MULTI-DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM

Ankit Yadav[1], Harsh Dwivedi[2], Gaurav Srivastava[3], Anshuman Pandey[4]

[1] *Undergraduate Student, Computer Science And Engineering, Institute Of Technology And Management GIDA Gorakhpur, India*
[2] *Undergraduate Student, Computer Science And Engineering, Institute Of Technology And Management GIDA Gorakhpur, India*
[3] *Undergraduate Student, Computer Science And Engineering, Institute Of Technology And Management GIDA Gorakhpur, India*
[4] *Undergraduate Student, Computer Science And Engineering, Institute Of Technology And Management GIDA Gorakhpur, India*

ABSTRACT

Machine learning techniques have revolutionized healthcare by enabling accurate and timely disease prediction. The ability to predict multiple diseases simultaneously can knowingly improve early diagnosis and behavior, leading to better patient outcomes and reduced healthcare costs. This research paper explores the application of machine learning algorithms in multi-disease prediction, focusing on their benefits, challenges and future directions. We provide an overview of the various machine learning models and data sources commonly used for disease prediction. Moreover, we discuss the importance of feature selection model estimation, and the integration of multiple data modes for heightened disease prediction. The research findings highlight the probable of machine learning in multi-disease prediction and its potential impact on public health Once more, I am applying machine learning model to identify that a person is affected by few diseases or not. This training model takes a sample data and train itself for predicting the ailment In the face of growing health problems, the project "Multi-Disease Prediction System Using Machine Learning" aims to solve a fundamental problem: the proactive identification and prediction of various diseases for effective healthcare management. Powered by the rising complexity of healthcare data and the need for personalized and timely interventions, this initiative aims to revolutionize disease prediction, prevention, and management through the integration of progressive machine learning procedures. The development of intellectual systems for disease detection and diagnosis has been made possible by the spread of machine learning techniques, which has made a significant involvement to the healthcare industry. This task proposes a complete way to deal with tending to the test of various illness recognition employing AI calculations. The essential objective is to plan an effective and exact framework equipped for recognizing and grouping different infections at the same time, giving an encircling perspective on a singular's wellbeing status. The diverse dataset that the proposed system makes use of includes apposite clinical information, imaging data, and medical records. A multi-modular organization is taken on, coordinating information from various sources to improve the strength and dependability of the recognition model. AI calculations, for example, convolutional brain organizations (CNNs), support vector machines (SVMs), and group strategies are utilized to learn complex examples and connections inside the information. The motivation behind this project stems from the growing drain of multiple diseases and the domineering to shift from reactive to hands-on healthcare strategies. Traditional healthcare models often fall short in anticipating and avoiding diseases, leading to increased healthcare costs and negotiated patient outcomes. By attributing the power of machine learning algorithms, such as cooperative methods and deep learning models, the project seeks to investigate diverse health datasets, including genetic information, lifestyle factors, and ancient medical records.

Keyword: - Multi-Disease Prediction, Machine Learning Algorithms, Healthcare Technology etc.

1. INTRODUCTION

Machine learning has emerged as a essential force in transforming the countryside of medical diagnostics as a result of the merging of healthcare and technology. This task discourses a critical step forward in medical services by donating a cutting edge Multi-Sickness Discovery Agenda. The essential goal is to saddle the capacities of AI with a complex stage prepared to instantaneously recognize and evaluate the weightiness of numerous illnesses inside people. A segmented understanding of an individual's health is produced by traditional diagnostic methods, which frequently involve isolated taxations for distinct medical conditions. This undertaking tends to this limit by utilizing AI calculations to dissect a different cluster of information sources, including experimental records, imaging information, and clinical data. By incorporating these different information modalities, our framework intends to give a all-inclusive and exact evaluation, empowering the concurrent recognition of a range of illnesses. As innovative heads keep on reshaping the medical services scene, the probable for early recognition and mediation has never been seriously encouraging. When it comes to deciphering intricate patterns and associations within complex datasets, machine learning algorithms, particularly those that make use of deep neural networks and ensemble approaches, offer proficiencies that are unparalleled. Through the precise preparation of our framework on a reaching dataset, we want to supply it with the capacity not exclusively to recognize sicknesses yet additionally to offer involvements into their movement and seriousness. This undertaking includes crucial parts like information reconciliation, highlight withdrawal, model preparation, and the improvement of an easy-to-understand interface custom-made for medical services experts. The synergistic joining of these machineries plans to make a hearty and solid device, engaging clinical authorities with a thorough comprehension of a singular's wellbeing status in a solitary affectionate cycle. The imagined result is a groundbreaking change in medical care, heartening early mediation, working with customized therapy plans, and at last working on quiet results.In the complementary segments, we will dive into the strategy, key highlights, and expected effect of our Multi-Illness Recognition Framework, giving a guide towards a future where medical services are more proficient, particular, and undeveloped.

