Early Dementia Identification with Machine Learning and Deep Learning Approach

Purushottam Panta, PhD student, Department of Computer Science, University of Kentucky

***Abstract*—As individuals age normally, a reduction in whole-brain volume initiates during early adulthood and progresses more rapidly during advanced aging [5][6]. This decline involves a particular loss of gray matter volume along with specific thinning of the cortex in certain regions [7]. Factors such as level of education, gender, socioeconomic status, and cardiovascular health play significant roles in this volume decline during advanced aging. This suggests that underlying health conditions, even when not clinically apparent, contribute to alterations in brain structure associated with aging. In this study we analyze OSSIS datasets of demented and nondemented peoples. We then build machine learning approaches such as Random Forest and Support Vector Machine, Long Short-Term Memory to build an efficient model, then train and predict (classify) the subjects based on various prominent features we picked from the source dataset. We measure the performance and accuracy of the models to identify the better model for the dataset. Early identification of the potentially demented subjects can help intervene the cause of dementia progression and alleviate symptoms or slow down the dementia progression.**

I. Introduction

Dementia can broadly be described as a set of symptoms that basically affects the human cognitive abilities including memory, thinking, reasoning which directly impacts the ability to perform daily activities. Alzheimer's disease is a common neurodegenerative condition primarily impacting older adults, causing dementia, yet its exact cause and development remain uncertain. It is a specific and progressive brain disorder that leads to memory loss, cognitive decline, and behavioral changes.

The first noticeable sign of Alzheimer's Disease is specific memory loss, and although treatments can alleviate certain symptoms, there's currently no cure. Magnetic resonance imaging (MRI) is employed to assess individuals suspected of having Alzheimer's Disease (AD). MRI results reveal both localized and widespread brain tissue shrinkage, as depicted in the visual representation of brain tissue. Research indicates that these MRI characteristics might forecast the pace of AD progression and potentially direct future treatment approaches.

The proposed machine learning model helps clinical professionals to identify the subjects of likely developing AD and Dementia in the future. his method can not only improve the classification performance but also facilitate the early intervention of AD. The new in our model is that we have built the Long Short-Term model (LSTM) and taking Socioeconomic Status with MMSE score and Education in our model.

II. Literature Research on Similar work

Zhou Y. et. al. [8] in their publication have provided an overview of research exploring methods and biomarkers utilized in forecasting Alzheimer's disease (AD) progression through multimodal MRI data. It initially scrutinizes and summarizes diverse approaches, encompassing machine learning, deep learning, regression models, and various MRI analysis methodologies, all tailored toward predicting the progression of the AD.

Kevin de Silva et. el. [9] research work aims to investigate the potential use of a convolutional neural network (CNN) as a diagnostic tool in predicting Alzheimer's Disease (AD) from magnetic resonance imaging (MRI). It focuses on utilizing the MIRIAD dataset (Minimal Interval Resonance Imaging in Alzheimer's Disease) and specifically examines the feasibility of using a single central brain slice for analysis. In this study, the predictive performance for Alzheimer's Disease diagnosis using MRI yielded impressive results: a Matthew's Correlation Coefficient (MCC) of 0.77, an accuracy and F1-score of 0.89 each, and an AUC of 0.92. Training a CNN for this task required less than 30 seconds with a GPU, while executing predictions took less than 1 second on a standard PC. The performance has been measured based on ROC curve analyses.

Diego et. el [10] trained and tested their classifier on Healthy Controls (HC) vs Mild Cognitive Impairment (MCI) vs Alzheimer's Disease (AD) taken from the combination of ADNI and OASIS datasets and obtained a balanced accuracy of 90.6%. The Matthew correlation coefficient (MCC) score was obtained as being 0.811. The classification decisions were primarily influenced by hippocampal features, contributing approximately 25–45%. Temporal regions followed at around 13%, while cingulate and frontal regions each contributed approximately 8–13%. This alignment with our existing knowledge of AD and its progression underscores the significance of these brain areas. Remarkably, the classifiers demonstrated consistent performance across various datasets and protocols. Interestingly, employing graph theory measures did not enhance the classification accuracy.

