myresearchgo Volume 2, Issue 6, June 2026, ISSN: 3107-3816 (Online)

Disease Prediction System Using Health Data

Shamika More

email id- smggaikar@gmail.com

Abstract

The widespread dissemination and accessibility of information have led to unprecedented amounts of information. A huge part of this information is random and untapped, while very little of it is regulated. This has led to the urgent need to regulate this impressive volume of information for use in many tasks such as corporate decision-making, increasing their competitiveness, etc. This has led to creating and developing algorithms with the ability to classify and organize data and extract knowledge from it. This facilitates the process of predicting, detecting, or preventing diseases, thereby preserving human capital, reducing expenditures, and keeping society healthy. This technology is a promising opportunity for investment and growth in various fields. The latest statistics of the Saudi General Authority for Statistics were in its report on the results of the survey. (Chronic diagnosed diseases among the Kingdom's population are among those aged 15 and over, at 16.4%. Chronic diseases increase significantly as age increases. The prevalence of chronic diseases among older persons in the 65-year age group is 7.7% higher than among young age groups (15–34 years, at 4.4%). This is what motivated the researcher to find the best means of predicting diseases and how they might help us understand the initial signals of diseases and avoid them. The study aims to review and analyze classification applications, clarify their uses and important features in the field of disease detection and the future of these classifications in the Kingdom, and compile the latest studies in disease prediction using classification algorithms. The researcher used a survey approach to answer the question of this research. This survey includes a review of previous studies from 2018 to 2021 in the Disease Prediction System using classification techniques. These studies have reached many of the results that we list as follows:

- Proper selection of attributes plays an important role in enhancing and increasing the accuracy of classification systems, especially as the same classifications have been used to determine disease-specific attributes.
- Guidance to the need for researchers to choose the right classification for their study where it gives faster and more accurate results.
- The use of macro-learning for big data, exploration and automation using classification techniques gives more accurate and sensitive results.

Based on the foregoing, we found that not all data will serve any purpose without discovering knowledge and that data mining helps to shape the perception of hidden patterns and trends in data sets for diseases that may not have been known before, as the results of the survey showed that the classification techniques used to predict heart disease were as follows: (Naive Bayes 84%, hybrid classification techniques using a total of

87.4% classification techniques, then the random forest 88.7%, and finally the most accurate percentage by study is the decision tree technique that gave 99.2%), and the classification techniques with diabetes disease came as follows: (The synthetic neural network is the most accurate and sensitive 98.4%, followed by the closest neighbors, support vectors, Naive Bayes, and finally the decision tree), and classification techniques to predict diseases in general: On the other hand, (the neural network had an accuracy of 84.5%. The most commonly used classifications were support vectors, followed by Naive Bayes technology). These classifications lead to clear and correct decisions that benefit the economy, health, and all areas of service. Indeed, in light of its blessed Vision 2030, the Kingdom of Saudi Arabia has achieved qualitative leaps in the use of classification applications. Sehaty, Tawakkolna, and Taba'd from the first apps that have been creative in predicting diseases through the use of mobile phones loaded with those smart apps.

Introduction

Healthcare organizations generate large volumes of patient data every day. Analyzing this data manually is challenging and time-consuming. Traditional diagnosis methods depend heavily on medical expertise and may sometimes fail to identify hidden patterns in patient records.

Machine Learning (ML) and Data Mining techniques provide efficient solutions for extracting meaningful information from healthcare datasets. Disease prediction systems use historical medical records and predictive algorithms to estimate the probability of a patient developing a particular disease.

The objective of this project is to develop a Disease Prediction System that uses health-related parameters to predict diseases accurately. Such systems can assist doctors, healthcare institutions, and patients by providing quick and reliable health assessments.

Identification of Research Problem

When we think about mining, the first thing that comes to mind is extracting gold, silver, or other valuable metals. In the world of data, the concept is not much different. You can extract information to discover and extract hidden resources and value (insights and knowledge) which is one of the most valuable assets, as knowledge is power. However, different metals are limited resources, while data is abundant, limitless, and reusable.

So, why was the disease prediction system using classification techniques chosen for this research?

