DISEASE PREDICTION USING MACHINE LEARNING

Seminar Report submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

By

NACHIKETA KUMAR	**17UECS0478**
K VARUN TEJA	**17UECS0370**
RANJAN JYOTI DAS	**17UECS0639**

Under the guidance

Dr. Paras Jain M.E., P

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SCHOOL OF COMPUTING

VEL TECH RANGARAJAN Dr. SAGUNTHALA R & D INSTITUTE OF SCIENCE AND TECHNOLOGY, CHENNAI 600 062, TAMILNADU, INDIA

(Deemed to be University Estd u/s 3 of UGC Act, 1956)

October, 2019

DISEASE PREDICTION USING MACHINE LEARNING

Seminar Report submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

By

NACHIKETA KUMAR	**17UECS0478**
K VARUN TEJA	**17UECS0370**
RANJAN JYOTI DAS	**17UECS0639**

Under the guidance of

Dr. Paras Jain M.E., Ph.D

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SCHOOL OF COMPUTING

VEL TECH RANGARAJAN Dr. SAGUNTHALA R & D INSTITUTE OF SCIENCE AND TECHNOLOGY, CHENNAI 600 062, TAMILNADU, INDIA

(Deemed to be University Estd u/s 3 of UGC Act, 1956)

October, 2019

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

This is to certify that the seminar entitled **"DISEASE PREDICTION USING MACHINE LEARNING"** submitted by NACHIKETA KUMAR (17UECS0478), K VARUN TEJA (17UECS0370) and RANJAN JYOTI DAS (17UECS0639) in partial fulfillment for the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the seminar report has not been submitted to any other University/Institute for the award of any Degree or Diploma.

Signature of Supervisor

Dr. PARAS JAIN M.E, Ph.D,

Associate. Professor,
Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology,
Avadi, Chennai-600062

Signature of Seminar Handling Faculty

Dr. M. KAVITHA M.Tech, Ph.D,

Associate Professor,
Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology,
Avadi, Chennai-600062.

Submitted for the partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering from Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology (Deemed to be University, u/s 3 of UGC Act, 1956).

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Nachiketa 21/10/19
(Signature with date)

NACHIKETA KUMAR
(17UECS0478)

K. Varun Teja 21/10/19
(Signature with date)

K VARUN TEJA
(17UECS0370)

RJDas 21/10/19
(Signature with date)

RANJAN JYOTI DAS
(17UECS0639)

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act. 1956)

APPROVAL SHEET

The Seminar report entitled **"DISEASE PREDICTION USING MACHINE LEARNING"** by NACHIKETA KUMAR, K VARUN TEJA and RANJAN JYOTI DAS is approved for the degree of B.Tech Computer Science and Engineering.

21/10/22

Signature of Supervisor

Dr. PARAS JAIN, M.E., Ph.D,

Associate. Professor,
Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology,
Avadi, Chennai-600062

Signature of HEAD & DEAN, SOC

R. V. SRINIVASA R, M.Tech Ph.D,

Professor,
Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology,
Avadi, Chennai-600062.

Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
Department of CSE

Date:

Place:

ACKNOWLEDGEMENT

We express our deepest gratitude to our respected **Founder Chancellor and President Col. Prof. Dr. R. RANGARAJAN B.E. (EEE), B.E. (MECH), M.S (AUTO). DSc. Foundress President Dr. R. SAGUNTHALA RANGARAJAN M.B.B.S.**, Chairperson Managing Trustee and Vice President.

We are very much grateful to our beloved **Vice Chancellor Prof. V. S. S. KUMAR, Ph.D,** for providing us with an environment to complete our seminar successfully.

We are obligated to our beloved **Registrar N.S PREMA,** for providing immense support in all our endeavours.

We are thankful to our esteemed **Director Academics Dr. ANNE KOTESWARA RAO, Ph.D,** for providing a wonderful environment to complete our seminar successfully.

We record indebtedness to our **Head of the Department/Dean Dr. V. SRINIVASA RAO, M.TECH, Ph.D,** for immense care and encouragement towards us throughout the course of this seminar.

