PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification

A Preprint

Ishu Gupta*
Cloud Computing Research Center
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
ishugupta23@gmail.com

Vartika Sharma
Department of Computer Applications
National Institute of Technology
Kurukshetra, India
136119
vsvartika.12sharma@gmail.com

Sizman Kaur
Department of Computer Applications
National Institute of Technology
Kurukshetra, India
136119
sizmankaur22@gmail.com

Ashutosh Kumar Singh
Department of Computer Applications
National Institute of Technology
Kurukshetra, India
136119
ashutosh@nitkkr.ac.in

Abstract

In this modern era of overpopulation disease prediction is a crucial step in diagnosing various diseases at an early stage. With the advancement of various machine learning algorithms, the prediction has become quite easy. However, the complex and the selection of an optimal machine learning technique for the given dataset greatly affects the accuracy of the model. A large amount of datasets exists globally but there is no effective use of it due to its unstructured format. Hence, a lot of different techniques are available to extract something useful for the real world to implement. Therefore, accuracy becomes a major metric in evaluating the model. In this paper, a disease prediction approach is proposed that implements a random forest classifier on Parkinson's disease. We compared the accuracy of this model with the Principal Component Analysis (PCA) applied Artificial Neural Network (ANN) model and captured a visible difference. The model secured a significant accuracy of up to 90%.

Keywords Parkinson's disease · Prediction · Learning · Feature selection · Random Forest Classification (RFC) · Principal component analysis (PCA) · Pyspark · Accuracy · Training · Testing · Artificial neural network (ANN)

1 Introduction

According to the fact sheets provided by the World Health Organization (WHO), the major cause that leads to disability and death all over the world is chronic diseases. Non-communicable diseases (NCD) kill around 40 million people each year, which is equal to 71% of all deaths worldwide. Heart diseases account for most of the NCD deaths, followed by cancers, respiratory diseases, and diabetes. This led to economic output loss of 47 US trillion dollars in the previous two decades. This loss represents globally around 75% of gross domestic product (GDP) in 2010 [1]. A study showed that in India around 25% of families with a member of cardiovascular disease and 50% of families suffering from cancer experience severe disastrous expenses and 10% and 25%, respectively, are affected due to poverty. Most of the estimations proved that the NCDs in India account for an overall economic loss in the range of 5-10% of GDP, which is a significant number and is thus slowing down GDP causing a loss in development [2].

~~People in India~~ have to spend more on treatment due to the lack of medical facilities and limited access to health insurance [3,4]. A lot of people do not buy insurance at the early stage of their life and then repent at an older age [5,6]. That's why it is very important to have an emergency fund saved separately. The healthcare issue of these types of diseases is crucial in other parts of the world as well [7, 8]. In America, over $10,000 per person is spent out of the pocket annually on health care, more than any other country in the world [9, 10]. The Partnership to Fight Chronic Disease estimates that around 83 million people in the U.S. will have 3 or more chronic health problems by 2030. Some behavioral factors including unhealthy eating, lack of exercise, excessive smoking, and use of alcohol are the main reasons for these chronic diseases [11, 12]. Therefore, it is essential to conduct risk analysis for chronic diseases. With a great increase in the world’s population, it would decline our life quality if we still depend on traditional systems of healthcare service [13, 14]. Such traditional systems of observing the real-time health condition of the person provided real quick assistance but with the rise in the number of data sources, there has been a flood of data in the healthcare sector [15, 16].

Due to such a big size of data records, it becomes very difficult to use the existing traditional techniques for effective results [17, 18]. Since Big data is a recent technology in the real world that can bring large benefits to business organizations, it becomes really necessary that various types of challenges and problems associated with adopting this technology are brought into the light [19, 20]. Detection of chronic diseases on early-stage helps in the early inception of preventive measures and on time-effective treatment at an initial stage has always been a better help for patients [21]. Currently, maintaining clinical data sets has become a crucial step in the medical field [22, 23]. The patient data that holds varied features and diagnostics related to a specific disease should be inserted in the dataset with the utmost care to provide the best quality results and services [24]. The data entered manually in medical data sets can contain a great amount of incomplete data and redundant values, mining such healthcare data becomes time-consuming [25]. As it can change the output of mining, it is important to incorporate good data preparation and data extraction prior to applying data mining algorithms. Prediction of the disease becomes faster and more efficient if data is accurate, consistent, and free from noises [26].

In this era of data explosion, a large amount of medical data is generated and updated daily [27]. The healthcare record, as well as any electronic variant of the traditional files, contains a proper identification of the patient [28, 29]. Medical data includes Electronic Health Records (EHR) which consists of clinical reports of patients, test reports from diagnosis, a prescription from a doctor, information from a pharmacist, information related to a person’s health insurance, social media posts such as blogs, tweets [30–32]. While the healthcare sector is moving towards using even more data sets into the daily routine diagnosis of patients suffering from chronic diseases, big data has truly become an effective tool in improving the patient’s treatment experience. Basic machine learning techniques with the assistance of big data technology have brought forward the concept of prediction in the medical sector [33]. Machine Learning capabilities have changed healthcare in various ways, improving diagnosis of treatment choices [34–36]. The predictive analysis implies doctors focus more on services and patient care. Prediction using classical models required data set of patients with disease and initial disease risk models used supervised machine learning techniques for training data [37, 38]. These models are helpful in clinical situations and are still studied globally [39, 40]. The features affecting a particular disease were selected through the doctor’s experience but they could not satisfy the changes in the disease later [41, 42].

