

Advanced Machine Learning Techniques for Alzheimer's Disease Detection: A PCA and Improved XGBoost Approach

¹Jala Shilpa, ²N. SatheeshKumar

¹Research scholar, Dept.of Computer Science and Engineering, Chaitanya Deemed to be University

²Professor, Dept. of Computer Science and Engineering, Malla Reddy College of Engineering and Technology, Hyderabad.

KEYWORDS

Machine learning, Alzheimer's Disease, Random ForestXGBoost, CNN

ABSTRACT

Alzheimer's Disease (AD) is one of the main fields in clinical medicine that contributes to the existing difficulties in research. The presented work is dedicated to the Machine Learning approach to the identification and detection of the Alzheimer's Disease, including the Image Enhancement techniques. The work of they also use Principal Component Analysis (PCA) together with the contemporary approach in improving the image quality of the brain images obtained from public databases. The focus of this research is the enhanced XGBoost classification model as applied with the help of two other classification methods to assure its effectiveness. A lot of tests were performed on the Alzheimer's Disease dataset with an analytical feature extraction procedure to enhance the model results. These proposed methodologies are tested against conventional algorithms with an emphasis on accuracy, precision, recall and F1-score. The first estimates suggest an increase in the level of AD detection accuracy and its superiority over conventional approaches. Apart from showing a correlation between PCA and new pre-processing methodologies, this study also underscores the enhanced diagnostic aptitude of the improved XGBoost classifier.

1. INTRODUCTION

Alzheimer's disease is the most prevalent type of dementia and is a neurological disorder that affects a person's ability to read or write, reason, or even remember. Earlier in 2018, Patterson et al. predicted that the diseases could be affecting as many as 152 million people in 2050. According to estimates made earlier in 2018, the price of the therapy reached \$1 trillion. Another study by Prince et al. in 2015 shows that the growth spurt of dementia cases is way a lot faster in LMIC compared to hi-income countries. Although extensive research is underway, no definitive or standard diagnostic approach for Alzheimer's disease or its treatment exists. Alzheimer's disease can be made much more tolerable if diagnosed early, thus having a reasonable quality of life for the patient. Usual survey techniques used to diagnose Alzheimer's are physical, neurological and psychological tests. Blood and urine tests, supplemented by genetic testing, in conjunction with other factors such as infections, nutritional deficiencies, and organ involution [Gounden et al., 2022] The number of Alzheimer's cases also increases every year, which a recent. The heritability of dementia is normally examined by genetic tests. In connection with these advancements in technology, a solution that has found a place as a diagnostic tool is the scans of the brains that have shrunk and therefore their structure differs depending on the stage of Alzheimer's disease. This type of scan mainly targets parts of the brain that include hippocampus, cerebrum, and speech, memory, judgment and thinking areas because changes related to Alzheimer's disease are known to occur. Figure 1 shows the factors that are causes of Alzheimer disease.

