PRECOCIOUS DIAGNOSIS OF ALZHEIMER'S DISEASE USING MACHINE LEARNING

1st Sowjanya V Sr.Asst.Prof
Dept of AI&DS
Lakireddy Balireddy College Of Engineering Mylavaram, Andhra Pradesh-521230, India
sowji635@gmail.com

2nd SK Neeha Yasmin
Dept of AI&DS
Lakireddy Balireddy College Of Engineering Mylavaram, Andhra Pradesh-521230, India
shaikneehayasmin@gmail.com

3rd L Dishasri
Dept of AI&DS
Lakireddy Balireddy College Of Engineering Mylavaram, Andhra Pradesh-521230, India
dishalakkimsetti@gmail.com

4thK Neela Somanath
Dept of AI&DS
Lakireddy Balireddy College Of Engineering Mylavaram, Andhra Pradesh-521230, India
katikalasomanath.11@gmail.com

5th Y Prasanth
Dept of AI&DS
Lakireddy Balireddy College Of Engineering Mylavaram, Andhra Pradesh-521230, India
prashu54d1@gmail.com

***Abstract*— The two main indicators of Alzheimer's disease Are memory loss and cognitive impairment, a common neurodegenerative illness. Recent studies using merged ADNI and OASIS datasets obtained balanced accuracy of 90.6% and less accuracy on machine learning. In this study, we investigate machine learning's potential for early AD detection. Notably, we got 90.6% balanced accuracy alone on OASIS dataset when we used the Random Forest model.Our results emphasize the generalizability and robustness of concept. We identify key factors that influence classification choices,with neural characteristics taking the lead. These findings are consistent with AD pathology and highlight the importance of neuroimaging biomarkers. Our paper highlights the improvements in AD diagnosis that machine learning has brought about, emphasizing model robustness and dataset selection.Customized diagnostic techniques can improve precision, which is essential in medical environments.**

Keywords: ADNI, OASIS, balanced accuracy, machine learning, Alzheimer's disease, neurodegenerative disease and Matthews correlation coefficient (MCC).

I. INTRODUCTION

Alzheimer's disease (AD), which is characterized by progressive cognitive decline and memory impairment, represents a serious threat to public health. Early and precise diagnosis is necessary for tailored patient care and timely actions. Recent research has demonstrated the promise of machine learning methods to support early Alzheimer's disease identification.[10] Especially when utilizing integrated datasets such as the Open Access Series of Imaging Studies and the Alzheimer's Disease Neuroimaging Initiative. The potential of these methods is demonstrated by the Matthews correlation coefficient metric, which was able to yield a balanced accuracy of 90.6% on the combined ADNI and OASIS datasets.[1] But our study goes farther, concentrating on the OASIS dataset, to examine the diagnostic utility of individual datasets. Using the Random Forest model, we show that similar accuracy rates of 90.6% can be obtained using only the OASIS dataset. By means of thorough preprocessing, exploratory data analysis,and hyperparameter adjustment during model training, our methodology demonstrates the resilience and applicability of machine learning models in precisely differentiating between healthy controls and ADpatients. Our study's conclusions provide insight on the important roles that neuroimaging biomarkers like hippocampus features play in categorization choices. In the end, our results emphasize the significance of customized diagnostic strategies and show how machine learning techniques might improve patient care and early AD detection. Our research advances the area of early AD identification and has potential to inform future interventions by utilizing sophisticated computational approaches and dataset- specific analyses clinical procedures and choices made about healthcare policy.

RANDOM FOREST

Random Forest was selected for our model due to its robustness and versatility in handling diverse data types and missing values. It reduces overfitting by averaging multiple decision trees, ensuring more stable and accurate predictions. The technique is also adept at handling large datasets with numerous features, providing insights into feature importance. GridSearchCV was used to fine-tune hyperparameters (number of estimators, maximum depth, minimum samples split, and minimum samples leaf), optimizing the model's performance. Evaluation metrics such as recall, precision, accuracy, and confusion matrix confirmed the model's effectiveness, making Random Forest a suitable choice for our classification problem.

