1: Summary of literature

Authors	Approach	Contributions
Shetty and Rao (2016) [5]	Researchers concentrated on gait signals in Parkinson's disease diagnosis and other neurological disorders using Support Vector Machine (SVM)	Outstanding results were obtained but researchers anticipate more improvements on the approach.
Prashanth and the team [6]	The study utilized non motor signals in the early diagnosis of Parkinson's disease by deploying Naïve Bayes, Support Vector Machine (SVM), Boosted Trees, and Random Forest.	The authors documented Support Vector Machine as an approach with the highest accuracy value of 96.40%.
Bloem in 2021 [7]	The authors implemented wavelet analysis Support Vector Machine as paired approach for efficient classification accuracy of 90.32%.	The study added values to the efforts in Parkinson's disease prediction studies.
Abiyev and Abizade [8]	The authors proposed Parkinson's Disease diagnosis methodology using Fuzzy Neural System (FNS) and Neural Network (NN).	The findings indicated that FNS is better off compared to NN.
Singh et al. [9]	Authors presented Parkinson's Disease detection using Support Vector Machine.	The study reported overall accuracy of 100% with Support Vector Machine.
Çimen and Bolat [10]	Generalized Regression Neural Network (GRINN) was proposed for Parkinson's Disease diagnosis.	The study documented Generalized Regression Neural Network (GRINN) as the best performed approach among Artificial Neural Network (ANN), Multilayer Perceptron (MLP), and Generalized Regression Neural Networks (GRNN) in the Parkinson's Disease diagnosis studies.
Gao et. al. [11]	Implemented several model-based and model-free Machine Learning techniques including the Random Forests (RF), the Logistic Regression (LR), the XGboost and the Support Vector Machines (SVM).	The study recorded a machine learning model with 80% accuracy.
Huang et. al. [12]	The study investigated the deep Convolutional Neural Network (CNN) framework.	Pre-trained model which was found to be able to successfully differentiate between patients with Parkinson's disease and normal subject.
Rehman et. al.	Evaluation of Support	The study trained

[13]	Vector Machine and Random Forests (RF) for Parkinson's disease predictions showed that Support Vector Machine performed better than Random Forest.	Random Forest and Support Vector Machine models with normalized data and were being evaluated using cross validation with area under the curve.

3 Methodology

Disease-Predictive Machine Learning framework was proposed in this study with the evaluation of the framework using Linear and Logistic Regression Analyses. The proposed framework is presented in Figure 1.

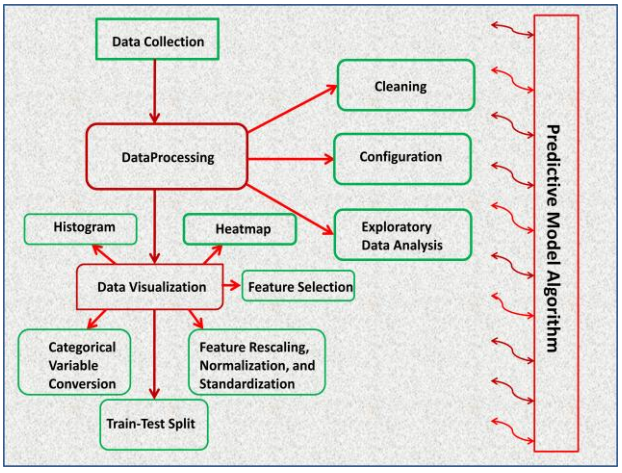

Figure 1: The proposed framework

3.1. Data collection

Datasets used in the evaluation of this framework include Parkinson's disease and Breast cancer datasets. The Parkinson's datasets were collected from UCI Machine Learning Repository consisting of *188* patients with PD *(107 men and 81 women)* with ages ranging from *33* to *87(65.1Â±10.9)* at the Department of Neurology in CerrahpaÅŸa Faculty of Medicine, Istanbul University [13]. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset consisting of *569* samples, having *32* attributes which include two class attribute labels (diagnosis: B= benign, M= malignant), ID number, and *32* numeric features were used.