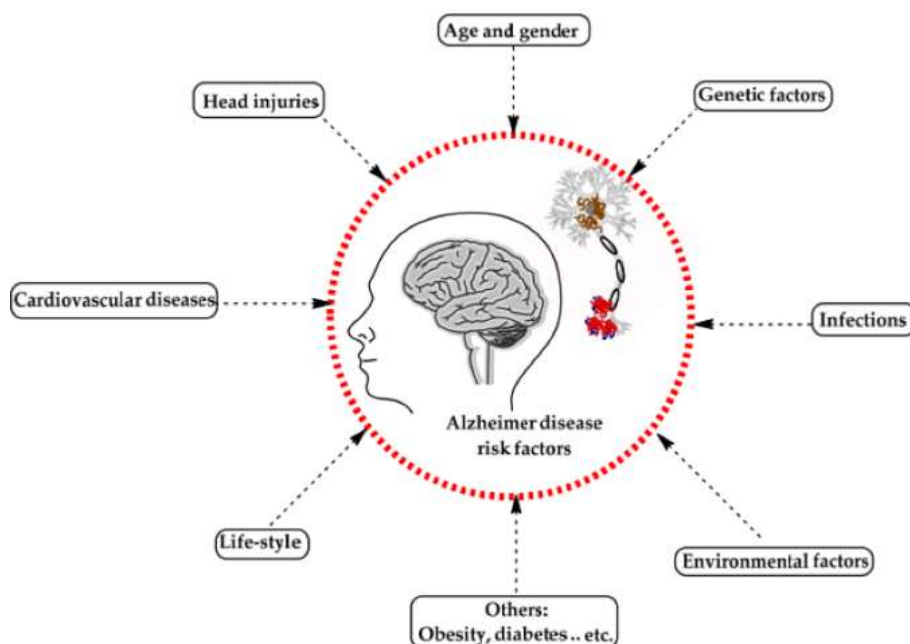


Fig.1 Factors influencing the occurrence of Alzheimer's disease [Breijyeh et al.2020]

Figure 1 shows that various factors have contributed to the development of the disease. These factors include cardio-vascular disease, brain or cranial surgery, genetic influences, infections, age, sex, environmental factors, and underlying conditions such as diabetes. In certain brain regions impacted by Alzheimer's disease, neuronal cell bodies contain both globular protein aggregates and fibrous protein structures.

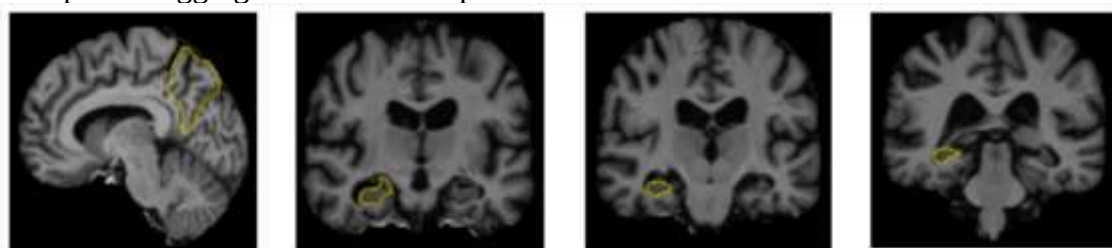


Fig.2 Areas impacted by Alzheimer's disease in brain MRI images [Choi et al.2020]

We can also refer to Figure 2, which illustrates the altered brain tissue associated with Alzheimer's disease. This is evident in regions of the brain where memory is impaired and where neurons are affected by plaques. To study Alzheimer's disease using brain MRI images, machine learning and deep learning techniques are primarily used. Most of these methods focus on classifying images or specific regions within them.

I. REVIEW OF LITERATURE

There are numerous potential benefits of an initial and exact analysis of AD for patients, society as a whole [Zeng et al., 2023]. One key advantage of detecting AD at the prodromal stage is the possibility of initiating treatment or interference with disease-changing therapies during medical trials, prior to the development of full dementia symptoms. Although this remains speculative due to the lack of effective therapies currently available, early diagnosis could still enhance the quality of life for patients. Many families experience relief upon receiving a diagnosis [Helaly et al., 2022]. Additionally, an early diagnosis may help patients gain better access to support services and care pathways while enabling them to plan for the future.

Several studies examining the advantages and challenges of earlier diagnosis of AD were identified over a literature review. While some included participants in the pre-dementia phase of AD, none specifically addressed diagnosis at the prodromal stage. One study

surveyed the psychological responses of patients and their matches upon obtaining a dementia identity. In contrast, three others concentrated on the stances of doctors or custodians rather than patients [Goenka et al., 2022]. These studies contained a survey of caregivers and primary care physicians (PCPs) regarding the scented risks and benefits of earlier dementia detection, research examining general practitioners' (GP) perspectives toward analysing dementia to comprehend low diagnosis rates, and a randomized managed trial evaluating the impact of an academic intervention on dementia diagnosis and control in primary care [Pruthviraja et al., 2023]. Two studies focused on memory clinic patients explored the impact of early diagnosis on institutionalization and mortality rates. Additionally, three studies utilized simulation models to evaluate the potential cost benefits of early judgment for dementia or Alzheimer's disease. The same authors, Kaabouch et al. (2019), probed a CNN-based standard that recorded a accuracy rate of 94.54% in categorizing both EMCI and LMCI subtypes. These results align with existing literature that used other deep learning methodologies for clinical image and text categorisation. This has been possible owing to the use of machine learning structures within CNNs that are project-specific. This study involved an analysis of health participants and participants suffering from EMCI and LMCI. However, this study faced some limitations which include; The research respondents were only 600 in total, and 200 each from each institution and this limited my analysis.

