JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Implementation of Random Forest, CNN, XGBOOST Algorithm of Machine Learning to Develop A System that can Detect the Human Disease Such as Alzheimer, Brain Tumor, Breast Cancer, Covid-19 & Heart Disease

**Akshay Parate[1], Sudershan Dolli[2], Niraj Patil[3],Kartik Chaudhari[4]**

Department of E&TC, SKNCOE, SPPU, Pune

Abstract— **Now a day one of the most significant subjects of society is human healthcare. It is looking for the best one and robust disease diagnosis to get the care they need as soon as possible. Other fields, such as statistics and computer science, are needed for the health aspect of searching since this recognition is often complicated. The task is challenging these disciplines, moving beyond the conventional ones. The actual number of new techniques makes it possible to provide a broad overview that avoids particular aspects. To this end, we suggest a systematic analysis of human diseases such as Alzheimer, Breast Cancer, Brain Tumor, Covid-19 & Heart Disease detection using machine learning algorithms such as CNN, XGBoost, Random Forest and one of the deep learning algorithm i.e. VGG-16 in order to make important predictions and help in decision-making**.

Keywords— **Human Healthcare, Random Forest, Recognition, CNN, XGBOOST, Systematic Analysis, Prediction, VGG-16**

## I. INTRODUCTION

Machine learning (ML) is used practically everywhere, from cutting-edge technology (such as mobile phones, computers, and robotics) to health care (i.e., disease diagnosis, safety). ML is gaining popularity in various fields, including disease diagnosis in health care. Many researchers and practitioners illustrate the promise of machine-learning-based disease diagnosis (MLBDD), which is inexpensive and time-efficient. Traditional diagnosis processes are costly, time-consuming, and often require human intervention. While the individual's ability restricts traditional diagnosis techniques, ML-based systems have no such limitations, and machines do not get exhausted as humans do. As a result, a method to diagnose disease with outnumbered patients' unexpected presence in health care may be developed. To create MLBDD systems, health care data such as images (i.e., X-ray, MRI) and tabular data (i.e., patients' conditions, age, and genders) are employed. Artificial intelligence can support providers in a variety of patient care and smart healthcare systems. Artificial intelligence techniques, from machine learning to deep learning, are widely used in healthcare for disease diagnosis, drug discovery, and patient risk identification. Full diagnosis of disease using artificial intelligence techniques such as ultrasound, magnetic resonance imaging, mammography, genomics, and computed tomography requires numerous medical data sources. In addition, artificial intelligence has largely improved the clinic experience and accelerated the readiness of patients to continue treatment rehabilitation at home. This article describes a comprehensive investigation based on artificial intelligence technology for diagnosing numerous diseases such as Alzheimer's disease, breast cancer, Covid-19 brain tumors, and chronic heart disease. Results are also compared using various quality parameters such as prediction rate, precision, sensitivity, specificity, precision of area under the curve, recall and F1 score, based on studies of different articles on disease diagnosis.

**This system is projected to attain subsequent goals:**

- To discover patterns in the user data and then make predictions on theses and intricate patterns for answering.
- To effectively predict if the patients suffers from the disease like Alzheimer, Breast Cancer, Brain Tumor, Covid-19 and Heart Disease.
- To make early decision for planning proper treatment and ensuring the well-being of patient.

## II. LITERATURE REVIEW

MIN CHEN et al, [1] proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, and Decision Tree. This proposed system had an accuracy of 94.8%.

Sayali Ambekar et al, [2] recommended Disease Risk Prediction and used a convolution neural network to perform the task. In this paper machine learning techniques like CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system uses structured data to be trained and its accuracy reaches 82% and achieved by using Naïve Bayes.

