2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3503-9311-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBE60311.2023.00066

Predicting the Onset of Dementia in Initially Healthy Individuals Using Demographic and Clinical Data

1st Nikolaos Ntampakis
Dept. of Information & Electronic Engineering
International Hellenic University
Sindos, Greece
ntampakisnik@gmail.com
MetaMind Innovations
Kozani, Greece
nntampakis@metamind.gr

2nd Konstantinos Diamantaras
Dept. of Information & Electronic Engineering
International Hellenic University
Sindos, Greece
kdiamant@ihu.gr

3rd Konstantinos Goulianas
Dept. of Information & Electronic Engineering
International Hellenic University
Sindos, Greece
gouliana@ihu.gr

4th Ioanna Chouvarda
School of Medicine
Aristotle University of Thessaloniki
Thessaloniki, Greece
ioannach@auth.gr

***Abstract*—In this study, we aim to predict health outcomes concerning dementia by leveraging data from individuals' first two doctor visits. Using the OASIS-2 dataset, we focus on early-stage, longitudinal data, which is especially crucial for patients initially diagnosed as healthy. This approach allows us to anticipate their future health trajectories more accurately. We propose and evaluate multiple machine learning models, finding the Extreme Gradient Boosting (XGBoost) algorithm to be particularly effective, with an accuracy rate of 100%. Our methodology provides a valuable resource for early interventions and preventive measures in dementia care.**

***Index Terms*—dementia, progression prediction, machine learning, OASIS**

I. Introduction

Dementia is a debilitating neurological incurable disorder affecting over 50 million people globally [1], with projections suggesting that this number will surge to approximately 81 million by 2040. About 6 million new cases emerge each year, the vast majority of which are attributed to Alzheimer's disease [2]. It's worth noting that dementia serves as an umbrella term, encompassing various cognitive disorders including Alzheimer's, which makes it particularly challenging to distinguish between them. Therefore, Alzheimer's and other forms of dementia are frequently treated as one in many diagnostic settings [3].

Given its irreversible nature, rapid and accurate diagnostic systems become crucial for early intervention. This study leverages clinical and demographic data from Open Access Series of Imaging Studies(OASIS-2) dataset [4] — including results from MRI scans, educational levels, and scores from the Mini-Mental State Exam (MMSE) [5]—to develop machine learning models aimed at predicting the likelihood of progression from a healthy state to dementia. Such predictive capabilities could be instrumental in facilitating timely diagnoses and preventive measures.

The structure of this paper unfolds as such: Section 2 delves into existing literature pertinent to predicting the progression of dementia. Section 3 introduces the dataset employed for the study. Experimental procedures and corresponding findings are detailed in Section 4. Finally, Section 5 encapsulates key takeaways and outlines avenues for future inquiry.

II. Related work

Prior research has illuminated the complex landscape of dementia progression. For example, studies by Espinosa et al. [6] reveal that approximately 50% of individuals diagnosed with no dementia or amnestic Mild Cognitive Impairment (aMCI) [7] transition to dementia within a three-year timeframe. Intriguingly, a study by Mitchell and Shiri-Feshki [8] points out that a subset of these individuals either maintains stable cognitive function or reverts to a normative state. Moreover, Yaffe et al. [9] have shown that the rate of conversion is influenced by several factors, including age, sex, neuropsychological test scores, and educational background. Given these heterogeneous clinical outcomes among patients, a nuanced approach that considers the unique constellation of risk factors for each individual is crucial for accurately predicting conversion to dementia.

The endeavor to predict the progression of dementia has engaged researchers and clinicians alike, yielding a variety of

2471-7819/23/$31.00 ©2023 IEEE
DOI 10.1109/BIBE60311.2023.00066

methods and approaches. One of the initial attempts to address this challenge involves screening through a noninvasive method, such as neuropsychological testing. Specifically, the Free and Cued Selective Reminding Test (FCSRT) [10] has gained prominence for its utility in evaluating amnesic syndromes. In a preliminary study conducted within the general population by Auriacombe et al [11], the FCSRT demonstrated significant results in sensitivity and specificity — 92% and 64%, respectively — in predicting Alzheimer's Disease over a five-year period. However, it is important to mention that the test had a relatively low positive predictive value of around 8%, which suggests limitations in its applicability. In a similar vein, the MMSE [5] was developed as another tool for screening the progression of dementia. Like the FCSRT, the MMSE is a brief, noninvasive neuropsychological test commonly employed in clinical settings.