In this paper, we proposed a **P**rincipal **C**omponent **A**nalysis based on **R**andom **F**orest (PCA-RF) model where we applied PCA on the medical data to extract the relevant features from the data set and along with that random forest algorithm is implemented as a classifier model. Principal Component Analysis (PCA) is an unsupervised learning algorithm employed for dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. Further, the accuracy of the model along with specificity and sensitivity is calculated. Finally, a comparison between the PCA-based Artificial Neural Network (ANN) and PCA-RF model is performed.

2 Related Work

The medical data contents can be maintained within the electronic format including images and may contain the patient’s identifiable details, like – films, digital imprints, or written conclusive outlines or interpretable findings. Medical data is generally represented in images or texts such as ECG, prescriptions, and MRI. Various researchers have suggested different technologies for feature selection from medical data for solving various problems like classification, regression, and retrieval. The prediction has various applications such as workload [43–45], security [46–48] and much more. In [49], the projected system used delivers a deep learning technique for efficient prediction of multiple distinctive diseases occurring in vulnerable areas which hold the high frequency of diseases. “It experimented with the modified estimate models over real-life medical data gained. It used a latent factor prototype to complete the missing data. It experimented on a regional chronic illness of cerebral infarction. It used machine learning Decision Tree algorithm and Map Reduce algorithm for data partitioning on structured and unstructured data taken from the hospital.”Compared to various crucial estimate algorithms, the accuracy of this proposed model reached 94.8%.

Various researchers have used different machine learning algorithms for feature extraction, classification of other chronic diseases such as Alzheimer disease [50, 51], heart disease [52–54], diabetes [55, 56]. In this paper, researchers used Indian Diabetes data taken from the UCI Repository. The system was implemented in Matlab. The Pima Indian Diabetes data set comprises approximately 768 real-life instances. The dataset contains the patient's summary and history and the output was predicted either as positive or negative. By analyzing the performance, it was perceived that among all the algorithms that were used for training, the Levenberg-Marquardt Algorithm gave the best results based on the epochs. In paper [57], a model was proposed to prove the outcome of a deep learning technique and to diagnose the heart disease the data set attained from the University of California, Irvine was used. The deep learning model was analyzed on the basis of performance and further was compared with the other four efficient machine learning algorithms for predicting the status of disease from data containing a record of 566 patients from a two-record set taken from the UCI database. This learning model accomplished an accuracy score of approximately 94% and an AUC score obtaining 0.964 in comparison to other models. The performance of this model and the non-linear machine learning algorithms was way better when compared to linear machine learning prediction models on increasing the data set size.

In the paper, [58], a hybrid learning prediction model was implemented with the help of missing value imputation (HPM-MI) that analyzes different techniques using K-means clustering and then applied with the most optimal imputation to a data set. This model was the first used amalgamation of K-means clustering along with Multi-layer Perceptron. To validate class labels for given data before applying classifier Kmeans clustering is used. This proposed system had greatly improved the quality of data by the use of the best imputation technique after the study of different eleven approaches. “The model is evaluated on the basis of prediction and classification system and is investigated on three criteria of medical data such as Pima Indians Diabetes, Wisconsin Breast Cancer, and Hepatitis dataset from the UCI Repository.”With respect to the performance matrix, the specificity, accuracy, sensitivity; kappa statistics, and the area under ROC were evaluated for best possible outcomes.

In paper [59], the investigation was done on the basis of the performance of different classification models. The breast cancer dataset which had 683 instances and 10 features were used for testing, based on classification accuracy. The breast cancer data found from the Wisconsin data set from the UCI repository was analyzed with the intention of developing an efficient prediction classifier for breast cancer disease with the help of various data mining approaches. In this observation, there were three classification methods that were implemented and comparison results showed that the Sequential Minimal Optimization (SMO) had a greater accuracy i.e. 96.2% than IBK and BF methods. There are a lot of other techniques used by researchers for predicting Parkinson's disease [60–65]. In this paper, we proposed a PCA-based decision tree model for the diagnosis of chronic disease at an early stage. Then, in the end, we have compared the ANN and Random Forest techniques for the same data set.

3 Proposed System

The proposed PCA-RF model uses a decision tree-based random forest algorithm in the core. In addition to that, it also exploits PCA that combines a large number of features into a comparatively small set of new principal components. Further, we pass these reduced features into our classifier that predicts the occurrence of disease. Finally, a comparative study is done using two classifying techniques – PCA along with ANN and PCA along with Random Forest based on various parameters. Fig. 1 represents the overall process of the proposed technique.