Mild cognitive impairment (MCI) patients at risk from getting Alzheimer's disease were identified in a study by Spasov et al. (2019). This approach was aimed at subtyping MCI and a differentiation between Alzheimer's disease and normal cerebral aging (NC). Consequently, the model is highly flexible in terms of imaging methods, including PET scans. Using CNN based approach, the technique was able to handle challenge of 3D imaging data input. The work also included a multimodal feature extractor, and the performance was evaluated by a 10-fold cross-validation approach. Further, the research responded to the problems such as overfitting in machine learning algorithms and the scarcity of medical data.

II. PROPOSED METHODOLOGY

The proposed methodology for detecting Alzheimer's Disease (AD) using machine learning algorithms consists of three main steps:

1. Rescaling of the image vectors
2. Classification through a combination of three machine learning algorithms: XGBoost, Random forest, Convolutional Neural Network (CNN).
3. Model validation

Every step taken aims at improving the accuracy of detecting AD based on the potential of various algorithm and sophisticated methods. The architecture of the proposed work is shown in the following figure.

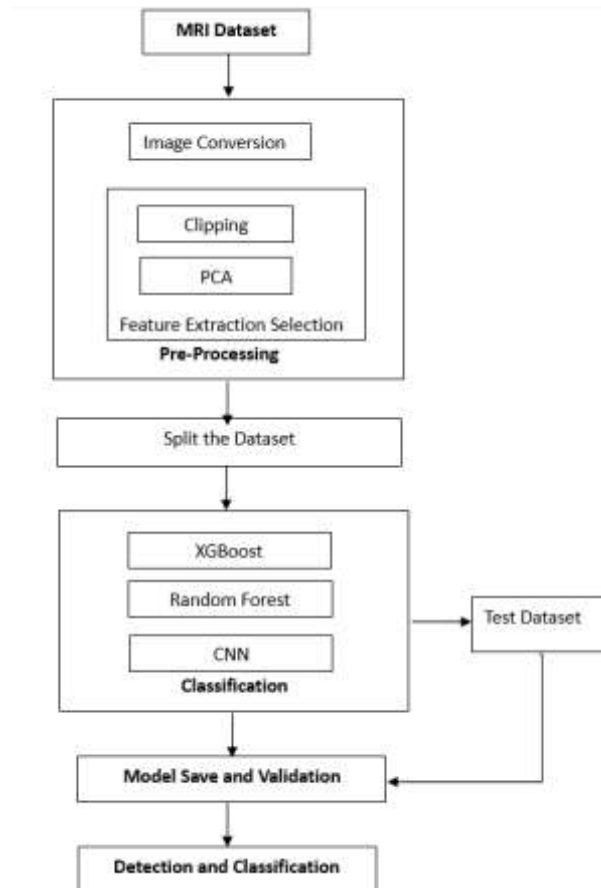


Fig.3 Proposed architecture

A) Image Pre-processing

Enhancing the quality of the Alzheimer's Disease images data allows for more effective classification. Pre-processing reduces dimensionality, eliminates noise, and improves feature extraction.

Steps in Pre-processing:

Data Acquisition:

The Kaggle dataset consists of approximately 5,000 MRI images, divided into four classes: Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented. These images are already split into training and testing sets, with a ratio of 70:30.

Data Cleaning:

Review the dataset to remove any corrupted, mislabelled, or low-quality images that might affect model performance.

Image Resizing:

Resize all images to a standardized size, such as 224x224 pixels, to ensure consistency, as CNNs generally require a fixed-size input.

Normalization:

Normalize pixel intensity values to fall within a range (e.g., [0, 1]) to reduce the influence of varying illumination conditions and ensure uniformity across the dataset.

Data Augmentation:

augmentation techniques like rotation, flipping, zooming, randomness cropping and others should be used to artificially enlarge the training dataset. This approach allows minimizing overfit, thus improving the generality of a model.

Dimensionality Reduction using PCA :

PCA aids in lowering the dimensionality of the dataset by concentrating on the most important features. For this 2D MRI dataset, PCA can minimize computational complexity by identifying and focusing on the most informative features.

