Early Dementia Indicators using Feature Selection: Data and Explainability Consideration

https://doi.org/10.63962/XHBH8993

Talib Alshehhi
School of Computer Science and Informatics
De Montfort University
Leicester, UK
P2602836@my365.dmu.ac.uk

***Abstract*— This paper pinpoints few key research issues surrounding the application of feature selection techniques for early dementia detection, with a focus on Alzheimer's Disease Neuroimaging Initiative (ADNI) data. Dementia, particularly Alzheimer's Disease (AD), impairs memory, language, and executive functions, necessitating early diagnosis for effective intervention. Drawing on four years of study utilizing cognitive assessments and biomarkers, this work highlights critical considerations in selecting relevant features to enhance data driven model performance and improve interpretability. Additionally, it addresses data-related challenges within ADNI datasets, emphasizing their impact on model development. By identifying these gaps, the paper aims to guide future research in the use of feature selection processes, thereby supporting clinicians in making more informed diagnostic decisions.**

Keywords*— *Artificial Intelligence; Dementia Detection; Classification Models; Feature Selection

I. Introduction and Problem

This paper scope features selection for dementia early diagnosis in order to highlight possible research directions in this area after 4-year study using cognitive assessments and other common biomarkers. The aim is to show where artificial intelligence (AI) techniques particularly machine learning (ML) are heading in this sensitive health application. More essentially, the aim of this paper is to highlight important research considerations that should be taken when using ML such as feature selection for early dementia detection.

Dementia is associated with deficiencies in language, memory, learning, problem solving and executive functions, which interferes with individuals' daily life (Chaves et al., 2011). There are millions of people with the condition worldwide in which Alzheimer's Disease (AD) accounts for more than 60% of dementia conditions. Dementia impacts physical, psychological, social, and economic impacts in which its subtype AD tends to worsen over time therefore discovering cognitive and non-cognitive features that can help clinicians to early detect the disease is significant for an intervention, which is useful for patients and caregivers.

One of the promising approaches to detect dementia conditions is ML techniques that provide healthcare professionals with models developed from historical data to predict the disease and its level (Wharton et al., 2019). These models are usually produced by training on a population of data subjects diagnosed as cognitively normal (CN), having Mild Cognitive Impairment (MCI), or demented. Over the last two decades, ML techniques have been applied in medical applications because the diagnostic outcome is produced using models learnt objectively from data, offering a less-biased result for physicians to assess in reaching a final diagnosis (Maroco et al., 2011).

Part of the ML process is a step called feature selection, which is a process that involves identifying and a subset of relevant features from a larger dataset (Büyükkeçeci and Okur, 2022). This process aims to reduce data dimensionality, improve model performance, reduce overfitting, and enhance computational efficiency by eliminating irrelevant or redundant features (Pudjihartono et al., 2022). By focusing on the informative feature specially in healthcare applications like dementia detection, feature selection can lead to simpler and more understandable models while maintaining or even improving predictive performance (Chandrashekar and Sahin, 2014).

By unfolding features ranking and their correlations from dementia dat and presenting them as simple to understand knowledge clinicians and other decision makers can determine key features and their correlations besides removing redundant features.

II. Background on Feature Selection and Dementia

A. Dementia Diagnosis

Dementia diagnosis involves a structured and lengthy process to establish clinical diagnosis thereby few diagnostic frameworks have been developed by healthcare professionals including the DSM-5 (American Psychiatric Association, 2013), ICD-11 (Tyrer et al., 2011), and criteria developed by the National Institute on Aging (NIA) and Alzheimer's Association

(AA) (Jack Jr et al., 2012). This comprehensive approach ensures accurate classification of dementia from other cognitive and degenerative conditions.

The DSM-5 classifies dementia under major and mild neurocognitive disorders (NCDs). The diagnosis starts with identifying cognitive decline in cognitive domains such as memory, executive function, learning, language, perceptual-motor skills, or social cognition. The decline is assessed through patient medical history, informant reports, and cognitive assessment using standardized methods. Major NCDs, or dementia, require major interference with independence in daily life activities, while mild NCDs involve lesser decline. The ICD-11 complements the DSM-5 by offering a global framework for dementia diagnosis, highlighting both cognitive symptoms and their impact on daily functioning. This classification includes subtypes such as AD, vascular dementia, and frontotemporal dementia. However, where resources allow, some biomarkers and neuroimaging techniques, such as computed tomography (CT) scans or magnetic resonance imaging (MRI), are used to detect any possible brain atrophy. The dementia diagnostic process typically follows these steps:

1. Clinical Assessment: Gathering patient history, conducting interviews, and performing clinical observations.

2. Standardized Testing: Utilizing tools like the Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA) (Folstein et al., 1975; Nasreddine et al., 2005) to quantify cognitive decline.

3. Functional Assessment: Evaluating the impact on daily living through informant-based questionnaires or direct observation.

4. Exclusion of Other Causes: Ruling out conditions like depression, or medication effects.

5. Biomarker Utilisation: Incorporating advanced imaging or cerebrospinal fluid (CSF) analysis, if available, to support diagnostic accuracy.

The diagnosis of AD involves additional specificity based on the NIA-AA criteria and biomarker research. Alzheimer's diagnosis often progresses through three key stages:

1. Preclinical Stage: This stage occurs before obvious symptoms and is characterized by biomarkers indicating amyloid beta plaques, and tau pathology. Biomarkers are detected through CSF analysis or imaging techniques.

2. MCI due to AD: Symptoms become evident but remain mild, including memory lapses and difficulty with some complex tasks. Diagnosis combines clinical assessment, cognitive testing, and biomarker evidence.

3. Dementia due to AD: This stage includes significant cognitive decline and functional impairment, confirmed through detailed neuropsychological testing and imaging.

B. Dementia and Feature Selection

Feature selection is an important step in the data process, particularly when conducting medical data analysis, due to its ability to reduce dimensionality by eliminating redundant, or noisy features. This step enhances the model's interpretability and may improve the predictive performance of machine learning (ML) models (Azhagusundari and & Thanamani, 2013; Barbieri et al., 2024). By focusing on the important features, the process of feature selection allows clinicians to better understand the underlying biological processes, providing useful information for healthcare professionals in application like dementia pre-diagnosis (Rajab et al., 2023).

Feature selection methods are typically classified into 3 types: filter, wrapper, and embedded methods. Filter methods use mathematical criteria, such as mutual information and correlation analysis, to sort features independently and they do not require any ML algorithms. These methods are efficient sine the computational time is minimal besides their results are highly objective since they do not require algorithms' tuning or user involvement. Wrapper methods, in contrast, evaluate subsets of features by their impact on model performance that has been derived by the ML algorithm.

While they may generate subsets that when processed by a ML algorithm produces an accurate model, wrapper methods are computationally intensive, particularly for high-dimensional datasets (Zhu et al., 2016). Lastly, embedded methods integrate feature selection with model training, balancing computational costs and predictive accuracy. This approach of feature selection can generate feature sets that are suitable for supervised learning tasks but may need more complex tuning.

In dementia pre-diagnosis, feature selection is particularly important given the complexity and dimensionality of medical datasets, which often include biomarkers, neuropsychological data, imaging data, and clinical measures. Accurate isolation of relevant features, such as cognitive, CSF or structural changes in brain regions, can impact the data driven models' performance in classifying dementia or detecting disease progression (Tohka et al., 2016). Data-driven approaches rely on feature selection to identify features that may be overlooked in traditional diagnostic processes. For example, Rajab et al. (2023) showed the effectiveness of ML models incorporating feature selection techniques in improving the classification of Alzheimer's-related pathologies, using both imaging and clinical datasets. In addition, (Thabtah et al., 2022) used neuropsychological data from ADNI to identify impactful features that may affect AD advancements.

Additionally, feature selection specially filter based methods may reduce decision maker bias and ensures the data driven model's generalizability. By excluding irrelevant and redundant features that may increase the chance of model's overfitting in ML feature selection in dementia and other health related applications is vital (Bron et al., 2015). In the context of dementia, this step is crucial for enhancing the reliability of predictions across diverse patient populations. For instance, Zhu et al. (2016) highlighted the importance of graph-based and embedded feature selection methods in analyzing imaging data, noting that these techniques significantly improve the sensitivity and specificity of dementia diagnoses.

The integration of feature selection into a data process for dementia pre-diagnosis not only can improve predictive accuracy but also aids in uncovering critical disease mechanisms. As demonstrated by Bron et al. (2015), embedding

feature selection methods into support vector machines (SVMs) enhanced the classification of dementia cases, enabling more precise differentiation between Alzheimer's and other dementia conditions. Furthermore, Tohka et al. (2016) when compared various feature selection methods, emphasised their role in optimizing ML models for brain neuro imaging data, with notable improvements in diagnostic accuracy. Recently, Thabtah et al., (2023) showed how feature selection when embedded in a ML algorithm more key cognitive and functional features that are related to early AD progression are identified and utilised successfully to develop more accurate data driven models that can be exploited by clinicians in clinical settings.

III. Discussion

The datasets used in most dementia research that involves feature selection and data driven techniques is called ADNI, e.g. ADNI-merge. While ADNI datasets are widely used in Alzheimer's research, they poses some limitations when used for feature selection studies aimed at identifying early AD indicators. This dataset includes neuroimaging data, such as T1-weighted MRI and FDG-PET scans, and clinical datasets like cognitive scores, CSF biomarkers, and genetic data. While these features are important, they may partly represent the entire dementia spectrum of features necessary for early AD detection. The reliance on specific imaging data limits the inclusion of complementary data sources, such as diffusion tensor imaging (DTI) or longitudinal cognitive assessments. Furthermore, standardized preprocessing and uniform sampling methods, while ensuring consistency, might not consider diverse dementia subgroups or other environmental elements. The dataset's focus on clinical trial participants also tends to exclude minority groups, such as those with varied socioeconomic backgrounds, limiting its generalizability. These limitations can effect dementia diagnostic models.

To address the above limitations, more research is required to expand the data in ADNI repository to include a broader range of specific items in neuropsychological assessments and neuroimaging data. Including neuropsychological assessments' items in ADNI could provide a more in depth understanding of cognitive decline and at which level of dementia or pre-dementia like MCI. Moreover, longitudinal data property can capture changes over time and that may improve the ability of AI and ML models to identify early dementia progression so the need to capture the disease progression is fundamental.

Moreover, neuroimaging features related to functional MRI (fMRI) and amyloid imaging, can offer insights into brain pathology, complementing other used structural imaging data. By integrating neuropsychological specific items, fMRI, and biomarker pathological features, the data driven model would address specific issues related to understanding features of AD at the early stages of the disease. Features like APOE, tau protein levels, and biomarkers could further enhance the models by linking molecular findings with imaging and cognitive features, providing a comprehensive view of AD progression.

These future dementia pathology data inclusions would benefit ML diagnostic models by enabling them to analyze multi-modal datasets, leading to better identification of early dementia indicators. Incorporating these advanced features would allow data driven models to improve accuracy of the diagnosis and reduce overfitting. For clinicians, these enhancements would translate into knowledge that goes beyond conventional methods, offering a n intelligent AI diagnostic framework. Moreover, a more inclusive dataset would ensure that diagnostic models are relevant to a wider range of clinical scenarios, contributing to effective healthcare computer aided tools. In the future, we will expand the work to investigate also a critical factor that may contribute to AD progression based on the disease stages to find out which features may have larger associations with dementia and at which stage.

AI models that are explainable in nature with feature selection in AD diagnosis application is also a topic that needs to be explored further in the near future. AI models that are explainable can greatly assist clinicians in the pre-diagnosis stage of AD, especially when combined with robust feature selection methods like RMAM. Explainable AI (XAI) provides information into how specific features contribute to predictions, ensuring that clinicians can understand the model's recommendations. By highlighting these features in an interpretable manner, XAI models allow clinicians to validate the information and use it in developing intervention plans, reducing uncertainty in clinical decisions. Feature selection further enhances this process by ensuring that only the most relevant and interpretable features are included in the model, minimizing further feature redundancy.

The integration of XAI and feature selection enables models to provide explanations, such as ranking in detecting AD. This explanation can aid the process of retaining indicators and support healthcare by connecting specific features to potential interventions. For example, if a model identifies tau protein levels and certain cognitive decline as critical factors, clinicians can prioritize diagnostic tests or treatments targeting these areas.

IV. Conclusion

This paper highlights the role of feature selection in developing data driven models for early dementia detection, with focus on studies from the ADNI datasets offering a rich array of cognitive, imaging, and biomarker features. While ADNI provides a diverse foundation, one limitation remains the absence of specific diagnostic progression labels, which constrains research focused on modeling the continuous stages of disease advancement. As such, there is a need for more longitudinal studies and enriched labeling within ADNI to better capture disease trajectories over time. Additionally, relying solely on feature selection is insufficient to address the complexities of dementia diagnostics. Integrating explainable AI (XAI) models, such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and counterfactual explanations, can enhance clinician trust by showing how specific features such as tau protein levels or hippocampal volume contribute to predictions. These models offer transparent, case-level insights, enabling clinicians to validate model outputs. Future research should focus on enhancing ADNI datasets with detailed progression labels and multi-modal data integration, alongside employing deep learning with XAI techniques. This combined approach can drive the development of interpretable, robust models that support early dementia diagnosis and inform clinical interventions effectively.

REFERENCES

[1] American Psychiatric Association. (2013). Diagnostic and statistical Büyükkeçeci, M., & Okur, M. C. (2022). A comprehensive review of feature selection and feature selection stability in machine learning. Gazi University Journal of Science, 36(4), 1506-1520.

[2] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & electrical engineering, 40(1), 16-28.

[3] Chaves, M. L., Godinho, C. C., Porto, C. S., Mansur, L., Carthery-Goulart, M. T., Yassuda, M. S., ... & Group Recommendations in Alzheimer's Disease Vascular DementiaBrazilian Academy of Neurology. (2011). Cognitive, functional and behavioral assessment: Alzheimer's disease. *Dementia & neuropsychologia*, *5*(3), 153-166.

[4] Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189-198.

[5] Jack Jr, C. R., Knopman, D. S., Weigand, S. D., Wiste, H. J., Vemuri, P., Lowe, V., ... & Petersen, R. C. (2012). An operational approach to National Institute on Aging–Alzheimer's Association criteria for preclinical Alzheimer disease. Annals of neurology, 71(6), 765-775.

[6] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes*, *4*, 1-14.

[7] Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., ... & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695-699

[8] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics, 2, 927312.

[9] Rajab, M. D., Jammeh, E., Taketa, T., Brayne, C., Matthews, F. E., Su, L., ... & Cognitive Function and Ageing Neuropathology Study Group. (2023). Assessment of Alzheimer-related pathologies of dementia using machine learning feature selection. Alzheimer's Research & Therapy, 15(1), 47.

[10] Thabtah, F., Ong, S., and Peebles, D. (2022) Detection of Dementia Progression from Functional Activities Data Using Machine Learning Techniques. Intelligent Decision Technologies, vol. Pre-press, no. Pre-press, pp. 1-16, 2022.

[11] Thabtah, F., Mohammad, H., Lu, Y., & Zhang, B. (2023). Neuropsychological features evaluation of data related to Alzheimer's disease progression using feature selection. Intelligent Decision Technologies, 17(4), 1161-1178.

[12] Tohka, J., Moradi, E., Huttunen, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics, 14, 279-296.

[13] Tyrer, P., Crawford, M., Mulder, R., Blashfield, R., Farnam, A., Fossati, A., ... & Reed, G. M. (2011). The rationale for the reclassification of personality disorder in the 11th revision of the International Classification of Diseases (ICD-11). Personality and Mental Health, 5(4), 246-259.

[14] Wharton, S. B., Wang, D., Parikh, C., Matthews, F. E., Brayne, C., Ince, P. G., & Cognitive Function and Ageing Neuropathology Study Group. (2019). Epidemiological pathology of Aβ deposition in the ageing brain in CFAS: addition of multiple Aβ-derived measures does not improve dementia assessment using logistic regression and machine learning approaches. *Acta Neuropathologica Communications*, *7*, 1-12.

[15] Zhu, F., Panwar, B., Dodge, H. H., Li, H., Hampstead, B. M., Albin, R. L., ... & Guan, Y. (2016). COMPASS: A computational model to predict changes in MMSE scores 24-months after initial assessment of Alzheimer's disease. *Scientific reports*, *6*(1), 34567.s