We thank our **Seminar handling Faculty Dr. M. KAVITHA** for the valuable information shared on proceeding with our seminar.

We also take this opportunity to express a deep sense of gratitude to Our **Internal Supervisor Dr. PARAS JAIN, M.E, Ph.D,** for his/her cordial support and guidance, he/ she helped us in completing this seminar through various stages.

A special thanks to our **Seminar Coordinators Mrs. S.HANNAH M.E., Mrs. D. FEMI M.E. and Mrs. T. ANJALI M.E.** for their valuable guidance and support throughout the course of the seminar. We thank our department faculty, supporting staff, parents, and friends for their help and guidance to complete this seminar.

NACHIKETA KUMAR (VTU10258) (17UECS0478)

RANJAN JYOTI DAS (VTU10703) (17UECS0639)

K VARUN TEJA (VTU10640) (17UECS0370)

TABLE OF CONTENTS

ABSTRACT

The study on machine learning gives us an insight of the different algorithms and its applications in various fields. There are various algorithms available in the categories of data mining, machine learning and statistics when we assemble your predictive analysis model. As we explore the data it becomes easier to take further decision. Algorithms like k-means, Naive Bayes, Support vector, Gradient-boosting methods, etc are available. The algorithms use different methods and approaches resulting in different percentage of accuracy and prediction. The time taken to predict the output accurately or to make the prediction better are also important factors. We do comparisons fairly to figure out the best and efficient algorithms. Later this will help us with our disease prediction. Disease prediction is one of the important research fields since the data available is vast considering the widespread of diseases and variety of reasons causing these diseases. Predicting diseases is helpful in preventing any complicated factors in future. Most importantly the prediction should be accurate enough to get medical treatment.

By the end of this we will be able predict the diseases, also we choose any particular disease to determine its chances of infecting or affecting us. Also predicting the chances of risks can be done using these algorithms. All we need is huge amount of data , training algorithm which learns to find the hidden pattern that can help in prediction, calculating the risk factor etc.

The final results of our report are to make you understand the various algorithms, analysis and prediction.

CHAPTER 1

INTRODUCTION

Machine learning is the scientific approach of studying algorithms and statistical models by relying purely on the patterns and inferences without giving any explicit instructions or coding to gather important information by learning and improving the experience it gained. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. We have some machine learning methods categorized as supervised and unsupervised.

Supervised machine learning algorithms is the part we train the machine by providing it with trained data set which gives it experience and ways to predict the outputs. This training happens till the machine is able to predict the correct output for a given training data set. Once it is able to, the machine is given with new data set to predict the output.

Unsupervised machine learning algorithms are used when the information used to train has no classification or labels which makes the machine to explore the data rather than to predict right output. It basically draws inferences from data sets to describe hidden structures from unlabeled data.

Semi-supervised machine learning algorithms falls between supervised and unsupervised since it uses both labeled and unlabeled data sets to train the machine. This is to increase its learning accuracy.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific field in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

The data sets are used for machine learning research and have been cited in peer-reviewed academic journals. Data sets are an integral part of the field of machine learning. Major advances in this field can result from advances in learning algorithms, computer hardware, and, less-intuitively, the availability of high-quality training data sets. High-quality labeled training data sets for

supervised and semi supervised machine learning algorithms are usually difficult and expensive to produce because of the large amount of time needed to label the data. Although they do not need to be labeled, high-quality data sets for unsupervised learning can also be difficult and costly to produce.

Disease prediction is a helpful concept considering the present condition of the world where disease outbreak is common. Keeping track of our health condition and predicting diseases can help us for better survival. The above-mentioned ML techniques can be used for disease prediction. A variety of datasets are available online to run simulations for training the algorithms. Choosing a appropriate dataset can help in focusing on the end result. It is important to diagnosis the patients with proper treatment when infected by a disease. By considering various parameters the algorithms predict the disease, determine the risk factors, suggest treatments, etc.

We can retrieve the necessary datasets from various online sources and use the recently popular and successful language python. Python has different versions; we opt to go with the python 3.7. It is easy to program using python since a lot of packages are available online which can be used by importing. A lot of preprocessing and training and testing has to be done have to be done.

CHAPTER 2

LITERATURE SURVEY

2.1 MACHINE LEARNING

A widely used term and method in computer science. As stated by the geeks for geeks machine learning[8] is a field of study that gives computer the capability to learn without being explicitly programmed. It can also be said as automation and also improves its productivity by continuously learning without any human assistance. We usually do traditional programming in which we give input and also give the logic which the computer computes and gives us the result but in machine learning only input is given and the machine tries to make the logic to get optimal results. Machine learning is comprised of different algorithms. The algorithms are classified as: Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning.

The life of machine learning program can be summarized in the following manner. Firstly, the problem is defined then the required data is collected which has to be visualized and then the algorithm is trained also there is a test algorithm which collects the feedback of each training and modifies the logic as the training algorithm loops till the results are satisfactory. After this all a prediction model is adopted. This is now tested on the new datasets to get the predictions. This applies to most of the techniques we use in machine learning. Now let us discuss about the supervised machine learning which is comprised of classification and regression methods.

Under classification supervised machine learning we have following algorithms:

Naive Bayes

Support Vector Machine

K-Nearest Neighbor (KNN)

Decision Tree

Random Forest

Now under regression supervised machine learning we have following algorithms:

Simple Linear Regression

Polynomial Regression

Support Vector Regression

Decision Tree Regression

Random Forest Regression

We have unsupervised machine learning it has the respective algorithms:

Clustering

K-Means Clustering

Hierarchical Clustering

Dimensionality Reduction

Principal Component Analysis

Singular Value Decomposition

Now when we discuss about reinforcement machine learning we have the following methods like:

Exploration/ Exploitation trade-off

Markov Decision Processes

Q- learning, policy learning and deep reinforcement learning

Value learning problem

Then comes the semi-supervised which works in between supervised and unsupervised.

2.2 NAIVE BAYE

As stated by the L. Jiang, C. Li[1] the Naive Bayes is a probabilistic classifier inspired by the Bayes theorem. The classification is conducted by deriving the maximum posterior which is the maximal P (Ci |X) with the above. Assumption applying to Bayes theorem. This assumption greatly reduces the computational cost. Simple algorithm to implement. It can be easily scalable to larger datasets since it takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

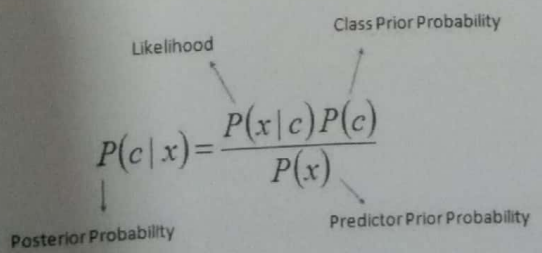

Fig:2.1 Naïve Bayes

2.3 SIMPLE LINEAR REGRESSION

Enrico R, Michel L[4], proposed a small theory on the linear regression which is as follows. Here we predict a target variable Y based on the input variable X. A linear relationship should exist between target variable and predictor and so comes the name Linear Regression.

$$Y = a + bX$$

Our aim is to find such values of coefficients which will minimize the cost function. The most common cost function is Mean Squared Error (MSE) which is equal to the average squared difference between an observation's actual and predicted values. The coefficient values can be calculated using Gradient Descent approach. While training and building a regression model, it is these coefficients which are learned and fitted to training data. The aim of training is to find a best fit line such that the cost function is minimized. The cost function helps in measuring the error. During training process, we try to minimize the error between actual and predicted values and thus minimizing cost function.

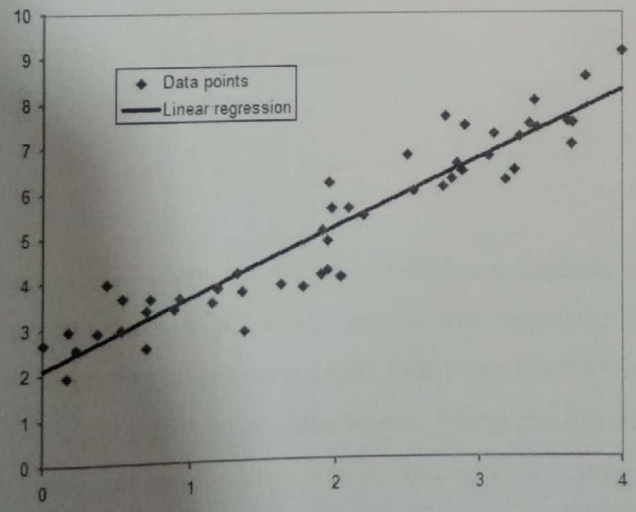

Fig:2.2 simple linear regression

In the figure, the red points are the data points and the blue line is the predicted line for the training data. To get the predicted value, these data points are projected on to the line.

2.4 k-MEANS CLUSTERING

K-means clustering as referred from W. Zhao[3], follows to be like, if we want to cluster our data points into k groups. A larger k creates smaller groups with more granularity, a lower k means larger groups and less granularity. With k-means clustering, we want to cluster our data points into k groups. A larger k creates smaller groups with more granularity, a lower k means larger groups and less granularity.

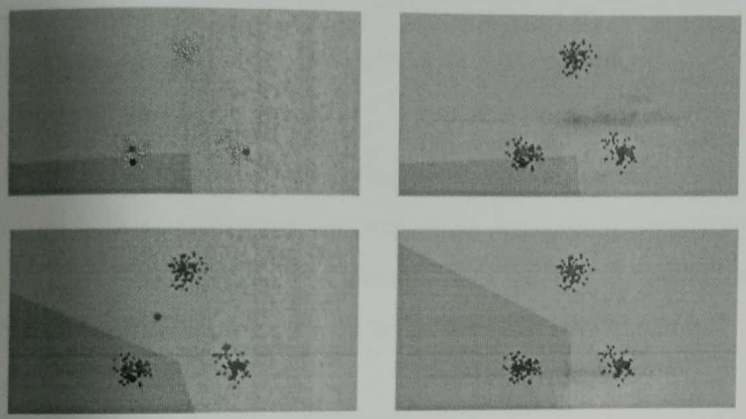

Fig:2.3 K-means clustering

2.5 LIGHT GBM

Pavan Kumar Mallapragada[2] proposed the Light GBM is a gradient boosting framework that uses tree-based learning algorithm. Light GBM grows tree vertically (leaf-wise) while other algorithm grows trees horizontally (level-wise). Light GBM can handle the large size of data and takes lower memory to run. Light GBM is sensitive to over fitting and can easily over fit small data. Light GBM is high speed and can handle large size of data and takes lower memory to run and it focuses on accuracy of results and also supports GPU learning. It has many parameters:

Control parameters
Core parameters
Metric parameters
IO parameters

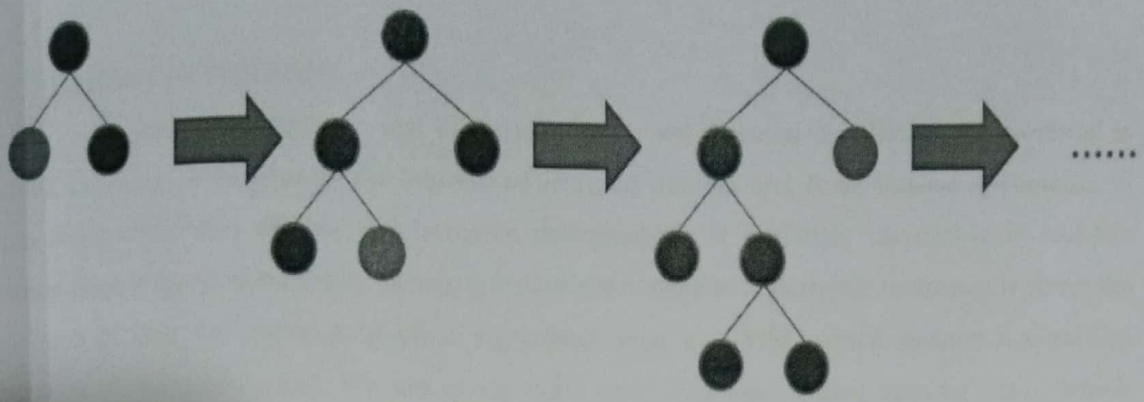

Fig:2.4 Light GBM Tree Growth

2.6 SUPPORT VECTOR MACHINE

Support vector machine[7], also called as support vector networks are supervised learning algorithm used for both classification and regression analysis. It classifies the data points plotted multi-dimension space into categories by parallel lines called the hyperplane. The classification of data points involves the maximization of margins between the hyperplane.

2.7 ADAPTIVE BOOSTING

Yoav Freund and Robert Schapire[9] is a machine learning algorithm used for classification and regression analysis. It happens to convert the weak classifier into a strong one using the ensemble technique. For this purpose, the prediction of each weak classifier is merged using weighted average or by taking into accounts their prediction accuracy as a metrics. Initially, all the attributes are given equal weights, and then the algorithm decides to assign a higher weightage to the inaccurate observation. After multiple iterations the prediction becomes accurate[10].

2.8 CONVOULTIONAL NEURAL NETWORK

In neural networks, Convolutional neural network also referred as Convent's or CNNs is one of the main categories to do images recognition, images classifications. Objects detections, recognition faces etc., are some of the areas where CNNs are widely used[14].

.9 K NEAREST NEIGHBORS

K-nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in attern recognition, data mining and intrusion detection[13]. It is widely disposable in real-life cenarios since it is non-parametric, meaning; it does not make any underlying assumptions about the istribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian istribution of the given data). We are given some prior data (also called training data), which lassifies coordinates into groups identified by an attribute.

.10 HANDLING BIOMEDICAL DATA

L Qiu, K. Gai and M. Qui in [11] stated how to handle data in large quantities in cloud. It is asically said that the data should be shared among all telehealth services over the cloud. The lisadvantages of sharing in cloud are network traffic, virtual machine switches. So, by sharing the lata it would be easy to access. Also, by focusing on transmission probabilities, network capabilities nd timing constraints.

As we can see above there are number of algorithms which shall be used for a wide spectrum f applications. These algorithms provide different results which have to conclude to design a better rediction model. Also, we saw the data handling and issues that can be seen in training the nachine. Later part we will be able to see the methodologies and results and conclusions of what hese algorithms are able to produce.

CHAPTER 3

METHODOLOGY

3.1 PROPOSED METHODOLOGY:

Initially we take disease dataset from UCI machine learning website and that is in the form of disease list with its symptoms. After that preprocessing is performed on that dataset for cleaning that is removing comma, punctuations and white places. And that is used as training dataset. After that feature extracted and selected. Then we classify that data using classification techniques such as Naïve Bayes, KNN, simple linear regression, etc. Based on machine learning we can predict accurate disease.

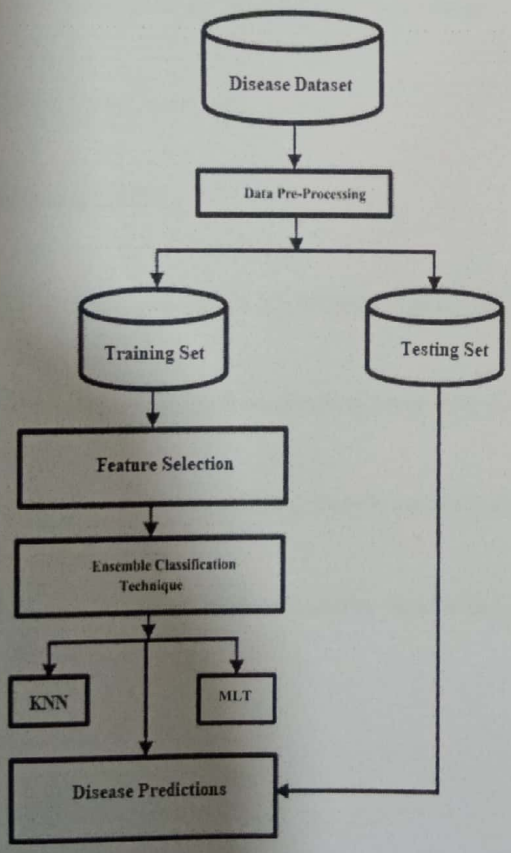

Fig:3.1 System architecture

As per the suggestions and knowledge gained from the article [12], we get system architecture (fig.no:2.5). So, we consider the two algorithms KNN and CNN.

3.2 ALGORITHMS USED:

1) K-Nearest neighbor (KNN)

Common Distance Metrics:

Euclidean distance (continuous distribution):

$d(a, b) = \sqrt{\Sigma(a_i - b_i)^2}$

Hamming distance (overlap metric):

bat (distance = 1)

toned (distance = 3) Discrete Metric (Boolean metric):

if x = y then d (x, y) = 0. Otherwise, d (x, y) = 1

Determine the class from *k* nearest neighbor list,

Take the majority vote of class labels among the k-nearest neighbors

Weighted factor:

w =1/d (generalized linear interpolation) or1/d2

2) Convolutional neural network (CNN)

Step 1: The dataset is converted into the vector form.

Step 2: Then word embedding carried out which adopt zero values to fill the data. The output of word embedding is convolutional layer.

Step 3: This Convolutional layer taken as input to pooling layer and we perform max pooling operation on convolutional layer.

Step 4: In Max pooling the dataset convert into fixed length vector form. Pooling layer is connected with the full connected neural network.

Step 5: The full connection layer connected to the classifier that is SoftMax classifier.

CHAPTER 4

RESULTS AND DISCUSSIONS

This section presents the performance of the KNN and CNN classification algorithms in terms of time required and accuracy. Fig 4.1 Shows accuracy Comparison of KNN and CNN algorithms for various Threshold. X-axis shows Algorithm & Y-axis shows accuracy in %. CNN gives more accurate disease prediction than KNN. Fig 4.2 shows the Time comparison of KNN and CNN algorithms for various size. The X-axis shows algorithms and Y- axis shows Time in ms. The CNN takes less time than KNN for classifying large dataset.

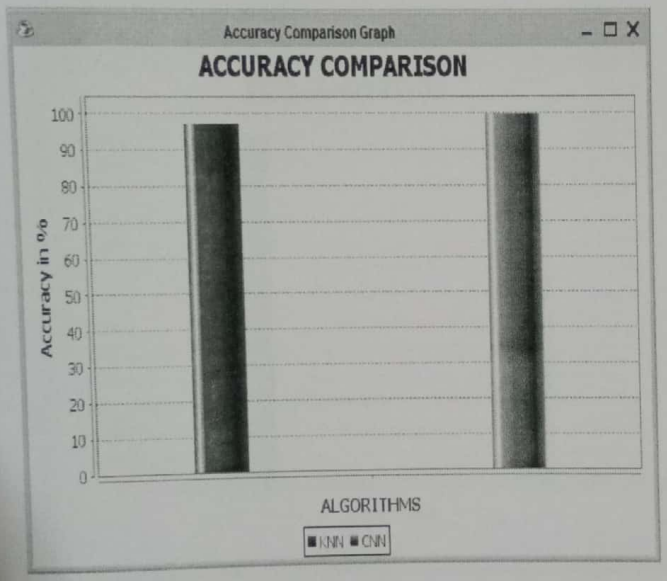

Fig:4.1 Accuracy Comparison Graph

We proposed general disease prediction system based on machine learning algorithm. We utilized KNN and CNN algorithms to classify patient data because today medical data growing very vastly and that needs to process existed data for predicting exact disease based on symptoms. We got accurate general disease risk prediction as output, by giving the input as patients record which help us to understand the level of disease risk prediction. Because of this system may leads in low time consumption and minimal cost possible for disease prediction and risk prediction. We compare the results between KNN and CNN algorithm in terms of accuracy and time and the accuracy of CNN algorithm which is more than KNN algorithm and time required for classification for CNN is less than KNN. So, we can say CNN is better than KNN in terms of accuracy and time.