2. LITERATURE SURVEY

It's true that machine learning and artificial intelligence have become integral parts of various industries, including the medical industry. Predictive models based on machine learning algorithms can help detect diseases accurately and quickly, allowing doctors to provide better treatment and care to patients. Your project to detect multiple diseases such as heart disease, liver disease, and diabetes using machine learning algorithms is a great initiative. Using algorithms such as Random Forest and K nearest neighbor (KNN) can help achieve maximum accuracy and improve the overall effectiveness of the predictive model. However, it's important to note that machine learning models are not always perfect and may have limitations. It's important to validate the accuracy of the model using real-world data and to have a medical expert validate the results to ensure the safety and well-being of patients. Overall, the use of machine learning and artificial intelligence in the medical industry has great potential and can lead to significant advancements in healthcare.

[1] It's great to see that you are proposing a system that can predict multiple diseases using machine learning algorithms and the Flask API. This system has the potential to improve the efficiency and accuracy of disease prediction as well as help doctors provide better care to their patients. By using machine learning algorithms and Tensor Flow, you can train models that can analyze multiple diseases simultaneously. You can also use the Flask API to create a disease parameters and the disease name, and then invoke the corresponding model to predict the disease status. The use of machine learning and the Flask API in this system has several benefits, including faster and more accurate disease prediction, early warning of potential health risks, and improved patient outcomes. It's also important to note that this system can be expanded to include other diseases in the future, which can further improve its utility and effectiveness. Overall, this system has the potential to revolutionize the way we diagnose and treat diseases and can help save countless lives by detecting diseases early and providing timely treatment.

[2]The use of computer-based technology in the healthcare industry has led to the accumulation of a large amount of electronic data. This made it difficult for medical personnel to appropriately analyze symptoms and detect diseases at an early stage. However, supervised machine learning algorithms have demonstrated significant promise for outperforming current illness diagnosis methods and supporting medical professionals in the early identification of high-risk disorders. Through the analysis of performance measures, this literature review sought to uncover trends in the use of supervised machine learning models for illness identification. Naive Bayes, Decision Trees, and K-Nearest Neighbor were the supervised machine learning algorithms that received the most attention. According to the results, support vector machines are the best at spotting Parkinson's illness and kidney disorders. Heart disease prediction was carried out using logistic regression. Finally, high accuracy predictions for breast diseases and common diseases were made by Random Forest and convolutional neural networks, respectively.

3. SYSTEM DESIGN

3.1 Existing System

A machine can predict diseases, but it cannot predict the subtypes of diseases caused by the occurrence of a single disease. It fails to predict all possible conditions of the people. Existing system handles only structured data. The prediction system are broad and ambiguous. Countless disease estimation classifications have been advanced and implemented in the current past. Permanent organizations organize a mixture of machine learning algorithms that are judiciously accurate in predicting diseases. However the restraint with the prevailing systems are speckled First, the prevailing systems are more expensive, which only rich people could pay for. And also, when it comes to folks, it becomes even higher. Second, the guess systems are non-specific and indefinite so far. So that the machine can imagine a positive disease, but it cannot expect subtypes of diseases and diseases caused by the existence of a single error. If diabetes is predicted to occur in a group of people, undoubtedly some of them could be at complex risk of heart viruses due to current diabetes. The remaining schemes cannot predict all possible tolerant neighborhoods.

