Original Article

A Novel Approach for Prediction of Human Disease using Symptoms by Multilayer Perceptron Algorithm to Improve Accuracy and Compared with Random Forest Algorithm

S.Avinash Prabhu[1], V.Parthipan[2]

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, TamilNadu, India, 602105.
[2]Project Guide, Corresponding Author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, TamilNadu, India, 602105.

Abstract

Aim : The aim of this paper is to improve Accuracy in Disease prediction using symptoms by a novel multilayer perceptron classifier in comparison with the Random Forest algorithm . **Materials and Methods :** Novel multilayer perceptron classifier and random forest sample size (N=10) to predict the accuracy percentage of predicted disease. G-power is calculated for two different groups, alpha (0.05), power (80%). **Results:** Based on the measurement of data, statistical analysis, and independent sample T-test, there is a statistically insignificant difference between the two study groups with value p=0.488 (p> 0.05). It was observed that the novel multilayer perceptron algorithm obtains accuracy as 95%. It appears to have better accuracy than the random forest (81%). **Conclusion:** The results prove that the novel multilayer perceptron algorithm there is a significant improvement in techniques with varied seed values in disease prediction using symptoms.

Keywords: Novel Multilayer perceptron, Random forest, Machine learning, Symptoms, Disease, Prediction.

DOI: 10.47750/pnr.2022.13.S04.079

INTRODUCTION

The purpose of this research work is to create a disease predictor model which is used to predict disease based on the symptoms entered by the user using a machine learning algorithm like the novel multilayer perceptron algorithm to improve accuracy (Bhoyar et al. 2021). Challenges faced by people in today's world are looking online for their treatment and health related information rather than rushing to hospitals and wasting their valuable time instead. So there is a need for such systems in today's modern world (Dahiwade, Patle, and Meshram 2019). Proposed model uses a novel multilayer perceptron algorithm which is used for better accuracy in predicting disease based on symptoms entered. Proposed model consists of a list of symptoms from which the user can choose any of the five symptoms to predict diseases (Grampurohit and Sagarnal 2020). Applications for this model deal with connection of databases (Harish and Gayathri 2019).

There are about 30 IEEE papers; 13 google scholar papers were published in recent years. From this research it has been found that machine learning and neural networks play a pivotal role in the heart disease prediction systems which can be used to get accurate and reliable results (Goel et al. 2019). The aim of this paper is to produce a model which is capable of predicting accurate outcomes, nowadays the evaluation field is filled with mixed data which is incapable of giving accurate outcomes to avoid these. SVM and naive bayes are two machine learning algorithms that are used and have an accuracy of 94% and 95% for prediction of disease using symptoms (Hamsagayathri and Vigneshwaran 2021). Several machine learning algorithms which are used in healthcare for prediction of disease, the model predicts only one disease based on the algorithms. There is no common system which can predict occurrences of diseases. The proposed model can predict multiple diseases based on the symptoms entered by the user (Yaganteeswarudu 2020). Data Mining, which is used to extract accurate data from the data collected to produce results in a reasonable way. Diabetes disease is predicted using the K nearest neighbor and naive bayes algorithm using these data mining techniques (Shetty et al. 2017). SVM kernels like linear, polynomial, sigmoid and RBF are employed in the diagnosis of diabetes in which RBF kernel performance

is said to be best and produce better accuracy than the other kernels. RBF can also be used to predict other diseases like thyroid cancer (Mohan and Jain 2020)
Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021). It has been proven that the machine learning algorithms in today's world are used to solve complex problems even when mixed or unordered data is used. Our main aim of this paper is to use the novel multilayer perceptron in prediction of diseases using symptoms and improve accuracy.

MATERIALS AND METHODS

The research work was performed in the OOAD Lab, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. Basically it is considered that two groups of classifiers are used namely Multilayer perceptron classifier and random forest. Group 1 is Multilayer perceptron algorithm with sample size of 10 and random forest algorithm is Group 2 with sample size of 10 and they are compared for a more accurate score for choosing the best algorithm. Pre- test analysis has been prepared using clinical.com by having a G power of 80% and threshold 0.05%.And a total of 20 samples in which standard deviation of MLP is .73391 and random forest is .90022 (Bhoyar et al. 2021).