3.2. Data processing

Python Data Analysis Library, Pandas, was used for the processing. Pandas is a data manipulation and analysis software built on top of the Python programming language. The datasets were preprocessed using Python Data Analysis Library, Pandas, making use of the its strength in providing extended data structures to hold different types of labelled and relational data. Moreover, Panda's libraries tool is considered useful in data cleaning and manipulation. The data preprocessing stage handles operations like merging, reshaping, joining, and concatenating of the datasets.

3.2.1. Data cleaning

Fixing or removing outliers, missing values, or duplicated values within the datasets were achieved at this phase of the framework. With the combination of many multiple data sources, there is always that possibility for data duplication or mislabelling which could be resolved seamlessly at the data cleaning phase of the framework.

3.2.2. Data configuration

Raw data are further prepared at this phase to ensuring further processing and analysis. Significant steps at this stage include collecting, cleaning, and labelling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.

3.2.3. Exploratory data analysis

With the EDA phase, in-depth understandings of the datasets were further achieved. Data manipulation techniques and necessary statistical tools were introduced in describing and understanding the relationship between variables and the effect of such on the results. Furthermore, Panda's libraries tool was invoked in ensuring thorough EDA output.

3.3. Data visualization (Analyzing the data)

The outputs obtained from EDA phase were analyzed carefully at the data visualization phase of the framework. Histogram, Heatmap, Categorical Variable Conversion, Feature Selection, Feature Rescaling, Normalization, and Standardization were achieved in the data visualization phase of the framework.

Whenever the histogram bars were seen higher, it shows that more data falls in that range. However, with heatmap, we were able to observe possible intersections of the values considered of higher data concentration than others. Heatmap greatly assists in seeing which intersections of the values have higher data concentration than others. The categorical variable conversion stage assists in the categorization of the variable numeric and the categorical features into possibly diagnosed and in the case of cancer, malignant or benign. Relevant features were selected out of the entire features in the datasets through the feature selection phase of the framework following the implementation of chi-square statistical analysis conditions for the study to specifically check the independence of two events. The algorithm may however have to observe count *X* for instance, and then predict count *Y* with the situation of data of two variables. In other to record optimum performances with the experiment's, the varied values of features in

the datasets undergo feature recalling form the range *0* to *1*, called, Normalization.

3.4. Train-test split

The data is normally split into two subsets: the training datasets and the testing datasets. The training dataset is the actual dataset that is used to train the model. 65% of the dataset was put into training set while 35% of the remaining dataset was used for the testing. The model sees and learns from this data. The predictive model algorithm proposed for this study is presented in Figure 2.

1.0. Start
1.1. Importation of the Dataset
1.2. Manipulating the dataset
1.3. Model the data into training set
1.4. Fit in the linear regression
1.5. Populate Results
1.6. Fit in the logistic regression
1.7. Populate Results
1.6. End

Figure 2: Proposed predictive model algorithm

4 Implementation

The implementation and the experiment were conducted in Python using the Jupyter Programming Language for developing and testing different ML models. The concept was based on statistical technique, which generally follows two distinct steps in data analysis. The first step involves descriptive statistics, which summarizes data using indexes such as mean and median. While the second step is inferential statistics, which concludes data using statistical tests such as the WBCD t-test. The validation dataset is used to evaluate the statistical accuracy of the model built. Necessary libraries required in the transformation of the data were imported such as import pandas as pd which tells the library pd as pandas, and import numpy as np which also means indicate np as numpy.

5 Results and discussion

Logistic Regression Analysis was observed better off as predictive algorithm with cancer datasets, recording better accuracy compared to Linear Regression Analysis. Similarly, with the few datasets of Parkinson's diseases used in this study, the framework was able to detect the presence of Parkinson's through status indication of *'1'* or its absence with the status *'0'* using logistic regression analysis. However, more efforts would be required to adapting the framework to be resourceful enough for firm prediction of Parkinson's Diseases. Some of the results are presented using Histogram charts, Density Plot which was actually achieved before and after Transformation processes, and also Multivariate Plot.