III. Dataset Details

A. Dataset description

Longitudinal MRI data from an Open Access Series of Imaging Studies (OASIS) **[0]**, accessible on their website and Kaggle is the data source for the models. This data presents an opportunity to train diverse machine learning models aiming to identify patients experiencing mild to moderate dementia. This OASIS longitudinal dataset contains the Medical Resonance Imaging (MRI) data of the subjects in the age of range 60 to 90. which had been prominent in the daily work plays an important role in cognitive analysis and so all the 150 subjects in the dataset are of being right-handed. Each subject was scanned at least once. Subjects are categorized further as 72 being *nondemented* and 64 being *demented*. The rest of the 14 subjects

are categorized as *converted*. The converted subjects were initially categorized as nondemented in their initial visit, and in the later visit they were categorized as demented. So, they are considered as converted.

B. Feature selection from the source data

Feature selection helps in building more efficient, accurate, and interpretable machine learning models by focusing on the most relevant information while discarding noise and irrelevant data. We basically identify prominent features from the dataset for further process and build the machine learning models on them.

Socioeconomic Status (SES): SES are determined with the help of various socioeconomic factors such as the occupation, collar-category of the jobs **[1].**

Clinical Dementia Rating (CDR): It basically serves as a comprehensive measure aimed at assessing the overall severity of dementia. It evaluates six distinct areas separately—memory, orientation, judgment and problem-solving, community affairs, home and hobbies, and personal care. The CDR ratings range from 0 to 5 on a scale where 0 signifies absence, 0.5 indicates uncertainty, 1 represents mild presence, 2 signifies moderate, 3 implies severe, 4 indicates profound, and 5 signifies terminal conditions. The cumulative score provides a global overview, allowing the use of CDR to categorize patients based on the severity of their dementia **[2]**.

Years of education (EDUC): EDUC represents the educational achievement of respondents, gauged by the highest level of schooling or degree completed.

Mini-Mental State Examination (MMSE): The MMSE serves as a rapid and uncomplicated screening instrument for assessing five cognitive domains **[3]**: orientation, immediate memory, attention, delayed memory, and language. The combined scores from these domains yield a total score ranging between 0 and 30. To mitigate the effects of age and educational variations, conversion tables may also be taken into consideration to determine the score.

Estimated Total Intracranial Volume (eTIV): Estimated Total Intracranial Volume (eTIV) is a measurement used in neuroimaging to estimate the total volume inside the skull that contains the brain, cerebrospinal fluid, and other structures. It's an essential metric in studying brain morphology, as it helps normalize brain volumes for individual differences in head size. Researchers and clinicians use eTIV to assess brain structure variations and to account for differences in brain sizes among individuals when analyzing neuroimaging data.

Atlas Scoring Factor (ASF): It is a scaling factor used in brain imaging, particularly in voxel-based morphometry (VBM) studies. VBM is a technique that allows for the comparison of brain anatomy, typically using MRI scans, by analyzing differences in the density or volume of brain tissues.

Normalize Whole Brain Volume (nWBV): It is the measure of the volume of the whole brain. In studies analyzing brain imaging data, normalizing the brain volume helps in comparing brain sizes across different individuals, accounting for variations in head size. This normalization process allows researchers to focus on relative differences in brain structures rather than absolute sizes, enabling more accurate comparisons between groups or populations.

No funding or financial assistantship was in place for this research and literature.

IV. Methods: Data Analysis and Preprocessing

In this section we are going to explore the dataset details and the preprocessing before we train and test the dataset in the model. We now analyze all the feature variables that we are taking consideration for our model:

Gender:

The dataset can be classified based on their gender (female:0, male:1) as below and we can observe that more males are having dementia than females.

