Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Prediction with ML paradigm in Healthcare System

Pradeep Jha
Assistant Professor, Department of CSE
Arya College of Engineering & Research Centre
Jaipur, India
pradeep.jha1988@gmail.com

Trisha Biswas
Department of CSE
Arya College of Engineering & Research Centre
Jaipur, India
trisha1812.2000@gmail.com

Utkarsha Sagar
Department of CSE
Arya College of Engineering & Research Centre
Jaipur, India
utkarshas123@gmail.com

Kiran Ahuja
Assistant Professor, Department of CSE
Arya College of Engineering & Research Centre
Jaipur, India
kiran.ahuja2011.ka@gmail.com

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532752

***Abstract*— Precise and synchronized research of any health-related issues is important for the eradication and curing of the illness. But the accurate prediction on the basis of symptoms becomes too difficult for doctors. A stronger proposal to medical-care is to eradicate a disease with initial interference instead of taking treatment after it is recognized. Medical-care produces an enormous quantity of data, with the assistance of disease data, ML finds invisible samples of details in this enormous dataset.**

The aim of this research is to enhance the precision of the prognosis. Enlarging a health treatment system centered on several ML algorithms for the prognosis of any disease can assist in an extra correct diagnosis. Our dataset consists of 4920 patient data identified with 41 diseases. An experimental target consists of 41 diseases. 95 of 132 manipulated features closely coupled to diseases were picked and improved. The result of this research relies on how correct the data set is. For prognosis prediction, this research work has integrated the Naive-Bayes, Decision Trees, Random Forest, KNN, SVM, Logistic Regression, and SGD. The accuracy of general disease prediction by using KNN is 97.32%, which is more than any other algorithms. Our diagnosis model can act as a doctor for the early diagnosis of a disease to ensure the treatment can take place on time and lives can be saved.

Keywords*— *Machine Learning, Naive-Bayes, Decision Trees, Random Forest, KNN, SVM, Logistic Regression, SGD, Disease Prediction.

I. INTRODUCTION

Health Disease: Medication and healthcare wellness are critical components of both economic growth and human development. Measles, tuberculosis, impetigo, diabetes, epilepsy, acid reflux, meningitis, and other chronic disease issues have a huge impact on one's nutrition and can even result in death if ignored[1].

The battle of COVID-19 will continue in 2021, with various countries around the world. In such scenarios where people need to take care of both being safe at home and medical assistance there needs to be a bridge that joins the patient with the healthcare system so that they can both be treated at home and kept safe at home without having the danger of getting affected with the virus.

Google designed an algorithm in 2017, with 89 percent accuracy, which detects illnesses such as cancer. Shortly afterward, by 2020, the main motto for smartphone-based mental health applications is automated behavioral solutions and machinery learning has just begun. ML aims to minimize the mistakes of humans without restricting the human factor in healthcare.

HDPS(Health Disease Prediction System) in ML: In the field of health care, machine learning is being used to combat disease. The prediction system helps enhance diagnosis, diagnose and reduce infections, crowdfund research, medicines, patient care improvements, etc. The safest idea is to keep the distance and to be at home in conditions like COVID-19 and EBOLA to end disease diffusion. [2].

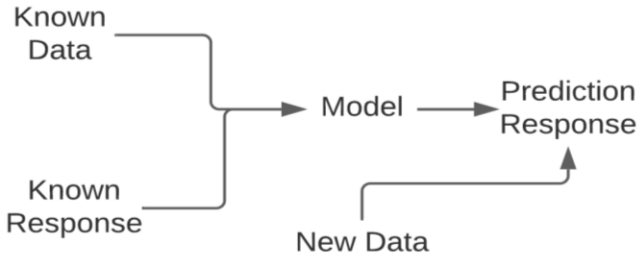
In the event of an emergency, the virtual doctors are certified health care professionals who do choose to practice digitally via multimedia and telephone consultations instead of on-the-go appointment scheduling. Models are notionally supercilious to mankind because, in the truancy of people's error, they can execute tasks more structurally and with a compatible target of precision. The illness predictive model is also known as a virtual doctor because it can anticipate the illness of just about any patient again with no human error.

II. OVERVIEW

A. Supervised Learning

It is the type of ML technique where the models are predicted through several "labelled" data to be trained,in keeping with that data, models predict the output. The categorized data is the way where some input data is until now identified with the accurate output. The motive of this algorithm is to discover a single valued function to join the dependent target(x) with the dependent target(y). In the

modern era, this algorithm can be used for Fraud Detection, Risk Assessment, spam filtering, etc [3].



B. Classification in ML

It is a technique of classifying a given cluster of data into classes, It can act on both structured or unstructured data[4].

C. Structured data classification

It is the process of categorizing data into a set variety of groups. Its ultimate objective is to identify the class/category within which fresh information might well fall.

D. Algorithms used for our research

1) ***Naive Bayes Classifier :*** This is a supervised classifier that is premised on Bayes' theorem, which further assumes predictor independence. In a simplified way, the existence of such a specific element in a group doesn't pertain to the existence of yet another character in a Naive Bayes classification.[5]It is the process of categorizing data into a set variety of groups. Its ultimate objective is to identify the class/category within which fresh information might well fall.

2) ***K-Nearest Neighbor****:* This is a facile linear classifier that collects all happenings in n-dimensional space relating to supervised learning. It focuses mostly on constructing a particular object state, but also on collecting training sample cases. Here 'k' is the number of neighbors it looks at[6].

3) ***Decision Tree:*** Concerning a tree structure, this algorithm builds prediction models. It uses the if-then regulations, which in categorization are consistently complete and frequently transferable. Deconstructing the statistics into small structures and connecting it to a progressive decision tree in a period is the methodology. The framework of the closure appears like something of a tree of leaves and nodes.

4) ***Random Forest:*** They are regression, classification, etc. categorized learning techniques. They address a variety of decision-making bodies at training and anticipate the group that is the category of the singular trees[7].

5) ***Logistic Regression:*** Here the output is a two-part variable which means that only 2 potential outcomes are obtained. Its objective is to forecast an optimal significance of the model and a cluster. The algorithms are sharper than other binaries as the nearest neighbor, seeing as they determine mainly categorization facets quantifiably.

6) ***Stochastic Gradient Descent***: The proposed legislation to integrate statistical parameters is immensely effective. Here When the test statistic is in a large amount. It involves the measurement and the derivatives from every example of training data.

7) ***Support Vector Machine:*** It is an extraction method that uses training data in the form of objects in time separated into classes. The gap in these classes is very large. To these points, the prediction of which class they come into and to which space they belong is then attached to space[8].

E. Advantages and Disadvantages of different Algorithms

1) ***Naïve Bayes Classifier:***
Advantage: Structured, not biased by outliers, works on non – linear problems, statistical approach[9].
Disadvantage: Based on the presupposition that the attributes have the same probabilistic pertinence[9].

2) ***K-Nearest Neighbor:***
Advantage: Easier to understand, secure and structured[9].
Disadvantage: Required to be hand-operated, pick the number of neighbours 'k'[9].

3) ***Decision Tree:***
Advantage: Accountable, not required for attributes spanning, applies on twain linear / non – linear problems[9].
Disadvantage*:* Deficient results on very low datasets, overfitting can simply occur[9].

4) ***Random Forest:***
Advantage: Strong and right, well performance on several problems, as well as non – linear[9].
Disadvantage*:* No accountability, overfitting can simply occur, required to pick the number of trees hand-operated[9].

5) ***Logistic Regression:***
Advantage: Statistical Approach, gives information about probabilistic importance of attributes[9].
Disadvantage: The assumptions of logistic regression[9].

6) ***Stochastic Gradient Descent:***
Advantage: Well structured and simplicity of implementation[9].
Disadvantage: Needed features of hyper-parameters and it is delicate to attribute scaling[9].

7) ***Support Vector Machine:***
Advantage: Effective, not biased by outliers, not delicate to overfitting[9].

Disadvantage: Not suitable for non-linear problems, not the better choice for a enhanced number of attributes[9].

III. PROPOSED SYSTEM

A. Workflow of Proposed System

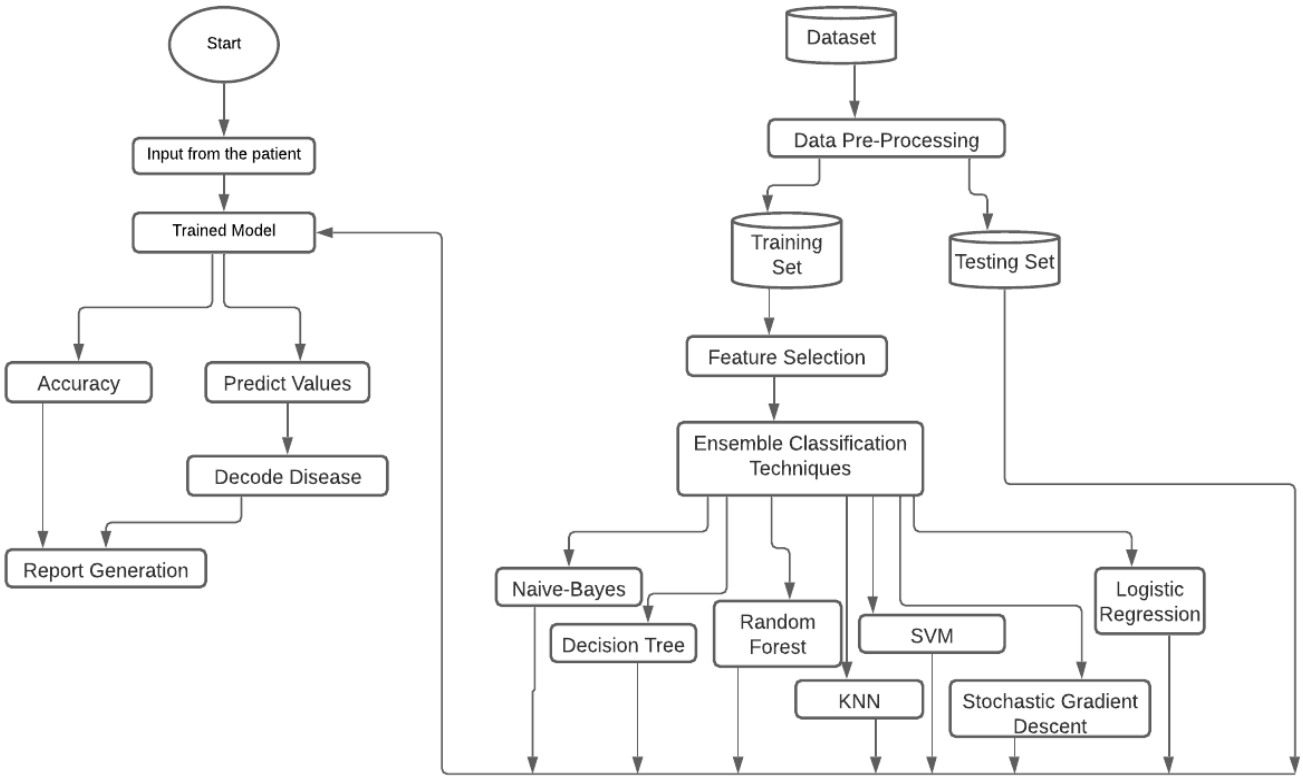


Fig 1: FLOWCHART OF PREDICTION MODEL

1) ***Input (user data):*** When interacting with the system for the first time, we assume that the user/ patient has a list of symptoms which he/ she is experiencing.

2) ***Data Preprocessing :***

- ***Collecting data :*** The data must be recorded in any format. This data will be used as training material by an algorithm to develop meaningful intelligence [10].
- For instance, we have a record of user-profiles and we now have to segregate them by incident and time stamp types[11].

  Data Cleansing or Data Munging: the data is cleaned by methods such as the supply of missing value so that discrepancies in the data can be resolved. This includes incomplete information like ignoring the tuples, fulfilling the missing values, noisy information includes binning, regression, and clustering[12].
- Filtering data through determined settings for the processing.
- ***Data Metamorphosis -*** This action is followed to process the information for the mining process in appropriate forms. That means standardization,

  selection of attributes, deconvolution, generation of concept hierarchy[13].

  The dataset we examined comprises 132 health problems, of which 41 diseases occur in conjunction or permutations [14]. Depending on 4,920 medical information, we sought to create a model for prediction which addresses the user's symptoms and predicts the most likely disease.

3) ***Feature Selection :*** This is essential for all statistical modeling and it is done to ensure multiple correlations and to remove any redundant and irrelevant features which are strongly associated with one another such that the model performs better.This is where we start with all the models' attributes, accompanied by a p-value elimination[15].

*4) **Model fitting and Testing :*** Only after the selection of features, SVM was employed, with the selected characteristic selection 7 classification algorithms including naive basins, Decisive tree, logistic regression, KNN, random forest, stochastic gradient descent, and the exact prediction was checked using the train/test split technique[16]

*5) **Training model :***For training the model, the dataset has been taken from Kaggle where the data itself was collected from various hospitals. After data preprocessing (removing redundancy, removing bad data), the data will be divided into two parts. The divided data will be saved as "training data" and "testing data". The ratio of training data to testing data is generally taken as 80:20 out of whole datasets. If the result of training data and testing data matches, then it proves that the model has been trained properly.

*6) **Value Prediction:*** If the data matches, the value of "1" is awarded, else "0" is the awarded value for a mismatch to the result. After proper prediction of the values, a disease is fetched as the result.

*7) **Decode disease:*** The disease will be a proper output with the majority of matches after having proper machine model training. The resulting disease will be provided as an output having perfect matches after several times getting trained by the model.

*8) **Report generation:*** The report generated may contain summary statistics like dashboards and analyses of treatment types, employees, including surgeons, nurses, administrators, etc.

Novelty Work: The novelty of this examination work is fostering a finding framework for numerous Health Diseases. In our idiosyncrasy, the framework utilized seven managed arrangement calculations for different Health Diseases. The framework was taken a stab at the Disease Prediction dataset observed from Kaggle to bunch various Health Diseases. Additionally, some unessential highlights reduced the demonstration of the affirmation structure and broadened the calculation time.

IV. RESULTS

A) Analysis on major machine learning algorithms

TABLE III. ANALYSIS OF ML ALGO.

Machine Learning Algorithms	Comparison of Features
Naive Bayes	**Outlier-** It is less pruned to outlier
	Online Learning- It can perform on online testing
	Overfitting and underfitting - It does not suffer underfitting and overfitting.
Decision Tree	**Outlier-** Outliers do not take part in a crucial play in integration of dataset.
	Online Learning- It does not support online testing.
	Overfitting and underfitting - It endures underfitting and overfitting.
Random Forest	**Outlier-** It handles outliers by essentially binning them.
	Online Learning- It does support online testing.
	Overfitting and underfitting - It is less prone to underfitting and overfitting.
KNN	**Outlier-**Here, the outliers are very sensitive ,because an individual wrongly labeled example drastically changes and interchanges the target boundaries.
	Online Learning- It supports online testing.
	Overfitting and underfitting - If the value of K is balanced then it does not suffer from underfitting and overfitting.
Logistic Regression	**Outlier-** Requires more assumptions than a tree and it is sensitive to outliers.
	Online Learning- It supports online testing.
	Overfitting and underfitting - It does suffer from underfitting and overfitting.
SVM	**Outlier-** It can handle outliers properly.
	Online Learning- Online training requires less time.
	Overfitting and underfitting - Perform better than underfitting and overfitting.
Stochastic Gradient Descent	**Outlier -** Highly prone to outliers.
	Online Learning - Online training requires more time.

	Overfitting and underfitting - Never Overfits and underfits on separable data.

B) ***Number of samples on Training data***

The classifier was tested using the medical records of 4920 patients who were predisposed to 41 diseases, which included the integration of multiple side effects. To mitigate underfitting and overfitting, our model selects 93 health problems from a total of 135.

TABLE IV. ACCURACY ON TRAINING DATA

Algorithm used	**Accuracy Score**
Naive - Bayes	87.43%
Random Forest	85.67%
Decision Tree	93.67%
K-Nearest Neighbour	95.54%
Logistic Regression	53.87%
Support Vector Machine	45.54%
Stochastic Gradient Descent	32.67%

C) Number of samples on Testing data

Once the training has completed, the model tried 41 newly patient records assuming symptoms of 95 people. The first table shows accuracy score and the second table shows confusion matrix is given by:

TABLE V. ACCURACY ON TESTING DATA

Algorithm Used	**Accuracy Score**
Naive - Bayes	90.43%
Random Forest	87.43%
Decision Tree	95.67%
K-Nearest Neighbor	97.32%
Logistic Regression	60.87%
Support Vector Machine	50.54%
Stochastic Gradient Descent	40.67%

TABLE VI. CONFUSION MATRIX OF TESTING DATA

Algorithm Used	**Confusion Matrix**	
	Correct Classifiers	**Incorrect Classifiers**
Naive - Bayes	38	3
Random Forest	35	6
Decision Tree	39	2
K-Nearest Neighbor	40	1
Logistic Regression	30	11
Support Vector Machine	21	20
Stochastic Gradient Descent	17	24

D) ***Accuracy Graph***

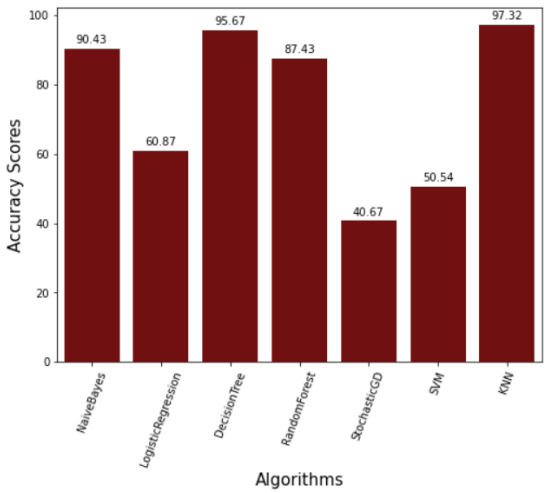


Fig 2. ACCURACY COMPARISON GRAPH

The above graph shows the comparison for different algorithms which can be used to predict disease, namely Naive Bayes, Logistic Regression, Decision tree, Random Forest, Stochastic Gradient Descent,Support Vector Machine, and

KNN. It is observed that the model accuracy is highest for the model which is designed using KNN and then Decision tree which has little less accuracy than that of KNN, and then Naive Bayes which has less accuracy than that of Decision Tree, and then Random Forest which has little less accuracy than that of Naive-bayes, and then Logistic regression which has little less accuracy than that of Random Forest, and then SVM which has little less accuracy than that of Logistic regression , and then comes Stochastic Gradient Descent with least accuracy.

TABLE VII. COMPARISON OF PREDICTED DISEASE AND ACTUAL DISEASE

Symptoms	**Accuracy Sore**	**Disease**
drying_and_tingling_lips, fast_heart_rate, chills , dizziness,dehydration,chest_pain	**Naïve Bayes -** 90%	**Predicted Disease-** Heart Attack
	Decision Tree- 95.67%	
	Random Forest- 24%	
	Logistic Regression- 31%	**Actual Disease-** Heart Attack
	SGD- 2%	
	SVM- 19%	
	KNN- 96.78%	
abdominal_pain, abnormal_menstruation, continuous_feel_of_urine,constipation	**Naïve Bayes -** 100%	**Predicted Disease-** Urinary_tract_infection
	Decision Tree- 100%	
	Random Forest- 28%	
	Logistic Regression- 51%	**Actual Disease- -** Urinary_tract_infection
	SGD- 10%	
	SVM- 17%	
	KNN- 100%	
continuous_sneezing,headache,cough,breathlessness	**Naïve Bayes -** 90%	**Predicted Disease-** Bronchial Asthma
	Decision Tree- 95.67%	
	Random Forest- 24%	
	Logistic Regression- 31%	**Actual Disease-** Bronchial Asthma
	SGD- 2%	
	SVM- 19%	
	KNN- 96.78%	
weakness_of_one_body_side,yellow_urine,yellowing_of_eyes	**Naïve Bayes -** 94.54%	**Predicted Disease-** Paralysis (brain hemorrhage)
	Decision Tree- 93.54%	
	Random Forest- 24%	
	Logistic Regression- 31%	**Actual Disease-** Paralysis (brain hemorrhage)
	SGD- 2%	
	SVM- 19%	
	KNN- 95.65%	
increased_appetite,history_of_alcohol_consumption,dark_urine	**Naïve Bayes -** 91.45%	**Predicted Disease-** Alcoholic hepatitis
	Decision Tree- 95.67%	
	Random Forest- 24%	
	Logistic Regression- 31%	**Actual Disease-** Alcoholic hepatitis
	SGD- 2%	
	SVM- 19%	
	KNN- 96.78%	

The table above shows the mixture of health conditions and, based on everyone's different ML algorithms illnesses, gives the precision that an illness is decoded based on this accuracy. And we compare this predicted disease with the actual disease which we took from the results of Google.

E) ***GUI Results:***

1) Interface

- Assign under qualified customers to be evaluated by machine learning.
- Work closely to avail the superior part of the information which is user-friendly and competent through machine learning.
- An effective process as it uses the contact to the customers.

2) Tools & Framework used: The web applications are demonstrated here through highlight driven turn of events. The web application has been created utilizing Django Framework of python. It uses the MVT(Model View Template) engineering.

Every action of highlight driven advancement is examined mind ancient rarities delivered during that action.

Fig 3. RESULTING GUI OF MODEL

The GUI developed takes in 5 symptoms from the user, as seen in the above figure. When the user clicks on the "Add the Symptoms" option, a list of symptoms appears, from which the user can select the symptoms. The visitor may be able to experience any set of symptoms.

After the side effects have been provided, the algorithms must be chosen. The symptoms have been analyzed as the algorithms are chosen, as well as the illness is scanned using the rule set characterized in the dataset segment.

Following are the symptoms entered by the patient named Rahul : "blood_in_sputum", "bloody_stool", "runny nose", "unsteadiness" , "foul_smell_of_urine".

Predictions by algorithms were:

- Naive-Bayes- Urinary Tract Infection
- Decision Tree- Urinary Tract Infection
- Random forest- Urinary Tract Infection
- KNN- Urinary Tract Infection
- SGD - Chronic cholestasis

- SVM - Tuberculosis
- Logistic Regression -Chronic cholestasis

As a result, the physician can rely on the large proportion of the results, indicating that the patient is much more probable to have **Urinary Tract Infection.**

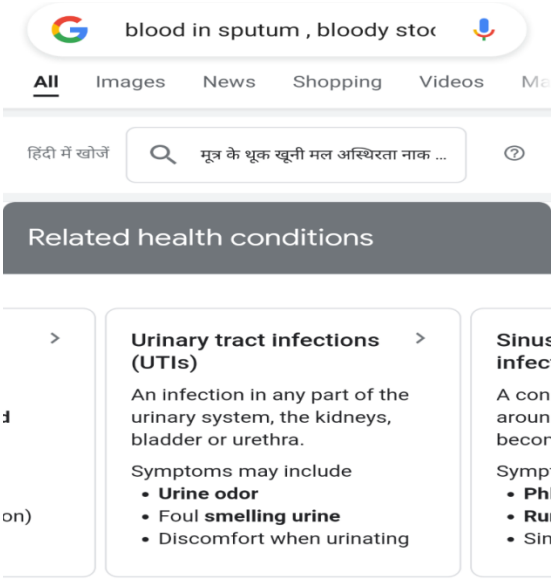


Fig 4. CROSS VERIFICATION OF RESULTED OUTPUT

The above image is of the google search engine, where we add symptoms "blood_in_sputum", "bloody_stool", "runny nose", "unsteadiness", "foul_smell_of_urine". After cross verifying from the internet, we got the disease as Urinary Tract Infection, which was same as the disease predicted by the model have been given as we have given in our GUI and through which we have been found that disease predicted is also Urinary Tract Infection (URIs) which is the same as disease predicted by our GUI.

Hence, we can say that our GUI decodes the accurate disease.

V. CONCLUSION

Medical care produces an enormous quantity of information bit by bit and can be brought into efficiently for prognosis predictions.KNN predicted best accuracy results in diagnosis.KNN also predicts other diseases flawlessly. It is less difficult, and performance is analogously faster. It works extremely well on complicated datasets.

The pros of our proposed system is that specialists depend on broad information for treatment. Any clinical experts can anticipate the infections with up to 67% precision whereas our proposed system gives 97.32% precision. By utilizing ML for clinical findings, this reenactment can possibly extraordinarily lessen the number of clinical mistakes and misdiagnoses. As here calculation does the undertaking consequently an all-around prepared model less will undoubtedly make blunders in anticipating sicknesses, in short exactness is improved what's more, along these lines it likewise saves time.

The outcome of this research ensures the use case of machine learning algorithms in diagnosing and prior observation of prognosis. From our foremost perception, the model prepared in accordance with the suggested way manifests good accuracy as compared to the present ones[14,16]. The research shows an extensive modification research of seven algorithms, and their performance on a health record and every algorithm capitulate an accuracy upto 97 percent. The prediction is analyzed with the help of accuracy score and a confusion matrix.

VI. FUTURE WORK

The limitations of our research work is that there's wariness among individuals in confiding in machines to settle on life-passing choices. An absence of trust in machines, focusing on people over insights are a portion of the social issues that may emerge from this recreation. Changing this dynamic interaction from a specialist to a calculation can prompt cases of disappointment because of ill-advised treatment because of some unacceptable calculation. Accordingly, this task ought to just be considered as an assistant device and ought to be completely confided in solely after future approval by incorporating more viable information.

The future area of work or the refinement of the research includes computerization or the robotization like Artificial Intelligence. Pipeline structure for data preprocessing could also helping in attaining better results. We will also propose to extend our algorithm to incorporate unstructured data as well. A new research on disease prediction with the help of deep learning is arising in coming days, because deep learning has the ability to work much better than the conventional algorithms.

REFERENCES

[1] G. K. Soni, A. Rawat, S. Jain and S. K. Sharma, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm with LSB Watermark Technique", Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, vol 141, pp. 483-492, 2020.

[2] Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang and Ninad Mehendale,"Disease prediction from various symptoms using machine learning", Social Science Research Network (SSRN), Vol.-, Issue-, Pp-, 2020.

[3] Rabah Kamal, et. al "How has U.S. spending on healthcare changed over time?", Peterson- KFF, 2020.

[4] B. Nithya et. al, "Predictive Analytics in HealthCare Using Machine Learning Tools and Techniques", Intelligent Computing, Information and Control Systems (ICICCS), 2017.

[5] Sathya Madhusudhanan ,Suresh Jagannathan and Jayashree LS, "Incremental Learning for Classification of Unstructured Data Using

ExtremeLearning Machine", Algorithms, Vol.-11, Issue-10, Pp-158, 2018.

[6] R. Lee and C. Chitnis, "Improving Health-Care Systems by Disease Prediction," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Pp- 726-731, 2018.

[7] Pooja Gupta, Pritesh Jain and Upendra Singh,"Review of Heart Disease Prediction using Supervise and Unsupervised Machine Learning Technique", International Journal of Scientific Development and Research (IJSDR), Vol.-3, Issue-6, Pp-218-224, 2018.

[8] Sotiris Kotsiantis, I. D. Zaharakis and P. E. Pintelas, " Supervised Machine Learning: A Review of Classification & Combining Techniques ", Artificial Intelligence Review (Artif Intell Rev), Vol.-26, Issue-3, Pp-159-190, 2007.

[9] https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-classification-models/.

[10] Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data -AI Integration Perspective", IEEE Transactions on Knowledge and Data Engineering (TKD), Vol.-33, Issue-4, Pp-1328-1347, 2021.

[11] Brett Lantz, "Machine Learning with R", Vol. 1, January 2013.

[12] Sapna Devi, Dr. Arvind Kalia, "Study of Data Cleaning & Comparison of Data Cleaning Tools ", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.- 4, Issue-3, Pp.-360– 370, 2015.

[13] Kaur, Rapinder, Vaishali Chauhan and Urvashi Mittal, "Metamorphosis of data (small to big) and the comparative study of techniques (HADOOP, HIVE and PIG) to handle big data ", International journal of engineering and technology (IJET), Vol.-7, Pp-2.27, 2018.

[14] S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), Pp. 1-7, 2020.

[15] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Pp. 1-4, 2018.

[16] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Pp. 1211-1215, 2019.