5.1. Histogram chart

The chart presented in Figure 3 shows that most of the features in the data were skewed to either left or right, while some were near normal and only the smoothness mean is perfectly distributed, in this case rescaling the data is required. The data distribution and skewness is presented in Figure 3.

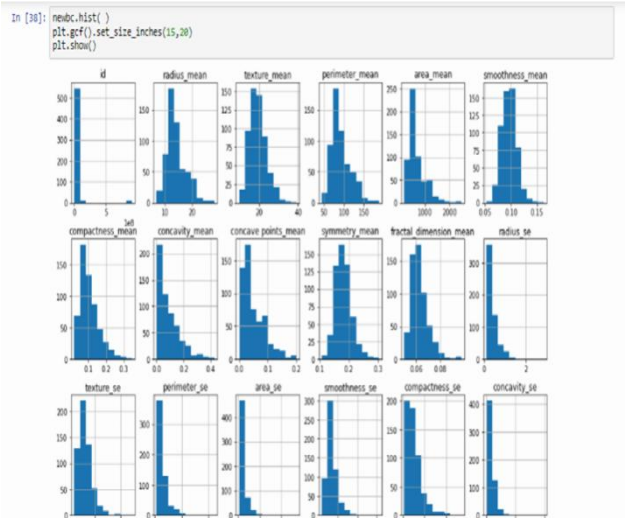

Figure 3: Histogram

5.2. Density plot

The plot was achieved before and after transformation processes. The function *plot ()* shows the situation if the data is normally distributed or skewed to the right or left-hand side. The density plot in Figure 4 shows that most of the features in the data were skewed to either left or right while some were near normal, just only the smoothness mean is perfectly distributed, in this case, rescaling the data would be required.

```
In [11]: # Univariate Density Plots
         newbc.plot(kind='density', subplots=True, layout=(10,5), sharex=False)
         plt.gcf().set_size_inches(40,50)
         pyplot.show()
```

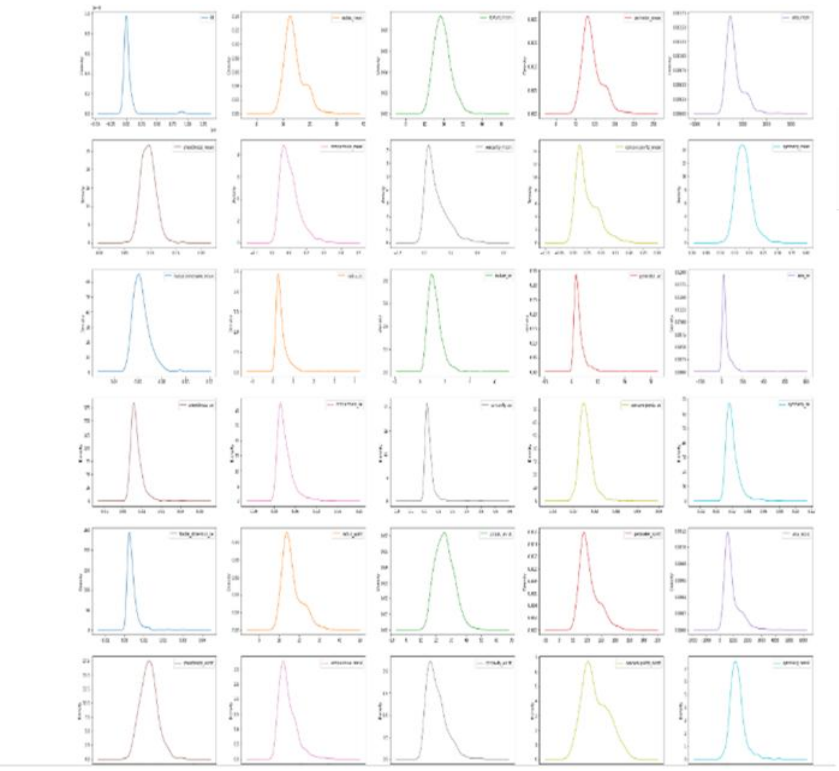

Figure 4: Density plot (before transformation process)

```
In [76]: # Univariate Density Plots after transformation
         newormbc.plot(kind='density', subplots=True, layout=(10,5), sharex=False)
         plt.gcf().set_size_inches(30,40)
         pyplot.show()
```