In recent years, the availability of more complex clinical data has enabled the development of advanced machine learning-based approaches. These models typically rely on comprehensive datasets that include cognitive and neuropathological measurements collected from large study cohorts [12]. Satone et al. [13] employed a Random Forest (RF) algorithm that utilized longitudinal features extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [14], demonstrating an 84.54% accuracy rate in predicting dementia progression over a two-year period. In Shaffel et. al. [15] study, Logistic Regression (LR) was used to evaluate how combining different types of data could improve the accuracy of predicting conversion to dementia. The data types considered were Magnetic Resonance (MR) imaging, Fluorodeoxyglucose Positron Emission Tomography (FDG PET), and cerebrospinal fluid (CSF) tests and clinical tests, achieving an accuracy of 90.62%.

The most recent research frontier seems to focus on leveraging data from multiple sources for a more comprehensive understanding of disease progression. Focused on predicting progression to dementia, during a 2 year period, in aMCI patients, Park et. al. [16], proposed a quantified Comprehensive Visual Rating Scale (CVRS) based on brain MRI. Building on this advancement, the CVRS score has been combined with clinical data to train a Light Gradient Boosting Machine model (Light-GBM), achieving an accuracy rate of 71.1%. Based on ADNI dataset, combining biomarkers with demographic information and behavioral tests, Albright [17] developed an all-pairs technique for predicting a patient's future cognitive state, training a Multi-Layer Perceptron (MLP) and achieving an accuracy of 86.6%.

The OASIS2 dataset [4] is one of the most pivotal open-source collections available for dementia research based on longitudinal MRI images. These datasets include beside MRI scans, demographic information and clinical data related to demented or not demented patients. To the best of our knowledge, the specific dataset has been primarily utilized for the classification of dementia. Shahina et. al. [18] proposed an XGBoost capable of achieving an 97.87% accuracy on predicting non-demented, demented and converted patients. Similarly, Morshedul et. al. [19] presented a Support Vector Machine (SVM) model that achieved an accuracy rate of 92.0% in distinguishing between demented and non-demented patients.

Building on the aforementioned research, our primary objective is to develop a predictive model capable of assessing the likelihood that a currently healthy individual will progress to dementia in the future.

III. Dataset description

TABLE I
OASIS-2 features [4]

Feature	Description
Subject ID	Identification code of each patient
MRI ID	Identification code of each patient's MRI
Group	The dementia group to which the subject belongs (Dementia, Non-Demented or Converted)
Visit	The ordinal number of the visit
MR Delay	The interval between the initial and the current MR session (days)
M/F	Sex (M or F)
Hand	The patient's significant hand of use
Age	Age at time of image acquisition (years)
EDUC	Years of education
SES	Socioeconomic status
MMSE	Mini-Mental State Examination score
CDR	Clinical Dementia Rating
eTIV	Estimated total intracranial volume (cm^3)
nWBV	Normalized whole-brain volume
ASF	Atlas scaling factor (unitless)

The scans included in the OASIS2 dataset [4] provide a comprehensive view of brain changes over multiple visits, facilitating a deeper understanding of neurodegenerative processes. Augmenting its value, the dataset not only includes a wide array of demographic and clinical information but also key cognitive metrics such as the MMSE, which is frequently utilized to assess the onset or progression of dementia.

Due to the longitudinal nature of OASIS-2 dataset, it includes at least two scanning sessions per participant, spaced by a minimum of one year. All scans were conducted using the same imaging equipment and identical sequencing protocols featuring 150 individuals ranging in age from 60 to 96. The total imaging events were 373. Clinical Dementia Rating (CDR) was employed to classify subjects as either nondemented or as having very mild, mild or moderate dementia.

All images offered under the OASIS-2 dataset are distributed in NIFTI1 format. Demographic, clinical, and derived imaging measures which are used under the needs of this study are available in XML and XLSX formats (oasis_longitudinal_demographics.xlsx). Table 1 shows the 15 features included in the XLSX file.

IV. Experimental Results & Evaluation

A. Experimental setup

Python served as the primary programming language for our experiments, utilizing the numpy and pandas libraries for effective data manipulation. The experiments were conducted

on a local machine, and the machine learning models were implemented using the scikit-learn library.

B. Data pre-processing

To meet the specific objectives of this study, which focus on predicting whether a patient will remain healthy or progress to dementia based on data from two consecutive visits, a comprehensive data pre-processing pipeline was designed. Patients with fewer than two visits were excluded to ensure the completeness of the longitudinal data. Features with a significant number of missing values or lacking predictive importance, such as "Hand" and "SeS," were eliminated. The data table was pivoted to encapsulate the sequence of two visits, appending the term "Next_" as a suffix to the variables from the second visit. The target value, "Progress" (0 = remain healthy, 1 = progress to dementia), was derived from the third visit's data. Concurrently, categorical features like "M/F" were appropriately encoded. Additionally, we filtered the dataset to include only those patients with two consecutive non-dementia values in the "CDR" column. All superfluous columns and rows containing missing values were removed, resulting in a dataframe with dimensions [47, 13]. Class distribution within the "Progress" column is further detailed in Fig 1.

