Original Article

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

S.AvinashPrabhu[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, 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 K nearest neighbor algorithm .
Materials and Methods : Multilayer perceptron classifier and naive bayes algorithm 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=0.768 (p>0.05) . It was observed that the novel Multilayer perceptron algorithm obtains the accuracy as 95%. It appears to have better accuracy than the K nearest neighbor (81%).
Conclusion: The results prove that the novel Multilayer perceptron algorithm approaches with varying seed value have significant improvement in disease prediction using symptoms.
Keywords: Novel Multilayer Perceptron, K-Nearest Neighbor, Machine Learning, Symptoms, Disease, Predict.

DOI: 10.47750/pnr.2022.13.S04.037

INTRODUCTION

The purpose of this research work is to identify disease type by symptoms using machine learning algorithms like novel multilayer perceptrons to improve accuracy (Bhoyar et al. 2021). In today's modern world with the emerging new technologies people are more sophisticated with virtual environments. So there is a need to develop such systems in healthcare and biomedical fields which can be less time consuming and prevent people from rushing to hospitals. ("DISEASE PREDICTION BASED ON SYMPTOMS USING CLASSIFICATION ALGORITHM" 2020a)("DISEASE PREDICTION BASED ON SYMPTOMS USING CLASSIFICATION ALGORITHM" 2020b)("DISEASE PREDICTION BASED ON SYMPTOMS USING CLASSIFICATION ALGORITHM" 2020a). Our proposed system deals with creating a user interface where users are provided with a list of symptoms and are asked to choose any of the five symptoms. Entered symptoms are given as input to our model where with the help of machine learning algorithms it predicts disease based on their probability and finally predicted disease is displayed to the user (Grampurohit and Sagarnal 2020). A database is created and is connected to the model . Details entered by users such as user name, user symptoms and algorithms which are used to predict accurate output are stored in the database (Harish and Gayathri 2019).

There are around 24 IEEE papers. 14 google scholar papers were published over the past few years As the adoption of new technology is increasing rapidly machine learning and artificial neural networks have a significant impact in healthcare and biomedical field (Goel et al. 2019). This paper proposed predictions of various diseases using symptoms where ML algorithms like SVM have got 94% accuracy and the naive bayes algorithm has 95% accuracy (Hamsagayathri and Vigneshwaran 2021). The aim of this paper is to develop a disease prediction system to predict

correct output with the help of medical data. Machine learning algorithms like CNN and K-Nearest Neighbor were used to predict diseases. It has been found that the CNN algorithm performance was better in terms of accuracy and time than the K-Nearest Neighbor algorithm (Dahiwade, Patle, and Meshram 2019). This study focuses on prediction of heart diseases using various machine learning classifiers like Random Forest, Kstar, Zeror, Voted novel Perceptron classifier. In which the random forest algorithm has got the highest accuracy of 97%, Kstar is 94.05%, ZeroR is 85.14%, Voted Perceptron is 94.39%. It was found that the random forest algorithm has higher accuracy than the remaining algorithms used (Roy, Mahmood, and Roy, n.d.). This paper refers to predictions of multiple diseases at a time. So that users need not be trying various prediction models for disease prediction this developed prediction model enables the user to enter input and users can select a particular symptom whether they are affected by certain diseases or not (Yaganteeswarudu 2020). In this paper machine learning algorithms like decision tree and map reduce are used which takes the structured and unstructured data together.Compared to several analytical algorithms which produces different outcomes the proposed model provides the accuracy of 94% which is better compared to existing algorithm in predicting the diabetes disease (S et al., n.d.)

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). By our top to bottom examination the machine learning algorithms are suitable enough and easy to implement these algorithms for prediction and pattern findings. And classification algorithms like K-Nearest Neighbor, CNN, MLP play a vital role in giving correct predictions according to data provided. Our proposed system deals with overcoming these problems by providing a better accuracy than the existing model.

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 K-Nearest Neighbor. Group 1 Multilayer perceptron algorithm with a sample size of 10 and K-Nearest Neighbor 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 the standard deviation of MLP is .81499 and KNN is.94285 (Bhoyar et al. 2021).

Multilayer Perceptron

The multilayer perceptron is a form of feed forward neural network. It uses supervised machine learning techniques like backpropagation and consists of three layers namely input, hidden, output layers.And input layer does not contain any neuron and remaining layers contain neurons. These neurons utilize non linear activation functions. It is composed of multiple layers of nodes in the form of a directed graph, where each layer is fully connected to the next one except for 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 steps 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 will be compared with the output and error is determined 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 : End

K Nearest Neighbor algorithm

K-Nearest Neighbor is a form of supervised machine learning algorithm and can be used to find pattern finding, data mining and also the best choice for addressing classification related problems. It is based on the idea of feature similarity instances. This algorithm is capable of handling a vast amount of data where there is a non linear decision division between classes. It doesn't perform preparation by any means and it doesn't seek after any unfair capacity from preparing information rather it holds the preparation dataset.