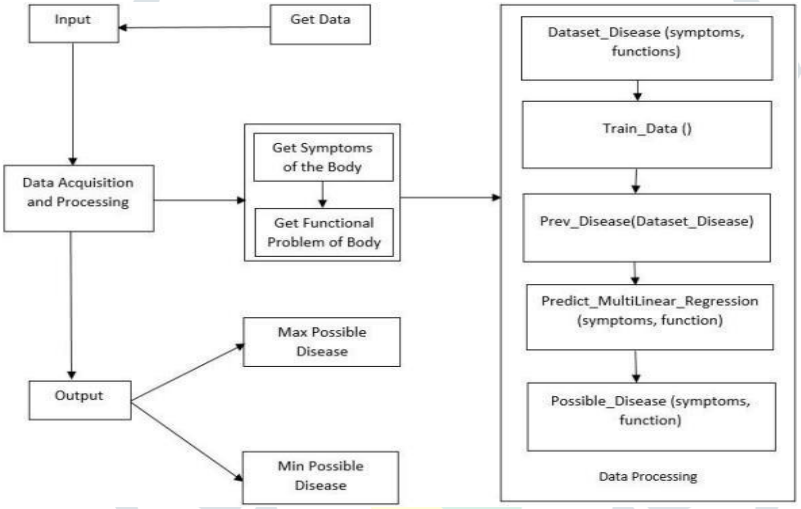
Naganna Chetty et al, [3] developed a system that gives improved results for disease prediction and used a fuzzy approach. And used techniques like KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. In this paper diabetes disease and liver, disorder prediction is done and the accuracy of Diabetes is 97.02% and Liver disorder is 96.13.

Dhiraj dahiwade et al, [4] designed a model for prediction of the disease using approaches of machine learning and used techniques like knn and cnn. This paper suggests disease prediction i.e. based on patient's symptoms. The accuracy of knn is 95% and the accuracy of cnn is 98%.

Lambodar Jena et al, [5] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.

Dhomse Kanchan B. et al, [6] studied special disease prediction utilizing principal component analysis using machine learning algorithms involving techniques like Naive Bayes classification, Decision Tree, and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for Heart disease

## BLOCK DIAGRAM




## IV. BLOCK DIAGRAM DESCRIPTION

**1. Input**: We are taking input from the user of the disease prediction system as a symptoms list.
**2. Get Data**: In this field, the user will provide data about their symptoms.
**3. Data Acquisition and Processing**: This field provides input for editing. Data Collection and Processing that performs two operations. The first collects the data, the second processing processes the data and extracts information based on the collected data.
**4. Get Symptoms of the Body**: It is a field that collects and analyzes physical symptoms. This information is used by the algorithm to help predict possible illnesses. Finding Physical Function Problems This panel collects physical function problems related to symptoms. This analyzes to get the possible diseases.

**5. Dataset Disease (symptoms, functions**): In this section has a predefined disease data set containing symptoms and features caused by diseases. This record is further used to match data received from the user, and if there is a correct match the system suggests a possible illness.
**6. Train Data ():** The system is trained in this area. Our disease prediction system is trained using the Random Forest Classifier and the XG Boost algorithm to solve problems related to heart disease and breast cancer disease prediction.
**7. Prev_Disease (Dataset Disease):** In this field, a disease record is provided as a parameter and processing is performed based on this record.
**8. Prediction (symptoms, function):** For this field, prediction is performed using a random forest classifier, the XG Boost algorithm, Symptoms and their functions in the user's body are involved in predictions.
**9. Possible Disease (symptoms, function)**: In this field the symptoms and features are passed as parameters and the possible

diseases are calculated based on these parameters.
**10. Data Processing:** This field includes the five data processing fields above and is a key part of our disease prediction system. It has all the fields you need to process your data.
**11. Output:** After Data Acquisition and Processing, possible diseases are generated as output.

## V. IMPLEMENTATION SYSTEM

The system is designed to detect diseases such as Alzheimer's disease, breast cancer, brain tumors, heart disease and Covid-19. Each disease has different signs and symptoms for patients. Various datasets are pulled from Kaggle's machine learning database to implement the disease detection system. The classification computation uses a random forest classifier algorithm in a disease detection system to detect diseases. It is a machine learning algorithm that leads to epidemic identification in disease detection with maximum accuracy, precision and recall. The Disease Detection Web App is built with Flask Framework support as a screening tool for doctors and medical professionals to easily identify patients with disease.

## VI. ALGORITHMS

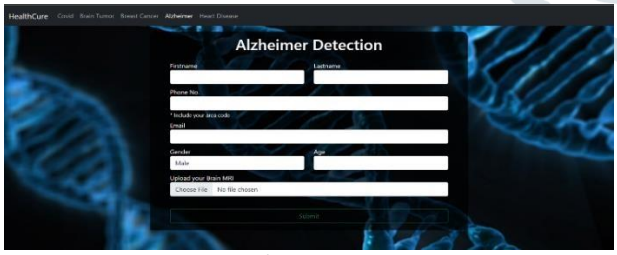
**1] XGBOOST:**
XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years.