3.1 Feature Extraction

More features can decrease the accuracy of the model since there is more data that needs generalization. To reduce the model's complexity and avoid over-fitting of data, feature selection or feature extraction can be used. Feature extraction deals with deriving information from the data set containing a large number of features to construct a new feature sub-space. Here, the PCA algorithm is used for feature extraction. PCA finds the direction of greatest variance in higher dimension data set and projects it onto a new sub-space with fewer dimensions than the original one. Mathematically, the covariance matrix for the available data is computed from the mean-subtracted data matrix in Eq. (1). Then, the Eigen vector-matrix V (must be unit Eigen-vectors) which diagonalizes C is calculated in Eq. (2), where D represents the diagonal matrix for Eigen-values of C. The Eigen-vector with the maximum Eigen-value is the principal component of the data. The corresponding Eigen-vectors give the components in order of their importance. The reduced set of principal components is passed to the classifier for prediction.

$$C = \frac{1}{n-1} B * B \tag{1}$$

$$V^{-1} C V = D \tag{2}$$

3.2 Network Training

Random Forest is a prototype consisting of an ample number of decision trees. This model makes use of two major theories suggested the name “random ”. First, a random specimen of trained data points is done on constructing trees.

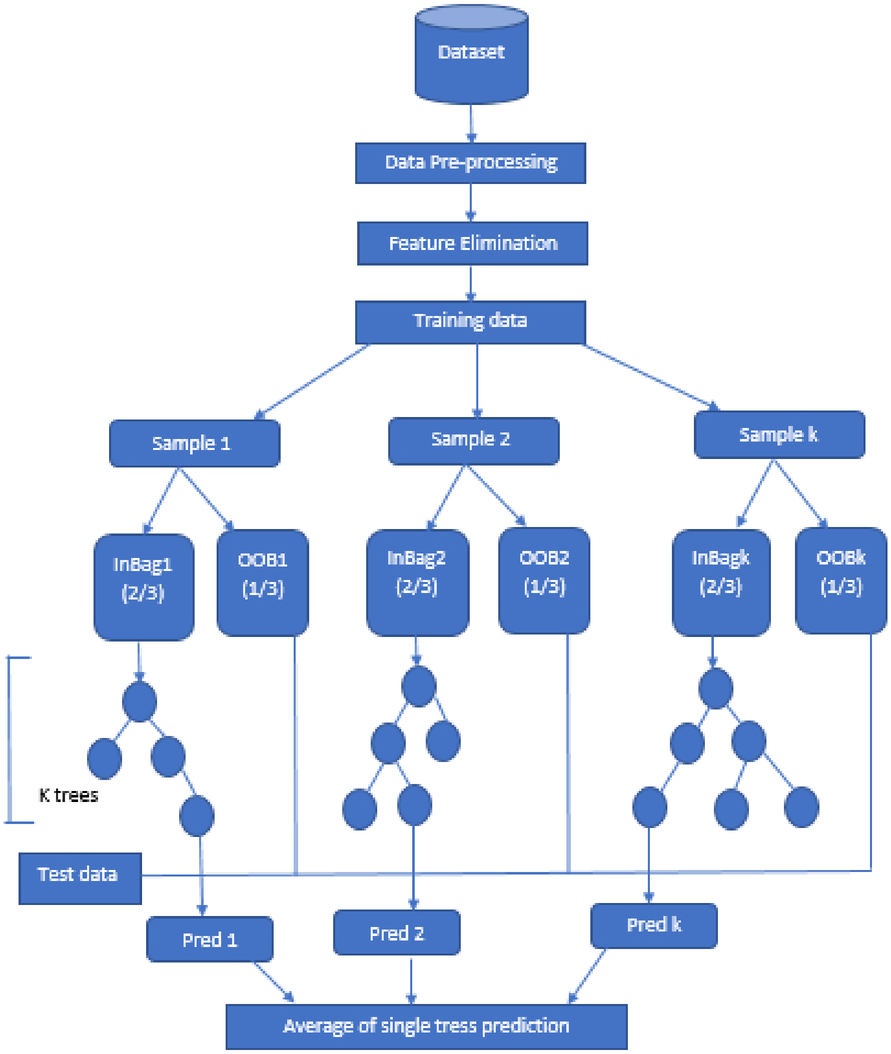

Figure 1: The workflow of the proposed disease prediction scheme

The next random divisions of features are taken into consideration when nodes split. During training, every tree in a random forest prototype understands from an arbitrary sample of the data points. The conclusive predictions of random forest are done by taking the average of the predictions of each distinctive tree. In order for random forest to make decisions like humans, it needs to learn things and for this learning, it needs to have a humongous amount of information in its training set. After learning from the training set, it is able to process the information and they can classify the given set of data into a predefined class.

3.3 Model Representation

The brick of a random forest is a decision tree which is an intuitive model. For example, a sequence of yes-no queries asked about the data gradually leading to a predicted class is a decision tree. It is an interpretable model because it constructs categorizations almost as humans do. This tree is constructed by shaping the questions via splitting of nodes as in Fig. 2. The answers from nodes lead to a high order reduction in Gini Impurity. Decision trees form nodes comprising a higher proportion of data points from a unit class by taking out values in the features that vividly divide the data into different classes.