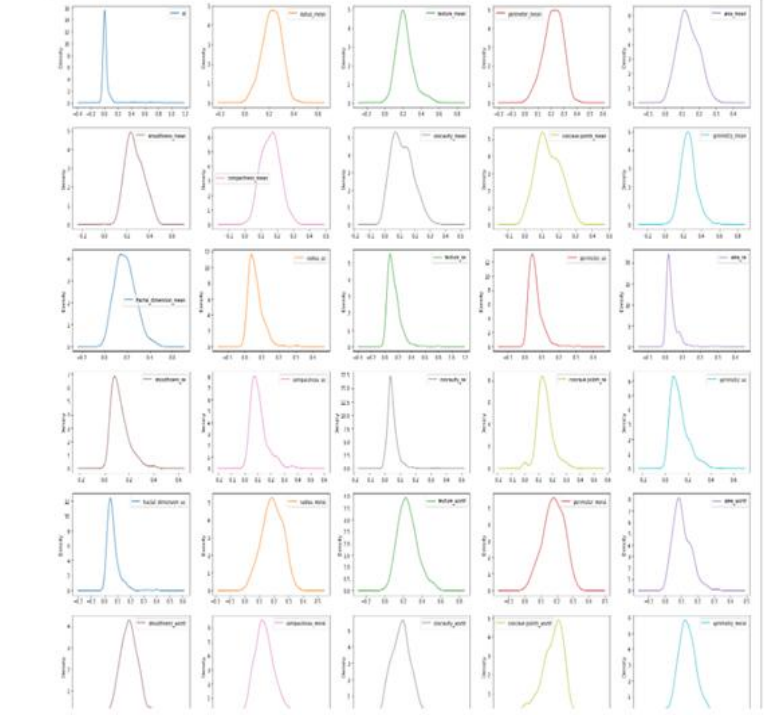

Figure 5: Density plot (after transformation process)

Figure 5 shows the distribution and skewness in the data. According to Figure 3 and 4, requiring rescaling to improve the features in the data, the density plot in Figure 5 shows an improved feature in the data where more than 15 features in the data were perfectly distributed, and some were near normal, while other features were little skewed to left or right.

5.3. Multivariate plot

The heatmap function *sb. heatmap ()* was invoked in expressing the various shades of the same hue for each value to be plotted, this illustrates the correlations between all the features in this data. With a range of *-0.6* to *1.0*, the lighter shades of the chart often denote greater values, which were preferable to the darker shades.

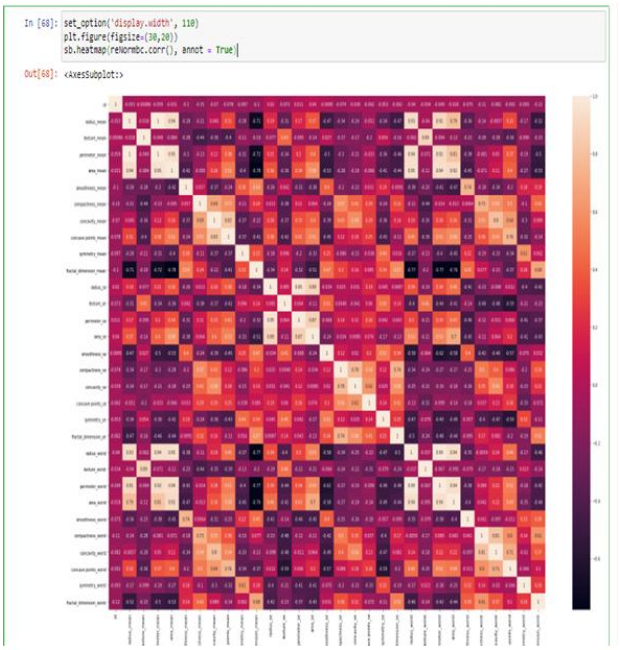

Figure 6: Heatmap

Darker values begin at *-0.6* and increase to a lighter value of *1.0* as presented in the heatmap in Figure 6. The left side is in Figure 7 and the right is in Figure 8 respectively.