The Kingdom of Saudi Arabia has placed significant importance on the healthcare sector, where local spending has exceeded 150 billion Saudi riyals. This increasing expenditure has led to the search for ways to reduce it through intensified efforts and cooperation between relevant entities in the Kingdom. Through the National Transformation Program, King Abdulaziz City for Science and Technology, in collaboration with several universities, research centers, and hospitals, has launched an initiative to localize and develop the latest technologies, including the health localization and development initiative. This project includes the development of an analytical predictive electronic platform for health data that meets the current and future health needs of the Kingdom. It aims to identify the relationship between lifestyle and diseases using machine

learning [2], and thus significantly reduce healthcare expenses by detecting diseases before critical stages and determining the optimal time for preventive examinations and the detection of diseases before their critical stages.

The importance of this research is due to the need to understand how classification applications and their uses in disease prediction can be used to reduce the high costs incurred by the state in this field. We also need to know to what extent we are keeping up with this development and the main research gaps in this field (Most of the research was on diabetes and heart disease, where many of the current prevalent and predictable diseases were not cared for and avoided some research. They did not provide clear figures on classification techniques; they simply mentioned the methodology used, the lack of research in this area, and limited them to certain diseases. Some research required a subscription to understand the research, and some research did not mention the research mechanism. Finally, there were few studies on the importance of classification techniques during the coronavirus crisis). Therefore, The problem of this research is due to the need to show the different studies on the most used and most accurate classification techniques "due to the lack of diversity of the diseases to be predicted" in order to answer the following main question: How did the applications of data mining and classification and their most important advantages and uses in predicting diseases begin and develop?

1. Introduction

When we think about mining, the first thing that comes to mind is extracting gold, silver, or other valuable metals. In the world of data, the concept is not much different. You can extract information to discover and extract hidden resources and value (insights and knowledge) which is one of the most valuable assets, as knowledge is power. However, different metals are limited resources, while data is abundant, limitless, and reusable.

So, why was the disease prediction system using classification techniques chosen for this research?

The Kingdom of Saudi Arabia has placed significant importance on the healthcare sector, where local spending has exceeded 150 billion Saudi riyals. This increasing expenditure has led to the search for ways to reduce it through intensified efforts and cooperation between relevant entities in the Kingdom. Through the National Transformation Program, King Abdulaziz City for Science and Technology, in collaboration with several universities, research centers, and hospitals, has launched an initiative to localize and develop the latest technologies, including the health localization and development initiative. This project includes the development of an analytical predictive electronic platform for health data that meets the current and future health needs of the Kingdom. It aims to identify the relationship between lifestyle and diseases using machine learning [2], and thus significantly reduce healthcare expenses by detecting diseases before critical stages and determining the optimal time for preventive examinations and the detection of diseases before their critical stages.

The importance of this research is due to the need to understand how classification applications and their uses in disease prediction can be used to reduce the high costs incurred by the state in this field. We also need to know to what extent we are keeping up with this development and the main research gaps in this field (Most

of the research was on diabetes and heart disease, where many of the current prevalent and predictable diseases were not cared for and avoided some research. They did not provide clear figures on classification techniques; they simply mentioned the methodology used, the lack of research in this area, and limited them to certain diseases. Some research required a subscription to understand the research, and some research did not mention the research mechanism. Finally, there were few studies on the importance of classification techniques during the coronavirus crisis). Therefore, The problem of this research is due to the need to show the different studies on the most used and most accurate classification techniques "due to the lack of diversity of the diseases to be predicted" in order to answer the following main question: How did the applications of data mining and classification and their most important advantages and uses in predicting diseases begin and develop?

2. Data Mining

Finding patterns in huge datasets by the application of intersecting techniques like machine learning, statistics, and database systems is known as data mining. It is the process of looking at data from several angles, finding imbalances, patterns, and correlations in datasets that are insightful and helpful in forecasting outcomes that support wise decision-making. The initial steps in the data mining process involve gathering data from multiple sources, preparing it, and centrally storing it. Despite having its roots in the 1990s, data mining has little to do with the actual process of searching for data. Sometimes called "knowledge discovery," the phrase "data mining" was not created until the 1990s. Its foundation is, nevertheless, entwined with a number of scientific fields, such as machine learning (algorithms that learn from data to make predictions), artificial intelligence (machines that exhibit intelligence akin to that of humans), statistics (the numerical study of data relationships), and the cognitive field of business.