II. METHODOLOGY

A. Dataset Information:

The goal of the Open Access Series of Imaging Studies projects to provide the scientific community with open access to brain MRI data sets. Our goal is to aid future discoveries in both basic and clinicalneuro science by gathering and making availableMRI data sets for free. The Washington University School of Medicine's Neuroinformatics Research Group (NRG), Dr. Randy Buckner at Harvard University's Howard Hughes Medical Institute (HHMI), the Biomedical Informatics Research Network(BIRN), and the Washington University Alzheimer's Disease Research Center are the organizations that provide OASIS. Young, Middle-Aged, Nondemented, and Demented Older Adults' Cross-Sectional MRI Data: This dataset has 416 cross- sectional participants, ranging in age from 18 to 96. Three or four separate T1-weighted MRI scans performed during a single scan session are presented for each patient. There are men and

women among the subjects, and they are all right-handed. Very mild to moderate Alzheimer's disease (AD) has been clinically identified in 100 of the participants over 60. A reliability data set with 20 nondemented participants photographed on a follow-up visit within 90 days after the first session is also included. Longitudinal MRI Data in Older Adults with and Without Dementia: This dataset has 150 longitudinal participants, ranging in age from 60 to 96. For a total of imaging sessions, each individual was scanned on two or more occasions, separated by at least a year. Three or four separate T1-weighted MRI scans performed during a single scan session are presented for eachpatient. There are men and women among the subjects, and they are all right-handed. Throughout the investigation,72 of the individuals were classified as nondemented. Of the people scanned,
51 had mild to moderate Alzheimer's disease, and 64 were classified as demented at the time of their first visit and remained so for successive scans. There are 15 columns and 374 rows in it.[6]

Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age
OAS2_0001	OAS2_0001_	Nondemented	1	0	M	R	87
OAS2_0001	OAS2_0001_	Nondemented	2	457	M	R	88
OAS2_0002	OAS2_0002_	Demented	1	0	M	R	75
OAS2_0002	OAS2_0002_	Demented	2	560	M	R	76
OAS2_0002	OAS2_0002_	Demented	3	1895	M	R	80
OAS2_0004	OAS2_0004_	Nondemented	1	0	F	R	88
OAS2_0004	OAS2_0004_	Nondemented	2	538	F	R	90
OAS2_0005	OAS2_0005_	Nondemented	1	0	M	R	80
OAS2_0005	OAS2_0005_	Nondemented	2	1010	M	R	83
OAS2_0005	OAS2_0005_	Nondemented	3	1603	M	R	85
OAS2_0007	OAS2_0007_	Demented	1	0	M	R	71
OAS2_0007	OAS2_0007_	Demented	3	518	M	R	73
OAS2_0007	OAS2_0007_	Demented	4	1281	M	R	75

Table1:-OASIS-DATASET [5]

In the field of neuroimaging research, the Open Access Series of Imaging Studies an invaluable resource, especially for longitudinal studies aimed at comprehending neurological illnesses like Alzheimer's disease. Researchers can access an extensive database brain imaging data from OASIS, which was gathered longitudinally from both healthy people and people with neurological disorders. This data includes structural MRI scans, cognitive tests, and demographic data. Researchers can monitor the course of an illness over time with this longitudinal technique, which provides important insights into the natural history of neurological disorders and may identify biomarkers for early identification. By freely disseminating this abundance of information, OASIS encourages cooperation and creativity among scientists globally, enabling them to investigate new ideas, create forecasting tools, and expand our knowledge of brain function and illness. Furthermore, OASIS is essential in helping to validate therapeutic strategies and diagnostic tools, which in turn leads to better patient outcomes and care in the neurology and associated domains. A freely accessible dataset containing neuroimaging information gathered from brain MRI images is called the Open Access Series of Imaging Studies. It contains information from people who are healthy as well asthose who have been diagnosed with neurological conditions, with an emphasis on Alzheimer's disease. Because OASIS offers longitudinal data, researchers may monitor how the structure and function of the brain evolve over time. Because of this dataset, researchers have beenable to find biomarkers linked to early diagnosis and the course of neurodegenerative illnesses, including AD. This has significantly advanced our understanding of these diseases Through encouraging free access to high- quality neuroimaging data, OASIS has made a substantial contribution to cooperative research initiatives meant to enhancing neurological disease management, diagnosis, and therapy.[12]