PCA Algorithm steps

Step 1: Flatten the 2D images into 1D vectors.

Step 2: Standardize these vectors by subtracting the mean and scaling to unit variance.

Step 3: Evaluate the covariance matrix of the standardized data.

Step 4: Evaluate eigenvalues and eigenvectors from the covariance matrix.

Step 5: Select the top k eigenvectors connected to the highest eigenvalues to form a new feature space.

Step 6: Transform the original image data into this reduced feature space.

B) CLASSIFICATION

To classify the pre-processed MRI images into four AD's categories using three machine learning techniques: XGBoost, Random Forest, and CNN.

a) Feature Extraction using XGBoost

In the context of detecting and classifying AD using brain MRI images, XGBoost can be effectively employed to classify these images into different phases of dementia. While Convolutional Neural Networks (CNNs) excel at capturing spatial hierarchies and extracting visual features from images, XGBoost is particularly effective in utilizing these extracted features—such as pixel intensity distributions, texture features, or principal component analysis (PCA) components—to achieve accurate classification. When combined with image preprocessing techniques, XGBoost can effectively identify various stages of Alzheimer's Disease by learning patterns from these extracted features.

Feature Extraction:

Before applying XGBoost for image classification, it is crucial to extract relevant features from MRI images. These may include PCA components, statistical measures like mean and standard deviation of pixel intensities, texture features such as Haralick textures or Gabor filters, or even features from the early layers of a CNN. These features are then fed into the XGBoost model as input data.

Initialization:

XGBoost begins by initializing a base model, often a simple decision tree or decision stump. This model may predict a continuous value, like the mean of the target variable in the training data.

Iterative Boosting Process:

In each iteration, XGBoost adds a new decision tree to the ensemble, aiming to correct the remaining errors (the differences between predicted and actual values) of the existing trees in the model.

Calculating Residuals:

For each training instance, XGBoost calculates the residuals or errors from the current model's predictions. The newly added tree is then trained to predict these residuals, improving the accuracy of the overall model by addressing the remaining errors.

Residual Calculation:

$$r_i = y_i - \hat{y}_i$$

Where:

r_i is the residual for the i -th example

y_i is the actual target value for the i^{th} example

$\hat{y}_i$ is the predicted value from the current ensemble of trees.

Gradient and Hessian Calculation:

- XGBoost leverages both the first derivative (gradient) and the second derivative (Hessian) of the loss function, allowing for more accurate updates to the model. This approach offers more precise adjustments compared to other boosting algorithms that rely solely on the gradient.

Gradient (g_i) and Hessian (h_i) calculation

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \quad g_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$$

Where:

L is the loss function.

Tree Construction:

A new decision tree is created to track the negative gradient of the loss function, with separately separated in the tree chosen to optimize the gain, which measures the reduction in residual errors.

Gain Calculation:

$$Gain = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma$$

Where:

- G_L and G_R are the sums of gradients for the left and right child nodes.
- H_L and H_R are the sums of Hessians for the left and right child nodes.
- λ is the regularization parameter.
- γ is the minimum loss reduction required to make a split.

Final Prediction:

$$\hat{y} = \sum_{m=1}^M \alpha_m f_m(x)$$

Where:

- M is the total number of trees in the ensemble.
- α_m is the learning rate (shrinkage parameter).
- $f_m(x)$ is the prediction from the m -th tree

Softmax Function for Multi-Class Classification:

In multi-class classification problems, such as Alzheimer's Disease detection, the final output from XGBoost is processed through a softmax function, which transforms the raw scores into class probabilities.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

Where,

z_j is the score for class j

K is the total number of classes

b) Random Forest (RF):

RF provides significant advantages due to its capability to handle high-dimensional feature spaces while being resistant to overfitting. By aggregating predictions from multiple decision trees, RF can effectively classify different stages of dementia, yielding reliable results. Moreover, its built-in feature importance scores help identify the most significant factors influencing the classification, offering insights into patterns associated with various stages of Alzheimer's disease.

Feature Extraction:

As with other algorithms, the process begins by extracting key features from the brain MRI images. These features can include pixel intensity, texture, shape, or more advanced features derived from PCA or CNNs, which are then used as input for the Random Forest model.

