An Efficient Techniques For Disease Prediction From Medical DataUsing Data Mining And Machine Learning

Dr. Bharat Bhushan Agrawal
Associate Professor
Department of Computer Science & Engineering
IFTM University
Moradabad, Uttar Pradesh
bharat_agarwal@iftmuniversity.ac.in

Husam H. Abdulmughni
Department of Information Technology
Faculty of computer and IT
Sana'a University
Sana'a, Yemen
amideast.hossam2016@gmail.com

Ali A. Al-Bakhrani
Department of Computer Science
Technique Leaders College
Sana'a, Yemen
Faculty of Administrative and Computer Sciences
Albaydha University
Yemen
albakhrani2017@gmail.com

Pranoti Durgadas Nage
PhD Scholar, Usha Mittal Institute of Technology, SNDT Women's University, Mumbai Juhu Campus, Santacruz (W) Mumbai
pranotinage.phd2019@umit.sndt.ac.in

Sushma Jaiswal
Assistant Professor
Department of Computer Science & Information Technology (CSIT),
Guru Ghasidas Vishwavidyalaya (A Central University)
Bilaspur (C.G.) India
jaiswal1302@gmail.com

Dr. Vikas Tripathi
Associate Professor
Department of Computer Science & Engineering
Graphic Era Deemed to be University
Dehradun, Uttarakhand, India,
vikastripathi.cse@geu.ac.in

2022 5th International Conference on Contemporary Computing and Informatics (IC3I) | 979-8-3503-9826-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/IC3I56241.2022.10072906

***Abstract*—Individuals are susceptible to numerous ailments in today's global setting, with people living highly automated lives under tremendous job pressure, both at home and at work. Such disorders have recently been on the rise at an alarming pace. As a result, the healthcare business must assume a prominent position soon, accountable for people's health, a better society, and a successful country at large. Healthcare prices are rising as the demand for health amenities grows. Healthcare facilities with improved detection, diagnostics, and treatment procedures are desperately needed. With growing digitization and computing techniques in place, a massive amount of data is being generated and used for diagnostic and detection approaches. This data might be used to obtain data for anticipating disease, commencing preventative measures, and improving treatment procedures long before the diseases progress. Powerful computing technologies must be used to build the finest intelligence of forecasting and decision-making abilities evaluated by an expert. The purpose of this paper is to examine the effectiveness of various classification techniques, such as Naive Bayes (NB), J48, REF Tree, Metaheuristic Optimization (SMO), and Linear regression (LR), on data sets pertaining to cardiovascular disease (CVD). Employing the WEKA tool, an analysis was designed to test the efficacy of several strategies using illness data sets obtained from UCI respiratory. Different categorization methods are used to compare the values of various metrics, such as the percentage of right classifications, the recall, the F-measure, and the time required.**

Keywords—Internet of things, Data mining, disease prediction, Artificial intelligence, Machine learning, health monitoring system, Sustainability.

I. INTRODUCTION

Data mining methods have aided medical research by simplifying the data processing process, allowing a medical professional to make an informed choice about commencing the appropriate therapy. As a result, it may spare the patient from unnecessary delays caused by various tests that must be performed before deciding on a treatment plan.[1] The proposed study centres around the utilization of information mining strategies, for example, choice trees, Credulous Bayes, irregular woods, and relapse examination, for diagnosing malignant growth and cerebrum cancers using AI utilizing Kaggle standard datasets. The detection performance measurements employing data mining approaches are completely adequate. The accuracy values of a cancer diagnosis on the standard dataset are 93.86 percent, 95.61 percent, 95.80 percent, and 98.25 percent, respectively, utilizing tree structure, Naive Bayes, randomized forest, then logistic regression. Overall accuracy values of identification of brain tumour illness on the standard dataset are 97.21 percent, 97.21 percent, 99.04 percent, and 98.14 percent, respectively, utilizing decision tree, Naive Bayes, randomized forest, or logistic regression.[2]
The examination planned to satisfy the accompanying goals:

- To understand machine learning in disease prediction
- To study data mining in disease prediction
- What are the techniques of data mining in diseases mining in medical data?
- To study the fundamentals of data mining
- To study issues and difficulties in disease prediction
- To study algorithms in machine learning in medical data
- To study areas of future in machine learning and data mining in the healthcare sector

II. LITERATURE REVIEW

The far and wide utilization of PC-based advancements in the medical services business has brought about a deluge of electronic information. Clinical specialists are attempting

to actually evaluate side effects and distinguish sicknesses at an underlying point because of enormous volumes of information. Moreover, standard managed learning (ML) advancements have shown a huge commitment to beating current disease demonstrative techniques and helping clinical professionals in the early distinguishing proof of high-risk issues. The objective of this examination is to distinguish designs in ailment finding across various types of managed AI by analyzing execution markers. Nave Bayes (NB), Decisions Trees (DT), and K-Nearest Neighbor waste is most often discussed in supervised ML techniques (KNN). According to the data, the Support Vector Machine (SVM) is indeed the best at diagnosing renal illnesses and Parkinson's disease. The Logistic Regression (LR) method predicted cardiac illnesses quite well. Finally, in precise breast illnesses & common illnesses, Random Forest (RF) or Convolutional neural network (CNN) predicted, correspondingly.

