IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Survey On Virtual Healthcare Prediction Using Machine Learning

B. KavyaSri, V. Ramya, P.S.G.V. Prasad, B. Pydi Naidu, V. Pavan Kumar, T. Ravi Kumar

Department of Computer Science & Engineering
Aditya Institute of Technology and Management,Tekkali

1. Abstract

In this digital world, human health care is one of the most important issues in society. The widespread use of technology in the healthcare industry has led to the growth of electronic data. With vast amounts of data, physicians face the challenge of accurately analyzing symptoms and recognizing disease early. However, supervised machine learning (ML) algorithms have helped medical professionals predict high-risk diseases early. A disease is a specific abnormal condition that adversely affects the structure or function of living organisms. It is very important to know early if you have the disease instead of discovering it later. In this way, a disease prediction system that predicts disease from symptoms plays an important role. The disease prediction system uses a machine learning algorithm called XG-Boost. In doing so, it predicts different kinds of diseases. Each disease presents a person with different signs and symptoms. This paper reviews various models based on such algorithms and techniques and analyzes their performance Models based on supervised learning algorithms such as Support Vector Machines, Gradient Boosting Classifier, Logistic regression, K- Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), Naive Bayes, Decision Tree, and Random Forest are found very popular among the researches.

Keywords: *Support Vector Machines; Gradient Boosting Classifier; Logistic regression; K-Nearest Neighbor (KNN); Convolutional Neural Network (CNN); Naïve Bayes; Decision Tree; Random Forest.*

2. Introduction

Machine learning is the study of computer systems learning from data and experience. Machine learning algorithms have two paths: training and testing. Predicting disease based on patient symptoms and machine-learning history has been a struggle for decades. Machine learning technology provides an excellent platform for efficiently solving health problems in the medical field. Machine learning technology allows you to build models to analyze data faster and deliver results faster.

Using machine learning technology, physicians can make informed decisions about patient diagnosis and treatment options, leading to improved patient care. Healthcare is the best example of how machine learning is being used in medicine. Machine Learning Disease Prediction also uses medical history and health data, applying various concepts such as data mining and machine learning techniques and algorithms. These approaches should predict the recurrence of specific diseases in advance.

Machine learning helps doctors and other medical professionals identify medical needs and solutions more quickly and accurately. Such early and accurate predictions can help improve individual health.

As the healthcare sector evolves into a new world of technology, many new developments are taking place. Approaches and applications based on artificial intelligence and machine learning are key to advances in this field, including increasing diagnostic speed, accuracy, and simplicity.

Today, billions of searches are performed every day, and the results returned may or may not be good. In these searches, thousands of searches are related to medical advice. People often want to know if they have a serious illness based on their signs and symptoms.

3. Dimensionality Reduction

Dimensionality reduction involves choosing a mathematical representation that captures most, but not all, of the variance in the given data. This ensures that only the most important information is included. The data considered for a task or problem can consist of many attributes or dimensions, not all of which affect the output equally. Many attributes or features can impact computational complexity, leading to overfitting and even degrading results. Dimensionality reduction is therefore a very important step to consider when creating a model. Dimensionality reduction is typically accomplished in two ways: feature extraction and feature selection.

Depending on the method used, dimensionality reduction can be either linear or nonlinear. The main linear method, called Principal Component Analysis or PCA, is discussed below.

This method was introduced by Karl Pearson. It works on the condition that the variance of the data in the low-dimensional space must be maximal while the data in the high-dimensional space are mapped to the data in the low-dimensional space.

A. Feature Extraction

The new feature set is derived from the original feature set. Feature extraction involves transforming features. This conversion is often irreversible because little or perhaps much useful information is lost in the process. This reduces the data in the high dimensional space to the low dimensional space. i.e., A Space with few dimensions

B. Feature Selection

In doing so, it tries to find a subset of the original set of variables or features to get a smaller subset that can be used to model the problem. It usually contains 3 options: 1. Filter, 2. Wrapper, 3. Embedded.