3.2 Proposed System

We built a system that uses algorithms and a variety of other tools to predict the patient's disease based on their symptoms, and we compare those symptoms to the system's dataset, which is already available, in the proposed system for predicting multiple diseases using machine learning. By taking those datasets and contrasting them with the patient's illness, we will foresee the exact rate of infection for the patient. The dataset and side effects go into the framework's prescient model, where the information is pre-handled for future reference, and afterward a highlight choice is made by the client, where he will enter/select the different side effects. Then, at that point, the characterization of this information is finished utilizing AI calculations like strategic relapse. Then, at that point, the information goes in the suggestion model, there it shows the gamble examination that is associated with the framework, and it additionally gives the likelihood assessment of the framework to such an extent that it shows the different likelihoods like how the framework acts when there are n number of expectations finished, and it likewise does the proposals for the patients from their eventual outcome and furthermore from their side effects, like it can show what to utilize and what not to use from the given datasets and the end-product. It predicts plausible illnesses by mining informational collections like Coronavirus, Constant Kidney Infection, and Coronary Illness. Apparently, in the space of clinical large information examination, none of the current work zeroed in on the two information types.

3.3 Architecture overview

We directed probes into four illnesses: heart disease, diabetes, mind growth, and Alzheimer. Importing the respective datasets for heart disease, diabetes, and liver disease from the UCI, PIMA, and Indian Over datasets is the first step. After the dataset has been stacked, each inputted piece of information is shown. After pre-handling the information for perception, which includes searching for exceptions, missing qualities, and scaling the dataset, the information is separated into preparing and testing. Then, we utilized the CNN, XG-Lift, and irregular woods calculations on the preparation dataset prior to applying what we realized about the order strategy to the testing dataset. We will choose the calculation with the most noteworthy precision for every illness subsequent to applying our ability. Then, we made a pickle document for every infirmity and consolidated it with the Jar structure to give the model's result on the site.

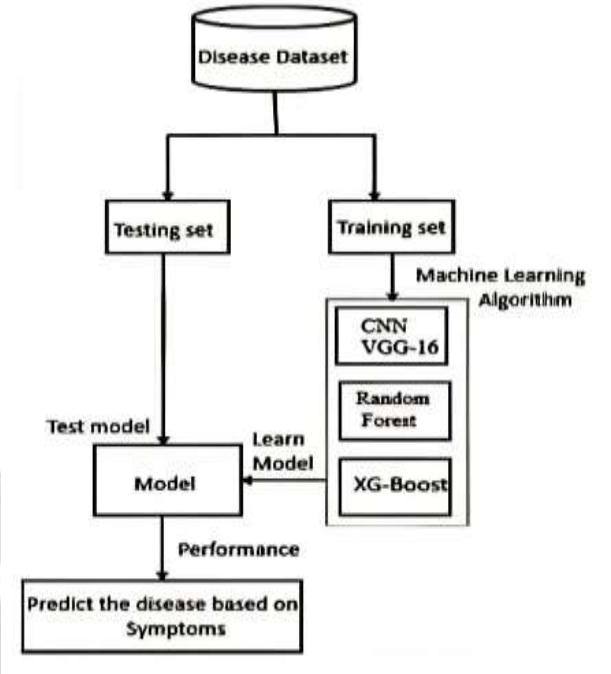


Fig 1. System Architecture

4. ALGORITHMS

4.1 KNN Algorithm

KNN is an AI strategy used for relapse as well as order. The calculation is considered computationally costly on the grounds that it includes various cycles to get the most ideal precision. This procedure is a directed AI strategy, which implies that the information is named and the calculation figures out how to foresee the result from the information. Even when the training data is large and contains noisy values, the algorithm still works perfectly. The dataset is divided into test and training datasets by the algorithm. The preparation dataset is utilized for model structure and

preparing. The test information is anticipated in light of the model fabricated. By and by, we figure the distance between the pre-arranged k-closest component regards and test point.