Bootstrapping:

The training data is sampled multiple times with replacement to generate several bootstrapped datasets. Each dataset is used to train an individual decision tree within the Random Forest, promoting diversity among the trees—a key factor in enhancing the ensemble's effectiveness.

Random Subset of Features:

At each decision tree node, Random Forest randomly chooses a subcategory of features and picks the best one for splitting the data. This randomization reduces correlation between the trees, boosting the overall model's performance.

Node Split:

$$Gini_Impurity(S) = 1 - \sum_{i=1}^C p_i^2$$

Where,

S is the subset of data at a node.

p_i is the probability of a sample belonging to class i in subset S

C is the number of classes.

Majority Voting

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$$

Where

$\hat{y}_n$ is the prediction from the n th decision tree

n is the total number of trees in the forest.

c) Feature Extraction using CNN (Convolutional Neural Network):

Description: CNNs are particularly effective for image classification tasks, such as MRI scan images, because they can automatically learn the spatial hierarchy of features.

CNN Algorithm:

In the field of AD detection and classification, CNNs are utilized to automatically learn and extract important features from brain MRI scans. This capability allows for the differentiation between various stages of dementia. CNNs are particularly well-suited for this task because of their hierarchical feature extraction abilities, which excel at identifying subtle patterns and textures in brain MRI images that indicate different stages of AD.

CNN Works in Brain Image Classification

Input Layer:

MRI images, typically grayscale, are input into the CNN as single-channel images. These images undergo pre-processing, such as resizing and normalization, to ensure consistency across the dataset. The input image, I (a grayscale MRI image), is then processed by the CNN in subsequent steps.

Convolutional Layers:

The early convolutional layers of the CNN focus on detecting basic features, like edges and corners in the MRI images. As the network deepens, later layers learn more complex features that represent intricate brain structures and regions associated with dementia.

The input image I is convolved with a set of filters K , producing a feature map F

$$F_k(i, j) = \sum_m \sum_n I(m, n) \cdot K_k(i - m, j - n)$$

Here, F_k is the feature map produced by the k -th filter

Activation using ReLU:

The ReLU activation function is applied to the feature map F_k to introduce non-linearity

$$A_k(i, j) = \max(0, F_k(i, j))$$

Where, A_k represents the activated feature map.

Pooling Layers:

Pooling layers are employed following each convolutional layer to downsample the feature maps. This process decreases the number of parameters and reduces computational requirements. As a result, it enhances the robustness of feature detection, making it less sensitive to variations in the input images.

For instance, a pooling operation like max pooling is applied to the activated feature map A_k to lower its dimensionality:

$$P_k(i, j) = \max_{(m, n) \in \text{patch}_{m, n}} A_k(m, n)$$

Where, P_k is the pooled feature map

Feature Hierarchies:

As the CNN processes through successive convolutional and pooling layers, it develops a hierarchical understanding of brain MRI images, evolving from detecting basic features to recognizing more intricate structures associated with Alzheimer's disease.

Fully Connected Layers:

In the end of this network the fully connected layers make use of all the features extracted to recognize the input image. This stage determines which of the four categories the image belongs to: Non-demented, Very mildly demented, Mildly demented or Moderately demented.

The last layers of the network involve a number of convolutional and pooling layers before the feature maps are flattened into a feature vector. This vector is then passed through the fully connected layers as output of this layer is a classification.:

$$z = W \cdot x + b$$

Where,

x is the flattened feature vector

W is the weight matrix

b is the bias vector

z is the output of the fully connected layer

Softmax for Classification:

The output z is passed through a softmax function to produce the probability distribution over the four classes:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

Where $\sigma(z)_j$ is the predicted probability for class j .

By training the CNN on labeled MRI images, the model learns the patterns and features that distinguish between different stages of Alzheimer's Disease, enabling accurate classification of new, unseen images

Output Layer:

The last layer in the CNN uses softmax activation function in order to provide probabilities for each of the four classes. Selecting the class with the highest possibility is the preceding shown category in the forecast of the input MRI image.

Process:

Input the normalized and resized MRI images into the CNN.