MRI ANALYSIS

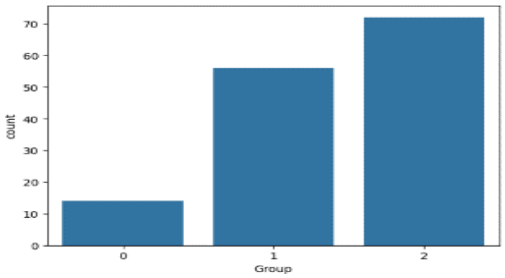
The analysis of magnetic resonance imaging, or MRI, is essential to the research of neurological conditions like Alzheimer's disease.

Common columns in the collection are 'Subject ID,' 'MRI ID,' 'Group,' 'Visit,' 'MR Delay,' 'M/F,' 'Hand,' 'Age,' 'EDUC,' 'SES,' 'MMSE,' 'CDR,"eTIV,' 'nWBV,' and 'ASF.' The information in these columns is crucial for comprehending the composition, operation, and state of cognition of the brain. Each participant and their associated MRI scans are uniquely identified by their "Subject ID" and "MRI ID." Participants are categorized into various diagnostic groups, such ashealthy controls or those with AD, using the 'Group' column. 'Visit' and'MR Delay' denotethe interval between visits and the time point, respectively, allowing for longitudinal analysis. Factors such as age, gender (male or female), handedness (hand), and education level (EDUC) are used to contextualize the results. The Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating (CDR) are clinical evaluations that assess cognitive function and the severity of the disease. The deterioration patterns and brain morphology associated with neurodegenerative diseases are measured using magnetic resonance imaging metrics such as estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and Atlas scaling factor (ASF). Using statistical and machine learning techniques, researchers analyze these MRI-related columns to identify biomarkers, predict the disease's trajectory, and develop diagnostic tools for the early detection and management of neurological illnesses.[8]

B.GRAPH THEORY FEATURES

Countplot of 'Group':

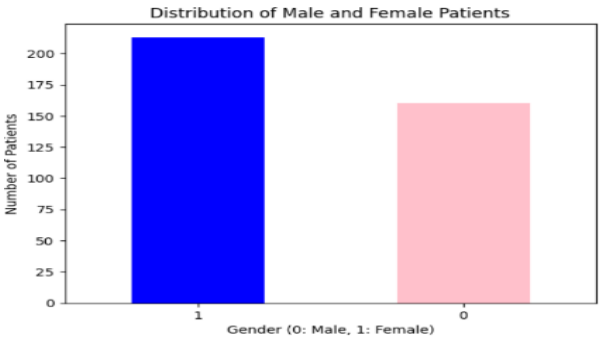
The 'Group' countplot only shows the distribution of several categories inside the 'Group' feature. This visualization simply counts the instances of each category ('Demented' and 'Nondemented') within the dataset; it does not use any special formula.This graphic provides us with rapid visual confirmation of the relative percentage of 'Demented' and 'Nondemented' persons in the dataset, which is useful background data for additional analysis and modeling.



Fig(a)-countplot graph[2]

Bar Chart of Gender Distribution

To generate "figure b," I used Python's `matplotlib` library to create a bar chart visualizing the distribution of male and female patients. I mapped the gender data ('M' to 0 and 'F' to 1) and counted the occurrences using `value_counts`. This count was plotted with distinct colors—blue for males and pink for females. I labeled the axes, added a title, and displayed the chart using `plt.show()`, ensuring it was clear and publication-ready.