Distance Metrics The distance between the data feature value and the test inputs can be determined using a variety of distance metrics.

S addresses the distance metric

1.Minkowski Distance:

--(10)

2.Euclidean Distance:

q=2

--(11)

3.Manhattan Distance:

q=1

--(12)

Euclidean methodology is the most widely used procedure to register the distance test and prepare information values.

Step by step instructions to pick a K worth

K shows the boundary, which is the quantity of the closest neighbors. Finding the best value to achieve the best accuracy of the model is a troublesome task. There is no pre-portrayed quantifiable technique to perceive the k worth to achieve astonishing accuracy. The main strategy to find k worth that achieves amazing exactness is to utilize Animal power technique, which implies we want to track down precision for various k worth. The K upsides of neighbors 1 to 20 and the neighbor that gives the most elevated exactness is considered for the expectation.

4.2 Logistic Regression

A strategic relapse calculation utilizes the strategic capability, so this calculation is named Strategic Relapse. The calculated capability is an "S" molded bend produced for measurable functionalities, and the bend is plotted somewhere in the range of 0 and 1. For the portrayal reason, strategic relapse utilizes conditions like straight relapse.

Calculated relapse condition

Y=1/(1+EXPO(-esteem))-- (1)

Input values (for the most part named as x) and co-efficients (Beta) are straightly joined to anticipate the worth of output(termed as y).

Equation of logistic regression y = EXPO(u0+u1*x)/(1+EXPO(u0+u1*x)). --(2)

y is anticipated result, a0 is catch or inclination, and a1 is single information coefficient esteem.

The probability of being first-class (sometimes referred to as the default class) is predicted by logistic regression models. For instance, if we are developing a model for predicting a person's gender based on their height, the default class might be male, and the formal expression for this would be P(gender=male|height). --(3)

For expectation probabilities should transform into twofold qualities, either 0 or 1. Probabilities are transformed into forecasts by utilizing the strategic capability. The model can be made as

y=EXPO(u0+u1*x)/(1+EXPO(u0+u1*x)). --(4)

Further addressing, we get the condition as follows:

ln(p(x)/1-p(x)) =u0+u1*x.- (5)

The left-hand size equation (ratio) is called chances of top of the line or default class. The chances are determined as the likelihood of an occasion isolated by the likelihood of its supplement occasion.

4.3 Random Forest

An irregular woodland might be built by consolidating N choice trees, and afterward creating predictions can be utilized

for each tree that was delivered in the initial step.

The arbitrary woodland works as follows:

Step-1: To start with, it will pick K pieces of information aimlessly from

The preparation set.

Step 2: Create decision trees that are linked to the selected data points (subsets) after selecting k data points.

Step 3: The next step is to select the Nth node for the decision trees you want to build.

Step-4: Rehash stages 1 and 2 in sync 4.

Step-5: Finding every choice tree's expectations and

apportioning the new information to the class with the

Most help is stage five.

4.4 VGG-16 Algorithm

Step 1: The 16 in VGG16 represents 16 weighted layers. Thirteen convolutional layers, five Max Pooling layers, three thick layers, and a sum of 21 layers make up vGG16, albeit just sixteen of them are weight layers, otherwise called learnable boundaries layers.

Step 2: The information tensor size for VGG16 is 224 or 244 and has three RGB channels.

Step 3: The most unmistakable element of VGG16 is that it focused on convolution layers of a 3x3 Modify with step 1 as opposed to numerous hyper-boundaries, and reliably utilized a similar cushioning and maxpool layer of a 2x2 Change with step 2.

Step 4: All through the entire plan, the convolution and max pool layers are consistently requested.

