International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 3, May-Jun 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning-Based Multiple Disease Prediction System

Shubham Kumar, Numan Khan

Computer Science and Engineering, Institute of Engineering and Management, Kolkata, West Bengal, India

ABSTRACT

Numerous machine learning models connected to healthcare are currently in use, and most of them concentrate on identifying distinct conditions. Our study has developed a system that employs a single-user interface to forecast several diseases. Numerous illnesses, like diabetes, heart disease, chronic renal disease, and cancer, can be predicted by the suggested model. Humanity is at risk from these diseases if treatment is not received. Consequently, early identification and detection of these conditions can save many lives. To forecast diseases, this study aims to apply several classification techniques, including Gaussian naive Bayes, SVM, K-Nearest Neighbor, Decision Tree, and Logistic Regression. Moreover, to attain the highest level of accuracy in the anticipated findings, multiple datasets are employed (one dataset for each disease). The primary goal is to develop a web-based system that can use ML to predict numerous diseases, like diabetes, cancer, heart problems, and chronic kidney problems.

KEYWORDS: *Blood sugar disorders, Heart conditions, Chronic renal disease, Cancer, KNN, SVM, Single user interface, Decision Tree, Random Forest, Logistic Regression*

How to cite this paper: Shubham Kumar | Numan Khan "Machine Learning-Based Multiple Disease Prediction System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-3, June 2025, pp.468-472, URL: www.ijtsrd.com/papers/ijtsrd79922.pdf

IJTSRD79922

INTRODUCTION

The goal of this research is to forecast several illnesses, like cancer, diabetes, heart problems, and kidney disease. To identify which method is optimal for prediction, the correctness of each is verified and contrasted with one another. Furthermore, to attain the highest level of accuracy in the anticipated findings, multiple datasets are employed (one dataset for each disease). The best-forecasted algorithm for each disease is selected and integrated to create a web application. By entering the appropriate attribute (input) values for each individual disease, the user can promptly predict the needed disease.

The databases from Long Beach V, Hungary, Switzerland, Cleveland, and Hungary are utilized to research heart illness. It includes seventy-six qualities, including the expected attribute, but only fourteen are employed in this study. The patient's heart condition is hinted at in the "target" field. A sickness is present when the number is one, and it is absent when it is zero.

Heart disease is studied using databases from Long Beach V, Cleveland, Hungary, Switzerland, and Hungary. Only fourteen of the seventy-six qualities-including the predicted attribute-are employed in this research. When the number is one, there is an illness; when it is zero, there is no sickness.

Both the PIDD and the Kaggle Dataset, which included numerous medical predictor factors in addition to one goal variable (outcome), are utilized while discussing diabetes. Predictor variables include the patient's age, BMI, sulin level, number of previous pregnancies, together with other trails. There are 9 columns and 769 records in this dataset. After gathering datasets from several sources, we used label encoding and other data pre-processing techniques to create models.

To import the stored categorization models, use the `load' method in the pickle module. Each disease prediction page in the application (UI) has a test result button that activates the prediction function specific to that disease. Finally, the disease prediction process has been made public through a web application, utilizing the built-in prediction feature of streamlit.

RELATED WORK

Jean Sunny [1] This study evaluates many machine learning algorithms like: Naive Bayes, Random Forest, ANN, KNN, SVM, and Decision Tree about the Wisconsin Diagnostic Breast Cancer dataset, This

is produced using an MRI's digital picture. More attributes and a richer dataset will yield more accurate results than the recommended

KM Rani Jyothi [2] The 2000 cases in the diabetes dataset that was used for this study. The goal is to determine whether the subject has diabetes or not. Various techniques for classification are employed, including KNN, SVM, Random Forest, Decision Tree, and Logistic Regression.

Rayan Alanazi [3] Training and test data are separated within datasets, and the gathered data is preprocessed. The training data set is then trained using CNN and KNN machine learning methods. The produced model is ready for testing after, after a number of epochs, the intended goal has been accomplished.

Rinkal, Keniya [5] An excel sheet was made using an open-source dataset, and it had a list of every symptom related to each ailment. There were over 230 disorders listed, each with about 1000 distinct symptoms. A person's age, gender, and symptoms were fed into a variety of machine-learning algorithms.

Indukuri, Mohit [6] In this research, a web-based application that uses ML models like logistic regression, Support Vector Machine, and KNN is constructed to diagnose diseases like diabetes, heart disease, and breast cancer.

