CSE3046 PROGRAMMING FOR DATA SCIENCE

THEORY DIGITAL ASSIGNMENT

Team members:

1. **Ashwath S - 20BDS0301**
2. **Vishaal T - 20BDS0271**
3. **R Shamini - 20BDS0350**
4. **Gokul V.S - 20BDS0078**
5. **Badrinarayan M - 20BDS0280**

Date: 25-10-22

Submitted to - Dr. Suresh P (Associate Professor Grade 2)

Topic:

TECHNIQUES FOR CLASSIFYING EARLY DETECTION OF ALZHEIMER'S DISEASE

ABSTRACT

Alzheimer's disease is one of the most prevalent dementias. Even though the symptoms start out mild, they eventually continue to deteriorate. The fact that there is no treatment for this illness makes it a challenging issue. Therefore, if the condition is identified early, its course or consequences may be slowed. In this study, we explored a few ensemble machine learning models to predict when Alzheimer disease would start to emerge.

1. Introduction:

Alzheimer's disease(AD) is a chronic condition that leads to the degeneration of brain cells leading to memory enervation. AD also occurs due to genetics, aging, and environmental factors.
A progressive neurologic disorder that causes brain shrinkage (atrophy) and cell death.This causes a continuous decline in behavioral and social skills that affects a person's ability to function independently.

Alzheimer's disease has a very high impact. Forgetting recent conversations or events is one of the disease's early symptoms. A person with Alzheimer's disease will develop severe memory impairment and lose the ability to perform daily tasks as the disease progresses.
Around worldwide 29.5 million people approximately suffered from Alzheimer's disease in 2015. At the age of 65, it most often begins in people, but 4% to 5% of cases are early-onset before these ages. Due to the cause of Dementia, 1.9 million deaths in the year 2015. AD is one of the most financial diseases in developed countries. In India, some form of Dementia is suffering by more than 4 million people. In 2050, the number of people suffering from AD will set to triple.

Approximately 5.8 million Americans aged 65 and older, according to a source, have Alzheimer's disease. Eighty percent of them are 75 or older. Between 60% and 70% of the estimated 50 million dementia sufferers worldwide are thought to have Alzheimer's disease. Alzheimer's disease has no cure. Complications such as dehydration, malnutrition, or infection occur in the advanced stages of the disease, leading to death [13].

This has a huge psychological and economic burden on people, society, and the country. No effective drug existed for a very long time. The recently first therapeutic drug, Aduhelm, was approved. This drug has not shown its efficiency though.

Machine learning methods have been explored and used in many medical sectors, such as lung cancer, skin cancer, breast cancer, etc., Right now, machine learning is playing a key role in health-related areas. Machine learning provides novel techniques to address high-dimensional data, integrate data from different sources, model the etiological and clinical heterogeneity, and discover new biomarkers. These directions have the potential to help us better manage the disease progression and develop novel treatment strategies.

The aim of this review is to detect Alzheimer's disease in the primitive stage and summarize different ML methods that have been applied to study AD.

2. Literature Survey:

Research paper no.	Work	Model Used	Future Work
1.	Investigate either the behavior of the main existing off-the-shelf CNNs or a deep ensemble-based strategy aimed at realizing a comprehensive CAD framework based on patient MRIs and fMRIs.	2D CNN, 3D CNN, Ensemble models,	Combining HC and CNN features. Handle unbalanced classes.
2.	To diagnose AD and MCI, the ISDL model uses joint deep feature extraction and critical cortical region identification.	RVM, CNN, Attention-CNN, en3D CNN, SCFR, 3D GCNet, 3D Efficient-B0, ISDL.	Deep learning model capable of dealing with inter-class similarity and intra-class variation.
3.	Systematically review the current state of using deep learning techniques in the diagnosis of Alzheimer's disease using neuroimaging data.	Multitask framework based on LSTM, multi-modal or multi-data approach, 2D CNN, 3D-CapsNet, and 3D-AEs.	Future efforts will be made to address poor performance brought on by small datasets due to the likelihood of overfitting occurrences.
4.	Deep learning model for auxiliary Alzheimer's disease diagnosis that simulates the clinical diagnostic process.	CNN, ANN, SVM, linear method.	NIL

5	Assessing the recent memory loss in interactions with people and between the virtual and physical worlds; abnormal recognition, expression, and understanding of diagnostic differences in language issues words.	IOT systems, Machine Learning Models and Deep Learning Models	Early Prediction of Alzheimer's Disease.
6	Multi-task learning approach based on hybrid feature maps and a high-order discriminative convolutional Boltzmann machine.	SVM,DBN-3,GDBM-2,FitNet-10,GoogleNet,CDBN etc.	If the discriminant version of DCssCDBM acts better, investigate it to see if it might be used for early disease identification, such as epilepsy detection.
7	Comparative analysis of around 100 publications published since 2019 that use generative models, CNNs, and other fundamental deep architectures for Alzheimer's Disease diagnosis	Deep feed forward neural networks (DFFNN),Convolutional neural networks (CNNs), Recurrent neural networks (RNNs) and Deep polynomial networks (DPNs)	NIL
8	Proposed a technique for analyzing medical data with the goal of identifying risk categories.	APRIORI algorithm and Generation of Association Rules	NIL
9.	Alzheimer's disease classification using SVM and Artificial Neural networks to distinguish various stages of the disease	SVM and ANN	Research work is needed to devise Deep Learning algorithms to integrate data from several early detection modalities.
10.	Proposing an efficient framework for identifying epistatic interactions between all pairs of nucleotides in a DNA sequence by Integrating Multifactor Dimensionality	MDR model using Deep Learning	NIL

	Reduction (MDR) with Deep Learning		
11.	Proposing a uni-data, a multi-model framework for Alzheimer's disease detection which is implemented using five-fold cross-validation which boosts the classification performance and thereby reduces overfitting	3D-ResNet, Random Tree Embedding,XGBoost,ensemble models	Extending the model to detect both AD and MCI. also plans to add brain images acquired from imaging modalities to increase the diversity of individual learners.
12.	A conditional deep triplet network model is used to overcome the limitation of lack of image data (limited image samples) and to provide higher accuracy with minimum samples	Deep Triplet network	NIL
13.	Early prediction of Alzheimer's disease.	Support vector machine and decision tree.	They are combining MRI scans with psychological parameters
14	Alzheimer’s disease in the ADNI goal is to investigate whether Positron Emission Tomography (PET), sequential Magnetic Resonance Image (MRI), and the biotic markers, neuropsychological and the objective evaluation are being connected.	Neural networks, Random forest, SVM, KNN, Gradient Boost	NIL
15	The preprocessed fMRI 4D data in Nifti format were concatenated across z and t axes and the converted to a stack of 2D images in JPEG using Neuroimaging package Nibabel and Python OpenCV. Next, images were labeled for binary classification of Alzheimer’s Vs Normal data. The labeled images were converted to lmdb storage Databases for high-throughput to be fed into Deep Learning platform. LeNet model which is based on Convolutional Neural Network architecture from Caffe DIGITS 0.2 - deep learning framework (Nvidia version) - was	Neural networks - CNN	Need to generalize this method for all age groups and extend this method for other stages of Alzheimer's disease as well.

	used to perform binary image classification.		
16	Six different machine learning models are used to find the five different stages of Alzheimer's disease using ADNI dataset.	K-NN, Naive Bayes, Decision Tree, Rule Induction, Generalized Linear Model, Deep learning models	The accuracy of the AD stages classification could be further improved by increasing the number of instances for EMCI and SMC classes so that the model can be trained with sufficient and balanced data for all classes.
17	They created a deep learning architecture with stacked auto-encoders and a softmax output layer to circumvent the issue and aid in the diagnosis of AD and its symptomatic stage (Mild Cognitive Impairment).	Auto encoders, softmax regression, ROI sensitivity evaluation	Fine tune parameters
18	Review paper based on Diagnosis of alzheimer's disease using Machine learning models	Paperwork based on major models such as ANN,SVM,DL,deep learning and ensemble learning were discussed	Novel variants of ANN can be used. More importance must be given to clinical interpretability of deep learning models
19	Developed models that can extract and classify digital EEG signal (dEEG) dataset patterns using an ML technique known as Support Vector Machines (SVM).	SVM models have been used to search patterns in EEG epochs	SVM parameters can be tuned
20	Modelling our brain using deep learning method in such a way that it differentiates normal brain and Alzheimer's disease affected brain.	SVM, DNN	Accuracy is only enough to test in a real clinic. Not enough to trust in the real-time process.

3. PROBLEM STATEMENT

Some studies suggest that MRI features may be used to guide therapy in the future and predict the rate of AD Deterioration. To get there, medical professionals will need to employ machine learning strategies that can precisely forecast a patient's progression from mild cognitive impairment to dementia.
We recommend developing a trustworthy model that can help clinicians achieve that and expect the onset of early Alzheimer's.

4. PROPOSAL OF METHODOLOGY

Initially we obtained a dataset regarding **Alzheirmer's disease prediction**. The dataset contained around 400 values and had 15 attributes out of which the "**Group**" attribute was taken to be the target attribute. The dataset contained many null values and because of this we had to Pre-Process the dataset by removing null values and also by removing the rows which contained the null values. Also, the redundant columns/attributes were dropped to train the model more efficiently. After Pre-Processing, the dataset was split into training and testing data. Then we applied 5 different Machine Learning models (Ensemble/Hybrid models) to fit into the dataset and the corresponding accuracies were obtained and compared.

The main goal of this paper is to determine/predict the presence of Alzheimer's disease given a set of attributes, so in order to achieve this, we had implemented 5 different Ensemble Models and obtained the accuracies of the models and finally concluded which was the **Best-Fitting** Ensemble Model.

4.1. *DATASET*

The used dataset was obtained from Kaggle website and the name of the dataset is "Detecting Early Alzheimer's" (Oasis Dataset). The dataset contained 374 different records with 15 attributes. The chosen target attribute was **"Group"**, which contained values "Demented", "Non-Demented" and "Converted".

4.2. *DATA PRE-PROCESSING*

The Data Pre-Processing stage is the most essential phase of the Data Analysis Life-Cycle, which makes the data clean and can be used to obtain accurate and efficient results.
The Pre-Processing steps that were implemented are as follows

1. Identifying Null Values.
2. Removal of Null values from the Dataset.
3. Identifying the Rows which contained the Null Values.
4. Removal of all the Rows that contained Null values.
5. Replacement of Values of Group Attribute from Converted to Demented.
6. Encoding the values of Target Attribute (Group) [Demented - 0, Non-Demented - 1]
7. Dropping of redundant columns.
8. Identifying Rows with missing values.
9. Removal of all the Rows with missing values.

4.3. *DATA SPLITTING*

The Pre-Processed data is now split into Training and Testing data as follows:

1. Splitting the Dataset into variables X and Y.
2. Here X variable contains the Predictor Attributes and Y variable contains the Target Attribute.
3. The X and Y variables were divided into Training and Testing sets respectively.
4. The split ratio that was implemented was 80:20.

4.4. DATA MODELING

The Training and Testing sets are then applied to fit various Machine Learning ensemble models to determine the appropriate model to be used for the classification of **early Alzheimer's disease**. The various models used are listed and discussed below:

4.4.1.Random Forest Classifier Model

The supervised learning method includes the well-known machine learning algorithm Random Forest. It can be applied to ML Classification and Regression issues. Its foundation is the idea of ensemble learning, which is the process of mixing various

classifiers to solve a challenging problem and enhance the performance of the model.Random Forest is a classifier that, as the name implies, "contains a number of decision trees on various subsets of the provided dataset and takes the average to enhance the predictive accuracy of that dataset." Instead of depending on a single decision tree, the random forest uses forecasts from all of the trees to anticipate the outcome based on which predictions received the most votes.

Algorithm:

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier()
clf.fit(X_train, y_train)
clf.score(X_test, y_test)
```

4.4.2.Gradient Boosting Classifier Model

Each prediction in gradient boosting aims to outperform the one before it by lowering the errors. Gradient Boosting's intriguing concept, however, is that it really fits a new predictor to the residual errors created by the preceding predictor, rather than fitting a prediction on the data at each iteration. The resulting technique, known as gradient-boosted trees, typically beats random forest when a decision tree is the weak learner. The construction of a gradient-boosted trees model follows the same stage-wise process as previous boosting techniques, but it generalizes other techniques by enabling the optimization of any differentiable loss function.

Algorithm:

```
from sklearn.ensemble import GradientBoostingClassifier
clf = GradientBoostingClassifier()
clf.fit(X_train, y_train)
clf.score(X_test,y_test)
```

4.4.3. AdaBoost Classifier Model

Yoav Freund and Robert Schapire proposed the Ada-boost or Adaptive Boosting ensemble boosting classifier in 1996. To improve classifier accuracy, it combines several classifiers. An iterative ensemble algorithm is AdaBoost. AdaBoost classifier combines a number of ineffective classifiers to create a strong classifier that has a high degree of accuracy. The fundamental idea underlying Adaboost is to train the data sample and set the

classifier weights in each iteration in a way that provides accurate predictions of uncommon observations. Any machine learning method that accepts weights from the training set can be used as the basis classifier.

Algorithm:

```
from sklearn.ensemble import AdaBoostClassifier
clf= AdaBoostClassifier(random_state=96)
clf.fit(X_train,y_train)
```

4.4.4.Extra Trees Classifier Model

Extremely Randomized Trees Classifier, also known as Extra Trees Classifier, is a form of ensemble learning technique that combines the findings of various de-correlated decision trees gathered in a "forest" to produce its classification outcome. The only way it differs conceptually from a Random Forest Classifier is in how the decision trees in the forest are built.

The initial training sample is used to build each decision tree in the Extra Trees Forest. Then, each tree is given a random sample of k features from the feature-set at each test node, from which it must choose the best feature to divide the data according to certain mathematical criterion (typically the Gini Index).

Algorithm:

```
from sklearn.ensemble import ExtraTreesClassifier
clf =  ExtraTreesClassifier(n_estimators=100, random_state=0)
clf.fit(X_train, y_train)
```

4.4.5.Voting Classifier Model

A voting classifier is a machine learning model that gains experience by training on a collection of several models and forecasts an output (class) based on the class with the highest likelihood of being the output.

To predict the output class based on the highest majority of votes, it merely averages the results of each classifier that was passed into the voting classifier. The concept is to develop a single model that learns from these models and predicts output based on their aggregate majority of voting for each output class rather than developing separate dedicated models and determining the correctness for each one.

Algorithm:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from itertools import product
from sklearn.ensemble import VotingClassifier

clf1 = DecisionTreeClassifier(max_depth=4)
clf2 = KNeighborsClassifier(n_neighbors=7)
clf3 = SVC(kernel='rbf', probability=True)
eclf = VotingClassifier(estimators=[('dt', clf1), ('knn', clf2), ('svc', clf3)],
                        voting='soft', weights=[2, 1, 2])

clf1 = clf1.fit(X, y)
clf2 = clf2.fit(X, y)
clf3 = clf3.fit(X, y)
eclf = eclf.fit(X, y)
```

4.5. IMPLEMENTATION

The ML algorithms are implemented using the standard libraries available such as SKLearn and its tools which are used in the direct application of various complex ML algorithms.

- Python, version 3.9.12 is a programming language that is open-source. Machine learning models were used for the implementation

- Scikit-learn, version 1.1 supports both Supervised and Unsupervised Machine Learning algorithms . Additionally, it offers a number of tools for data preprocessing, model selection, model evaluation, and many other utilities. Ensemble models are used from this library

Performance Analysis:

Machine learning tasks are associated with evaluation metrics. For classification and regression tasks, various metrics are available. Some metrics, such as precision, and recall, are useful for a variety of tasks. Classification and regression are examples of supervised learning, accounting for most machine

learning applications. We should be able to increase our model's overall predictive power using various metrics for performance assessment before deploying it for production on unrecognized data. When the respective model is deployed on unseen data, failing to rigorously evaluate the Machine Learning model using different evaluation metrics and relying solely on accuracy can lead to problems and poor predictions.

This is useful because it provides a rough target for a machine learning engineer or data scientist to work forward into. However, the evaluation metric might alter over time due to the nature of experimentation. In this project, we evaluated a model using scikit-learn evaluation metrics. In this project, the built-in score method of the estimator, the scoring parameter, and problem-specific metrics functions are used to evaluate Scikit-learn models or estimators.

Accuracy, the area under the ROC curve, the confusion matrix, and the classification report are metrics for evaluating classification models. The Goldilocks model is what we're looking for. One that performs well not only on our dataset but also on previously unseen examples. We could use a validation set to test different hyperparameters, but since we don't have much data, we'll use cross-validation. K-fold cross-validation is the most common type of cross-validation. It entails dividing your data into k-folds and then testing a model on each of them. Although the cross-validated accuracy is preferred, we still take it into consideration even though the mean accuracy is higher.

Using the baseline model, we attempted to improve and evaluate the model through hyperparameter tuning. Each model we use has a set of dials that can be turned to control how it performs. Changing these values may improve or degrade model performance. The techniques used in this project include hyperparameter tuning by hand, gridsearchCV, and randomsearchCV. Although we attempted to improve our model through hyperparameter tuning, the accuracy is lower than the baseline model accuracy, but there are improvements in other metrics such as precision, recall, and f1-score.

RandomForest model :

Evaluation metrics	**Non-Cross-Validated**	**Cross-Validated**
Accuracy	89.66 %	90.15 %
Precision	1.0	0.96
Recall	0.833	0.814

Hyperparameter tuning - RandomForest :

Evaluation metrics	**Manual**	**RandomSearchCV**	**GridSearchCV**
Accuracy	80.95 %	82.76 %	82.76 %
Precision	1.00	1.00	1.00
Recall	0.64	0.72	0.72

Gradient Boosting model :

Evaluation metrics	**Non-Cross-Validated**	**Cross-Validated**
Accuracy	86.11 %	84.51 %
Precision	0.833	0.8752
Recall	0.882	0.814

AdaBoost model :

Evaluation metrics	**Non-Cross-Validated**	**Cross-Validated**
Accuracy	86.11 %	83.15 %
Precision	0.8095	0.8584
Recall	0.9444	0.814

Extra-trees classifier model :

Evaluation metrics	**Non-Cross-Validated**	**Cross-Validated**
Accuracy	8.76 %	88.77 %
Precision	0.882	0.96
Recall	0.833	0.814

Voting Classifier model :

Evaluation metrics	**Non-Cross-Validated**	**Cross-Validated**
Accuracy	96.55	85.86
Precision	1.0	0.881
Recall	0.971	0.828

5. Link for Implementation:

https://github.com/ashwath7/Early-detection-of-Alzheimer-disease

6. CONCLUSION:

We have hence performed Alzheimer's disease detection in the early stages of the disease. We can see that Random forest and Voting classifiers have the highest accuracy among other classifiers. We have performed our analysis based on the dataset where the dataset is derived from MRI images(Open Access Series of Imaging Studies (OASIS) is the organization that has converted MRI images into longitudinal MRI data). Hence our model is very efficient with data in the format of our dataset. Data preprocessing is required with our implementation as we have to convert image data into the format of our dataset. In the case of raw MRI images, we would have to go with deep-learning models such as CNN, DNN, RNN, etc., Since they are more efficient than normal Machine learning classifiers in terms of images. Our model can be used to solve the current economic stability of AD and solve the AD problems to a good extinct.

7. REFERENCES:

1. Deep learning based pipelines for Alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method Andrea Loddo *, Sara Buttau, Cecilia Di Ruberto

2. Iterative sparse and deep learning for accurate diagnosis of Alzheimer's disease Yuanyuan Chen a , b , Yong Xia a , b , *

3. Early diagnosis of Alzheimer's disease based on deep learning: A systematic review Sina Fathi a, Maryam Ahmadi a,*, Afsaneh Dehnad b

4. Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease Fan Zhang a , b , *, Zhenzhen Li a , Boyan Zhang c , Haishun Du a , b , Binjie Wang d , Xinhong Zhang b , e , **

5. Intelligent diagnosis of Alzheimer's disease based on internet of things monitoring system and deep learning classification method Yuxin Zhou a,b, Yinan Lu a,*, Zhili Pei b

6. Early Alzheimer’s disease diagnosis based on EEG spectral images using deep learning Bi Xiaojun a, Wang Haibo b,*

7. Deep learning for Alzheimer's disease diagnosis: A survey M. Khojaste-Sarakhsi a,b, Seyedhamidreza Shahabi Haghighi b,*, S.M.T. Fatemi Ghomi b, Elena Marchiori a

8. Application of Machine Learning in medical data analysis illustrated with an example of association rules. Beata Butryna, Iwona Chomiak-Orsaa, Krzysztof Haukea, Maciej Pondela*, Agnieszka Siennickab

9. Machine learning and deep learning algorithms used to diagnosis of Alzheimer’s: Review Sridevi Balne a,⇑, Anupriya Elumalai b

10. Discovering epistasis interactions in Alzheimer's disease using deep learning model Marwa M. Abd El Hamid a,b,*, Yasser M.K. Omar b, Mohamed Shaheen b, Mai S. Mabrouk c

11. Ensemble learning using traditional machine learning and deep neural network for diagnosis of Alzheimer’s disease Dong Nguyen a,b,1, Hoang Nguyen a,b,1, Hong Ong a,b, Hoang Le c, Huong Ha a,b,*, Nguyen Thanh Duc d,e,**, Hoan Thanh Ngo a,b,*

12. Alzheimer’s disease detection from structural MRI using conditional deep triplet network MaysamOrouskhania,b,*, ChengchengZhub, SaharRostamianc, FiroozehShomalZadehb, MehrzadShafieib, YasinOrouskhanid

13.Alzheimer Disease Prediction using Machine Learning Algorithms

14. Diagnosis of Alzheimer’s Disease using Machine Learning

15. Classification of Alzheimer’s Disease Using fMRI Data and Deep Learning Convolutional Neural Networks

16. Classification of Alzheimer's Disease using Machine Learning Techniques

17. EARLY DIAGNOSIS OF ALZHEIMER’S DISEASE WITH DEEP LEARNING Siqi Liu1, Sidong Liu1, Student Member, IEEE, Weidong Cai1, Member, IEEE, Sonia Pujol2, Ron Kikinis2, Dagan Feng1, Fellow, IEEE

18. Machine Learning Techniques for the Diagnosis of Alzheimer’s Disease: A Review M. TANVEER, B. RICHHARIYA, and R. U. KHAN, Discipline of Mathematics, Indian Institute of Technology Indore, Simrol, Indore

19. Improving Alzheimer's Disease Diagnosis with Machine Learning Techniques Lucas R. Trambaiolli, Ana C. Lorena, Francisco J. Fraga, Paulo A.M. Kanda, Renato Anghinah and Ricardo Nitrini

20. Automated detection of Alzheimer's Disease using Deep Learning in MRI