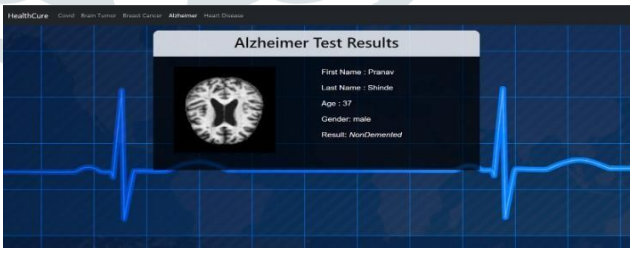
**2] Random Forest**:
Random forest is an ensemble model using bagging as the ensemble method and decision tree as the individual model. Let's take a closer look at the magic of the randomness:
Step 1: Select n (e.g. 1000) random subsets from the training set Step
2: Train n (e.g. 1000) decision trees
1. One random subset is used to train one decision tree
2. The optimal splits for each decision tree are based on a random subset of features (e.g. 10 features in total, randomly select 5 out of 10 features to split)
Step 3: Each individual tree predicts the records/candidates in the test set, independently.
Step 4: Make the final prediction for each candidate in the test set, Random Forest uses the class (e.g. cat or dog) with the majority vote as this candidate's final prediction.

**3] Convolutional Neural Networks**:
A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex.A ConvNet is able to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters.The role of the ConvNet is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction

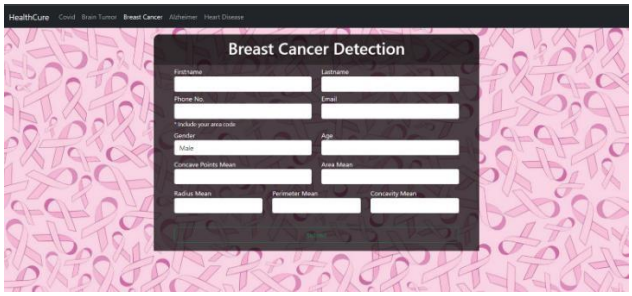
## VII. RESULT

**1 Alzheimer Detection**:
We use CNN (Sequential) model for the Alzheimer disease detection, as we give the input in the form of image, model can predict the outcome either Demented or Non-Demented.



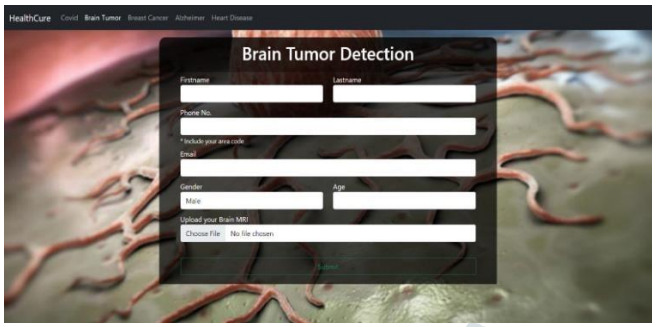

**2 Breast Cancer Detection:**
We use Random Forest model for the Breast Cancer disease detection, as we give the input model can predict the outcome either Benign or Malignant.

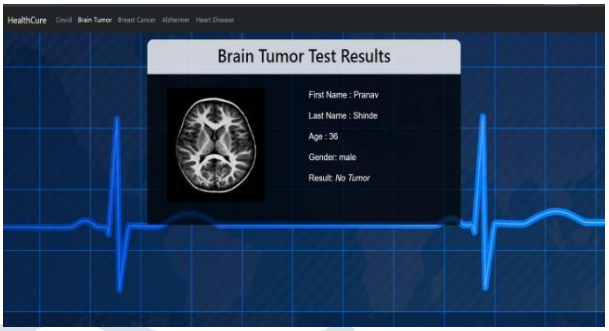




**3 Brain Tumour Detection:**

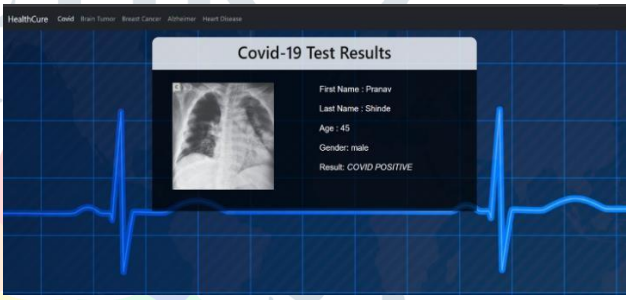
We use CNN VGG-16 model for the Brain Tumour detection, as we give the input in the form of image, model can predict the outcome either Tumour or No Tumour.