Apply a series of convolutional layers with various filters to extract hierarchical features, including edges, textures, and patterns related to different stages of dementia.

III. RESULTS AND DISCUSSIONS

We have used the GPUs available on Kaggle to train our model. We experimented with a variety of epochs to achieve the best possible results. In this part, we evaluate and analyze the overall performance of our suggested method using the Alzheimer's Dataset, which was obtained from a publicly available dataset on Kaggle.

DATASET

The dataset is collected from the publicly available Kaggle repository. It includes data that has been hand-collected from various websites, with each label thoroughly tested. The dataset comprises 5,000 MRI images and contains four classes of images, which are present in both the training and testing sets:

- Mild Demented
- Moderate Demented
- Non-Demented
- Very Mild Demented

Metric of evaluation

The proposed detection approach in performance domains is evaluated using a widely used metric that includes accuracy, sensitivity, specificity, and precision. These measurements calculated

using

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)

- A True Positive (TP) occurs when the model correctly identifies a patient as having Alzheimer's disease.
- A True Negative (TN) occurs when the model correctly identifies a patient as not having any form of dementia (i.e., non-demented).
- A False Positive (FP) occurs when the model mistakenly classifies a patient as having Alzheimer's disease (or a specific type of dementia) when they are actually non-demented.
- A False Negative (FN) occurs when the model fails to detect Alzheimer's disease (or a specific type of dementia) in a patient who indeed has the condition.

Accuracy measures the overall correctness of a model by considering both true positives (TP) and true negatives (TN).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

Precision indicates the proportion of correct identifications among all positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

Recall, also known as Sensitivity or True Positive Rate, measures how effectively the model identifies positive cases.

$$Recall = \frac{TP}{TP + FN}$$

The **F1-Score** is the harmonic mean of Precision and Recall, offering a balance between the two metrics:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Table.1 The performance of proposed work in various metrics

	Precision	Recall	f1-score	support
NonDemented	0.96	1.00	0.98	639
VeryMildDemented	1.00	1.00	1.00	635
MildDemented	0.93	0.89	0.91	662
ModerateDemented	0.90	0.91	0.90	624

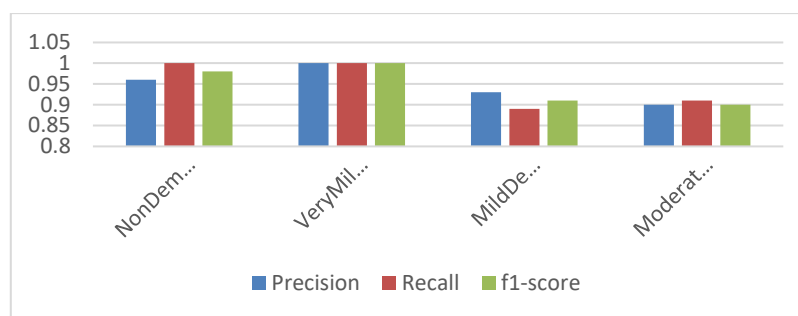


Fig.4 Proposed work Performance in the scenario of Precision, Recall, and F1-score

As illustrated in Figure 4 above, the X-axis represents the performance metrics: The P, R, and F1-Score for each of the Alzheimer's Disease categories, including Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented. On the Y-axis is shown performance of the values of these metrics given is a percentage hence ranging from 0.90 to 1.00 for each class..