Figure 1: Data mining areas

Data mining, often known as "knowledge discovery in databases," is a long-standing technique used to uncover hidden correlations and forecast future trends. While the term "data mining" was not created until the 1990s, statistics, machine learning, and artificial intelligence are the three interconnected scientific fields that form its basis. We have been able to move past laborious and time-consuming manual procedures to swiftly and conveniently analyse data over the past ten years because to rapid and repeated advancements in processing power and speed. The possibility of finding pertinent insights increases with the complexity of the data sets gathered. Data mining is used by a variety of businesses, including banks, insurance companies, manufacturers,

retailers, and telecommunications service providers, to find connections between various factors such as demographics, competitive dynamics, and social media, and how these factors affect their business models, revenues, operations, and relationships with customers. Thus, what makes data mining crucial? The amount of generated data doubles every two years, as evidenced by the startling statistics. Ninety percent of the digital world is made up entirely of unstructured data. However, having more data does not equate to having more knowledge. You can: use data mining

- Identify all the random and repetitive noise in your data.
- Understand what is relevant and use that information effectively to evaluate potential outcomes.
- Accelerate the pace of informed decision-making [3].

The importance of data mining technology has emerged as a result of the significant development in database usage in the latter half of the 1990s, along with the need for what is called knowledge discovery, which allowed for dealing with a large quantity of data. Hence, the importance of data exploration technology has been recognized, thereby providing accurate and correct information [4] for the benefit of different institutions and organizations. We can summarize the stages of knowledge discovery using data mining technology into eight stages "Figure 2": data discovery, data filtering and cleansing, data integration, data selection, data transformation, data mining, model evaluation, and knowledge presentation.

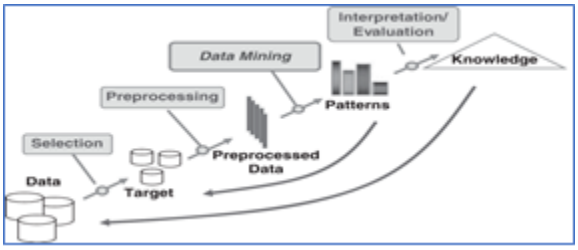


Figure 2: Stages of knowledge discovery

Types of classification algorithms Data analysis techniques such as classification algorithms are used to identify models that precisely characterise significant data categories and classifications:

- **Logistic regression**
- **Nayef Baez**
- **Random regression:**
- **K- The closest neighbours**
- **Decision Tree:**
- **Random forest**
- **Transport truck support**

- **Industrial neural networks**
- **Synthetic neural networks**

Types of classification algorithms

1. Logistical regression

Logistical regression is a calculation used to predict a binary result: something happens or not [6]. Logistic regression is designed for this purpose (classification),

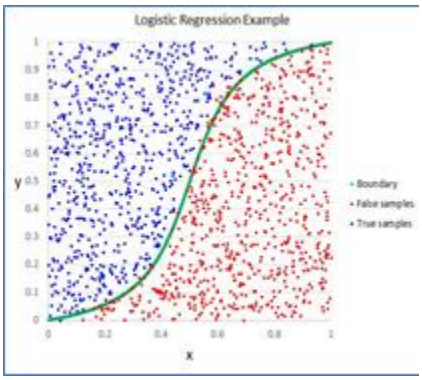


Figure 3: Logistic regression classification method

2. Nayef Baez

The Bayes theory is a mathematical formula that allows us to calculate "reversed" policing possibilities [7].

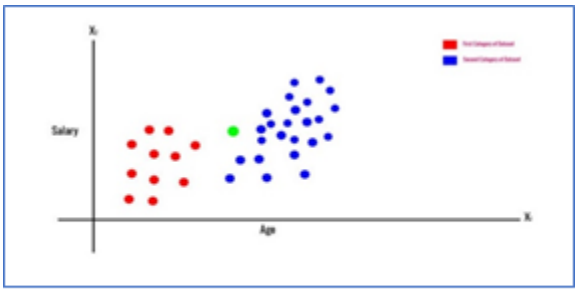


Figure 4: How the Naif Bayes classification works

3. k-nearest neighbors algorithm:

This algorithm is one of the machine learning algorithms that operates with the supervisor and not the closest neighbor of the predictive and descriptive classification algorithms and has the capabilities to generate local estimates of the point, [8].