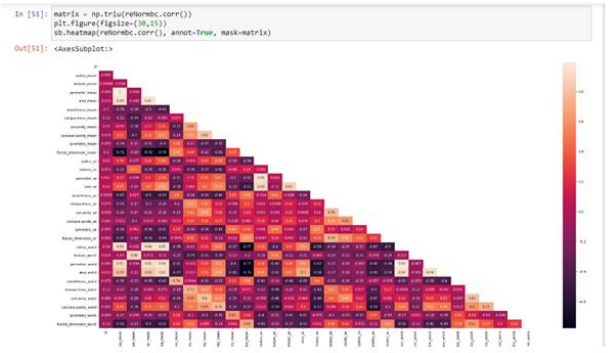

Figure 7: Heatmap (left side)

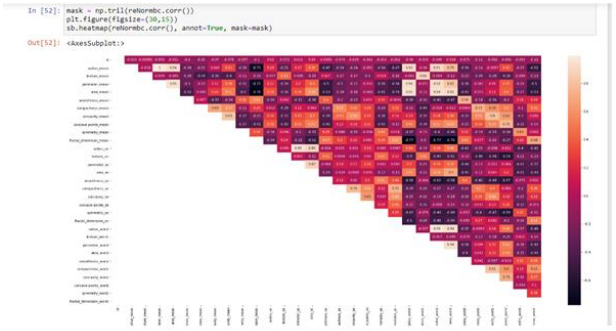

Figure 8: Heatmap (right side)

Similarly, trying to look at a plot that shows the interactions between multiple variables in a dataset, we used scatter plots. Scatter plots are useful for spotting structured relationships between variables. Looking into whether to summarize the relationship between two variables with a line right as presented in Figure 9.

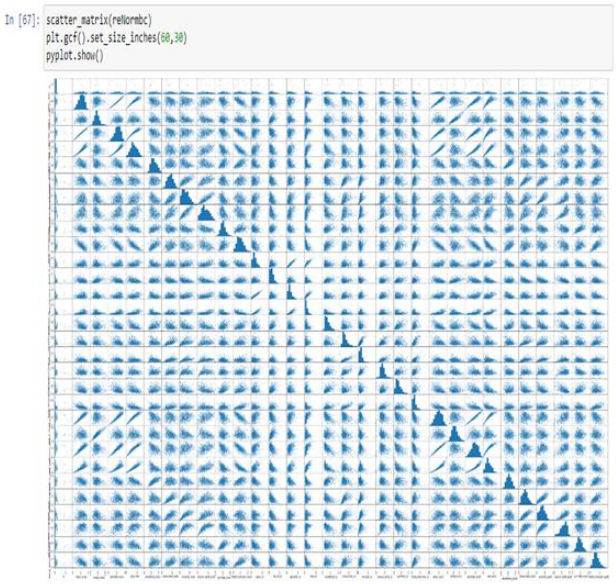

Figure 9: Scatter plot

Evaluations showing the accuracy of results of linear regression analysis, using seed 7 in comparison with the logistic regression analysis at the same seed number are presented in Figure 10 and Figure 11 respectively.

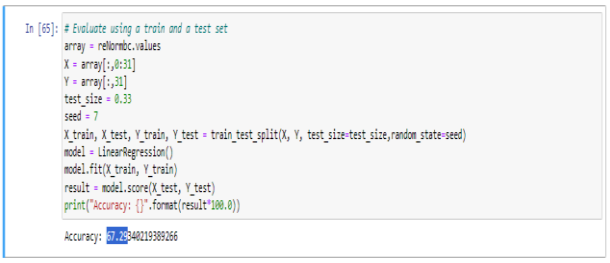

Figure 10: Linear regression test result

Figure 11: Logistic regression first test result

Further evaluation of the predictive model algorithm for logistic regression analysis with seed 5 and seed 6 are presented in Figure 12 and Figure 13 respectively.