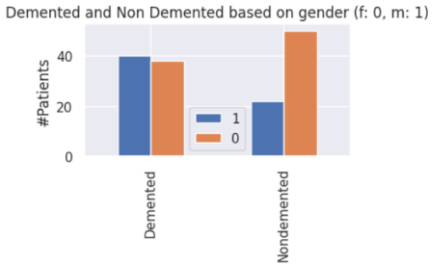


Fig1: Female (0) and Male (1) are classified based on Demented and Nondemented.

ASF, eTIV, nWBV:

By plotting the graph for ASF, eTIV and nWBV we can observe that the nondemented (0) group has higher brain volume ratio than Demented group (1). The following figures (Fig. 2, Fig. 3 and Fig. 4) illustrates that the Nondemented group exhibits a higher brain volume ratio than the Demented group. This is likely due to the diseases causing a reduction in brain tissue, leading to shrinkage.

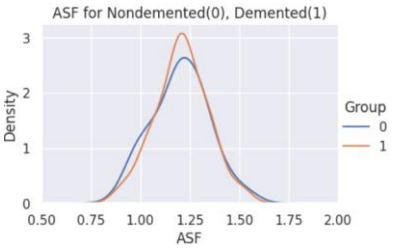


Fig2: ASF Scores distribution over Nondemented (0) and Demented (1) subjects.

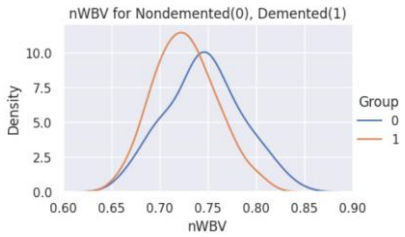


Fig3: nWBV scores distribution over Nondemented (0) and Demented (1) subjects.

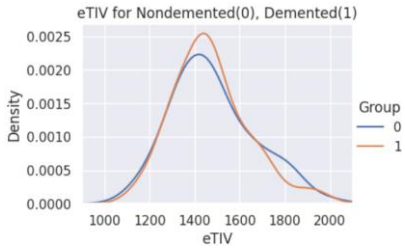


Fig4: eTIV Scores distribution over Nondemented (0) and Demented (1) subjects.

Age:

Age is a significant factor in Alzheimer's disease. While Alzheimer's can affect individuals of various ages, it primarily occurs in older adults. The risk of developing Alzheimer's increases as a person gets older. Most individuals with Alzheimer's are 65 years of age or older, and the likelihood of developing the condition doubles roughly every five years after the age of 65. However, it's essential to note that not everyone who reaches old age will develop Alzheimer's, but age is considered one of the strongest risk factors for the disease. There are cases of early-onset Alzheimer's that can appear in individuals in their 40s or 50s, but these are relatively rare compared to the late-onset form associated with aging.

As observed in the following figure (fig. 5), the Demented patient group shows a greater concentration of individuals aged 70-80 compared to the nondemented patients. This suggests that individuals affected by this disease may have a lower survival rate, which explains the fewer instances of individuals aged 90 years old within this group.

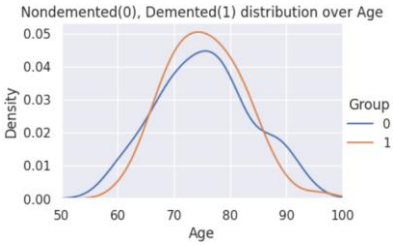


Fig5: Age feature distribution over Nondemented (0) and Demented (1) subjects.

EDUC:

Research suggests that higher levels of education may have a correlation with a reduced risk of developing Alzheimer's disease or experiencing its symptoms later in life. Engaging in lifelong learning, formal education, and intellectually stimulating activities might contribute to building cognitive reserve, which is the brain's ability to withstand neurological damage. People with more years of formal education often exhibit better cognitive abilities and have more mentally stimulating occupations or hobbies, potentially leading to a reduced risk of Alzheimer's. However, while higher education seems to correlate with a decreased risk, it doesn't guarantee immunity from the disease. It's important to note that Alzheimer's is a complex condition influenced by various factors, and while education might have a protective effect, it's just one aspect among many that can contribute to an individual's risk profile for developing the disease. We can observe the similar relationship in the fig. 6 below.