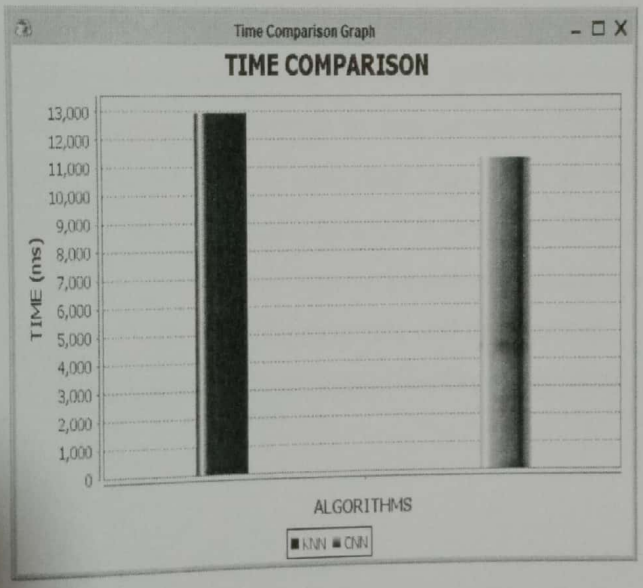

Fig:4.2 Time Comparison

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENTS

.1 CONCLUSION:

We can conclude that the algorithms we used were working effectively with the given amount of data but it has to be noted that they are many other algorithms which can have a better result. The algorithms we used on the datasets are KNN and CNN only. The datasets used are also optimized and made sure to be perfect without any redundant data. A final conclusion is same as discussed in the results that the CNN performed better than KNN.

.2 FUTURE ENCHANCEMENTS:

The future scope and improvement of the project involve automation of the steps such as data mining, feature selection and model fitting for best prediction accuracy. Use of pipeline structure for data pre-processing could further help in improvement. The methods of extraction of data and preprocessing it and clearing all the redundant data must be well optimized for better prediction of data. In future more algorithms can be used to understand the level of prediction and accuracy. Also, the datasets handling processes have to be automated and quality of data should be increased to have the best results in terms of effectiveness and optimization. The ONNX happens to provide the interchangeable AI models which can help in deploying the models for a wide variety of use in medical field.

APPENDICES

REFERENCES

[1] L. Jiang, C. Li, June, 2016, "Deep feature weighting for naive Bayes and its application to text classification", Journal of Engineering Applications of Artificial Intelligence, Vol 52, Pp: 26-39.

[2] Pavan Kumar Mallapragada, November, 2009, "Semi Boost: Boosting for Semi-Supervised Learning", Journal of IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol 31, Pp: 11.

[3] W. Zhao, H. Ma and Q. He, 2009, "Parallel k-Means Clustering Based on Map Reducer", Vol 5931, Springer Berlin / Heidelberg, Pp: 674-679.

[4] Enrico R, Michel L, July, 2016, "The Counter, a Frequency Counter Based on the Linear Regression", IEEE Transactions on Ultrasonic Ferroelectrics, Vol 63, Issue 7, Pp: 961-969.

[5] S. Patel and H. Patel, March, 2016, "Survey of data mining techniques used in healthcare domain", Institution Journal of Information Science and Technology, Vol. 6, Pp: 53-60,

[6] "UCI Machine Learning Repository" [ONLINE], Available: https://archive.ics.uci.edu/ml/index.php.

[7] V. Jakkula, 2006, "Tutorial on Support Vector Machine (SVM)", Sch.EECS, Washington State University, Pp. 1-13.

[8] "Geeks for Geeks online study material" [ONLINE], Available: https://www.geeksforgeeks.org/machine-learning/

[9] Y. Freund and R. Schapire, 2013, "A Tutorial on Boosting", Pp. 1-35.

[10] Pahulpreet Singh Kohli, Shriya Arora, 2018, "Application of Machine Learning in Disease Prediction", 4th International Conference on Computing Communication and Automation (ICCCA).

[11] L.Qiu, K. Gai and M. Qui, "Optimal big data sharing approach for telehealth in cloud computing", in Proc, IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, Pp: 184-189.

[12] Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram, 2019, "Designing Disease Prediction Model Using Machine Learning Approach", Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC).

[13] "Medium website", https://medium.com/@chiragsehra42/k-nearest-neighbors-explained-easily-c26706aa5c7f

[14] "Medium website", https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148