Step 5: There are 64 channels in the Conv-1 Layer, 128 channels in Conv-2, 256 channels in Conv-3, and 512 filters in Conv-4 and Conv-S.

Stage 6: A pile of convolutional layers is trailed by three Completely Associated (FC) layers; the third leads a 1000-way ILSVRC game plan and has 1000 channels. There are 4096 channels (one for each class) in each of the first two FC layers. The delicate max layer is the final remaining one.

4.5 CNN Algorithm

Step 1: The dataset is first changed into a vector design.

Step 2: Following that, word implanting was completed, using no qualities to fill in the information. Word implanting produces a convolutional layer as its outcome.

Step 3: We lead the largest pooling procedure on the convolutional layer in the wake of accepting it as a contribution to the pooling layer.

Step 4: For Max pooling, convert the dataset into a fixed-length vector. The whole associated brain network is

coupled to the pooling layer.

Step 5: The classifier, a SoftMax classifier, is connected to the total association layer.

4.6 XG Boost Algorithm

The XG-Lift calculation works as follows:

Step 1: The first step is to make a tree with just one leaf.

Step 2: Subsequent to registering the normal of the objective variable as an expectation for the principal tree, we should decide the residuals utilizing the predefined misfortune capability. The residuals for future trees are then gotten from the expectation from the principal tree.

Step 3: Utilizing the recipe to decide the likeness score.

Similitude Score = (Residuals)

N = Number of Residuals

A=Regularization Boundary

Stage 4: Utilizing the closeness score, we pick the CO 4/6 hub. Greater homogeneity is seen when the likeness score

is more prominent.

Step 5: The fifth step includes applying the likeness score to the data procured. Data gain uncovers how much homogeneity is gotten by separating the hub at a particular spot and assists with recognizing old and

new likenesses. The formula for calculating it is:

Data Gain = LS + RS-Comparability for Roots

Where LS = Left Comparability and RS = Right Similitude

Stage 6: Utilizing the strategy, you might prune and regularize the tree to the proper length by changing the

regularization hyperparameter.

Step 7: Utilizing the Choice Tree you made, we can then estimate the leftover qualities.

Step 8: The learning rate is utilized to compute the new arrangement of residuals.

Step 9: The new residue equals the old residue plus 8 predicted residues. The next step is to go back to Step 1 and repeat it for each tree.

4.7 Support Vector Machine (SVM)

A supervised method is SVM. This calculation is used for both orders and relapse studies. Data are plotted in n-dimensional space using coordinates in this algorithm, and SVM can be classified as linear or nonlinear. In our work, we use information that is straightly divisible, so we utilize the direct SVM classifier. The grouping of classes is finished by finding the ideal hyperplane. Hyperplanes are the limits that partition the classes into classifications. In two dimensions, the line is a hyperplane. In two-dimensional space, the line is sufficient to distinguish the classes. For instance, consider the condition S0+(S1*U1)+(S2*U2)=0 B0; B1 are coefficients, and B2 is the block of the line. K1, K2, are input focuses. This line is utilized for the grouping.

The data value falls into the category "0" because the value returned by the equation is greater than zero above the line. Beneath the line, the worth returned by the situation is under nothing, and the information point has a place with the class "1." A point that yields a value near zero is hard to order. Edge is alluded to as the distance between the adjacent piece of information and the line. If it has the greatest advantage, the ideal line can separate the classes. This line is known as a maximal edge hyperplane. This edge is enlisted by using the erect distance between the closest feature of the line and the line. Support vectors are the data values, and these points are essential for describing the line and the classifier structure. Hyperplanes are upheld and characterized by help vectors.