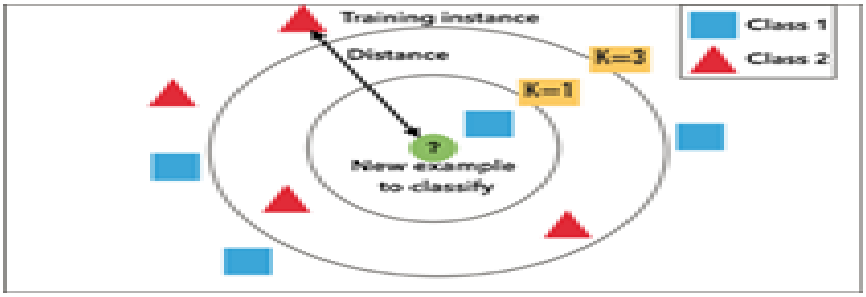


Figure 5: How to classify K-Nearest Neighbors

4. The decision tree:

Decision tree is a predictive modeling technique used in statistics, data mining and machine learning. The decision tree is a simple representation of the classification of examples [7].

Figure 6: How the decision tree classification works

5. Random Forest:

It is a machine learning algorithm that is developed based on a set of decision trees, [9].

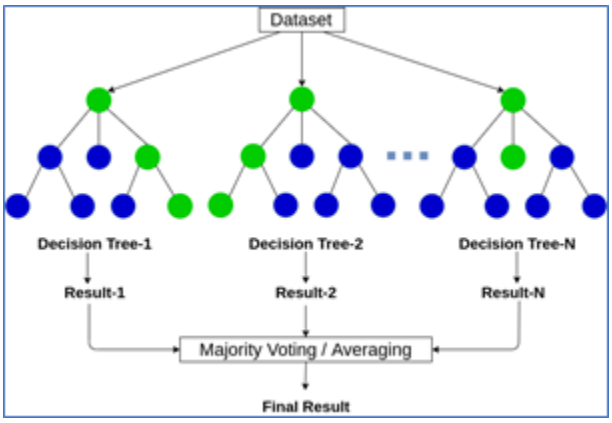


Figure 7: How random forest classification works

6. Support Vector

This algorithm finds a linear or non-linear, flat or set of surfaces in another dimension that differs in length from the feature vector .

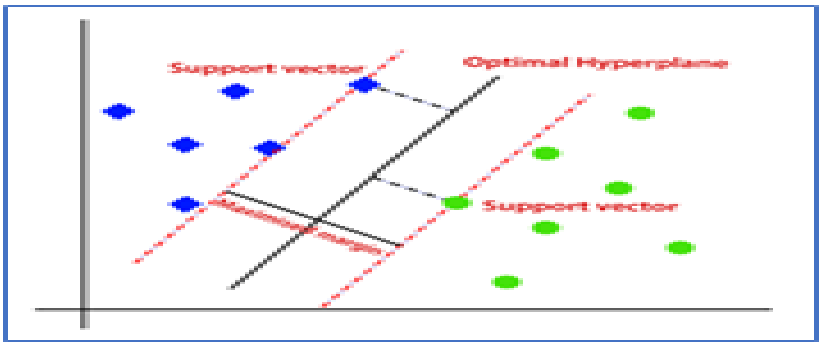


Figure 8: How support vector classification works

7. Artificial neural network

Deep learning techniques are based on artificial neural networks (ANNs), also known as simulated neural networks (SNNs), which are a subset of machine learning. The human brain served as the model for both its name and structure.

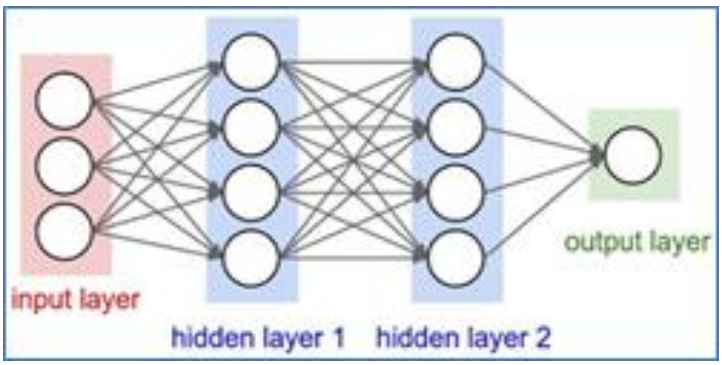


Figure 9: Neural network diagram

8. Recurrent Neural Networks (RNNs)

Because RNNs are made up of neurons with trainable weights and biases, they resemble classic neural networks quite a bit. The output is the sum of all the classes, and each neuron takes in some inputs and computes some dot products. Here, some of the computation methods used in conventional neural networks are still applicable.