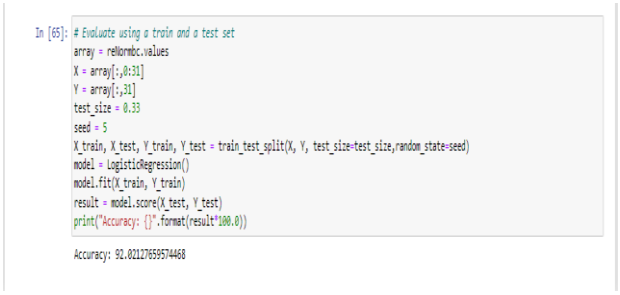

Figure 12: Logistic regression third test result

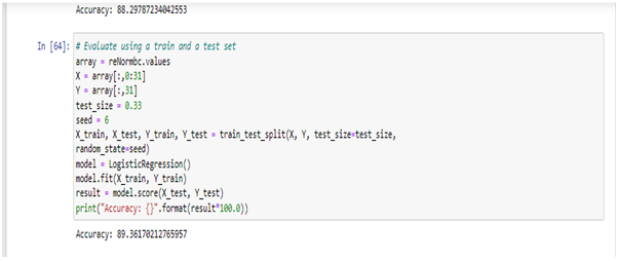

Figure 13: Logistic regression second test result

Table 2 and Table 3 show the evaluations of logistic and linear algorithms using Precision, Recall, F1-Score and Accuracy metrics.

Table 2: Seed 6 Evaluation

Seed 6					
Regression Analysis		Precision	Recall	F1-Score	Accuracy
Logistic	1	0.93	0.96	0.92	
Linear		0.74	0.73	0.74	0.72

Table 3: Seed 7 Evaluation

Seed 7				
Regression Analysis	Precision	Recall	F1-Score	Accuracy
Logistic	0.98	0.88	0.93	0.88
Linear	0.89	0.68	0.77	0.67

6 Discussion

Ability to successfully predict a disease using revolutionary computational tools is one of the greatest hopes in disease detection, diagnosis and therapy management. Diseases of the progressive nervous system disorder that affects movements, such as Parkinson's, currently the fastest growing neurodegenerative disease still remain challenges. Apparently, submissions till date, still confirm Parkinson's disease challenging when it comes to diagnosing, as there are no specific tests to confirm the disease and many symptoms overlap with those of other neurodegenerative disorders such as essential tremor, a condition that causes involuntary and rhythmic shaking seen similar to Parkinson's. Unlike other previously proposed approaches, the framework proposed in this study was intended to use auto-selectivity, auto-productivity, auto-adaptivity and self-learning qualities of artificial intelligence in the development of its algorithms as alternate resourceful technique for early detection, characterization, prediction and tracking of Parkinson's disease, and also possibly other neurodegenerative diseases.

7 Conclusion and future work

With this study, we have been able to design and develop a framework to function in multimodal capacities using Logistic Regression Analysis and Linear Regression Analysis in the evaluation. Under the same experimental conditions, Logistic Regression algorithm recoded *88.298%* with seven seeds in the first prediction, recorded *89.362%* when the number of seed is reduced to six for the second prediction, while *92.021%* with the prediction at five seeds. Linear Regression algorithm was rated poor with low prediction of *67.293%* with the seed of *7* whereas logistic regression predicts *88.298%* with the same number of seed.