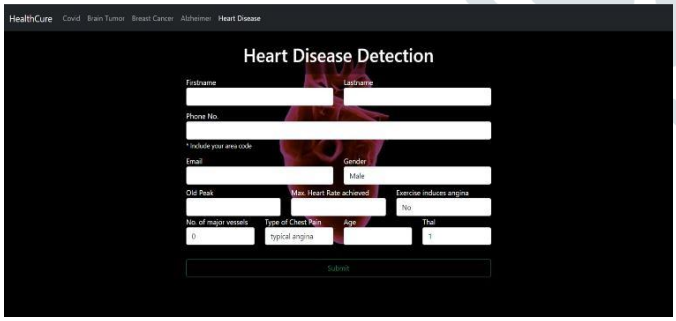

**4 Covid-19 Detection:**

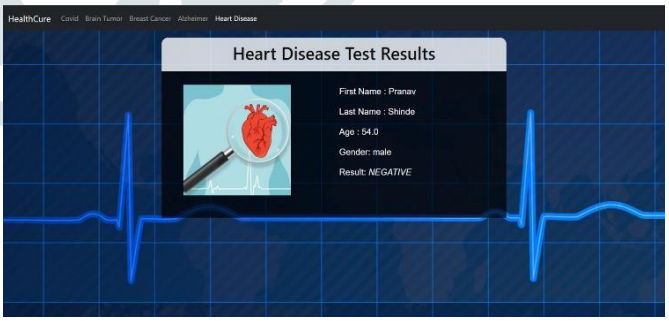
We use CNN model for the Covid-19 disease detection, as we give the input in the form of image, model can predict the outcome either Covid Positive or Covid Negative.


**5 Heart Disease Detection:**

We use XGBOOST model for the Heart disease detection, as we give the input, model can predict the outcome either Positive or Negative.



| Disease | Alternate Models | Alternate Model Accuracy | Final Model Use | Model Accuracy |
|---|---|---|---|---|
| Covid19 | CNN (Sequential) | 93.13% | CNN (Sequential) | 93.13% |
| | Functional API | 91.25% | | |
| | Data Augmentation | 89.51% | | |
| Brain Tumor | CNN,VGG16 | 95.23% | CNN,VGG16 | 95.23% |
| | Functional API | 94.18% | | |
| | Data Augmentation | 92.31% | | |
| Breast Cancer | Random Forest | 94.15% | Random Forest | 94.15% |
| | Decision Tree | 91.22% | | |
| | KNeighbors Classifier | 84.21% | | |
| Alzheimer | CNN (Sequential) | 73.54% | CNN (Sequential) | 73.54% |
| | Functional API | 72.52% | | |
| | Data Augmentation | 69.47% | | |
| Heart disease | XGBoost | 86.96% | XGBoost | 86.96% |
| | Decisioin Tree | 67.39% | | |
| | KNeighbors Classifier | 65.22% | | |

## VIII. CONCLUSION

Based on the various algorithm implementations, we have concluded that algorithms such as CNN, Random Forest, XGBoost and VGG-16 are highly useful for disease diagnosis on the given data set. We got accuracies for Alzheimer is 73.54%, Breast Cancer 94.15%, Brain Tumor 95%, Covid-19 93%, Heart Disease 86.96%. Based on the objective, we can conclude that all goals have been met, with the disease being predicted based on the input symptoms, using multiple algorithms.

## IX. REFERENCES

[1] ] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities” IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.
[2] Sayali Ambekar, Rashmi Phalnikar, “Disease Risk Prediction by Using Convolutional Neural Network” IEEE, 978-1-5386-5257-2/18, 2018.
[3] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, “An Improved Method for Disease Prediction using Fuzzy Approach” IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015.
[4] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, “Designing Disease Prediction Model Using Machine Learning Approach” IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019.
[5] Lambodar Jena and Ramakrushna Swain, “Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers” IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.
[6] Dhomse Kanchan B. and Mahale Kishor M., “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis” IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2016.
[7] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Disease Prediction” IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.
[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, ” Diabetes Disease Prediction Using Data Mining” IEEE, 978-1-5090-3294-5/17, 2017.
[9] Rashmi G Saboji and Prem Kumar Ramesh,“A Scalable Solution for Heart Disease Prediction using Classification Mining Technique” IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.
[10]Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.