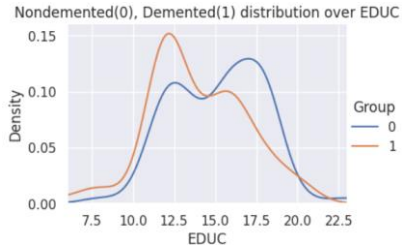


Fig6: EDUC feature distribution over Nondemented (0) and Demented (1) subjects.

MMSE Score:

We further analyzed how the MMSE score has been distributed over the demented and nondemented subjects. We can see that Demented population has much lower MMSE score than the Nondemented population.

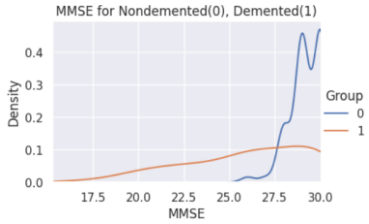


Fig7: MMSE score distribution over Nondemented (0) and Demented (1) subjects.

A. *Preprocessing*

Data preprocessing ensures that the data provided to machine learning models is well-structured, clean, and optimized for effective learning, ultimately improving the model's accuracy and performance. We basically enforce the following data preprocessing steps to improve the accuracy and quality of model outcome:

a. Data Integration and load: Integrating the data source and loading it into the python library for further processing is the initial step of the preprocess.
b. Data reduction: Eliminating the unnecessary features (or attributes) from the input dataset basically helps further processing in many ways such as by improving the processing performance, simplifying the overall model structure.
c. Data cleanup: Eliminating the rows missing values or having null attributes.
d. Data transformation (Normalization, Aggregation and Generalization)
e. Data Imputation: Since socioeconomic factors are discrete variables, we simply take median as a constant for data imputation.
f. Data visualization and analysis: We have plotted various data attributes in graph to better understand relationship between selected features and control features.

B. *Performance Measures*

Our primary performance gauge is the area under the receiver operating characteristic (ROC) curve (AUC). In medical diagnostics for non-life-threatening terminal illnesses like most neurodegenerative diseases, a high true positive rate is vital for early identification of all Alzheimer's patients. Simultaneously, it's crucial to minimize the false positive rate to avoid incorrectly diagnosing healthy adults as having dementia and initiating unnecessary medical interventions. Therefore, we've chosen AUC as an ideal performance measure to strike a balance between high sensitivity in detecting the disease early and minimizing false positives.

TABLE I. TABLE TYPE STYLES

		ACTUAL	
		Positive	***Negative***
PREDICTED	***Positive***	True Positive	False Positive
	Negative	False Negative	True Negative

True Positives (TP): data points labelled as positive that are positive. *False Positives (FP):* data points labelled as positive that are actually negative. *True Negatives (TN):* data points labelled as negative that are actually negative. *False Negatives (FN):* data points labelled as negative that are actually positive.

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

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

$$F1\ Score = \frac{TP}{TP + 1/2(FP + FN)}$$

V. Machine Learning Models

We have taken random forest and support vector machine and long short-term memory algorithms to classify the demented and nondemented subject in our dataset. We then will be evaluating the performance of each of them with AUC score and recall score.

A. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to make more accurate and robust predictions. We build the random forest model with the following outlined algorithm:

Algorithm1: Random Forest

For a tree T_i in range R, and features in F_i, and depth D_i:

Build a Random Forest Model
Compute cross validation score
Find the best cross-validation score for more accuracy.
Pick the random forest model with the best score.
Fit the model and predict.

Where, T_i is i^{th} tree in Range R = (5, 15, 20). F_i is the max number of features.

```
---Random Forest---
Based on picked parameters Test accuracy 0.8421052631578947
Accuracy on validation set: 0.7770750988142293
We picked the parameters having best performance respectively
# of Trees to combine, M:  5
Max depth of tree, m :  1
Max # of features, d :  7
Test AUC: 0.8444444444444443
Test recall: 0.8
```

B. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm used mainly for classification tasks, but it can also be employed for regression and outlier detection. Its primary purpose is to find a decision boundary (or hyperplane) that best separates data points belonging to different classes in a dataset. The best regularization parameter, c was picked out of the list (0.001 to 1000, stepping in 10x fold) to minimize misclassification during training.

The gamma parameter defines the influence range of a single training example, affecting the smoothness of the decision boundary. Specifically, in the RBF kernel, a small gamma value means a wider Gaussian kernel, implying that points farther away from the decision boundary are considered in calculating the decision. Conversely, a larger gamma value makes the decision boundary more dependent on points close to it, making the decision boundary more specific to the training data.

So, gamma parameter is also picked the best one out of (0.001 to 1000, stepping in 10x fold) to achieve better performance.

```
---Support Vector---
Accuracy on validation set is: 0.7687747035573123
Gamma value:  0.1
Regularization parameter (c):  100
Kernel Type:  rbf
Cross validation score:  0.8157894736842105
Test AUC 0.8222222222222222
Test recall: 0.7
```

C. Long Short-Term Memory (LSTM):

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to handle the issue of vanishing or exploding gradients in traditional RNNs, enabling them to effectively capture long-range dependencies in sequential data. *Memory Cells*, *Gates* (*Forget*, *Input* and *Output*), *Cell State* and *Hidden State* are the primary components of a LSTM model.

LSTMs are specifically engineered to address the problem of preserving and utilizing information over long sequences, which is crucial in various tasks involving sequential data like time series prediction, natural language processing, speech recognition, and more. The architecture and mechanisms of LSTM enable it to capture and remember long-term dependencies by selectively remembering or forgetting information at different time steps. This capability is particularly useful in scenarios where understanding context over longer sequences is essential. The longitudinal dataset of this research can take advantage of the long retaining information and can provide better classification outcome.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 20)                2320

 dense (Dense)               (None, 2)                 42

=================================================================
Total params: 2362 (9.23 KB)
Trainable params: 2362 (9.23 KB)
Non-trainable params: 0 (0.00 Byte)
```

Fig8: LSTM model setup summary

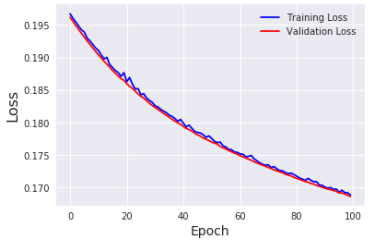


Fig9: Training loss in our LSTM epoch (100).

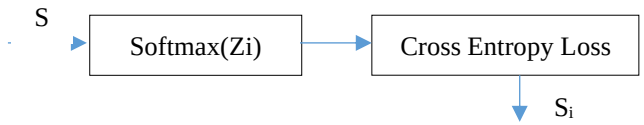
This validation is executed using the final 20% of the data, which was segregated from the initial 80% of the dataset. We set up the batch size 30 and epoch 100 to fit the model. We finally got the Test AUC score of 0.845, which is very thinly better than what we got in Random Forest. We still have to calculate the Recall score as our future work.

We use SoftMax to squash the vector in range (0, 1) in the output layer to calculate the relative probabilities.

$$Softmax(Zi) = \frac{\exp(Zi)}{\sum_j \exp(Zj)}$$

Cross-entropy is a measure used in the context of evaluating the difference between two probability distributions or, more commonly, between the predicted probability distribution and the actual probability distribution of outcomes. We have been using cross-entropy to calculate the loss in the classification as:

$$Cross\ Entropy\ Loss = -\sum_{i}^{C} Ti\ \times \log\left(f(s)i\right)$$



D. Results and Future works

TABLE II. PERFORMANCE OF THE CLASSIFICATION

ML System	***AUC***	***Recall***
Random Forest (RF)	0.844	0.8
Support Vector Machine (SVM)	0.82	0.7
Long Short-term Memory (LSTM)	0.845	-

Fig. 1. Three different classification models were tested and measured the performance in terms of AUC

The result shows that the LSTM model seems to be outperforming very thinly than the others. The Random Forest is also performing close enough to the LSTM. The Recall score that measures the overall performance is better in Random Forest. We still need to compute the recall score for LSTM as our future work.

As some future development, we further need to examine the models in larger dataset. Also, there are pretrained models now available in the hugging face where we highly likely get better performance and accuracy. Moreover, we may use Transformer models that now have self-attention mechanism which can give more importance to some features and are likely to produce more desirable outcomes.

REFERENCES

[1] Räisänen S, Cartwright R, Gissler M, Kramer MR, Heinonen S. The burden of OASIS increases along with socioeconomic position--register-based analysis of 980,733 births in Finland. PLoS One. 2013 Aug 27;8(8):e73515. doi: 10.1371/journal.pone.0073515. PMID: 24013645; PMCID: PMC3754956.)

[2] Philip D. Harvey, Richard C. Mohs, 5 - Memory Changes with Aging and Dementia, Editor(s): PATRICK R. HOF, CHARLES V. MOBBS,Functional Neurobiology of Aging, Academic Press, 2001, Pages 53-63, ISBN 9780123518309,https://doi.org/10.1016/B978-012351830-9/50007-X.

[3] Salis F, Costaggiu D, Mandas A. Mini-Mental State Examination: Optimal Cut-Off Levels for Mild and Severe Cognitive Impairment. Geriatrics (Basel). 2023 Jan 12;8(1):12. doi: 10.3390/geriatrics8010012. PMID: 36648917; PMCID: PMC9844353.

[4] Fulton, L.V.; Dolezel, D.; Harrop, J.; Yan, Y.; Fulton, C.P. Classification of Alzheimer's Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50. *Brain Sci.* 2019, *9*, 212. https://doi.org/10.3390/brainsci9090212

[5] C. DeCarli, J. Massaro, D. Harvey, J. Hald, M. Tullberg, R. Au, A. Beis er, R. D'Agostino, P.A. Wolf, Measures of brain morphology and infarction in the framingham heart study: establishing what is normal Neurobiol Aging, 26 (4) (2005), pp. 491-510.

[6] A.F. Fotenos, A. Snyder, L. Girton, J. Morris, R. Buckner, Normative estimates of cross-sectional and longitudinal brain volume decline in aging and ad Neurology, 64 (6) (2005), pp. 1032-1039.

[7] N. Raz, A. Williamson, F. Gunning-Dixon, D. Head, J.D. Acker, Neuroanatomical and cognitive correlates of adult age differences in acquisition of a perceptual-motor skill Microsc Res Tech, 51 (1) (2000), pp. 85-93

[8] Zhou Y, Song Z, Han X, Li H, Tang X. Prediction of Alzheimer's Disease Progression Based on Magnetic Resonance Imaging. ACS Chem Neurosci. 2021 Nov 17;12(22):4209-4223. doi: 10.1021/acschemneuro.1c00472. Epub 2021 Nov 1. PMID: 34723463.

[9] Kevin de Silva, Holger Kunz, Prediction of Alzheimer's disease from magnetic resonance imaging using a convolutional neural network, Intelligence-Based Medicine, Volume 7, 2023, 100091, ISSN 2666-5212, https://doi.org/10.1016/j.ibmed.2023.100091.(https://www.sciencedirect .com/science/article/pii/S2666521223000054)

[10] 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