III. RESEARCH METHODOLOGY

Healthcare is always a big worry in any technological advancement that humans makes. This recent Coronavirus onslaught, which has partially devastated the economy, is a good instance of the increasing need for health insurance. In areas where the virus has spread, it is always advisable to monitor persons by ls using remote health monitoring tools.ML algorithms are particularly susceptible to mistakes.[3] To begin, it is dependent on the quality or selection of datasets, which would be critical for making accurate and impartial conclusions. Second, ML algorithms rely largely on the appropriate selection of characteristics derived from the dataset, thus proving challenging, time intensive, and computationally demanding. These issues impede the learning model's effectiveness and result in catastrophic mistakes that risk patients' lives.

In comparison, contended that normal statistical approaches, job experience, and medical physicians' intuition all contributed to unfavorable biases and inaccuracies in recognizing disease-related risks[4]. With the massive increase of health-related electronic data, physicians are finding it difficult to effectively diagnose illnesses at an early stage. As a result, powerful computational approaches such as MLX techniques were developed to identify relevant patterns or hidden information in data that may be utilized to make vital decisions. As a result, the load on medical personnel was reduced, while patient survival rates improved.

Standards of machine learning for diseases prediction are shown in table no 1:

Table 1- Performance of Machine Learning algorithm

Percent (%)	Standards of machine learning for disease prediction
12-20%	Speed and audio with efficient accuracy
25-30%	Analytics and computer vision
40-60%	Content generation and deep learning with NLP

A. Data mining in disease prediction

Significant information technology advances have resulted in an overabundance of data in health care bioinformatics.

Public healthcare informatics data comprises hospital information, patient information, illness information, or cost of treatment. These massive amounts of data are created from many sources and formats.[5]

It may include extraneous qualities as well as missing data. Using data mining tools to extract insights from enormous amounts of illness data is a critical strategy. There are several methods in which data mining may be used to get insights from large collections of sickness data. It is feasible to utilize information mining methods like arrangement, bunching, or rule mining to break down the information to remove applicable data from it. Among the most fundamental data mining applications in the healthcare system include predicting future treatment outcomes using previous information accumulated from medical illnesses, disease diagnosis using patient data, analyzing treatment costs and resource demand, pre-processing of additive noise, missing information, and minimizing the time for waiting for disease diagnosis. Data mining tools also including Weka, Rapid miner, as well as Orange, are being used to analyze and predict improved outcomes in health care data. [6] New and contemporary data mining methods and technologies are employed in disease diagnosis or health care bioinformatics to enhance health care services while lowering illness diagnostic time.

B. Techniques for Data Mining

In disease information examination, information mining techniques like order, bunching, and affiliation rules are usually utilized.

a) Classification

Classification is a data mining process that is based on machine learning. Classification is the process of categorizing each piece of information in a batch of data into a specific preset set of categories or classes. It classifies data using mathematical approaches including decision trees, linear programming, neural networks, or statistics.[7]

Modern categorization approaches provide further sophisticated strategies for illness prediction. The help vector machine, discriminate investigation, guileless based, choice trees, and straight &nonlinear relapse are instances of arrangement drawing near.

b) Clustering

Cluster is a data mining approach that uses an automated technique to create clusters of items with similar characteristics. Clustering establishes classes and places things in them when the

category is not preset. K-means, Fuzzy Cleans (FCM), Rough Cleans (RCM), Rough-Fuzzy Cleans (RFCM), Robust RFCM (RCM), and hierarchical or Gaussian mixture are examples of cluster approaches.[8]

c) *Mining for Association Rules*

Association rule mining is a well-known and well-studied approach for discovering intriguing relationships between various types of data in huge datasets. Its goal is to detect well-built rules revealed in databases by using various techniques of significance dependent on the set of input data.[9] The data mining procedure of discovering rules, frequent patterns, connections, correlations, and other causal structures amongst groups of items that may control relationships and causality objects amongst sets of things is known as association rule mining. Discover client purchasing behaviors by identifying relationships and correlations here between various things in their "shopping basket." Basket's data analysis, cross-marketing, and catalog design are some of the most common uses of association rule mining. The data mining methods described above may be utilized to diagnose illnesses.[10]

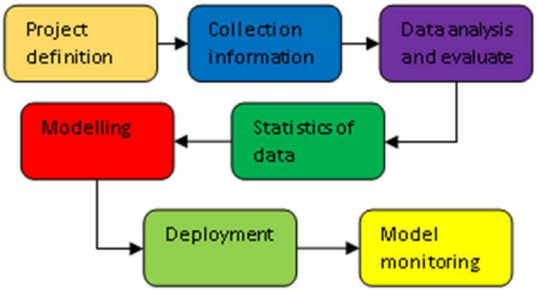

Figure 1. Risk prediction using machine learning

C. *Data Mining Fundamentals*

Information mining is the procedure of computationally getting obscure data from monstrous measures of information. It is basic to separate significant data from monstrous informational indexes and give dynamic results to disease determination and treatment. By assessing and estimating various ailments, information mining might be used to acquire information. Medical care information-digging offers huge potential for revealing secret examples inside clinical area information sets.[11] There are a few information mining strategies open, with not entirely set in stone by the medical care data. Utilizations of information mining in medical care have colossal commitment and adequacy. It computerizes the system of finding prescient information in huge datasets. Information mining depends vigorously on sickness expectations. Different tests should be performed on the patient to analyze a condition. On the opposite side, the use of data mining technologies may reduce the total number of tests required. Performance and time are impacted significantly by this reduced test set. [12]Data mining in health care is critical because it enables physicians to identify which characteristics, such as age, weight, symptoms, and so forth are most important for determining a diagnosis. This one will allow physicians to identify the condition more quickly.The process of discovering relevant information & patterns within the data is often known as knowledge discovery using databases. Data mining may be used to discover knowledge in datasets. [13]

Using data mining mostly in the medical industry is a difficult challenge in the medical profession. Data mining in medical research starts with a theory, and outcomes are altered to meet the hypothesis. This varies from the typical data mining approach, which begins with datasets and no obvious hypothesis. Patterns and themes in datasets are mostly concerned with conventional data mining, but they are not respected in medical data mining. Clinical decisions are often made based on the doctor's intuition.[14] Unwanted prejudice, mistakes, also, over-the-top clinical costs affect the nature of care given to people. Information mining can possibly give an information-rich climate. It can possibly improve the meaning of clinical choices.

The three-machine learning supervised learning techniques are employed in the study of, the cardiovascular disease dataset was analyzed using these methods.[15] This algorithm's Classification Performance should be examined. This research should be expanded to predict heart disease with fewer features. In the surveys, cardiovascular disease is forecasted using a frequent patterns data mining approach. The author presented a method that employs a search restriction to reduce the number of constraints. In the future, this research should be expanded by using fuzz learning models to determine the precision of time in a bid to reduce the number of rules. There isa survey in which the author introduced a novel approach for categorization that employs a weighted association rule. This study may be expanded in the future by using the association rule concealing approach in data mining.[16] The author presented the smallest group of variables for heart disease predictions in his survey. This study may be developed and improved in the future for the automated prediction of heart diseases. Real data of health care agencies and organizations should indeed be obtained to compare the optimal accuracy including all data mining techniques. In the investigation, the author anticipates the characteristics of a diabetic patient developing a cardiac disease. As a consequence of using the Weka tool, the Bayes model was capable of correctly identifying 74% of the input examples. This study will be expanded in the future by including additional data mining approaches.[17]

IV. RESULTS AND ANALYSIS

Medical diagnoses are becoming more dependent on machine learning. Patient survival rates have increased dramatically because of advances in disease classification and detection techniques, which provide data that aid medical professionals in the early detection of life-threatening disorders. Disorders of the blood vessels, such as coronary arterial disease (CAD), congenital heart abnormalities, heart valve disorders, heart infection, and

heart muscle are just few of the many conditions that may affect the heart. Damage to the body's ability to absorb and utilize glucose from diet is what causes diabetes. Autoantibodies, pancreatic damage, genetics, stress, obesity, hypertension, smoking, low HDL cholesterol, and lack of physical activity are all risk factors for developing diabetes. Liver enlargement, portal hypertension, irregular bleeding, intense itching, excessive fatigue, and yellowing of the skin and eyes are all symptoms often associated with liver disease. The kidneys, a crucial organ, are situated in one's lower back. Reduced urine output, swelling in the legs, chronic nausea, and shortness of breath are all symptoms of kidney failure. This section discusses the outcomes of several categorization strategies, including as Naive Bayes, J48, REF Tree, SMO, Multi-Layer Perceptron, and LR algorithm, applied to data sets pertaining to various illnesses.

A. KNN

KNN The illness will be predicted by the user in the Health Sector. In this approach, the user may anticipate whether or not the sickness will be detected. In the proposed method, illness is classified into numerous groups that indicate whichever disease will occur based on symptoms. With each regression and categorization problem, the KNN rule is employed. The KNN algorithm is constructed on the feature similarity method.

B. NAIVE BAYES

Naive Bayes is a simple yet very effective rule for prognosticative modeling. The 'naive' assumption permits decomposing joint probability into a composite of marginal probabilities. Naive Bayes is the name given to this basic Bayesian classifier. The Naive Bayes classifier posits that the existence of one character in a class is independent of the presence of another. It is simple to construct and beneficial for huge datasets. A supervised learning method is Naive Bayes.[18]

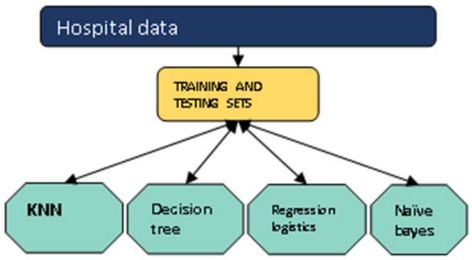

Figure 2. Algorithm and Systemic Architecture in Machine Learning

C. Regression in Logistcs

Logistic regression is a supervised learning classification technique used to estimate the likelihood of a disease target variable. Because the nature of an objective or variable is separated, there are only two viable groups.
In simple terms, the variables are binary by nature, with information represented as either 1 (for success / yes) or 0 (for failure / no). A logistic regression model predicts (y=1) as a function of x.

D. The Decision Tree

A tree is a structure that might be utilized to successfully part a major assortment of information into more modest arrangements of records by utilizing a progression of fundamental choice trees.

The membership of the resultant sets grows more similar to one another with each consecutive division. A decision tree model comprises a set of rules for splitting a big diverse community into shorter, more homogenous (mutually exclusive) groups about a certain aim.

The variable is often categorical, as well as the decision tree has been used to either: calculate the likelihood that a particular record belongs to each of the categories or classify the recorded by designating it to the most probable class (or category).

Dataset characteristics include date of birth, sex, hypertension type (4 values), resting pulse rate, cholesterol level in mg/dl, fasting sugar levels > 120 mg/dl, resting electro physiologic scores (values 0, 1, 2), maximum heart rate attained, exercise-induced chest pain, old peak = ST anxiety triggered by physical activity relative to rest, the slope of the maximum exercise ST segment, the number of major vessels (0-3), colored by fluoroscopy, thal: 3 = normal; 6 = fixed When it comes to heart disease, random forest is the most effective classification method. It is 83.77 percent accurate in its classifications. The numerical results of the simulations are shown in table 2.

Table 2- Accuracy assessment of various Algorithm on Heart Disease

	Correctly classified instance/ incorrectly classified instance	Kappa statistics	Precision	Recall	F-Measure
NB	227/45	.67	.84	.82	.80
MLP	213/61	.57	.77	.75	.76
SMO	230/44	.68	.86	.88	.78
RF	*222/55*	.72	.71	.73	.74
LR	*208/65*	.87	.84	.79	.81

*NB-NaiveBayes, MLP- Multilayer Perception, SMO- Sequential minimum optimization, RF- Random forest, LR- Linear Regression

Ensemble learning is used by the random forest algorithm to find answers to difficult questions. Random forest is an effective tool for analyzing heart disease characteristics, since it is similar to decision trees but includes many decision trees. In a machine learning application using RF signals, the result is predicted using a tree-based method. The table 2 shows that the F measure of LR performs better than its competitors by a margin of around 81%.

V. CONCLUSION & FUTURE SCOPE

Certain data mining strategies perform rather well when used to the study of cardiovascular diseases. After conducting a thorough analysis of the relevant published research, we have determined that support vector machines (SVMs) and naïve bayes (NBs) are the most well-known and widely used algorithms for the forecasting of sickness. Both approaches have an accuracy that is much superior than that of other algorithms. Complex algorithms like as KNN, SMO, and Random Forest are also used, despite the fact that they have not achieved broad acceptance and preference for the purpose of sickness prediction. The incapacity of statistical models to deal with significant volumes of data. The processing of large data sets is made much easier with the help of data mining. A literature review of the different approaches of sickness prediction that are presently in use is presented in the first section of this study. In addition, a range of respiratory data sets from the University of California, Irvine is analyzed using Weka using the data mining techniques of Naive Bayes, J48, REF Tree, SMO, Multilayer Perceptron, and LR. Instances that have been correctly categorized, accuracy, recall, and F-measure are the metrics that are used to assess the effectiveness of algorithms. According to the findings of the simulations and the accompanying debate, linear regression offers the best accuracy when it comes to predicting cardiovascular disease. Recent research has encouraged the use of quantum-based algorithms and deep learning techniques to large data analysis, which might inform our future projects.

REFERENCES

[1] "Analysis and prediction of heart disease using machine learning and data mining techniques," Canadian Journal of Medicine, 2021.

[2] A. Jain and A. Kumar Pandey, "Modeling and optimizing of different quality characteristics in electrical discharge drilling of titanium alloy (grade-5) sheet," Materials Today: Proceedings, vol. 18, pp. 182–191, 2019. [3]"Comparative study of data mining techniques on heart disease prediction system: A case study for the," *International Journal of Science and Research (IJSR)*, vol. 5, no. 5, pp. 1564–1571, 2016.

[4] C. Catal, "Software mining and fault prediction," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 5, pp. 420–426, 2012.

[5] V. Poornima and D. Gladis, "Analysis and prediction of heart disease aid of various data mining techniques: A survey," *International Journal of Business Intelligence and Data Mining*, vol. 1, no. 1, p. 1, 2018.

[6] A. Jain, A. K. Yadav, and Y. Shrivastava, "Modelling and optimization of different quality characteristics in electric discharge drilling of titanium alloy sheet," *Materials Today: Proceedings*, vol. 21, pp. 1680–1684, 2020.

[7] "Intelligent cardiovascular disease prediction using data mining techniques," *International Journal of Pharmaceutical Research*, vol. 10, no. 4, 2018.

[8] A. Jain and A. Kumar Pandey, "Modeling and optimizing of different quality characteristics in electrical discharge drilling of titanium alloy (grade-5) sheet," *Materials Today: Proceedings*, vol. 18, pp. 182–191, 2019.

[9] V. Panwar, D. Kumar Sharma, K. V. Pradeep Kumar, A. Jain, and C. Thakar, "Experimental investigations and optimization of surface roughness in turning of EN 36 alloy steel using response surface methodology and genetic algorithm," *Materials Today: Proceedings*, vol. 46, pp. 6474–6481, 2021.

[10] A. Jain, C. S. Kumar, and Y. Shrivastava, "Fabrication and machining of fiber matrix composite through electric discharge machining: A short review," *Materials Today: Proceedings*, vol. 51, pp. 1233–1237, 2022.

[11] "A review: Prediction on chronic kidney disease using data mining methods," *International Journal of Pharmaceutical Research*, vol. 12, no. sp1, 2020.

[12] V. Panwar, D. Kumar Sharma, K. V. Pradeep Kumar, A. Jain, and C. Thakar, "Experimental investigations and optimization of surface roughness in turning of EN 36 alloy steel using response surface methodology and genetic algorithm," *Materials Today: Proceedings*, vol. 46, pp. 6474–6481, 2021.

[13] "Heart disease prediction with data mining clustering algorithms," *International Journal of Computing, Communication and Instrumentation Engineering*, vol. 4, no. 1, 2017.

[14] A. Jain and A. Kumar Pandey, "Modeling and optimizing of different quality characteristics in electrical discharge drilling of titanium alloy (grade-5) sheet," *Materials Today: Proceedings*, vol. 18, pp. 182–191, 2019.

[15] "Data Mining Apriori algorithm for heart disease prediction," *International Journal of Computing, Communication and Instrumentation Engineering*, vol. 4, no. 1, 2017.

[16] E. Saleh and M. F. Bin Abd Kadir, "Prediction of chronic kidney disease using data mining techniques," *SSRN Electronic Journal*, 2022.

[17] "Data mining classification algorithms for heart disease prediction," *International Journal of Computing, Communication and Instrumentation Engineering*, vol. 4, no. 1, 2017.

[18] N. Briones and V. Dinu, "Data mining of high density genomic variant data for prediction of alzheimer's disease risk," *BMC Medical Genetics*, vol. 13, no. 1, 2012.

[19] R.Anupriya, P.Saranya, and R.Deepika, "Mining Health Data in Multimodal Data Series for Disease Prediction," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 6, no. 2, pp. 96–99, 2018.

[20] N. - and D. P. Mittal, "Healthcare data analysis using data mining techniques for disease prediction," *Indian Journal of Computer Science and Engineering*, vol. 12, no. 5, pp. 1224–1237, 2021.