# A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches

**F.M. Javed Mehedi Shamrat, Md. Asaduzzaman, A.K.M. Sazzadur Rahman, Raja Tariqul Hasan Tusher, Zarrin Tasnim**

**Abstract**: Objective: The primary objective of the study is to inspect the exhibition of three supervised algorithms for improving Parkinson disease analysis by detection. **Methods:** I utilized three AI methods for the detection of Parkinson disease datasets. SVM, KNN, and LR were utilized for the forecast of Parkinson Disease. The exhibition of the classifiers was assessed via recall, precision, f 1 extent, and precision. **Results:** SVM shows the accuracy level of 100% for Parkinson disease prediction. LR achieved the second-highest classification accuracy of 97%. Also, as far as precision for dissecting Parkinson illness datasets, KNN acquired the worst performance (i.e. 60%). **Conclusion:** My findings showed that the SVM obtained the highest performance for analyzing the Parkinson datasets. This perusal has emphasized the current of Parkinson research aptitude and scope in connection to clinical research fields by machine learning techniques. That will be a viable effect in the field of Parkinson disease.

**Keywords**: Machine Learning, Classification, Parkinson, prediction.

## 1. INTRODUCTION

Parkinson diseases are the most critical causes of death and disability worldwide. According to the Parkinson disease foundation, The affected peoples worldwide of Parkinson disease is projected that the 1 million people are Living by 2020 in the USA [1]. The medical treatment of Parkinson disease can be endorsed on Neuropathology and Histopathology [2] [3]. Medical diagnostic detection of Parkinson Disease should be possible across the board choice basing on the affectability and particularity of the trademark Parkinson sickness highlights. In this manner, Parkinson Disease is expected to investigate the clinical, pathologic, and nosology studies grounded on the recurrence of event, attributes, and including danger components of tests [4][2]. Parkinson usually affects a large part of worldwide patients over the age of 50, which has affected up to now [5]. Still now there is no known cause of Parkinson disease, however, it is very likely possible to assuage symptoms knowingly in the early stage of the subjective patients [6]. Approximate 90% of the patients affected with voiced damage a study appealed this [7]. The Parkinson treatment is likely very costly. This causes most of the patients cannot afford the cost of the Parkinson disease. Nowadays, Parkinson disease prediction is most critical matter for clinical practitioners to take accurate decision of such disease. It's a great exercise at present time, machine learning based extensive platform can detect Parkinson disease. Medical data has grown a vast scale of volume from different clinical areas including health care services. To handle this data and attaining insights from this data there is a need for Big Data analysis through Machine learning that aims to solve a diverse medicinal and clinical issue [8] [9]. Officially, a significant number of investigations demonstrate that machine learning methods have picked up genuinely superior in classification based medical issues. Be that as it may, supervised learning-based strategies are one of the best techniques for the exploration network of research and real-life applications in clinical fields [10] [11]. This works main objective is to improve the detection and diagnosis techniques of Parkinson disease treatment. Therefore, my study can be playing an important role in detecting Parkinson disease with machine learning algorithms. In recent, machine learning algorithms have generated a significant influence and commitment in the Parkinson research community for the detection of Parkinson disease. Moreover, machine learning techniques are specified more precise results in disease prediction as compared to other data taxonomy techniques [10][12][13]. Motivated by this, the authors have used three prominent machine learning methods for recognition and appropriate finding of Parkinson patients. The primary objective of this study is to look at the exhibition estimation of different conspicuous classification methods for this study I used three supervised learning techniques were used including KNN, Support Vector Machine and Logistics Regression. Moreover, the performance of the three classifiers was evaluated using different methods.The rest of the paper discusses the literature review in section 2, the Methodology (Experimental Setup, Data Collection, Data Preprocessing, Evaluation Criteria) consist of section 3, in section 4 discussed Result & Discussion, in section 5 short brief on Conclusion.


- *F. M. Javed Mehedi Shamrat is currently pursuing Bachelor's degree program in Software Engineering at Daffodil International University, Bangladesh. E-mail: javedmehedicom@gmail.com*
- *Md. Asaduzzaman is currently pursuing Bachelor's degree program in Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: asaduzzaman15-7279@diu.edu.bd*
- *A.K.M Sazzadur Rahman is currently pursuing Master's degree program in Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: sazzad433@diu.edu.bd*
- *Raja Tariqul Hasan Tusher is currently pursuing Bachelor's degree program in Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: tusher.cse@diu.edu.bd*
- *Zarrin Tasnim is currently pursuing Bachelor's degree program in Software Engineering at Daffodil International University, Bangladesh. E-mail: zarrint25@gmail.com*


## 2 LITERATURE REVIEW

Through related work, 17 studies were done on applying and using different machine learning approaches to determine the detection of Parkinson Disease. Previous work also introduces a set of studies-based detection of Parkinson diseases using machine learning algorithms. However, the outcomes of the 17 articles on machine learning used in disease prediction as follows: Tarigoppula et al. (Sriram, Rao, Narayana, Kaladhar, & Vital, 2013) presented a comparative study between Naïve Bayes, Random Forest, Logistics Regression, Support Vector Machine to detect Parkinson disease. SVM (i.e. 88.9%) has shown good performance to compared NB (i.e. 69.23%), and RF (90.26%) shown the compared to SVM for the Parkinson detection. Moreover, LR (i.e. 83.66%) showed quite good performance. 86%). And the SVM and LDA have superior

sensitivity in comparison to other classifiers. The contribution of this study is to the analysis of voice data to understand the presence of Parkinson disease. So as to extra improve the demonstrative precision for the identification of Parkinson Disease, the study (Chen et al., 2013) proposed a fuzzy-based KNN model to predict Parkinson. Their study was shown to the best accuracy (96.07%) obtained by the proposed algorithm including 10-fold cross-validation. Another study (Chen et al., 2016) also considers a hybrid model of detection Parkinson with compared to the existing methods and their proposed model has achieved brilliant accuracy through 10-fold cross-approval investigation, the highest precision of 96.47% and quite good accuracy of 95.97%. Moreover, The experimental (Hariharan, Polat, & Sindhu, 2014) results show that the maximum classification accuracy of 100% for the Parkinson's dataset via feature pre-processing. Hanzel et al.(Hazan, Hilu, Manevitz, Ramig, & Sapir, 2012) presented a new prediction system that can detect of Parkinson from voice data seems to be possible and precise with results approaching (90%) in two different data sets. Another hybrid method (Ma, Ouyang, Chen, & Zhao, 2014) named SCFW-KELM has been presented for the analysis of Parkinson disease. The result of the proposed method is effective for Parkinson detection by MAE for the Total-UPDRS and Motor-UPDRS were accomplished correspondingly MAE = 0.4656 and MAE = 0.4967 (Nilashi, Ibrahim, Ahmadi, Shahmoradi, & Farahmand, 2018). Moreover, A study (Ozcift, 2012) uses kernel Support Vector Machine for their classification and Neural Network classification scheme. Thus, the prediction performances of the 2 classifiers respectively are 91.4% and 92.9%. Hence, one study (Geetha, Professor, Head, & Sivagami, 2011) found they showed into their study that the Random Forest obtained the highest performance. But, another study showed SVM reaches an upright precision of 83.33% (Shetty and Rao, 2017). Ferdous et al. (Wahid, Begg, Hass, Halgamuge, and Ackland, 2015) exhibited a relative report between various classifiers. Their analysis showed that the RF attained the accuracy of 92.6% after standardizing gait data using the multiple regression method, competed to 80.4% (Support Vector Machine) and 86.2% (Kernel Fisher Discriminant). Hence, the study (Yadav, Kumar, & Sahoo, 2012) compared to different classifiers and showed the results LR obtained the highest performance than others.

## 3 METHODOLOGY

### 3.1. Environment Setup

In this study, this section represents the experimental process (figure 1) of the experiment including machine learning techniques. Parkinson Disease data sets have been considered in this work. Firstly, we focused on preparing and combined data from the main datasets. Moreover, we extracted 30 features from the Parkinson datasets. Then, we checked the missing values and co-related values. Secondly, Data set parting (splitting) is a significant errand of these machine learning-based fields. Figure 3.1 shows the Parkinson data set was split into train sets and test sets. After that, 3 supervised based classifiers performed the operation. After successfully executed these algorithms SVM obtained the highest performance.

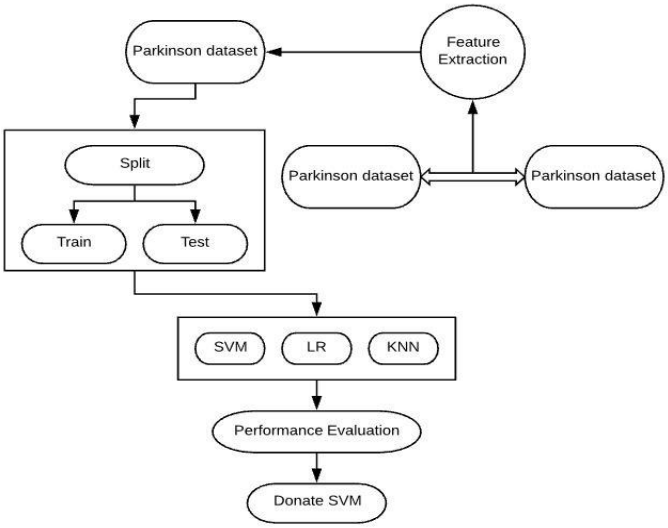


Fig. 1. The Experimental Setup.

- SVM

SVM includes a supervised learning technique that takes a gander at information and sorts it into one of two classes. An SVM yields a guide of the arranged information with the edges between the two as far separated as could reasonably be expected. SVMs are utilized in content order, picture arrangement, penmanship acknowledgment, and technical studies.

For SVM, training involves the minimization of the error function:

$$\frac{1}{2}w^t w + c\sum_{i=1}^{N} \in i \quad (1)$$

Subject to the imperatives:

$$y_i(w^t \emptyset(x_i) + b) \geq 1 - \zeta_i \, and \, \zeta_i \geq 0, i = 1, \ldots, N \quad (2)$$

Where C is the limit steady, w is the vector of coefficients, b is a consistent, and $\zeta_i$ speaks to parameters for taking care of no distinguishable information (inputs). The list I name the N preparing cases. Note that y∈±1 speaks to the class names and xi speaks to the free factors.

- LR

Logistic regression was utilized in the natural sciences in the mid-twentieth century. It was then utilized in numerous sociology applications. Logistic Regression is utilized when the reliant variable (target) is categorical.

The strategic bend relates the free factor, X, to the moving mean of the DV, P (Y). The recipe to do so might be composed,

$$p = \frac{e^{a+bX}}{1+e^{a+bx}} \quad (3)$$

- KNN

KNN makes predictions dependent on the result of the K neighbors nearest to that point. Accordingly, to make forecasts with KNN, we have to characterize a measurement for estimating the separation between the question point and cases from the model's test. One of the most mainstream decisions to quantify this separation is known as Euclidean.

Euclidean Formula,

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (4)$$

### 3.2. Data Collection

- Parkinson Disease Datasets

In this study, we used the Parkinson disease data provided by the UCI Machine Learning Repository [14]. In addition, this dataset is comprising of 62 individuals with Parkinson disease and 15 people groups were sound. The author's utilized three kinds of recording are taken, for example, static winding test, dynamic winding test, and dependability test score. However, we have chosen the particular features for data analysis which are below presented,

1. No of strokes
2. Stroke speed
3. Velocity
4. Acceleration
5. Jerk
6. Horizontal velocity/acceleration/jerk
7. Vertical velocity/acceleration/jerk
8. Number of changes in velocity direction
9. Number of changes in acceleration direction
10. Relative NCV
11. Relative NCA
12. In air time
13. On surface time
14. Normalized in-air time
15. Normalized on-surface time
16. In air/on the surface ratio

### 3.3. Data Preprocessing

In this section, we extracted features from the Parkinson disease datasets. Then, we picked the 30 columns and 77 entries of data and conducted several experiments to checking missing values, redundant values. Figure 2 has shown that the 30 features from the dataset which are selected for our analysis.