Proposed Model Accuracy

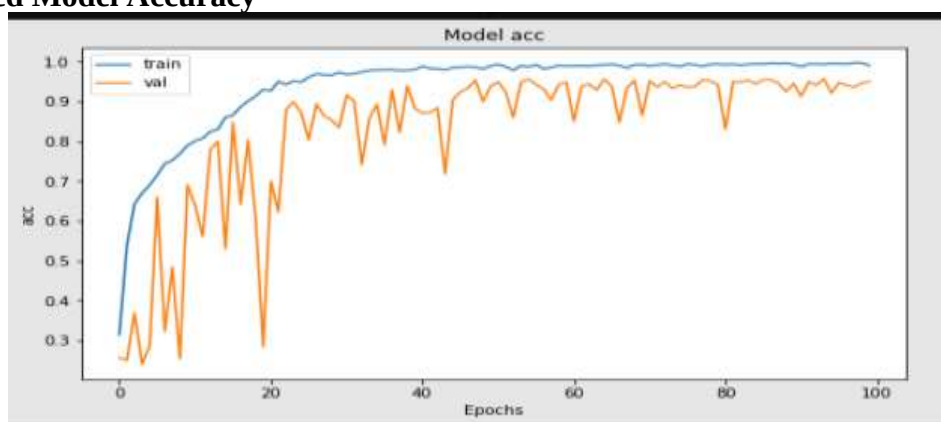


Fig 5. Model Training Accuracy

In Figure 5, the X-axis represents the training epochs, which range from 0 to 100. Each epoch corresponds to one complete pass through the training dataset during the model's training process. The Y-axis reflects the accuracy of the method, showing the percentage of accurate predictions made by the model throughout the training. By the end of the training, the model achieved an accuracy of 94.84%, demonstrating its capability to accurately classify cases of Alzheimer's Disease.

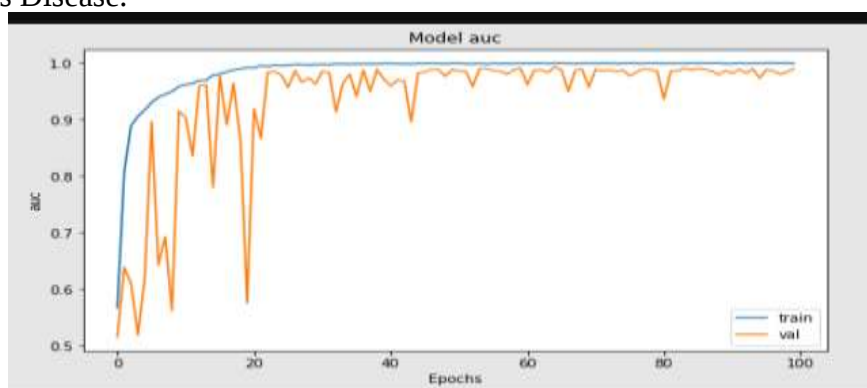


Fig 6: Model validation Accuracy

the X-axis represents the epochs, which range from 0 to 100, indicating the number of complete passes through the dataset during training. The Y-axis signifies the proof accuracy, which measures how well the model achieves on unseen validation data at each epoch. As training progresses through each epoch, the validation accuracy improves, indicating that the model is effectively generalizing beyond the training data. The final validation accuracy reaches 94.84%, demonstrating that the model performs consistently well on both the training and validation sets.

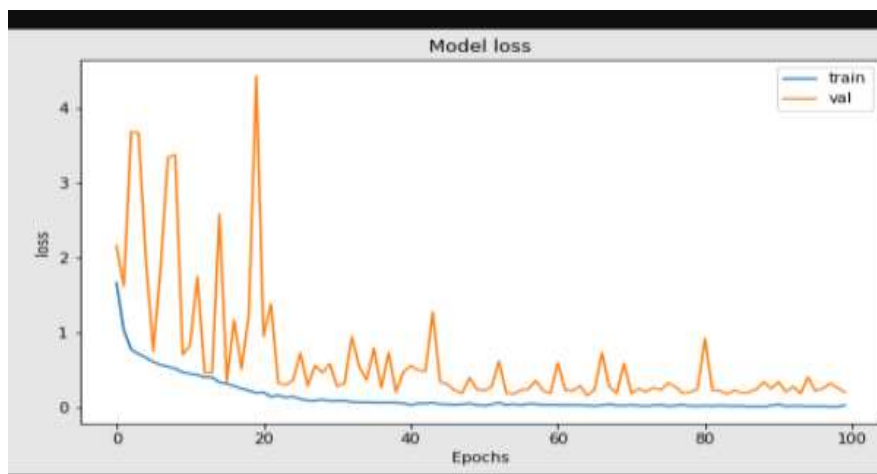


Fig.7 Model Loss

In Figure 7, the X-axis denotes the epochs, which range from 0 to 100. What is more, each epoch means going through all the training data for making one iteration through it. Y-axis is model loss, which measures the efficiency of the model, and the lower the losses the better is the model.