Algorithm for KNN

Step 1 : Load training dataset.
Step 2 : Prepare data by scaling, dimensionality reduction as required missing value treatment.
Step 3 : Choose K value .
Step 4 : Calculate Euclidean distance between all of the training data samples.
Step 5 : Sort and save the distances in an ordered list.
Step 6 : From the sorted list, select top K entries.
Step 7 : Majority of classes contained in selected points are used to label the test point.
Step 8 : End.

Dataset 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 analysis. It's a data-analysis statistical tool. For comparing the MLP and K-Nearest Neighbor algorithms, an independent sample T-test was performed with a maximum of 10 samples and the projected accuracy was logged for each iteration in order to anticipate the correct accuracy. The dependent variables are symptoms and the independent values are various types of diseases. Finally the value obtained from this iteration from these samples. T- test was performed and a graph was plotted to know the exact difference between MLP algorithm and K-Nearest Neighbor algorithm.

RESULTS

Table 1. Shows the statistical difference between novel multilayer perceptron and KNN algorithms and sample size of 10 has been taken to calculate mean accuracy. Accuracy comparison of KNN and Gaussian naive bayes 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 95.03 % and that of the naive bayes 81.38% with a standard deviation of .81499 and .94285 and Standard error mean of .25772 and .29816. The results for the independent sample t-test are shown in Table 3 with their mean difference whether they are 2 tailed or not . A comparison bar chart of MLP and KNN mean accuracy has been shown in Fig. 1. Graphical representation of the bar graph is plotted using groupid as X-axis novel Multilayer perceptron and KNN. Y-Axis displaying the error bars with a mean accuracy of detection +/- 1 SD.

DISCUSSION

From this an accuracy of ~95% is obtained. And added a way to deal with storing the data entered by the customer-like name, symptoms and algorithm used. Our system furthermore has an easy to utilize interface. It has an alternate visual depiction of data assembled and results achieved by utilizing sqlite3 database which can be used by users for their medical history purposes.

This research work deals with the prediction of heart diseases based on the symptoms which were analyzed from several patients. This data was collected and used to predict whether a person can be affected by heart disease or not. Machine learning algorithms like decision tree and data mining techniques were used (Prabakaran and Kannadasan 2018). Heart disease prediction models where algorithms like decision tree, SVM and artificial neural networks were

used to predict heart disease. In order to prevent misjudgement by one algorithm, grouping of three algorithms were used by the voting system through which accurate results can be obtained (Wenxin 2020). Diabetes disease prediction model which uses naive bayes and k nearest neighbor algorithms which helps in getting better accuracy. This model uses a data mining approach to predict disease where data can be taken from this dataset provided with all the information related to diabetes disease (Shetty et al. 2017). Kernels of SVM like RBF, linear, polynomial and sigmoid are used to predict diabetes diseases. Analysis of performance using these four kernels was calculated where it is found that the SVM RBF kernel has got the highest accuracy than other kernels in predicting accurate results (Mohan and Jain 2020).

The limitations for this model is that the proposed system deals with less data related to the diseases and their symptoms which may result in partial disease predictions. Future scope of this project is to collect more data related to new diseases and their symptoms which results in accurate prediction of diseases.

CONCLUSION

Prediction of disease using symptoms has been successfully developed. Our current study focused on different machine learning algorithms like Multilayer perceptron algorithm and KNN algorithm where outcome for this study has proved that Multilayer perceptron algorithm has higher accuracy of 95% than the naive bayes algorithm of 81%.

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 the study.

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

REFERENCE

1. Adhinarayanan, Rajesh, AravindhRamakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and DamodharanDillikannan. 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. "DISEASE PREDICTION BASED ON SYMPTOMS USING CLASSIFICATION ALGORITHM." 2020a. *JOURNAL OF XI'AN UNIVERSITY OF ARCHITECTURE & TECHNOLOGY*. https://doi.org/10.37896/jxat12.04/1055.———. 2020b. *JOURNAL OF XI'AN UNIVERSITY OF ARCHITECTURE & TECHNOLOGY*. https://doi.org/10.37896/jxat12.04/1055.
9. 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.
10. 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.
11. 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.
12. 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.
13. Jayanth, BellappuVenkat, Melvin Victor Depoures, GopalKaliyaperumal, DamodharanDillikannan, DilipsinghJawahar, KumaranPalani, and Ganesha Prasad MeravanigeeShivappa. 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.
14. 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.
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. Roy, Diti, Md Ashiq Mahmood, and Tamal Joyti Roy. n.d. "An Analytical Model for Prediction of Heart Disease Using Machine Learning Classifiers." https://doi.org/10.36227/techrxiv.14867175.
21. 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.
22. Shanmugam, Rajasekaran, DamodharanDillikannan, 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 GanodermaLucidum 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. Comparing accuracy values between multilayer perceptron and naive bayes algorithm with sample size n=10. Accuracy of MLP (95.03) and naive bayes (81.31).

S.No.	Multilayer Perceptron	K Nearest Neighbor
1	95	81
2	94	80
3	93	81
4	95	80
5	94	80
6	95	82
7	95	80
8	94	81
9	95	82
10	93	81

Table 2. Shows group statistics Multilayer perceptron algorithm with KNN algorithm by grouping this iteration with sample size 10, got a mean 95.0340, standard deviation .81499 and standard error mean .25772. 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	95.0340	.81499	.25772
DT	10	81.3890	.94285	.29816

Table 3. Independent Sample Test of Accuracy p value achieved is 0.768 ($p>0.05$), which shows it is a statistically insignificant difference between the study groups. Therefore, by detailed analysis MLP and naive bayes are significantly different from each other and both the algorithms have mean difference 13.6450 with standard error difference of .39410.

Dependent variables	Assumptions	F	sig.	t	df	sig.(2-tailed)	Mean Difference	Std.Error Difference	Lower	Upper
Accuracy	Equal variance assumed	0.90	.768	34.623	18	.000	13.64500	.39410	12.81702	14.47298
	Equal variances not assumed			34.623	17.631	.000	13.64500	.39410	12.81578	14.47422

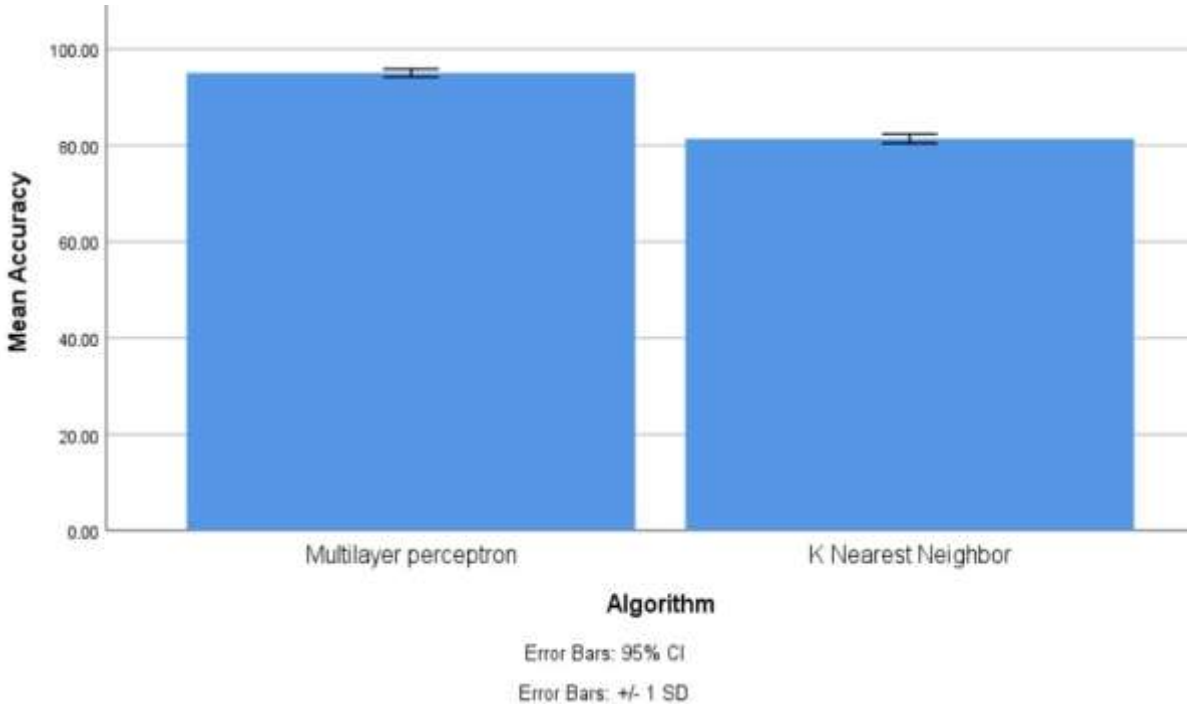

Fig. 1. Comparison of mean accuracy of multilayer perceptron and KNN algorithm it has been shown that there is a difference between two algorithms and multilayer perceptron algorithm has more accuracy than the KNN. Graphical representation of the bar is plotted where the group id represents X -axis labels and the 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.