Multilayer Perceptron
Multilayer perceptron is a form of feed-forward neural network. It uses supervised machine learning techniques like backpropagation and it is made up of three layers: input, concealed, and output. Input layer does not contain any neuron and the remaining layers contain neurons. Neurons utilize nonlinear activation functions. It is composed of various layers of nodes in the form of a directed graph, where each layer is fully connected to the next one except for the input data where each node is a neuron which collides with a non linear activation function. It is also called a standard linear perceptron.

Algorithm for MLP
Step 1 : Load the training dataset which is propagated through MLP input layers.
Step 2 : Inputs are pushed through MLP which takes a dot product between the input layer and hidden layer.
Step 3 : It utilizes activation functions which are calculated at each layer.
Step 4 : Calculated values are put together through any of the activation functions.
Step 5 : Move the other layer in the MLP and repeat Step 1.
Step 6 : Repeat step 3 and 5 until the output layer is reached.
Step 7 : Once it reaches the output layer, the calculations use a back propagation method which corresponds to activation functions.predicted output would be compared to the actual output, and an error percentage will be calculated and an accuracy score is calculated if satisfied stop else go to
Step 6.
Step 8 : Finally decisions will be made based on the output.
Step 9 : Stop.

Random Forest
Random forest algorithm is one of supervised machine learning algorithms which can be used for both classification and regression.It is flexible, easy to use and can be used to overcome limitations of the decision tree algorithm.

Algorithm
Step 1 : Load the dataset.
Step 2 : Firstly it chooses random data samples from a dataset.
Step 3 : It constructs decision trees for every sample dataset chosen.
Step 4 : At this step every predicted result will be compiled and voted on.
Step 5 : Finally, the most voted prediction will be selected and presented as a result of classification.
Step 6 : Stop.
Data collection for this research has been taken from a study of the University of Columbia performed at New York Presbyterian Hospital during 2004. Testing setup has all the components to do our test process. It has 2 types of configurations, Hardware configuration, and Software configuration. Hardware configurations include Intel core i3 5th generation processor, 8 GB RAM (Random Access Memory), 64-bit Windows OS. Software configuration includes Windows OS.

Statistical Analysis
IBM SPSS version 21 was used to conduct the study. It's a data-analytical tool. Ten iterations with a maximum of 10 samples were used to compare the MLP and Random Forest algorithms, and the projected accuracy for each iteration was documented. Symptoms are the dependent variables, while diseases of various forms are the independent variables. Finally, the value obtained from these iterations of the independent sample T- test was performed and a graph was plotted to know the exact difference between MLP algorithm and Random Forest algorithm (Bhoyar et al. 2021).

RESULTS

Table 1 shows statistical differences between multilayer perceptron and Random Forest algorithm; the sample size of 10 has been taken to calculate mean accuracy. Accuracy comparison of Random forest and random forest algorithms is shown in this table. Table 2 shows group statistical differences between two algorithms where it is found that mean accuracy for MLP is 95.43 % and that of Random Forest is 91.72% with a standard deviation of .81499. Results for the independent sample t-test are shown in Table 3.Comparison of MLP and Random forest mean accuracy has been shown in Fig.1. Graphical representation of the bar graph is plotted using groupid as X-axis Multilayer perceptron and Random forest. Y-Axis displaying the error bars with a mean accuracy of detection +/- 1 SD.

DISCUSSION

In this research for prediction of diseases using symptoms using MLP algorithm and have achieved an accuracy of ~95%. Moreover, a user interface model from which the user can pick any of the five symptoms listed in this model and can predict disease. Proposed model is connected with a database using sqlite3 which can be used for storing details entered by the user and can be used for their future medical history purposes.
With the detailed history of patients from various hospitals data is collected. By using a multi multi-process method where decision tree techniques and clustering methods prediction models for cardiac disease have been developed (Prabakaran and Kannadasan 2018). Based on this model ensemble, the heart disease prediction model has been developed which utilizes three ML algorithms like SVM, ANN and decision tree and can be used to give an accurate outcome (Wenxin 2020). For the early diabetes prediction machine learning algorithms like SVM, naive bayes and decision tree algorithms are used to build the model. The Naive Bayes algorithm gives the highest accuracy of 74.28% compared to other algorithms (Shafi and Ansari, n.d.). Various ML algorithms like logistic regression, random forest, KNN and naive bayes are used for prediction of various diseases related to heart cancer. Logistic regression has the highest accuracy of 92% than the other ML algorithms in this prediction model (Kumar 2021). With the structured and unstructured data by using the machine learning algorithms like decision tree map reduce which helps in predicting the diabetes disease prediction. This model gains better accuracy of 94% which is better than existing algorithms which are used for the analytics purposes (S et al., n.d.).
The limitations for this research is that it uses less data for prediction of diseases where it needs to update new diseases and symptoms which can predict the latest diseases more accurately. Future scope of this research is to collect more data related to the unknown diseases and their symptoms and update the dataset for further improvements.