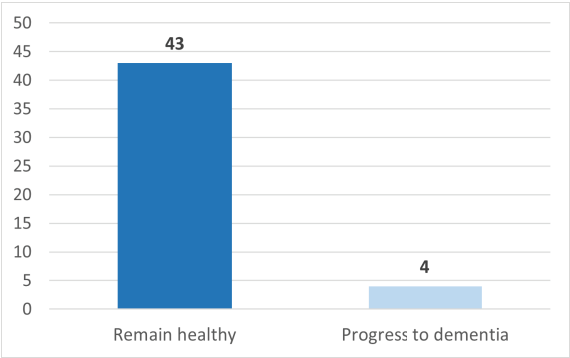

Fig. 1. Progress distribution

To mitigate the issue of varying feature scales across the dataset, we employed standard scaling methods to normalize the attributes. To tackle the class imbalance in the "Progress" column, we utilized the Synthetic Minority Over-sampling Technique(SMOTE) [20] technique to achieve a more uniform class distribution in the training set. The 'augmented' training set consisted of the same amount of samples of each class. Moreover, to ensure that both the training and test sets contained samples from each class, we performed the dataset splitting in an 80/20 ratio and enabled the 'stratify' parameter.

C. Results

1) Evaluation results: In the context of our binary classification problem, which aims to determine whether a patient will remain healthy or progress to dementia, we utilized key evaluation metrics such as the confusion matrix, accuracy and F1 score. The confusion matrix offers a structured framework that quantifies the predictive performance of our model. It displays the true positives (TP), which represent cases where our model correctly identifies a patient's progression to dementia; true negatives (TN), which are cases where the model accurately predicts that a patient will remain healthy; false positives (FP), where the model incorrectly flags a healthy patient as progressing to dementia; and false negatives (FN), where the model fails to identify an actual progression to dementia.

Based on the results from the confusion matrix, we calculated key evaluation metrics, including accuracy (Equation 1) and the F1 score (Equation 2). Given the imbalanced distribution of the dataset, special emphasis was placed on the macro-F1 score during the experimental evaluations. The macro-F1 score offers a balanced assessment by equally weighing the performance on both the minority and majority classes, making it a particularly appropriate metric for our problem.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2}$$

During the experimentation phase, an array of machine learning models was tested to identify the most effective algorithm for our study. The models under scrutiny included eXtreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP) and Logistic Regression (LR). To optimize the performance of each model, hyperparameter tuning was conducted using grid search techniques. This search ensured that the best possible set of hyperparameters was chosen for each model, thus maximizing their predictive accuracy. The evaluation results for the best-performing instance of each algorithm are summarized in Table 2.

TABLE II
EVALUATION OF MODELS

Model	macro-F1	Accuracy
XGBoost	**100.00%**	**100.00%**
SVM	47.37%	90.00%
RF	37.50%	60.00%
MLP	60.00%	70.00%
LR	60.00%	70.00%

As we can observe from the Table 2 and further corroborated by the confusion matrix in Fig 2, XGBoost excels with a perfect macro-F1 score and accuracy of 100%. This indicates its robust capability to accurately classify both minority and majority classes. MLP and LR show promise in classifying the minority class, as evidenced by their macro-F1 scores of 60%. However, their accuracy of 70% points to limitations in correctly classifying all instances of the majority class. In contrast, SVM and RF notably struggle with the minority class, which is reflected in their low macro-F1 scores.