3.4 Gini Impurity

The weighted average of Gini Impurity keeps on decreasing as we proceed further to the lower levels of the tree. The Gini Impurity of a node is the prospect that a sample preferred randomly in a node would be wrongly categorized if it

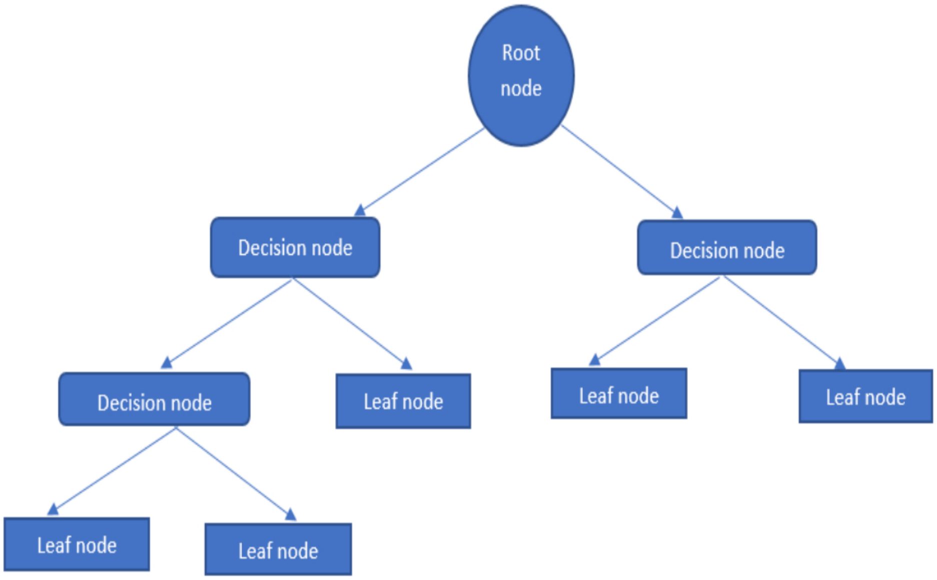

Figure 2: Predictive model's network representation

got categorized by the allocation of experiments in the node. Eq. (3) defines the formula for finding a Gini Impurity of any node n. At every node, the tree looks out for all the characteristics for finding the value to split on that can result in the severe drop in Gini Impurity. Then it reiterates the split process in a greedy, recursive manner until it reaches a maximum depth or each node has only samples from one class.

$$I_G(n) = 1 - \Sigma_{i=1}^{J}(p_i)^2 \tag{3}$$

4 Performance Evaluation

The implementation of the proposed model is described in this section followed by accuracy results.

4.1 Experimental Set-up and Benchmark Dataset

For implementing the model, the Pyspark platform is used with python language. Pyspark is used for structured and semi-structured data. It can also read data from various data sources having different file formats with the help of an API. The dataset is attained from the deep learning repository which is from the University of California(UCI) [66]. The patient's ages were ranging from 33 to 87 who was suffering from Parkinson's disease. The task is of classification and the dataset is of multivariate characteristics. The number of attributes is 754 and it has no missing values. The class values are 0 and 1, depicting the occurrence of disease or not. The database contains information like age etc. Forecast Accuracy, sensitivity, specificity, precision, and F1 Score are evaluated for the random forest classifier to validate the performance. The evaluation of these metrics for the proposed scheme is done using a confusion matrix.

4.2 Experimental Results

To evaluate the model, a 2×2 confusion matrix is used. The confusion matrix of size 2 × 2 is the measure of correct and wrong predictions which are abridged with count values. The model is evaluated using the following various performance metrics that are computed from the confusion matrix where TN reflects a true negative case, where TP reflects truly-positive cases, FP shows false-positive cases and FN holds the false-negative cases.

- **Sensitivity:** Sensitivity can be defined as the model's ability to accurately detect the patients who actually have the disease.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

- **Specificity:** Specificity can be defined as the model's ability to accurately identify the people who are healthy and who don't have the condition.

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

- **Accuracy:** Accuracy is delineated as to the frequency of correct result predictions out of the total predictions from the dataset.
$$Accuracy = \frac{\text{\#Number of correct predictions}}{\text{\#Total number of predictions}} \tag{6}$$
- **Precision:** Precision tells us how precise the prototype is out of the predicted positive and what number of those are really positive.
$$Precision = \frac{TP}{TP + FP} \tag{7}$$
- **F1 Score:** This score is the harmonic mean between two vital stats which are sensitivity and precision.
$$F1Score = \frac{2TP}{2TP + FP + FN} \tag{8}$$
- **ROC Curve:** Receiver's Operating characteristics curve is a performance metric for evaluating a classification model's performance. It is a probability curve that is plotted as TPR on the vertical dimension Y-axis and FPR on the horizontal dimension X-axis.