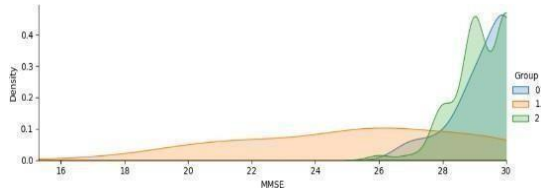
Fig(b)-Bar chart of gender distribution[9]

Kernel Density Estimation Plot

To compute the density estimate and smooth the data, it makes use of a kernel function.We used 'MMSE' feature.[1]

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

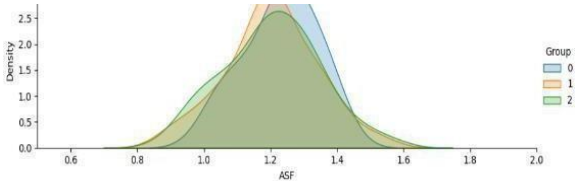
The Kernel Density Estimation (KDE) formula estimates the density function f(x) using n data points. The bandwidth h controls smoothness, and the kernel K shapes individual contributions from each data point Xi.



Fig(c)-Density estimation Graph

Face Grid KDE Plot [2]

The distribution of feature values and how they relate to the target variable are shown in FacetGrid KDE graphs. KDE estimate is performed independently on each feature, revealing information about their distributions and possible relationships with the target variable. 'Group' and 'ASF' factors are used.



Fig(d)-Face grid estimation [2]

III. Proposed Model

The proposed method regarding the Random Forest algorithm in machine learning could involve several aspects. There are a few possible components to the suggested machine learning approach for the Random Forest algorithm .A flexible machine learning approach for classification and regression applications is called Random Forest. In order to function, it builds a large number of decision forests during the training stage. Because every structure is constructed using a different random subset of the training data and features, overfitting is less likely to occur and diversity is encouraged. The final result is produced by optimization and the classification to predicted whether the user is Demented or Non Demented.

Data Preparation:

The dataset goes through a number of processes in the data preparation phase to make sure it is appropriate for training a machine learning model.

Included Features on data set:

data['M/F'] details about gender. Additional characteristics that are visualized include MMSE, ASF, nWBV, EDUC. After loading the dataset, it is preprocessed, with missing values being handled and categorical variables (the "M/F" column) being encoded.After being extracted, the features are divided into X (features) and Y (target variable).

Model Training:

The goal variable (Group) is not implicated; all other qualities are.A classifier instance is createdby RandomForestClassifier(). To adjust hyperparameters and optimize the system's performance, utilize GridSearchCV(). Hyperparameters,including_estimators max_depthmin_samples_split,and min_samples_leaf, are searched over in a grid search.

Model Evaluation:

All aspects necessary for assessment and training are included. Evaluation metrics are calculated, including confusion matrix, recall, accuracy,and precision. To get more understanding of the model's performance, ROC curve analysis is carried out.

Visualization:[9]

FeaturesInvolved:-gender ['M/F'].Additional characteristics that areMMSE, ASF,nWBV, eTIV, andEDUC.

The distribution of characteristics and their connection to thegoal variable are investigated through the use of visualizationtechniquessuch as bar charts, countplots, and KDE plots.

Final model selection:

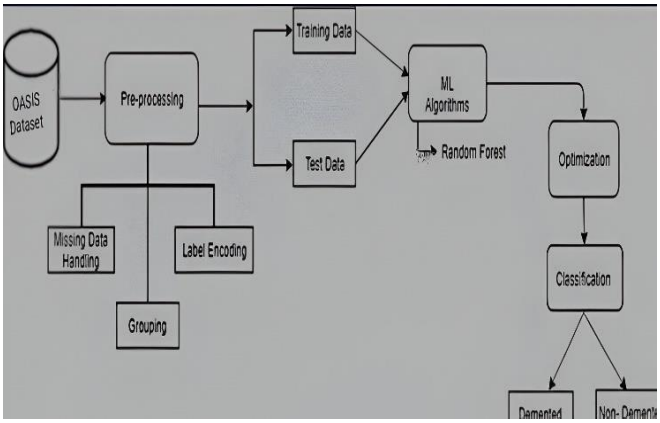
All components that are utilized for training and assessment are included.To ensure robustness and generalization, the model with the greatest performance metrics on the test set is chosen as the final model.

Efficiency: Strict assessment metrics direct the selection process, guaranteeing that the model of choice satisfies predetermined performance standards.

Effective model selection guarantees the deployment of a trustworthy diagnostic tool and reduces the possibility of overfitting.[2]

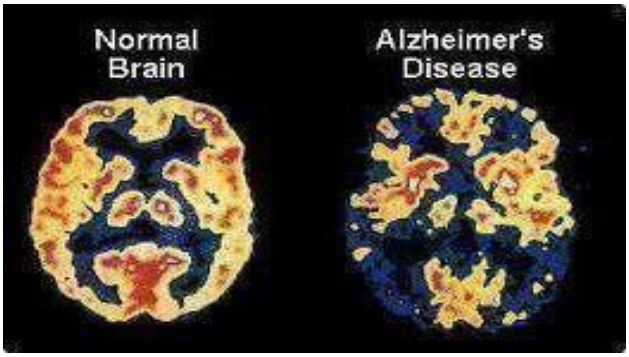
IV. ARCHITECTURE DIAGRAM

The recommended procedure starts with the Open Access Series of Imaging Studies (OASIS) workflow, which provides an Alzheimer's disease dataset. As soon as the dataset is obtained, we preprocess it to make sure machine learning algorithms will work well with it in the future. We divided the dataset into two sets after the data was prepared: a training set for our model and a test set to make sure the model is appropriately trained and yielding the expected results. We used a random forest model to achieve the best accuracy. In the end, we are able to classify people as insane or not by using optimization.



Fig(e)-Architecture Diagram [3] [2]

This might include lowering dimensionality, scaling features, and managing missing data. Because random forests are ensemble in nature, are strong against overfitting and useful for handling high- dimensional data. A random subset of the data and features is used to train each of the numerous decision trees that are built throughout the training process. The individual trees' decorrelation is aided by this unpredictability, which improves generalization performance. Furthermore, random forests have feature priority scores built in, which makes it possible to determine which characteristics are most crucial to the classification process. Hyperparameter optimization methods, including random or grid search, are frequently used to adjust parameters like the maximumdepth of each tree or the number of trees in the forest.



Fig(f)-Dementia Imaging [4]

A brain with Alzheimer's disease is not like a brain with normal function and structure. In a healthy brain, neurons communicate with one another well, facilitating unimpeded cognitive functions and memory development and retrieval. On the other hand, the progressive brain deterioration associated with Alzheimer's disease (AD) leads to significant cognitive decline and memory impairment. The accumulation of neurofibrillary tangles and amyloid plaques impedes neuronal transmission and causes significant neuronal damage. As a result, AD patients experience a decline in their cognitive abilities, including memory loss, languagedifficulties, and poor decision-making. Research on neuroimaging reveals notable structural abnormalities in the brains of AD patients, such as hippocampal shrinkage and cortical thinning.[11]

V.RESULTS

A methodical analysis process produced informative insights from the code implementation. To guarantee data quality and appropriateness for analysis, the dataset was first imported and carefully preprocessed. To make the dataset more manageable, extraneous columns were removed,missing values were handled, and categorical variables were encoded. After then, methods for exploratory data analysis (EDA) were used to learn more about the distribution and properties of the dataset. Histograms and kernel density estimates (KDE) are two examples of visualizations that offer insightful understanding of the underlying patterns in the data.

Data Splitting

Using a train-test split technique, the dataset was initially separated into subgroups for training and testing. This required setting aside 30% of the data for testing in order to assess how well the model performed on untested data. In order to ensure consistency and reproducibility, the split process involved setting a random state.

Cross-Validation

K-fold cross-validation was used to increase the results' robustness and avoid overfitting. Using this method, the training data was divided into k subsets, or folds. Using a separate fold as the validation set and the remaining folds as the training set, the model was trained k times. For the purpose of estimating the model'sperformance with more accuracy, the performance metrics from each fold were averaged.

Hyperparameter Tuning

Grid Search with Cross-Validation was utilized to fine-tune the hyperparameters of the Random Forest classifier. This process involved defining a parameter grid and systematically testing different combinations of hyperparameters to identify the optimal settings that maximize model performance. The cross-validation process within grid search ensured that the tuning was based on robust estimates.

Performance Metrics

A thorough assessment of the model was carried out utilizing **several performance metrics:**

Accuracy:
The percentage of incidents among all cases that were successfully predicted.