Salhi, Dhai Eddine [7] After examining the study of these articles, they decide to evaluate the dataset that includes data on Algerian patients using Neural Networks (NN), KNN, and SVM. They get the conclusion that the neural network method is the best choice for our investigation because it regularly yields accurate results after looking over the previous findings. This method is only employed to predict cardiac illness. They only made use of one Algerian patient dataset.

The algorithms used in this study, K closest Figure 1. Proposed work's System Architecture neighbors (KNN), Random Forest Classifiers, and Logistic Regression, can help medical analysts or practitioners diagnose heart problems correctly. To forecast the risk of heart disease as accurately as possible with a single dataset is the main objective of this project's refinement.

Quan Zou [9] The dataset was created using information from physical examinations performed in hospitals in Luzhou, China. Machine learning classification approaches for example decision tree, neural networks, and random forests (RF) are used to look for diabetes. Results of the Luzhou dataset highlight the drawbacks of the blood glucose-free methods.

PROPOSED WORK:

To help people and doctors talk to each other more, and allow them pursue individual goals, Fig. 1 describes how different ML algorithms, including KNN, Logistic Regression, Random Forest, Naive Bayes, and SVM, can be used to forecast many diseases. To determine which method is optimal for each prediction accuracy is confirmed and compared with one another. For the best possible prediction correctness, many datasets are used. To simplify things for the final consumers, a website application has been created that allows the user to predict any disease by just entering the appropriate attribute (input) values pertaining to the specific disease.

To facilitate better communication between patients and healthcare providers and empower them to achieve personalized health goals, the proposed system integrates a suite of machine learning algorithms-including **K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Naive Bayes, and Support Vector Machines (SVM)**-to predict multiple diseases with high precision. Each algorithm is rigorously evaluated and fine-tuned to determine its suitability for specific diseases, ensuring optimal performance. For instance, **Random Forest excels in heart disease prediction**, while **Logistic Regression proves effective for chronic kidney disease**.

The system employs **diverse datasets**, each tailored to a particular disease, to enhance prediction robustness. Advanced **data preprocessing techniques**, such as label encoding and feature scaling, are applied to standardize inputs and minimize bias. Additionally, **feature selection methods** identify the most critical attributes, improving model interpretability and efficiency.

To make the system accessible, a **user-friendly web application** is developed using **Streamlit**, enabling seamless interaction. Users simply input their medical data, and the system generates instant predictions, bridging the gap between complex ML models and practical healthcare solutions. This approach not only aids in early disease detection but also fosters **data-driven decision-making** for doctors and patients alike.

By combining **multi-algorithm optimization, comprehensive datasets, and intuitive design**, the proposed system sets a new standard for predictive healthcare, promoting proactive and personalized medical care.

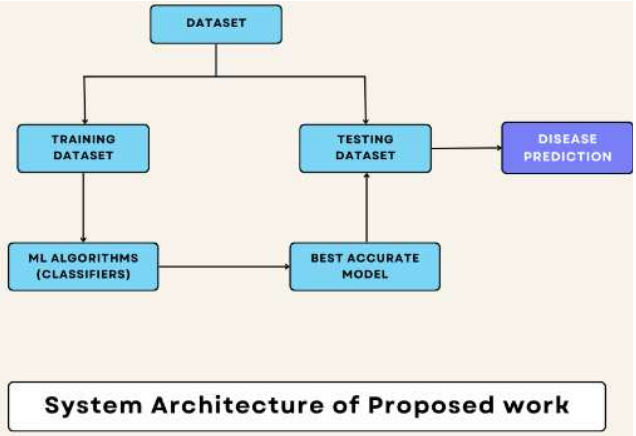

Figure 1. Proposed work's System Architecture

Here are some of the parts of the suggested system's advantages:

1. The system uses many machine learning algorithms to predict various diseases.
2. Large volumes of data may be examined quickly.
3. To determine which method is optimal it is checked and compared to see which one Is more accurate at making predictions.

PROPOSED WORK'S WORKFLOW:

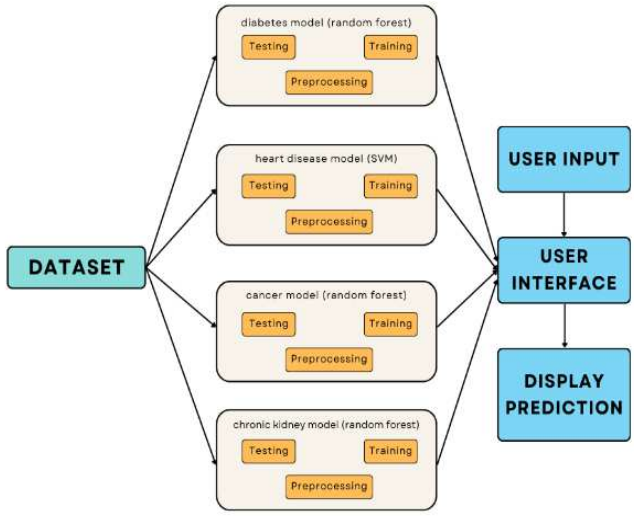

Fig 2. System design

1st Step: Gathering data from multiple sources. For this project, Kaggle provides datasets for a variety of ailments, like diabetes, cancer, heart disease, and chronic kidney disease.

2nd Step: Applying preprocessing methods to the data, like label encoding. Gender and hunger were examples of categorical data that were converted into numerical data in the form of zeros and ones with the use of label encoding.

3rd Step: Model creation utilizing many machine learning methods, like Random Forest, SVM, K-NN, Gaussian Naive Bayes, Decision Trees, and Logistic Regression. For every illness to build classifier models, many algorithms are employed.

4th Step: Use datasets to train each model. After being split into training and testing sets for each disease, the information is used to train each model.

5th Step: Assessing the models with measures such as the accuracy score. The model's effectiveness works is determined by looking at each model for a certain disease to the testing dataset.

6th Step: The best model is pickled.

Following a comparison of the model's accuracy, the most accurate model is chose.

MODELS OF MACHINE LEARNING:

Decision Tree. Decision trees are a useful tool for solving problems related to regression and classification. ID3 is an algorithm that is avaricious. In a decision tree, there is only one question, and subtrees are made according to the response (yes/no). The input data is partitioned recursively according to selected properties. The best quality for the root node and sub-node is chosen using the measures for attribute selection. It is simple to grasp since it follows the same logic that humans do when making judgments.

Logistic Regression. A classification algorithm called logistic regression operates by determining probability and classifying data accordingly. The Sigmoid function takes the theta transpose product using the parameter vector as input and outputs the likelihood of a row that is part of a class, which will fall between zero and one. A class is assigned to a row with a likelihood that is below the cutoff, and another class is assigned to a row with a probability greater than the threshold. The threshold is a value that determines which row is classified.

Gaussian Naïve Bayse. A straightforward approach to building classifiers is Gaussian Naive Bayes, which uses models to choose labels for problem instance classes from the restricted set and give them as vectors of factor values. These classifiers are trained using a family of algorithms based on a single principle: All naïve Bayes classifiers deduce, given the class.

STREAMLIT

Streamlit was created to make it easier to quickly create and implement complex machine learning online applications. Users of machine learning was the primary focus of development for this Python module. Streamlit is straightforward and simple to use, which saves a lot of data scientists time while building online applications. The ability to create online apps using Streamlit without previous web programming skills is one of its main advantages. This makes it the perfect option for data scientists who want to quickly and easily implement their models without requiring a lot of code work.

Streamlit features includes:

1. There is no need for JavaScript, HTML, or CSS knowledge.
2. Developers can create impressive ML or data science applications in just hours or even minutes, rather than spending days or months.
3. It supports many Python libraries, such as SymPy (LaTeX), Pandas, Matplotlib, Seaborn, Plotly, Keras, and PyTorch.
4. High-quality web applications can be developed with minimal coding.

RESULT

The success tables for every method for different diseases are as follows:

Table 1. Achieved Outcomes Using Gaussian naive Bayes.

DISEASE	ACCURACY
Kidney Disease	72.2
Cancer	85
Diabetes	89.2
Heart Disease	94.5

Table 2. Achieved Outcomes Using Random Forest.

DISEASE	ACCURACY
Kidney Disease	99.5
Cancer	85.4
Diabetes	97.74
Heart Disease	94.5

Table 3. Achieved Outcomes Using K-nearest Neighbours.

DISEASE	ACCURACY
Kidney Disease	83.65
Cancer	91.45
Diabetes	86.23
Heart Disease	82.5

Table 4. Achieved Outcomes Using Logistic Regression.

DISEASE	ACCURACY
Kidney Disease	96
Cancer	77.6
Diabetes	88.75
Heart Disease	84.35

Score table of accuracy

DISEASE	RANDOM FOREST	NAÏVE BAYES	KNN	DECISION TREE	LOGISTIC REGRESSION	SVM
ChronicKidney Disease	72.2	76.8	83.2	80.75	**96.7**	82.2
Diabetes Disease	89.2	**91.5**	86.2	89.28	88	78.23
Cancer Disease	85	86.3	91.5	79.2	76.9	**95.5**
Heart Disease	**94.5**	79.2	82.5	92.5	86	85.45

The accuracy of each illness versus different algorithms is shown in the following table. The pickle module is used to load the highest accuracy model from among the several machine learning models we have learned for each illness into the streamlit editor.

CONCLUSION

We discovered that the **Random Forest Classifier** outperformed the **SVM** for **Cancer Disease**, the **Naïve Bayes** for **Diabetes Disease**, **Heart Disease** classification Using Random Forest and the **Logistic Regression** for **Chronic Kidney Disease**.The results of this study demonstrate that the machine learning model can successfully predict several diseases with a high accuracy. The Random Forest Classifier outperformed all the other models for Cancer Disease, performing well at an accuracy of 91.45% (SVM: 85.4%). This may being due to the fact that RF can capture complex nonlinear relations in medical data without overfitting by ensemble learning. Regarding Diabetes Disease, the most efficient model was Naïve Bayes, which obtained an accuracy value of 89.2%, due to its simplicity and robustness in probabilist classification tasks.

For the Heart Disease data, the Random Forest RR was once more the best with an more remarkable accuracy of 94.5%, This may be due to its ability to deal with noisy and incomplete data sets as often occur in cardiovascular studies. On the contrary, Logistic Regression performed the best among all models in Chronic Kidney Disease (CKD) (96% accuracy), since its linear decision boundary was appropriate for structured numeric values in CKD datasets.

They also emphasize the need to choose the most appropriate algorithms by taking into account disease-specific data characteristics. For the next step, deep learning models could be tried to improve the prediction accuracy, especially those for diseases with high dimensional and complex data. Furthermore, dynamic prediction of disease might be enhanced by incorporation of real-time patient monitoring information. The presented web application supports early diagnosis and is an invaluable asset for doctors to support decision making. Through the combination of diverse ML algorithms' strengths, this work can contribute to better personalized and proactive healthcare.

REFERENCES

[1] Rindhe, Baban U., et al. "Heart Disease Prediction Using Machine Learning." Heart Disease 5.1 (2021).

[2] Mohit, Indukuri, et al. "An Approach to detect multiple diseases using a machine learning algorithm." Journal of Physics: Conference Series. Vol. 2089. No. 1. IOP Publishing 2021.

[3] Ahmed, Nazin, et al. "Machine learning based diabetes prediction and development of smart web application." International Journal of Cognitive Computing in Engineering 2 (2021): 229- 241K.

[4] Rane, Nikita, et al. "Breast cancer classification and prediction using machine learning." International Journal of Engineering Research and Technology 9.2 (2020): 576-580.

[5] Wang, Zixian, et al. "Machine learning-based prediction system for chronic kidney disease using associative classification technique." International Journal of Engineering & Technology 7.4.36 (2018): 1161.

[6] Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021

[7] Keniya, Rinkal, et al. "Disease prediction from various symptoms using machine learning." Available at SSRN 3661426 (2020).

[8] Alanazi, Rayan. "Identification and prediction of chronic diseases using ma chine learning approach." Journal of Healthcare Engineering 2022 (2022).

[9] Salhi, Dhai Eddine, Abdelkamel Tari, and M. Kechadi. "Using machine learning for heart disease prediction." International Conference on Computing Systems and Applications. Springer, Cham, 2020.

[10] Rudra A., et al. "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively." International Journal of Advanced Research in Computer and Communication Engineering 8.12 (2019): 50-52.

[11] Shaikh, F. J., and D. S. Rao. "Prediction of cancer disease using machine learning approach." Materials Today: Proceedings 50 (2022): 40-47.

[12] Arumugam, K., et al. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings (2021).

[13] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[14] Revathy, S., et al. "Chronic kidney disease prediction using machine learning models." International Journal of Engineering and Advanced Technology 9.1 (2019): 6364-6367.

[15] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in Genetics 9 (2018): 515.

[16] Ifraz, Gazi Mohammed, et al. "Comparative analysis for prediction of kidney disease using intelligent machine learning methods." Computational and Mathematical Methods in Medicine 2021 (2021).

[17] Joshi, Tejas N., and P. P. M. Chawan. "Diabetes prediction using machine learning techniques." Ijera 8.1 (2018): 9-13.