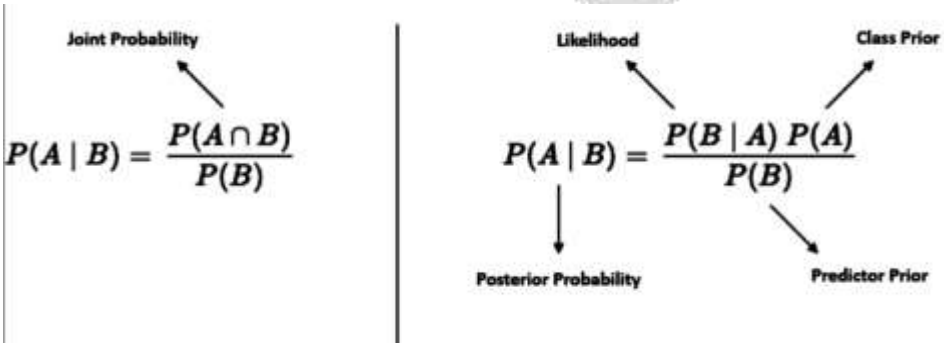
4.8 Naïve Bayes Theorem

In The Bayes in Gullible Bayes comes from Bayes' Hypothesis. Assuming you focused on likelihood and measurements in your numerical class, there's little opportunity you haven't proactively known about Bayes' hypothesis. How about we review it:

Based on prior knowledge of the conditions that might be related to the event, Bayes' Theorem describes the probability of that event.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

This equation is derived from formula of Conditional Probability given below:



P(A) or Class Prior is the response variable's prior probability; P(B) or Predictor Prior is the evidence or probability of training data; and P(AJB) or Posterior Probability is the conditional probability of the response variable having a particular value given the input attributes.

• P(B|A) or Probability is essentially the fire up the back likelihood or the probability of preparing information

Stepwise Bayes Hypothesis

Stage 1-Gather crude information

Stage 2-Convert information to a recurrence table(s)

Stage 3-Work out earlier likelihood and proof.

Step 4: Add probabilities to the equation for the Bayes Theorem.

Innocent Bayes accepts restrictive freedom over the preparation dataset. The classifier isolates information into various classes as indicated by the Bayes' Hypothesis. however makes the assumption that each class's input features have an independent relationship. Consequently, the model is called innocent.

This aides in working on the estimations by dropping the denominator from the equation while accepting freedom:

$$\textit{Bayes Theorem} \longrightarrow P(A \mid x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n \mid A)\, P(A)}{P(x_1, \ldots, x_n)}$$

$$\downarrow$$

$$\textit{Naïve Bayes} \longrightarrow P(A \mid x_1, \ldots, x_n) = P(x_1 \mid A) \cdot P(x_2 \mid A) \cdot P(x_i \mid A)\, P(A)$$

5. RESULT AND DISCUSSION

5.1 Experimental setup

Anaconda tools are used to develop all of the experimental cases in a crowded environment in Python. The contending characterization approach and different component extraction procedures are likewise utilized, and the framework is designed with an Intel Center iS-6200U processor running at 2.30 GHz and 8GB of Smash.

5.2 Dataset

Github,Kaagle, and other ML websites provide access to the disease dataset. Furthermore, according to industry guidelines train set and test are ready. Using the Scikit learn train, test, split method, the data are divided into 70% for training and 30% for testing.

Illustration of Diabetic Sickness:

Diabetes feature train, diabetes feature test, diabetes label train, diabetes label test=train-test split (diabetes features, diabetes label, test size = 0.3, train size = 0.7)

5.3 Evaluation method

Method to begin with, we recognize Genuine Positive (TP), Bogus Positive (FP), Genuine Negative (TN), and Misleading Negative (FN). Genuine positive alludes to the quantity of cases effectively anticipated as the need might arise, bogus positive alludes to the quantity of occasions erroneously anticipated as required, etc. Coming up next are the four estimations that might be acquired: exactness, accuracy, review, and Fl-measure.