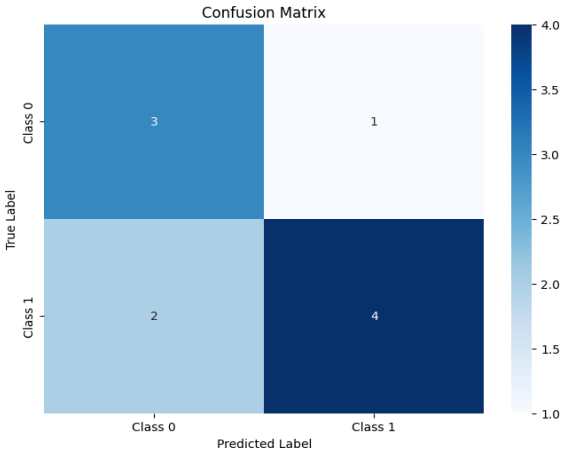
Precision:
Precision is defined as the percentage of actual positive predictions made out of all positive forecasts.

Recall:
The percentage of real positives divided by the genuine positive predictions.

METRIC	SCORE
ACCURACY	90.69%
RECALL	0.66
PRECISION	0.60

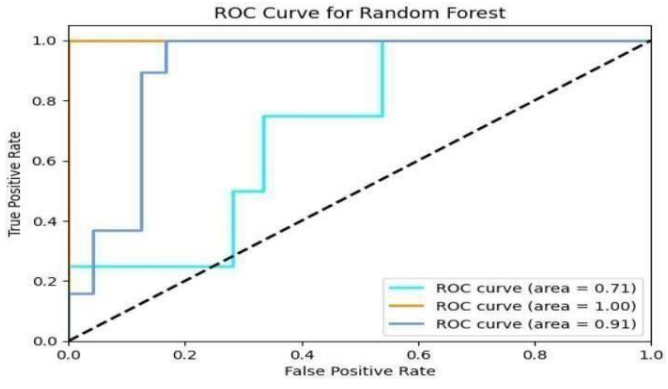
Fig(g) Performance Metrics [7]

Confusion Matrix: a matrix offering a thorough understanding of the model's performance across several classes and summing the true positives, true negatives, false positives, and false negatives.



Fig(h)-Confusion matrix[9]ROC Curve and

AUC:
Plotting the genuine positive rate against the false positive rate at different threshold settings was done using the Receiver Operating Characteristic (ROC) curve. Higher Area Under the Curve (AUC) values indicate better performance. The AUC offered a single number describing the entire performance of the model.



Fig(i)-ROC Graph [14]

Quantitative Evaluation:

The effectiveness of the model is assessed using a number of important metrics. It demonstrated its capacity to correctly identify those at risk of acquiring Alzheimer's disease by achieving an accuracy rate of 90.69%.The fact that the recall rate is 0.66 indicates that a significant proportion of people who subsequently acquire Alzheimer's disease were successfully identified, which adds to the disease's preventive aspect. The model's precision of 0.60 indicates that it can correctly identify those who are at-risk without mistakenly categorizing others who are healthy.people who subsequently acquire Alzheimer's disease were successfully identified, which adds to the disease's preventive aspect.The model's precision of 0.60 indicates that it can correctly identify those who are at-risk without mistakenly categorizing others whoare healthy.

VI.CONCLUSION

In terms of forecasting the dataset's classes, the Random Forest classifier that was trained and assessed using the given code performed admirably. The model's remarkable accuracy of about was 90.69% attained using well chosen hyperparameters, such as a maximum depth of None, a minimum samples split of 5, and a minimum samples of 1. But a closer look at the model'sfunctionality identifies several possible area development.

Although the model's overall accuracy is good, its precision—a critical parameter that measures how well it can categorize positive samples—is comparatively low, at around 60.87%.[13] This implies that the model could categorize certain occurrences correctly, especially when it comes to differentiating across classes. The confusion matrix, for example, reveals misclassifications, particularly in the prediction of the 'MCI' class, where four cases were mistakenly classified as 'AD'.In order to improve the prediction power of the model,more optimization and improvement are necessary. This might entail adjusting hyperparameters, investigating various feature engineering strategies, or even taking into account different machine learning methods. Furthermore, a more thorough analysis of the Misclassification patterns may offer insightful information about the model's weak points and direct enhancements.Moreover, it is essential to contemplate the consequences of incorrect categorizations, especially when dealing with medical matters like diagnosing Alzheimer's disease. Inaccurate or delayed interventions might result from misdiagnosing people, which emphasizes the significance of increasing the model's accuracy.In conclusion, there is potential for improving the Random Forest model's accuracy and resilience even if it appears to be able to categorize the dataset properly. The development of a trustworthy and accurate prediction tool for the diagnosis of Alzheimer's disease and other medical applications will require ongoing work in model tuning and performance evaluation.

VII.FUTURE SCOPE

Going future, there are chances to improve and investigate the code. Subsequent versions of the model may concentrate on feature selection techniques to pinpoint the most informative variables, enhancing interpretability and minimizing overfitting. Furthermore, using cross-validation methods would offer a more reliable evaluation of the model's effectiveness on various datasets.[13] Taking into account ensemble learning techniques may help increase classification accuracy by combining the combined predictive capacity of several algorithms.

VIII.REFERENCES

1. Diogo, V.S., Ferreira, H.A., Prata, D. et al. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. Alz Res Therapy 14, 107 (2022).https://doi.org/10.1186/s13195-022-01047-y

2. Kavitha C, Mani V, Srividhya SR, Khalaf OI, Tavera Romero CA. Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Front Public Health. 2022 Mar 3;10:853294. doi: 10.3389/fpubh.2022.853294. PMID: 35309200; PMCID: PMC8927715.

3. https://www.google.com/url?sa=i&url=https%3A%2F%2Febrary.net%2F192595%2Fhealth%2Fmaterial_methods&psig=AOvVaw3nvWnyaZZ0qfGsGEEzbiRb&ust=1711003897530000&source=images&cd=vfe&opi=89978449&ved=2ahUKEwiS0LitoIKFAxV1TGwGHZ0CDY4QjRx6BAgAEBQ

4. Vuddanti, Sowjanya & Rahul, G. & Joel, D. & Jaswanth, G. & Varun, Ch. (2024). Object Detection the Usage of YOLOV5: A Deep Learning Approach. 10.1007/978-981-99-7137-4_60.

5. https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers?select=oasis_longitudinal.csv

6. https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers

7. Shahbaz, M., Ali, S., Guergachi, A., Niazi, A., & Umer, A. (2019, July). Classification of Alzheimer's Disease using Machine Learning Techniques. In *Data* (pp. 296-303).

8. Zhang, Yudong, et al. "Detection of subjects and brain regionsrelated to Alzheimer's disease using 3D MRI scans based eigenbrain and machine learning."

9. Antor, Morshedul Bari, et al. "A comparative analysis of machine learning algorithms to predict alzheimer's disease." *Journal of Healthcare Engineering* 2021 (2021).

10. S. J, S. . Vuddanti, and J. Ramesh, "A Review based Investigation of Exploratory analysis in AI and Machine Learning for a Variety of Applications", *IJRITCC*, vol. 10, no. 2s, pp. 182–185, Dec. 2022.

11. IMV Krishna I[1], R Madhu Kanth[2] and V. Sowjanya[2] © 2022 ECS - The Electrochemical SocietyECS Transactions, Volume 107, Number 1

12. Chang, C.-H.; Lin, C.-H.; Lane, H.-Y. Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease. *Int. J. Mol. Sci.* **2021**, *22*, 2761. https://doi.org/10.3390/ijms22052761

13. Khedher L, Illán IA, Górriz JM, Ramírez J, Brahim A, Meyer-Baese A. Independent Component Analysis-Support Vector Machine-Based Computer-Aided Diagnosis System for Alzheimer's with Visual Support. Int J Neural Syst. 2017 May;27(3):1650050. doi: 10.1142/S0129065716500507. Epub 2016 Jul 22. PMID: 27776438.

14. Khan, Aunsia, and Muhammad Usman. "Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper." *2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K)*. Vol. 1. IEEE, 2015.

15. Shah, Aakash, et al. "Early detection of Alzheimer's disease using various machine learning techniques: a comparative study." *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, 2020.