4. Algorithms and Techniques Used

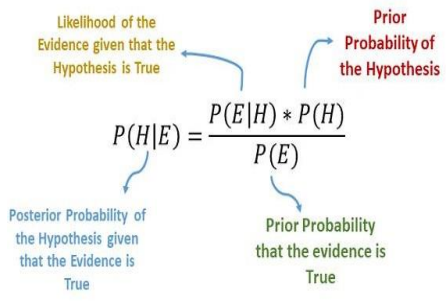
A. Naive Bayes:

The Naive Bayes algorithm is a supervised learning algorithm that uses the Bayes theorem to solve classification problems. The Nave Bayes Classifier is a simple and effective Classification algorithm that supports the development of fast machine learning models that can provide quick predictions. It is a probabilistic classifier, which means it predicts grounded on an object's possibility.

Naive: It is termed Naive because it considers that the presence of one trait is unrelated to the occurrence of others. For example, if the fruit is classified based on color, shape, and flavor, then a red, spherical, and delicious fruit is identified as an apple. As a result, each characteristic helps to identify it as an apple independently of the others.

Bayes: It is termed Bayes because it is based on the Bayes Theorem. The Bayes theorem, often known as the Bayes Rule or the Bayes law, is used to calculate the probability of a hypothesis given past knowledge. It is determined by conditional probability**.**

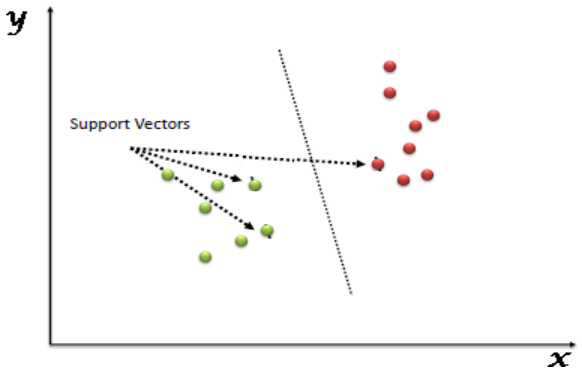
Bayes' theorem can be expressed as follows:



In [2], Naive Bayes Classifier calculates the probability of the disease. In [4], the results obtained by the model using naive bayes have achieved an accuracy of 92.9%. In [6], Naive Bayes has achieved an accuracy of 100% for training and testing datasets. In [9], Naive Bayes has achieved an accuracy of 96.9986% with 52 most significant features. In [10], Gaussian Naive Bayes had an accuracy of 16.8 %. Kernel Naive Bayes had an accuracy of 16.8 %. In [11], Naive Bayes has achieved an accuracy of 77.5% for heart disease in diabetic patients. In [12], Naive Bayes has achieved an accuracy of 95.21% for the disease dataset. In [13], provides support for multiple sickness prediction treatments with completely different Machine Learning algorithms like naive bayes. In [14], The naive bayes model fits the training data and gives an accuracy of 100% on all parameters in the classification problem. In [15], The accuracy of Multinomial Naive Bayes on the chosen data set was 92%. In [17], Naive Bayes with an accuracy of 76.98%.

B. Support Vector Machine

Support Vector Machine, or SVM, is a famous Supervised Learning technique that is used for both classification and regression problems. In Machine Learning, however, it is primarily used for Classification problems. The SVM algorithm's main goal is to find the finest line or decision boundary for categorizing n-dimensional space so that we may simply place fresh data points in the proper category in the future. A hyperplane is the optimal choice boundary. SVM selects the extreme points/vectors that help in the creation of the hyperplane.



In [3], There are 5 diseases for each disease there is a dataset. The experimental results for Breast Cancer prediction using SVM have achieved an accuracy of 97.66%, for heart disease prediction has achieved an accuracy of 51.65%, for kidney disease prediction has achieved an accuracy of 72.50%, for diabetes prediction has achieved an accuracy of 83.33% and for liver disease prediction has achieved an accuracy of 71.18%. In [5], indicates SVM has a high accuracy of 83. In [6], the SVM model has achieved an accuracy of 100% for training and testing datasets. In [7], we have analyzed the accuracy of this system for 5 different diseases and our accuracy can go up to 87%. In [8], the SVM model showed superiority in accuracy at most times for kidney diseases and PD because of its reliability in handling high-dimensional, semi-structured, and unstructured data.