Precision = (TP+TN)/(TP÷TN+FP+FN)

Accuracy = TP/(TP+FP)

Review = TP/(TP•FN)

F1 Score = TN/(TN+FP)

5.4 Result

The results for all the ML models and of final completed project are shown in the tables:

SN.	Disease Name	Algorithm Name	Existing system accuracy	Proposed system accuracy
1	Diabetes	SVM Classifier	76%	78%
2	Heart disease	Logistic Regression	80%	85%
3	Parkinson's disease	SVM Classifier	71%	87%
4	Kidney disease	TensorFlow and keras	-	97%
5	Breast cancer	TensorFlow andkeras	-	96%

The current framework doesn't have a kidney illness and bosom malignant growth forecast framework. That is the reason we leave "- " in the current framework precision for kidney infection and bosom malignant growth. expectation framework. that is the reason we leave "- " in the current framework for precision for kidney infection and bosom malignant growth.

6. CONCLUSION

This paper gives an examination of the numerous explorations done in this field. Our proposed framework targets connecting holes among specialists and patients, which will help the two classes of clients accomplish their objectives.

➢ This framework offers help for numerous infection expectations, utilizing different AI calculations.

➢ The current methodology of numerous frameworks centers just around robotizing this cycle, which needs to construct the client's confidence in the framework.

➢ By giving a specialist's proposal in our framework, we guarantee client's trust next to each other, guaranteeing that the specialists won't feel that their business is getting impacted because of this framework.

7. ACKNOWLEDGEMENT

We might want to offer our thanks to our school, Organization of Innovation, and the executives, Gida, Gorakhpur, for furnishing us with the necessary resources to set up a venture regarding the matter of "Various Sickness Expectation Framework Utilizing AI," as well as our mentor, Sir Ajay Gupta, for providing us with the time and resources to conduct the necessary research. Additionally, I would like to express my gratitude to Mr. Ashutosh k. Rao, the head of the Computer Science Department, and Assistant Professor Mr. Nitin Dixit for their assistance during our study, which would have been difficult without their motivation, unwavering support, and insightful suggestions. Moreover, without the participation, counsel, and help of our loved ones, the intricacy of this study article could not have possibly been reachable.

8. REFERENCES

[1] Trends in Coronary Illness: The Study of Disease Transmission

[2] Community for Infectious Prevention and Counteraction (Coronary Illness Realities).

[3] The Global Trend in Cancer Mortality: Asian Pacific Journal: A 25-year study.

[4] Seasons of India: Disease cases rise 10% in 4 years to 13.9 lakh.

[5] Global Diabetes Alliance: Use and passings connected with diabetes.

[6] The study of disease transmission of diabetes: A report of Indian Heart Affiliation.

[7] "Prediction of Diabetes Using Machine Learning Classification Algorithms" by Naveen Kishore G, V. Rajesh, A. Vamsi Akki Reddy, and K. Sumedh, and T. Rajesh Sai Reddy.

[8] Gavin Pinto, Sunil Jangid, Radhika Desai, "Understanding the Way of Life of Individuals to Recognize the Reasons of

Diabetes utilizing information mining".

[9] "Analysis of Heart Disease Prediction Using Machine Learning Techniques," by M. Marimuthu, S. Deivarani, and R. Gayatri.

[10] Dr. Kanak Saxena, Purushottam, and Richa Sharma, "Efficient Heart Disease Prediction System."

[11] Adil Hussain She, Dr. Pawan Kumar Chaurasia, "A Survey on Coronary Disease Forecast utilizing AI

Procedures".

[12] M. Chinna Rao, K. Ramesh, G. Subbalakshmi, "Decision Backing in Coronary Disease Expectation Framework utilizing Credulous

Bayes".

[13] Ch Shravya, ,K.Pravallika, and Shaik Subhani, "Expectation of malignant growth utilizing administered AI

Calculations".

[14] "Breast cancer classification and prediction using machine learning" by Nikita Rane, Jean Sunny, Rucha Kanade, and Sulochana Devi.

[15] Dr. "Breast Cancer Prediction Using Machine Learning Algorithms" by B.Santhosh Kumar, T. Daniya, and Dr. J. Ajayan

[16] Utilizing Machine Learning Algorithms for Cancer Prediction and Detection: A

Comparitive Review".

[17] Coronary illness dataset" by UCI.