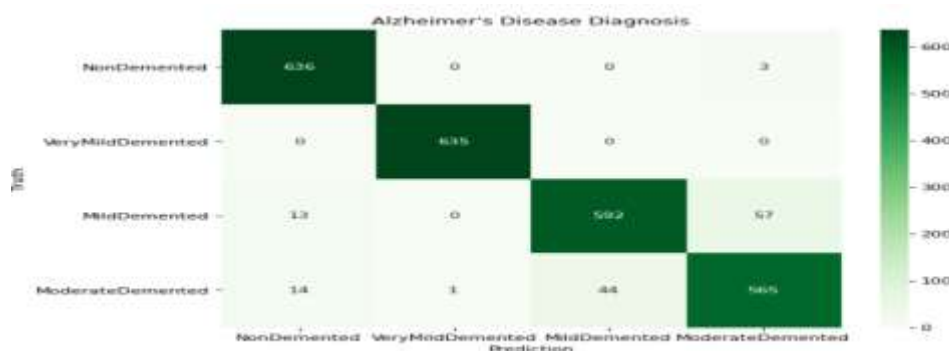


Fig.8 Confusion matrix

The confusion matrix depicted in figure 8 further distinguishes the capacity of the model categorizing the four types of the Alzheimer's disease. As for the both Figures, the True Labels are represented at the Y-axis, and the Predicted Labels at the X-axis. In this regard the four categories are non-demented, very mildly demented, mildly demented and moderately demented.

IV. CONCLUSION

In this paper, we described a machine learning system for detection and classification of Alzheimer's disease using Magnetic Resonance Imaging images of the brain from Kaggle data set. This research used RF, improved XGBoost, and CNN, as well as PCA preprocessing to mitigate the problem of data dimensions and to optimize feature extraction from two-dimensional brain MRI images. The performances of these models were assessed using conventional metrics of classification, specifically accuracy, precision, recall, and F1-score, the best accuracy score obtained was 94.84% by the CNN model. Our proposed methodology showed that integrating highly accurate classifiers with the most beneficial preprocessing methodologies can enhance the identification and differentiation of the different phases of Alzheimer's Disease. This is especially necessary to differentiate between non-demented elderly, very mild, mild, moderate dementia states. Compared with the empirical results for Our proposed CNN model presented a higher accuracy than XGBoost and Random Forest and its precisions, recalls and F1-score indicated its appropriateness in dealing with complex medical image taxonomy settings.

REFERENCES

1. C.Patterson, 2018, "The State of the Art of Dementia Research: New Frontiers; Alzheimer's Disease".
2. Prince, M and Wimo, A, 2015, "The global impact of dementia: An analysis of prevalence, incidence, cost, and trends".
3. Gounden, V, I. "Renal Function Tests; StatPearls Publishing: Treasure Island", 2023. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK507821/> (accessed on 21 December 2022)
4. N. Zeng and Y. Peng,2023, "A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease", pp. 11599-11610.
5. R. Karaman and Z. Breijyeh, 2020, "Comprehensive Review on Alzheimer's Disease: Causes and Treatment".
6. H.A. Helaly and A.Y. Haikal, 2022, "Deep learning approach for early detection of alzheimer's disease", pp. 1711-1727.
7. N. Goenka, A. Goenka, S. 2022, "Tiwari Patch-based classification for alzheimer disease using sMRI", Institute of Electrical and Electronics Engineers Inc. (2022), pp. 1-5.
8. D. Pruthviraja and N.A. Alkhaldi, 2023, "Detection of alzheimer's disease based on cloud-based deep learning paradigm Diagnostics", p. 2687
9. S. Savaş, 2022, "Detecting the stages of Alzheimer's disease with pre-trained deep learning architectures "Arabian J Sci Eng, pp. 2201-2218
10. Choi, J.Y.; Lee, B, 2020, "MRI Segmentation and Classification of Human Brain Using Deep Learning for Diagnosis of Alzheimer's Disease: A Survey".
11. H.T. Gorji and N. Kaabouch, 2019, "A deep learning approach for diagnosis of mild cognitive impairment based on MRI images Brain" p. 217.
12. S. Spasov, A. Duggento, 2019, "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease", pp. 276-287.
13. M. Liu, Shen, et al. "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease" Neuroimage, 208 (2020), Article 116459