CONCLUSION

Prediction of disease using symptoms has been successfully developed. Current study focused on different machine learning algorithms like Multilayer perceptron algorithm and random forest algorithm where the outcome of this study proved that Multilayer perceptron algorithm has higher accuracy of 95% than the random forest algorithm of 91%.

Declarations
Conflict of Interests
No conflict of interest

Authors Contribution
Author NNC was involved in data collection, data analysis, and manuscript writing. Author DV was involved in the Action process, Data verification and validation, and Critical review of the manuscript.

Acknowledgments
The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this research work successfully.

Funding: We thank the following organization for providing financial support that enabled us to complete this study.

1. Softeon Pvt.Ltd,Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Adhinarayanan, Rajesh, Aravindh Ramakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and Damodharan Dillikannan. 2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, June, 1–18.
2. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." Biomass Conversion and Biorefinery, August. https://doi.org/10.1007/s13399-020-00938-0.
3. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique." Biomass Conversion and Biorefinery, May. https://doi.org/10.1007/s13399-021-01551-5.
4. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." Fuel 304 (November): 121490.
5. Bhoyar, Sakshi, Nikki Wagholikar, Kshitij Bakshi, and Sheetal Chaudhari. 2021. "Real-Time Heart Disease Prediction System Using Multilayer Perceptron." 2021 2nd International Conference for Emerging Technology (INCET). https://doi.org/10.1109/incet51464.2021.9456389.
6. Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. 2019. "Designing Disease Prediction Model Using Machine Learning Approach." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). https://doi.org/10.1109/iccmc.2019.8819782.
7. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." Biomass Conversion and Biorefinery, February. https://doi.org/10.1007/s13399-021-01312-4.
8. Goel, Sakshi, Abhinav Deep, Shilpa Srivastava, and Aprna Tripathi. 2019. "Comparative Analysis of Various Techniques for Heart Disease Prediction." 2019 4th International Conference on Information Systems and Computer Networks (ISCON). https://doi.org/10.1109/iscon47742.2019.9036290.
9. Grampurohit, Sneha, and Chetan Sagarnal. 2020. "Disease Prediction Using Machine Learning Algorithms." 2020 International Conference for Emerging Technology (INCET). https://doi.org/10.1109/incet49848.2020.9154130.
10. Hamsagayathri, P., and S. Vigneshwaran. 2021. "Symptoms Based Disease Prediction Using Machine Learning Techniques." 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). https://doi.org/10.1109/icicv50876.2021.9388603.
11. Harish, S., and K. S. Gayathri. 2019. "Smart Home Based Prediction of Symptoms of Alzheimer's Disease Using Machine Learning and Contextual Approach." 2019 International Conference on Computational Intelligence in Data Science (ICCIDS). https://doi.org/10.1109/iccids.2019.8862163.
12. Jayanth, Bellappu Venkat, Melvin Victor Depoures, Gopal Kaliyaperumal, Damodharan Dillikannan, Dilipsingh Jawahar, Kumaran Palani, and Ganesha Prasad Meravanigee Shivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, June, 1–18.
13. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." Process Biochemistry 99 (December): 36–47.
14. Kumar, Anand. 2021. "Disease Prediction and Doctor Recommendation System Using Machine Learning Approaches." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.36234.
15. Mohan, Narendra, and Vinod Jain. 2020. "Performance Analysis of Support Vector Machine in Diabetes Prediction." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). https://doi.org/10.1109/iceca49313.2020.9297411.
16. Prabakaran, N., and R. Kannadasan. 2018. "Prediction of Cardiac Disease Based on Patient's Symptoms." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). https://doi.org/10.1109/icicct.2018.8473271.
17. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." Fuel 277 (October): 118166.
18. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." Fuel 278 (October): 118315.
19. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." Biomass Conversion and Biorefinery, March. https://doi.org/10.1007/s13399-021-01427-8.

20. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." Environmental Progress & Sustainable Energy 40 (6). https://doi.org/10.1002/ep.13696.
21. Shafi, Salliah, and Gufran Ahmad Ansari. n.d. "Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3852590.
22. Shanmugam, Rajasekaran, Damodharan Dillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 43 (23): 3064–81.
23. Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. 2017. "Diabetes Disease Prediction Using Data Mining." 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). https://doi.org/10.1109/iciiecs.2017.8276012.
24. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of Ganoderma Lucidum Using Saccharomyces Cerevisiae." Fuel 306 (December): 121680.
25. S, Vinitha, S. Vinitha, S. Sweetlin, H. Vinusha, and S. Sajini. n.d. "Disease Prediction Using Machine Learning Over Big Data." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3458775.
26. Wenxin, Xu. 2020. "Heart Disease Prediction Model Based on Model Ensemble." 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD). https://doi.org/10.1109/icaibd49809.2020.9137483.
27. Yaganteeswarudu, Akkem. 2020. "Multi Disease Prediction Model by Using Machine Learning and Flask API." 2020 5th International Conference on Communication and Electronics Systems (ICCES). https://doi.org/10.1109/icces48766.2020.9137896.

TABLES AND FIGURES

Table 1. Shows the statistical difference between the multilayer perceptron and random forest algorithms .Sample size (n=10) were taken and compared the accuracy between the two algorithms Therefore, by detailed analysis MLP and naive bayes are significantly different from each other.

S.No.	Multilayer Perceptron	Random Forest
1	95	91
2	94	90
3	95	91
4	94	92
5	93	92
6	93	91
7	94	91
8	93	92
9	95	90
10	94	92

Table 2. Group statistics of the Multilayer perceptron algorithm with the random forest algorithm by grouping the iterations with sample size 10, mean95.4350, standard deviation .73391 and standard error mean .23208. Descriptive independent sample test of accuracy and precision is applied for this dataset which is in SPSS.And here it specifies Equal variances with and without assuming a T-Test Score of two groups with each sample size of 10.

Group	N	Mean	Std. Deviation	Std. Error Mean
MLP	10	95.4350	.73391	.23208

RF	10	91.7200	.90022	.28468

Table 3. Independent Samples T-test shows significance value achieved is p=0.488 (p>0.05), which shows that two groups are statistically insignificant.

Dependent variables	**Assumptions**	**F**	**sig.**	**t**	**df**	**sig.(2-tailed)**	**Mean Difference**	**Std.Error Difference**	**Lower**	**Upper**
Accuracy	Equal variance assumed	0.502	.488	7.392	18	.000	2.71500	.36729	1.94335	3.48665
	Equal variances not assumed			7.392	17.298	.000	2.71500	.36729	1.94110	3.48890

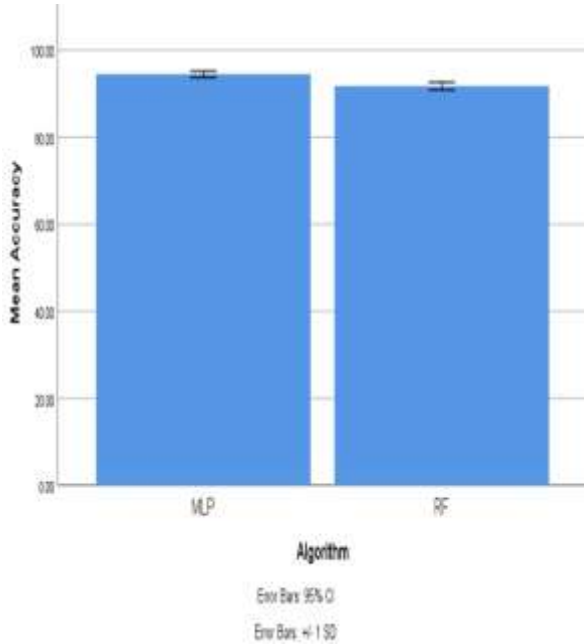

Fig. 1 : Comparison of mean accuracy of multilayer perceptron and Random forest algorithm it has been shown that there is a difference between two algorithms and multilayer perceptron algorithm has more accuracy than the Random forest .Graphical representation of bar is plotted were group id represents X -axis labels and mean accuracy in Y-axis. Graphical representation of this bar is plotted where the group id represents X -axis labels and mean accuracy in Y-axis with +/- 1 SD.