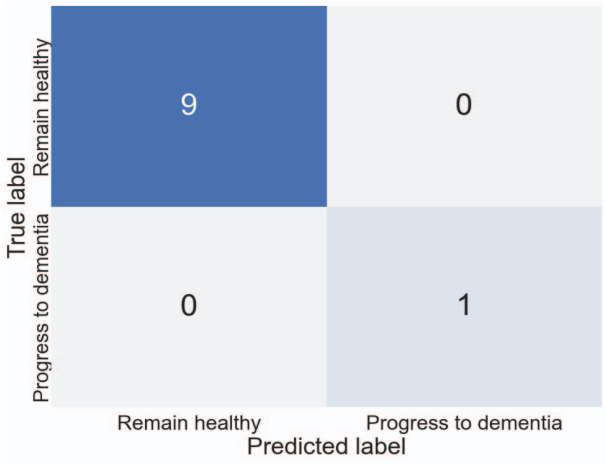

Fig. 2. Confusion matrix

2) Validation: To further verify the robustness of our model, which achieved a perfect score of 100% accuracy and macro-F1 on the test set, we embarked on an evaluation using various k-fold cross-validation techniques, as shown in Fig 3. With 3-fold cross-validation, the validation macro-F1 was 95.57%, while for 4-fold, it was slightly lower at 94.09%. As the number of folds increased, such as 6 and 9, the validation macro-F1 hovered around 94%. There was a slight dip to around 92% with 7-fold 13-fold validation. Despite the variations in the validation macro-F1 scores with different folds, a consistent trend emerged in the test results: our model consistently achieved a perfect 100% macro-F1 score across all k-fold tests.

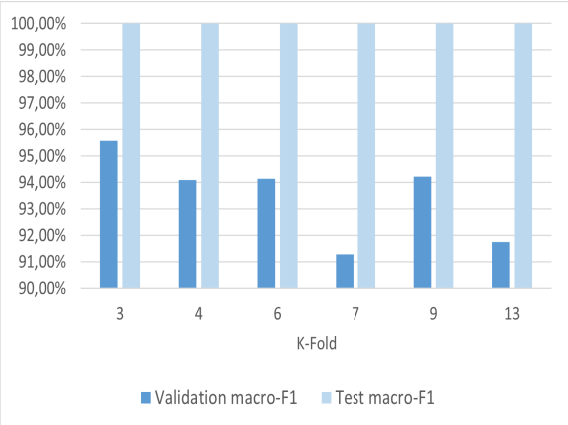

Fig. 3. K-fold cross validation

Seeking to delve deeper into the evaluation metrics of our binary classification model, we employed explainable AI techniques using SHapley Additive exPlanations (SHAP) [21], as shown in Fig 4. The model's reliance on key features aligns well with clinical insights [9]. eTIV, representing the Estimated Total Intracranial Volume, exhibits a nuanced relationship with dementia risk, with both high and low values playing pivotal roles. The MMSE score, indicative of cognitive function, also exerts varied influences on the predictions. Lastly, Age stands out as a multifaceted determinant. Their clinical relevance, combined with the model's accuracy, accentuates the model's robustness in predicting dementia progression.

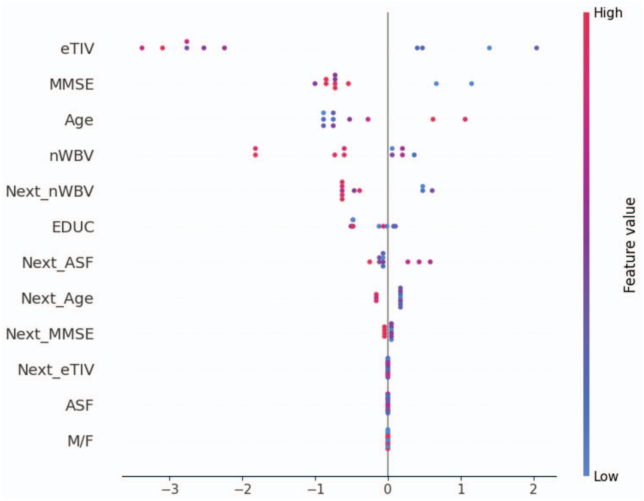

Fig. 4. Impact on model prediction using SHAP values

V. Conclusion & Future Work

Our study leveraging the OASIS-2 dataset to predict dementia progression using the XGBoost model has proven to be highly effective, achieving a macro-F1 score of 100%. The robustness of our model was not only emphasized through various k-fold cross-validation techniques but was also further validated using the SHAP explainable AI method. This approach spotlighted essential features such as eTIV, MMSE, and Age, which harmonize perfectly with well-established clinical insights.

Looking ahead, the horizon for this research is vast. To enhance model universality, dataset incorporation spanning diverse demographics and regions is essential. The potential in the longitudinal data can be harnessed using temporal modeling techniques like recurrent neural networks, capturing the dynamic nature of dementia over patient visits.

Transitioning from theory to practical application, real-world testing emerges as a crucial step. Collaborative clinical trials with health institutions can offer insights into the model's real-world applicability and areas for improvement. Engaging with neurologists, geriatricians, and specialists can ensure that technological innovations are deeply anchored in the complexities of dementia.