In [9], the SVM has achieved an accuracy of 95.6250% with 52 most significant features. In [11], SVM has achieved an accuracy of 87.9% for heart disease in diabetic patients. In [13], provides support for multiple sickness prediction treatments with completely different Machine Learning algorithms like SVM. In [14], the SVM model fits the training data and gives an accuracy of 100% on all parameters in the classification problem. In [16], the SVM has achieved an accuracy of 83%. In [17], the SVM with linear kernel generates the best performance with an accuracy of 89.21%. In [18], the SVM has accuracy with a standard scale of 96%.

C. Gradient Boosting Classifier

Gradient boosting is used in Machine Learning to resolve classification and regression problems. Gradient Boosting is a supervised learning approach that combines various weak models to generate a robust classifier. It is a sequential ensemble learning approach in which the model's performance increases over iterations. The model is created in stages using this procedure. The model is inferred by allowing the optimization of an absolute differentiable loss function. As each weak learner is added, a new model is formed that provides a more exact estimation of the response variable.

In [3], There are 5 diseases for each disease there is a dataset. The experimental results for Breast Cancer prediction using Gradient boosting classifier have achieved an accuracy of 97.66%, for heart disease prediction has achieved an accuracy of 79.12%, for kidney disease prediction has achieved an accuracy of 97.50%, for diabetes prediction has achieved an accuracy of 86.84% and for liver disease prediction has achieved an accuracy of 71.18%.

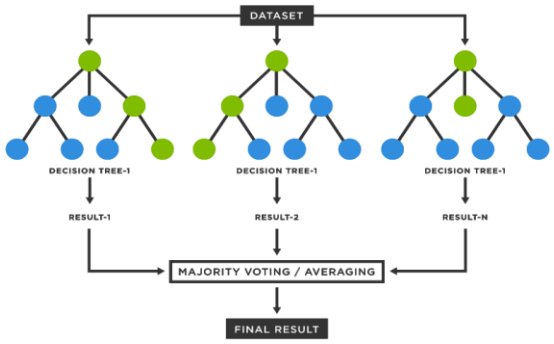
D. Logistic Regression

Logistic regression is a commonly used method for machine learning that belongs to the Supervised Learning approach. It is used to predict the category-dependent variable based on a group of independent variables. The production of a categorical dependent variable is predicted through logistical regression. As a result, the result must be a categorical or discrete value. It can be Yes or No, 0 or 1, true or False, and so on, but instead of presenting precise values like 0 and 1, it presents the probability values that fall between 0 and 1. In [2], By using linear regression we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis. In [3], There are 5 diseases for each disease there is a dataset. The experimental results for Breast Cancer prediction using logistic regression have achieved an accuracy of 95.91%, heart disease prediction has achieved an accuracy of 80.22%, kidney disease prediction has achieved an accuracy of 90.00%, diabetes prediction has achieved an accuracy of 81.14% and liver disease prediction has achieved an accuracy of 69.41%. In [5], It has been found that the ROC value of LR is 86%. In [8], the LR algorithm proved to be the most reliable in predicting heart diseases. In [9], logistic regression has achieved an accuracy of 98.8790% with 124 most significant features. In [15], As for Logistic Regression, it was merely 89%. In [17], the Logistic regression classifier achieved a competitive result with an accuracy of 85.61%.

E. Random Forest

Random Forest is a well-known machine learning algorithm again from the supervised learning approach. It may be applied to both classification and regression problems in machine learning. It is based on the theory of ensemble learning, which is a method that involves merging several classifiers to solve a complicated issue and enhance the model's performance.

Instead of depending on a single decision tree, the random forest collects the predictions from each tree and predicts the final output based on the majority vote of predictions. The bigger the number of trees in the forest, the higher the accuracy and the lower the risk of overfitting.

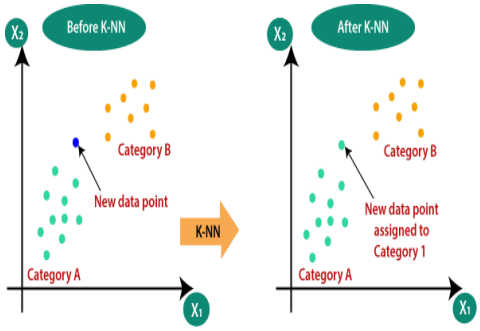


In [4], the results obtained by the model using a random forest have achieved an accuracy of 97.64. In [6], Random Forest has achieved an accuracy of 100% for training and testing datasets. In [9], Random Forest has achieved an accuracy of 80.8566% with 52 most significant features. In [12], Random Forest has achieved an accuracy of 95.11% for the disease dataset. In [13], provides support for multiple sickness prediction treatments with completely different Machine Learning algorithms like Random Forest. In [14], The random forest classifier model fits the training data and gives an accuracy of 100% on all parameters in the classification problem.

In [18], Random Forest has accuracy with a standard scale of 75%.

F. K-Nearest Neighbor (KNN)

K-Nearest Neighbor is one of the most known Supervised Machine Learning techniques. The K-NN algorithm assumes similarity between the new case/data and available cases and places the new case in the category that is most similar to the available categories. The K-NN algorithm saves all available data and classifies a new data point based on its similarity. This signifies that when fresh data comes, it may be quickly sorted into a well-suited category using the K- NN method.

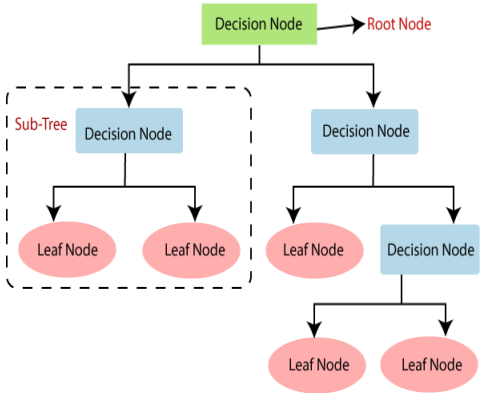


In [9], KNN has achieved an accuracy of 96.0116% with 52 most significant features. In [10], The Weighted KNN model gave the highest accuracy of 93.5 % for the prediction of diseases. In [12], KNN has achieved an accuracy of 95.12% for the disease dataset. In [13], provides support for multiple sickness prediction treatments with completely different Machine Learning algorithms like KNN. In [14], The k-nearest neighbors model fits the training data and gives an accuracy of 100% on all parameters in the classification problem. In [16], After performing the machine learning approach for testing and training we find the accuracy of the KNN is much more efficient as compared to other algorithms, the values have been calculated and it is concluded that KNN is best among them with 87% accuracy. In [18], KNN has accuracy with a standard scale of 57%.

G. Decision Tree

A Decision Tree is a Supervised learning technique that may be used for both classification and regression problems, however, it is most commonly used for classification. It is a tree classifier in which internal nodes contain features of datasets, branches represent decision rules, and each leaf node represents the result. The Decision tree contains two nodes: the Decision Node and the Leaf Node. Decision nodes are used to make decisions and have various branches, whereas Leaf nodes represent the results of those decisions and do not have any further branches.

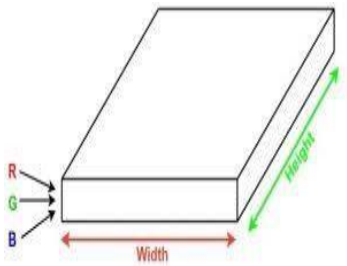
The decisions or tests are based on the characteristics of the provided dataset. It is a graphical representation of all possible solutions to a problem/decision based on certain conditions. It is labeled as a decision tree because, similar to a tree, it begins with the root node and then branches out to form a tree-like structure. The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to form a tree. A decision tree simply asks a question and based on the answer (Yes/No), it further splits the tree into sub-indices.



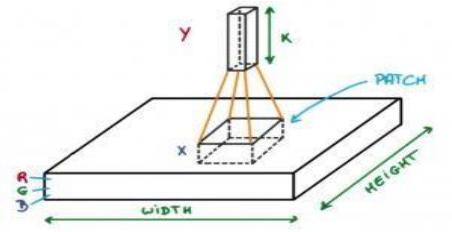
In [3], There are 5 diseases for each disease there is a dataset. The experimental results for Breast Cancer prediction using Decision Tree have achieved an accuracy of 91.81%, for heart disease prediction has achieved an accuracy of 76.92%, for kidney disease prediction has achieved an accuracy of 97.50%, for diabetes prediction has achieved an accuracy of 82.89% and for liver disease prediction has achieved an accuracy of 62.94%. In [4], the results obtained by the model using a decision tree have achieved an accuracy of 93.85. In [11], The decision tree consistently outperformed the naive Bayes and support vector machine models, so we fine-tuned it for optimal performance for predicting the probability of heart disease in diabetic patients because the decision tree achieved an accuracy of 90% whereas the naïve bayes and support vector machine models achieved an accuracy of 77.5% and 87.9%. In [12], the decision tree has achieved an accuracy of 95.12% for the disease dataset. In [13], provides support for multiple sickness prediction treatments with completely different Machine Learning algorithms like Decision Trees. In [15], The highest accuracy among all these algorithms was demonstrated by Decision Tree which was 97%. In [16], Decision Tree has achieved an accuracy of 79%.

H. CNN

Convolutional Neural Networks, often called convnet, are neural networks that share parameters. Assume you have a picture. It may be represented as a cuboid with length, width (image dimension), and height (as images generally have red, green, and blue channels).



Consider taking a small region of this image and running a small neural network with, say, k outputs on it, and representing it vertically. Slide the neural network across the whole image, and we'll produce a new image with varied width, height, and depth.
Instead of having the R, G, and B channels, we now have more channels but with a smaller width and height. This operation is called Convolution. If the patch size is the same as the picture size, the neural network is a normal neural network. We have fewer weights as a result of this small patch.



In [1], the accuracy obtained by using CNN is 99.8%.

I. Linear regression

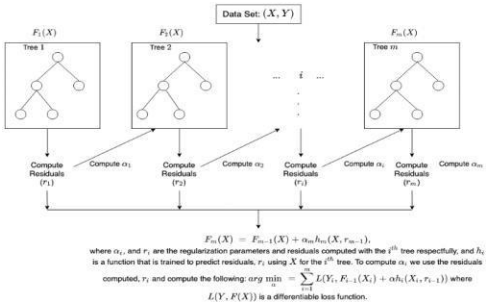
Linear regression is a form of supervised machine learning. It tries to apply relationships that will predict the result of an event based on data points from independent variables. The relation is often a straight line that matches the various data points as closely as feasible. The result is continuous, i.e., a numerical value. For example, the output might be revenue or sales in currency, the number of products sold, and so on. The independent variable in the above example might be single or multiple. Linear Regression is so called because it represents a linear connection between a dependent (y) variable and one or more independent (y) variables. This means it determines how the value of the dependent variable varies when the value of the independent variable changes. A straight line with a slope connects the independent and dependent variables.

In [16], Linear Regression has achieved an accuracy of 78%.

J. AdaBoost classifier

XG Boost produces new models regularly and combines them into an ensemble model. Build the first model and compute the error for each observation in the dataset. Then you create a new model that will predict those residuals (errors).

The prediction from this model is then added to the ensemble of models.
XG Boost is preferable to gradient boosting algorithms because it provides a good balance of bias and variation (Gradient boosting is only optimized for the variance so tends to overfit training data while XG Boost offers regularization terms that can improve model generalization).



In [18], we employed the machine learning classifier algorithms on the Wisconsin Breast Cancer (original) datasets and try to compare the efficiency and effectiveness of those algorithms to find the best classification accuracy, where the XGBOOST classifier is giving us the maximum accuracy.

k. Recurrent neural network (RNN)

A Recurring Neural Network (RNN) is a type of artificial neural network that works with temporal series or sequential data. These deep learning techniques are often used for ordinal or temporal problems like language translation, natural language processing (NLP), speech recognition, and image captioning; they are used in popular apps like Siri, voice search, and Google Translate. Recurrent neural networks, like feedforward and convolutional neural networks (CNNs), recurring neural networks use training data to learn. They are identified by their "memory," which allows them to impact the current input and output by using information from previous inputs. While traditional deep neural networks assume that inputs and outputs are independent of one another, the output of recurrent neural networks is dependent on the sequence's prior elements. While future events would also be useful in predicting the outcome of a given sequence, unidirectional recurrent neural networks cannot account for them in their predictions.

In [19], To extract features from unstructured facts RNN algorithm can be used.

5. Conclusion

This important project aims to identify an illness from its symptoms. It is believed that applying ML technology to disease diagnostics is advantageous.
Early diagnosis and therapy are advantageous to the patients. The system is designed so that the device takes the user's symptoms as input and outputs a disease prediction. Using the xg-boost algorithm, a forecast accuracy probability of 99.9% is typically achieved.

References

[1] Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. EktaaMeshram, "Designing Disease Prediction Model, Us Machine Learning Approach", IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

[2] Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, Prof. Abhijeet Karve, "Disease Prediction using Machine Learning", e-ISSN: 2395-0056 p-ISSN: 2395-0072 [2019]

[3] Anika Shreya Pawar, "Disease Prediction using Machine Learning", ISSN No: -2456-2165

[4] Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A.D. Prajapati, "Disease Prediction using Machine Learning", e-ISSN: 2395-0056 p-ISSN: 2395-0072

[5] Raja Krishnamoorthi, Shubham Joshi, Piyush Kumar Shukla, Hatim Z. Almarzouki, Ali Rizwan, C. Kalpana, and Basant Tiwari "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Volume 2022, Article ID 1684017, 10 pages.

[6] Chaube, N., Mansuri, F., Sonkar, K., & Singh, S. Disease Prediction using Machine Learning.

[7] Md. Ehtisham Farooqui, Dr. Jameel Ahmad, "Disease Prediction System using Support Vector Machine and Multilinear Regression", ISSN: 2347-5552, Volume- 8, Issue- 4, July- 2020

[8] Marouane FethiFerjani, "Disease Prediction Using Machine Learning"DOI:10.13140/RG.2.2.18279.47521

[9] Kriti Gandhi, Mansi Mittal, Neha Gupta, ShafaliDhall "Disease Prediction using Machine Learning", ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429

[10] RinkalKeniya, Aman Khakharia, Vruddhi Shah, VrushabhGada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, NinadMehendale "Disease prediction from various symptoms using Machine Learning".

[11] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2021). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings.

[12] Kumar, A., & Pathak, M. A. (2021). A machine learning model for early prediction of multiple diseases to cure lives. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(6), 4013-4023.

[13] Godse, R. A., Gunjal, S. S., Jagtap, K. A., Mahamuni, N. S., & Wankhade, S. (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively. International Journal of Advanced Research in Computer and Communication Engineering, 8(12), 50-52.

[14] Takke, K., Bhaijee, R., Singh, A., & Patil, M. A. Medical Disease Prediction using Machine Learning Algorithms.

[15] Patil, K., Pawar, S., Sandhyan, P., & Kundale, J. (2022). Multiple Disease Prognostication Based on Symptoms Using Machine Learning Techniques. In ITM Web of Conferences (Vol. 44, p. 03008). EDP Sciences.

[16] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

[17] Le, H. M., Tran, T. D., & Van Tran, L. A. N. G. (2018). Automatic heart disease prediction using feature selection and data mining techniques. Journal of Computer Science and Cybernetics, 34(1), 33-48.

[18] Sinha, N. K., Khulal, M., Gurung, M., & Lal, A. (2020). Developing a web-based system for breast cancer prediction using the xgboost classifier. International Journal of Engineering Research Technology (IJERT), 9.

[19] Jadhav, S., Kasar, R., Lade, N., Patil, M., & Kolte, S. (2019). Disease prediction by machine learning from healthcare communities. International Journal of Scientific Research in Science and Technology, 5, 8869-8869