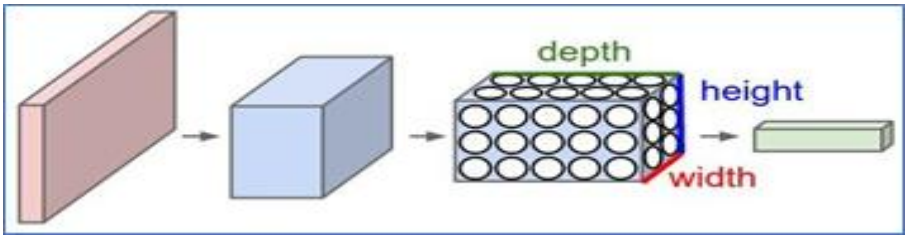


Figure 10: Recurrent Neural Networks diagram

9. Synthetic neural network

The synthetic neural network is made up of multiple layers; some layers require parameters, while others do not; the inputs and outputs of the network are both three-dimensional.

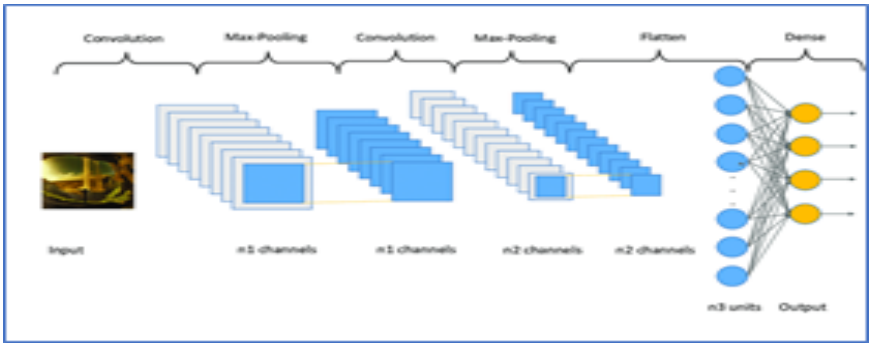


Figure 11: How Synthetic neural network classification works

10. Critical reviews of previous studies

Most studies have agreed that the choice of the algorithm for data mining in terms of features is important and necessary to give more accurate and clear disease forecasting results and have been interested in improving these techniques and making comparisons between them in terms of the most used and applied to the most accurate and sensitive, and have indicated their improvement and development besides diversifying applications of the algorithm for data mining to predict diseases, especially chronic diseases (Diabetes, heart) of the decision tree, and Nave Baez, The neural network, the random forest and, groups classification, and supporting transport trucks where the neural network algorithm distinguished itself from the rest of the past by taking into account a person's living habits and screening information with features for early prediction and accurate diagnosis of diseases. I refer to the coronavirus crisis and its relationship with chronic diseases such as diabetes and its impact on them, and I urge the use of exploration techniques to accurately select features to predict diabetes early to avoid the risks of this pandemic.

In a study , various machine learning techniques were used to predict diabetes, and the results showed that the random forest algorithm can achieve higher accuracy when using all features. In a study which reviewed studies on feature selection for predicting chronic diseases, it emphasized the importance of selecting appropriate features to enhance classification accuracy. Applying classification algorithms to disease datasets led to promising results in developing these systems. Since most studies in this research were conducted they focused on improving the effectiveness of prediction and identifying important features using data mining techniques for heart disease. In this study, grouping was used to improve the accuracy of weak algorithms and was also applied to medical datasets to demonstrate its usefulness in early disease prediction. The study also indicated that clustering techniques such as oversampling and boosting are effective in improving prediction accuracy, and improving feature accuracy results in more accurate predictions, which is consistent with the study . Feature selection and data extraction techniques can improve the accuracy of predicting heart and vascular diseases by using different feature sets and classification techniques. The results showed that the heart disease prediction model developed using important and specified features and data extraction techniques performed the best with accuracy.

However, in study , hybrid techniques were used for more effective prediction of heart disease, where a prediction model was presented using different sets of features and various known classification techniques. The output was more accurate through the heart disease prediction model with random forests.

In study , which proposed a general disease prediction based on patient symptoms, and to predict the disease, a convolutional neural network algorithm was used for accurate prediction, which requires a set of disease symptom data and takes into account the person's lifestyle habits and examination information. It was clarified that this system is capable of giving the risks associated with the disease in general, which reduces the risks of infection.

Using algorithms as feature selection techniques shows improvement in results by showing prediction accuracy. About the prediction of diabetes using machine learning techniques.

A. Using Machine Learning Techniques to Predict Diabetes Mellitus

In order to predict diabetes, decision trees, random forests, and neural networks were employed in this study. The 14 features in the dataset are physical examination results from a hospital in Luochu, China. Five-fold cross-validation was employed in this investigation to confirm the models. Some methods with greater performance were chosen for independent testing experiments to confirm the methods' global applicability. Data from diabetic patients and 68,994 randomly chosen healthy people were utilised as training sets, respectively. The average of the five trials was obtained after the data was randomly extracted five times due to data imbalance. To minimise dimensions, this study used minimum redundancy maximum relevance (mRMR) and principal component analysis (PCA). The findings demonstrated that, when all features are used, random forests can produce predictions with the highest accuracy (ACC = 0.8084). Study in a study on systems of selection and classification of traits for predicting chronic diseases: a review.

B. A review of feature selection and classification methods for predicting chronic diseases

Improving the classification systems' accuracy is largely dependent on the right feature selection. Reducing dimensionality also aids in enhancing machine learning algorithms' overall performance. The use of classification algorithms to disease datasets yields promising outcomes in the form of intelligent, automated, and adaptable chronic disease diagnostic systems. It is possible to expedite the procedure and improve the computing efficiency of the outcomes by using parallel categorization systems. This book offers a thorough analysis of several feature selection techniques along with an explanation of their benefits and drawbacks. Next, in order to predict chronic diseases, analysis is done on parallel and adaptive categorization schemes. Study [14] on improving the accuracy of predicting heart disease risks based on ensemble classification techniques.

C. Increasing the precision of heart disease risk prediction using ensemble classification methods

This study looks into an approach known as ensemble classification, which combines several classifiers to increase the accuracy of weak algorithms. This tool was used to run experiments on a dataset of heart disorders. A comparative analytical method was used to figure out how to use ensemble approaches to increase heart disease prediction accuracy. In order to show the algorithm's value in early disease prediction, this research focuses on both improving the accuracy of weak classifiers and applying it using medical datasets. The study's findings reveal that aggregation methods, such bagging and boosting, perform satisfactorily in predicting heart disease risks and are useful in increasing prediction accuracy for weak classifiers. With the aid of ensemble classification, weak classifiers saw a maximum gain of 7%. Feature selection was used to further enhance process performance, and the outcomes demonstrated a notable increase in prediction accuracy. Study on effective prediction of heart diseases using hybrid machine learning techniques.