Integrating the framework with suitable machine learning algorithms is quite resourceful in the prediction of diseases. Experimental evaluations of the framework found it appropriate for multiple algorithms of different fundamental concepts, though using it with algorithms of these same concepts is still possible even with only a particular algorithm, thereby proving the framework's scalability in usage. Furthermore, though considering the sensitivity and the robustness of the framework, it is found resourceful in Parkinson's disease analysis but integration of certain extensive features into the framework would make it completely adaptive to Parkinson's disease detection, characterization and predictions, and with possible extension for other neurodegenerative diseases. Hence, future work will focus more on ensuring firm use of the framework for intended multimodal disease predictions and analyses.

Acknowledgement

This study is supported by Multimedia University, Malaysia through MMUI/230025 grant. Authors would like to thank all the anonymous reviewers for their constructive comments.

References

[1] Adeshina, A.M., Hashim, R., Khalid, N.E.A., Abidin. Locating abnormalities in brain blood vessels using parallel computing architecture. Interdiscip Sci Comput Life Sci **4**, 161–172 (2012).

[2] Kumar, D., Jain, N., Khurana, A., Mittal, S., Satapathy, S. C., Senkerik, R., & Hemanth, J. D. (2020). Automatic Detection of White Blood Cancer from Bone Marrow Microscopic Images Using Convolutional Neural Networks. IEEE Access, 8, 142521–142531.

[3] Sridhar, K., Yeruva, A. R., Renjith, P. N., Dixit, A., Jamshed, A., & Rastogi, R. (2022). Enhanced Machine Learning Algorithms Lightweight Ensemble Classification of Normal versus Leukemic Cells. Journal of Pharmaceutical Negative Results, 13, 496–505.

[4] Huljanah, M., Rustam, Z., Utama, S., & Siswantining, T. (2019). Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. IOP Conference Series: Materials Science and Engineering, 546(5).

[5] Shetty, S., and Y. Rao (2016). SVM-Based Machine Learning Approach to Identify Parkinson's Disease using Gait Analysis. In Proceedings of the Inventive Computation Technologies (ICICT). IEEE. Coimbatore, India, August. Vol. 2. pp. 1–5.

[6] Prashanth, R., Roy, S. D., Mandal, P. K., and Ghosh, S. (2016). Highaccuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning. International Journal of Medical Informatics. Vol. 90.

[7] Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. The Lancet, 397(10291), 2284-2303.

[8] Abiyev, R. H. and Abizade, S. (2016). Diagnosing Parkinson's Diseases using Fuzzy Neural System," Computational and Mathematical Methods in Medicine. Vol. 2016, Article ID 1267919.

[9] Singh, G., Vadera, M., Samavedham, L., and Lim, E. C.-H. (2016). Machine Learning-Based Framework for Multi-Class Diagnosis of Neurodegenerative Diseases: A Study on Parkinson's Disease. Vol. 49, Issue 7, Pages 990-995.

[10] Çimen, S. and Bolat, B. (2016). Diagnosis of Parkinson's disease by using ANN. In Proceedings of the Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). International Conference on, pp. 119–121, IEEE, Jalgaon, India.

[11] Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N.I., Müller, M.L.T.M., Herman, T., Giladi, N., Kalinin, A., Spino, C., Dauer, W., Hausdorff, J.M., Dinov, I.D. (2018). Model-based and Model-Free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. Sci Rep 8, 7129.

[12] Huang, G.H., Lin, C.H., Cai, Y.R., Chen, T.B., Hsu, S.Y., Lu, N.H., Chen, H.Y. and Wu, Y.C. (2020). Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13(5), pp.508-523.

[13] Rehman RZU, Del Din S, Shi JQ, Galna B, Lord S, Yarnall AJ, Guan Y, Rochester L. (2019). Comparison of Walking Protocols and Gait Assessment Systems for Machine Learning-Based Classification of Parkinson's Disease. Sensors (Basel). Dec 5;19(24):5363. doi: 10.3390/s19245363. PMID: 31817393; PMCID: PMC6960714.

[14] Sakar, C.O., Serbes, G., Gunduz, A., Tunc, H.C., Nizam, H., Sakar, B.E., Tutuncu, M., Aydin, T., Isenkul, M.E. and Apaydin, H. (2018). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Applied Soft Computing.