The performance in terms of accuracy, sensitivity, specificity, precision, and F1 Score of PCA-RF model without PCA and with PCA is depicted in Table 1. It can be seen that the PCA-RF model secures 89.9% and 76.7% accuracy, 70.2% and 55.6% sensitivity, 96.5% and 80.6% specificity, 70.2% and 35.1% precision, 77.7%, and 43% F1 Score without and with PCA respectively. It is observed that the performance without PCA is better compared to with PCA for every parameter accuracy, sensitivity, specificity, precision, and F1 Score since the performance greatly relies on the amalgamation of the feature reduction technique with the proposed classifier model.

Table 1: Performance metrıcs for PCA-RF model

Performance Metrics	Without PCA	With PCA
Accuracy	89.867	76.651
Sensitivity	70.175	55.555
Specificity	96.470	80.628
Precision	70.175	35.087
F1 Score	77.669	43.010

we have plotted the ROC curve for Parkinson's disease that is depicted in Fig.3. The figure shows the deflection of the model curve from the baseline where the horizontal axis (x-axis) demonstrates a false positive rate and the vertical axis (y-axis) depicts the true positive rate. It can be seen that the true positive rate of the proposed model is high which validates its performance.

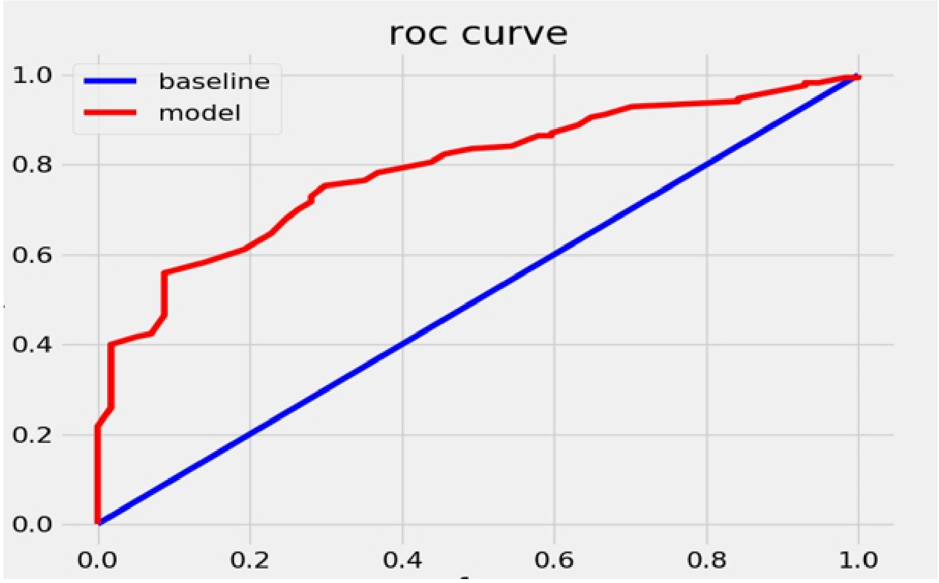

Figure 3: ROC curve of the proposed model

4.3 Comparative Analysis

Random forest classifier, when compared with Artificial Neural Network, shows a great deflection in the accuracy. Random forest works well with tabular data and being an intuitive model, it works on decision trees. At every step, each attribute plays a significant role in arriving at the desired result. Whereas Artificial Neural Network is an advanced algorithm of machine learning where a group of attributes is selected that participate in the classification process to predict the result. ANN has a large number of layers, each layer's output being input to the next layer. The accuracy of both the models is demonstrated in Table 2 when applied with and without a feature reduction technique. Table 2 shows the performance in terms of accuracy, sensitivity, specificity, precision, F1 score, and the comparison among ANN and Random Forest with PCA and without PCA respectively. The depicted results show a diverge deflection when applied with and without PCA. When applied with PCA the random forest classifier has low performance compared to ANN. Whereas, in the case of Artificial Neural Network when PCA has applied the accuracy greatly increases. While random forest classifier has high performance whereas ANN as comparatively low performance in case of without PCA. It depends on the importance of features and the dimension of the dataset. Dimensionality reduction is required for removing the redundancy in data that reduces time and storage. It becomes easy to visualize the data by reducing it to a low dimension.

Table 2: Comparison among ANN and Random Forest

Metrics	with PCA		without PCA	
	ANN	**Random Forest**	**ANN**	**Random Forest**
Accuracy	97.354	76.651	79.470	89.534
Sensitivity	95.454	55.555	64.516	70.175
Specificity	97.931	80.628	83.333	96.470
Precision	93.333	35.087	50.000	70.175
F1 Score	94.382	43.010	56.338	77.669

5 Conclusion and Future Scope

The work elaborated in this paper is based on the evaluation of Random forest classification on the particular dataset which is a high dimensional data containing 754 attributes. In terms of tabular data, random forest works well and so far the particular data we have opted for this classifier is the most appropriate selection. The paper also deals with how accuracy deflects when we pass the dataset to the proposed random forest model and when passed to the artificial neural network along with PCA. There is a visible difference between both models. Accuracy is high for the ANN model since a feature reduction technique is used along with it. It depends on the classifier model used along with the feature reduction technique that ensures good accuracy. Therefore, the accuracy of the proposed model is approximately 90%. The paper deals with the prediction of only one type of disease, so further diseases can be predicted simultaneously for early detection. Furthermore, the accuracy of the various models can be contrasted and the best classifier model can be selected for that particular disease.

References

[1] Kundu, M.K., Hazra, S., Pal D., Bhattacharya M.: A review on Noncommunicable Diseases (NCDs) burden, Indian Journal of Public Health, 62:4:302-4 (2018)

[2] P. Godha, S. Jadon, A. Patle, I. Gupta, B. Sharma, and A. K. Singh, "Flooding and Forwarding Based on Efficient Routing Protocol," in *International Conference on Innovative Computing and Communications*, vol. 1166. Singapore: Springer Singapore, 2021, pp. 215–223, advances in Intelligent Systems and Computing.

[3] Treating chronic diseases requires a lot more than hospitalization: Know the cost, The Indian Express, Nov 30 (2019)

[4] K. Kaur, I. Gupta, and A. K. Singh, "A Comparative Study of the Approach Provided for Preventing the Data Leakage," *International Journal of Network Security & Its Applications*, vol. 9, no. 5, pp. 21–33, 2017.

[5] P. Tiwari, S. Mehta, N. Sakhuja, I. Gupta, and A. K. Singh, "Hybrid Method in Identifying the Fraud Detection in the Credit Card," in *Evolutionary Computing and Mobile Sustainable Networks*, vol. 53. Singapore: Springer Singapore, 2021, pp. 27–35, data Engineering and Communications Technologies.

[6] I. Gupta, S. Mittal, A. Tiwari, P. Agarwal, and A. K. Singh, "TIDF-DLPM: Term and Inverse Document Frequency based Data Leakage Prevention Model," 2022.

[7] I. Gupta and A. K. Singh, "A Holistic View on Data Protection for Sharing, Communicating, and Computing Environments: Taxonomy and Future Directions," 2022.

[8] V. Sharma, S. Jalwa, A. R. Siddiqi, I. Gupta, and A. K. Singh, "A Lightweight Effective Randomized Caesar Cipher Algorithm for Security of Data," in *Evolutionary Computing and Mobile Sustainable Networks*, vol. 53. Singapore: Springer Singapore, 2021, pp. 411–419, data Engineering and Communications Technologies.

[9] Waters, H., Graf, M.: Chronic diseases are taxing our healthcare system and our economy, May 31 (2018)

[10] I. Gupta and A. K. Singh, "Dynamic Threshold based Information Leaker Identification Scheme," *Information Processing Letters*, vol. 147, pp. 69 – 73, 2019.

[11] A. K. Singh and I. Gupta, "Online Information Leaker Identification Scheme for Secure Data Sharing," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 31 165–31 182, November 2020.

[12] A. Acharya, H. Prasad, V. Kumar, I. Gupta, and A. K. Singh, "Host Platform Security and Mobile Agent Classification: A Systematic Study," in *Computer Networks and Inventive Communication Technologies*, vol. 58. Singapore: Springer Singapore, 2021, pp. 1001–1010, data Engineering and Communications Technologies.

[13] I. Gupta, T. K. Madan, S. Singh, and A. K. Singh, "HISA-SMFM: Historical and Sentiment Analysis Based Stock Market Forecasting Model," 2022.

[14] A. Kesharwani, A. Nag, A. Tiwari, I. Gupta, B. Sharma, and A. K. Singh, "Real-Time Human Locator and Advance Home Security Appliances," in *Evolutionary Computing and Mobile Sustainable Networks*, vol. 53. Singapore: Springer Singapore, 2021, pp. 37–49, data Engineering and Communications Technologies.

[15] I. Gupta and A. K. Singh, "An Integrated Approach for Data Leaker Detection in Cloud Environment," *Journal of Information Science and Engineering*, vol. 36, pp. 993–1005, Sep. 2020.

[16] R. Verma, V. Gautam, C. P. Yadav, I. Gupta, and A. K. Singh, "A Survey on Data Leakage Detection and Prevention," in *Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020*. SSRN, Elsevier, May 2020, pp. 1–7.

[17] I. Gupta and A. K. Singh, "GUIM-SMD: Guilty User Identification Model using Summation Matrix-based Distribution," *IET Information Security*, vol. 14, pp. 773–782, November 2020.

[18] S. Jalwa, V. Sharma, A. R. Siddiqi, I. Gupta, and A. K. Singh, "Comprehensive and Comparative Analysis of Different Files Using CP-ABE," in *Advances in Communication and Computational Technology*, vol. 668. Singapore: Springer Singapore, 2021, pp. 189–198, electrical Engineering.

[19] I. Gupta, H. Mittal, D. Rikhari, and A. K. Singh, "MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of A Day," 2022.

[20] K. Kaur, I. Gupta, and A. K. Singh, "E-Mail Protection System to Prevent Data Leakage," *Vigyan Prakash*, vol. 16, pp. 30–36, 2018.