In essence, the exploration with the OASIS-2 dataset is just a starting point, with the overarching goal being to utilize machine learning's prowess to offer hope and solutions for individuals confronting dementia.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101095435 (REALM).

REFERENCES

[1] World Health Organization, "Fact sheets of dementia," Geneva, Switzerland: World Health Organization, 2022.
[2] Brayne C., Ferri C., Prince M., "Global prevalence of dementia: a delphi consensus study," in Lancet, 2005, vol. 366 (9503), p. 2112–26117.
[3] Sims I. M., "2009 Alzheimer's disease facts and figures," in Alzheimer's and Dementia, 2009, vol. 5 (3), p. 234–270.
[4] Csernansky G. J., Morris C. J., Buckner L. R., Marcus D. S., Fotenos F. A., "Open access series of imaging studies: Longitudinal MRI data in non demented and demented older adults," in Journal of Cognitive Neuroscience, 2010, vol. 22 (12), p. 2677–2684.
[5] Woodward M., Galea M., "Mini-mental state examination (mmse)," in Aust J Physiother, 2005, vol. 51(3), p. 198.
[6] Valero S., Vinyes-Junqu e G., Hern andez I,. Maule on A., Espinosa A., Alegret M., "A longitudinal follow-up of 550 mild cognitive impairment patients: evidence for large conversion to dementia rates and detection of major risk factors involved," in J. Alzheimer's Dis. 34, 2013, p. 769–780.
[7] De Rotrou J., Fabrigoule C., Pasquier F., Legrain S., Sarazin M., Berr C., "Amnestic syndrome of the medial temporal type identifies prodromal ad: a longitudinal study," in Neurology, 2007, vol. 69, p. 1859–1867.
[8] A. J. Mitchell and M. Shiri-Feshki, "Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies," in Acta. Psychiatr. Scand. 119, 2009, p. 252–265.
[9] Petersen R. C., Lindquist K.-Kramer J., Yaffe, K. and Miller B., "Subtype of mild cognitive impairment and progression to dementia and death," in Dement Geriatr. Cogn. Disord. 22, 2006, p. 312–319.
[10] Crystal H., Bang S., Dresner R., Grober E., Buschke H., "Screening for dementia by memory testing," in Neurology, 1988, vol. 69(19), p. 1859–1867.
[11] Amieva H., Berr C., Dubois B., Dartigues JF., Auriacombe S., Helmer C., "Validity of the free and cued selective reminding test in predicting dementia: the 3c study," in Dement Geriatr. Cogn. Disord., 2010, vol. 74(22), p. 1760–1767.
[12] Khan R. U., Rashid A. H., Khanna P., Prasad M., Lin C. T., Tanveer M., Richhariya B., "Machine learning techniques for the diagnosis of Alzheimer's disease: A review," in ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, vol. 16, pp. 1–35.
[13] Faghri F., Nalls M. A., Singleton A. B., Campbell R. H., Satone V., Kaur R., "Learning the progression and clinical subtypes of Alzheimer's disease from longitudinal clinical data," 2018.
[14] Aisen P., Beckett A. L., Clifford R. J., Jagust W., Weiner M., Dallas P. V., "The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception," 2013.
[15] Sheldon F. C., Choudhury K. R., Calhoun V. D., Coleman R. E., Doraiswamy P. M., Shaffer J. L., Petrella J. R., "Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers," in Radiology, 2013, vol. 266(2), pp. 583–591.
[16] Joo G., Yeshin K., Seongheon Ki., Gihwan B., Park S. W., Hosseinzadeh P., Park C., Jang J., "Predicting progression to dementia with "comprehensive visual rating scale" and machine learning algorithms," in Front. Neurol, 2022, vol. 13.
[17] Albright J., "Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm," in Alzheimer's Disease Neuroimaging Initiative, 2019.
[18] Nayeemulla K. A., Shanmuga S. V. E., Shahina A., "Dementia prediction on oasis dataset using supervised and ensemble learning techniques," in International Journal of Engineering and Advanced Technology (IJEAT), 2020, vol. 10 (01).
[19] Mamtaz M., Monirujjaman K. M., Morshedul B. A., Shafayet J. A., "A comparative analysis of machine learning algorithms to predict Alzheimer's disease," in Journal of Healthcare Engineering,, 2021.
[20] Bowyer K., Hall L. O., Kegelmeyer W. P., Chawla N., Nitesh V., "Smote: synthetic minority over-sampling technique," in Journal of artificial intelligence research, 2002, vol. 16, pp. 321–357.
[21] Shapley S. L., ""A value for n-person games," in Contributions to the Theory of Games, 1953, vol. 2 (28), pp. 307–317