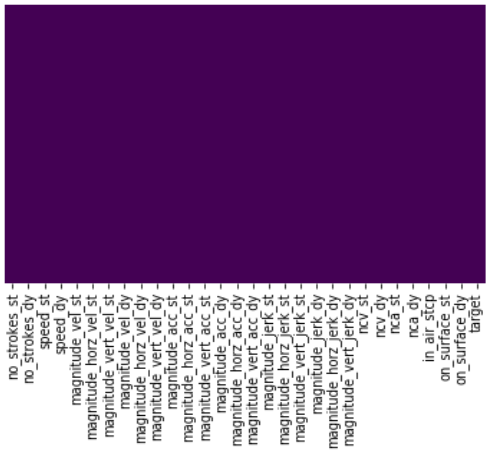


***Fig. 2**. Feature extraction from Parkinson Data.*

Therefore, analyzing the attributes of the selected Parkinson's datasets, some of them presented very few values whereas others appeared not correlated with the specific medical event. There were no missing values exist in this dataset. Figure 3 shows the number of missing values is empty. Moreover, Parkinson's datasets were also checked to verify the correlation of parameters. The heatmap appeared in figure 4 seems to have some corresponded parameters.

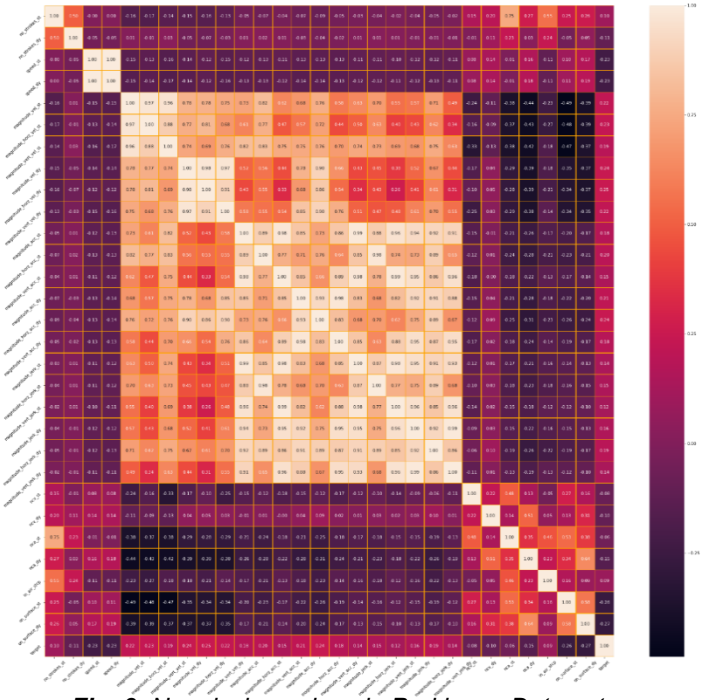

***Fig. 3**. No missing values in Parkinson Data sets.*

```
In [40]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 77 entries, 0 to 76
Data columns (total 30 columns):
no_strokes_st              77 non-null float64
no_strokes_dy              77 non-null float64
speed_st                   77 non-null float64
speed_dy                   77 non-null float64
magnitude_vel_st           77 non-null float64
magnitude_horz_vel_st      77 non-null float64
magnitude_vert_vel_st      77 non-null float64
magnitude_vel_dy           77 non-null float64
magnitude_horz_vel_dy      77 non-null float64
magnitude_vert_vel_dy      77 non-null float64
magnitude_acc_st           77 non-null float64
magnitude_horz_acc_st      77 non-null float64
magnitude_vert_acc_st      77 non-null float64
magnitude_acc_dy           77 non-null float64
magnitude_horz_acc_dy      77 non-null float64
magnitude_vert_acc_dy      77 non-null float64
magnitude_jerk_st          77 non-null float64
magnitude_horz_jerk_st     77 non-null float64
magnitude_vert_jerk_st     77 non-null float64
magnitude_jerk_dy          77 non-null float64
magnitude_horz_jerk_dy     77 non-null float64
magnitude_vert_jerk_dy     77 non-null float64
ncv_st                     77 non-null float64
ncv_dy                     77 non-null float64
nca_st                     77 non-null float64
nca_dy                     77 non-null float64
in_air_stcp                77 non-null float64
on_surface_st              77 non-null float64
on_surface_dy              77 non-null float64
target                     77 non-null float64
dtypes: float64(30)
memory usage: 18.1 KB
```

***Fig. 4.** Heat map for checking correlated columns in Parkinson data sets.*

### 3.4. Evaluation Criteria

In this work, we used three supervised learning strategies for the identification of Parkinson disease. Therefore, the performance measurements of the classifiers are evaluated by various measurable methods. For example, Recall, Precision,

f1- measure, etc. Hence, the calculation technique for the estimation considerations are as pursues [15],

Accuracy = $(TP + TN) / (TP + FP + TN + FN)$
(5)
Recall or sensitivity = $TP / (TP + FN)$
(6)
Precision = $TP / (TP + FP)$
(7)
f1 = $2 * (Recall * Precision) / (Recall + Precision)$
(8)

## 4 RESULT AND DISCUSSION

### 4.1. Analysis of the Result

In this segment, we directed different experiments to assess the three machine learning supervised algorithms for recognition of Parkinson Disease. The investigation of three classification techniques was evaluated for the exposure of Parkinson disease data. Figure 5 shows the accuracy of three supervised techniques. Here, SVM outperformed than LR and

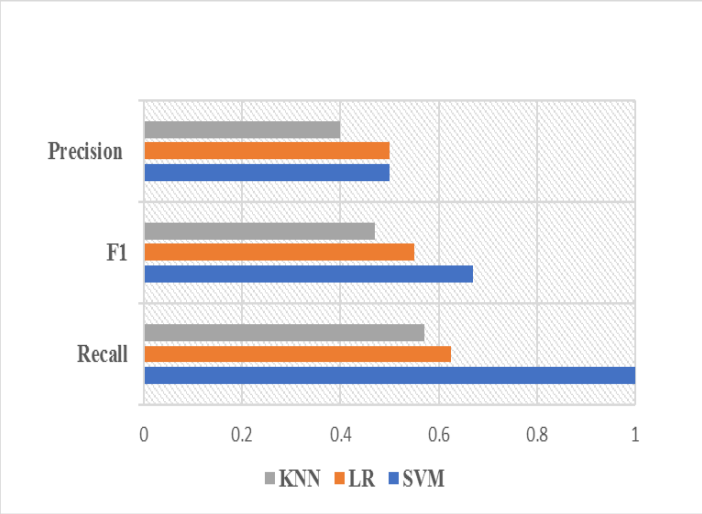


***Fig. 5****. Performance of six supervised classification techniques.*

However, the LR achieved 70% accuracy and KNN obtained 60% accuracy. Table 1and shows the classification performance measurements of six classification techniques.

***TABLE 1***
*Classification Performance Measurements*

| | Recall | F1 | Precision |
|---|---|---|---|
| SVM | 1 | 0.67 | 0.5 |
| LR | 0.625 | 0.55 | 0.5 |
| KNN | 0.57 | 0.47 | 0.4 |

According to the performance measurements of three classification algorithms are displayed in figure 6. The results evidently show that the SVM reached the maximum recall (100%). LR achieved the highest F1, it's 67%. KNN obtained the worst performance in terms of f1 measure (i.e. 0.47) and LR accomplished the second-most noteworthy score (i.e. 0.550. Then KNN also attained the worst precision (40%). Considering precision, SVM and LR show the same performance, it's around 50%, respectively. Finally, SVM is the highest performer by overall performance.

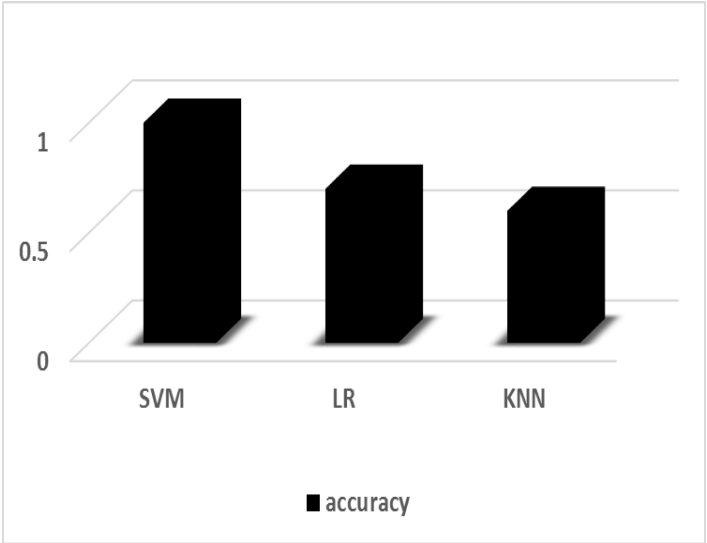


***Fig. 6.*** *Classification performance of three classifiers..*

## 5 CONCLUSION

In this analysis, we have represented three supervised learning machine learning approaches. A while later, the performance of the three classifiers which are utilized in the prediction of Parkinson disease and assessed their exhibition utilizing diverse statistical methods. The tentative performance demonstrations that the SVM has achieved the highest performance than the other two classifiers within the Parkinson datasets. It is 100%. This analysis has utilized three machine learning methods for the exposure of Parkinson disease in view of a few parameters. In accumulation, this work is part of a project that has the aim to cultivate an automated application to give more accurate action to normal occurrences and make a greater decision to multifaceted situations. The application will be able to detect in Parkinson disease in very few minutes and notify the dangerous probability of having the disease. This application can be outstandingly helpful in peoples, where is a lack of medical institutes and as well as particular physicians. In my experiments, each classification algorithms were prepared and assessed on a training set that includes both positive and negative samples. Moreover, the work can be supportive of Parkinson disease detection by collecting data from different clinical and medical centers and can provide more accurate results for disease prediction and diagnosis. In my research goal, there are several directions for future work in this area of research. We have only investigated three popular supervised algorithms; it can be preferring more algorithms for developing the precise model of these Parkinson disease prediction and performance can be more improved. In synopsis, our study painted the research objective besides opportunity with respect to Parkinson disease area by machine learning approaches, which has an arising impression in health fields

## ACKNOWLEDGMENT
The authors are grateful and pleased to all the researchers in this research study.

## REFERENCES
[1] C. Marras et al., "Prevalence of Parkinson's disease across North America," npj Park. Dis., vol. 4, no. 1, p. 21, Dec. 2018.
[2] D. Gelb, E. Oliver, S. G.-A. of neurology, and undefined 1999, "Diagnostic criteria for Parkinson disease," jamanetwork.com.
[3] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, and S. Sapir, "Early diagnosis of Parkinson's disease via machine learning on speech data," 2012 IEEE 27th Conv. Electr. Electron. Eng. Isr. IEEEI 2012, pp. 29–32, 2012.
[4] D. Aarsland, K. Andersen, J. L.-A. of neurology, and undefined 2003, "Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study," jamanetwork.com.
[5] "Parkinson's Disease Information Page | National Institute of Neurological Disorders and Stroke." [Online]. Available: https://www.ninds.nih.gov/disorders/all-disorders/parkinsons-disease-information-page. [Accessed: 08-Apr-2019].
[6] N. Singh, V. Pillay, Y. C.-P. in neurobiology, and undefined 2007, "Advances in the treatment of Parkinson's disease," Elsevier.
[7] "Speech impairment in a large sample of patients with Parkinson's disease," hindawi.com.
[8] R. Hossain, S. M. H. Mahmud, M. A. Hossin, S. R. Haider Noori, and H. Jahan, "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques," Procedia Comput. Sci., vol. 132, pp. 1068–1076, Jan. 2018.
[9] R. Das, I. Turkoglu, A. S.-E. systems with applications, and undefined 2009, "Effective diagnosis of heart disease through neural networks ensembles," Elsevier.
[10] A. K. Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," Neural Comput. Appl., pp. 1–9, Apr. 2017.
[11] A. Aljaaf, D. Al-Jumeily, … H. H.-2018 I. C., and undefined 2018, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," ieeexplore.ieee.org.
[12] S. M. H. Mahmud, M. A. Hossin, M. R. Ahmed, S. R. H. Noori, and M. N. I. Sarkar, "Machine Learning Based Unified Framework for Diabetes Prediction," in Proceedings of the 2018 International Conference on Big Data Engineering and Technology - BDET 2018, 2018, pp. 46–50.
[13] M. Razu Ahmed, S. M. Hasan Mahmud, M. Altab Hossin, H. Jahan, and S. Rashed Haider Noori, "A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms," 2018.
[14] "UCI Machine Learning Repository: Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Parkinson+Disease+Spiral+Drawings+Using+Digitized+Graphics+Tablet. [Accessed: 14-Jun-2019].
[15] A. D.-N. C. and Applications and undefined 2016, "Performance evaluation of different machine learning techniques for prediction of heart disease," Springer.