[21] I. Gupta, P. K. Yadav, S. Pareek, S. Shakeel, and A. K. Singh, "Auxiliary Informatics System: an Advancement towards a Smart Home Environment," 2022.

[22] I. Gupta and A. K. Singh, "A Framework for Malicious Agent Detection in Cloud Computing Environment," *International Journal of Advanced Science and Technology (IJAST)*, vol. 135, pp. 49–62, Feb 2020.

[23] K. Kaur, I. Gupta, and A. K. Singh, "Data Leakage Prevention: E-Mail Protection via Gateway," *Journal of Physics: Conference Series*, vol. 933, p. 012013, jan 2018.

[24] P. Godha, S. Jadon, A. Patle, I. Gupta, B. Sharma, and A. Kumar Singh, "Architecture, an Efficient Routing, Applications, and Challenges in Delay Tolerant Network," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 824–829.

[25] I. Gupta and A. K. Singh, "A Hybrid Technique for the Detection of Data Leakage in Cloud computing Environment," in *Ist International Conference on Science in Hindi*, August 2017, vigyan Prakash.

[26] K. N. Kaur, Divya, I. Gupta, and A. K. Singh, "Digital Image Watermarking Using (2, 2) Visual Cryptography with DWT-SVD Based Watermarking," in *Computational Intelligence in Data Mining*, vol. 711. Singapore: Springer Singapore, 2019, pp. 77–86, advances in Intelligent Systems and Computing.

[27] Jain, D., Singh, V.: Feature selection and classification systems for chronic disease prediction: A review, Egyptian Informatics Journal, 19 (3), 179-189 (2018)

[28] A. Acharya, H. Prasad, V. Kumar, I. Gupta, and A. K. Singh, "MACI: Malicious API Call Identifier Model to Secure the Host Platform," in *Proceedings of the Seventh International Conference on Mathematics and Computing*. Singapore: Springer Singapore, 2022, pp. 309–320.

[29] K. Gupta and I. Gupta, "A Comprehensive Study on Architecture, Security issues and Challenges in Cloud Computing," *International Journal of Scientific & Engineering Research*, vol. 7, no. 12, pp. 128–131, Dec. 2016.

[30] Kumar, M., N., Manjula R.: Role of Big data analytics in rural health care – a step towards svasth bharath, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 7172-7178 (2014)

[31] P. Agarwal, S. Mittal, A. Tiwari, I. Gupta, A. K. Singh, and B. Sharma, "Authenticating Cryptography over Network in Data," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 632–636.

[32] A. Nag, A. Kesharwani, B. Sharma, I. Gupta, A. Tiwari, and A. K. Singh, "Potential and Extention of Internet of Things," in *Second International Conference on Computer Networks and Communication Technologies (ICCNCT)*, vol. 44. Cham: Springer International Publishing, 2020, pp. 542–551.

[33] A. K. Singh, I. Gupta, R. Verma, V. Gautam, and C. P. Yadav, "A Survey on Data Leakage Detection and Prevention," in *Proc. Int. Conf. Innov. Comput. Commun.*, 2020.

[34] I. Gupta, N. Singh, and A. Singh, "Layer-based Privacy and Security Architecture for Cloud Data Sharing," *Journal of Communications Software and Systems (JCOMSS)*, vol. 15, no. 2, 2019.

[35] I. Gupta and A. K. Singh, "SELI: Statistical Evaluation based Leaker Identification Stochastic Scheme for Secure Data Sharing," *IET Communications*, vol. 14, pp. 3607–3618, December 2020.

[36] Khushbu, P. Nishad, V. Kashyap, and I. Gupta, "A Classification and Distribution Model for Data Leakage Prevention and Detection," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 2, pp. 348–354, Feb. 2021.

[37] G. Batra, H. Singh, I. Gupta, and A. K. Singh, "Best Fit Sharing and Power Aware (BFSPA) Algorithm for VM Placement in Cloud Environment," in *2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall)*. IEEE, 2017, pp. 1–4.

[38] U. Arora, S. Verma, I. Gupta, and A. K. Singh, "Implementing Privacy using Modified Tree and Map Technique," in *2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall)*. IEEE, 2017, pp. 1–5.

[39] I. Gupta, "A Comparative Study of the Approach Provided for Preventing the Data Leakage," *Other Topics Engineering Research eJournal*, vol. 9, no. 5, September 2017.

[40] I. Gupta and K. Gupta, "Evaluation of Intrusion Detection Schemes in Wireless Sensor Network," *IOSR Journal of Computer Engineering*, vol. 18, no. 2, pp. 60–63, Mar-Apr. 2016.

[41] Khushbu, P. Nishad, V. Kashyap, I. Gupta, and A. K. Singh, "An Organized Study on Data Divulge Elimination and Discernment," in *Computer Networks and Inventive Communication Technologies*. Singapore: Springer Singapore, 2021, pp. 569–578.

[42] I. Gupta and K. Gupta, "Review on Intrusion Detection System Architectures in WSN," *International Journal of Scientific & Engineering Research*, vol. 7, no. 12, pp. 111–115, Dec. 2016.

[43] I. Gupta and A. K. Singh, "A Probabilistic Approach for Guilty Agent Detection using Bigraph after Distribution of Sample Data," *Procedia Computer Science*, vol. 125, pp. 662 – 668, 2018.

[44] I. Gupta, R. Gupta, A. K. Singh, and R. Buyya, "MLPAM: A Machine Learning and Probabilistic Analysis Based Model for Preserving Security and Privacy in Cloud Environment," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4248–4259, 2021.

[45] I. Gupta and A. K. Singh, "A Probability based Model for Data Leakage Detection using Bigraph," in *Proceedings of 7th International Conference on Communication and Network Security (ICCNS)*, ser. ICCNS 2017. New York, NY, USA: Association for Computing Machinery (ACM), 2017, p. 1–5.

[46] D. Saxena, I. Gupta, J. Kumar, A. K. Singh, and X. Wen, "A Secure and Multiobjective Virtual Machine Placement Framework for Cloud Data Center," *IEEE Systems Journal*, pp. 1–12, 2021.

[47] I. Gupta and A. K. Singh, "A Confidentiality Preserving Data Leaker Detection Model for Secure Sharing of Cloud Data using Integrated Techniques," in *2019 7th International Conference on Smart Computing Communications (ICSCC)*. Curtin University, Sarawak Malaysia: IEEE, 2019, pp. 1–5.

[48] K. Kaur, I. Gupta, and A. K. Singh, "A Comparative Evaluation of Data Leakage/Loss Prevention Systems (DLPS)," in *Proc. 4th International Conference Computer Science & Information Technology*, 2017, pp. 87–95.

[49] Vinitha, S., Sweetlin, S., Vinusha, H., Sajini, S.:Disease Prediction Using Machine Learning Over Big Data, Computer Science & Engineering: An International Journal (CSEIJ), Vol.8, No.1, February (2018)

[50] A. Alberdi et al., :Smart Home-Based Prediction of Multidomain Symptoms Related to Alzheimer's Disease, IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 6, pp. 1720-1731, Nov., doi: (2018)

[51] Kruthika, K., R., Rajeswari, Maheshappa, H., D.:Multistage classifier-based approach for Alzheimer's disease prediction and retrieval, Informatics in Medicine Unlocked, Volume 14, pp. 34-42, ISSN 2352-9148(2019)

[52] Panda, D., Dash, S.R.:Predictive System: Comparison of Classification Techniques for Effective Prediction of Heart Disease, In: Satapathy S., Bhateja V., Mohanty J., Udgata S. (eds) Smart Intelligent Computing and Applications. Smart Innovation, Systems and Technologies, vol 159. Springer, Singapore (2020)

[53] Ching-seh, W., Mustafa, B., Bhagwat, V.:Heart Disease Prediction Using Data Mining Techniques, 7-11. (2019)

[54] Singh, P., Singh, S., Pandi-Jain, G., S.:Effective heart disease prediction system using data mining techniques, International Journal of Nanomedicine, 15 March Volume 2018:13(T-NANO 2014 Abstracts) Pages 121—124.(2018)

[55] Jerjawi, E., Samer, N., Naser, A., Samy S.:Diabetes Prediction Using Artificial Neural Network, International Journal of Advanced Science and Technology, 121:54-64 (2018)

[56] Marie-Sainte S., L., Aburahmah, Almohaini, Saba,.:Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences. 9 (2019)

[57] Saji, S., A., Balachandran, K.:Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction, 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, pp. 201-206 (2016)

[58] Purwar, A., Singh, S., K.:Hybrid prediction model with missing value imputation for medical data, Expert Systems with Applications,Volume 42, Issue 13, 5621-5631 (2015)

[59] Chaurasia, V., Pal, S.: Novel Approach for Breast Cancer Detection Using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January (2014)

[60] Rastegar, A., D., Ho, N., Halliday, G., M.: Parkinson's progression prediction using machine learning and serum cytokines. npj Parkinsons Dis. 5, 14 (2019)

[61] Grover, S., Bhartia, S., Akshama, Yadav, A., Seeja K., R.: Predicting Severity Of Parkinson's Disease Using Deep Learning, Procedia Computer Science, Volume 132, Pages 1788-1794 (2018)

[62] Challa, K., N., R., Pagolu, V., S., Panda, G., Majhi, B.:An improved approach for prediction of Parkinson's disease using machine learning techniques, 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, pp. 1446-1451 (2016)

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

[64] Gokul S., Sivachitra, M., Vijayachitra, S.:Parkinson's disease prediction using machine learning approaches, 2013 Fifth International Conference on Advanced Computing (ICoAC), Chennai, 2013, pp. 246-252 (2013)

[65] Naghavi, N., Wade, E..:Prediction of Freezing of Gait in Parkinson's Disease Using Statistical Inference and Lower–Limb Acceleration Data. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 27. 947 - 955. (2019)

[66] Parkinson disease dataset from UCI repository : https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification