iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46166

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII August 2022- Available at www.ijraset.com

Chronic Disease Prediction Using Machine Learning

Vijay Kumar Korke[1], Vaibhav Chowdhary[2], Sagar M Keri[3], Miss Pallavi Patil[4], Dr. Suvarna Nandyal[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi, India.*

Abstract: ***Technological advancement, including machine learning, has a significant impact on health by allowing for more accurate diagnosis and treatment of various chronic diseases. Accurate prediction is critical in the biomedical and healthcare communities for determining the risk of disease in patients. The only way to overcome chronic disease mortality is to predict it earlier so that disease prevention can be implemented. Such a model is a Patient's requirement for which Machine Learning is highly recommended. However, a doctor finds it difficult to make an exact forecast based just on symptoms. The most challenging task is making an accurate diagnosis of a disease. Data mining is crucial in helping to predict the sickness and solve this issue. Based on a dataset for chronic diseases from the UCI machine learning data warehouse, this study assesses chronic diseases using machine learning techniques. In order to create accurate prediction models for various chronic diseases using data mining approaches, we employ datasets for heart disease, kidney disease, cancer disease, and diabetes disease. To increase accuracy and shorten training time, the dataset's most pertinent features are chosen. The system evaluates the user's symptoms as input and outputs the likelihood that the disease will occur. The implementation of Logistic Regression is used to predict disease. Prediction of diseases like diabetes, heart disease, cancer, and kidney disease using logistic regression, random forest, and decision trees are performed. Different models, methodologies, and algorithms are utilized to forecast and analyses each chronic disease. The study includes a conceptual model that includes the prediction of the majority of chronic diseases.***
Keywords: Disease Prediction and Accuracy, Logistic Regression, Chronic Diseases, Machine Learning

I. INTRODUCTION

Machine learning is the process of programming computers to optimise their performance based on previous data or examples. The study of computer systems that learn from data and experience is known as machine learning. ML can be supervised (i.e., output variables are predicted from input variables) or unsupervised (i.e., output variables are not predicted from input variables) (i.e., deals with clustering of different groups for a particular intercession). Complex models are determined using machine learning, and medical information is extracted using ML, revealing innovative ideas to professionals and specialists. In clinical practise, machine learning predictive models can be used to highlight stronger rules when making decisions about individual patient treatment. These are also capable of making independent diagnoses of many diseases based on clinical guidelines. Incorporating these models into medicine prescriptions can save doctors time and money while also providing new medical prospects for identification. Machine learning has been demonstrated to be useful in assisting decision-making and forecasting from enormous amounts of data generated by the healthcare business. We optimise machine learning methods for accurate chronic illness outbreak prediction. Various research only provides a sliver of what can be done with machine learning to forecast disease. We present a unique strategy that uses machine learning techniques such as the K-Nearest Neighbour Algorithm (KNN), Decision Trees (DT), Logistic Regression, Random Forest, and Naive Bayes (NB) to uncover meaningful characteristics, resulting in improved disease prediction accuracy. To improve the accuracy of the learning process, several such algorithms are used. It can then be put to the test using the datasets that are available. The prediction model is introduced using a variety of feature combinations and well-known classification approaches.

II. LITERATURE SURVEY

Arthur Samuel created the term "machine learning" in 1959. "Machine learning is the study of computer algorithms that allow computer programmes to automatically improve via experience," says Tom Mitchell. It's a collection of linkages and correlations. The majority of existing machine learning algorithms are concerned with discovering and/or exploiting relationships between datasets. When Machine Learning Algorithms are able to detect specific connections, the model can either use these links to forecast future observations or generalize the data to highlight intriguing patterns. Linear Regression, Logistic Regression, Naive Bayes Classifier, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, SVM (Support Vector Machines), K-means Algorithm, Random Forest, and others are examples of algorithms.

Machine learning is the research and development of algorithms that can learn from and predict data. It is closely connected to (and frequently overlaps with) computational statistics, which focuses on making predictions using computers as well. It has strong linkages to mathematical optimization, which provides the discipline with methods, theory, and application domains. Unsupervised learning is a subset of machine learning that focuses on exploratory data analysis and is commonly confused with data mining.
Typically, machine learning tasks are divided into numerous categories:

1) *Supervised learning:* The machine learning task of learning a function that translates an input to an output based on example input-output pairs is known as supervised learning. It infers a function from a set of training examples and labelled training data. Each example in supervised learning is made up of an input object (usually a vector) and a desired output value (also called the supervisory signal).
2) *Unsupervised learning:* Unsupervised learning is a sort of machine learning that searches a data set for previously unnoticed patterns with no pre-existing labels and minimal human observation. Unsupervised learning, also known as self-organization, allows for the modelling of probability densities across inputs, comparable to supervised learning, which often uses human-labeled data.

III. THE PRESENT PROCESS

Chronic diseases are becoming one of the leading causes of death around the world. A growing percentage of the world's population is suffering from the negative health repercussions of living. In general, doctors must thoroughly examine the patient's reports in order to make a disease diagnosis. It can be difficult for clinicians to treat patients efficiently when the diagnosis is manual.
The number of persons affected by chronic diseases is steadily increasing. The traditional health-care system is a passive one. Patients may die as a result of a lack of effective treatment during crises such as cardiac arrest related to this type. The key to increasing health-care efficiency is to lower mortality rates due to a lack of effective treatment and to turn a passive health-care programme into a continuous, low-cost one.

IV. PROPOSED SYSTEM

Due to the slow progression of Chronic Diseases, it is critical to make an early diagnosis and administer efficient treatment. As a result, it's critical to develop a decision model that may aid in the diagnosis of chronic diseases and the prediction of future patient outcomes.
While there are other approaches to this in the field of AI, the current research focuses specifically on machine learning predictive models used in the diagnosis of Chronic Diseases.
Our project's major goal is to make hospital activities easier and to produce an effective and practical software that will replace the manual prediction system with an automated healthcare management system.
This initiative helps healthcare practitioners increase operational efficiency, decrease medical errors, and save time. If an illness can be predicted, patients can receive early treatment, which reduces the risk of death and saves lives. Early detection can also help to lower the expense of disease treatment.
The diagnosis is based on a variety of Machine Learning Classification Models, which includes

1) KNN algorithm,
2) Naive Bayes Classification, and
3) Logistic Regression

V. SYSTEM DESIGN

A. *Design Objectives*

The design goals are a collection of several designs that we've used in our "Chronic Disease Prediction Using Machine Learning" system. Data flow diagrams, sequence diagrams, class diagrams, use case diagrams, and activity diagrams are all used to construct this system. Our system is set up in such a way that the registration procedure is handled completely by the administrator. Users, such as doctors, can log into the system using their credentials after completing the registration process. Doctors will be able to forecast chronic disease based on the inputs/attributes provided.

B. *Architecture of the System*

An architecture diagram is a graphical representation of a group of architectural concepts, such as principles, elements, and components. The diagram depicts the system software in the context of a system overview.

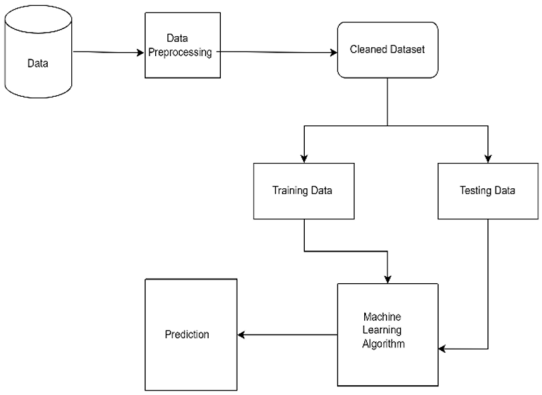


Figure 1. System Architecture

C. Activity Diagram

Figure 2 depicts the activity diagram. It denotes the sequence in which a system task is completed in order to produce a result. The Administrator is in charge of the User/Doctor registration process. After completing the registration process, the user, in this case a doctor, will log into the system using the credentials provided by the administrator. When a user logs in successfully, the system directs him to the appropriate page based on his specialism. The user must enter the qualities (independent variables) in the appropriate order to obtain the desired forecast. To generate the appropriate predictions and visualization, the system uses a Machine Learning Model that is built using accessible datasets and several ML methods (classification algorithms).

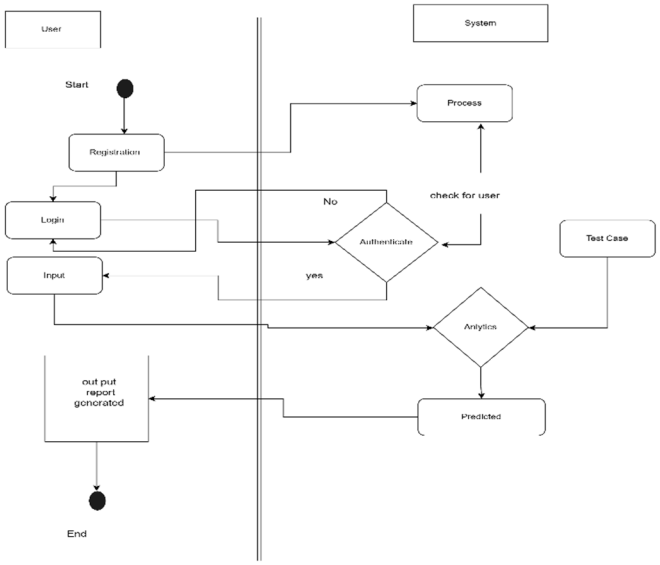


Figure 2. System Architecture

VI. ALGORITHM

A. KNN

K Nearest Neighbor (KNN) is a basic, easy-to-understand, adaptable, and one of the most advanced machine learning algorithms. The user will be able to predict the disease in the Healthcare System. The user can forecast whether or not an illness will be detected using this approach. The proposed method divides diseases into distinct classifications, indicating which disease will occur based on symptoms. For each classification and regression problem, the KNN rule was utilised. Based on a feature comparison technique, the KNN algorithm was developed. A case is classified by a majority vote of its neighbours, with the case being allocated to the most common class among its K closest neighbours as determined by a distance function. If K is equal to 1, the instance is simply placed in the category of its closest neighbour.

$$\text{Euclidean Distance}=\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

It's also worth noting that each of the three distance measurements is only valid for continuous variables. The Hamming distance must be employed when categorical variables are present. It all adds up to the difficulty of standardizing numerical variables between zero and one when the dataset contains both numerical and category variables.

$$\text{Hamming Distance} = \sum_{i=1}^{k} |x_i - y_i|$$

B. Naïve Bayes

For prognosticative modelling, Naive Bayes is a simple yet incredibly powerful rule. One of the easiest methods is to choose the most likely hypothesis based on the facts we have, which we may utilize as past information about the subject.

The Bayes' Theorem explains how we can determine the likelihood of a hypothesis based on the information we already have. The presence of a certain feature in an extremely class is unrelated to the presence of the other feature, according to the Naive Bayes classifier. The Bayes theorem allows you to calculate the posterior probability P (b|a) from P (b), P (a), and P (a|b). Take a look at the following equation:

$$\text{P (bVa)} = \frac{P(aVb)\,P(b)}{P(a)}$$

Above all,

1) The posterior chance of class (b, target) given predictor is P (b|a) (a, attributes).
2) P (b) is the class prior probability.
3) P (a|c) denotes the probability of a predictor in a particular class.
4) P (a) denotes the predictor's prior probability.

C. Logistic Regression

Logistic regression is a supervised learning classification technique used to predict the likelihood of a disease target variable. Because the nature of the target or variable is separated, there are only two possible groups. In simple terms, the variable is binary in nature, with information represented as either 1 (meaning success) or 0 (meaning failure). A logistic regression model predicts P(y=1) as a function of x.

Logistic regression can be expressed as:

$$\log(p(X)/(1 - p(X)) = \beta_0 + \beta_1 X$$

Where the logiest or log odds function is on the left, and p(x) / (1-p(x)) is on the right. The odds are the ratio of the chances of success to the chances of failure. As a result, in logistic regression, a linear combination of inputs is translated to the log (odds), with the output sufficient to 1.

VII. DISCUSSION AND RESULTS

The metrics listed below provide insight into the quality of the results obtained in this investigation.

Precision, also known as positive predictive value, is the proportion of patients who have chronic diseases who are also predicted to have chronic diseases (true positive and false positive).

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall, also known as sensitivity, is the ratio of the number of patients with chronic diseases who are accurately recognized to the total number of chronic disease patients.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-Measure: It assesses the test's precision. It's the harmonic mean of memory and precision.

$$\text{F - Measure}=2 * \frac{Recall*Precision}{Recall+Precision}$$

The ratio of accurately anticipated output cases to all cases in the data collection is called accuracy.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

Table I: Accuracy results for correctly classified and wrongly classified occurrences.

Disease	Accuracy	Correctly Classified Instances	Incorrectly Classified Instances
Cancer	80.8246	107	23
Heart	98.8243	565	1
Diabetes	79.9272	207	54
Kidney	76.2311	307	92

Table II: Precision, Recall, and F – Measurement Results.

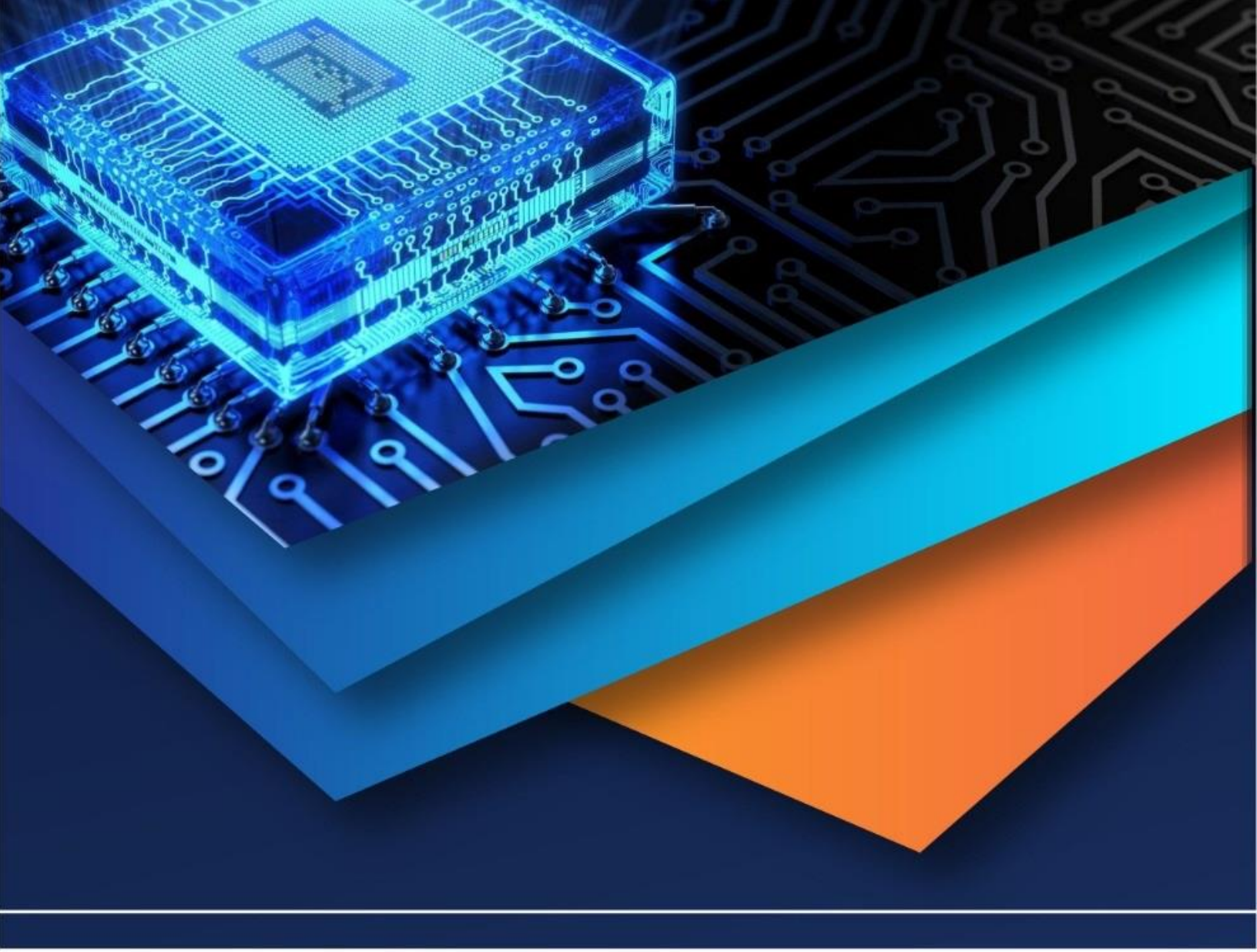
Disease	Precision	Recall	F - Measure
Cancer	1.000	0.332	0.500
Heart	1.000	0.994	0.997
Diabetes	0.811	0.898	0.852
Kidney	0.856	0.665	0.748

VIII. CONCLUSION AND IMPLICATIONS FOR THE FUTURE

The healthcare industry has benefited greatly from machine learning. The challenging and life-critical tasks such as chronic disease diagnosis are made easy and reliable with the help of Machine Learning. It has resulted in significant improvements to hospital, clinic, and laboratory practices. Doctors can forecast a patient's future condition by studying historical and real-time data. Various datasets for heart, kidney, cancer, and diabetes disorders have been used to test our technique. The study's major goal was to predict chronic disease utilizing features while retaining a high level of accuracy (here we obtain an accuracy of about 90 percent). In addition, our algorithm creates a report that includes the likelihood of disease occurrence. The outcomes indicate the robustness of the proposed method. For better chronic disease prediction, future research should examine several supervised and unsupervised machine learning techniques with additional performance indicators.

REFERENCES

[1] Hamet P., Tremblay J. Artificial intelligence in medicine. Metabolism.2017;69: S36S40.doi:10.1016/j.metabol.2017.01.011. [PubMed] [CrossRef] [Google Scholar].
[2] Johnson K.W., Soto J.T., Glicksberg B.S., Shameer K., Miotto R., Ali M., Dudley J.T. Artificial intelligence in cardiology. J. Am. Coll. Cardiol.2018;71:2668–2679.doi:10.1016/j.jacc.2018.03.521. [PubMed] [CrossRef] [Google Scholar].
[3] Bini S. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J. Arthroplast. 2018; 33:2358–2361. doi: 10.1016/j.arth.2018.02.067. [PubMed] [CrossRef] [Google Scholar].
[4] Kotsiantis S.B., Zaharakis I., Pintelas P. Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 2007; 160:3–24. [Google Scholar].
[5] Deo R.C. Machine Learning in Medicine. Circulation. 2015; 132:1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593. [PMC free article] [PubMed] [CrossRef] [Google Scholar].
[6] Battineni G., Sagaro G.G., Nalini C., Amenta F., Tayebati S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Machines. 2019; 7:74. doi: 10.3390/machines7040074. [CrossRef] [Google Scholar].
[7] Polat H., Mehr H.D., Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. J. Med. Syst. 2017; 41:55. doi: 10.1007/s10916-017-0703-x. [PubMed] [CrossRef] [Google Scholar].



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY