Cognitive States Classification Analysis

VIRGINIA VALCHEVA, OLGA GEORGIEVA
Faculty of Mathematics and Informatics,
Sofia University "St. Kliment Ohridski",
BULGARIA

Abstract: - Alzheimer's disease is a chronic, prolonged, and irreversible neurodegenerative disease of unknown cause. In recent years growing research interest assumes that by processing data of essential factors effective models can be defined for recognizing and predicting the disease development. The present article aims to propose classification models for the diagnosis of Alzheimer's disease cognitive states. For this aim medical data of biomarkers and cognitive assessment data are used. The novelty of the paper is to explore both the Amyloid/TAU/ Neurodegeneration framework and the biologically determined process of delay between the brain impairment and visibility of its appearances by incorporating these concepts in the model development procedure. The study explores the ability of three classifiers – Random Forest, Extreme Gradient Boosting, and Logistic Regression. Conclusion results have been done by comparison of the grouping abilities in different data spaces. The practical result of the study is helping to determine medical examinations that give accurate results for the diagnosis and prediction of the progression of the disease in possible earlier stages of the disease development.

Key-Words: - Data Analysis, Machine Learning, Data Mining, Classification, Medical Data Analysis, Alzheimer's Disease.

Received: August 23, 2023. Revised: May 29, 2024. Accepted: July 16, 2024. Published: September 3, 2024.

1 Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease of unknown cause. The disease is a severe, prolonged, and irreversible condition that compromises social and professional functioning. Various factors such as genetic burden, lifestyle, and environment can contribute to its appearance and development, [1], [2].

In recent years growing research interest assumes that by processing data of essential factors effective models can be defined for recognizing and predicting disease development. The factor dependence models can help professionals in searching for unknown factors' relationships and disease knowledge. Having such models at earlier disease stages can be helpful for developing prevention strategies and help in managing the problems of the sick, [3], [4], [5].

A large part of the investigations in this direction are focused on the analysis of brain Magnetic Resonance Imaging (MRI) as a good and reliable data source about the presence of the disease. A sophisticated statistical analysis procedure was implemented on diffusion-weighted MRI to detect changes in the white matter regions of the brain, [6]. Based on logistic regression analysis of genotype data it is concluded that Alzheimer's disease has a significant polygenic component, which has predictive utility for the disease risk, [7].

In answering the aim for identification of the dependency model a number of recent publications show the applicability and benefit of machine learning methods. The risk of Alzheimer's disease is analyzed based on data from various demographic, clinical examinations, and genetic factors, showing that age, cognitive function assessments, and specific biomarkers are important in the disease diagnosis. Three different machine learning approaches – Support Vector Machine (SVM), eXtreme gradient boosting of decision trees, and Artificial neural network are used to identify blood biomarkers used to improve the model predictivity for incident dementia, [8]. Classification models that analyze speech patterns detect early signs of Alzheimer's disease by analyzing features such as pauses, hesitation, and word-finding difficulties in speech samples to predict the possibility of Alzheimer's disease, [1]. The study [9] uses the kernel combination method of SVM to discriminate between AD or Mild cognitive impairment (MCI) and healthy controls using three modalities of biomarkers. Another study compares the different performances of three machine learning algorithms - Random Forest, Gradient Boosting, and eXtreme

Gradient Boosting algorithms using biomarkers of MCI classified factors to predict MCI to AD conversion. The highest accuracy was achieved using neuropsychological and Alzheimer-related biomarkers and cognitive tests, [10]. Authors of [11] show that the SVM algorithm successfully separate patients with AD from healthy aging subjects. It concludes that a combination of MRI features and demographics could predict AD with high accuracy.

Other studies rely on classification techniques to recognize disease cognitive groups by dealing with different data sets. Thus, several classifiers - GaussianNB, Decision Tree, Random Forest, XGBoost, Voting Classifier, and GradientBoost have been explored to predict Alzheimer's disease and demonstrate the potential of this approach, [3]. To train the models the authors use the Open Access Series of Imaging Studies (OASIS) data set and show the beneficial outcome with the voting classifier. The research of [5] employs a convolution NN for training and a Random Forest Classifier, KNeighborsClassifier, XGBClassifier, and Logistic Regression for testing and classification algorithms. This study looks at how different types of machine learning algorithms can be used to solve AD diagnostic challenges using a range of imaging modalities employed to diagnose Alzheimer's disease. Our recent investigation confirms the best performance of three classifiers namely Random Forest, Extreme Gradient Boosting, and Logistic Regression for AD diagnosis, [12]. It could be summarized that classification algorithms are successful tools for the recognition and prediction of Alzheimer's disease using different types of data, including MRI images, EEG signals, and biomarkers.

At the same time, there are still open questions that can be solved by machine learning methods. Thus, a deep and wide understanding of the existing interdependence between disease factors, disease symptoms, and appeared cognitive states as well as the respective description models is still under ongoing investigation purpose. In such aim, in recent years, a growing consensus on the critical importance of the timing of intervention and the need to initiate antiamyloid treatment during the presymptomatic stages of the disease has emerged, [13].

The present paper aims to investigate classification models for the recognition of cognitive states of Alzheimer's disease. A novelty of the paper is in exploring the concept of Amyloid/Tau/Neurodegeneration (A/T/N) framework improving the feature selection of the classification model. In addition, the biologically determined process of delay between the brain impairment and visibility of its appearances is also accounted for in the feature selection improving the model accuracy. The study explores the ability of three classifiers – Random Forest, Extreme Gradient Boosting, and Logistic Regression, that have already been proven to perform better than others for cognitive impairment recognition. Conclusion results have been done by comparison of the grouping abilities in the different data spaces formed. The practical result of the proposed investigation is effective models for diagnosis and prediction of the illness progression in possible earlier stages of its development

2 Data Set

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, [14]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease database, [14]. ADNI provides open access data of a wide range of clinical data collected over the years that are related to Alzheimer's disease and its inherent cognitive disorders. Nevertheless, ADNI has been primarily initiated to research the disease according to the brain image data as MRI, here our focus is on three different types of Alzheimer's examinations namely demographic, biomarker, and cognitive data. The good reason for this search is last medical investigations show that brain proteins serve as biomarkers for the disease and in combination with some demographic parameters they are disease preconditions, [13], [15], [16]. On the other hand, cognitive examinations are commonly used in medical practice being a solid base for the diagnosis and prediction of cognitive impairments. These examinations are not invasive, do not need special medical equipment, and are easily applicable. The present study uses the following data types.

- The data of demographic parameters as information on participants' age (AGE), gender (PTGENDER), and education (PTEDUCAT) as well as genetic risk data as Body Mass Index (BMI) and gene APOE4.

- Biomarkers information based on Cerebrospinal fluid and plasma analysis of the proteins β-amyloid (ABETA), total tau (TAU), and phospho-tau (PTAU), as well blood examinations of fluorodeoxyglucse (FDG) of glucose metabolism measure are known as most significant markers that indicate the disease presence.
- Cognitive examinations of various neuropsychological and neuropsychiatric tests of specific questions and observations estimate the cognitive functions of the different domains - memory, visuospatial, executive, and language. The most used is MMSE for neurodegenerative assessment as a commonly accepted test of cognitive function. Clinical Dementia Rating Scale (CDRS) and Clinical Dementia Rating Sum of Boxes (CDRSB) are widely used for assessing the severity of dementia in patients. Activities Questionnaire (FAQ) measures the ability to perform everyday activities. Alzheimer's Disease Assessment Scale (ADAS) is a set of tests that assess various aspects of cognitive function such as memory, language, and orientation. Data from the Long Delay Free Recall Total (LDELTOTAL) test for the memory and neuropsychological test Rey Auditory Verbal Learning Test (RAVLT) are also used as data for neurodegenerative assessments.

Data on the clinical diagnosis of the cognitive state - normal cognition (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD), are also provided in ADNI. At only first visit the participants were diagnosed with five cognitive states: MCI is distinguished as Early mild cognitive impairment (EMCI) and Late mild cognitive impairment (LMCI). Significant memory concern (SMC) is a condition noted as a cognitive problem, but not diagnosed as Alzheimer's. At their next visits, the subjects from SMC are relegated to CN.

3 Methodology

The applied research approach follows a data mining procedure consisting of the following successive steps: data preprocessing, feature selection, data classification, and result analysis. The specificity of each stage and the particular techniques applied are presented below.

3.1 Data Preprocessing

Despite the data described in the previous section being a subset of ADNI data still preprocessing is important to apply. The problem of missing examination data and data of diagnosis is accomplished by filtering to ensure a fully processible data set. The remaining amount of data for further processing varies within the data spaces formed after the feature selection stage discussed in the next subsection.

Data normalization and transformation of categorical to numeric data are other tasks of the preprocessing stage. Min-max normalization is applied in order to solve the scaling problems. Categorical data are diagnosis data and some demographic data such as PTGENDER and PTEDUC. Appropriately a respective numerical value is written instead.

3.2 Feature Selection

The importance of this stage is determined by the need to select significant attributes that form a data space, where the cognitive groups could be well separated. There is no full information about the dependency between the features or their role in determining the cognitive state. Due to the existing diversity and amount of disease factors and biomarkers a feature selection algorithm needs to be applied to find features most relevant to the classification task.

In this study, we extend the feature selection investigations by forming and investigating different feature spaces in seeking the most informative one. First, we apply the standard approach to this task. The feature selection algorithm SelectKBest selects the best k features that are most informative for predicting the target variable of disease diagnosis. The evaluation function assesses the relevance of each feature by calculating an ANOVA F-value, that measures the linear dependency between the feature and the target value in the classification task.

The disadvantage of this approach is that feature selection is done according to the medical diagnosis. In medical practice most of the diagnosis rely on data of cognitive tests, which are not expensive and not invasive examinations. These examinations do not present the current brain impairment but the disease appearance. In answering this problem, we extend the feature search by adopting the Amyloid/Tau/Neurodegeneration framework as a valuable evidence of the biological state of AD, [13], [15], [16]. Amyloid-beta (ABETA) is a protein fragment that is produced naturally in the brain, but in Alzheimer's disease, it tends to accumulate and form plaques, that disrupt communication between

brain cells. Elevated level of ABETA is considered one of the disease biomarkers. In Alzheimer's disease, tau proteins, which play a crucial role in stabilizing neuronal structures, undergo modifications, such as phosphorylation. Phosphorylated tau (PTAU) forms tangles inside brain cells that disrupt normal neuronal function and contribute to cognitive impairment. In [15] the neurodegenerative status is estimated by MRI analysis. However, often in the medical practice the Neurodegenerative status is examined by combining assessments of cognitive tests, [10], [17]. By taking advantage of these results and trying to avoid the expensive and difficult-to-apply examinations here we adopt the cognitive data in the A/T/N framework. Thus, the space formed by ABETA/PTAU/Cognitive assessments is investigated for being an informative data space of cognitive group classification.

In forming the informative data space, we explore as well other knowledge for Alzheimer's disease. The dynamics of the disease, including the asymptomatic period, proceed with the deposition of the amyloid-β peptide in the brain, triggering the so-called "amyloid cascade", [13]. Obviously, the time delay in the onset of the disease relative to the asymptomatic accumulation of amyloid plaques must be considered. To answer of this, we investigate the classification abilities of space formed by ABETA/PTAU/Cognitive assessments, where the cognitive tests are done in late time then biomarkers examinations. Figure 1 summarizes the three approaches for forming the data spaces that are further investigated for classification analysis.

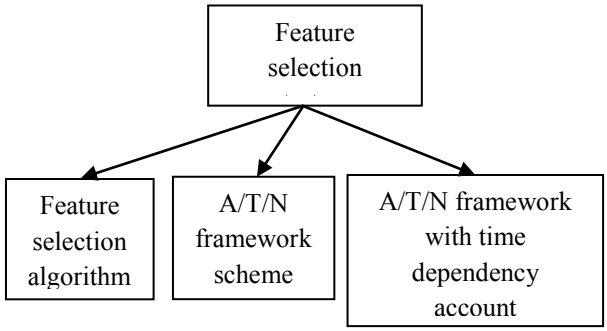

Fig. 1: Strategies for data space definition

3.3 Classification

Our recent investigation [12] based on the considered data set shows that three classifiers among seven ones, covering at large the diversity of the known classification approaches, are most presented. The two of them - Random Forest (RF), and Extreme Gradient Boosting (XGB) are based on the decision tree classification concept but with respective substantial improvement. RF is an ensemble learning method of multiple decision trees aggregating their predictions. XGB applies gradient boosting algorithm. The third method is extended version of Logistic Regression (LR) that deals with multiclassification task of statistical estimation of relationship between the features and the diagnosis outcome. Here, those three classifiers are used to solve the research aim.

Training the classifiers allow to learn the patterns and relationships between selected features and target diagnosis It is based on training dataset that consists a part of the available data. Adjusting hyperparameters of each classifier such as learning rate or tuning optimize the model performance. Cross-validation by StratifiedKFold algorithm is applied in order to ensure reliable training avoiding the imbalance of the data in the distinct classes. It provides such that each split contains approximately the same proportion of instance data of each class as the full data set.

Each classification model has to be further assessed for predicting ability by classification of the test data – data that are not used for training. This proves the model's applicability to new data. In order to form the test data, we took the next visits data. Each classifier is run several times for randomly generated and in an equal ratio of training and test sets.

3.4 Accuracy Evaluation of the Classification Models

Assessment of the performance of each trained classification model is evaluated by metrics Precision (P), Recall (R), $F1$ score, and Average Accuracy (AA):

$$P = TP/(TP + FP) \quad (1)$$

$$R = TP/(TP + FN) \quad (2)$$

$$F1 = 2 * P * R/(P + R) \quad (3)$$

$$AA = (TN + TP)/(TN + FP + TP + FN), \quad (4)$$

where the accuracy metrics are counted by the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) cases.

The accuracy assessment results for the training data runs and for the testing data runs are respectively averaged. The best performed classifier is considered in terms of all metrics and of both data sets.

Area under the curve (AUC) is also used as an accuracy measure. ROC AUC compares the relation between the True Positive Rate and the False Positive Rate. It typically includes the Precision rate

calculated by equation (1) on the ordinate and the False Positive Rate (FPR), where FPR=1-*P*, on the X-axis. In order to evaluate the accuracy of multi-class classifiers the One-vs-the-Rest multiclass strategy, also known as one-vs-all, is applied. It consists of computing the ROC AUC curve for each of the classes. The larger area under the ROC AUC curve means better classification.

4 Data Analysis Results

Initially, amount of 2370 participants' data was examined at their first visit for all considered features. The results of the preprocessing stage are discussed in the frame of the investigations of the respective data space.

4.1 Data Spaces

Following the considerations of the previous section feature selection algorithm is applied for first-visit data where diagnosis data were given for the three cognitive states – CN, MCI, and AD. By setting k=5 of the SelectKBest algorithm implemented by Python five features of cognitive assessments are selected. They are cognitive test data of MMSE, CDRSB, FAQ, ADAS13, and LDELTOTAL that define the Feature Space A (**FS_A**). The selection of only cognitive tests as significant features could be explained with the applied selection function. It finds attributes most correlated with the diagnosis. Bearing in mind that cognitive tests are most used in the practice for Alzheimer's diagnosis it could be supposed that diagnoses are much correlated with cognitive assessments.

The second data space to be investigated is determined by the A/T/N framework. We consider feature space defined by the proteins' biomarkers and some cognitive assessment of the first visit data. MMSE cognitive test is one of most applicable cognitive assessments. Thus, Feature space B (**FS_B**) formed by ABETA, PTAU, and MMSE is the second examined data space.

In order to set a data space by cognitive data obtained in late time we first investigated which time delay period is most appropriate for this aim. It could be seen that the number of changes of diagnosis is most often in 24-nd month (Table 1). Data from 481 subjects at their 24-month visit are used for investigation of the data space. In this space, the late cognitive assessment values in regards to the biomarkers values are taken. Thus, Feature space C (**FS_C**) is formed by the first visit data of ABETA, PTAU, and 24-th month assessments of one of the cognitive tests MMSE, CDRSB, FAQ noted as MMSE_24, CDRSB_24, FAQ_24, respectively. The three tests have been discovered as significant ones by the feature selection algorithm.

4.2 Classification Analysis

The three classification models were trained for the three defined data spaces. As at all visits except the first one the participants were diagnosed in three groups the classifiers were trained to distinguish the three classes namely CN, MCI, and AD. The trained classifiers were evaluated in regard to their ability to classify the test data sets. The test data sets were formed by the data of the 12-th month visit. As far as some participants do not have examinations at this visit the test set has been accordingly reduced for each examined data space.

Table 1. Number of the changed diagnosis

Diagnoses changed from CN to MCI		**Diagnoses changed from MCI to AD**	
Period /months/	Number of subjects	Period /months/	Number of subjects
6	12	6	46
12	9	12	72
24	**23**	18	36
36	10	**24**	**75**
48	12	36	48
72	10	48	29
108	16	72	10
120	7	108	20

Data space **FS_A** is defined by the estimations of cognitive tests MMSE, CDRSB, FAQ, ADAS13, LDELTOTAL. It consists data of from 2320 participants. They were divided into 2088 training and 232 testing sets used for the training stage. The trained classifiers were further applied to the test data. Table 2 presents the result (rounded values) of the accuracy metrics (1)-(4) obtained through the three classifiers for both training and test sets. The corresponding averaged metrics values are shown as well. The maximal metrics values are given in bold.

According to the A/T/N framework, the investigated data space is **FS_B** which is formed by ABETA, PTAU, and MMSE examination data. After filtering due to missing diagnosis and examinations 1541 data remain for the training. Data of 320 participants examined at the 12th month visit serve as a test set. The accuracy metrics values and respective their averaged values are presented in Table 3.

We to pay a special attention to the third discussed data space noted as **FS_C**. It is formed according to the novelty concept of exploring both the A/T/N framework and accounting for the biologically determined process of delay between

brain impairment and visibility of its appearances. Thus, the three classifiers were trained in the space **FS_C** that have been formed by varying different cognitive test data. Training results for each of the interested classifiers in the spaces formed by ABETA, PTAU data of the first visit and by respectively: a) MMSE_24 having 1064 data; b) CDRSB_24 (1054 data); c) FAQ_24 (1045 data) and d) by two estimations CDRSB and CDRSB_24 (1054 data) are presented at Table 4.

Table 2. Accuracy metrics values of classifiers' performance in **FS_A** data space

Classifier	***P***	***R***	***F1***	***AA***
Accuracy metrics values for the training data set				
RF	**0,944**	**0,943**	**0,943**	**0,943**
LR	0,917	0,916	0,915	0,916
XGB	0,929	0,928	0,928	0,928
Accuracy metrics values for the testing data set				
RF	0,813	0,833	0,820	0,818
LR	**0,818**	**0,835**	**0,824**	**0,821**
XGB	0,814	0,833	0,821	0,820
Average accuracy metrics values				
RF	**0,878**	**0,888**	**0,882**	**0,880**
LR	0,867	0,875	0,87	0,868
XGB	0,872	0,881	0,875	0,874

Table 3. Accuracy metrics values of classifiers' performance in **FS_B** data space

Classifier	***P***	***R***	***F1***	***AA***
Accuracy metrics values for the training data set				
RF	0,611	0,605	0,605	0,605
LR	**0,640**	**0,621**	**0,622**	**0,621**
XGB	0,614	0,607	0,607	0,607
Accuracy metrics values for the testing data set				
RF	0,607	0,603	**0,603**	0,601
LR	**0,689**	0,563	0,543	**0,611**
XGB	0,588	**0,616**	0,592	0,588
Average accuracy metrics values				
RF	0,609	0,604	**0,604**	0,603
LR	**0,665**	0,592	0,583	**0,616**
XGB	0,601	**0,611**	0,599	0,597

Table 4. Accuracy metrics values of classifiers' performance in **FS_C** data space

RF classification results				
Data space	***P***	***R***	***F1***	***AA***
a) ABETA, PTAU, MMSE_24	0,574	0,571	0,569	0,571
b) ABETA, PTAU, FAQ_24	0,653	0,650	0,648	0,650
c) ABETA, PTAU, CDRSB_24	**0,828**	**0,8198**	**0,820**	**0,8198**
d) ABETA, PTAU, CDRSB, CDRSB_24	**0,860**	**0,854**	**0,853**	**0,854**
LR classification results				
Data space	***P***	***R***	***F1***	***AA***
a) ABETA, PTAU, MMSE_24	0,595	0,594	0,579	0,594
b) ABETA, PTAU, FAQ_24	0,711	0,6996	0,686	0,6996
c) ABETA, PTAU, CDRSB_24	**0,813**	**0,806**	**0,802**	**0,806**
d) ABETA, PTAU, CDRSB, CDRSB_24	**0,848**	**0,842**	**0,840**	**0,842**
XGB classification results				
Data space	***P***	***R***	***F1***	***AA***
a) ABETA, PTAU, MMSE_24	0,575	0,567	0,567	0,567
b) ABETA, PTAU, FAQ_24	0,653	0,647	0,647	0,647
c) ABETA, PTAU, CDRSB_24	**0,822**	**0,814**	**0,815**	**0,814**
d) ABETA, PTAU, CDRSB, CDRSB_24	**0,846**	**0,8396**	**0,839**	**0,8396**

5 Results Analysis and Discussion

Comparison concerning the classifier's ability to distinguish the data space shows close partition performance of the three investigated classifiers. However, it should be underlined that Random Forest outperforms the rest two classifiers having accuracies over 0,94 for the training data set and the highest averaged metrics values in **FS_A** data space (Table 2). The Random Forest model is also best performed in **FS_C** data space (Table 4). Logistic Regression outperforms in the testing task of space **FS_A** (Table 2) and in the training task of **FS_B** (Table 3). However, testing and averaged accuracy metrics of **FS_B** space do not show any favorite performance.

The experimental results give information about the classification and predictability characteristics of the three examined feature spaces. It is a base to draw conclusions about the applicability of the three studied strategies for feature selection. The feature space **FS_A** formed by the cognitive tests examinations is the most informative one as it outperforms the accuracy values that are over 0,8 for both training and testing sets (Table 2). However again, it should be underlined that assessing only by

the cognitive tests means to diagnose the disease at the time of its visible appearance and not at its early stage.

On the other hand, the accuracy results of the data space **FS_C** for data spaces c) and d) consisting of CDRSB test as neurodegenerative assessment are fully commensurable with those of **FS_A** as the accuracy presented is also over 0,8 (Table 4). The accuracy is maximized if two data of CDRSB taken in different times examinations are used to form the data space. This proves the vitality of the idea using A/T/N framework with accounting for the delay of the cognitive tests data with respect to the biomarkers data to diagnosis and predicting Alzheimer's disease.

The obtained results are confirmed also by metrics of the Area under the curve. It is represented for RF classification for the different **FS_C** data spaces. Figure 2, Figure 3, Figure 4, Figure 5 and Figure 6 show the entire accuracy (Micro-average) and accuracy reached for each class (the three classes 0, 1, and 2 are shown, respectively). The areas of the curves in Figure 5 and Figure 6 are maximal. The obtained results confirm the preference of data space defined by biomarkers ABETA and PTAU data and CDRSB test data obtained 24 months after than biomarkers data.

Besides these achievements, it should be emphasized the good classification ability of the CDRSB test in comparison with the rest investigated. This could be explained by its properties as it tries to assess all aspects of the cognitive impairment. The recommendation is that it be used alone and not in the battery of cognitive tests as usual, [18].

6 Conclusion

The study presents a supervised data mining procedure for answering the important task of early-stage recognition of Alzheimer's disease for the need for planning and effective care. ADNI data is applied as a reliable medical data set for the classification analysis.

The practical result of the study is helping to determine medical examinations that give accurate results for the diagnosis and prediction of the progression of the disease in possible earlier stages of the disease development. For this, the feature selection stage is deeply considered a crucial stage in determining an informative data space where the cognitive data groups could be reliably distinguished. The vitality of the A/T/N framework applied to form the data space is shown. In addition, the novelty concept to improve the model accuracy by accounting for the delay in the cognitive tests' information with respect to the biomarkers data is proved. It is shown that the data space formed by the two important biomarkers ABETA and PTAU and cognitive test CDRSB data obtained 24 months after the biomarkers presents significant accuracy of the cognitive group distinguishing.

The comparison analysis of three known and well-performed classifiers Random Forest, Logistic Regression, and XGBoost for being classification models is investigated. The preferences of the Random Forest classifier are shown.

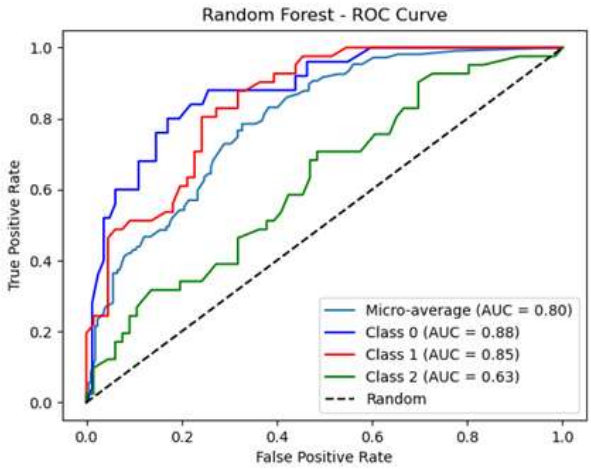

Fig. 2: AUC-ROC curve of Random Forest classifier applied to ABETA, PTAU, MMSE_24 data space

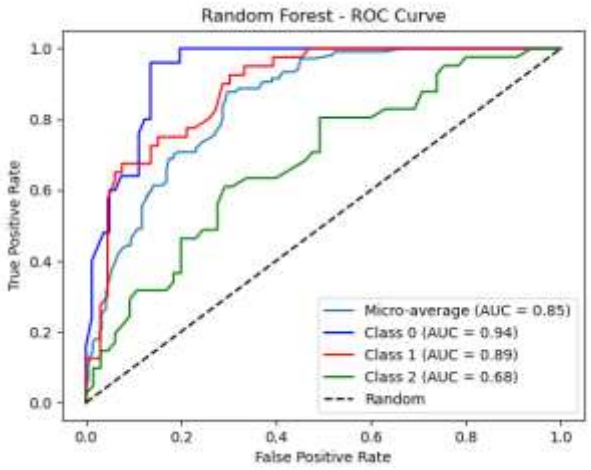

Fig. 3: AUC-ROC curve of Random Forest classifier applied to ABETA, PTAU, FAQ_24 data space

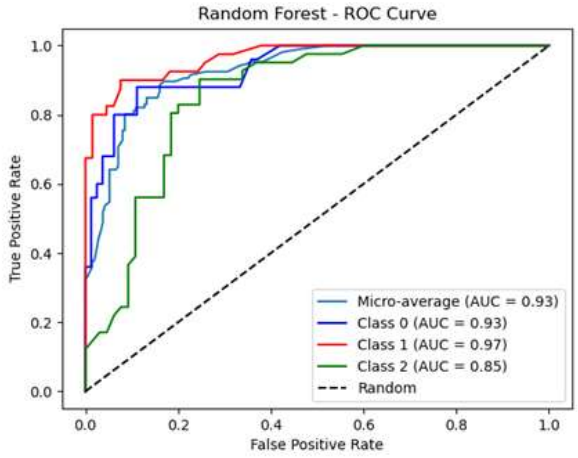

Fig. 4: AUC-ROC curve of Random Forest classifier applied to ABETA, PTAU, CDRSB_24 data space

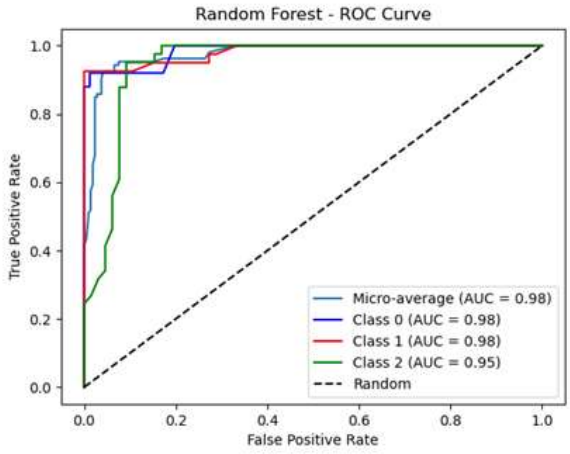

Fig. 5: AUC-ROC curve of Random Forest classifier applied to ABETA, PTAU, FAQ_24, CDRSB_24 data space

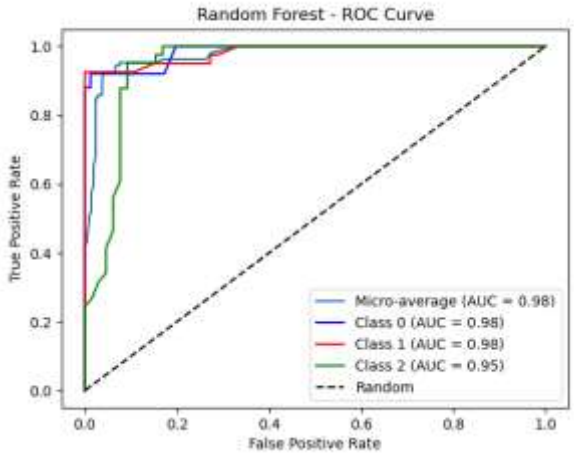

Fig. 6: AUC-ROC curve of Random Forest classifier applied to ABETA, PTAU, CDRSB, CDRSB_24 data space

Acknowledgement:
This research work has been supported by GATE project, funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 program under grant agreement no.857155 and by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT BG-RRP-2.004-0008-C01 and funded by Science Fund of Sofia University by project no. 80-10-137/2024.

References:

[1] Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's disease: JAD*, 49(2), 407–422. https://doi.org/10.3233/JAD-150520.

[2] Hugo, J., & Ganguli, M. (2014). Dementia and cognitive impairment: epidemiology, diagnosis, and treatment. *Clinics in geriatric medicine*, 30(3), 421–442. https://doi.org/10.1016/j.cger.2014.04.001.

[3] Uddin, K. M. M., Alam, M. J., Jannat-E-Anawar, Uddin, M. A., & Aryal, S. (2023). A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. *Biomedical materials & devices (New York, N.Y.)*, 1–17. Advance online publication. https://doi.org/10.1007/s44174-023-00078-9.

[4] Shrivastava, R.K., Singh, S.P., Kaur, G. (2023). Machine Learning Models for Alzheimer's Disease Detection Using OASIS Data. In: Koundal, D., Jain, D.K., Guo, Y., Ashour, A.S., Zaguia, A. (eds) Data Analysis for Neurodegenerative Disorders. Cognitive Technologies. Springer, Singapore, 111-126. https://doi.org/10.1007/978-981-99-2154-6_6.

[5] Sentamilselvan, K., Swetha, J., Sujitha, M., Vigasini, R. (2022). Alzheimer's Disease Detection Using Machine Learning and Deep Learning Algorithms. In: Abraham, A., et al. Innovations in Bio-Inspired Computing and Applications. IBICA 2021. Lecture Notes in Networks and Systems, 419. Springer, Cham. https://doi.org/10.1007/978-3-030-96299-9_29.

[6] Zhang, Y., Schuff, N., Ching, C., Tosun, D., Zhan, W., Nezamzadeh, M., Rosen, H. J., Kramer, J. H., Gorno-Tempini, M. L., Miller, B. L., & Weiner, M. W. (2011). Joint assessment of structural, perfusion, and diffusion MRI in Alzheimer's disease and

frontotemporal dementia. *International journal of Alzheimer's disease*, 2011, 546871. https://doi.org/10.4061/2011/546871.

[7] Escott-Price, V, Sims, R, Bannister, C, Harold, D, Vronskaya, M, Majounie, E, Badarinarayan, N, Morgan, K, Passmore, P, Holmes, C, Powell, J, Brayne, C, Gill, M, Mead, S, Goate, A, Cruchaga, C, Lambert, JC, Duijn, C, Maier, W, Ramirez, A, Holmans, P, Jones, L, Hardy, J, Seshadri, S, Schellenberg, GD, Amouyel, P, Williams, J, Gerad, P & Consortia, I 2015, Common polygenic variation enhances risk prediction for Alzheimer's disease, *Brain*, 138, pp. 3673-3684. https://doi.org/10.1093/brain/awv268.

[8] Lin, H., Himali, J. J., Satizabal, C. L., Beiser, A. S., Levy, D., Benjamin, E. J., Gonzales, M. M., Ghosh, S., Vasan, R. S., Seshadri, S., & McGrath, E. R. (2022). Identifying Blood Biomarkers for Dementia Using Machine Learning Methods in the Framingham Heart Study. *Cells*, 11(9), 1506. https://doi.org/10.3390/cells11091506.

[9] Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3), 856–867. https://doi.org/10.1016/j.neuroimage.2011.01.008.

[10] Franciotti, R., Nardini, D., Russo, M., Onofrj, M., Sensi, S. L., Alzheimer's Disease Neuroimaging Initiative, & Alzheimer's Disease Metabolomics Consortium ADMC (2023). Comparison of Machine Learning-based Approaches to Predict the Conversion to Alzheimer's Disease from Mild Cognitive Impairment. *Neuroscience*, 514, 143–152. https://doi.org/10.1016/j.neuroscience.2023.01.029.

[11] Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Jr, Ashburner, J., & Frackowiak, R. S. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(Pt 3), 681–689. https://doi.org/10.1093/brain/awm319.

[12] Valcheva V., Georgieva O., (2023). Data Classification Analysis for Alzheimer Disease Diagnostic, *27th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, Rhodes Island, Greece, 2023, 153-159. IEEE. https://doi.org/10.1109/CSCC58962.2023.00032.

[13] Jack, C. R., Jr, Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., Sperling, R., … Contributors (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 14(4), 535–562 https://doi.org/10.1016/j.jalz.2018.02.018.

[14] Alzheimer's Disease Neuroimaging Initiative, [Online]. https://adni.loni.usc.edu (Accessed Date: July 1, 2024).

[15] Calvin, C. M., de Boer, C., Raymont, V., Gallacher, J., Koychev, I., & European Prevention of Alzheimer's Dementia (EPAD) Consortium (2020). Prediction of Alzheimer's disease biomarker status defined by the 'ATN framework' among cognitively healthy individuals: results from the EPAD longitudinal cohort study. *Alzheimer's research & therapy*, 12(1), 143. https://doi.org/10.1186/s13195-020-00711-5.

[16] Jack, C. R., Jr, Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., Hampel, H., Jagust, W. J., Johnson, K. A., Knopman, D. S., Petersen, R. C., Scheltens, P., Sperling, R. A., & Dubois, B. (2016). A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*, 87(5), 539–547. https://doi.org/10.1212/WNL.0000000000002923.

[17] Chaves, M. L. F., Godinho, C. C., Porto, C. S., Mansur, L., Carthery-Goulart, M. T., Yassuda, M. S., Beato, R., & Group Recommendations in Alzheimer's Disease and Vascular Dementia of the Brazilian Academy of Neurology (2011). Cognitive, functional and behavioral assessment: Alzheimer's disease. Dementia & neuropsychologia, 5(3), 153–166. https://doi.org/10.1590/S1980-57642011DN05030003.

[18] ADNI3 Procedures Manual, Version 3.0, Alzheimer's disease neuroimaging initiative 3: Defining Alzheimer's disease, Keck School of Medicine of USC, [Online]. https://adni.loni.usc.edu/wp-content/uploads/2024/02/ADNI3_Procedures_

Manual_v3.0_29Feb2024.pdf, July 2024 (Accessed Date: July 5, 2024).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.