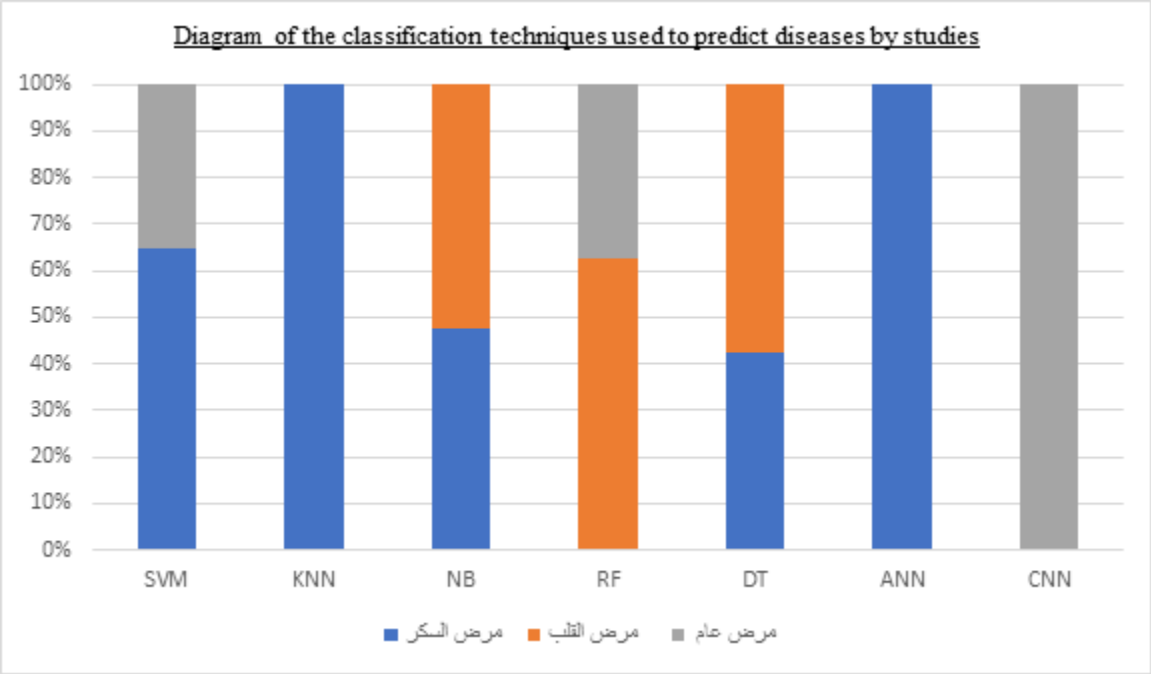
D. Effective Prediction of Heart Disease Through Hybrid Machine Learning Approaches

There are few studies that concentrate on significant characteristics that are essential in predicting heart and circulatory disorders. The selection of an appropriate subset of features is critical in order to enhance prediction model performance. The goal of this research is to pinpoint important characteristics and data mining strategies that can improve the precision of heart and cardiovascular disease prediction. Different feature sets and seven classification methods—k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network, and Voting—were used to create prediction models. Voting is a hybrid method that combines Naive Bayes with Logistic Regression. According to the experiment results, the best data mining technique (voting) combined with the identified significant features to create a prediction model for heart illnesses has an accuracy of 87.4%. Study on designing a disease prediction model using a machine learning approach.

E. Using Machine Learning to Design a Disease Prediction Model

The hardest part is predicting diseases accurately. In order to solve this issue, data mining is crucial to the prediction of illness. Large amounts of medical data can contain hidden pattern information that can be found through data mining. Based on the symptoms of the patient, a broad disease prediction is suggested. We employ the K-Nearest Neighbour (KNN) algorithm to forecast the illness. A collection of disease symptoms is

necessary for disease prediction. For an accurate general illness prognosis, the individual's lifestyle choices and examination data are considered. CNN outperforms KNN in the general disease prediction accuracy, coming in at 84.5%. Additionally, KNN requires more memory and time than CNN. This system can reduce the chance of contracting a general or higher disease by providing the related hazards of the general disease after the general disease prediction.



Limitations & Future Scope:

Limitations

1. Prediction accuracy depends on dataset quality and size.
2. The system may not account for all medical conditions.
3. Real-world healthcare data may contain inconsistencies.
4. Limited disease categories are included in the current model.

Future Scope:

1. Integration with Electronic Health Records (EHR).
2. Deployment as a web or mobile healthcare application.
3. Use of Deep Learning techniques for enhanced accuracy.
4. Real-time disease monitoring through IoT healthcare devices.

Conclusion:

The Disease Prediction System Using Health Data demonstrates the potential of machine learning in healthcare applications. By analyzing patient health parameters, the system provides accurate disease predictions and assists in early diagnosis. Among the evaluated algorithms, Random Forest achieved the highest accuracy and proved to be the most effective model. The project highlights the importance of predictive analytics in improving healthcare services, reducing diagnosis time, and supporting medical professionals in clinical decision-making. Future enhancements can further improve system reliability and applicability in real-world healthcare environments.

References

1. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.
2. Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
3. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
4. UCI Machine Learning Repository. Healthcare Datasets.
5. Kaggle Healthcare Datasets Repository.
6. Patel, V. L., Shortliffe, E. H., Stefanelli, M., et al. (2009). The Coming of Age of Artificial Intelligence in Medicine. Artificial Intelligence in Medicine.
7. Deo, R. C. (2015). Machine Learning in Medicine. Circulation.
8. Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques.