Thesis: Supervised Machine Learning Assessment of Dementia Using Feature Selection Filter Methods

Mohammed Dabash A Rajab

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

March 2023

The University of Sheffield

Faculty of Engineering

Department of Computer Science

Supervisors: Professor Dennis Wang, Dr. Maria-Cruz Villa-Urioll

Declaration

Three manuscripts have been integrated into this thesis. I am listed as the primary contributor to three of them. Details of my manuscript contributions are noted at the beginning of each chapter. Neither the University of Sheffield nor any other university or educational institution has approved this thesis for use in a degree or qualification application.

Mohammed Dabash A Rajab

Mohammed Rajab

March, 2023

Acknowledgements

I spent almost 4 years working on this dissertation from conception to completion. It involved many meetings, conferences, and collaborations that wouldn't have been possible otherwise. Firstly, I would like to express my gratitude towards my primary supervisor, professor Dennis Wang, for his exceptional guidance and unwavering support throughout the entire process. This is even though the latter half was completed during the COVID years. This made everything more difficult and slower. My supervisor provided me with the chance to embark on this journey and provided me with constant mental support during countless meetings and discussions. Without his compassionate care and unwavering encouragement for my academic growth and well-being, I wouldn't have been able to reach this level of achievement. I couldn't have asked for a better supervisor who is both involved and supportive, and who has placed their confidence and time in me.

Additionally, I would like to express gratitude for all the discussions, and problem solving we had with my past and present lab mates. In addition to the conferences, the training, the teaching, and the manuscript submissions, the people in my research group made all these experiences worthwhile and educational. Acknowledgement should also be made to the contributions by the government of Saudi Arabia, Ministry of Education, who sponsored and offered me a scholarship for this project.

The most important thing for me is to acknowledge my family's contribution, not only to this PhD journey, but also to all my achievements. I would not have been able to do so without their confidence, help, support, and love. My brother Dr. Khairan D. Rajab, I would not have succeeded without his trust, help, support, and love. He always exhibited a positive attitude, even when I did not seem to be thinking positively. There is no one I am more grateful to than my mother “Rahma”. I am at a loss as to how I could ever thank her adequately for the countless things she has done for me. She consistently encouraged me to excel academically while also providing unwavering emotional support. The love and support of my cherished spouse “Alhanouf” and children “Sultan, Fahad, and Ilyas” have been the greatest source of inspiration throughout my journey, driving me to pursue excellence and make positive strides. I wouldn’t be here without them. I dedicated my PhD to my family and late father, as their love and support were my constant inspirations. Without them, I wouldn't have been able to make it this far.

Table of contents

Table of Contents

List of publications

Chapter 2

Manuscript 1

Rajab, Mohammed, and Dennis Wang. "**Practical Challenges and Recommendations of Filter Methods for Feature Selection**." *Journal of Information & Knowledge Management* 19.01 (2020): 2040019. https://doi.org/10.1142/S0219649220400195

Conference abstract

Rajab, Mohammed, and Dennis Wang. "**Practical Challenges and Recommendations of Filter Methods for Feature Selection**." (Special Issue, presented at *ICSDAA-2019, The International Conference on Intelligent Computing Systems and Data Analytics Applications, Jadara, Jordan*)

Chapter 3

Manuscript 2

Rajab, M.D., Jammeh, E., Taketa, T. *et al.* Assessment of Alzheimer-related pathologies of dementia using machine learning feature selection. *Alz Res Therapy* 15, 47 (2023). https://doi.org/10.1186/s13195-023-01195-9

Conference abstract

Mohammed Rajab, et al. "**Neuropathological Assessments of Dementia Using Machine Learning Feature Selection**." (Poster presented at *LOD 2021, 7th International Conference on Machine Learning, Optimization & Data Science, October 4-8, 2021 - Grasere, Lake District, UK*)

Conference Co-speaker

Mohammed Rajab, Teruka Taketa, Dennis Wang. "**Neuropathological Assessments of Dementia Using Machine Learning Feature Selection**." (presented at *The Khouribga IBRO Neuroscience School (KIBROS III 2021), The International Brain Research Organization, 2021 - Casablanca, Morocco*)

Chapter 4

Manuscript 3

Mohammed D Rajab, Teruka Taketa, Carol Brayne, Fiona E Matthews, Paul G Ince, Stephen B Wharton, Dennis Wang and on behalf of the Cognitive Function and Ageing Neuropathology Study Group & Alzheimer's Disease Neuroimaging Initiative. "**Machine Learning Feature Selection: Feature-feature correlation biases ranking of dementia features in machine learning studies**" (submitted to GigaScience)

Other contribution

Varsha Gupta, Sokratis Kariotis, **Mohammed Rajab**, Niamh Errington, Elham M Alhathli, Emmanuel Jammeh, Martin Brook, Naomi Meardon, Paul Collini, Joby Cole, James Wild, Steven Hershman, Roger Thompson, Thushan de Silva, Euan Ashley, Dennis Wang, Allan Lawrie, ***Unsupervised machine learning identifies and associates trajectory patterns of COVID-19 symptoms and physical activity measured via a smart watch*** (Submitted to *NPJ Digital Medicine*)

Non-Manuscript Abbreviations

AI	Artificial intelligence
AD	Alzheimer’s disease
ADNI	Alzheimer's disease neuroimaging initiative
ADAS-Cog	Alzheimer's disease assessment scale-cognitive subscale
CFAS	Cognitive function and ageing studies
CDT	Clock drawing test
DLB	Dementia with Lewy body
GR	Gain ratio
FAST	Functional assessment staging
FCBF	Fast correlation-based filter
FD	Frontotemporal dementia
ML	Machine learning
MMSE	Mini-mental state examination
MoCA	Montreal cognitive assessment
MCI	Mild cognitive impairment
MRI	Magnetic resonance imaging
mRMR	Minimum redundancy maximum relevance
MRLR	Maximum relevance less redundancy
PET	Positron emission tomography
SU	Symmetrical uncertainty
R^2	R-squared
VD	Vascular dementia
TemRMR	Temporal minimum redundancy maximum relevance

Abstract

The prevalence of dementia is increasing globally. Due to the massive resources required, this issue is pressuring governments and private healthcare systems. Accurate diagnosis by clinicians on the cause of dementia, such as Alzheimer's disease (AD), is difficult because of the time and assessments needed like neuropathological. The issue becomes more challenging when considering if various brain lesions contribute to the pathological assessment of dementia, the relationship of these lesions to the various dementia conditions, how they interact, and how to quantify them. Thereby, systematically assessing neuropathological measures by their degree of association with dementia, especially AD, may lead to better diagnostic systems and treatment targets. One promising approach that can answer these challenges is to develop data-driven solutions with core functions of feature evaluation and automatic subject classification based on machine learning (ML).

Recent research studies in medical diagnosis, including dementia research, reveal that ML techniques, when used with feature selection, can identify critical features of Alzheimer-related pathologies and their association with the disease's diagnosis and prognosis. The feature selection removes noisy features from the dementia data to increase the predictive performance and improve interpretability while reducing the dimensionality and computational complexity. However, filter-based feature selection methods can generate dissimilar feature rankings and may be sensitive to the correlations among themselves.

This thesis investigates dementia with a focus on AD neuropathological assessments from a data-driven perspective to develop mechanisms to assist pathologists during these clinical assessments. The thesis investigation comprises phases such as feature ranking, feature-feature correlation, and classification. The work determines the impact of neuropathological feature-features correlations on the feature ranking for better biomarker identification.

The investigation assesses real datasets related to dementia, the Cognitive Function and Aging Studies (CFAS) and the Alzheimer's Disease Neuroimaging Initiative (ADNI), using filter methods and classification techniques. The results showed that classification models generated from the CFAS and ADNI sets of chosen neuropathological features were strong in terms of sensitivity, accuracy, and other measures when mined by different classification techniques. In the ADNI dataset results, the significant neuropathological features contributing to AD included neocortical neuritic plaques, Braak stage, Thal phase, diffuse plaques, and cerebral amyloid angiopathy (CAA), all of which showed a high correlation with AD's diagnostic label. In the CFAS dataset, the results were consistent with those derived from the ADNI dataset. Moreover, among the filter methods considered, reliefF had the strongest correlation with feature-feature correlations in both ADNI and CFAS datasets, less sensitive to feature-feature correlations. However, no filter method had clear dominance over ADNI results. More essentially, the results indicated limited consistency in feature rankings between ADNI and CFAS. However, reliefF had the most agreement, while the Gain Ratio method had less consistency in ranking the features in both datasets.

In summary, this thesis provided valuable insights into the application of filter methods and neuropathology data for developing classification models for dementia conditions' diagnosis. The study demonstrated the significance of considering feature-feature correlations when selecting influential features and the impact of different filter methods on feature ranking and classification performance. These findings suggest that the proposed approach could effectively minimise the discrepancy of feature ranking and generate an impactful set of features for classification algorithms. These results had practical implications for pathologists in improving the understanding of AD pathology. Furthermore, the study has highlighted the potential for future research to leverage diverse filter methods to identify more reliable biomarkers and enhance the detection of dementia, particularly for AD.

Chapter 1 - Introduction

1.1. Study Background

The study focuses on a specific healthcare application domain: dementia pathology and diagnosis. Dementia is a broad term that describes a group of symptoms with decline in cognitive function that interferes with daily activities [1,2]. Some symptoms accompanying dementia include memory loss, difficulty with language, disorientation, mood swings, and problems with motivation and self-care [3,4]. The most prevalent cause of dementia is AD. About 60%–70% of dementia is considered AD, which makes it the most common cause of dementia [2]. In the UK, 650 thousand people were estimated to have dementia as of 2015, costing £23.0 billion, and expected to be 1.3 million in 2040, costing £80.1 billion [5]. As of 2018, 50 million people were living with dementia worldwide [6], with the number expected to increase to 82 million by 2030 [7] and to 150 million by 2050 [6].

The relationship between the cognitive assessment of dementia and neuropathology assessment of brains is vital to understand the progression of dementia [8]. Cognitive assessment is a crucial way to diagnose dementia, which involves observing and measuring a person's ability to think, remember, and reason [9]. However, neuropathological assessment, which entails examining brain tissue and cells under a microscope, is essential for studying the changes in the brain caused by the condition [9]. Therefore, this research study focuses on neuropathological assessments by examining data collected from tissue samples from the cadavers of individuals with dementia, mostly AD, to identify the underlying cause. I used Machine Learning (ML) techniques such as feature selection filter methods and classification algorithms to conduct my investigation.

To determine effective neuropathological indicators, which was one of the aims of this research work, the study used feature selection considering the associations between each feature ranking when considering the diagnosis class, and pairs of feature-feature correlations. Through analysing the correlation between each feature's ranking and its correlation with other features, it is possible to identify and eliminate any potential redundancies, as well as identify which filter method was particularly sensitive to feature-feature correlations. To achieve this aim, a data process using two different datasets was used from longitudinal population-based studies: the Cognitive Function and Ageing Studies (CFAS) [10–12] and Alzheimer's Disease Neuroimaging Initiative (ADNI) [13]. CFAS and ADNI collect neuropathological assessments related to AD.

The purpose of this study was to investigate ML approaches for classifying dementia. Details of the aims and research questions are given later in this chapter.

1.2. Dementia

As stated earlier, 'dementia' is a term used to describe a group of symptoms usually associated with decline in cognitive ability, and sometimes with functional impairment, that can interfere with an individual's activities of daily living [1,2]. In earlier times, it was widely accepted that ageing caused dementia and considered it part of an unavoidable natural process [14]. However, in 1906, Alois Alzheimer examined the brain tissues of a 50 years old who had died with dementia, showing age may not be the determinant factor. This form of dementia was later named AD by Emil Kraepelin [15,16]. The other common forms of dementia are vascular dementia (VD), dementia with Lewy body (DLB), frontotemporal dementia (FD), and other dementias. Symptoms of dementia can include [2,17]:

- Memory loss, particularly for recent events
- Difficulty with language, such as finding the right word

- Disorientation in time and space
- Challenges with problem-solving and planning
- Difficulty with coordination and motor functions
- Changes in mood and behaviour
- Difficulty with self-care and performing daily activities

Dementia, recognized as a major neurocognitive disorder, is a term used to include conditions that are formed because of abnormal changes in a human's brain structure that gradually reduces the individual's cognitive ability, and is not typically caused by ageing [18]. Dementia appears to be of considerable international concern, with repeated estimates predicting significant global increases in the following decades [19]. According to the World Health Organization, dementia is a major cause of dependency among ageing communities worldwide and a primary cause of death [2,20].

There have been many improvements in medical services for people with dementia, so retaining consistency in diagnostic and methodological practice is challenging, which may contribute to further shifts in occurrence and incidence. An early and accurate diagnosis of common dementia such as AD can help individuals and their families plan for the future and access support and services [2,20]. There is no cure for dementia, and interventions are focused on managing symptoms and helping individuals maintain their independence for as long as possible [21,22].

Dementia research is a challenging field due to the complexity of the contributing factors and a lack of understanding of the underlying causes. One challenge is that dementia involves multiple symptoms that can occur as a result of various underlying conditions, making it difficult to develop effective interventions that target the underlying cause [23–25]. Another challenge is that dementia often occurs in the elderly, who may have other health conditions and take several medications [14]. This can make it difficult to determine the specific cause, and to develop targeted treatments.

The diagnosis of dementia diseases can also be challenging because symptoms can be similar to other medical conditions and there is no single test that can differentiate dementia [23]. This makes identifying people with dementia early, when intervention may be more effective, a difficult task. Additionally, there is a lack of funding for dementia research posing a challenge to develop effective treatments [26–29]. Ongoing research efforts are helping to improve our understanding of dementia and to develop new treatments that can improve not only the lives of individuals with the disease, but also their families' lives [2,30]. This study targets the issue of dementia pathology on real neuropathological datasets to develop an accurate classification system, and to isolate a small number of influential neuropathological indicators.

1.2.1. Cognitive assessment of dementia

The assessment of a patient's cognition involves considering multiple features including the patient's history, and using cognitive tests such as the Mini-Mental State Examination (MMSE) for the screening of dementia [31]. Each person has a unique degree of intellect and education that may contribute to greater cognitive reserve; therefore, several cognitive techniques used to identify cognitive decline for dementia and other cognitive impairment conditions [23]. Some research studies reported that people who have spent more time in education are more likely to have a ceiling effect on many cognitive tests, making it harder to tell whether they have dementia or mild cognitive impairment (MCI), which is a precursor of dementia [32]. MCI is diagnosed when symptoms are considered not severe enough to interfere with an individual's daily life [33].

- The MMSE test is a 30-point questionnaire that assesses a person's cognitive abilities in areas such as memory, language, and attention, which is usually conducted in a clinical setup by a trained healthcare professional. Other cognitive tests used in the assessment of dementia include: the Montreal Cognitive Assessment (MoCA) [34]he Clock Drawing Test (CDT) [35]he Functional Assessment Staging (FAST) Scale [36]he Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) [37].

In general, the clinician may use one or more cognitive tests to assess different cognitive areas to decide whether the person may exhibit dementia.

1.2.2. Neuropathology assessment of brains

Neuropathological assessments of brains involve the examination of brain tissue samples taken during an autopsy, which shows the structure and biochemical changes that occur in the brain as a result of illnesses such as dementia conditions. Neuropathologists utilise diverse techniques to examine abnormal protein clusters and cellular damage patterns in brain tissue via microscopic and imaging methods where the information is used to describe the changes that take place in the brain as a result of specific diseases or conditions.

There are characteristic neuropathological changes in AD, often characterised by the accumulation of amyloid-beta peptide (Aβ) in the medial temporal lobe and neocortical structures of the brain. This accumulation leads to the formation of neuritic plaques and neurofibrillary tangles (**Figure 1.1**), ultimately causing neurodegeneration [38,39]. AD is characterised by distinct neuropathological changes in the brain. The primary hallmarks include amyloid plaques, which are extracellular deposits of amyloid β protein, and neurofibrillary tangles, formed by the accumulation of hyperphosphorylated tau protein inside neurons. Additionally, there's noticeable brain atrophy, particularly in the hippocampus, accompanied by neuronal and synaptic loss. Neuroinflammation, marked by the activation of glial cells releasing inflammatory substances, further contributes to the degeneration observed in AD [39–43].

Several forms of dementia, such as VD, FD, DLB and others, have specific neuropathological characteristics.

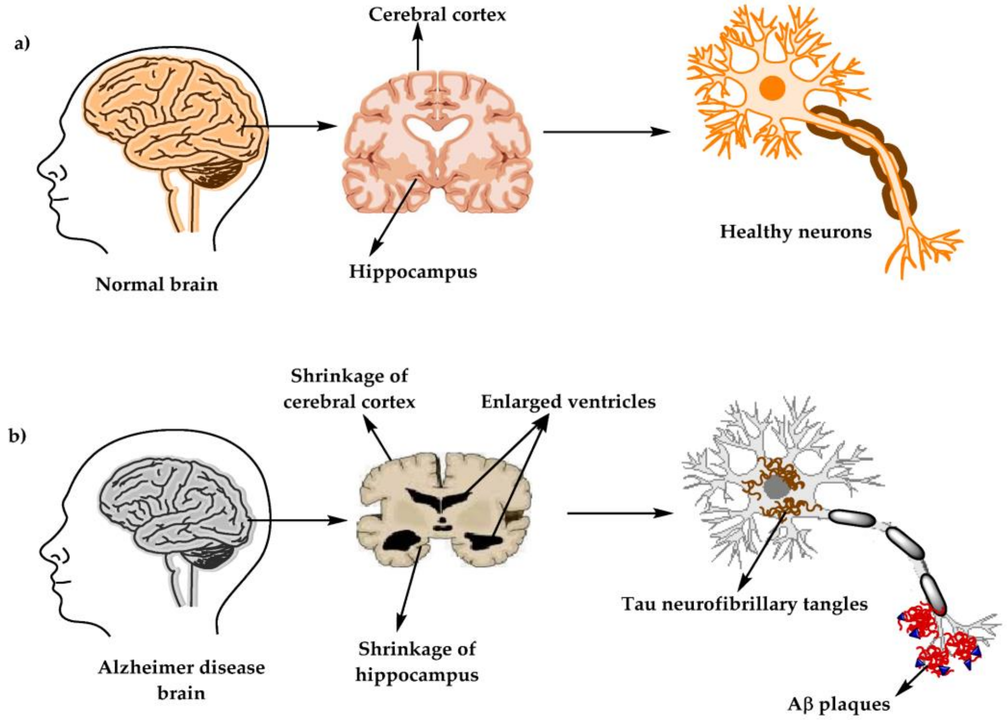


Figure 1.1: The physiological structure of the brain and neurons in (**a**) healthy brain and (**b**) Alzheimer’s disease (AD) brain [38].

1.2.3. Neuropathological tests related to dementia

Neuropathology is a subfield of pathology that deals with the study of diseases of the nervous system [44]. This includes both structural and functional changes in the brain and spinal cord due to various pathological conditions [45]. As stated earlier, the diagnosis of dementia conditions like AD is complex, often requiring a combination of clinical, and neuropathological tests. Neuropathological tests help to determine the underlying cause(s) of dementia and to identify specific brain changes [46].

There are various tests used in neuropathology for assessing neuropathological features (**Table 1.1**). This thesis is concerned with the neuropathological features that are carried out post-mortem, some of which are depicted in **Table 1.1**, and all are described in **Chapter 3** (**Table 1** – Section 3.3 excluding demographic, and general non-neuropathological features). These tests help to diagnose various neurodegenerative diseases such as AD, Parkinson's disease, and multiple sclerosis. The combination of different tests provides a comprehensive evaluation of the nervous system and helps to determine the most appropriate treatment strategies. Some of the tests related to dementia diagnosis are described below.

- **Autopsy:** This is an invasive technique. A post-mortem examination of the brain and nervous system to identify any abnormalities, injuries or diseases.

- **Magnetic resonance imaging (MRI):** A imaging technique that uses a powerful magnetic field and radio waves to create detailed images of the brain and spinal cord [47].

- **Computed tomography (CT):** A diagnostic non-invasive imaging test that uses X-rays and computer technology to create cross-sectional images of the body’s internal structures [47].

- **Positron emission tomography (PET):** A type of imaging test that uses a radioactive tracer to visualise and measure physiological processes in the body, often used to detect cancer, heart problems, and brain disorders [47].

- **Cerebrospinal Fluid Analysis (CSF):** This is an invasive technique. A clear fluid that circulates in the brain and spinal cord. Analysis of the CSF helps to identify various neuropathological changes such as inflammation, infection, and the presence of abnormal proteins in the nervous tissue [48].

Table 1.1: Descriptions of neuropathological and general pathological features

Feature	Description
Braak stage	The Braak stage is a system used to describe the progression of AD based on the spread of neurofibrillary tangles, which are a hallmark of the disease. The Braak stages range from 0 to VI, with stage 0 indicating no tangles and stage VI indicating widespread tangles in the brain [42,49].
Thal phase	The Thal phase is a specific stage in the Braak staging system for AD. It refers to the stage where neurofibrillary tangles are found in the hippocampus, a region of the brain involved in memory and learning [50,51].
CAA	Cerebral amyloid angiopathy (CAA) is a condition in which amyloid protein deposits build up in the walls of small and medium-sized blood vessels in the brain [12,51].
Brain atrophy	Brain atrophy refers to a decrease in the size of the brain, which can occur due to various factors such as ageing, neurodegenerative diseases, injury, or lack of oxygen.
Microinfarcts	Microinfarcts are small, localised areas of tissue damage in the brain that can occur as a result of decreased blood flow to a specific area. They are often considered a sign of underlying cerebrovascular disease and have been associated with an increased risk of dementia and cognitive decline [52].
TSA	Subpial thorn-shaped astrocytes are a type of glial cells that are found in the brain and spinal cord. They are unique in their shape and arrangement and have been implicated in AD and other neurological disorders [53–56].
BrainNet tau stage	The BrainNet Tau stage is a system used to stage the progression of tau pathology in the brain, particularly in AD. Tau is a protein that helps maintain the structure of neurons but in certain conditions, it can become abnormal and form clumps called neurofibrillary tangles. The BrainNet Tau stage ranges from 0 to 5, with higher stages indicating more advanced tau pathology.and the level of tau protein in the cerebrospinal fluid [50].
Lewy bodies	Lewy bodies are abnormal structures that form inside the brain cells of individuals with certain neurodegenerative disorders. Lewy bodies are made up of a protein called alpha-synuclein, and their accumulation in the brain is thought to play a role in the development of these disorders.
Neuronal loss	Neuronal loss refers to the reduction in the number of neurons in the brain.
Aβ stage typical	Aβ stage typical refers to the typical stage of amyloid-beta (Aβ) accumulation in the brain in AD. Aβ is a protein that is involved in the formation of plaques in the brain. The stage of Aβ accumulation ranges from 0 to 5, with higher stages indicating more advanced Aβ

	accumulation.
PART	Primary age-related tauopathy (PART) refers to a group of neurodegenerative disorders characterised by the accumulation of abnormal tau protein in the brain, which is associated with ageing. Tau protein is involved in maintaining the structure of neurons and is essential for normal brain function [57].
Infarcts and lacunes	Infarcts and lacunes are types of brain lesions that can occur in various neurodegenerative disorders, including stroke and Parkinson's disease.
Argyrophilic grain disease	Argyrophilic grain disease is a neurodegenerative disorder characterised by the accumulation of abnormal protein deposits in the brain, known as argyrophilic grains. These deposits are composed of tau protein and are often found in areas of the brain that are involved in memory and learning.
Diffuse plaque	Diffuse plaque refers to the accumulation of beta-amyloid protein in the brain, a hallmark of AD. Beta-amyloid protein is a fragment of a larger protein that accumulates in the brain and forms clumps, or plaques, between nerve cells.
Arteriolar sclerosis	Arteriolosclerosis is a type of arteriosclerosis, which is a disease that affects the walls of arteries. Arteriolosclerosis specifically affects the arterioles, which are small arteries that branch off from the larger arteries and supply blood to the capillaries.
Atherosclerosis	Atherosclerosis is a disease that affects the arteries and is characterised by the buildup of fatty deposits, known as plaques, in the arterial walls. These plaques can restrict blood flow and increase the risk of serious health problems, such as heart attack, stroke, and peripheral artery disease.
Neocortical neuritic plaques	Neocortical neuritic plaques are formed by the accumulation of amyloid beta (Aβ) protein in the brain. These plaques cause damage to neurons and disrupt communication between brain cells.
Haemorrhage	Haemorrhage, also known as bleeding, can occur anywhere in the body and can be caused by a variety of factors, including injury, disease, or abnormalities in the blood vessels.
Gliosis	Gliosis is the process by which glial cells, which are the supportive cells in the brain, increase in number and size in response to injury, disease, or other forms of brain damage.

1.2.4. Neuropathology assessment of post-mortem brains

Neuropathology is the study of disease and injury of the nervous system using techniques such as microscopic examination of brain tissue, both fresh and fixed [58,59]. Neuropathological assessment of brains is a crucial component of the diagnosis of various neurological disorders, including dementia and other degenerative conditions, traumatic brain injury, and infectious diseases [46,59]. The goal of neuropathological assessment is to identify and describe specific changes in the brain that are indicative of a particular disease or condition [46].

As described in Section 1.2.3, neuropathological assessment starts with a thorough search of the patient's medical history, followed by a clinical examination [45]. Next obtaining a sample of brain tissue, either through an autopsy or a biopsy procedure following the patient's death. The tissue is fixed in formalin and embedded in paraffin for sectioning and staining, or it may be frozen for a later time. The tissue is examined under a microscope to help identify specific changes in the brain. Changes in the brain can include presence of amyloid plaques and neurofibrillary tangles in AD, or presence of inflammation or neuronal loss in other types of dementia. Other techniques, such as immunohistochemistry and electron microscopy, can also be used to provide more detailed information about the changes that are present in the brain [58]. Overall, neuropathological assessment of the post-mortem brain is a complex and multi-step process requiring expertise in both pathology and neuroscience. It can, however, provide valuable information about the underlying causes of many neurological disorders, which is important for the diagnosis of these disorders.

1.2.5. Clinical Diagnosis of Dementia Gold Standard

The DSM-5 framework [60] explains the standards used to establish an AD possible or AD probable diagnosis. Initially, a level of neurocognitive disorder or dementia is established. Major ND entails that the individual has a major deterioration over time in at least one of the six cognitive areas: executive function, complex attention, language, perceptual-motor, learning and memory, or social cognition. The deterioration can be reported by an informant, physician, or the patient. Cognitive decline is then tested using one or more cognitive tests—such decline must affect the independence of the individual while performing daily activities.

To establish minor ND, a reasonable deterioration in cognitive domains is observed over time, although the independence of the individual in performing daily activities is not impacted. A diagnosis of Minor ND or Major ND also requires that the cognitive deterioration is not noticed when the individual is delirious, and that it cannot be better elaborated by other mental conditions. Once Minor or Major ND is established, the DSM-5 framework can be used to determine whether the cognitive decline is caused by AD (Possible or Probable). To diagnose Possible or Probable AD, besides genetic testing and family history, the memory and learning domains must show gradual decline in addition to at least one other cognitive area, to be considered. More details on the differences between Possible and Probable AD can be found in [33].

For a final diagnosis of dementia, including AD, more difficult and in-depth criteria are considered. For example, for a diagnosis of AD in the ADNI study, subjects have to initially be diagnosed with probable AD according to the measures used by the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [61]. Probable or Possible AD are established based on specific criteria outlined earlier by the American Psychiatric Association (2013) in which a level of ND must be established first based on six cognitive areas in addition to evaluating the level of independence of the subject when performing daily activities.

1.3. Biomedical Applications of Machine Learning

ML is a subfield of the scientific field of Artificial Intelligence (AI) that concentrates on how computers learn from provided data [62–64], without being explicitly programmed. One of the crucial and efficient technologies for processing complicated medical data is ML [65]. Feature selection is a technique used in ML and a critical processing step of a data process to pinpoint a relevant set of features that maximise the performance of predictive models specially for classification benchmarks [66,67]. Feature selection is an important step that real world data may require because most real datasets are messy, unstructured and contain high dimensionality hence pre-processing the data to make the learning process feasible is a necessity [68]**.**

According to Alelyani et al. (2013); Dong and Liu (2018), selection of features is a process that directly affects the learning algorithm performance in processing classification tasks [69,70]. The quality of the data, which depends on the input features, may affect the learning process based on the levels of noise [71,72]. Noisy data involving redundant features may hinder the classification model performance in terms of predictive rate and time. Thereby, using feature selection on data to achieve desirable outcomes, is essential [71–73].

ML algorithms can analyse large medical datasets that consist of patient information, electronic health records, medical assessments, etc., to identify patterns and predict outcomes. The outputs of the learning algorithms can help researchers identify patients at risk of certain conditions, such as dementia, and develop more cost–effective prevention strategies. By analysing medical images, such as magnetic resonance imaging (MRI) and positron emission tomography (PET) scans, ML algorithms can identify knowledge that cannot be seen by the human eye and is fundamental to decision making. Since the development of accurate diagnostic tools based on the knowledge discovered, researchers can identify early signs of diseases like AD. Moreover, ML can be used to develop decision support systems that help physicians and other healthcare providers make better decisions about patient care.

1.3.1. Feature selection

Due to the complexity, unstructured nature, and high dimensionality of real datasets, pre-processing these datasets is needed to improve learning outcomes [68]**.** Features selection is defined as reducing the dimensionality of a dataset by eliminating irrelevant, redundant, or noisy features [74,75]. Additionally, the process of feature selection may be beneficial in terms of improved interpretability, reduced overfitting, and shorter training times [76–79]. One of the main advantages of feature selection in medical data is improved interpretability [71–73]. When reducing the number of features, the generated solution by the learning algorithm becomes easier to understand by the end-user.

In the case of healthcare applications like for classifying dementia, the results of the models need to be transparent and easily explainable to the clinicians and pathologists. Additionally, interpretability would help to identify the vital features in the data where those features would provide insights into the underlying biology and inform further studies. It is also possible to improve generalisation in medical data with feature selection by reducing overfitting [79], where a model is too complex and fits too closely to the training data, resulting in biassed performance on unseen data [80,81]. This is particularly important in dementia pathological assessment applications where the amount of available data is limited, as overfitting can lead to misleading predictions.

According to Chowdhury and Turn (2020), feature selection can reduce the risk of bias in medical data analysis [82]. A number of features may also be associated with demographic variables or other confounders that may lead to bias in analyses. In order to reduce the risk of bias in the analysis, feature selection can be used to select only relevant features and to remove features that are associated with demographics or other confounders.

In recent years, several feature selection methods have been proposed including filter methods, wrapper methods, and embedded methods [74,83,84] (**Figure 1.2**). Filter methods are based on pre-defined feature ranking criteria, such as mutual information, chi-squared test, or correlation, to select the most informative features. Filter methods are fast and computationally efficient, but they do not consider the classifier's performance [85–87]. Wrapper methods, on the other hand, use the classifier performance as a criterion for feature selection, by evaluating the performance of the classifier on different feature subsets. Wrapper methods are computationally expensive and time-consuming, but they tend to provide more accurate results yet they may become infeasible when the dimensionality of the data is high [86,88,89]. Embedded methods integrate the feature selection and classifier training processes into a single framework. Embedded methods are less computationally expensive compared to wrapper methods and can lead to improved results by considering the classifier's performance yet they are more complex and still less efficient than filter methods [90].

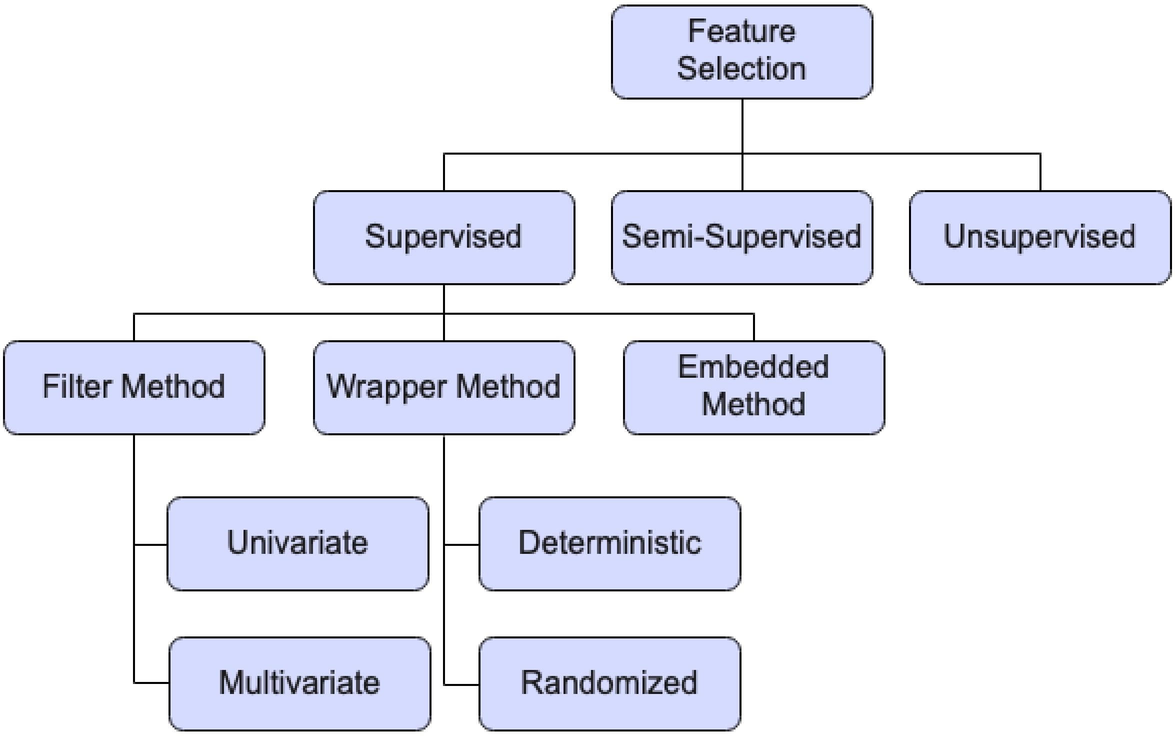


Figure 1.2: Categorization of Feature Selection Algorithms [91]. Univariate filter methods consider each feature independently, and rank them according to some metric such as correlation with the target variable, mutual information, or statistical significance. These methods are fast and computationally inexpensive, but may miss important interactions or dependencies between features, such as chi-squared test, correlation-based feature selection. Multivariate filter methods, on the other hand, consider the joint distribution of all features and their relationship with the target variable. These methods are more powerful than univariate methods in capturing complex interactions, but can be computationally expensive and may require larger datasets, such as principal component analysis (PCA), linear discriminant analysis (LDA).

Each feature selection type has its strengths and weaknesses, and the choice of the ideal method depends on the study's specific requirements and the data's characteristics [74,78,92]. Further research is needed for feature selection methods specially for medical applications because of the density of the features and the data dimensionality besides addressing the existing methods' limitations [78]. Therefore, the thesis focuses on ML feature selection related to filter methods since such methods are non–biased, quick, and they are not dependent on classification algorithms on performance metrics.

It can be challenging to identify the significant pathological factors in the diagnosis of dementia, because pathologists may interpret the brain's features differently [42,93,94]. This is essential for improving the investigation of dementia cases. According to Thabtah *et al.* (2022) and Rajab *et al.* (2022), identifying the specific sets of features associated with dementia pathology is difficult due to the large number involved in the diagnosis of dementia, including biological markers and other brain imaging results. These features are collected by healthcare professionals during pathological procedures, and are used to determine dementia pathology. However, there is a lack of consensus on which features are essential for identification of dementia conditions such as AD, which is one of the main aims of this research [95,96].

Since the diagnosis of dementia is complex, costly and resource demanding it has become crucial to employ data related methods. In recent years, there has been a growing trend in the use of feature selection to select relevant features for accurate dementia classification [97,98]. Bharati *et al.* (2022) explored influential features for the detection of AD. The authors used feature selection methods with learning algorithms to detect the relevant features from MRI images, and clinical data to enhance AD detection rate generated by classification algorithms. The study concluded that feature selection when used before classification step in a data process the predictive rate of the models developed improved when compared to models generated without feature selection or models of traditional non–driven medical methods [97]. A study by Mahendran and Vincent PM (2022) presents a deep learning framework that incorporates embedded-based feature selection for early detection of AD. The authors evaluated the deep learning methods in identifying AD by analysing data related to MRI images. According to the results of the study, the deep learning method outperformed other ML methods in the early diagnosis of AD [98].

1.4. Current Challenges

The presence of high dimensional data has rendered the task of feature selection challenging due to the need to process a vast number of features, which poses efficiency and quality-related difficulties. However, these difficulties present opportunities for exploring and developing innovative intelligent techniques to produce a meaningful and concise set of features. In this section, I discussed various challenges that researchers and domain experts may face when designing, employing, or developing filter methods for data processing.

1.4.1. Automated models for dementia pathology

Currently, the diagnosis of dementia conditions, including AD, is a lengthy and labour-intensive process that involves medical assessments and in-depth investigation. This process is often expensive and time-consuming and can be challenging for medical professionals and patients alike [99]. Additionally, the accuracy of current diagnostic methods is not always reliable, which can lead to misdiagnosis, particularly in the early stages of the disease when symptoms may be subtle [100,101]. By introducing an automated system into the dementia diagnostic process, it would be possible to improve the accuracy, affordability, equipment, medical staff, and data [102–104]. Automated systems are also able to process and analyse a large volume of data quickly and efficiently, which can help to reduce the amount of time and resources needed to diagnose dementia. In addition, automated systems can be easily integrated with existing medical systems, allowing for greater accessibility to data and a more streamlined process for diagnosis. Developing an automated system for dementia diagnosis can revolutionise how dementia is diagnosed and treated, providing a more objective, and reliable diagnostic process [102–104].

1.4.2. Relationship between neuropathological features and dementia

Dementia is a growing healthcare concern among the elderly, and an accurate and timely diagnosis may provide opportunities for treatment. However, dementia, especially in older people, is associated with multiple brain pathologies, making it challenging to assess interactions among them [105,106]. By analysing the neuropathological features of post-mortem brains, ML methods can identify cases where dementia status and neuropathological features differ, such as those related to Aβ-related assessments and tau. It would be valuable to identify cases of dementia with inadequate pathology when certain features are not informative. This can help to reduce resources, such as time, cost, and effort, during pathological assessments and highlight the need for more extensive clinical evaluations.

ML algorithms and feature selection techniques have enabled automated ways of classifying heart and skin diseases, and studies investigating dementia involving brain imaging have utilised ML algorithms for the diagnosis of AD and VD [107–109]. The CFAS studies focus on cognition, and neuropathological research has investigated the correlation between dementia phenotypes and pathological characteristics in the brain, such as measures of tau and beta-amyloid (Aβ) pathologies [110]. The analysis of brains donated from the CFAS showed considerable overlap in the burden of lesions between participants dying with and without dementia [105,106]. Attributable risk showed the importance of many other pathologies in the brain [44,111].

1.4.3. Features ranking

Feature ranking refers to the process of selecting '*n*' number of features based on their computed scores. The scores are normally computed based on a feature's relevancy to the class variable. According to Venkatesh and Anuradha (2019), feature ranking is an independent evaluation process of the available features as per their importance to eliminate potentially irrelevant features [83]. The majority of filter methods evaluate the features based on scores computed using statistics, information theory, or some functions of the classifier's output. Gain Ratio (GR), Symmetrical Uncertainty (SU), and ReliefF methods are examples of filter methods that use a ranking function to sort features. Feature ranking is a crucial step in feature selection, commonly used by domain experts to determine the best feature subsets. However, filter methods do not provide the number of features to be selected, leaving the decision to the user's experience and knowledge.

Existing filter methods typically display features with their ranks, adopting a rudimentary approach that requires careful consideration and accuracy, often resulting in a time-consuming process. A significant challenge in filter methods is the discrepancy in results obtained from applying different methods to the same dataset [84,96]. This challenge arises due to the use of different mathematical models by filter methods to compute the weights per feature in the dataset. These models typically use a contingency table that holds the frequency of the feature and feature-class together, including observed and expected probabilities, among others. Gómez-Ramírez *et al.* (2020) aims to improve the accuracy of MCI prediction by identifying the most relevant features from a large set of self-assessed variables. The study suggests that different feature selection methods may produce different results, indicating the challenge of selecting the most appropriate method for a given dataset. Additionally, the study highlights the challenge of interpreting the results of feature ranking and making meaningful clinical inferences from the selected features [112]. A study by Haider *et al.* (2020) evaluates the potential of paralinguistic acoustic features for detecting Alzheimer's dementia in spontaneous speech. The study highlights several challenges associated with feature ranking. These include the large number of available features and the potential redundancy among them, which can lead to difficulty in selecting the most informative features. They also note the importance of selecting appropriate feature selection methods and avoiding overfitting. Furthermore, the authors note that the selected features may not be clinically meaningful, and there may be challenges in interpreting the results and translating them into clinical practice [113].

1.4.4. Feature-to-feature correlations

Most of the available filter methods do not consider feature-to-feature correlation when determining the optimal subsets during feature analysis. Valuing this is important because it helps to reduce the number of features and provides a set that contains features that do not overlap in data instances and are different from each other. This will be vital in medical applications like dementia pathology in which several neuropathological, cognitive and biomarkers are investigated by healthcare professionals like pathologists and clinicians. Identifying dissimilar related features to dementia diagnosis will allow healthcare professionals to focus on the critical dementia indicators thus minimising time and valuable medical resources. In cases where predictor variables exhibit a high degree of correlation, a phenomenon known as "multicollinearity," the ability to discern how a model makes its predictions and which variables are most influential in determining the outcome can be complicated. This is especially problematic in medical contexts where interpretability is crucial for clinical decision-making. In a recent investigation conducted by Lombardi *et al.* (2022), the impact of multicollinearity on the dependability and consistency of explainable artificial intelligence markers for mild cognitive impairment and AD was examined. The study found that multicollinearity can exert a significant negative influence on the performance of ML models, particularly with regards to their stability and reliability [114]. Specifically, when predictor variables are highly correlated, the model may struggle to distinguish between their individual contributions to the outcome variable, resulting in less accurate predictions and less trustworthy feature importance rankings.

Limited research investigations highlighted the importance of identifying feature-to-feature correlation to enhance the performance of the overall feature selection process. The study by Yu and Liu (2004a) is one such attempt that addressed the need to incorporate a redundant feature analysis process as relevancy is insufficient to determine the best subsets [115]. The authors introduced a novel mechanism called fast correlation-based filter (FCBF) by selecting relevant features and then identifying predominant features from the selected set to enhance the selection process through a redundancy analysis. Another attempt was mRMR method [116], which defines relevant features as those with minimum redundancy with each other while maintaining the maximum relevance with the class label with mutual information as a parameter [116,117].

Other research studies have used metrics to identify the intercorrelation among the features to produce optimal feature subsets such as the study of Radovic *et al.* (2017), which enhanced the mRMR method by dealing with temporal data (TemRMR) [118]. TemRMR uses the value of F-statistics across different time as the parameter to compute the temporal information and relevancy among features; this is by applying a dynamical time-warping approach to handle temporal gene expression data in an effective manner. Temporal gene expression occurs when data encoded within the gene is turned into a function (product) at a specific time [119].

F-statistics values determine redundant features by identifying features with small and large inter-class variances. Gu *et al.* (2012) presented a more relevance less redundancy method (MRLR) that uses mutual information, conditional mutual information, and relevance degree to eliminate redundant features [120].

1.4.5. Dementia feature importance and consistency

One of the major issues in dementia pathology research using data driven methods is the reliability of the results obtained on the features reported especially with the limited datasets available for researchers that contain neuropathological indicators. Increasing the reliability of the features ranking results obtained from different datasets can indeed be an indicator of the data goodness and the features consistency degree when measuring neuropathological indicators found in different datasets. The problem is obvious when researchers may obtain different results in regards to feature significance to dementia diagnosis when using the same neuropathological features during the feature engineering phase of the data process.

To ensure that the study under consideration is non-biased toward a single data repository, and there is some agreement on features consistency in ranking, we adopted two neuropathological cohorts of subjects in an attempt to measure the consistency of feature ranking results obtained. Specifically, dementia pathology datasets related to CFAS and ADNI [10–13] have been sourced with a concentration on neuropathological features. Due to the different correlation and covariance structures found in different studies, applying filter methods to dementia datasets may result in biassed ranking results for the features. Therefore, investigating the consistency in the neuropathological ranking derived by filter methods from two different cohorts is an essential step that increases the validity and reliability of the reported results.

1.4.6. Filter methods sensitivities

There is a complex relationship between the degree of similarity between the features and the correlation between each feature and the class label in classification datasets. The former shows features, which are similar, and the latter reveals features that are more critical to the class, e.g. features that are critical to dementia pathology in our case. There are mathematical methods like chi square testing, Spearman correlation, and others that can measure feature-feature correlation and feature-class correlation as described earlier. However, there may still be a lack of research studies on investigating filter method sensitivity to the ranking of features obtained by these methods.

Filter method ranks the relevance of features in a dataset based on their impact on the target variable. However, the sensitivity of the filter method to various factors can impact its reliability and performance. One such factor is the feature-feature correlations, which refers to the extent to which the correlation between the features in a dataset affects their ranking by the filter method [121]. In cases where there is a high degree of correlation between features, the filter method may assign high ranks to multiple features that are highly correlated with each other. This may occur even if only one of them is truly relevant to the target variable. Such overfitting can lead to poor generalisation performance of the resulting model, which is a crucial aspect of model evaluation.

Alirezanejad et al. (2020) studied the impact of feature correlations on filter-based feature selection methods for medical datasets [122]. They used three heuristic filter feature selection methods and found that the performance of classification models was sensitive to feature correlations, with highly correlated features leading to a decrease in accuracy. Different feature selection methods also responded differently to feature correlations. A study by Remeseiro et al. (2019) reviewed feature selection methods in medical applications, focusing on the sensitivity of filter methods [78]. The authors noted that filter methods are sensitive to feature distribution and correlation, often selecting correlated features together which can cause redundancy and overfitting. Filter methods can also be biassed towards high or low variance features depending on the chosen criterion. To mitigate these issues, the authors recommended preprocessing the data to remove correlated features and standardise the features. They also suggested using a combination of filter methods to reduce bias and improve robustness. Another study by Khagi et al. (2019) evaluated the performance of different ML techniques and feature selection methods for AD classification based on the Clinical Dementia Rating (CDR) level [123]. The authors observed that the performance of the classification models was sensitive to the ranking of features obtained by the feature selection methods, particularly, the accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve of the classification models varied significantly with different feature ranking orders. The study highlighted the sensitivity of filter methods to the ranking of features and the need for careful consideration when selecting feature selection methods for this task.

One of the aims of this study is to unfold the association between feature ranking computed by filter methods and feature-feature correlation to determine which filter methods are more sensitive to feature ranking, and using two different dementia datasets. Specifically, the study reveals filter methods that are less sensitive when considering the similarities between the neuropathological features themselves to reduce any feature ranking discrepancy hence utilising the less redundant subset of features by the classification algorithm during construction of the predictive models of dementia.

1.5. Hypothesis, Aims, and Research Questions

The hypothesis of this thesis is that developing an automated model using ML with feature selection for AD diagnosis would minimise the disparity in feature ranking, account for feature-feature correlations, and produce a succinct set of significant features for classification algorithms. I aimed to develop and test the approach on two different dementia neuropathological datasets from longitudinal population-based studies CFAS and ADNI with the following aims:

A. **Apply filter methods to assess AD-related pathologies in a large cohort of elderly individuals, after conducting an in-depth review (Chapter 2 & 3).** The aim is to investigate filter methods, and then apply them intelligently on dementia-related data to identify critical features of Ad-related pathologies. I used classification techniques and filter methods for feature ranking to compare neuropathological features and their relationship to dementia status in a cohort of 186 individuals from CFAS. This provides valuable insights into the potential of using ML techniques for the assessment of AD-related pathologies of dementia.

B. **Compare sensitivity of filter methods to different neuropathology datasets (Chapters 4).** Discrepancies can occur when we use different filter methods to rank features. This is especially problematic when we examine patients from two different studies, such as CFAS and ADNI. I aimed to develop a single feature score that reduces the volatility in generating different scores by filter methods.

C. **Measure the impact of feature-feature correlation on the ranking of features (filter method sensitivity to feature ranking), and then test classification approaches to determine whether they can better classify dementia (Chapter 3 & 4).** The ranked features obtained from filter methods were evaluated using classification algorithms in order to determine whether the classifiers can explain cognitive decline using neuropathology features. The models when used by pathologies are able to detect, and explain dementia pathology better than conventional medical methods.

The thesis addresses the following research question:

1. How can we rank the various dementia condition features in an unbiased way to facilitate ML besides determining redundant information?
2. What is the smallest subset of neuropathological features needed in an ML model to explain dementia using real data?
3. Which filter methods are less sensitive to feature-feature correlations?
4. Is there a difference between two cohorts of ageing individuals (ADNI and CFAS) in terms of the association between feature-feature scores and feature rankings?

1.6. Thesis Structure

Chapter 2 reviews practical challenges for feature selection filter methods and background of neuropathological assessments and features. The chapter outlines the practical challenges associated with implementing filter methods for feature selection to examine the challenges that end-users may encounter when using filter methods for feature selection. The review examined the difficulties and limitations of these methods, as well as suggested recommendations based on previous experiments and studies. (The Chapter has been disseminated in *Journal of Information & Knowledge Management, 23 March 2020*).

Chapter 3 details the contribution to the assessment of Alzheimer-related pathologies in dementia through the use of feature selection filter methods and other ML techniques. The chapter specifically focuses on the application of these methods to the CFAS cohort for assessing neuropathological measures and their degree of association with dementia. To achieve this, filter methods were utilised to evaluate AD-related pathologies in a large group of elderly individuals. We identified essential features that are associated with AD-related pathologies and contribute to dementia diagnosis. The Chapter further describes the use of classification techniques and filter methods for feature ranking to compare neuropathological features with dementia status in the cohort of 186 individuals from CFAS. This analysis provided significant insights into the potential of using machine learning techniques for the assessment of AD-related pathologies in dementia. (The Chapter has been published by the *Journal of Alzheimer's Research & Therapy, 10 March 2023*). Furthermore, **Chapter 4** focuses on the investigation and analysis of feature-feature correlation on rankers in the CFAS and ADNI datasets, along with the identification of the filter method that is less sensitive to feature-feature correlations. (The Chapter has been submitted to the *Journal of GigaScience, 2023*).

The results, and their analysis in this thesis are distributed in **Chapters 3 & 4**. Specifically, in **Chapter 3**, the use of ML techniques and filter methods has been evaluated to measure the ranking of neuropathological features, and to compare neuropathological features with dementia status in the cohort of 186 individuals from CFAS. This analysis provided significant insights into the potential of using ML techniques for the assessment of AD-related pathologies in dementia research. More essentially, empirical results that compare the sensitivity of filter methods to various neuropathology datasets (CFAS, ADNI) and emphasise the potential discrepancies that can arise when using different filter methods to rank features, are discussed at the end of **Chapter 4**. These possible discrepancies are challenging when comparing patients from two separate studies, such as CFAS and ADNI. Hence, in **Chapter 4** we tested a single feature score that reduces the volatility in generating different scores by filter methods, and can reveal sensitive methods to feature ranking. This approach can help mitigate the discrepancies in feature ranking, account for feature-feature correlations, and produce a concise set of influential features for ML algorithms. All ML techniques are compared using known performance measures in ML such as sensitivity, specificity, predictive accuracy, and others.

Chapter 5 of the thesis covers the conclusion and future directions section. The chapter provides an overview of the contents and results of each previous chapter and offers final concluding remarks. It also highlights the limitations of the study and identifies potential avenues for future exploration. The concluding remarks emphasise the importance of the study's contribution to the growing body of research on machine learning techniques and feature selection filter methods for dementia assessment. Finally, the chapter outlines potential directions for future research in this area, such as expanding the sample size and incorporating additional features to enhance the effectiveness and generalizability of the developed techniques.

Chapter 2 - Practical Challenges and Recommendations of Filter Methods for Feature Selection

2.1. Background

The main goal of this article was to explore the difficulties and suggestions regarding the use of filter methods for feature selection in data analysis. We detailed the benefits and drawbacks of filter methods and explored various obstacles such as choosing the right feature selection techniques, the consequences of data dimensionality, and the impact of feature correlation. Additionally, we provide recommendations for overcoming these challenges, such as employing multiple methods for feature selection, scaling features, and handling multicollinearity. We provide useful insights for researchers and practitioners who want to utilise filter methods for feature selection in data analysis.

2.2. Contribution

The following version of the accepted manuscript was published in the *Journal of Information & Knowledge Management* 19.01 (2020): 2040019. For this publication I was the first author who conceptualised, conducted critical analysis and writing of the manuscript. My supervisor Dennis Wang assisted with the conceptualisation and editing of the manuscript.

2.3. Manuscript 1

Practical Challenges and Recommendations of Filter Methods for Feature Selection

Mohammed Rajab
Department of Computer Science
The University of Sheffield, Sheffield, UK
mdrajab@gmail.com

Dennis Wang
Department of Computer Science
The University of Sheffield, Sheffield, UK
Sheffield Institute for Translational Neuroscience, Sheffield, UK
NIHR Sheffield Biomedical Research Centre, Sheffield, UK
dennis.wang@sheffield.ac.uk

Abstract

Feature selection, the process of identifying relevant features to be incorporated into a proposed model, is one of the significant steps of the learning process. It removes noise from the data to increase the learning performance while reducing the computational complexity. The literature review indicated that most previous studies had focused on improving the overall classifier performance or reducing costs associated with training time during building of the classifiers. However, in this era of big data, there is an urgent need to deal with more complex issues that makes feature selection, especially using filter-based methods, more challenging. This in terms of dimensionality, data structures, data format, domain experts' availability, data sparsity, and result discrepancies, among others. Filter methods identify the informative features of a given dataset to establish various predictive models using mathematical models. This paper takes a new route in an attempt to pinpoint recent practical challenges associated with filter methods, and discusses potential areas of development to yield better performance. Several practical recommendations, based on recent studies, are made to overcome the identified challenges and make the feature selection process simpler and more efficient.

Keywords: Feature selection; Filter methods; Machine Learning; Data Imbalance; Ranking methods

Introduction

The curse of dimensionality is one of the challenges that domain experts often face when dealing with massive amounts of data (Town & Thabtah, 2019). Feature selection is a critical processing step that directly affects the success of machine learning algorithms by reducing space dimensionality through identifying the relevant set of features to be used (Hall, 2000). It also involves simplifying the classification process by strengthening the decision rules of the feature selection algorithm (Kamalov & Thabtah, 2017). Feature selection plays a vital role in classification because a robust feature selection mechanism can reduce the computational complexity associated with the learning process and improve its generalisation capabilities (Maldonado *et al.*, 2014). Domains characterised with a large number of features and small number of samples benefit immensely through feature selection mechanisms. For instance, domains such as biochemistry, bioinformatics, text mining, medical diagnosis, and biomedicine require robust feature selection algorithms to improve the performance and comprehensibility of the models; these are often established based on a few samples and a large number of features (Yu & Liu, 2004a; Saeys *et al.*, 2008; Thabtah & Peebles, 2019).

Filter, wrapper and embedded are the three primary types of feature selection methods used for learning purposes. The filter method is the most common and involves selecting features without utilising a classification algorithm. Basically, this method involves filtering out irrelevant features using various selection principles such as information gain (IG) (Rajab, 2017). Filter methods use selection criteria to assign scores for the available features in the training dataset and then invoke a ranker search method to rank each individual feature based on the computed scores (Tang *et al.*, 2014). Informative features usually gain higher scores and uninformative features gain lower scores. Finally, the complete features, ranked on computed scores, are offered to the end user for subset selection. Based on the selection principles used, there are various filter-based feature selection methods such as IG (Quinlan, 1986), Pearson Correlation (Hall, 1999) and Fisher's Score (Gu *et al.*, 2012), among others. Wrapper methods consider using a machine learning algorithm to identify classifiers for each possible subset in the input dataset. Hence, this kind of feature selection offers the best outcome yet suffers from a lengthy, exhaustive search, particularly when the input data is highly dimensional (Thabtah *et al.*, 2018). Lastly, embedded methods use a combination of filter and wrapper methods to select an ideal set of features. This research is concerned only with filter-based methods.

Several research studies have evaluated filter-based methods, i.e. Thabtah *et al.*, (2011, 2018), Rajab, (2017), Zhang, *et al.*, (2014), Estevez *et al.*, (2009), Hall, (2000), Zhao *et al.*, (2018), Kamalov & Thabtah, (2017), and Hancer *et al.*, (2017). However, most of these investigated functional issues with filter methods such as the impact on predictive performance, or enhancing training efficiency; few covered practical challenges related to the basis on which features are selected and how results can be interpreted (Cherrington *et al.*, 2019). For example, a drawback of the filter methods, such as result dependencies, which make it hard for the end user to decide which features to choose prior to the learning process, has been investigated by few scholars. These combine results of multiple filter-based methods to reduce results variability, i.e. Labani, Moradi *et al.*, (2018); Gao *et al.*, (2018); Rahmaninia and Moradi, (2017). Despite this effort, recent research (Cherrington *et al.*, 2019) pinpointed that there is a need for a domain expert to manually check the outcomes of filter-based methods to recommend the final set of features needed; this can be resource demanding. More importantly, the authors indicated that there is no fine line to discriminate among features in the results sets which can also be a serious issue. Hence, this research covers practical challenges in filter-based methods and presents viable recommendations to overcome these issues. Particularly, this research builds upon previous efforts and possible research directions rarely covered including feature ranking, results discrepancies, thresholding, feature-to-feature correlation, domain expert involvement, and data imbalance.

The paper consists of multiple sections. The Introduction provides an overall understanding of the feature selection process, filter-based methods, aims, objectives, and the outline of the paper. The second section further explains the research problem and previous related work by various scholars. Discussion, the third section, critically analyses the potential challenges of filter-based feature selection methods with practical recommendations to overcome identified challenges. The conclusion wraps up the information provided with suggestions on future work.

Problem and Literature Review

Filter-based feature selection is a research topic that has attracted the attention of many scholars and experts in multiple domains. **Figure 1** shows filter methods in the learning process. The filter method involves carrying out feature selection as a pre-processing step without an induction algorithm. Training data is processed through a mathematical criterion to compute and assign scores to features in the training dataset, then a feature score is used to rank the features. These feature scores vary based on the type of the filter method used, and all the feature scores/rankings are offered to the end-user to make relevant decisions. Domain experts, or the end-user, decide the features to be used in the learning process based on their computed scores. The optimum threshold between selected and eliminated features is determined by the end-user based on knowledge and experience. Finally, a machine learning approach is employed to process the results set of the features and produce the classifier. The accuracy and the performance of the established classifier is evaluated by applying the model on sample data.

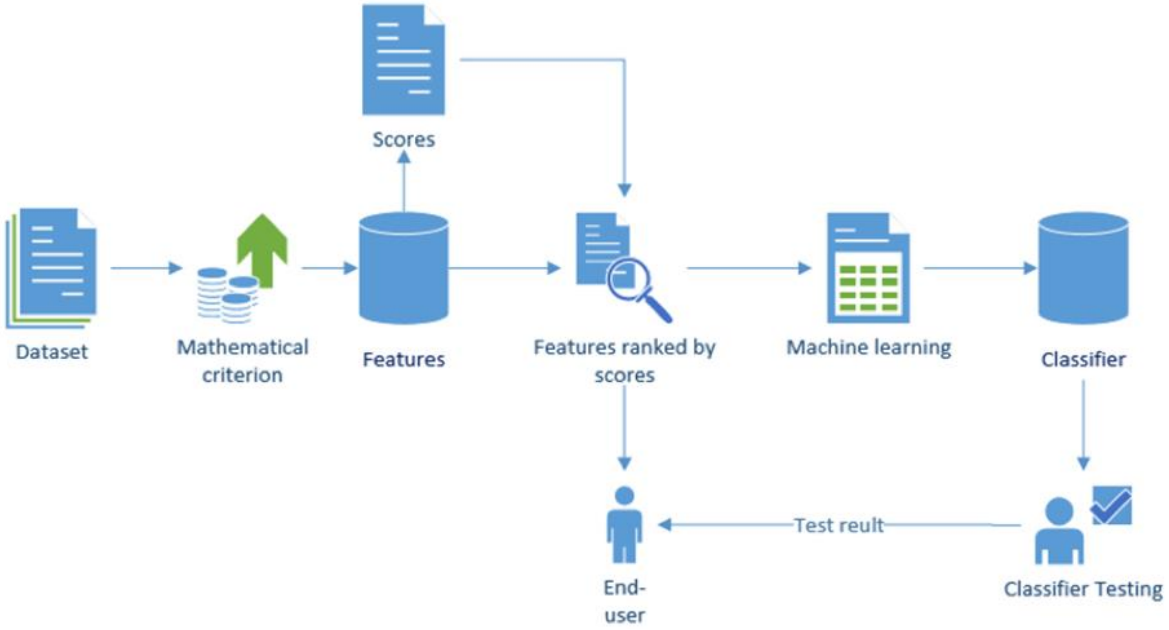


Figure 1: Filter Method as Part of the Learning Process.

Thabtah *et al.*, (2019a) introduced an observed frequency-based feature selection method called Least Lost (L2) to reduce the dimensionality of data by eliminating noisy data from the datasets while maintaining a healthy classifier performance. It is a more simplified and in-built approach that involves ranking of each variable in ascending order based on the L^2 distance between observed and expected variables and class labels. The scores are computed based on observed and expected probabilities of the available features. Tests conducted using datasets from the University of Irvine Repository (UCI) reported that L^2, when applied in the pre-processing phase, results in fewer features being obtained. When these are further processed by a machine learning algorithm, they derive competitive classifiers in terms of accuracy. L^2 implementation in Java can be accessed at https://github.com/suhelhammoud/L2.

Zhao *et al.*, (2018) proposed the Redundant Penalty between the Feature Mutual Information algorithm (RPFMI), a filter-based feature selection mechanism, to identify optimal features in terms of redundancy, relationship between classifier and the selected features, and the correlation between selected features and the class labels and small data samples. The experimental results of the study suggested that the proposed RPFMI is highly effective in selecting an optimal set of features for Intrusion detection as it demonstrated a high accuracy.

Gao *et al.*, (2018) introduced the Dynamic Change of Selected Feature (DCSF), with the class a linear filter feature selection method, which takes dynamic information changes of the selected features with the class labels into account in the feature selection process; this to yield more accurate and efficient results. This novel model uses conditional mutual information between candidate features and class labels to identify the most informative features; the other conventional filter methods use mutual information to compute the relevancy of the candidate features to the select optimal feature subset. The experimental results implied that DCSF has the highest average classification accuracy of all the other compared methods.

Another filter mechanism presented by Hancer *et al.*, (2017) is quite unique. These authors focus on selecting features based on their true rankings obtained by applying ReliefF (Robnik-Šikonja & Kononenko, 2003) and Fisher Score (Bishop , 1995) rather than focusing on their mutual redundancies. MIRFFS (Mutual Information, ReliefF, and Fisher Score), the proposed mechanism used Differential Evolution (DE) (Marinaki & Marinakis, 2013) as the search strategy and it has two parts: one mechanism to be applied on single-objective problems and the other on multi-objective problems.

Labani *et al.*, (2018) introduced multivariate relative discrimination criterion (MRDC), a novel filter-based feature selection mechanism to enhance the performance of the text classification process. This is accomplished by diminishing the dimensionality in feature space using minimal-redundancy and maximal-relevancy (mRmR) (Peng *et al.*, 2005). MRDC involves identifying the most relevant features using relative discrimination criterion (RDC) (Rehman *et al.*, 2015). Since, RDC is not capable of classifying the irrelevant features, it utilises the Pearson correlation matrix to perform that task.

Kamalov and Thabtah (2017) used three robust filter methods in combination to produce a new feature selection mechanism (vectors of scores/ *V*-score) to select the most relevant features of a given dataset while eliminating the shortcomings and maximising the advantages. They used information gain (Quinlan, 1986), chi-squared statistic (Liu & Setiono, 1995), and inter-correlation methods (CFS) (Hall, 1999) together to stabilise each feature's ranking score; they were able to reap more accurate prediction results rather than when applying them individually.

OSFSMI (Online Stream Feature Selection Method based on Mutual Information) and OSFSMI-k is another mutual information-based online streaming feature selection method, presented by Rahmaninia and Moradi (2017), to distinguish between the most informative and uninformative features. This is done by computing the correlation between features and their relevancy to the class labels where the number of instances increases exponentially (for example, social networks, finance analysis applications, and traffic network monitoring systems). The general framework followed by the proposed OSFSMI model, comprises two unique phases: online relevancy analysis to compute the relevancy of each newly arriving feature, and online redundancy analysis to estimate the effectiveness of each selected feature and eliminate any with effectiveness below the average. OSFSMI-k is a modified version of OSFSMI, developed to address the issues arising due to the continuously increasing nature of features. To end this, OSFSMI-k keeps selecting the correlated features until the size of the selected feature subset reaches a constant value (*k*).

A research by Estevez *et al.*, (2009) proposed a normalised mutual information feature selection (NMIFS), to evaluate the relevancy and redundancy in the features of a given dataset. Researchers have used three mutual information-based feature selection methods: Battiti's mutual information feature selector (MIFS), MIFS-U (Battiti, 1994), and min-redundancy max-relevance (mRMR) (Peng *et al.*, 2005) criteria to develop NMIFS by enhancing their individual strengths and minimising their weaknesses. They also present the Genetic algorithm, guided by mutual information for feature selection (GAMIFS), a hybrid version of both the filter and wrapper methods that combines NMIFS and genetic algorithms to fine-tune their performance.

Filter Methods Challenges

High dimensional data have made feature selection difficult as it necessitates dealing with a large number of features during data processing creating multiple challenges related to efficiency and quality. These challenges can be opportunities to learn and investigate new intelligent techniques to generate a meaningful concise set of features. In this section, we discuss various challenges that researchers and domain experts may face when designing, employing, or developing filter methods for data processing.

Results discrepancies

Results discrepancy is one of the obvious challenges in filter methods as different results may be obtained from the same dataset when applying different methods. To demonstrate this issue, we applied three different filter methods: IG, Correlation, and ReliefF (keeping Ranker as the search method) on a nursery database (Bohanec *et al.*, 1997) using WEKA 3.8 (Hall *et al.*, 2009). **Table 1** shows the features extracted by the three considered filter methods and their ranks based on the assigned weights. The nursery dataset is developed to sort the parents' applications for nursery school using a number of features related to the parents' job, the financial situation of the family, family structure, health status of the family, and social aspect.

Table 1 clearly shows differences in the results generated by the filter methods, especially the ranking. For instance, if we consider the results derived by the IG and Correlation methods, after the third ranked feature there is a discrepancy in the results for the remaining features ranked 4–8. This discrepancy arises mainly because of the different mathematical models used by the considered filter methods to compute the weights per feature in the dataset. All these mathematical models primarily employ a contingency table that holds the frequency of the feature and that of the feature-class together, besides observed and expected probabilities among others. For example, IG uses entropy as a base metric to compute the weights; this relies on the information of the feature and the class in the dataset, whereas the chi-square method uses the observed and expected probabilities. These differences in computing the weight assigned to each feature in the mathematical model can clearly impact the order in which the final feature sets are offered to the end-user. Consequently, when these feature sets are processed by the learning algorithm, performance may also be impacted such as the predictive accuracy of the models derived.

Table 1: Ranking results generated by each feature selection method

Ranking	IG Features	Correlation Features	ReliefF Features
1	Health	Health	Health
2	Has_nurs	Has_nurs	Has_nurs
3	Parents	Parents	Parents
4	Social	Housing	Housing
5	Housing	Social	Social
6	Children	Finance	Finance
7	Form	Children	Form
8	Finance	Form	Children

Few studies have addressed this issue and presented viable solutions to stabilise the knowledge discovery process through robust feature selection methods. For example, Kamalov and Thabtah (2017) pinpointed the results discrepancy in filter methods and showed that this problem can lead to selecting the wrong feature subsets thus impacting the performance of the classification models derived by the learning algorithm. The authors suggested a filter mechanism that involves combining and normalising IG, Inter-correlation, and CHI feature scores to produce one unified score that can be assigned to each available feature. The term 'normalizing' refers to the introduction of one unified feature score range instead of several that vary according to the feature selection method used. For instance, feature selection methods like IG produce data scores ranging from 0 to 1, whereas methods like CHI produce feature scores between (-1) and (+1). The experimental results demonstrated that the normalisation of feature scores, and then integrating these into one unified score, is highly effective in reducing the volatility in the feature selection outcomes.

A similar approach that deals with the results discrepancy of filter methods was proposed by Rajab (2017). The author presented a method that combines the score of IG and CHI after normalising the initial scores computed by both methods. The new feature selection method was applied on a cybersecurity application for detecting phishing websites and contrasted with other common filter methods. Results reported that Rajab's (2017) method indeed reduced the dimensionality of the dataset and selected features sets, and when processed, using decision trees and rule induction classification techniques, improved the detection rate of phishing websites.

Feature ranking

Feature ranking refers to the process of selecting '*n*' number of features based on their computed weights/scores. The weights are normally computed based on a feature's relevancy to the class variable. According to Duch *et al.* (2003), feature ranking is an independent evaluation process of the available features as per their importance to eliminate potentially irrelevant features. All filter-based feature selection methods use a "Ranker" to evaluate the features based on scores computed using statistics, information theory, or some functions of the classifier's output. IG, Gain Ratio (GR), Symmetrical Uncertainty (SU), CHI, IG and ReliefF methods are examples of filter methods that use Rankers in feature selection. IG ranks the features based on the amount of information relevant to the class variable, reflected by each candidate feature, whereas GR uses the prediction capabilities of each candidate feature to determine their individual rankings (Novakovic *et al.*, 2011).

Feature ranking is used by domain experts as a basic way of determining the best feature subsets; however, Ranker search methods do not provide the number of features to be selected, instead leaving the domain expert to decide. Most existing ranking search methods employ an elementary approach to display features along with their rank. More importantly, they leave the decision of which features to select up to the users' experience and knowledge, which subsequently requires time, care, and accuracy. Therefore, there is a need to develop a new intelligent Ranker search method that specifically recommends the features that should be chosen and the ones to ignore. The new Ranker should act as a recommendation to the feature selection process, be totally independent, and not filter-based-method specific. This will enable the Ranker to be embedded with any filter methods without dependency or data sensitivity and thus act as a generic search method.

A number of research studies have evaluated the performance of available feature ranking methods. Most concluded that there is no one Ranker method that is intelligent enough to distinguish influential features from redundant ones without domain expert involvement (Hu *et al.*, 2003; Duch *et al.*, 2004; Novakovic *et al.*, 2011; Cherrington et al., 2019). Further, none of the studies found an intelligent solution for ranking within filter methods, hence, more research and investigation is needed to develop more advanced Rankers that can be used effectively with any feature selection method.

Optimum threshold and domain expert involvement

Determining the optimal threshold between good and useless features is another vital issue related to feature selection. Most of the available filter methods do not distinguish the cut-off value which could help these methods provide a small subset of features rather than relying on the domain expert. Distinguishing between features is a difficult task because of the diverse nature of datasets, their characteristics, and filter methods' mathematical metrics used to calculate weights for each feature among others (Thabtah *et al.*, 2018). This difficult task relies on the knowledge of the domain expert, requiring additional time, care, and resources.

Let us assume that there is a dataset with over 1,000 features, and IG or CHI is used to determine the influential features. Both of these filter methods will return a feature set of 1,000 ranked on the assigned weights of the filter methods. Then, the user will have to choose possibly the top 5, top 10, top 30, top 100, etc. based on his/her requirements and experience. The process of selecting which features is lengthy and difficult with a high chance that the user may miss prominent features. Having an automated threshold embedded within the filter method to offer the domain expert a small subset of features would be advantageous. This threshold is important since it represents a boundary between features to be selected and features to be eliminated. Using irrelevant features and eliminating relevant features would negatively impact the performance of learning algorithms and possibly lead to confusing and false predictions.

More research and development are recommended to establish an automated feature selection technique that has an inbuilt metric to identify the optimal threshold between informative and uninformative features without having to rely on a domain expert, dataset characteristics, and mathematical equations as used in the filter method.

Feature-to-feature correlation

Most of the available feature selection-based filter methods do not consider feature-to-feature correlation when determining the optimal subsets during feature analysis. Valuing this is important as it helps to reduce the number of features and then offers a set that does not overlap in data instances and is different from each other yet correlated with the class. One of the successful methods that dealt with this issue was mRMR (Peng *et al.*, 2005) and its extensions. mRMR ranks each candidate feature based on its relevance to the class identifying the redundant features (those correlated with each other). According to Cai *et al.* (2012), mRMR defines relevant features as those with minimum redundancy with each other while maintaining the maximum relevance with the class label. Mutual information (MI) is the parameter used by mRMR to measure the mutual dependencies between features and class labels to identify the redundant and the relevant features. Fast-mRMR and mRMRe (De Jay *et al.*, 2013; Ramírez-Gallego *et al.*, 2016) are extensions of mRMR that were developed to overcome computational complexities of traditional mRMR and make it more efficient.

Limited research investigations have been conducted to highlight the importance of identifying feature-to-feature correlation to enhance the performance of the overall feature selection process. The study by Yu and Liu (2004a) is one such attempt that addressed the need to incorporate a redundant feature analysis process as relevancy is insufficient to determine the best subsets. The authors introduced a novel mechanism called fast correlation-based filter (FCBF). This involves first selecting relevant features and then identifying predominant features from the selected set to enhance the selection process through a relevance and redundancy analysis. Yu and Liu (2004b) also discussed the importance of identifying and eliminating redundant features in gene expression microarray data analysis to classify diseases or phenotypes accurately.

Various studies have used different mathematical metrics to identify the intercorrelation among the features to produce optimal feature subsets. Radovic *et al*. (2017) proposed the temporal mRMR (TmRMR), a filter approach which uses the value of F-statistics across different time steps as the parameter to compute the temporal information and relevancy among feature; this is by applying a dynamical time-warping approach to handle temporal gene expression data in an effective manner. F-statistics values determine redundant features by identifying features with small and large inter-class variances.

Another research by Gu *et al*. (2012) presented a novel approach called more relevance less redundancy (MRLR) that uses mathematical metrics such as information amount, conditional mutual information, and relevance degree to eliminate redundant features. Mutual information is one of the most common parameters used in identifying feature-to-feature correlation in most of the literature. Cai *et al*. (2012) also used the mutual information value to rank features and identify redundant features. In a former study by Yu and Liu (2004a, b), the linear correlation coefficient is suggested as a viable mathematical metric to determine the goodness of the features. The authors describe this as a successful method as it helps to identify the features with near zero correlation with the class and it helps to eliminate the redundant features through identifying those with high correlation to each other. **Table 2** shows mathematical metrics used to identify feature-to-feature relevancy.

Table 2: Mathematical metrics used in feature selection approaches to derive feature-to-feature correlation.

Literature	Filter Method	Mathematical Metrics	Equation
Radovic *et al.* (2017)	TmRMR	F-Statistics	$F(g_j, c) = \frac{1}{T}\sum_{t=1}^{t} \quad F(g_j^{(t)}, c)$
Gu *et al.* (2015)	MRLR	information amount, conditional mutual information, and relevance degree	$NMI(f_i;\ f_s) = \frac{MI()}{min\{H(f_i).H(f_s)\}}$
Cai *et al.* (2012)	mRMR	Mutual Information	$I(X,Y) = \int \quad \int \quad p(x,y)\ log\frac{p(x,y)}{p(x)\,p(y)}$
Yu & Liu, (2004a, b)	FCBF	Linear Correlation Coefficient	$r = \frac{\sum_i \quad (x_i - \underline{x_i})\,(y_i - \underline{y_i})}{\sqrt{\sum_i \quad (x_i - \underline{x_i})^2}\sqrt{\sum_i \quad (y_i - \underline{y_i})^2}}$

Data imbalance

The class imbalance is a critical challenge observed in datasets with extremely different class distributions, often encountered in the classification tasks, which may result in generating results that favour the dominant class in the dataset (the class label with higher frequency) (Japkowicz and Stephen, 2002). Data is said to be imbalanced when the majority of the classification instances belong to one class and only a few instances belong to a minority class, especially in medical applications (Thabtah *et al.*, 2019b). For instance, if we have data of 1,000 instances, where only 10 of them have been diagnosed with autism, if we consider “Autism” and “No Autism” as two class values, this dataset is highly imbalanced. It will be imperative to distinguish the features that are related to autism in this dataset, which is difficult as most instances belong to the “No Autism” class. Hence, scholars proposed a solution that is mainly data-driven to balance the data before feature selection and learning phases such as under-sampling and oversampling (Wasikowski and Chen, 2010; Yin *et al*., 2013).

Machine learning algorithms are sensitive to data with imbalanced class labels since they produce classifiers that are biased to the majority class and overlook the minority class label. This is because data instances fed into the learning algorithm tend to assume the unavailable points to make predictions by generalising the available points to the entire population. Because of that, the classifier would demonstrate a poor prediction accuracy on the minority class (Wasikowski and Chen, 2010).

A study by Wasikowski and Chen (2010) compared different schemes that include sampling and feature selection techniques to evaluate which technique performed better in dealing with imbalanced class data. The study revealed that feature selection with signal-to-noise correlation coefficient (S2N) (Gailey *et al.*, 1997) and feature assessment by sliding thresholds (FAST) (Chen and Wasikowski, 2008) techniques are highly effective on class imbalanced data. But feature selection methods used for balanced data may not perform as well on the imbalanced data, so the feature selection method should focus more on identifying features that help to predict the minority classes rather than the majority classes. A major issue that is encountered is locating a threshold to distinguish between relevant and irrelevant features. In feature selection, various ratios are used to rank the features based on their relevancy to the target class labels, but when most of the data belongs to one class, the results tend to be biased towards the features relevant to the majority class, ignoring those with more potential to predict the minority classes (Pant and Srivastava, 2015).

Many studies have been conducted on determining the most appropriate feature selection method to be used on class imbalanced data to yield a better classifier performance (Japkowicz and Stephen, 2002; Wasikowski and Chen, 2010; Yin et al., 2013; Maldonado *et al.*, 2014; Thabtah *et al.*, 2019b). Most of them investigated the impact of class imbalance data on classifier performance, but little research addresses the impact on the feature selection process of imbalanced classes. Yin *et al.* (2013) addressed this problem and presented two feature selection approaches to overcome the issue. One approach is based on class decomposition (Maimon and Rokach, 2002), which involves the partition of majority classes into small class subsets before feature selection, and the other is based on Hellinger distance

(Beran, 1997); this measures the distribution divergence of each class to evaluate its goodness for feature selection. The results showed that the proposed two approaches outperformed most of the available conventional feature selection methods. In an experiment carried out on protein function data, Al-Shahib *et al.* (2005) showed that under-sampling the majority class prior to feature selection significantly increases the classifier performance on imbalanced data.

Recommendations and Conclusions

A high level of noise is a major problem that makes managing data difficult, and most often this noise is generated from the technology used in collecting data or the source of data itself. Dimensionality reduction through filter-based feature selection is a commonly used solution to eliminate this problem. However, in the era of big data in which we have different feature types, sparse data, and unstructured data, among others, filter methods face practical challenges that have been rarely addressed in recent research. This paper critically analysed challenges of filter-based methods associated with results quality and performance including results discrepancies, ranking of features in the results set, absence of clear threshold between good and bad features, handling imbalanced data, and feature-to-feature correlation.

Different feature selection methods deliver different selection outcomes as a result of the mathematical models used to compute the feature scores based on feature-to-feature frequencies, feature-to-class frequencies, and expected and observed frequencies of the features. Therefore, if two different feature selection methods are employed on the same dataset, the end user can get two different outcomes for the most relevant feature subsets. The paper highlights the importance of addressing this challenge as the credibility and reliability of the final learning algorithm depend enormously on the feature subsets selected through the employed filter method. Use of normalised feature scores is recommended to yield more static, reliable, feature selection outcomes. Further research to develop more normalised advanced feature scoring mechanisms is vital.

All the filter methods use simple rankers to weigh the features based on their importance or the relevancy to the class labels. These rankers are very primitive and do not provide information on how many features are to be selected or eliminated. Therefore, the number highly depends on the end-user's knowledge and level of expertise, requiring an excessive amount of time, effort, and care. Hence, there is a need for an advanced Ranker that intelligently offers the subset of features by creating a fine line to differentiate good features from useless ones. Hence, the end user will not have to scan the entire features within the results set, rather just take that ordered by the Ranker.

Absence of a clear threshold between good and bad features is also another challenge pinpointed in the paper that makes conventional filter-based feature selection over-dependent on the end-user/domain experts' involvement. Determining the cut-off between relevant and irrelevant features is essential as using irrelevant features in induction models can hinder the learning process significantly. Hence, the importance of developing an automated threshold embedded into traditional filter methods is emphasised.

Disregarding the feature redundancies is one of the main drawbacks of filter-based feature selection. Identifying the feature-to-feature correlation is of utmost importance as it helps to eliminate features that overlap. Therefore, to overcome this challenge, a viable approach that determines the feature-to-feature correlation and automatically eliminates the redundant features should be embedded into existing filter methods.

Some data characteristics such as uneven distribution can also make the feature selection process biased and inaccurate. Feature selection requires data that is perfectly balanced to generate unbiased accurate results. But it is not always practical to have perfectly balanced data, therefore, the paper highlighted the need for a valid mechanism to balance imbalanced data prior to the feature selection process to yield better results. Smart automated sampling techniques are recommended to be integrated into filter methods to identify class imbalanced data and to balance this without changing the original data.

Further research and investigation are advised to produce more intelligent automated feature selection techniques that mitigate the identified challenges and make the feature selection process more effective and efficient. In the near future, we are going to examine a number of filter methods on pathological datasets related to dementia in order to determine high effective attributes that may have correlations with dementia at different levels. Feature selection can provide a bottom-up approach of exploring datasets to reveal hidden useful patterns; in the case of diagnosing dementia, features that are hidden from the eyes of a pathologist but have clear impact on detecting dementia can be identified. This bottom-up approach of recommending features to domain-experts, such as pathologists, must also demonstrate that the features are interpretable to clinicians and can reduce observer bias. Features that achieve this are much more likely to be adopted by the clinical community and used as valuable biomarkers for diagnosing and stratifying patients into subgroups. Further work is needed to investigate the determinants of influential features, especially within application domains to pinpoint factors that influence feature interpretability and bias. While we highlight general best practices for feature filtering, understanding their impact in different research domains will be critical for these to have true value.

References (Manuscript 1)

Al-Shahib, A, R Breitling and D Gilbert (2005). Feature selection and the class imbalance problem in predicting protein function from sequence. Applied Bioinformatics, 4, 195–203.

Battiti, R (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4), 537–550.

Beran, R (1997). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3), 445–463.

Bishop, C (1995). *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

Bohanec, M, V Rajkovic and B Zupan (1997). Applications of qualitative multi-attribute decision models in health care. *International Journal of Medical Informatics*, 58–59, 191–205.

Cai, Y, T Huang, L Hu, X Shi, L Xie and Y Li (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids*, 42, 1387–1395.

Chen, X-W and M Wasikowski (2008). FAST: A ROC-based feature selection metric for small samples and imbalanced data classification problems. *In Proceeding KDD '08 Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 124–132. New York: ACM.

Cherrington, M, F Thabtah, J Lu and Q Xu (2019). Feature selection: Filter methods performance challenges. *In International Conference on Computer and Information Sciences (ICCIS)*. doi:10.1109/ICCISci.2019.8716478.

De Jay, N, S Papillon-Cavanagh, C Olsen, N El-Hachem, G Bontempi and B Haibe-Kains (2013). mRMRe: *An R package for parallelized mRMR ensemble feature selection. Bioinformatics*, 29(18), 2365–2368.

Duch, W, T Wieczorek, J Biesiad and M Blachni (2004). Comparison of feature ranking methods based on information entropy. *In IEEE International Joint Conference on Neural Networks*. doi:10.1109/IJCNN.2004.1380157.

Duch, W, T Winiarski, J Biesiada and A Kachel (2003). Feature ranking, selection and discretization. *In IEEE International Joint Conference on Neural Networks (IJCNN)*. doi:10.1109/IJCNN.2004.1380157.

Estevez, P, M Tesmer, C Perez and J Zurada (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2), 189–201.

Gailey, P, A Neiman, J Collins and F Moss (1997). Stochastic resonance in ensembles of nondynamical elements: The role of internal noise. *Physical Review Letters*, 79(23), 4701–4704.

Gao, W, L Hu, P Zhang and F Wang (2018). Feature selection by integrating two groups of feature evaluation criteria. *Expert Systems with Applications*, 110, 11–19.

Gu, Q, Z Li and J Han (2012). Generalized Fisher score for feature selection. *Machine Learning*, 256–269.

Hall, M (1999). *Correlation-Based Feature Selection for Machine Learning*, Hamilton, New Zealand: Waikato University.

Hall, M (2000). *Correlation-Based Feature Selection for Discrete and Numeric Class*, Hamilton, New Zealand: Waikato University.

Hall, M, G Holmes and E Frank (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.

Hancer, E, B Xue and M Zhang (2017). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, 140, 1–17.

Hu, K, Y Lu and C Shi (2003). Feature ranking in rough sets. *AI Communications*, 16(1), 41–50.

Japkowicz, N and S Stephen (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449.

Kamalov, F and F Thabtah (2017). A feature selection method based on ranked vector scores of features for classification. *Annals of Data Science*, 4(4), 483–502.

Labani, M, P Moradi, F Ahmadizar and M Jalili (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 70, 25–37.

Liu, H and R Setiono (1995). Chi2: Feature selection and discretization of numeric attribute. *In Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence*, pp. 388–391. New York: IEEE.

Maimon, O and L Rokach (2002). Improving supervised learning by feature decomposition. *Foundations of Information and Knowledge Systems*, 2284, 178–196.

Maldonado, S, R Weber and F Famili (2014). Feature selection for high-dimensional classimbalanced data. *Information Sciences*, 286, 228–246.

Marinaki, M and Y Marinakis (2013). An island memetic differential evolution algorithm for the feature selection problem. *Nature Inspired Cooperative Strategies for Optimization (NICSO)*, pp. 29–42. Berlin: Springer.

Novakovic, J, P Strbac and D Bulatovic (2011). Towards optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 21(1), 119–135.

Pant, H and R Srivastava (2015). A survey on feature selection methods for imbalanced datasets. *International Journal of Computer Engineering and Applications*, 9(2), 197–204.

Peng, H, F Long and C Ding (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226–1238.

Quinlan, J (1986). Induction of decision trees. *Machine Learning*, 1, 81–106.

Radovic, M, M Ghalwash, N Filipovic and Z Obradovic (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, doi:10.1186/s12859-016-1423-9.

Rahmaninia, M and P Moradi (2017). OSFSMI: Online stream feature selection method based on mutual information. *Applied Soft Computing*, 68, 1568–4946.

Rajab, K (2017). New hybrid features selection method: A case study on websites phishing. *Security and Communication Networks*. doi:10.1155/2017/9838169.

Ramírez-Gallego, S, L Lastra, D Martíne, V Bolón-Canedo, J Benítez, F Herrera and A Alonso-Betanzos (2016). Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data. *International Journal of Intelligent System*s, 32(2), 134–152.

Rehman, A, K Javed, H Babri and M Saeed (2015). Relative discrimination criterion — A novel feature ranking method for text data. *Expert Systems with Applications*, 42(7), 3670– 3681.

Robnik-Šikonja, M and I Kononenko (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53(1–2), 23–69.

Saeys, Y, I Inza and P Larranaga (2008). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.

Tang, J, S Alelyani and H Liu (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37. doi:10.1201/b17320-3.

Thabtah, F, F Kamalov, S Hammoud and S Shahamiri (2019a). A new feature selection method based on simplified observed and expected likelihoods distance. Available at https://github.com/suhelhammoud/L2.

Thabtah, F, N Abdelhamid and D Peebles (2019b). A machine learning autism classification based on logistic regression analysis. *Health Information Science and Systems*, 7(1), 12.

Thabtah, F and D Peebles (2019). A new machine learning model based on induction of rules for autism detection. *Health Informatics Journal*. doi:10.1177/1460458218824711.

Thabtah, F, F Kamalov and K Rajab (2018). A new computational intelligence approach to detect autistic features. *International Journal of Medical Informatics*, 117, 1386–5056.

Thabtah, F, W Hadi, N Abdelhamid and A Issa (2011). Prediction phase in associative classification. *Journal of Knowledge Engineering and Software Engineering*, 21(6), 855– 876.

Town, P and F Thabtah (2019). Data analytics tools: A user perspective. *Journal of Information and Knowledge Management*, 18(1), 1950002.

Wasikowski, M and X-W Chen (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1388–1400.

Yin, L, Y Ge, K Xiao and X Wang (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing*, 105, 3–11.

Yu, L and H Liu (2004a). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205–1224.

Yu, L and H Liu (2004b). Redundancy based feature selection for microarray data. *In KDD '04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 737–742. New York: ACM.

Zhang, X, Y Hu, K Xie, S Wang, E Ngai and M Liu (2014). A causal feature selection algorithm for stock prediction modelling. *Neurocomputing*, 142, 48–59.

Zhao, F, J Zhao, X Niu, S Luo and Y Xin (2018). A filter feature selection algorithm based on mutual information for intrusion detection. *Journal of Applied Science*, 8(9), 1–20.

Chapter 3 - Assessment of Alzheimer-related Pathologies of Dementia Using Machine Learning Feature Selection

3.1. Background

This chapter employs feature selection methods to evaluate Alzheimer-related pathologies to a population cohort of ageing individuals and investigate the relationship between neuropathological features and dementia status. The study employed feature selection filter methods, referred to as ‘feature ranking methods’ in this article, and classification algorithms to analyze the data and identify the optimal set of features for accurately diagnosing AD. This chapter serves as a preliminary investigation before creating the data process described in chapter 4.A step-wise ML approach was employed to rank and select Alzheimer-related pathologies and assess the ability of different measures, such as those related to Aβ-related assessments and tau, to inform about dementia condition status. The chapter also identifies clusters of highly correlated measures in the dataset and tested several classification algorithms using various subsets of ranked features to examine the impact of ranking. Additionally, the chapter suggests that more specific neuropathology features for specific brain regions should be utilized to identify the pathophysiological processes associated with dementia in individual patients.

The chapter will try to reveal misclassification cases if any, indicating discordance between neuropathology and dementia, where some demented individuals had no known pathology, and some non-demented individuals had pathology. The expected results will provide valuable insights into the potential of using ML techniques for assessing Alzheimer-related pathologies of dementia. The findings will highlight there is a need for further research to enhance the performance of dementia classification and develop better diagnostic systems and treatment targets for dementia patients.

3.2. Contribution

The following version of the accepted manuscript was published in the *Journal of Alzheimer's Research & Therapy, 15, 47 (2023).* This publication involves my contribution as the first author in the process of data analysis, manuscript drafting, and editing. Specifically, I was the main contributor in producing the initial draft of the manuscript and supplementary with the assistance of editing from Emmanuel Jammeh, Teruka Taketa, and Dennis Wang. Moreover, I was responsible for generating most of the code, figures, and tables used in the study, except for the running feature signatures, which were completed by Emmanuel Jammeh to demonstrate the association of non-standard pathologies and demographic features with clusters. The contributions of other co-authors are duly recognized in the "Contributions" section of the paper.

3.3. Manuscript 2

Assessment of Alzheimer-related Pathologies of Dementia Using Machine Learning Feature Selection

Mohammed D Rajab[1,2], Emmanuel Jammeh[1], Teruka Taketa[1], Carol Brayne[3], Fiona E Matthews[4], Li Su[1], Paul G Ince[1], Stephen B Wharton[1], Dennis Wang[1,2,5,6] and on behalf of the Cognitive Function and Ageing Neuropathology Study Group

1. Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield S10 2HQ, UK.
2. Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK
3. Cambridge Public Health, Cambridge CB2 1PZ, UK

4. Population Health Sciences Institute; Newcastle University, Newcastle upon Tyne NE4 5PL, UK
5. Singapore Institute Clinical Sciences, A*STAR, Singapore, 117609, Singapore
6. National Heart and Lung Institute, Imperial College London, London, SW3 6LY, UK

Correspondence: dennis.wang@imperial.ac.uk

Abstract

Although a variety of brain lesions may contribute to the pathological assessment of dementia, the relationship of these lesions to dementia, how they interact and how to quantify them remains uncertain. Systematically assessing neuropathological measures by their degree of association with dementia may lead to better diagnostic systems and treatment targets. This study aims to apply machine learning approaches to feature selection in order to identify critical features of Alzheimer-related pathologies associated with dementia. We applied machine learning techniques for feature ranking and classification to objectively compare neuropathological features and their relationship to dementia status during life using a cohort (n=186) from the Cognitive Function and Ageing Study (CFAS). We first tested Alzheimer's Disease and tau markers, and then other neuropathologies associated with dementia. Seven feature ranking methods using different information criteria consistently ranked 22 out of the 34 neuropathology features for importance to dementia classification. Although highly correlated, Braak neurofibrillary tangle stage, Beta-amyloid and cerebral amyloid angiopathy features were ranked the highest. The best-performing dementia classifier using the top eight neuropathological features achieved 79% sensitivity, 69% specificity, and 75% precision. However, when assessing all seven classifiers and the 22 ranked features, a substantial proportion (40.4%) of dementia cases was consistently misclassified. These results highlight the benefits of using machine learning to identify critical indices of plaque, tangle and cerebral amyloid angiopathy burdens that may be useful for classifying dementia.

Keywords— Dementia, Alzheimer's, Feature Selection, Machine Learning, Neuropathology, Beta-amyloid

Introduction

Dementia is a significant healthcare concern among the elderly, and the number of people with dementia will reach 131.5 million worldwide by 2050 [1]. There is no cure for this syndrome, but an accurate and timely diagnosis of dementia may create opportunities for patients to access symptomatic and potentially disease-modifying therapies. As defined in the Diagnostic and Statistical Manual of Mental Disorders 5th edition, cognitive and daily activity decline defines the syndrome, often measured using cognitive and functional tests along with medical history reported by the patient or caregiver [2]. In clinical settings, further investigations are performed primarily on younger onset dementias focused on anatomical and, sometimes, functional changes measured by magnetic resonance imaging (MRI) and positron emission tomography (PET) scans, and increasingly cerebrospinal fluid (CSF) samples taken from a lumbar puncture are considered to be dementia subtype biomarkers. However, dementia, as it most often manifests in older people, is associated with multiple brain pathologies [3,4]. Research remains challenging when assessing the interactions among multiple brain factors related to the syndrome as it manifests during life.

The Cognitive Function and Ageing Studies (MRC CFAS, CFAS I, CFAS II) were longitudinal population-based ageing studies focusing on cognition. This analysis focused on brains donated from the original MRC CFAS. More than 550 participants from CFAS voluntarily donated their brains to the study after their death in order to undergo a comprehensive pathological assessment [5,6]. Neuropathological investigations have explored the relationship of pathological features in the brain to dementia phenotypes, including various measures related to tau and beta-amyloid (Aβ) pathologies [7]. These studies showed considerable overlap in the burden of lesions between participants dying with and without dementia [3,4]. Attributable risk showed the importance of many other pathologies in the brain [8,9].

Machine learning (ML) classification algorithms and feature selection techniques have enabled automated ways of classifying heart and skin diseases, and identified the most informative combination of predictors of those diseases [10,11]. Studies investigating dementia involving brain imaging utilized three supervised ML algorithms (neural network, support vector machine, and adaptive neuro-fuzzy inference system) for the diagnosis of Alzheimer's disease (AD) and vascular dementia (VD) [12]. These algorithms used ranked MRI features based on their performance in identifying dementia cases within the dataset. Their results showed that categorizing AD and VD profiles using ML had high discriminant power with a classification accuracy of more than 84% in some cases. ML feature selection approaches were applied to enable the identification of neuropsychological measures and MRI features for the classification of AD [13]. ML using demographic and clinical features as predictors had also been used to predict dementia and neuropathology [14], but this assumes the predictors were stable over time. Alternatively, ML techniques could assess the relationship between dementia status and the neuropathological features of post-mortem brains, and identify cases where they disagreed. Feature selection could also find which features are most informative of dementia. Where features are not informative, it could be interesting to reveal cases of dementia with insufficient pathology. Identifying informative features could help reduce resources, such as time, cost, and effort utilized during pathological assessment and highlight a need for more profound clinical assessments.

In order to distinguish related indices such as plaque, tangle and CAA burdens, we needed an objective approach to rank these pathologies and identify a combination of features useful for classifying dementia. In this chapter, we evaluate whether ML feature ranking can identify a subset of neuropathological features ordered by their relative contribution to dementia.

The evaluation is performed using Alzheimer-related and other dementia-related pathologies measured in a population-representative sub–cohort of CFAS [6,15–18]. There were 34 features determined by pathologists, including Aβ features, cerebral amyloid angiopathy (CAA) features and plaque scores. These features were automatically ranked, filtered and included in ML classifiers of dementia. We also reported the limits of ML classification of dementia using neuropathology factors and discussed possible reasons for these limitations.

Material and methods

Overview of the feature selection approach

The selection of neuropathology features that were informative of dementia involved several steps (**Figure 1**). We first obtained access to and downloaded the CFAS dataset following review and ethics approval by the CFAS management committee. Since the dataset contains general features, a re-labelling of the available neuropathological and other types of features was performed to assist user understanding, e.g. sample of the used features types: tau, Aβ, demographics, etc.

We then applied supervised learning and feature selection techniques based on multiple filter-based methods. Features were ranked based on their importance and the most informative features were determined. The smallest subset of features that can classify dementia most accurately were identified using several ML classifiers. Finally, we examined misclassified cases in relation to the neuropathology features and linked the associations with other non-standard pathologies.

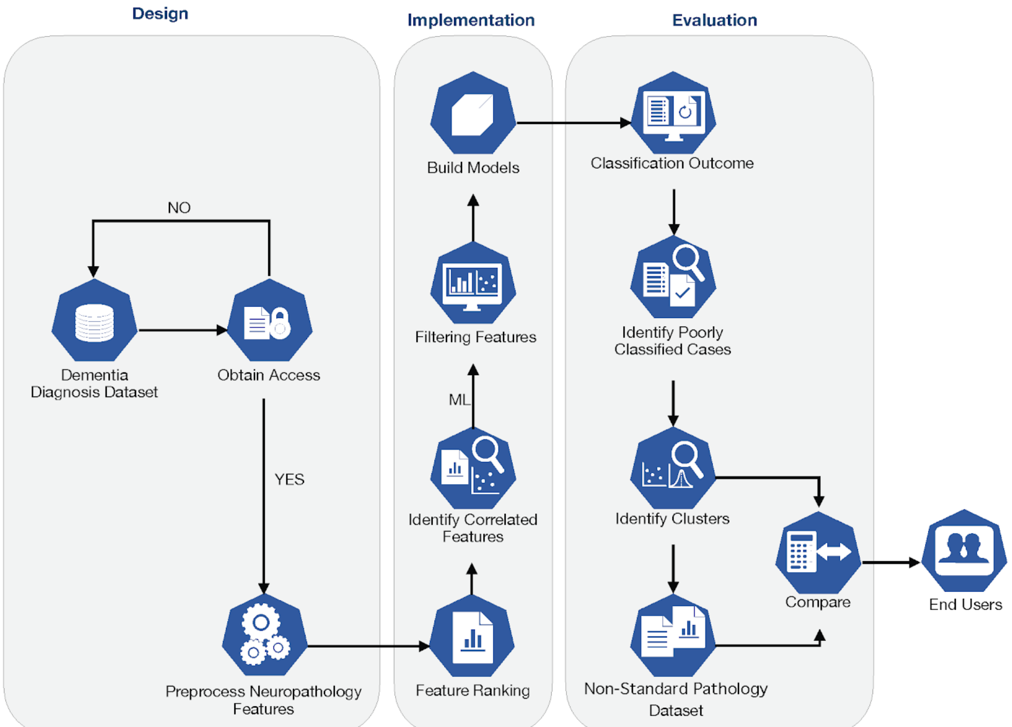

Design
Implementation
Evaluation
NO
Dementia Diagnosis Dataset
Obtain Access
YES
Preprocess Neuropathology Features
Build Models
Filtering Features
ML
Identify Correlated Features
Feature Ranking
Classification Outcome
Identify Poorly Classified Cases
Identify Clusters
Compare
End Users
Non-Standard Pathology Dataset

Figure 1: Methodology for classification of dementia. The methodology for the classification of dementia followed three stages: design, implementation, and evaluation. First, we pre-processed and assessed feature-feature correlation after acquiring access to neuropathology and clinical data from CFAS. We then applied feature ranking methods to rank and filter all neuropathology features. Next, classifiers benchmarked with different subsets of features were selected according to their rankings. Finally, we compared cases that were consistently misclassified and evaluated brain attributes associated with these cases in order to improve machine learning.

Neuropathology features in the CFAS cohort

The CFAS cohort used for this study included data from two centres (Cambridge and Newcastle), totalling 186 subjects with 34 neuropathology features, plus age and brain weight, as shown in **Table 1**. Immunohistochemical detection of Aβ in formalin-fixed, paraffin-embedded sections (5 μm) as previously described [19]. Assessment of Aβ phase was performed according to the Thal scheme, and BrainNet Europe approach [20,21]. Neurofibrillary tangles were assessed by the Braak stage [22] and plaques were assessed using the CERAD method [23]. The features included basic neuropathological measures for each subject, including Braak neurofibrillary tangle (NFT) stage, Brain-Net Europe protocol for tau pathology, hippocampal tau NFT stage [24], Thal phase, primary age-related tauopathy (PART), cerebral amyloid angiopathy (CAA), thorn-shaped astrocytes (TSA) [17] and microinfarct stage [25] (**Table 1**).

Table 1: Description of the neuropathology features in addition to the age and brain weight features

No	Feature	Feature Description	Type	Control		
				Dementia (n=107)	**No Dementia** (n=70)	**Missing** (n=9)
1	Braak NFT stage	Braak Stage refers to the Braak neurofibrillary tangle (NFT) stage (0-VI) [22,26].	Nominal	107	70	0
2	Thal phase	Thal phase refers to the Thal Aβ phase, which is the new BrainNet stage for Aβ to detect immunopositive amyloid in cortical and subcortical areas and differentiate five phases [20,21].	Nominal	107	70	0
3	Aβ stage typical	Aβ stage typical indicates the Aβ stage typical and atypical [18].	Nominal	107	70	0
4	PART-definite	PART relates to the new primary age-related-tauopathy concept. PARTdefinite as cases having no Aβ pathology (Thal 0) and with Braak NFT stage I-IV [27].	Nominal	50	47	80 (45.2%)
5	PART-all	Those cases with mild Aβ pathology (Thal I-II) and with Braak NFT stage I-IV [27].	Nominal	71	63	43 (24.3%)
6	CAA areas	The number of brain areas examined that have CAA (number of areas out of 9 maximum) [19].	Numeric	107	70	0
7	CAA type	As defined by Thal where CAA type 1 are cases with capillary amyloid and 2 only in larger vessels and type 0 no CAA [15,19].	Nominal	107	70	0
8	CAA parenchymal	CAA severity score according to Love et al [28] leptomeningeal and parenchymal vascular amyloid in four neocortical areas- So in any area CAA can be 1, 2, or 3 and the score ranges from 0 to 12 [18].	Nominal	107	70	0
9	CAA meningeal	CAA severity meningeal has the same scoring system as CAA parenchymal with the score ranging from 0 to 12 [18].	Nominal	107	70	0
10	CAA total severity	The scores for parenchymal and leptomeningeal amyloid were summed in four areas, scores range from 0 (minimum) to 24 (maximum) for severity in cortical areas [19].	Numeric	107	70	0
11	CAA frontal	CAA in frontal cortex (Present or Absent) [24].	Nominal	107	70	0
12	CAA temporal	CAA in temporal cortex (Present or Absent) [24].	Nominal	107	70	0
13	CAA parietal	CAA in the parietal cortex (Present or Absent) [24].	Nominal	107	70	0
14	CAA occipital	CAA in occipital cortex (Present or Absent) [24].	Nominal	107	70	0
15	CAA hippocampus	CAA in hippocampus and occipitotemporal gyrus (Present or Absent) [24].	Nominal	107	70	0
16	CAA cerebellum	CAA in cerebellum (Present or Absent) [24].	Nominal	106	69	2 (1.13%)
17	BrainNet tau stage	BrainNet tau stage, refers to BrainNet Europe Protocol for tau pathology, a six-stage scheme that uses neuropil threads and is proposed by the BrainNet Europe Consortium [21].	Nominal	107	69	1 (0.6%)
18	Hippocampal tau NFT stage	Hippocampal tau neurofibrillary tangles (NFT) stage	Nominal	56	35	86 (48%)
19	subpial TSA in expanded cortex	The subpial thorn-shaped astrocytes (TSA) in the expanded cortex.	Nominal	107	69	1 (0.6%)
20	subpial TSA in mesial temporal lobe	The subpial thorn-shaped astrocytes (TSA) in the mesial temporal lobe.	Nominal	107	69	1 (0.6%)
21	subpial TSA in brainstem	The subpial thorn-shaped astrocytes (TSA) in the brainstem.	Nominal	107	67	3 (1.7%)
22	TSA-any	Thorn-shaped astrocytes (TSA) in any brain area. (Present or Absent)	Nominal	107	69	1 (0.6%)
23	TSA-total	The number of areas in the brain with thorn-shaped astrocytes (TSA) [29–32].	Numeric	107	69	1 (0.6%)
24	Tufted astrocytes	The tufted parenchymal astrocytes in any brain area.	Nominal	107	69	1 (0.6%)

25	Subpial mesial temporal	The subpial tau neurites in the mesial temporal lobe.	Nominal	107	69	1 (0.6%)
26	Subpial brainstem	The subpial tau neurites in the brainstem/subcortical region.	Nominal	107	67	3 (1.7%)
27	Argyrophilic grains	The argyrophilic grains disease.	Nominal	107	69	1 (0.6%)
28	Cortical stage	The cortical microinfarcts stage which distinguishes the number of cortical areas that have microinfarcts.	Numeric	106	70	1 (0.6%)
29	Subcortical stage	Subcortical lacune stage which distinguishes the number of subcortical areas that have microinfarcts.	Numeric	106	70	1 (0.6%)
30	Microinfarct stage	The total microinfarct stage which differentiate the number of total areas that have microinfarcts.	Numeric	106	70	1 (0.6%)
31	Frontal microinfarct	Frontal microinfarct [25].	Nominal	106	70	1 (0.6%)
32	Temporal microinfarct	Temporal Microinfarct [25].	Nominal	106	70	1 (0.6%)
33	Parietal microinfarct	Parietal microinfarct [25].	Nominal	106	70	1 (0.6%)
34	Occipital microinfarct	Occipital Microinfarct [25].	Nominal	106	70	1 (0.6%)
35	Age	Patient's age at death.	Numeric	107	70	0
36	Brain weight	Patient's brain weight.	Numeric	91	59	27 (15%)
37	Gender	Sex	Nominal	107	70	0
38	Virchow-Robin space expansion	Virchow-Robin spaces (VRS) are cavities filled with cerebrospinal fluid surrounding small penetrating cerebral arterioles with extensions of the subarachnoid space.	Nominal	106	70	1 (0.6%)
39	Lewy bodies in substantia nigra	The Lewy body is a distinguishing neuronal inclusion. This is always found in the substantia nigra and brain regions in Parkinson's disease, which occurs wherever there is excessive loss of neurons.	Nominal	105	68	4 (2.3%)
40	Neuronal loss in hippocampus	Neuronal loss in hippocampus	Nominal	106	70	1 (0.6%)
41	Neuronal loss in substantia nigra	Neuronal loss in substantia nigra	Nominal	105	68	4 (2.3%)
42	Tangles in temporal lobe	Tangles in temporal lobe	Nominal	106	70	1 (0.6%)
43	Parenchymal CAA in frontal lobe	Parenchymal CAA in frontal lobe	Nominal	106	70	1 (0.6%)
44	Gliosis in hippocampus	Gliosis in hippocampus	Nominal	106	70	1 (0.6%)
45	Dementia Status	Class Label (Dementia or No dementia) Status of a patient	Binary	107	70	0

Dementia status

Dementia status at death for each respondent was determined based on interviews/assessments during the last years of the respondent's life. This included using the full Geriatric Mental State-Automated Geriatric Examination for Computer Assisted Taxonomy diagnostic algorithm, the Diagnostic and Statistical Manual of Mental Disorders (third edition - revised), interviews with the informants after the respondent's death and the cause of death. Respondents were assessed as having no dementia at death if they had not been identified with dementia at their last interview less than six months before death or if they did not have dementia identified at the last interview and the retrospective interview showed no dementia at death. Bayesian analysis was used to estimate the probability of dementia when last interviews were more than six months before death, and no record of having dementia at the interview and no retrospective informant interview (RINI) [5,33]. A total of 107 of the 186 subjects had a diagnosis of dementia, which represented approximately 58% of the cohort. Of these 107 cases, 72 were women and 35 were men; their median ages were 89 and 88 respectively. There was a balanced gender ratio (37 females and 33 males) for participants dying without dementia (median age 85, 79 respectively). The Consortium to Establish a Registry for Alzheimer's disease (CERAD) criterion determined that in 64 out of the 107 cases (60.0%), Alzheimer’s disease was the definite, probable or possible cause of the observed symptoms.

Ranking neuropathology features

We used several filter-based feature selection methods to determine the relevance of each feature to dementia in order to gain preliminary insight. These included Chi-square (CHI) [34], gain ratio [35], information gain (IG) [36], reliefF [37,38], symmetrical uncertainty [39], least loss [40] and variable analysis [41,42]. Generally, filter-based methods use different mathematical models to compute feature relevance. These methods are efficient feature selection tools that employ mathematical models to derive scores for each feature based on correlations between the features and class labels in the input dataset. There can be discrepancies in the ranking of features based on such scores due to the different mathematical models used [42,43]. The CFAS cohort consisting of 186 post-mortem and 34 neuropathology features was used for feature ranking. In addition to the 34 neuropathology features, age and brain weight were included. Using SciPy.stats v1.5.4 in Python3, we used z-score to adjust brain weight based on sex.

CHI utilizes the difference between observed and expected frequencies of the instances, as shown in Equation (1).

$$X^2 = \frac{(O-E)^2}{E} \tag{1}$$

O and *E* are the Observed and Expected frequencies for a specific feature, respectively. IG employs Shannon entropy to measure the correlation between a feature and dementia status (Equations 2 & 3).

$$IG\ (S, A)\ =\ Entropy\ (S)\ - \sum \quad ((\mid S_v \mid\ \div \mid S \mid)\ \times\ Entropy\ (S_v)) \qquad (2)$$

where Entropy (*T*) $= - \sum \quad P_c P_c$ (3)

P is the probability that *S* belongs to class label *c*. *S*$_v$ is the subset of *S* for which *a* feature has value *v*. |*S*$_v$| is the number of data instances in *S*$_v$, and |*S*| is the size of *S*.

A gain ratio is a normalized form of IG, which is estimated by dividing the IG by the Entropy of the feature with respect to the class (Equations 4 and 5).

Gain Ratio = $\frac{IG}{ENT(S,F)}$ (4)

$$ENT(S,F) = - \sum \quad \frac{S_i}{S} log_2 \frac{S_i}{S} \qquad (5)$$

where IG denotes the information gain, and *ENT* is the Entropy of feature F over a set of examples S.

Symmetrical uncertainty deals with the bias of IG that occurs due to a large number of distinct values for the feature and presents a normalized score (Equation 6).

$$SU(A, B)\ =\ \frac{2 \times IG(A|B)}{E(A) + E(B)} \qquad (6)$$

where $IG(A|B)$ denotes the information gained by A after knowing the class. E(A) and E(B) are the Entropy values of A and B, respectively.

ReliefF calculates the scores of each available feature with the class using the differences between the neighboring data instances and the target instances (Equation 7).

W[A] = W[A] - $\frac{\left(diff\frac{A,R_i,H}{m}\right)}{\left(diff\frac{A,R_i,M}{m}\right)}$ (7)

where W[A] is the feature weights, A is the number of features, and m is the number of random training data instances out of the ‘n’ number of training data instances used to amend W.

R_i = A randomly chosen test instance, and H/M is the nearest hit and nearest miss

Least loss is computed per feature based on the simplified expected and observed frequencies of the features (Equation 8), and Variable Analysis employs a vector of scores of both CHI and IG results, normalizes the scores, and then computes the vector magnitude (V_score) (See Equations 9 & 10).

$$L^2(Y,X) = \sum_{i,j} \quad [P\big(Y_{i,}X_j\big) - P(Y_i)P(X_j)]^2 \tag{8}$$

where X is the independent feature class, Y is the class label, $P(Y_i)$ is the theoretical marginal distribution of Y, and $P\big(X_j\big)$ is the theoretical marginal distribution of X, $P\big(Y_{i,}X_j\big)$ is the theoretical joint probability distribution of X and Y.

$$V_a = \left(\frac{IG_x}{CST_x}\right) \tag{9}$$

$$|V_a| = \sqrt{(IG)^2 + (TST)^2} \tag{10}$$

where V_a is the square root of the sum of the square of its CHI and IG results of a feature.

The V_score and the Correlation Feature Set results [44] are then integrated to represent a new measure of goodness to select relevant features.

$$IG\ (S,A)\ =\ Entropy\ (S)\ -\sum \qquad ((\ |\ S_v\ |\ \div\ |\ S\ |)\ \times\ Entropy\ (S_v)) \tag{2}$$

The number of samples used in the feature selection process was 177 out of 186 after removing the nine missing values in the diagnostic class and 36 features (34 neuropathology features plus brain weight and age features). All filter-based feature selection was conducted using Waikato Environment for Knowledge Analysis (WEKA version 3.9.1) [45]. The percentage contribution of each feature was calculated by averaging the total weights assigned by all filter methods to each feature after normalizing weights scores.

Dementia conditions classification

We attempted the classification of dementia status in 146 samples after removing missing values from the 177 that were used in the feature selection process. The 146 samples had a slight class imbalance, with 89 demented versus 57 non-demented patients. Before training our models, we randomly selected 57 patients from the demented group using the sample() function from the random module in Python3. Then, the rows were shuffled using sklearn.utils version 0.22.2.post1. As a result, 114 samples were utilized after balancing the class label. The 32 samples were held-out for final assessment. The hippocampal tau stage feature, which had 50% missing values, was dropped during the training process. Age and brain weight were removed before training the models, ending up with 22 features and 114 samples for classification. The dataset was split into a training set of 70% (80 samples) and a testing set of 30% (34 samples).

Seven classification algorithms were trained to classify individuals' dementia status from the 22 top-ranked features. Scikit-learn version 0.22.2.post1 was used to implement and train the ML classifiers, and then measure their classification performance. Logistic regression was implemented using the *sklearn.linear_model* package where penalty was set to 12, the regularization parameter C was set to 1, the maximum number of iterations taken for the solvers to converge was set to 2000, and other parameters were set to default values. A decision tree classifier was implemented using the *sklearn.tree* package. K-nearest neighbors classifier was implemented using the *sklearn.neighbors* with the number of neighbors set to 5, the function "uniform weights" used for prediction, the "Minkowski" distance metric utilized for the tree, and with other parameters were set to default values. The linear discriminant analysis classifier was implemented using the *sklearn.discriminant_analysis* package with singular value decomposition for solver hyperparameter and other parameters were set to default values. The Gaussian naïve Bayes classifier was implemented using *sklearn.naive_bayes.* The support vector machine with a radial basis function kernel (SVM-RBF) was implemented using *sklearn.svm* with the regularization parameter C set to 1, the kernel coefficient gamma= "scale", and other parameters were set to default values. The support vector machine with a linear kernel (SVM-LINEAR) was implemented using the *sklearn.svm* package with regularization parameter C set to 1, with a "linear" kernel, gamma coefficient "scale", and other parameters were set to default. The *sklearn.metrics* package was used to report classification performance. Training and performance evaluation were performed 500 times, from which the average performance measure was calculated as overall performance. Accuracy, balanced accuracy, F1-score, precision, sensitivity, and specificity utilizing regression plots, were measures used for performance. ML models and feature selection libraries were built using Python 3.7.3.

Classification with multiple feature sets

We created subsets of neuropathological features from the 22 top-ranked features in a stepwise manner to identify the smallest subset that included features with at least 5% contribution towards the classifier model. We initially created a feature set that contained the single top-ranked feature *N(1),* which was used to train the ML algorithms to classify dementia and calculate their classification performances. Then, the second top-ranked feature was added to the feature subset to generate a feature set with *N(1)+1* features. The ML classifiers were trained using the new feature subset, and the classification performances were calculated. This process was repeated in descending rank order until a feature set containing all ranked features was included in the feature set. This process resulted in 22 feature sets that ranged in size from

1 to 22 features, with the performance of each feature subset in classifying dementia calculated. The best subset of features was determined as a compromise between performance and size. The data was split into a 30% test set and a 70% training set for each feature set.

Evaluation of classification performance

We formulated the prediction of dementia as a binary classification problem (Dementia, Control); therefore, evaluation metrics, such as accuracy, F1-score, balanced accuracy, precision, specificity, and sensitivity, were used to measure the performance of the subsets of features. The following evaluation metrics were used:

- True positives (TP): Number of dementia cases that were correctly classified.
- False positives (FP): Number of healthy subjects incorrectly classified as dementia cases.
- True negatives (TN): Number of healthy subjects correctly classified.
- False negatives (FN): Number of dementia cases incorrectly classified as healthy subjects.
- Accuracy (%): The proportion of correct classifications among total classifications:

$$Accuracy = \frac{TP+TN}{n} \tag{11}$$

where *n* is the number of total classifications per test.

- Sensitivity (%): The proportion of correctly classified dementia cases.

$$Sensitivity = \frac{TP}{TP+FN} \tag{12}$$

- Specificity (%): The proportion of correctly classified healthy subjects.

$$Specificity = \frac{TN}{TN+FP} \tag{13}$$

- Precision: The proportion of subjects classified as dementia cases who have dementia.

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

- F1-score (F-measure) (%): Harmonic mean of precision and sensitivity.

$$F1 = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision} = \frac{TP}{TP + (FP + FN)/2} \qquad (15)$$

Identifying misclassified cases

Leave-one-out cross-validation was used for training and performance evaluation of trained classifiers using Scikit-learn version 0.22.2.post1 [46] in Python3. A *split()* function was used to enumerate training and test sets for evaluation. The classification algorithms trained the classical AD features using the top-ranked 22 subsets and 114 samples, where one feature was added at a time creating 22 subsets of features for each classifier. All samples were clustered into true positive & true negative, false positive and false negative based on the performance of each classification run, and visualized using a heatmap to highlight the differences. The "*clustermap*" function in Seaborn package version 0.11.0 [46] was used for hierarchical clustering. The linkage method was set to average, and the distance metric was euclidean.

Explaining misclassified cases

To identify pathological and demographic features distinguishing the three clusters of classification performance, we used robust feature selection based on recursive feature elimination (RFE) with a linear SVM as the estimator [47] to identify the smallest set of non-standard pathological features for each of the three clusters [48]. This technique balances performance and computational cost [49]. The linear SVM was initially trained using the complete feature set of the training data with the C-parameter set to one. The absolute weights in the weights vector of the hyperplane of the trained model were used to rank features according to importance, and the worst-performing feature was pruned from the feature set. This process was repeated until the required number of features in the signature was achieved.

For a dataset with J samples and K features, M=100 subsamples were randomly sampled, feature selection was carried out in each subsample, and classification performance was calculated. For each cluster, different sizes of signatures ranged from one to the complete feature set. Each feature set was used to train an XGBoost model to classify the cluster against the rest [50]. The best signature of features for each cluster was chosen as a trade-off between signature size and classification performance. Accuracy and F1-score were used as classification metrics. ML models and feature selection libraries were built using Python 3.8.5, Scikit-learn 24.2, and Jupyterlab 2.2.6. We used the 114 samples and a 'leave-one-out' cross-validation for training and performance evaluation of trained classifiers.

Code availability

Links for python script codes in GitHub (https://github.com/mdrajab/CFAS-ranking-code) for the processes of ranking neuropathology features and classification models and (https://github.com/emmanueljammeh/cfas) for feature signatures showing association of the non-standard pathologies and demographics features with clusters.

Results

Distribution of neuropathology feature scores across dementia cases

Figure 2 depicts the distribution of values of participants dying with and without dementia across all neuropathological features in our study containing 186 samples and 34 attributes. In addition to the 34 neuropathological features, age and brain weight were included. People between 80 and 89 years had a higher frequency of dementia than other age sub-groups. The proportion of individuals with dementia increased with increasing Braak NFT stage, Thal phase and hippocampal tau stage. This validates previous findings from multivariable regression models of dementia and neuropathology [19]. The measures of CAA across subjects revealed that the proportion of dementia cases increased as the number of brain areas with CAA increased. Microinfarct features, in the frontal, occipital and parietal regions, were observed in individuals who died with dementia. A similar observation was seen with Aβ stage typical and Argyrophilic grains, which may limit classifiers from differentiating subjects using these features.

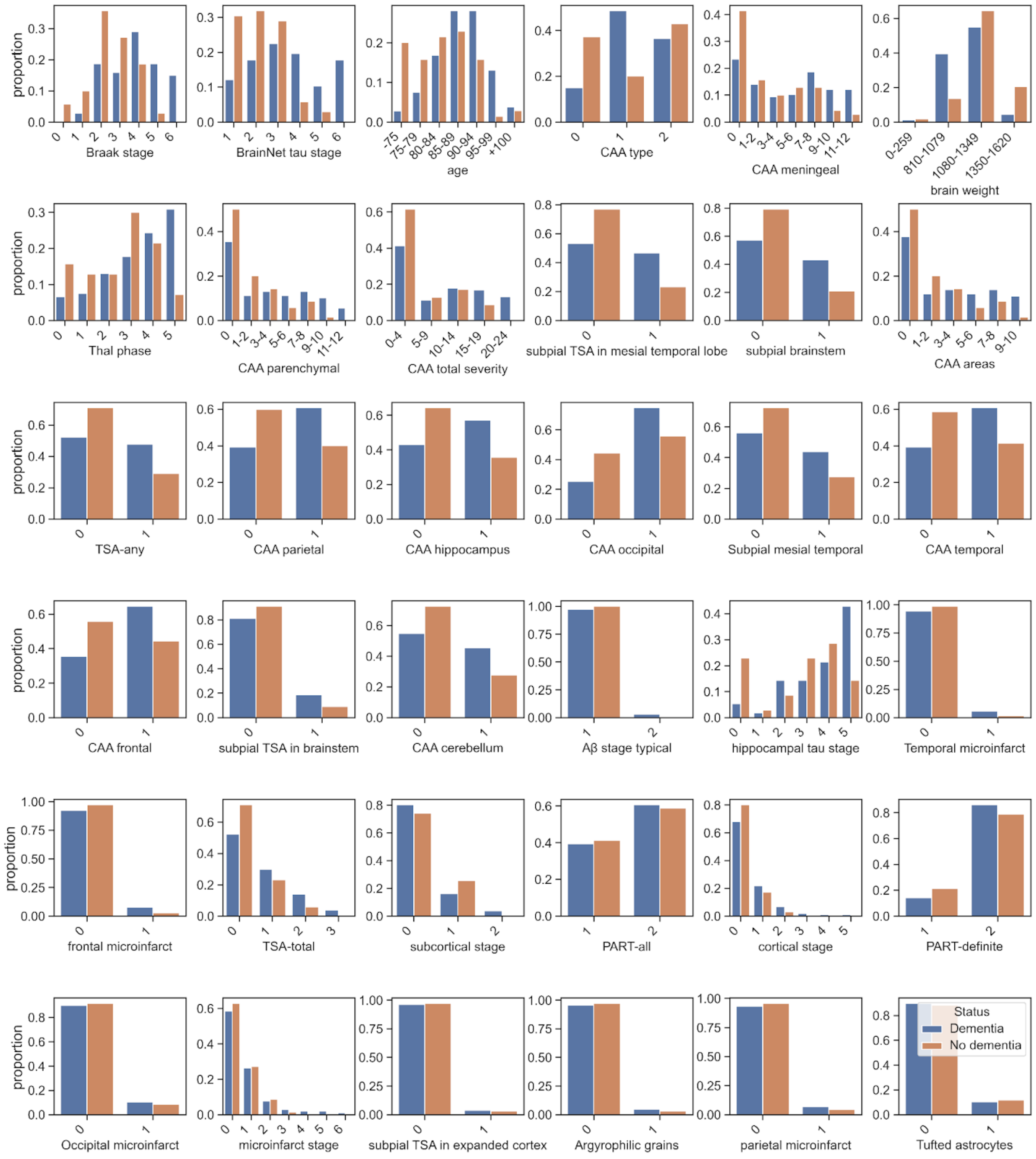


Figure 2: CFAS Neuropathology features distribution. The figure depicts neuropathology features distribution including age and brain weight (proportion of individuals with and

without dementia of the CFAS neuropathology Dataset). All features shown were based on the ranking features list, from left to right. Most features were categorical, but some were ordinal, such as age, CAA total severity, brain weight, CAA areas, TSA-total, cortical stage, subcortical stage and Microinfarct Stage.

Highly correlated neuropathology features

The comparison of features identified highly correlated features (Spearman rho > 0.7), such as CAA-related features. Since CAA-related features, including CAA type, CAA areas, and CAA total severity (CAA meningeal, CAA parenchymal), were shared among the top features presented by the different feature selection methods (**Supplementary Table 1**), we needed to ensure that only distinct features were chosen by minimizing feature-to-feature correlations. We identified three main clusters of highly correlated features (**Figure 3**) when comparing all neuropathology features in our study. Hence, some of these features may be redundant for assessing dementia based on neuropathological features.

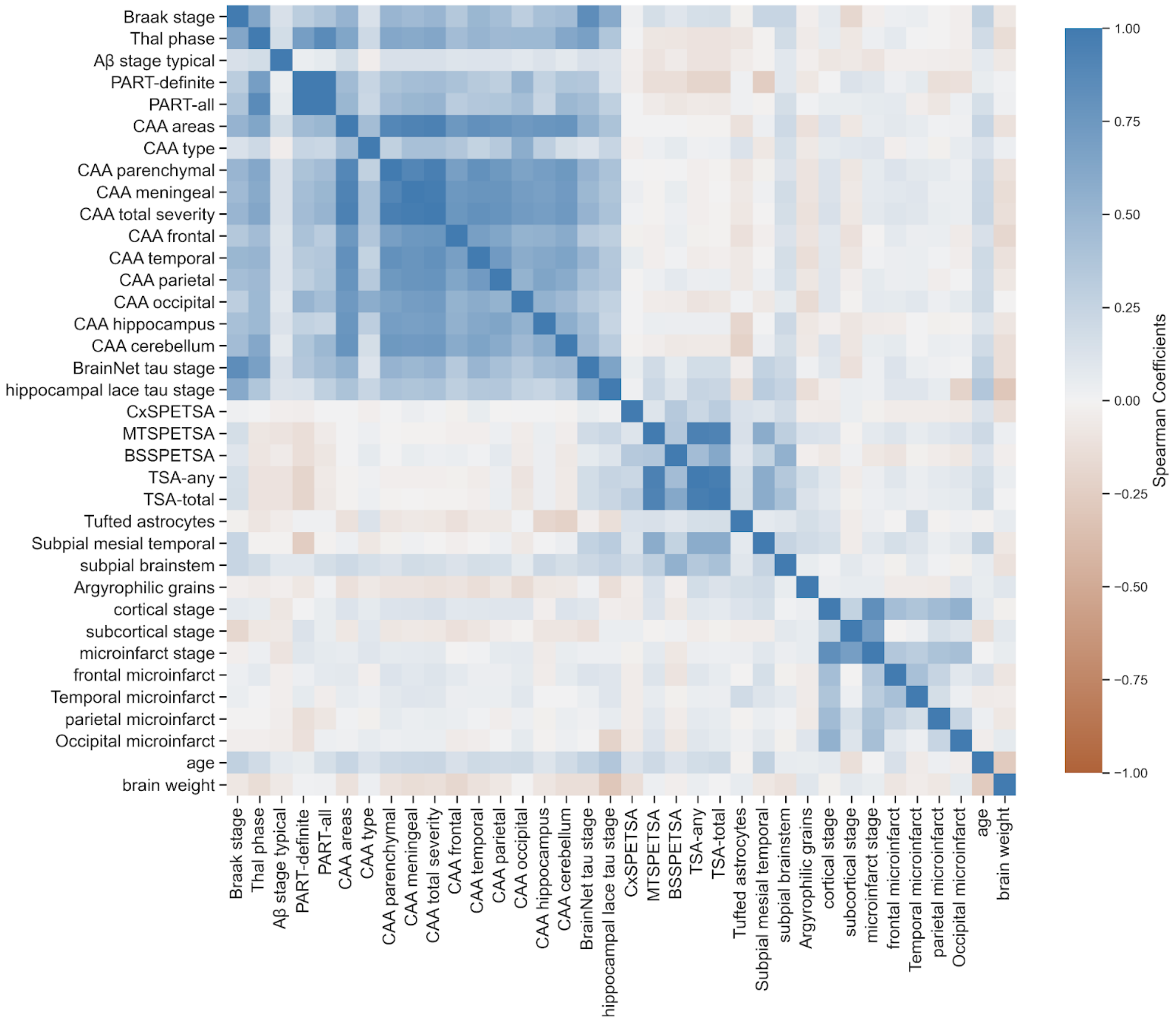


Figure 3: Spearman correlation of the complete CFAS neuropathological data set. Heat map Spearman correlation of the complete CFAS neuropathological data set 34 neuropathology features in addition to age and brain weight features as a benchmark, 36 features in total and 186 samples. A coefficient close to 1(blue colour) means a high positive correlation between the two variables. The diagonal line is the same variable, i.e. spearman rho 1.

Ranking of neuropathology features

The ranking of neuropathology features was conducted to estimate each feature's contribution to dementia using seven feature ranking methods (**Supplementary Table 1**). A high ranking of the Braak NFT stage, which showed the neurofibrillary tangle stage (0–VI), supported it as a highly relevant feature for dementia pathology [22]. All ranking techniques

(CHI, gain ratio, information gain, reliefF, symmetrical uncertainty, least loss and variable analysis) ranked the Braak NFT score in the top six, making it useful for human and computer-aided dementia diagnosis, and should be considered a primary attribute. Different feature selection techniques reported different rankings of the features, however, the most commonly used features were consistently highly ranked. For example, Braak stage, BrainNet tau stage, CAA type, Thal phase, subpial brainstem, and subpial TSA in the mesial temporal lobe were consistently ranked in the top 12 (out of 36) notwithstanding which ranking method was used.

BrainNet tau stage appeared as the top of ranked features, and it had been previously found to be highly correlated with the Braak NFT stage as tangles and neuropil threads seemed to progress together [17]. BrainNet tau stage, a six-stage scheme that uses neuropil threads and was proposed by the BrainNet Europe consortium [51], and has been used to predict dementia in recent research studies. CAA-related features, including CAA type, CAA areas, and CAA total severity, were common among the top features presented by the different feature selection methods (**Supplementary Table 1**). We believed this may be partly due to the high correlation among these CAA-related features (**Figure 3**). Therefore, we evaluated these features to ensure that only dissimilar features were chosen by minimizing feature-to-feature correlations. Lastly, subpial TSA in the mesial temporal lobe appears frequently in the results of all feature selection methods with a high rank. This indicated that the presence of subpial TSA in the mesial temporal lobe had a strong association with dementia.

All 34 neuropathology features, in addition to age and brain weight, and 186 samples were assessed using seven ranking methods (**Supplementary Table 1**; **Figure 4**). We calculated each feature's contribution percentage based on each ranker's weights. We did this by taking each feature's average of the total weight assigned by all filter methods. All features, except parietal microinfarct and Tufted astrocytes, were estimated by one ranking method to have at least 1% contribution to dementia classification. We found a subset of 25 features where all ranking methods estimated a percentage of contribution and at least 5% contribution. In order to assess the utility of neuropathology features to classify dementia, we removed the non-neuropathology features (age and brain weight) and hippocampal tau stage due to high missingness, leaving 22 top ranked features.

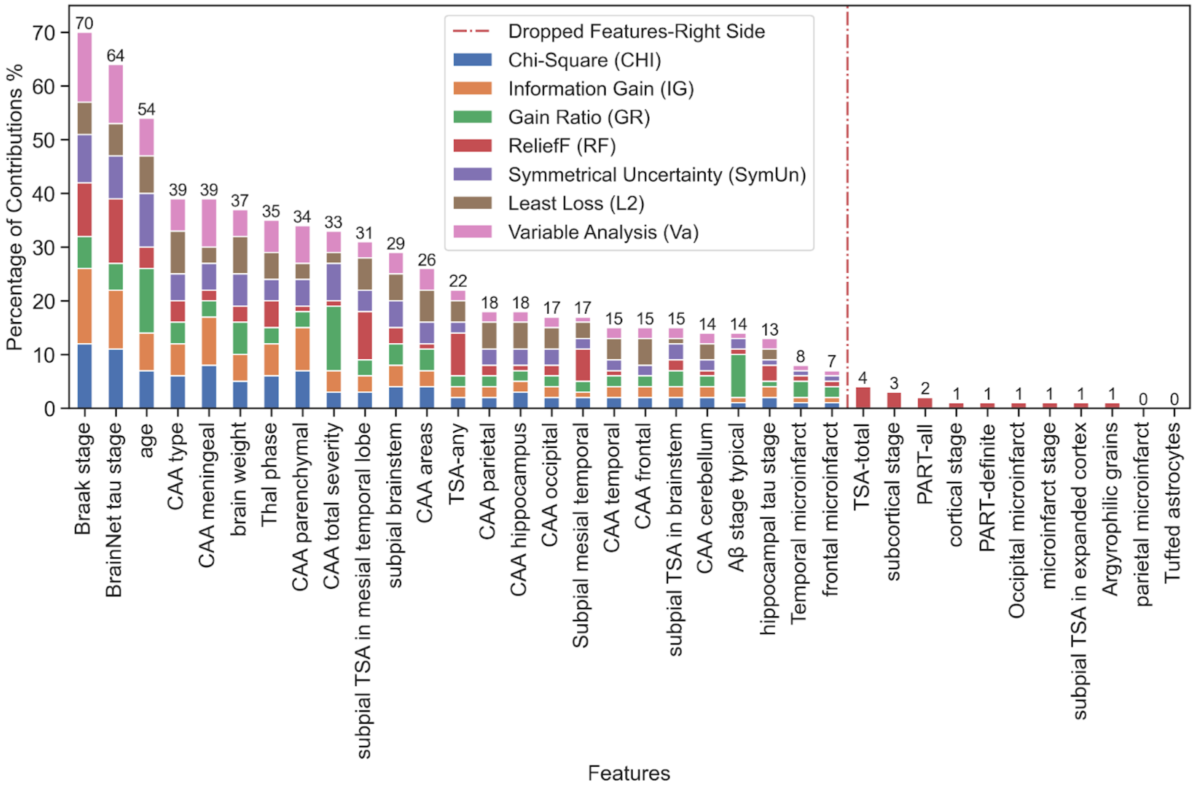


Figure 4: Ranking of neuropathology features. Ranking 34 neuropathology features plus age and brain weight using seven filter methods. After normalizing the weight scores of each feature, the percentage contribution of each feature was calculated by averaging the total weights assigned to each feature by all filter methods. The dotted line indicates features to be dropped, which features percentage contribution show less than 7%.

Classification of the ranked neuropathology features

We further investigated subsets of the top 22 ranked neuropathological features and 114 samples using ML classification. A single feature was successively added from the 22 top ranked feature set to create subsets with sizes ranging from 1 to 22 (from top to lower-ranked features). The dataset was randomly split into a training set containing 70% of the samples and the remaining 30% was used for testing. The training set was used to train classification models using logistic regression, decision tree, k-nearest neighbors, linear discriminant analysis, gaussian naïve Bayes, SVM-RBF, and SVM-LINEAR classification algorithms. The performance of each trained model was evaluated using the test set for prediction. **Supplementary Figure 1** depicts the F1-score performance of all subsets of features (by forward and backward order of ranked features) in classifying dementia status for the seven ML classifiers considered. In the F1-score, the top eight features had the highest performance of 74% using the algorithms SVM-RBF and logistic regression. For comparison with a traditional univariate approach, we trained each neuropathology feature using the seven classifiers and reported their F1-scores. The Thal phase was found to have achieved 69% F1-score using SVM-LINEAR (**Supplementary Figure 2**). The results were supported by the accuracy and balanced accuracy that showed the top eight features' achieving 74% with most classifiers (**Supplementary Figures 3 & 4**). There was no significant improvement in classification beyond the use of eight features. As the number of features was increased beyond eight, most of the trained models performed slightly worse in identifying dementia patients, possibly due to overfitting. We also showed sensitivity and specificity for all models to explain why some of the forward-ranking performances increased when adding the last three features (**Supplementary Figures 6 & 7**). Some of these had class imbalance, resulting in high specificity but low sensitivity. For example, in the linear discriminant analysis classifier, the last five features achieved 84% sensitivity but 50% specificity.

Limits to the accuracy of classification of neuropathology features

Classification results of different feature subsets using the seven classifiers, 114 samples and 22 top-ranked neuropathology features showed that 40.4% of patients were misclassified out of 114 individuals using cross-validation. Further, we investigated the cause of the high misclassification rate. Heatmaps used to visualize the classification of each patient revealed that some cases were misclassified as false positives or negatives, irrespective of the machine learning algorithm used. **Supplementary Figure 7** shows the clustering of patients' classifications from seven classification techniques using multiple subsets of features in order to identify similarities in their performance. Three clusters were identified, containing cases classified correctly, and misclassified as a false positive or false negative. The false positive cluster denoted cases where neuropathology features classified them as having had dementia when in actuality, they did not. Conversely, the false negative cluster denotes cases classified as not having dementia, but in reality, they did. Perhaps, this cluster could correspond to cases of dementia with insufficient neuropathology changes [52].

For each misclassified case (false positive or false negative), we looked at the Mini-Mental State Exam (MMSE) scores at baseline and final interviews (**Supplementary Figure 8**). For false negatives, there were observations of more moderate and severe cases at the final interview compared with baseline. On the other hand, the false positives were evenly distributed as normal, mild and moderate at baseline, with no severe cases. Then we performed further analyses to determine which features were associated with cases where the ranked neuropathology features alone could not explain dementia. Since the classical markers of neuropathology features summarizing the prevalence of plaques and tangles did not classify a large proportion of patients, we hypothesized that non-standard pathologies for rarer dementia syndromes and regional markers could be more helpful. These less common and 'disregarded' pathologies have been described across the CFAS cohort [53]. The non-standard features used were based on more granular neuropathology features in different regions in the brain, such as neuronal loss, gliosis, pick bodies, Lewy bodies, spongiform changes, superficial gliosis, tangles, virchow-robin space expansion and ballooned neurons and some demographic features such as gender, age, and brain weight features.

Our best-performing model for non-standard features, SVM-RFE, effectively removed irrelevant and redundant features to achieve good generalization. The level of each non-standard feature was compared to the classification performance of the classifiers using standard neuropathology (**Figure 5**). We found that the mean age for false negative cases was the highest, with a mean of 89.3 years. In contrast, the false positive mean age was 84.5, and the true positive & true negative mean ages were (88.5 and 80.6) respectively. We also found that the mean brain weight was lower in the false negative cases than in the false positives, true positives, and true negatives. Lewy bodies in the substantia nigra, neuronal loss in the hippocampus, neuronal loss in the substantia nigra, tangles in the temporal lobe, parenchymal CAA in the frontal lobe, and gliosis in the hippocampus could all be combined to explain the classification performance of standard neuropathology (**Supplementary Figure 9**). However, a high proportion of misclassifications occurred where there was a lack of any pathology (**Supplementary Figure 10**). A t-test of each feature also demonstrated no difference in the values of non-standard pathology features between false positives and negatives (**Supplementary Table 2**).

For further evaluation, we combined the top eight classical neuropathological features with the ten non-standard features associated with classifier performance. Together, we tested subsets of the 18 features to classify dementia status. When using classical features, we observed that 40.4% of cases were misclassified; however, when the feature sets were combined, the misclassified cases decreased to 35.1% (**Supplementary Figure 10**). The decrease in misclassification was observed in individuals of at least 85 years old (46.3% to 40.3%) and in those younger than 85 years (31.9% to 27.7%). Of the 32 cases held out, we observed a sensitivity of 68.8% (logistic regression) using the top eight neuropathology features. In contrast, the combined standard and non-standard neuropathological features achieved a better sensitivity of 81.3%.

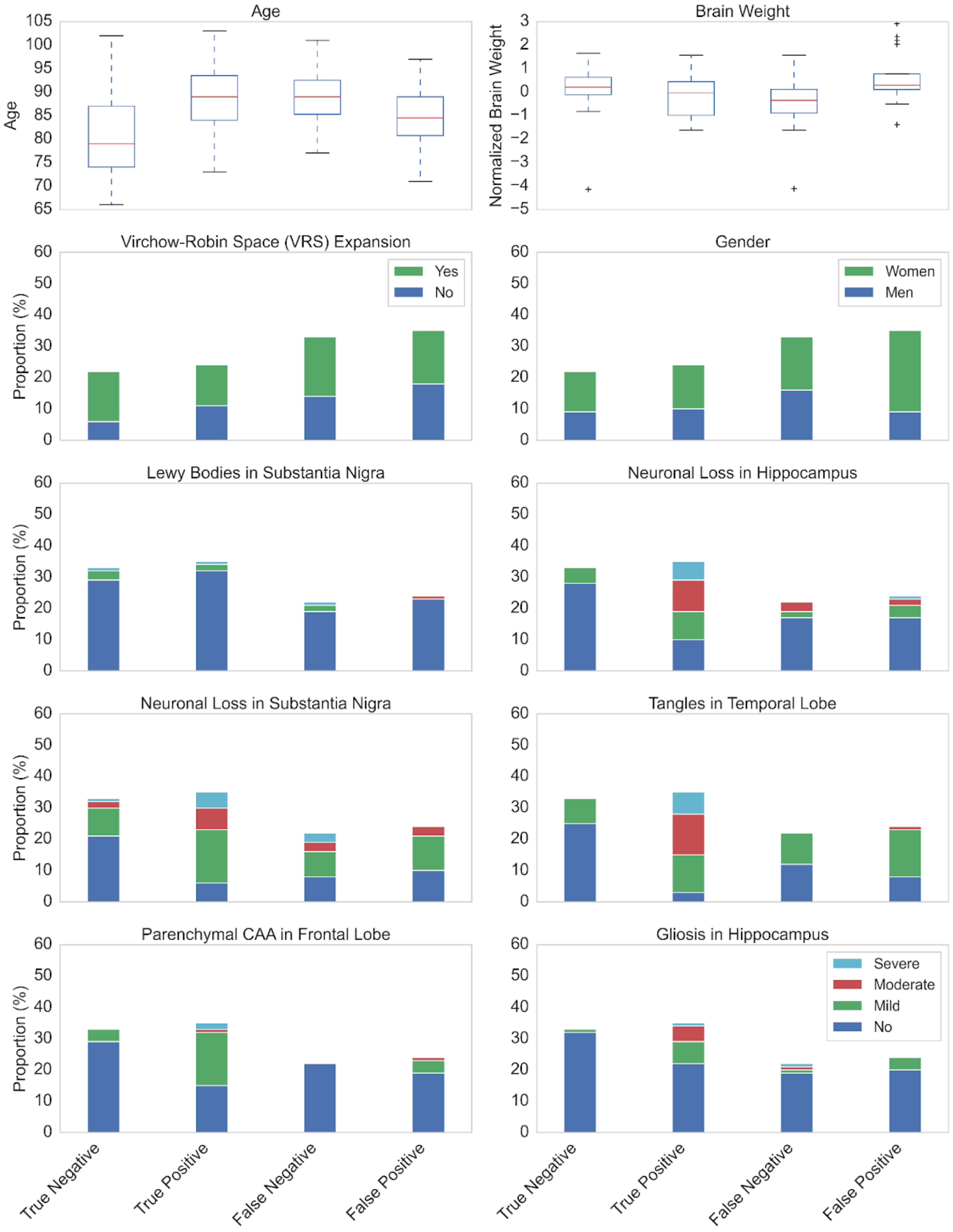


Figure 5: Associations of standard and non-standard neuropathological and demographic features. Non-standard neuropathological and demographic features were associated with misclassified and correctly classified cases by classifiers that used the standard neuropathology features.

Discussions

In this study, we introduced an ML approach to describe how neuropathological features at the end of life were related to dementia. Our step-wise ML approach to rank and select Alzheimer-related pathologies allowed us to investigate how the different measures, such as those related to Aβ-related assessments and tau, can inform about dementia status. The different feature ranking methods resulted in a slightly different ordering of the features in terms of their association with dementia status. However, the top-ranked features were consistent across methods. For example, the Braak NFT and BrainNet tau stages were the top two selected features in line with previous studies [6,17,18,54,55]. However, our results also showed that subpial TSA in the mesial temporal lobe was highly ranked, presenting a contradictory finding from prior studies [6]. Additionally, we identified three clusters of highly correlated measures in the dataset, CAA, TSA, and microinfarct-related, demonstrating that some measures were redundant. Removing these redundant features may reduce collinearity and improve the performance of feature selection and classification accuracy [56–60].

In order to examine the impact of ranking, we tested seven classification algorithms using different subsets of ranked features. Cross-validation during classifier training yielded a maximum classification accuracy of, at most, 74%, using the top eight ranked features. Two subgroups of misclassified participants were identified (false positives and negatives), accounting for 21.2% and 19.3%, respectively. These individuals were consistently misclassified across all classification algorithms. In order to improve classification accuracy, we also considered whether more specific neuropathology features for particular brain regions, which were collected in addition to the standard assessment, could help with classification.

Consistent with previous reports, dementia was most associated with age and brain weights [4]. We further found that the classification of dementia using AD pathology differed between younger and older individuals [8]. Our results suggested that imaging and body fluid biomarkers for a range of pathological changes should be used to identify pathophysiologic processes associated with dementia in individual patients [61–64]. The feature ranking and filtering approaches could be applied to these other sources of pathology data.

The high proportion of misclassifications (35.1%) also indicated discordance between neuropathology and dementia, where some demented individuals had no known pathology and some non-demented individuals with pathology. An explanation for the poor classification performance is that some cases express dementia during life without classical neuropathological changes [52]. Corrada et al. reported that 22% of demented participants did not have sufficient pathology to account for cognitive loss [65]. Using the Vantaa 85+ cohort, Hall et al. showed that cognition and education predicted dementia but not AD or amyloid-related pathologies in the elderly [14]. When combining the top eight neuropathology features with the non-standard pathologies' features, the discordance was less for older individuals (85 years old and above).

The results can be further investigated using other ML techniques, such as embedded feature selection and additional cohorts with the same pathology features and clinical outcomes. Alzheimer's Disease Neuroimaging Initiative [66] or the Rush Memory and Ageing Project [67] could be cohorts to validate our findings from CFAS. However, this requires adjusting for demographic and measurement differences between these other cohorts. Another challenge in relating neuropathology assessments to the clinical diagnosis of dementia was the time lapse between the last assessment of dementia and the post-mortem assessment of the brain. Further follow-up reports on the participant's cognitive status could be collected from those who knew the individual up to the time of death. Pathological features may differ between different types of dementia, such as AD, frontotemporal dementia, vascular disease, and Lewy body dementia [68–70]. There is a need to quantify measures of other key age-related brain pathologies, particularly vascular disease, synuclein staging and age-related Transactive response DNA-binding protein 43 (TDP43) pathology (limbic predominant age-related TDP43 encephalopathy). By doing so, we could link pathology with other symptoms related to dementia. Rather than assessing associations between one feature and an outcome at a time, it would be helpful to investigate whether combinations of features were associated with dementia [71–75].

This study provided a new approach to understanding how much cognitive classification of dementia can be explained by pathological features of the brain. The application of ML as a means of robust evaluation of neuropathological assessments and scores for 186 subjects and 34 neuropathology features from the CFAS cohort highlighted key indices of Alzheimer-related pathologies that may contribute to dementia. While we found that as many as 22 neuropathology features could be independently associated with dementia, tau-related assessments were most informative for ML classifiers of dementia. We hope that further neuropathology studies using multiple feature ranking techniques can lead to identifying more robust biomarkers and enhance the early detection of disease.

References (Manuscript 2)

1. Prince M, Wimo A, Guerchet M, Ali GC, Wu YT, Prina M. Alzheimer’s disease international (2015). world alzheimer report 2015: The global impact of dementia: An analysis of prevalence, incidence, cost and trends. Alzheimer’s Disease International, London [Google Scholar]. 2018;

2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub; 2013.

3. Lancet. Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales [Internet]. The Lancet. 2001. p. 169–75. Available from: http://dx.doi.org/10.1016/s0140-6736(00)03589-3

4. Matthews FE, Brayne C, Lowe J, McKeith I, Wharton SB, Ince P. Epidemiological pathology of dementia: attributable-risks at death in the Medical Research Council Cognitive Function and Ageing Study. PLoS Med. journals.plos.org; 2009;6:e1000180.

5. Brayne C, Nickson J, McCracken C, Gill C, Johnson AL. Cognitive function and dementia in six areas of England and Wales: the distribution of MMSE and prevalence of GMS organicity level in the MRC CFA Study. Psychol Med. CAMBRIDGE UNIV PRESS 32 AVENUE OF THE AMERICAS, NEW YORK, NY 10013-2473 USA; 1998;28:319–35.

6. Wharton SB, Brayne C, Savva GM, Matthews FE, Forster G, Simpson J, et al. Epidemiological neuropathology: the MRC Cognitive Function and Aging Study experience. J Alzheimers Dis. content.iospress.com; 2011;25:359–72.

7. Boyle PA, Yu L, Wilson RS, Leurgans SE, Schneider JA, Bennett DA. Person-specific contribution of neuropathologies to cognitive loss in old age. Ann Neurol. 2018;83:74–83.

8. Savva GM, Wharton SB, Ince PG, Forster G, Matthews FE, Brayne C. Age, neuropathology, and dementia. N Engl J Med. Mass Medical Soc; 2009;360:2302–9.

9. Boyle PA, Yu L, Leurgans SE, Wilson RS, Brookmeyer R, Schneider JA, et al. Attributable risk of Alzheimer's dementia attributed to age-related neuropathologies. Ann Neurol. 2019;85:114–24.

10. Shilaskar S, Ghatol A. Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. Expert Syst Appl. Elsevier; 2013;40:4146–53.

11. Verma AK, Pal S, Kumar S. Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study. Appl Biochem Biotechnol. Springer; 2020;190:341–59.

12. Castellazzi G, Cuzzoni MG, Cotta Ramusino M, Martinelli D, Denaro F, Ricciardi A, et al. A Machine Learning Approach for the Differential Diagnosis of Alzheimer and Vascular Dementia Fed by MRI Selected Features. Front Neuroinform. frontiersin.org; 2020;14:25.

13. Thapa S, Singh P, Jain DK, Bharill N, Gupta A, Prasad M. Data-driven approach based on feature selection technique for early diagnosis of Alzheimer's disease. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. p. 1–8.

14. Hall A, Pekkala T, Polvikoski T, van Gils M, Kivipelto M, Lötjönen J, et al. Prediction models for dementia and neuropathology in the oldest old: the Vantaa 85+ cohort study. Alzheimers Res Ther. Springer; 2019;11:11.

15. Thal DR, Rüb U, Orantes M, Braak H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD. Neurology. AAN Enterprises; 2002;58:1791–800.

16. Murray ME, Lowe VJ, Graff-Radford NR, Liesinger AM, Cannon A, Przybelski SA, et al. Clinicopathologic and 11C-Pittsburgh compound B implications of Thal amyloid phase across the Alzheimer's disease spectrum. Brain. academic.oup.com; 2015;138:1370–81.

17. Wharton SB, Minett T, Drew D, Forster G, Matthews F, Brayne C, et al. Epidemiological pathology of Tau in the ageing brain: application of staging for neuropil threads (BrainNet Europe protocol) to the MRC cognitive function and ageing brain study. Acta Neuropathologica Communications. actaneurocomms.biomedcentral …; 2016;4:11.

18. Wharton SB, Wang D, Parikh C, Matthews FE, Brayne C, Ince PG. 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. BioMed Central; 2019;7:1–12.

19. Wharton SB, Wang D, Parikh C, Matthews FE, Brayne C, Ince PG, et al. 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 Neuropathol Commun. 2019;7:198.

20. Thal DR, Rüb U, Orantes M, Braak H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD [Internet]. Neurology. 2002. p. 1791–800. Available from: http://dx.doi.org/10.1212/wnl.58.12.1791

21. Alafuzoff I, Thal DR, Arzberger T, Bogdanovic N, Al-Sarraj S, Bodi I, et al. Assessment of β-amyloid deposits in human brain: a study of the BrainNet Europe Consortium [Internet]. Acta Neuropathologica. 2009. p. 309–20. Available from: http://dx.doi.org/10.1007/s00401-009-0485-4

22. Braak H, Alafuzoff I, Arzberger T, Kretzschmar H, Del Tredici K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol. Springer; 2006;112:389–404.

23. Mirra SS, Heyman A, McKeel D, Sumi SM, Crain BJ, Brownlee LM, et al. The Consortium to Establish a Registry for Alzheimer’s Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology. AAN Enterprises; 1991;41:479–86.

24. Lace G, Savva GM, Forster G, de Silva R, Brayne C, Matthews FE, et al. Hippocampal tau pathology is related to neuroanatomical connections: an ageing population-based study. Brain. academic.oup.com; 2009;132:1324–34.

25. Ince PG, Minett T, Forster G, Brayne C, Wharton SB, Function MRCC, et al. Microinfarcts in an older population-representative brain donor cohort (MRC CFAS): Prevalence, relation to dementia and mobility, and implications for the evaluation of cerebral Small Vessel Disease. Neuropathol Appl Neurobiol. Wiley Online Library; 2017;43:409–18.

26. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. Springer; 1991;82:239–59.

27. Crary JF, Trojanowski JQ, Schneider JA, Abisambra JF, Abner EL, Alafuzoff I, et al. Primary age-related tauopathy (PART): a common pathology associated with human aging. Acta Neuropathol. Springer; 2014;128:755–66.

28. Love S, Chalmers K, Ince P, Esiri M, Attems J, Jellinger K, et al. Development, appraisal, validation and implementation of a consensus protocol for the assessment of cerebral amyloid angiopathy in post-mortem brain tissue. Am J Neurodegener Dis. 2014;3:19–32.

29. Ikeda K. Glial fibrillary tangles and argyrophilic threads: Classification and disease specificity [Internet]. Neuropathology. 1996. p. 71–7. Available from: http://dx.doi.org/10.1111/j.1440-1789.1996.tb00158.x

30. Ikeda K, Akiyama H, Arai T, Nishimura T. Glial Tau Pathology in Neurodegenerative Diseases: Their Nature and Comparison with Neuronal Tangles [Internet]. Neurobiology of Aging. 1998. p. S85–91. Available from: http://dx.doi.org/10.1016/s0197-4580(98)00034-7

31. Ikeda K, Akiyama H, Kondo H, Haga C, Tanno E, Tokuda T, et al. Thorn-shaped astrocytes: possibly secondarily induced tau-positive glial fibrillary tangles [Internet]. Acta Neuropathologica. 1995. p. 620–5. Available from: http://dx.doi.org/10.1007/bf00318575

32. Nishimura M, Namba Y, Ikeda K, Oda M. Glial fibrillary tangles with straight tubules in the brains of patients with progressive supranuclear palsy [Internet]. Neuroscience Letters. 1992. p. 35–8. Available from: http://dx.doi.org/10.1016/0304-3940(92)90227-x

33. Marioni RE, Matthews FE, Brayne C, MRC Cognitive Function and Ageing Study. The association between late-life cognitive test scores and retrospective informant interview data. Int Psychogeriatr. research.ed.ac.uk; 2011;23:274–9.

34. Huan Liu, Setiono R. Chi2: feature selection and discretization of numeric attributes. Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. 1995. p. 388–91.

35. Kononenko I. On biases in estimating multi-valued attributes. Ijcai. Citeseer; 1995. p. 1034–40.

36. Quinlan JR. Induction of decision trees [Internet]. Machine Learning. 1986. p. 81–106. Available from: http://dx.doi.org/10.1007/bf00116251

37. Robnik-Šikonja M, Kononenko I. Machine Learning [Internet]. 2003. p. 23–69. Available from: http://dx.doi.org/10.1023/a:1025667309714

38. Novakovic J, Strbac P, Bulatovic D. Toward optimal feature selection using ranking methods and classification algorithms [Internet]. Yugoslav Journal of Operations Research. 2011. p. 119–35. Available from: http://dx.doi.org/10.2298/yjor1101119n

39. Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. J Mach Learn Res. 2004;5:1205–24.

40. Thabtah F, Kamalov F, Hammoud S, Shahamiri SR. Least Loss: A simplified filter method for feature selection. Inf Sci . Elsevier; 2020;534:1–15.

41. Rajab KD. New Hybrid Features Selection Method: A Case Study on Websites Phishing. Security and Communication Networks [Internet]. Hindawi; 2017 [cited 2019 Nov 12];2017. Available from: https://www.hindawi.com/journals/scn/2017/9838169/abs/

42. Kamalov F, Thabtah F. A Feature Selection Method Based on Ranked Vector Scores of Features for Classification [Internet]. Annals of Data Science. 2017. p. 483–502. Available from: http://dx.doi.org/10.1007/s40745-017-0116-1

43. Rajab M, Wang D. Practical Challenges and Recommendations of Filter Methods for Feature Selection. J Info Know Mgmt. World Scientific Publishing Co.; 2020;2040019.

44. Hall MA. Correlation-based Feature Selection for Machine Learning. 1999.

45. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data

mining software: an update. ACM SIGKDD Explorations Newsletter. ACM; 2009;11:10–8.

46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. JMLR. org; 2011;12:2825–30.

47. Lin X, Li C, Zhang Y, Su B, Fan M, Wei H. Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics. Molecules [Internet]. mdpi.com; 2017;23. Available from: http://dx.doi.org/10.3390/molecules23010052

48. Xia J, Sun L, Xu S, Xiang Q, Zhao J, Xiong W, et al. A Model Using Support Vector Machines Recursive Feature Elimination (SVM-RFE) Algorithm to Classify Whether COPD Patients Have Been Continuously Managed According to GOLD Guidelines [Internet]. International Journal of Chronic Obstructive Pulmonary Disease. 2020. p. 2779–86. Available from: http://dx.doi.org/10.2147/copd.s271237

49. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. academic.oup.com; 2007;23:2507–17.

50. Chen T, Guestrin C. XGBoost [Internet]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. Available from: http://dx.doi.org/10.1145/2939672.2939785

51. Alafuzoff I, Arzberger T, Al-Sarraj S, Bodi I, Bogdanovic N, Braak H, et al. Staging of neurofibrillary pathology in Alzheimer's disease: a study of the BrainNet Europe Consortium. Brain Pathol. Wiley Online Library; 2008;18:484–96.

52. Serrano-Pozo A, Qian J, Monsell SE, Blacker D, Gómez-Isla T, Betensky RA, et al. Mild to moderate Alzheimer dementia with insufficient neuropathological changes. Ann Neurol. Wiley Online Library; 2014;75:597–601.

53. Keage HAD, Ince PG, Matthews FE, Wharton SB, McKeith IG, Brayne C, et al. Impact of less common and "disregarded" neurodegenerative pathologies on dementia burden in a population-based cohort. J Alzheimers Dis. IOS Press; 2012;28:485–93.

54. Lace G, Ince PG, Brayne C, Savva GM, Matthews FE, de Silva R, et al. Mesial temporal astrocyte tau pathology in the MRC-CFAS ageing brain cohort. Dement Geriatr Cogn Disord. 2012;34:15–24.

55. Keo A, Mahfouz A, Ingrassia AMT, Meneboo J-P, Villenet C, Mutez E, et al. Transcriptomic signatures of brain regional vulnerability to Parkinson's disease. Commun Biol. 2020;3:101.

56. Jain I, Jain VK, Jain R. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification [Internet]. Applied Soft Computing. 2018. p. 203–15. Available from: http://dx.doi.org/10.1016/j.asoc.2017.09.038

57. Mwadulo MW. A review on feature selection methods for classification tasks [Internet]. Citeseer; 2016 [cited 2021 Apr 6]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1075.7828&rep=rep1&type=pdf

58. Shi H, Li H, Zhang D, Cheng C, Cao X. An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification. Computer

Networks. Elsevier; 2018;132:81–98.

59. Gómez Flores W, Pereira WC de A, Infantosi AFC. Improving classification performance of breast lesions on ultrasonography. Pattern Recognit. Elsevier; 2015;48:1125–36.

60. Agarwal B, Mittal N. Prominent feature extraction for review analysis: an empirical study. J Exp Theor Artif Intell. Taylor & Francis; 2016;28:485–98.

61. Dallaire-Théroux C, Beheshti I, Potvin O, Dieumegarde L, Saikali S, Duchesne S, et al. Braak neurofibrillary tangle staging prediction from in vivo MRI metrics. Alzheimers Dement. 2019;11:599–609.

62. Lantero-Rodriguez J, Snellman A, Benedet AL, Milà-Alomà M, Camporesi E, Montoliu-Gaya L, et al. P-tau235: a novel biomarker for staging preclinical Alzheimer's disease. EMBO Mol Med. 2021;13:e15098.

63. Banerjee G, Ambler G, Keshavan A, Paterson RW, Foiani MS, Toombs J, et al. Cerebrospinal Fluid Biomarkers in Cerebral Amyloid Angiopathy. J Alzheimers Dis. content.iospress.com; 2020;74:1189–201.

64. Kim HJ, Park D, Yun G, Kim H, Kim H-G, Lee KM, et al. Screening for cerebral amyloid angiopathy based on serological biomarkers analysis using a dielectrophoretic force-driven biosensor platform. Lab Chip. pubs.rsc.org; 2021;21:4557–65.

65. M. Corrada M, J. Berlau D, H. Kawas C. A Population-Based Clinicopathological Study in the Oldest-Old: The 90+ Study. Curr Alzheimer Res. ingentaconnect.com; 2012;9:709–17.

66. Weiner MW, Aisen PS, Jack CR Jr, Jagust WJ, Trojanowski JQ, Shaw L, et al. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimers Dement. Wiley; 2010;6:202–11.e7.

67. Bennett DA, Schneider JA, Buchman AS, Mendes de Leon C, Bienias JL, Wilson RS. The Rush Memory and Aging Project: study design and baseline characteristics of the study cohort. Neuroepidemiology. karger.com; 2005;25:163–75.

68. Elahi FM, Miller BL. A clinicopathological approach to the diagnosis of dementia [Internet]. Nature Reviews Neurology. 2017. p. 457–76. Available from: http://dx.doi.org/10.1038/nrneurol.2017.96

69. Barker WW, Luis CA, Kashuba A, Luis M, Harwood DG, Loewenstein D, et al. Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the State of Florida Brain Bank. Alzheimer Dis Assoc Disord. journals.lww.com; 2002;16:203–12.

70. Geldmacher DS, Whitehouse PJ. Evaluation of dementia. N Engl J Med. Mass Medical Soc; 1996;335:330–6.

71. Hoque A, Galib S, Tasnim M. Mining pathological data to support medical diagnostics. Workshop on Advances on Data Management: Applications and Algorithms, Department of Computer Science and Engineering, BUET, Dhaka. academia.edu; 2013. p. 71–4.

72. Kherif F, Muller S. Neuro-Clinical Signatures of Language Impairments: A Theoretical Framework for Function-to-structure Mapping in Clinics. Curr Top Med Chem.

ingentaconnect.com; 2020;20:800–11.

73. Allen TA, Schreiber AM, Hall NT, Hallquist MN. From Description to Explanation: Integrating Across Multiple Levels of Analysis to Inform Neuroscientific Accounts of Dimensional Personality Pathology. J Pers Disord. 2020;34:650–76.

74. Gaiteri C, Mostafavi S, Honey CJ, De Jager PL. Genetic variants in Alzheimer disease—molecular and brain network approaches. Nat Rev [Internet]. nature.com; 2016; Available from: https://www.nature.com/articles/nrneurol.2016.84.pdf?origin=ppub

75. Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med. Elsevier; 2010;48:139–52.

Declarations

Funding

This work was supported by the Medical Research Council (MRC/G9901400, U.1052.00.0013, G0900582). SBW is also supported by the Alzheimer's Society (AS-PG-17-007 and AS-PG-14-015). Work in the individual CFAS centres is supported by the UK NIHR Biomedical Research Centre for Ageing and Age – awarded to Newcastle-upon-Tyne Hospitals Foundation Trust; Cambridge Brain Bank supported by the NIHR Cambridge Biomedical Research Centre; Nottingham University Hospitals NHS Trust; University of Sheffield, Sheffield Teaching Hospitals NHS Foundation Trust and NIHR Sheffield Biomedical Research Centre; The Thomas Willis Oxford Brain Collection, supported by the Oxford Biomedical Research Centre; The Walton Centre NHS Foundation Trust, Liverpool. DW and EJ received support from the Academy of Medical Sciences Springboard (SBF004/1052). MR is supported by the Saudi Arabia Ministry of Education. We would like to acknowledge the essential contribution of the liaison officers, the general practitioners, their staff, and nursing and residential home staff. We are grateful to our respondents and their families for their generous gift to medical research, which has made this study possible.

Ethical Approval and consent to participate

For the CFAS dataset, fully written informed consents were obtained from all participants or their authorized representatives, and the study was conducted in accordance with the ethical standards of the Declaration of Helsinki. The study was undertaken with ethical approval from a UK Multicentre Research Ethics Committee (10/H0304/61).

Consent for publication

Not Applicable

Availability of data and materials

Data from the CFAS study is accessible via application to the CFAS (http://www.cfas.ac.uk/cfas-i/data/#cfasi-data-request), under the custodianship of FM and CB. The authors declare that they have no competing interests.

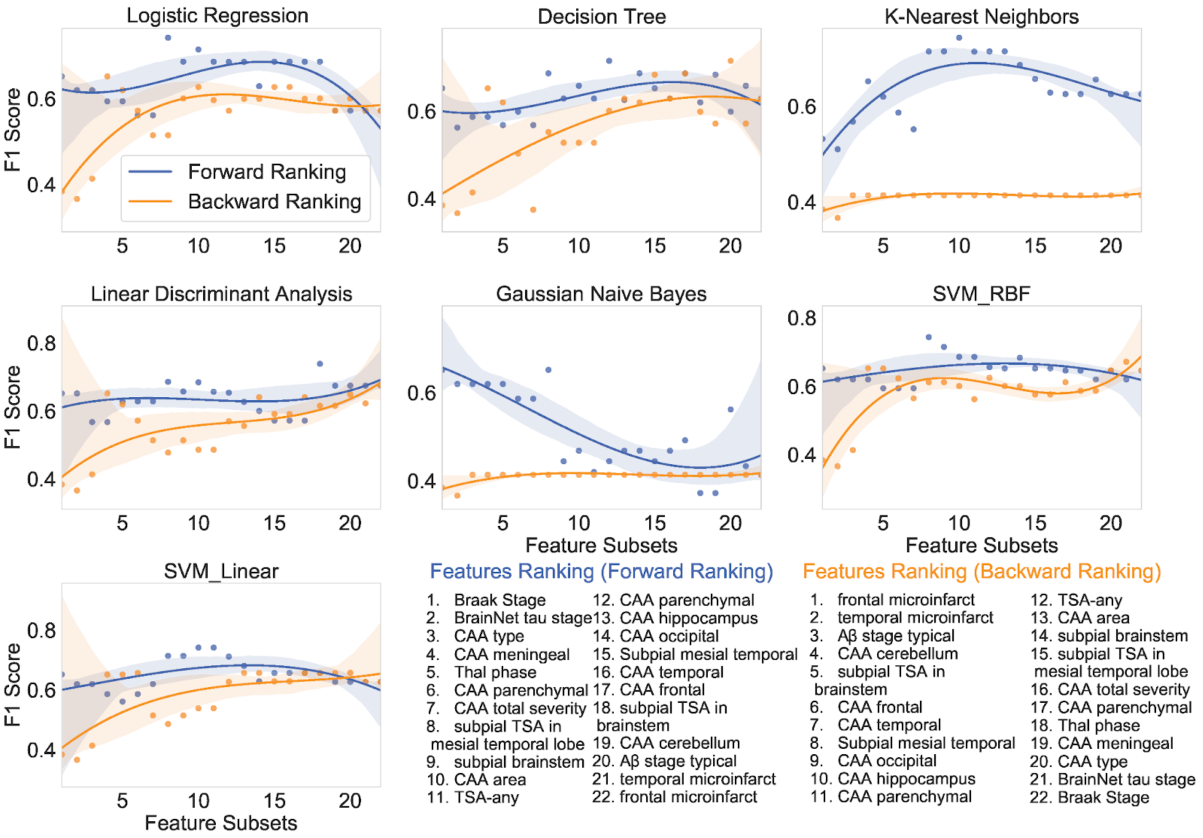
Contributions

Study design and assessment of tissue sections; SBW, PGI. Data analysis; MR, EJ, TT, DW. Writing of first draft MR, EJ, TT, DW. Data oversight and analysis results interpretation; LS, FM, CB. Contribution to interpretation and to the final manuscript; all authors. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Supplementary Materials (Manuscript 2)



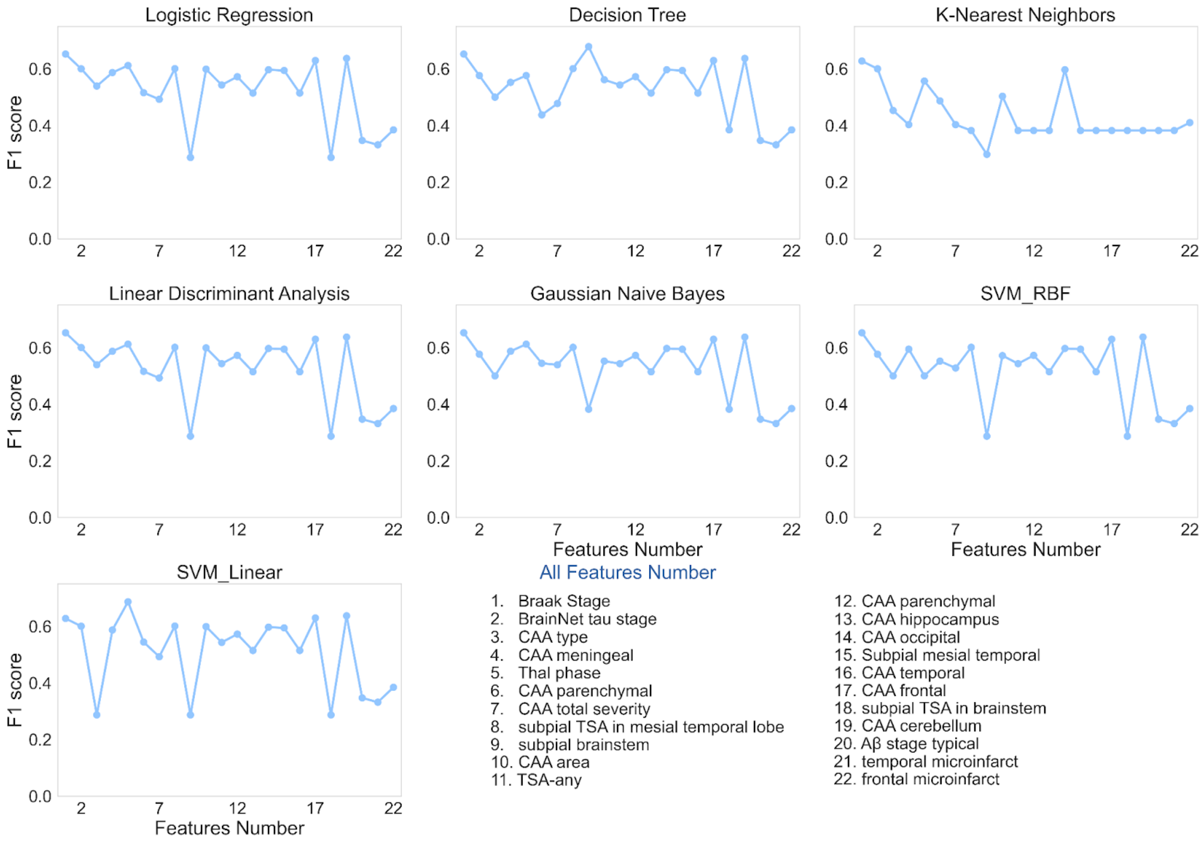
Features Ranking (Forward Ranking)

1. Braak Stage
2. BrainNet tau stage
3. CAA type
4. CAA meningeal
5. Thal phase
6. CAA parenchymal
7. CAA total severity
8. subpial TSA in mesial temporal lobe
9. subpial brainstem
10. CAA area
11. TSA-any
12. CAA parenchymal
13. CAA hippocampus
14. CAA occipital
15. Subpial mesial temporal
16. CAA temporal
17. CAA frontal
18. subpial TSA in brainstem
19. CAA cerebellum
20. Aβ stage typical
21. temporal microinfarct
22. frontal microinfarct

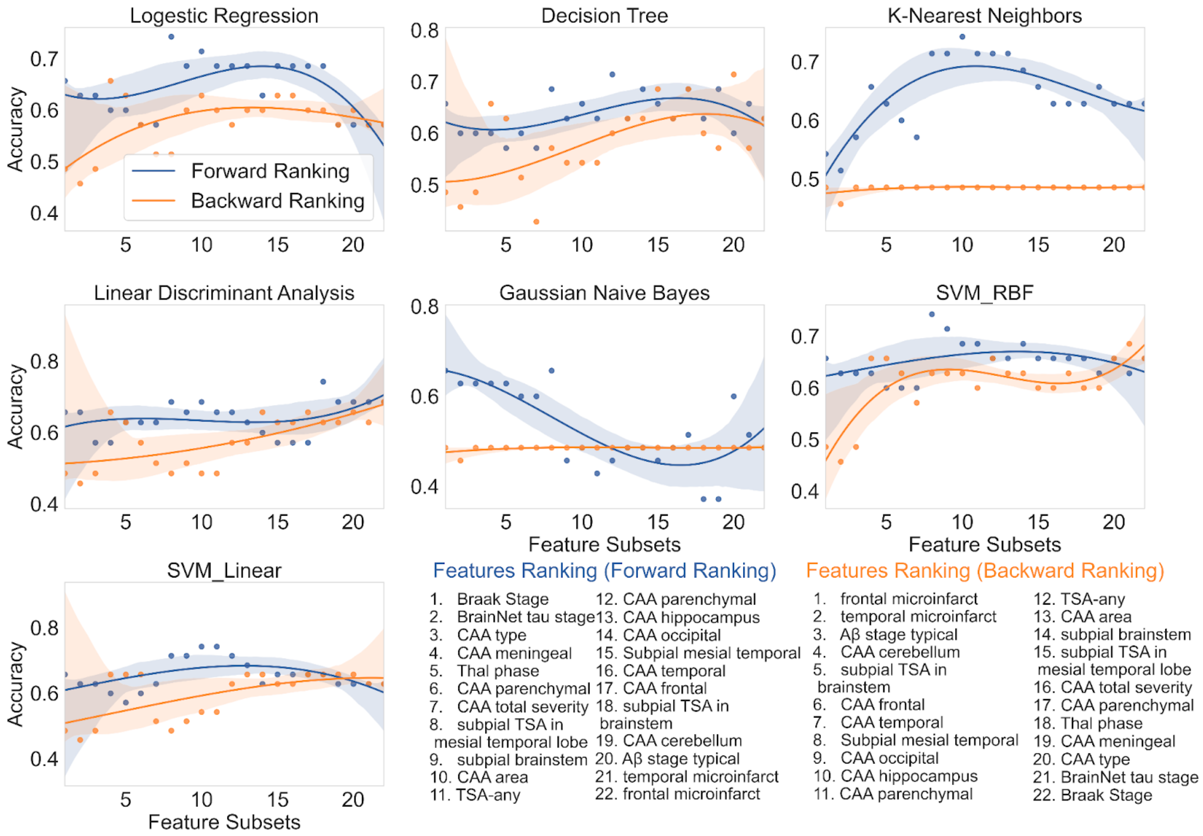
Features Ranking (Backward Ranking)

1. frontal microinfarct
2. temporal microinfarct
3. Aβ stage typical
4. CAA cerebellum
5. subpial TSA in brainstem
6. CAA frontal
7. CAA temporal
8. Subpial mesial temporal
9. CAA occipital
10. CAA hippocampus
11. CAA parenchymal
12. TSA-any
13. CAA area
14. subpial brainstem
15. subpial TSA in mesial temporal lobe
16. CAA total severity
17. CAA parenchymal
18. Thal phase
19. CAA meningeal
20. CAA type
21. BrainNet tau stage
22. Braak Stage

Supplementary Figure 1. Performance of all subsets of neuropathology features. F1-score performance of all subsets of neuropathology features from the rank list forward and backward rankings. Forward ranking (blue) adds to the classifier model from the top feature to the lowest feature, while backward ranking (orange) adds to the model from the lowest feature to the top feature. Seven classifiers were utilized in this investigation: logistic regression, decision tree, k-nearest neighbors, linear discriminant analysis, gaussian naive bayes, support vector machines with radial basis function kernel, and support vector machines with linear kernel. Please see (**Supplementary: Figures 3-6**) for other metrics such as accuracy, balanced accuracy, sensitivity, and specificity.



Supplementary Figure 2. F1-score performance of each single neuropathology feature from the rank list. This is to show comparison of a traditional univariate approach. Seven classifiers were utilized in this investigation with top score in each: Logistic Regression (Braak stage: 0.65), Decision Tree (Subpial Brainstem: 0.68), k-Nearest Neighbors (Braak stage: 0.63), Linear Discriminant Analysis (Braak stage: 0.65), Gaussian Naive Bayes (Braak stage: 0.65), Support Vector Machines with Radial Basis Function kernel (Braak stage: 0.65), and Support Vector Machines with Linear kernel (Thal phase: 0.69).



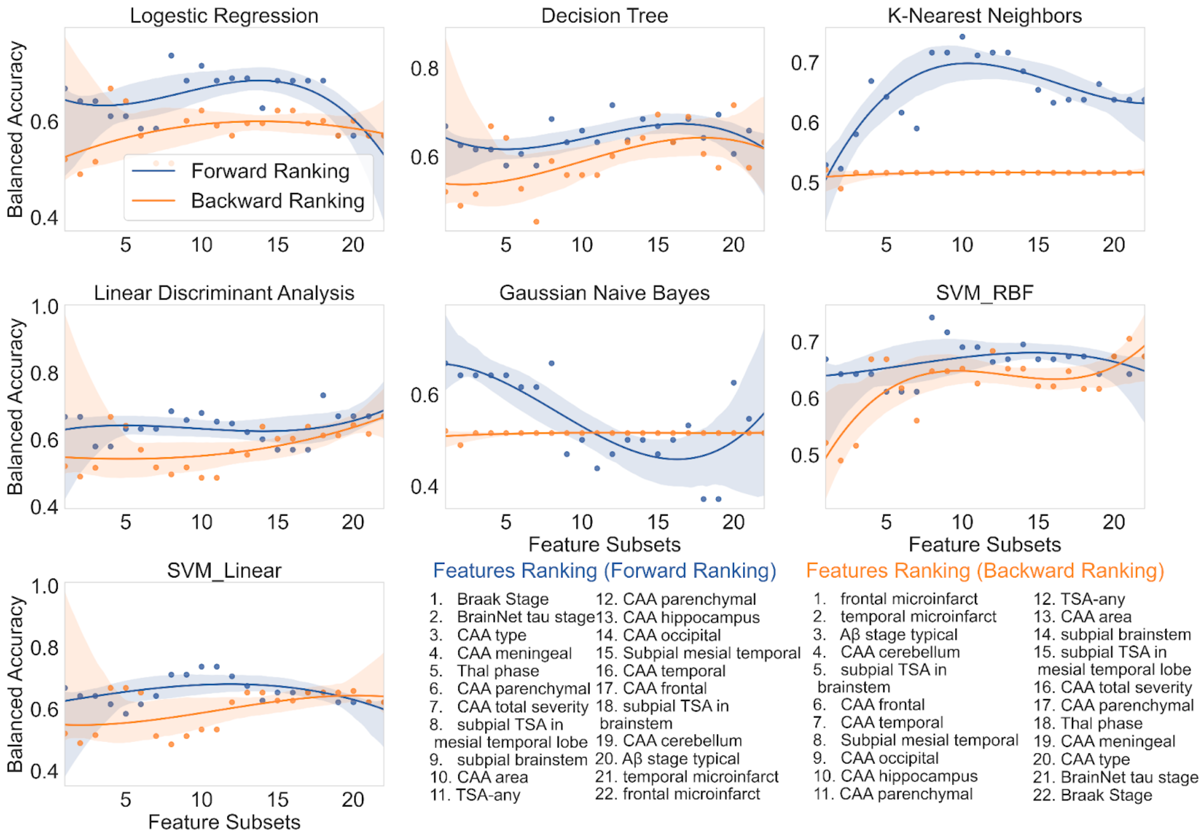
Features Ranking (Forward Ranking)

1. Braak Stage
2. BrainNet tau stage
3. CAA type
4. CAA meningeal
5. Thal phase
6. CAA parenchymal
7. CAA total severity
8. subpial TSA in mesial temporal lobe
9. subpial brainstem
10. CAA area
11. TSA-any
12. CAA parenchymal
13. CAA hippocampus
14. CAA occipital
15. Subpial mesial temporal
16. CAA temporal
17. CAA frontal
18. subpial TSA in brainstem
19. CAA cerebellum
20. Aβ stage typical
21. temporal microinfarct
22. frontal microinfarct

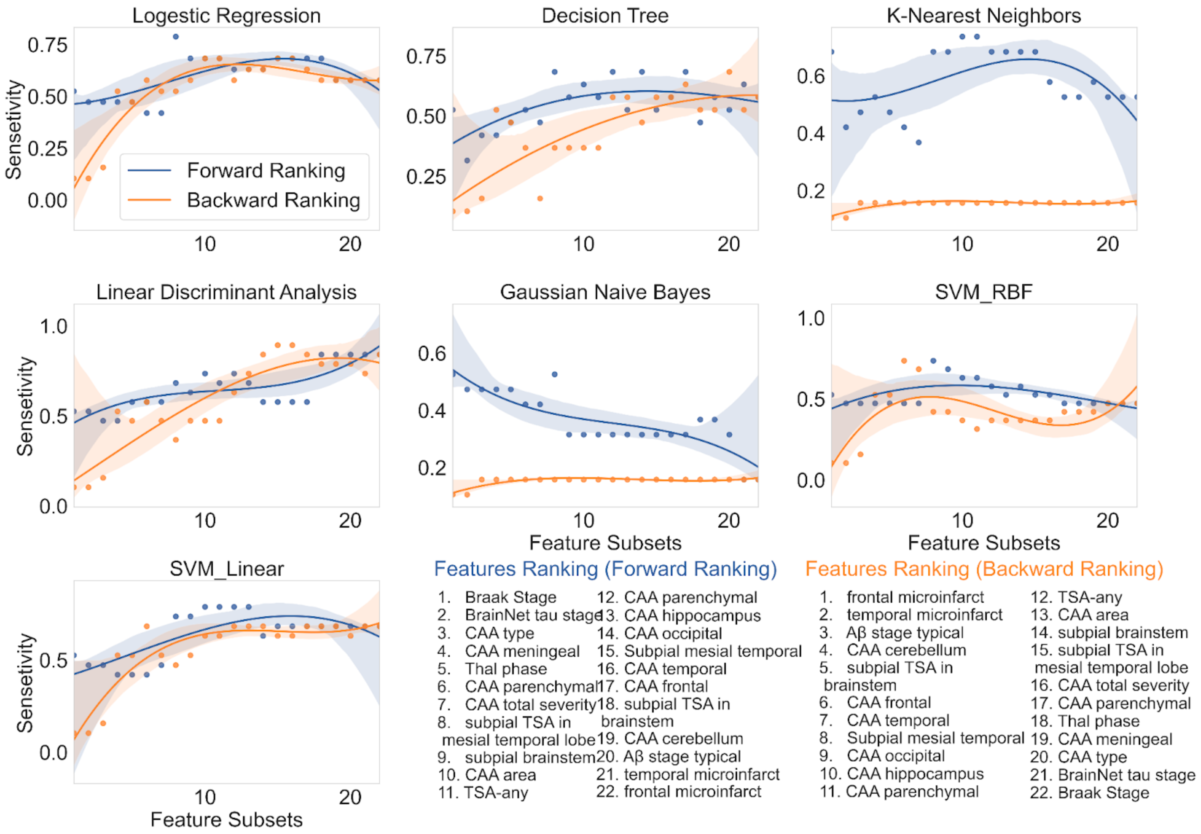
Features Ranking (Backward Ranking)

1. frontal microinfarct
2. temporal microinfarct
3. Aβ stage typical
4. CAA cerebellum
5. subpial TSA in brainstem
6. CAA frontal
7. CAA temporal
8. Subpial mesial temporal
9. CAA occipital
10. CAA hippocampus
11. CAA parenchymal
12. TSA-any
13. CAA area
14. subpial brainstem
15. subpial TSA in mesial temporal lobe
16. CAA total severity
17. CAA parenchymal
18. Thal phase
19. CAA meningeal
20. CAA type
21. BrainNet tau stage
22. Braak Stage

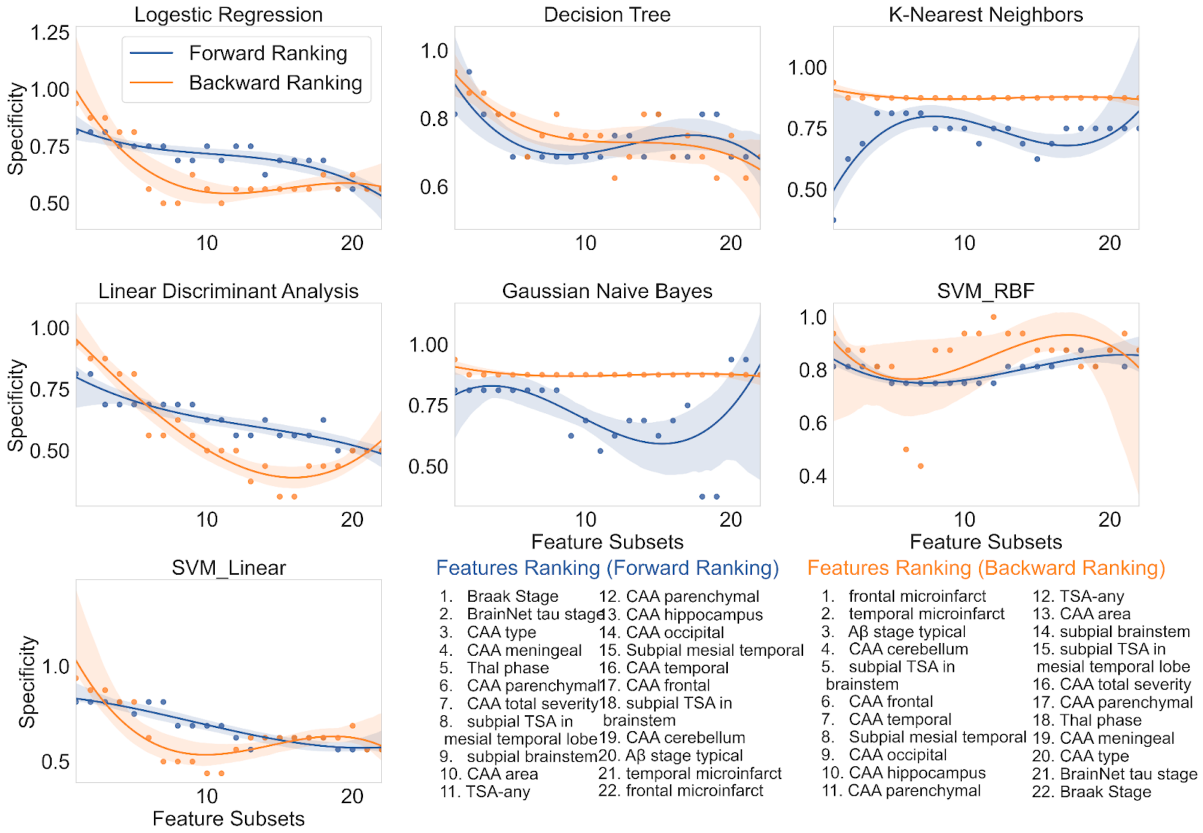
Supplementary Figure 3. Accuracy performance of all subsets of neuropathology features from the rank list forward and backward rankings. Forward ranking (blue) adds to the classifier model from the top feature to the lowest feature while the backward ranking (orange) adds to the model from the lowest feature to the top feature. Seven classifiers were utilized in this investigation: Logistic Regression, Decision Tree, k-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, Support Vector Machines with Radial Basis Function kernel, and Support Vector Machines with Linear kernel.



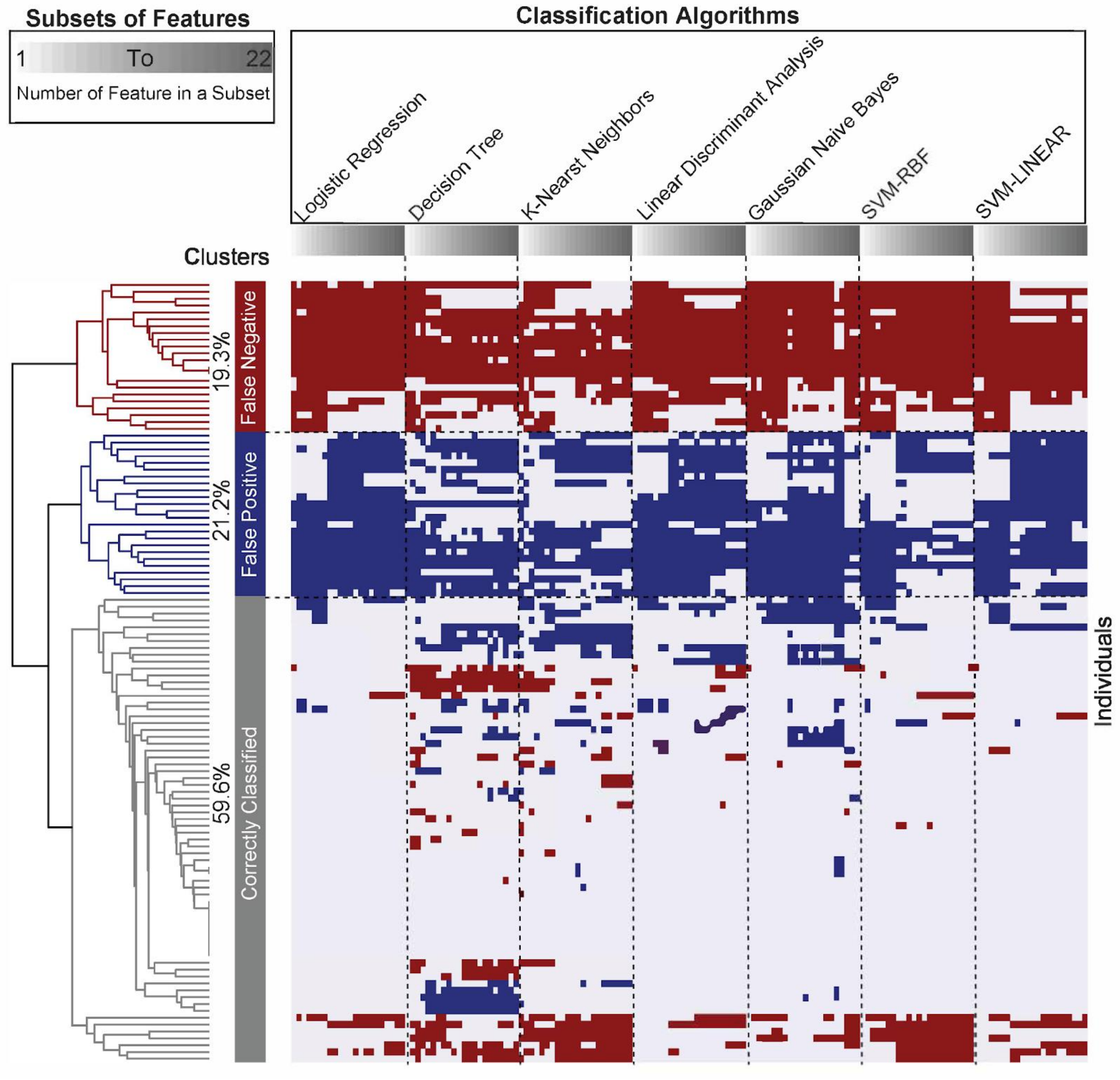
Supplementary Figure 4. Balanced Accuracy performance of all subsets of neuropathology features from the rank list forward and backward rankings. Forward ranking (blue) adds to the classifier model from the top feature to the lowest feature while the backward ranking (orange) adds to the model from the lowest feature to the top feature. Seven classifiers were utilized in this investigation: Logistic Regression, Decision Tree, k-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, Support Vector Machines with Radial Basis Function kernel, and Support Vector Machines with Linear kernel.



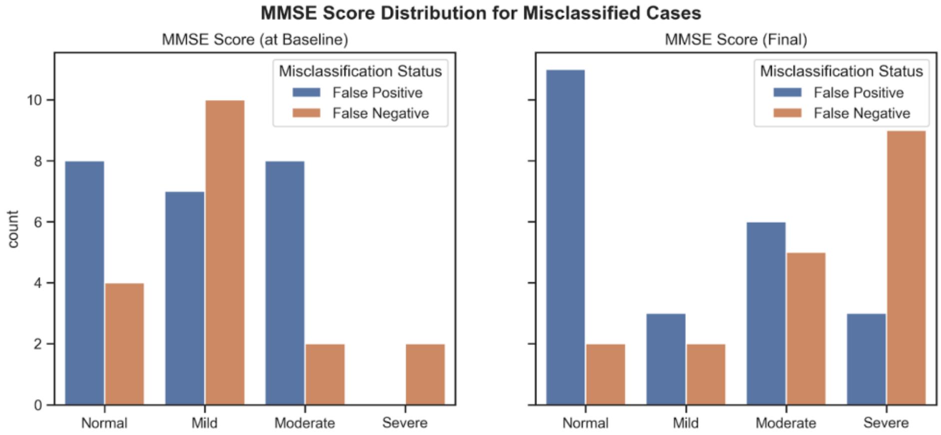
Supplementary Figure 5. Sensitivity performance of all subsets of neuropathology features from the rank list forward and backward rankings. Forward ranking (blue) adds to the classifier model from the top feature to the lowest feature while the backward ranking (orange) adds to the model from the lowest feature to the top feature. Seven classifiers were utilized in this investigation: Logistic Regression, Decision Tree, k-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, Support Vector Machines with Radial Basis Function kernel, and Support Vector Machines with Linear kernel.



Supplementary Figure 6. Specificity performance of all subsets of neuropathology features from the rank list forward and backward rankings. Forward ranking (blue) adds to the classifier model from the top feature to the lowest feature while the backward ranking (orange) adds to the model from the lowest feature to the top feature. Seven classifiers were utilized in this investigation: Logistic Regression, Decision Tree, k-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, Support Vector Machines with Radial Basis Function kernel, and Support Vector Machines with Linear kernel.

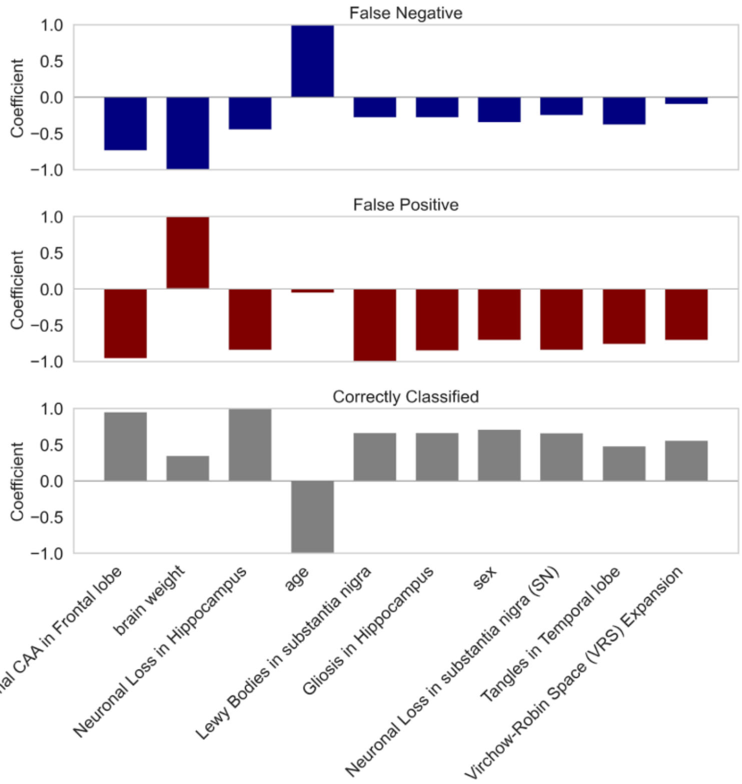


Supplementary Figure 7. Clustering of classification performance. Clustering of classification performance from leave one out cross-validation on 114 CFAS participants and top 22 ranked standard neuropathology features. Each cluster illustrates a classification that was given to individuals consistently or nearly consistently, irrespective of what classification algorithm was used. Evaluation of 7 classifiers revealed 24 individuals (blue) were mostly misclassified as a false positive, 22 individuals (red) were mostly misclassified as false negative, and 68 individuals (grey) were mostly correctly classified as true positive or true negative. Each algorithm evaluated subsets of ranked features from 1 (top feature) to 22 features (all ranked features).

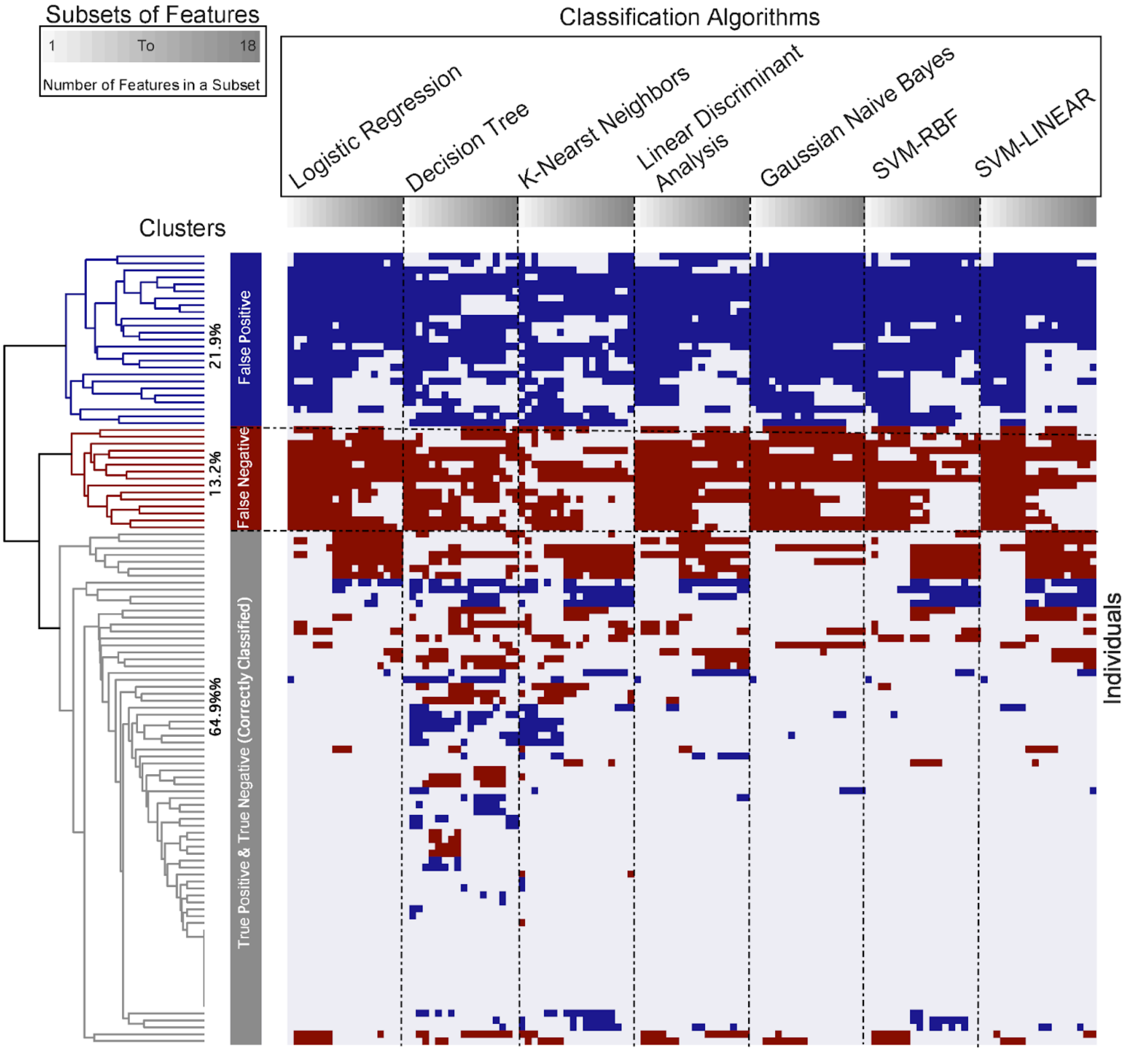


Normal(26-30), Mild(22-25), Moderate(18-21), Severe(0-17)

Supplementary Figure 8. Distribution of MMSE scores at baseline and final for all misclassified cases (False positive and False Negative).



Supplementary Figure 9. Non-standard neuropathological and demographic features that were associated with misclassified and correctly classified cases by the classic neuropathology features. The coefficients shown for each variable were extracted from the most predictive support vector machine classifiers: demographics features such as age, brain weight and sex.



Supplementary Figure 10. Clustering of cross-validation classification performance on 114 CFAS participants and subsets of 18 features, including eight top-ranked neuropathology features and ten non-standard neuropathology features. Each cluster illustrates a classification that was given to individuals consistently, or nearly consistently, irrespective of what classification algorithm was used. Evaluation of 7 classifiers revealed 23 individuals (blue) were mostly misclassified as false positive, 15 individuals (red) were mostly misclassified as false negative, and 76 individuals (grey) were mostly correctly classified as true positive or true negative.

Supplementary Table 1. Ranking of the CFAS Dataset Features using Different Feature Selection Techniques Seven feature-ranking methods were presented: Chi-Square (CHI), Gain Ratio (GR), Information Gain (IG), ReliefF (RF), Symmetrical Uncertainty (SmyUn), Least Loss (L2) and Variable Analysis (Va).

NO	Chi-Squares	Gain Ratio	Information Gain	ReliefF	Symmetrical Uncertainty	Least Loss	Variable Analysis
1	BraakStage	age	BraakStage	BrainNetStage	age	CAAType	BraakStage
2	BrainNetStage	CAATotalSev	BrainNetStage	BraakStage	BraakStage	brain weight	BrainNetStage
3	CAAMeningeal	AbStageTypical	CAAMeningeal	MTSPETSA	BrainNetStage	age	CAAMeningeal
4	CAAParenc	brain weight	CAAParenc	TSAAny	CAATotalSev	BrainNetStage	CAAParenc
5	age	BraakStage	age	SubpialMesTemp	brain weight	MTSPETSA	age
6	ThalStage	BrainNetStage	ThalStage	ThalStage	CAAType	BraakStage	ThalStage
7	CAAType	CAAAreas	CAAType	age	CAAMeningeal	CAAAreas	CAAType
8	brain weight	SubpialBrainstem	brain weight	CAAType	CAAParenc	CAAHippocampus	brain weight
9	SubpialBrainstem	CAAType	CAATotalSev	TSATotal	SubpialBrainstem	SubpialBrainstem	SubpialBrainstem
10	CAAAreas	MTSPETSA	SubpialBrainstem	SubpialBrainstem	CAAAreas	CAAParietal	CAATotalSev
11	CAATotalSev	BSSPETSA	CAAAreas	HippocTauStage	ThalStage	ThalStage	CAAAreas
12	MTSPETSA	CAAMeningeal	MTSPETSA	SubcorticalStage	MTSPETSA	CAAFrontal	MTSPETSA
13	CAAHippocampus	TempMicroinf	CAAHippocampus	brain weight	BSSPETSA	CAATemp	CAAHippocampus
14	CAAParietal	CAAParenc	CAAParietal	BSSPETSA	CAAHippocampus	CAAOccipital	CAAParietal
15	CAAFrontal	ThalStage	CAAFrontal	CAAOccipital	CAAParietal	TSAAny	CAAFrontal
16	CAAOccipital	CAAHippocampus	BSSPETSA	CAAParietal	CAAOccipital	CAACerebellum	CAAOccipital
17	HippocTauStage	CAAOccipital	CAAOccipital	PARTall	CAAFrontal	CAAMeningeal	HippocTauStage
18	CAATemp	CAAParietal	HippocTauStage	CAAMeningeal	CAATemp	SubpialMesTemp	BSSPETSA
19	BSSPETSA	CAAFrontal	CAATemp	CAAHippocampus	TSAAny	CAAParenc	CAATemp
20	TSAAny	CAATemp	TSAAny	MicroinfarctStage	CAACerebellum	CAATotalSev	TSAAny
21	CAACerebellum	TSAAny	CAACerebellum	CAACerebellum	AbStageTypical	HippocTauStage	CAACerebellum

22	SubpialMesTemp	FrontalMicroin	SubpialMesTemp	CAAParenc	SubpialMesTemp	BSSPETSA	SubpialMesTemp
23	AbStageTypical	CAACerebellum	AbStageTypical	CAATotalSev	HippocTauStage	FrontalMicroin	AbStageTypical
24	TempMicroinf	SubpialMesTemp	TempMicroinf	CAAAreas	TempMicroinf	TempMicroinf	TempMicroinf
25	FrontalMicroin	HippocTauStage	FrontalMicroin	CxSPETSA	FrontalMicroin	PARTdefinite	FrontalMicroin
26	PARTdefinite	ArgyrGrains	PARTdefinite	PARTdefinite	ParMicrin	AbStageTypical	PARTdefinite
27	ParMicrin	ParMicrin	ParMicrin	ArgyrGrains	PARTdefinite	ParMicrin	ParMicrin
28	ArgyrGrains	PARTdefinite	ArgyrGrains	TempMicroinf	ArgyrGrains	OccipMicroing	ArgyrGrains
29	OccipMicroing	CxSPETSA	OccipMicroing	CorticalStage	OccipMicroing	ArgyrGrains	OccipMicroing
30	CxSPETSA	OccipMicroing	CxSPETSA	AbStageTypical	CxSPETSA	PARTall	CxSPETSA
31	TuftedAst	TuftedAst	TuftedAst	FrontalMicroin	TuftedAst	TuftedAst	TuftedAst
32	PARTall	PARTall	PARTall	OccipMicroing	PARTall	CxSPETSA	PARTall
33	MicroinfarctStage	CorticalStage	SubcorticalStage	CAATemp	SubcorticalStage	CorticalStage	MicroinfarctStae
34	TSATotal	TSATotal	MicroinfarctStage	ParMicrin	CorticalStage	SubcorticalStage	TSATotal
35	SubcorticalStage	SubcorticalStage	TSATotal	CAAFrontal	TSATotal	TSATotal	SubcorticalStage
36	CorticalStage	MicroinfarctStage	CorticalStage	TuftedAst	MicroinfarctStage	MicroinfarctStage	CorticalStage

Supplementary Table 2. T-test and p-values for all non-standard and demographic features. The result shows that there statistically significant differences in the values of non-standard features between false positives and false negatives

No	Feature	T test	P-value
1	Age	3.132	0.00
2	Brain weight	-3.741	0.001
3	Virchow-Robin Space (VRS) Expansion	0.607	0.547
4	Gender	-0.842	0.404
5	Lewy Bodies in Substantia Nigra	0.328	0.744
6	Neuronal Loss in Substantia Nigra	0.478	0.635
7	Neuronal Loss in Hippocampus	0.541	0.591
8	Tangles in Temporal Lobe	-1.046	0.301
9	Parenchymal CAA in Frontal Lobe	-1.734	0.090
10	Gliosis in Hippocampus	0.644	0.523

Chapter 4 - Feature-feature Correlations Biases Ranking of Dementia Features in Machine Learning Studies

4.1. Background

The aim of this study is to identify the effective filter methods for detecting dementia and reducing the negative impacts of the disease. The study assesses two real datasets related to dementia conditions, namely, CFAS and ADNI. We investigated the correlation between dementia conditions' levels, features, and filter methods, focusing on the associations between feature ranking obtained by filter methods and correlations among the features themselves. The research identified which filter methods were less sensitive to similarities among the neuropathological features and the impact of varying features' rankings between different data cohorts on dementia conditions' prediction models.

We investigated filter methods sensitivity to feature-feature correlation in dementia conditions' diagnosis. We applied seven filter methods to two cohorts of aging individuals, ADNI and CFAS, then applied Kendall's tau to detect the agreement between results. We developed a multiple regression classification model to mathematically model the association between the results of the classification model and feature ranking for each filter method. We evaluated the system with classification algorithms to indicate if the diagnosis models developed were competitive.

4.2. Contribution

The following version of the accepted manuscript has been submitted to the *Journal of GigaScience*. This manuscript details my primary contribution as the lead author in the data analysis process, manuscript drafting writing. My supervisor, Dennis Wang, provided exceptional guidance and unwavering support throughout the entire process, editing, analysing and manuscript drafts. In particular, I was the main contributor in producing the initial draft of the manuscript, while also taking responsibility for the generation of code, figures, and tables used in the study. Additionally, Teruka Taketa's contribution to this paper involved referencing a common features and features dictionary, which she developed during her Master's degree.

4.3. Manuscript 3

Feature-feature Correlation Biases Ranking of Dementia Features in Machine Learning Studies

Mohammed D Rajab[1,2], Teruka Taketa[1], Dennis Wang[1,2,5,6], on behalf of the Cognitive Function and Ageing Neuropathology Study Group, and for the Alzheimer's Disease Neuroimaging Initiative*

1. Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield S10 2HQ, UK.
2. Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK
3. Cambridge Public Health, Cambridge CB2 1PZ, UK
4. Population Health Sciences Institute; Newcastle University, Newcastle upon Tyne NE4 5PL, UK
5. Singapore Institute Clinical Sciences, A*STAR, Singapore, 117609, Singapore
6. National Heart and Lung Institute, Imperial College London, London, SW3 6LY, UK

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Correspondence: dennis.wang@imperial.ac.uk

Abstract

Background

The prevalence of dementia is increasing globally. Due to the significant resources required to treat the condition, governments and private healthcare systems are experiencing pressures. Early diagnosis of dementia diseases, such as Alzheimer's disease, is difficult because of the time and resources needed to perform neuropsychological and pathological assessments. Given the increasing use of machine learning methods to evaluate neuropathology features in the brains of dementia patients, it is important to investigate how the selection of features may be impacted and which features are most important for the classification of dementia.

Results

The study investigated feature ranking, feature-feature correlation, multiple regression, and classification in two real dementia datasets, the Cognitive Function and Aging Studies (CFAS) and Alzheimer's Disease Neuroimaging Initiative (ADNI). The ReliefF filter method was the most biased by feature-feature correlation but its ranking was also most consistent in both ADNI and CFAS. The least-loss and gain ratios filter methods were the least impacted by feature-feature correlations in both datasets. A Random Forest classifier achieved high performance in classifying dementia status with 94.4% accuracy, 90.0% sensitivity, and 98.4% specificity in ADNI while in CFAS Naive Bayes performed best at 70.6% accuracy, 81.3% sensitivity, and 54.3% specificity.

Conclusions

The study showed that for selecting relevant neuropathology features related to dementia, feature-feature correlation impacts feature rankings obtained by filter methods and can vary between cohort studies. By examining bias in filter methods, we can reduce discrepancies in feature ranking and identify a minimal set of features for accurate classification of dementia.

Keywords— Dementia, Alzheimer’s, Feature Selection, Machine Learning, Neuropathology, Feature-feature Correlation

Introduction

Dementia poses a significant global challenge, affecting the lives of individuals, their families, and caregivers [1]. The economic burden of dementia was estimated to exceed $818 billion in 2015, and the number of people living with dementia is expected to surpass 75 million by 2030 [2]. Early diagnosis and intervention are crucial in mitigating the negative impact of dementia [3]. However, identifying the determinants of dementia can be difficult due to its complex spectrum of characteristics, encompassing various disorders with distinct pathologies. Alzheimer's disease (AD) is the most common form of dementia, characterised by the presence of amyloid plaques and neurofibrillary tangles in the brain [4].

Feature selection methods play a vital role in biomedical data analysis, helping to identify the most relevant features contributing to a health outcome while eliminating noise, redundancy, and irrelevant factors [5–7]. Biomedical datasets collected from human biosamples often contain many features, some of which may be irrelevant to the outcome of interest. Analysing all features can lead to overfitting, reduced accuracy, and a less concise understanding of the underlying biological processes [5,8,9]. Filter methods measure the relevance of features based on their correlation with the outcome. While commonly used to select features for downstream analysis or machine learning of biomedical datasets, there lacks a systematic comparison of filter methods when used to study complex human disorders, such as dementia.

Previous studies have employed filter methods to identify features related to AD [10–12]. Gómez-Ramírez et al. (2020) focused on self-reported data, investigating demographics and other relevant factors associated with the development of dementia from Mild Cognitive

Impairment (MCI). Permutation-based methods were employed as a filter to identify significant cognitive decline features. Subsequently, the Random Forest algorithm was applied to identify features strongly correlated with cognitive impairment [13]. Thabtah et al. (2022) conducted a comprehensive analysis of feature selection methods using continuous features derived from MRI images to detect dementia. The study compared popular methods such as mutual information gain [14], Pearson correlation [15], and Symmetrical Uncertainty [16]. Univariate feature selection and recursive feature elimination techniques were also employed to identify the most informative features correlating with AD using the Functional Activities Questionnaire (FAQ), a common neuropsychological assessment [17,18]. The authors further investigated the relationships between cognitive and functional features across different levels of dementia progression.

While the relationship between cognitive function and neuropathology features has been extensively explored, less attention has been given to how feature correlation might impact machine learning of dementia. Given the diversity of filter methods and features, it is essential to identify the methods that are less sensitive to similarities or differences between neuropathological features in order to minimise discrepancies in feature rankings. To address these issues, we focused on data from two large dementia studies in the UK and USA, and hypothesised that there would be associations between feature-feature correlations and the ranking scores computed by filter methods. Several questions arise regarding the ranking of neuropathology features: 1) Which filter methods are less sensitive to feature-feature correlations? 2) Are there differences in feature-feature correlations and rankings between the separate dementia cohorts? 3) How do variations in feature rankings between the cohorts impact dementia prediction models? To investigate these questions, we applied seven filter methods to the two cohort datasets [19–21] to generate feature rankings and observed how they varied depending on the degree of similarity between features. We are able to identify the best performing feature selection techniques for neuropathology data and assess the level of reproducibility in the associations with dementia found in the two studies.

Material and Method

Overview of Feature Ranking Analysis

We examined the correlation structure of neuropathology and its relationship with dementia (as depicted in **Figure 1)**. Following a comprehensive review and subsequent ethics approval from the management committees, we downloaded the pathological assessments from the Cognitive Function and Ageing Study (CFAS) [21] and the Alzheimer's Disease Neuroimaging Initiative (ADNI)[22]. After conducting pre-processing on both datasets, we pinpointed features that were present in both, ensuring their compatibility in terms of features and data types whenever possible. In both datasets, neuropathological features were evaluated and ranked by utilising a range of feature selection techniques centred around various filter methods. We then gauged the ranking disparities between the neuropathological features of CFAS and ADNI, in addition to the consistency between both datasets for each filter method applied. To discover the relationship between a given feature and the remaining ones, we delved into feature-feature correlations using the R^2 metric, which is based on multiple regression analyses. We considered both the correlation among the features and their rankings, which was achieved by implementing classification algorithms and noting accuracy, sensitivity, and specificity values. From these insights, we inferred that certain feature subsets can be classified as dementia.

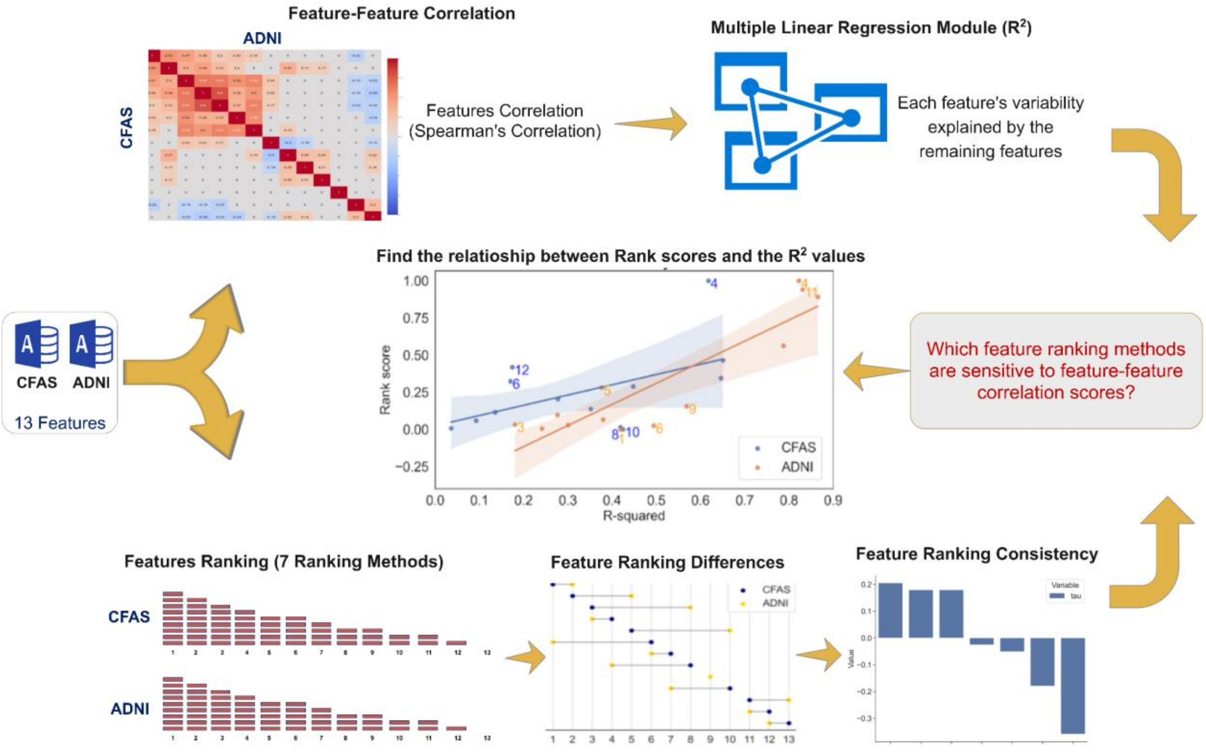

Feature-Feature Correlation
ADNI
CFAS
Features Correlation
(Spearman's Correlation)
Multiple Linear Regression Module (R²)
Each feature's variability
explained by the
remaining features
Find the relatioship between Rank scores and the R² values
Rank score
R-squared
1.00
0.75
0.50
0.25
0.00
−0.25
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CFAS
ADNI
Which feature ranking methods
are sensitive to feature-feature
correlation scores?
CFAS
ADNI
13 Features
Features Ranking (7 Ranking Methods)
CFAS
ADNI
Feature Ranking Differences
CFAS
ADNI
1 2 3 4 5 6 7 8 9 10 11 12 13
Feature Ranking Consistency
Variable
tau
Value
0.2
0.1
0.0
−0.1
−0.2
−0.3

Figure 1: Methodology for Dementia Classification using CFAS and ADNI Datasets. The dementia classification methodology was developed and executed in three key stages: design, implementation, and evaluation. After acquiring neuropathology data, we carried out preprocessing and determined the correlation between different features. Utilising seven filter methods, we ranked all neuropathological features. Subsequently, we explored the connection between feature-feature correlation and feature ranking across all applied filter methods. Thereafter, classifiers were evaluated using various feature subsets, depending on their interrelations.

CFAS Cohort

This study considered the donated brains of 186 participants, and 13 neuropathological features were assessed (**Table 1**). These features constituted fundamental neuropathological assessments for each participant, including Braak neurofibrillary tangle (NFT) stage, Thal phase, and cerebral amyloid angiopathy (CAA). Of the total participants, 107 (equivalent to 58%) had been diagnosed with dementia. The participant pool consisted of 72 women and 35 men, with respective median ages of 89 and 88. Among those participants who passed away without a dementia diagnosis (with median ages of 85 for females and 79 for males), the gender distribution was evenly balanced with 37 females and 33 males [23].

ADNI Cohort

The data utilised for the creation of this article were acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), which was established in 2003 as a collaboration between public and private entities. The neuropathology data version utilised in the ADNI database was NEUROPATH_07_06_21. The ADNI cohort consisted of 1,736 individuals, including 85 clinical features across ADNI-1, ADNI-GO, and ADNI-2. For this study, we specifically focused on 80 post-mortem brains, which exhibited 13 neuropathological features, as detailed in **Table 1**. These features involved fundamental measures of neuropathology for each subject, such as Braak neurofibrillary tangle (NFT) stage, Thal phase, and cerebral amyloid angiopathy (CAA). Within the cohort, 77.5% of participants (62 out of 80) were diagnosed with dementia, while 12.5% had mild cognitive impairment (MCI), and 10% were cognitively normal (CN). Among the 62 dementia cases, 16 were women and 46 were men, with median ages of 79 and 81.5, respectively. The MCI participants exhibited a gender ratio of 1 female to 9 males (with a median age of 85 for both genders), while those who passed away without a dementia diagnosis had a gender ratio of 5 females to 3 males (with median age of 84 for females and 79 for males). To ensure consistency with the CFAS dataset comparisons, we excluded the 10 participants diagnosed with MCI, leaving us with 70 participants diagnosed with dementia or CN for this study [24–26].

Table 1: Neuropathology features from the CFAS and ADNI cohorts considered for feature selection.

No	Feature	Feature Description	CFAS				ADNI			
			Type	Dementia (n=107)	No Dementia (n=70)	Missing (n=9)	Type	Dementia (n=62)	No Dementia (n=8)	Missing (n=0)
1	Braak stage	Braak neurofibrillary tangle (NFT) stage[26,27].	Nominal	107	70	0	Nominal	62	8	0
2	Thal phase	Detects immunopositive amyloids in cortical and subcortical areas[28,29].	Nominal	107	70	0	Nominal	62	8	0

3	Cortical atrophy	A condition in which the brain’s cortex—the outer layer of the cerebrum—thins and shrinks in size.	Binary	103	67	7(4.0%)	Nominal	57	8	5(7.1%)
4	Hippocampus atrophy	Characterised by a decrease in the size of the hippocampus, the area of the brain responsible for the formation and recall of memories.	Nominal	81	39	57(32.2%)	Nominal	57	8	5(7.1%)
5	Atherosclerosis	A condition that is characterised by the hardening and narrowing of arteries due to a buildup of fatty deposits known as plaque.	Nominal	98	65	14(7.9%)	Nominal	54	7	9(12.9%)
6	haemorrhage	A medical condition in which there is a loss of blood.	Binary	47	27	103(58.2%)	Binary	62	8	0
7	Neocortical neuritic plaques	Accumulation of amyloid beta peptides in the brain in the form of neuritic plaques, comprising dense deposits of amyloid beta protein.	Nominal	88	56	33(18.6%)	Nominal	62	8	0
8	Neuronal loss in substantia nigra	Characterised by the death of neurons in the substantia nigra, a part of the brain associated with movement and coordination.	Nominal	105	68	4(2.3%)	Nominal	61	8	1(1.3%)
9	Argyrophilic grains disease	A type of tauopathy which is a class of neurodegenerative diseases caused by an accumulation of the tau protein in the brain.	Binary	107	69	1(0.6%)	Binary	27	7	36(51.4%)
10	Cerebral amyloid angiopathy (CAA)	A form of cerebrovascular disease in which amyloid protein deposits accumulate in the walls of small blood vessels in the brain.	Numeric	84	42	51(28.8%)	Nominal	62	8	0
11	Infarcts and lacunes	Types of brain lesions that are commonly associated with a stroke. Infarcts are areas of tissue death caused by a lack of oxygen due to a blockage of the brain's blood vessels. Lacunes are small cavities that develop when parts of the brain become damaged or die due to a lack of oxygen or other factors. Often caused by small strokes or other vascular changes.	Binary	51	29	97(54.8%)	Binary	62	8	0
12	Arteriolar sclerosis	A condition in which the walls of the arterioles become stiff and thickened	Nominal	105	69	3(1.7%)	Nominal	62	8	0

		due to deposits of fatty material.								
13	Diffuse plaques	Caused by deposits of amyloid-beta proteins accumulating in the brain. These proteins form clumps that disrupt normal cell functioning, leading to inflammation and damage.	Binary	107	70	0	Nominal	62	8	0
14	Diagnostic	Class label (dementia or no dementia) status of a patient	Binary	107	69	0	Binary	62	8	0

Feature Pre-processing

In the ADNI dataset, certain features were represented as individual columns, whereas in CFAS, multiple columns were employed to capture related attributes such as 'infarcts and lacunae' and 'diffuse plaques'. Furthermore, CFAS distinguished between infarcts and lacunae separately, whereas ADNI combined them. To address this disparity, our study unified infarcts and lacunae into a single category, treating them as binary indicators of pathology within CFAS. We encountered a similar challenge with the diffuse plaques feature in CFAS, where columns were grouped based on their presence or absence. We refrained from encoding feature measures in both datasets to ensure unbiased feature ranking and analysis. For instance, CFAS utilised a binary representation for cortical atrophy, while ADNI employed an ordinal scale consisting of four categories: no atrophy, mild, moderate, or severe. Upon encoding cortical atrophy as a binary feature, we discovered a 100% correlation between this attribute and certain other features. A similar issue was encountered with ADNI's diffuse plaques feature. Following the preprocessing stage, we analysed a total of 177 post-mortems from CFAS and 70 post-mortems from ADNI, examining 13 neuropathology features for feature ranking.

Ranking Neuropathology Features

To gain preliminary insight and highlight influential neuropathological features of dementia, we used a variety of feature selection filter methods to measure each feature’s relevance. Chi-square (CHI) [31], gain ratio [32], information gain (IG) [14], reliefF [33,34], symmetrical uncertainty [16], least loss [35], and variable analysis [36,37] were included in the analysis. Scores varied according to the mathematical criteria and type of filter method used. Due to the different models, there may be discrepancies in the ranking of features based on such scores [37,38]. Details of the mathematical formulation of the considered filter methods follow:

CHI utilises the difference between observed and expected frequencies of the instances as shown in Equation (1).

$$X^2 = \frac{(O-E)^2}{E} \quad (1)$$

where *O* and *E* are the Observed and Expected frequencies for a specific feature, respectively. IG employs Shannon entropy to measure the correlation between a feature and dementia status (Equations 2 and 3).

$$IG\ (S, A)\ =\ Entropy\ (S)\ - \sum \quad ((\mid S_v \mid\ \div \mid S \mid) \times\ Entropy\ (S_v)) \quad (2)$$

where Entropy (*T*) =$- \sum \quad P_c P_c$ (3)

P is the probability that *S* belongs to class label *c*. S_v is the subset of *S* for which *a* feature has value *v*. $|S_v|$ is the number of data instances in S_v, and $|S|$ is the size of *S*.

Gain ratio is a normalised form of IG which is estimated by dividing the IG with the entropy of the feature with respect to the class (Equations 4 and 5).

$$\text{Gain ratio} = \frac{IG}{ENT(S,F)} \quad (4)$$

$$ENT(S, F) = - \sum \quad \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (5)$$

where IG denotes the information gain and *ENT* is the Entropy of feature F over a set of examples S.

Symmetrical uncertainty deals with the bias of IG that occurs due to a large number of distinct values for the feature and presents a normalised score (Equation 6).

$$SU(A,B) = \frac{2 \times IG(A|B)}{E(A) + E(B)} \tag{6}$$

where $IG(A|B)$ denotes the information gain of A after knowing the class. E(A) and E(B) are the entropy values of A and B, respectively.

ReliefF calculates the scores of each available feature with the class using the differences between the neighbouring data instances and the target instances (Equation 7).

$$\mathrm{W[A]} = \mathrm{W[A]} - \frac{\left(diff\frac{A,R_i,H}{m}\right)}{\left(diff\frac{A,R_i,M}{m}\right)} \tag{7}$$

where, W[A] are the feature weights, A is the number of features, m is the number of random training data instances out of ‘n’ number of training data instances used to amend W.

R_i = A random chosen test instance and H/M is nearest hit and nearest miss

Least loss is computed per feature based on the simplified expected and observed frequencies of the features (Equation 8), and variable analysis employs a vector of scores of both CHI and IG results, normalises the scores, and then computes the vector magnitude (V_score) (See Equations 9 and 10).

$$L^2(Y,X) = \sum_{i,j} \quad [P(Y_{i,}X_j) - P(Y_i)P(X_j)]^2 \tag{8}$$

where X is the independent feature class, Y is class label, $P(Y_i)$ is the theoretical marginal distribution of Y, and $P(X_j)$ is the theoretical marginal distribution of X. $P(Y_i\, X_j)$ is the theoretical joint probability distributions of X and Y.

$$V_a = \left(\frac{IG_x}{CST_x}\right) \tag{9}$$

$$|V_a| = \sqrt{(IG)^2 + (TST)^2} \tag{10}$$

where V_a is the square root of the sum of the square of its CHI and IG results of a feature.

The V_score and the Correlation Feature Set results [39] are then integrated to represent a new measure of goodness by which to select relevant features.

$$IG\ (S, A)\ =\ Entropy\ (S)\ - \sum \quad ((\ |\ S_v\ |\ \div |\ S\ |)\ \times\ Entropy\ (S_v)) \tag{2}$$

The experiment-related filter-based feature selection was conducted using Waikato Environment for Knowledge Analysis (WEKA version 3.9.1) [40]. The percentage contribution of each feature was calculated by averaging the total weights assigned by all filter methods to each feature after normalising the weight scores.

Measuring Filter Methods Consistency

Kendall's tau, a measure of correlation between two ranking lists, provides insights into the level of agreement or disagreement between them. Values closer to 1 indicate a stronger agreement, while values closer to -1 indicate a stronger disagreement. A value of tau = 0 suggests no association between the ranking lists. To compare the feature rankings between CFAS and ADNI datasets for each filter method, we utilised the kendalltau() function from the Python3 machine learning package (scipy.stats version 1.7.3). Specifically, we employed this function, available in version v1.9.3, to assess the correlation between the CFAS and ADNI cohorts. The comparison involved seven filter methods and 13 distinct features.

Imputing Missing Values

Due to the limitations of the considered cohorts (ADNI = 70 samples, CFAS = 177 samples) and the tendency of machine learning models to encounter errors when encountering NaN values, addressing missing values became necessary. To handle this, we adopted an iterative imputer approach utilising the Scikit-learn version 0.22.2.post1 [41] library in Python3. This approach allowed us to impute missing values for both numerical and categorical features. For numerical and categorical values, we employed the IterativeImputer from the sklearn.impute package to perform the imputation transformation. To replace missing numeric values, we utilised the RandomForestRegressor from the sklearn.ensemble [42] package as an estimator. The missing values were initially initialised with the mean and underwent a maximum of five iterations. Similarly, for categorical values, we constructed a model employing the RandomForestClassifier from the sklearn.ensemble package. The missing values were initialised with the mean, and the imputation process followed a maximum of five iterations. All the machine learning models and feature selection libraries utilised in this study were developed using Python 3.7.3, ensuring consistency across the analysis.

Measuring Feature-Feature Correlation

In our analysis of CFAS, a total of 177 subjects were included. However, nine subjects had to be excluded from the analysis due to missing values in the class label. Regarding the ADNI dataset, individuals with mild cognitive impairment (MCI) were excluded, leaving us with a cohort of 70 out of 80 participants who were classified as either cognitively normal (CN) or diagnosed with dementia. To investigate the relationship between each feature, treated as a dependent variable, and the remaining features, considered as independent variables, we utilised multiple linear regression models. These models were implemented on both the ADNI and CFAS neuropathology cohorts. The coefficients R^2 obtained from the models (specifically Equations 11–13) were used to describe the relationships between the features. To ensure consistency in the analysis, we applied feature normalisation to the numerical features. This was achieved using the minmaxScaler package from scikit-learn version 0.22.2.post1 [41]. For the linear regression models, we employed the ordinary least squares method with the statsmodels.formula.api package version 0.13.2.

The determination of coefficient (R^2) represents the similarity of the dependent feature with the independent features by showing to which level the remaining features can explain the variability of the feature at hand.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{11}$$

where RSS is the sum of squares residuals (Equation 12) and TSS is the total sum of squares (Equation 13).

$$RSS = \sum_{i=1}^{n} \quad (y_i - f(x_i))^2 \tag{12}$$

where n is the upper limit of summation, y_i is the i^{th} value of the feature to be predicted, and $f(x_i)$ is the predicted value of y_i.

$$TSS = \sum_{i=1}^{n} \quad (y_i - \underline{y})^2 \tag{13}$$

where n is the number of the observation, y_i is the the value in a sample, and $\underline{y}$ is the mean value of a sample.

Evaluation of Feature Ranks Against Feature-feature Correlation

To normalise the scores of the CFAS and ADNI features, we utilised the minmaxScaler package from scikit-learn version 0.22.2.post1 [41]. This scaling process ensured that the feature scores were within the range of [0, 1]. Consequently, we performed linear regression analyses to examine the relationship between the feature scores and their corresponding R^2 values for each filter method. For data visualisation and fitting linear regression models, we employed the regplot() function from the Seaborn package version 0.11.0 [43]. The function creates a scatterplot with a linear regression model fit. It employs the Least Squares method to estimate the linear regression coefficients, minimising the sum of the squares of the differences between the observed and predicted values. The function also computes and plots a 95%

confidence interval for the regression line, which estimates the uncertainty around the line of best fit. This interval is calculated using bootstrapping with 1000 iterations by default, a resampling method that generates an empirical representation of the sampling distribution and quantifies the uncertainty of the estimate. This allowed us to plot the data and visualise the linear relationship. To calculate the correlation coefficients and corresponding p-values, we utilised the pearsonr() function from the SciPy.stats package version 1.7.3 [44]. This statistical analysis provided valuable insights into the strength and significance of the correlations between the variables.

Dementia Classification

In the CFAS dataset, a total of 177 subjects were initially included. However, nine subjects had to be excluded due to missing values in the class label. For the ADNI dataset, individuals with mild cognitive impairment (MCI) were removed, and the remaining participants with cognitive impairment or dementia were retained. Given the imbalance in the class label of the ADNI dataset, where there were 62 instances of 'Dementia' and only 8 of 'No Dementia,' we utilised the Synthetic Minority Oversampling Techniques for Numerical and Categorical Features (SMOTE-NC) [45,46]. This method was applied using the imbalanced-learn toolbox, version 0.9.1 [47]. This technique involved generating synthetic data instances for the minority class label using the k-Nearest Neighbours classification algorithm with k=5. After balancing the ADNI dataset, we were left with 124 samples and 13 features. To train and evaluate the classifiers, we utilised scikit-learn version 0.22.2.post1 in Python3. The evaluation was performed using the "leave-one-out" cross-validation approach, ensuring robustness in the analysis.

For further assessment of the neuropathological features, we employed various supervised learning techniques, primarily Random Forest (RF) [13] and Gaussian Naive Bayes

(GNB) [48]. The default parameter settings were used for both RF and GNB. Specifically, RF was configured with 100 estimators (the number of trees in the forest), and the quality of the split was measured using the Gini impurity function. The minimum number of samples required to split a node (min_samples_split) and the minimum number of samples required to be a leaf node (min_samples_leaf) were both set to 1.

Evaluation of Classification Performance

In this study, we approached the prediction of dementia as a binary classification problem, with the two classes being "Dementia" and "No dementia." To assess the performance of the feature subsets, we employed evaluation metrics such as accuracy, sensitivity, and specificity. These metrics provided valuable insights into the effectiveness of the selected features in predicting dementia. The following evaluation metrics were utilised for performance assessment:

- True positives (TP): Number of dementia cases that were correctly classified
- False positives (FP): Number of healthy subjects incorrectly classified as dementia cases
- True negatives (TN): Number of healthy subjects correctly classified
- False negatives (FN): Number of dementia cases incorrectly classified as healthy subjects
- Accuracy (%): The proportion of correct classifications among total classifications:

  $$Accuracy = \frac{TP+TN}{n} \quad (14)$$

  where *n* is the number of total classifications per test
- Sensitivity (%): The proportion of dementia cases correctly classified

  $$Sensitivity = \frac{TP}{TP+FN} \quad (15)$$
- Specificity (%): The proportion of healthy subjects correctly classified

$$Specificity = \frac{TN}{TN+FP} \tag{16}$$

Results

Distribution of Neuropathology Feature Scores Across Dementia Cases

Examining the distribution of neuropathology feature scores among dementia cases was crucial for gaining deeper insights into these features. We conducted an analysis to detect any dissimilarities in the feature distributions between cohorts and to provide plausible explanations for these variations. For this purpose, we plotted the distributions of all neuropathology features for the CFAS and ADNI cohorts, comprising 186 and 70 individuals, respectively (**Figure 2**). An interesting observation was made regarding the ADNI dataset diffuse plaques feature, which posed a similar challenge as the infarcts and lacunae feature in CFAS. In both cases, we had to group the columns corresponding to diffuse plaques based on their presence or absence. Our findings revealed notable differences in the distributions of certain features between the two cohorts. For instance, cortical atrophy, represented as a binary feature in CFAS and ordinally in ADNI, exhibited distinct distributions. Additionally, other features such as atherosclerosis, neocortical neuritic plaques, neuronal loss in the substantia nigra, argyrophilic grains disease, and diffuse plaques demonstrated different distribution patterns. We attribute these differences in feature distributions to a combination of factors, including the varying number of cases in CFAS (n=186) and ADNI (n=70) and the contrasting class distributions within the datasets. Notably, the CFAS dataset displayed a relatively balanced distribution, with 60.5% classified as dementia and 39.5% as non-dementia, as shown in **Table 1**. On the other hand, the ADNI dataset exhibited an imbalanced distribution, with 88.6% classified as dementia and 11.4% as non-dementia. These findings underscore the importance of considering the dataset characteristics, including sample sizes and class distributions when interpreting and comparing the distributions of neuropathology features across cohorts.

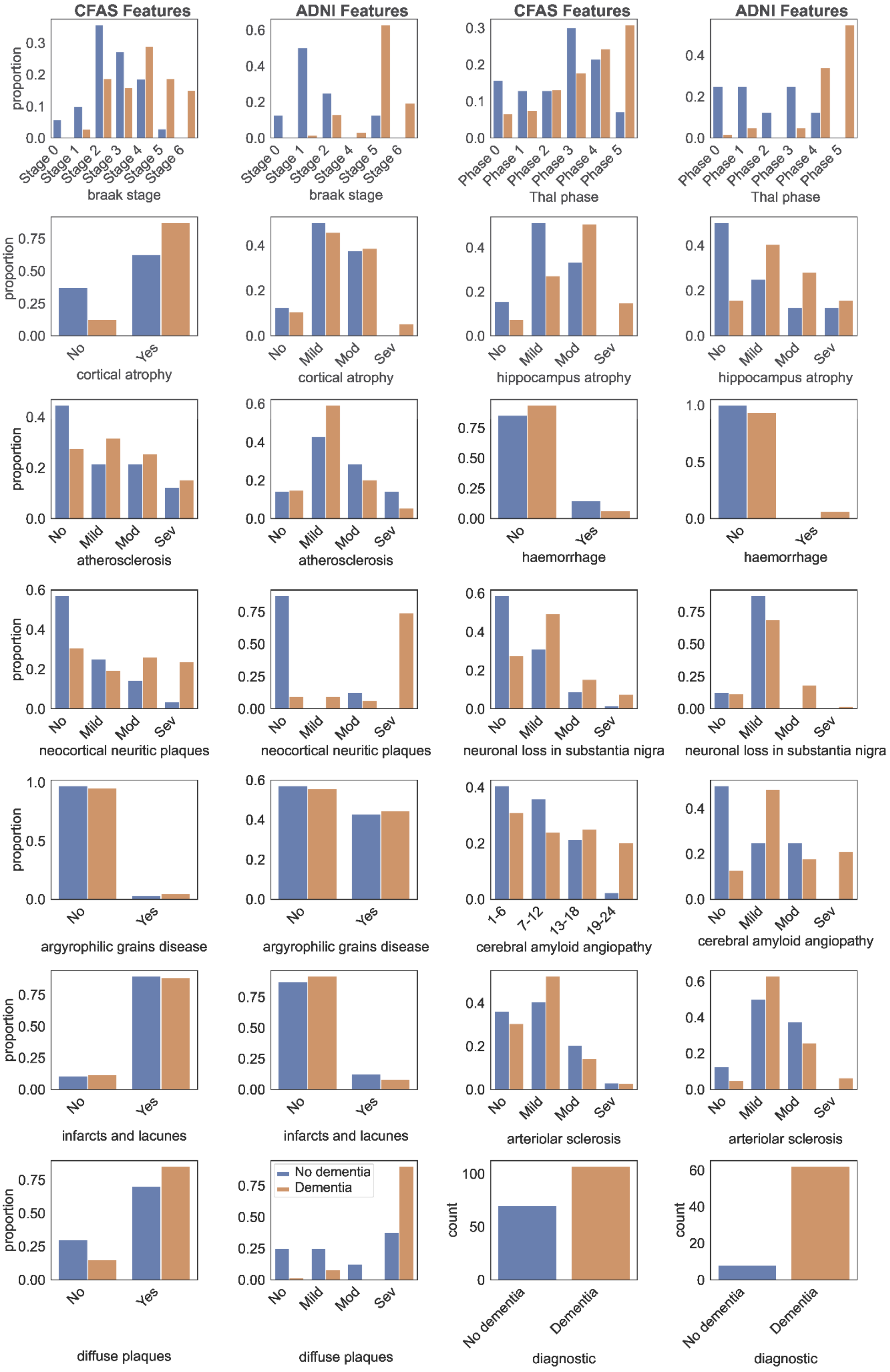


Figure 2: Distribution of Neuropathology Features in CFAS and ADNI Datasets. The distribution of individuals with and without dementia was examined in both the CFAS and ADNI neuropathology datasets. The features presented in the table were arranged based on their ranking in the features list, moving from left to right. It is important to note that all features, except for cerebral amyloid angiopathy, were categorical in nature. In CFAS, cerebral amyloid angiopathy was the only feature that had numeric values.

Ranking of Neuropathology Features

To examine the utility of filter methods on dementia-related features, we conducted a feature selection analysis using two neuropathological datasets, CFAS and ADNI. We aimed to rank the features consistently across both datasets and derive valuable insights for improving dementia diagnosis and treatment. To achieve unbiased and comprehensive results, we employed multiple filter methods to assess the sets of neuropathological features in each dataset. By applying these methods, we calculated feature scores based on the models generated for each filter method (**Figure 3**). The ranking of features in descending order based on their scores provided a comprehensive and cross-dataset comparison, aiding the medical profession in better understanding dementia pathology.

The consistent findings across both datasets revealed the significance of certain features in contributing to dementia. The Braak stage emerged as the most influential pathological feature, demonstrating strong correlations, particularly in the CFAS dataset (**Figure 3A**). In the ADNI dataset, other features such as neocortical neuritic plaques, Thal phase, diffuse plaques, and cerebral amyloid angiopathy were also highly correlated with dementia (**Figure 3B**). Notably, these results were consistent with those obtained from the CFAS dataset, which identified the Braak stage, Thal phase, and cerebral amyloid angiopathy as relevant factors associated with dementia. To further investigate the consistency of feature ranking across the filter methods in both CFAS and ADNI datasets, we conducted a detailed analysis. The overall outcomes of our study provided crucial insights into the neuropathological features that play a role in dementia. Furthermore, employing multiple filter methods ensures generalizability and reduces the risk of biased outcomes, emphasising the importance of considering diverse approaches in feature ranking.

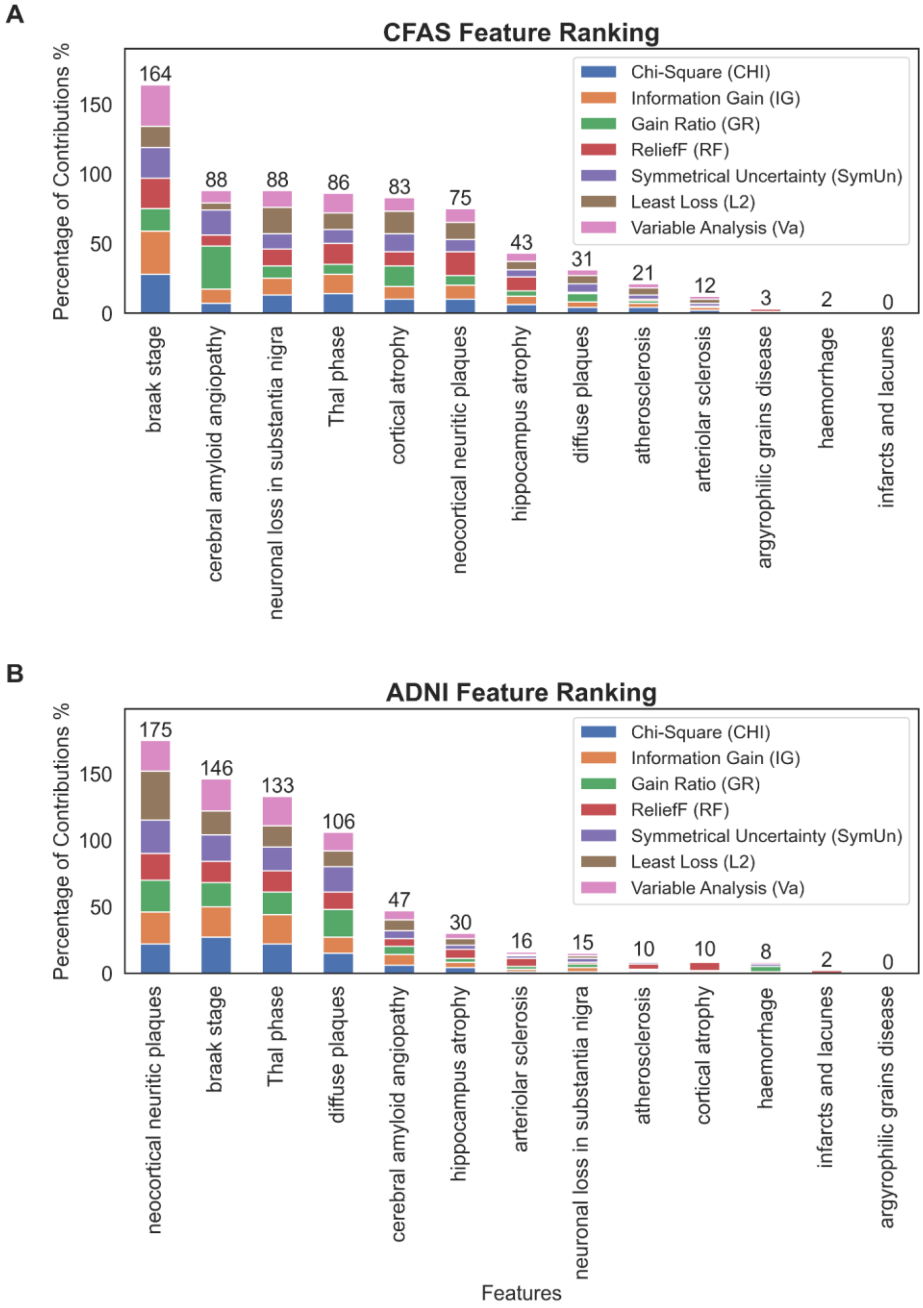


Figure 3: Ranking of neuropathology features in order of association to dementia status as estimated by filter methods in (A) CFAS and (B) ADNI. The cumulative contributions of 13 neuropathology features to dementia status in the CFAS and ADNI datasets are estimated by seven filter methods. The weight scores of each feature were normalised, and the percentage contribution of each feature was calculated by averaging the total weights assigned to it by the filter methods.

Consistency in the Ranking of Features Between Studies

In this study, pathological feature rankings from the CFAS and ADNI datasets were assessed using various filter methods to identify quantitative discrepancies between these rankings. We compared the positional differences in feature rankings between the two cohort datasets (**Figure 4A**).

Both datasets consistently positioned the top two features (Braak stage and Thal phase) and the ninth feature (atherosclerosis) identically. However, significant statistical divergence was observed for other feature rankings. For instance, neuronal loss in the substantia nigra and cortical atrophy occupied the third and fifth positions in CFAS, yet these were ranked four and six positions higher in ADNI. In contrast, diffuse plaques and arteriolar sclerosis, ranked eighth and tenth in CFAS, were higher in ADNI, sitting at fourth and eighth positions.

Evaluation of the filter methods, including information gain, reliefF, symmetrical uncertainty, and least loss, revealed considerable variations in feature ranking. Of note, the most prominent discrepancies were identified using the least loss filter method, with the top two features in CFAS and ADNI descending six and nine positions, respectively.

To summarise, there is considerable variability in pathological feature rankings in CFAS and ADNI datasets across the examined filter methods. However, certain features, like the Braak stage, demonstrated consistent patterns irrespective of the filter method employed. The discrepancy in ranking positions might result from differing models used by the filter methods to compute feature-to-class correlations. Despite normalising average scores to a unified scale to mitigate deviations, some features displayed diverse rankings.

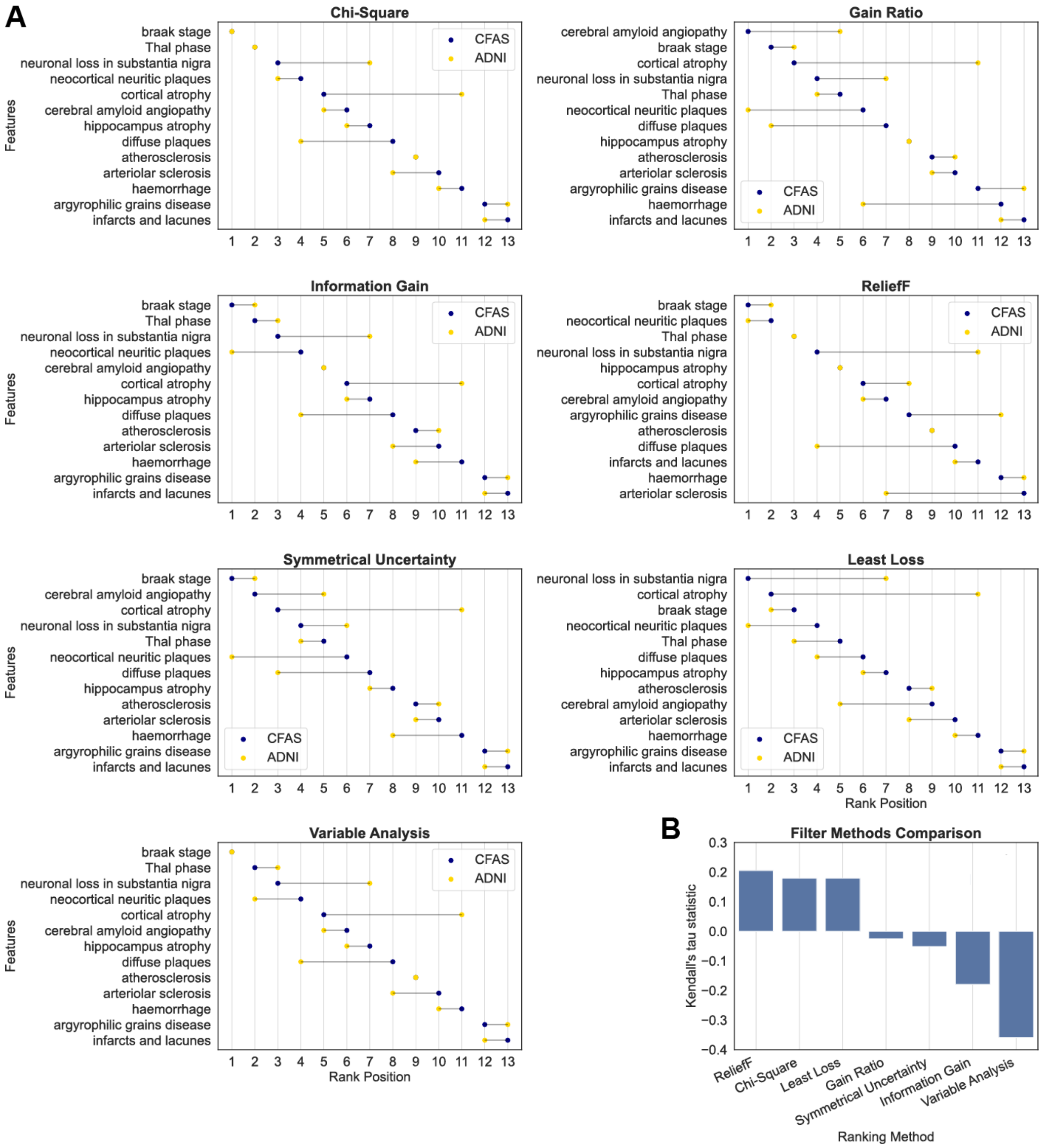


Figure 4: Comparison of feature rankings in CFAS and ADNI. (A) Relative difference in the ranking of each neuropathology feature as estimated by each filter method. (B) Kendall's tau measure of correlation between CFAS and ADNI feature rankings from each filter method.

The study used **Figure 4B** to compare different filter methods, aiming to showcase the consistency of each method when applied to two datasets, CFAS and ADNI. Kendall's tau measure was used to evaluate the level of consistent feature ranking within each dataset by each filter method. Kendall's tau measures the correlation between two ranking lists, with values near 1 signalling agreement, and values near -1 indicating disagreement. Filter methods were assessed based on their statistical relationships between the ranked features in both datasets.

We show in **Figure 4B** that the reliefF, Chi-Square, and least loss filter methods exhibited positive correlations in feature rankings for both ADNI and CFAS datasets. These results were in line with the earlier feature ranking for these filter methods. For instance, the reliefF, Chi-Square, and least loss methods showed similar rankings for features such as Thal phase, Braak stage, neuronal loss in substantia nigra, neocortical neuritic plaque, and cortical atrophy, with minor variations, indicating their reproducibility in feature ranking. Three methods maintained some consistency in feature rankings for the ADNI dataset, including the Braak stage, Thal phase, and neocortical neuritic plaque. Conversely, negative correlations were observed for variable analysis and information gain filter methods with -0.36, -0.18, respectively, when applied to ADNI and CFAS datasets. Additionally, the gain ratio and symmetrical uncertainty filter methods showed only slight consistency in feature ranking between the ADNI and CFAS datasets, with Kendall correlation coefficients (-0.03 and -0.03, respectively) close to zero. Of all methods, reliefF displayed the highest agreement between CFAS and ADNI feature rankings with Kendall correlation coefficients of 0.21, while variable analysis exhibited the most significant disagreement with Kendall correlation coefficients of -0.36. Overall, reliefF, Chi-square, and least loss filter methods showed some degree of consistency with Kendall correlation coefficients 0.21, 0.18, and 0.18 respectively, whereas the remaining methods resulted in inconsistent feature ranking between the two datasets.

High Correlations Between Alzheimer’s Disease Pathological Features

Our study conducted a thorough analysis of the CFAS and ADNI datasets, which included 13 shared features. The analysis involved plotting the feature-feature correlation to identify highly correlated features. For instance, in neocortical neuritic plaques, we observed correlation coefficients of 0.55, 0.52, 0.63, and 0.73 for diffuse plaques, CAA stage, Braak phase, and Thal phase, respectively. **Figure 5A** shows a positive correlation of Thal phase with diffuse plaques, CAA, and Braak stage with values of 0.49, 0.63, and 0.63, respectively. Additionally, The CAA was positively correlated with diffuse plaques with a value of 0.65, as reflected in **Figure 5B**. Furthermore, atherosclerosis demonstrated positive correlations with diffuse plaques, CAA, and arteriolar sclerosis of 0.80, 0.64, and 0.61, respectively. CAA also exhibited positive correlations with neuronal loss at substantia nigra and Braak stages of 0.53 and 0.47, respectively. These correlations suggest a possible cluster of features that include the Thal phase, Braak stage, CAA, neocortical neuritic plaques, and diffuse plaques.

Our study aimed to investigate whether the feature-feature correlations significantly influence the feature ranking determined by filter methods and which filter methods were most sensitive to these associations. To achieve this, we excluded the diagnostic class and considered each feature as a dependent variable, and the rest of the features as independent variables. We then utilised multiple regression models to determine the coefficient (R^2), which measures the similarity of the available feature to the rest of the dataset by identifying how the remaining features can explain the feature's variability. In both CFAS and ADNI, the Thal phase, neocortical neuritic plaque, and Braak stage showed the highest R^2 scores, indicating their potential significance in these datasets, with scores of 66%, 65%, and 61% for CFAS and 87%, 83%, and 82% for ADNI, respectively. Therefore, to investigate the impact of feature-feature correlations on feature ranking using filter methods, it is necessary to analyse the association between feature correlations and feature ranking scores.

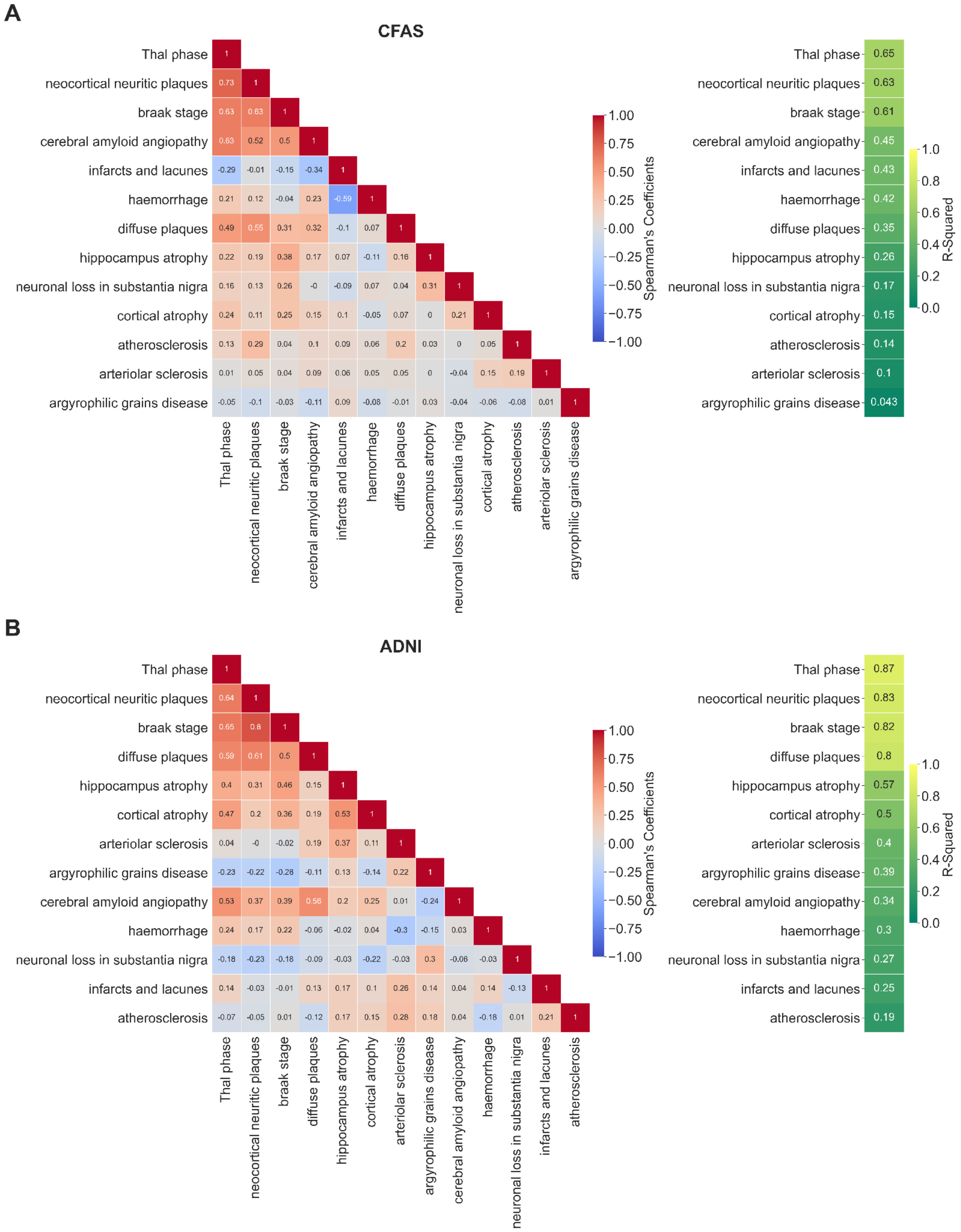


Figure 5: Spearman Correlations and R^2 of Pathological Features from the ADNI and CFAS Datasets. **(**A) A heat map of Spearman correlation for the CFAS neuropathological dataset. (B). Spearman correlation for the ADNI neuropathological dataset. A correlation coefficient close to 1 (red) indicates a very strong positive correlation between the two variables, while a correlation coefficient closer to -1 (blue) indicates a strong negative correlation. Generally, the lighter the colour, the closer it is to white (zero), and the weaker the correlation. On the right-hand side of panels, A and B, the R^2 values range from [0-1].

Impact of Feature-Feature Correlations on Feature Ranking

The study investigated discrepancies and inconsistencies in feature rankings obtained by filter methods, as some methods resulted in inconsistent feature ranking between two neuropathology datasets. The investigation considered the impact of feature-feature correlations on feature ranking scores. The aim was to identify filter methods that were less sensitive to similarities among the neuropathological features themselves in order to reduce any feature ranking discrepancy.

Pearson correlation was used to measure the association between the feature ranking scores and R^2 since all values were numeric, and a Min-Max normalisation technique was applied on feature ranking scores to ensure that all values were on the same scale. For the CFAS dataset, the results indicated a weak positive relationship between feature ranking scores and R^2 for all filter methods as shown in **Figure 6**. However, a weak relationship was observed for the gain ratio filter method. Most features were located below the relationship line, indicating that this filter method is not sensitive to the relationship between feature ranking and feature-to-feature correlation. A similar pattern was observed for the remaining filter methods, though with a slightly higher positive correlation.

The results of feature rankings obtained from the ADNI dataset were consistent with those from the CFAS dataset. There was a significant positive correlation between feature ranking scores and R^2 values, which ranged between r=0.79 and r=0.92 in ADNI. CFAS had a slightly lower positive correlation ranging between r=0.23 and r=0.63. Consequently, ADNI demonstrated a stronger correlation between feature ranking and feature-feature correlation represented by R^2 than CFAS for all feature selection methods considered. As the coefficients of ADNI data were somewhat similar, no filter method stood out from the rest.

Pearson correlations (coefficients and p-values) were used to determine which filter method was affected by feature-feature correlation when ranking features. Statistically significant correlations were found between ReliefF and CFAS (r=0.61, p-value=2.81e-02) and ADNI (r=0.92, p-value=8.85e-06). Conversely, the gain ratio and the least loss ranked last two, respectively, in terms of correlation coefficients, with the most sensitivities (CFAS: r=0.28, p-value=3.46e-01) and (ADNI: r=0.88, p-value=6.32e-05) for Gain Ratio, (CFAS: r=0.23, p-value=4.46e-01) and (ADNI: r=0.79, p-value=1.34e-03) for Least Loss, based on feature-feature correlations and feature rankings.

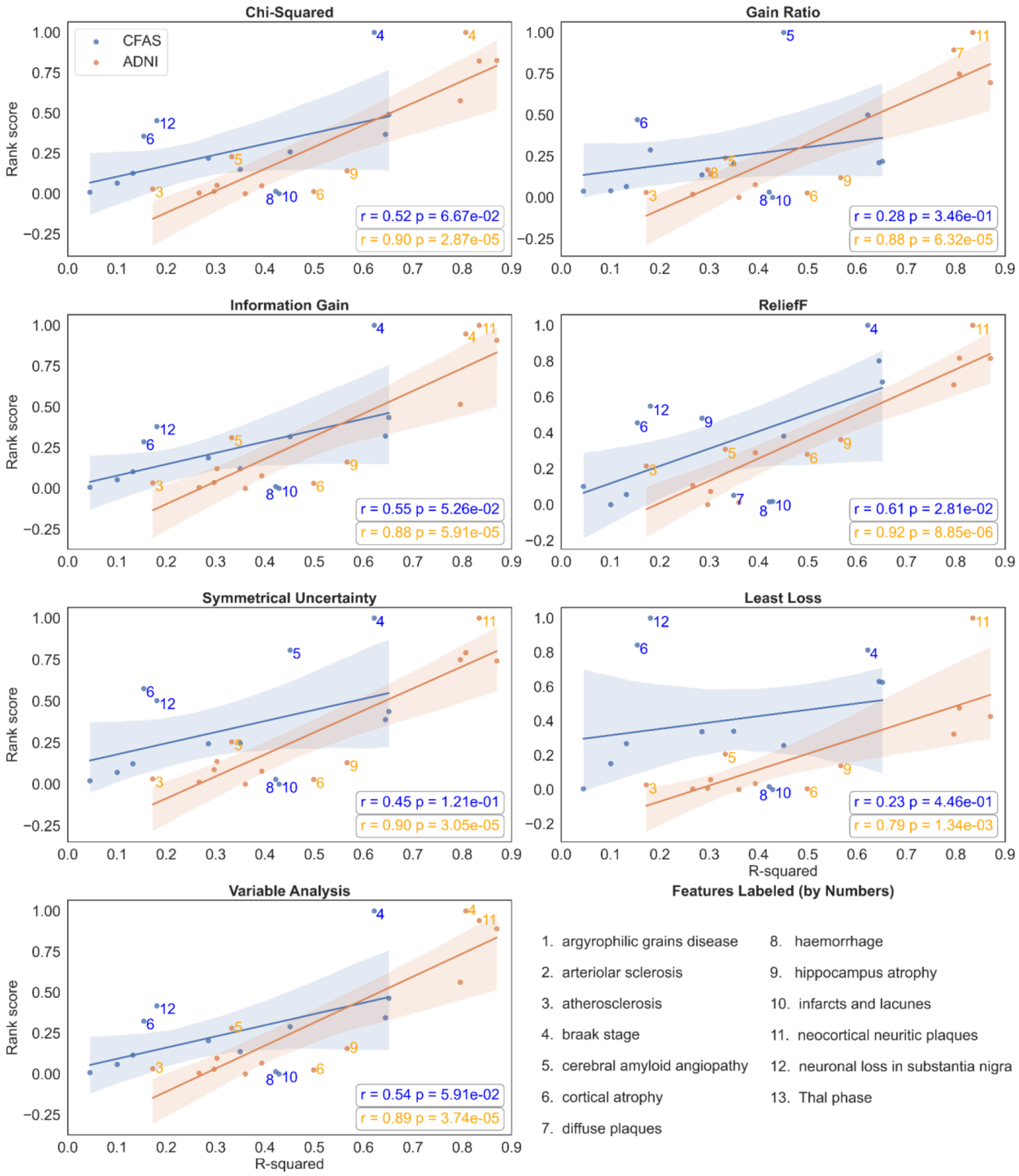


Figure 6: The relationship between feature-feature correlation and feature ranking obtained from filter methods using ADNI-pathology and CFAS datasets. Best fit linear models with confidence intervals (shading) describe the relationship. Pearson's correlation coefficients (r) and p-values (p) are reported.

Classification Using Highly Ranked Neuropathology Features

The study employed classification models, specifically the Random Forest and Naive Bayes classifiers [13], to evaluate the efficacy of selected neuropathological features. We assessed the performance of dementia classification on specific subsets of data from both ADNI and CFAS datasets. The features selected for evaluation in the CFAS and ADNI datasets were presented in **Table 2**.

Table 2: Selected sets of features from the CFAS and ADNI Datasets based on feature ranking and feature-feature correlations

Description	CFAS	ADNI
All features	All 13 features	All 13 features
Features ranking higher than expected (RHE)[+]	1. Braak stage 2. Cortical atrophy 3. Neuronal loss in substantia nigra	1. Atherosclerosis 2. Braak stage 3. CAA 4. Neocortical neuritic plaques
Features ranking lower than expected (RLE)[−]	1. Haemorrhage 2. Infarcts & Lacunes	1. Hippocampus 2. Cortical atrophy 3. Argyrophilic grain disease

[+] Features ranking higher than expected: Features that rank higher than expected, based on feature-feature correlation, are denoted as the features that fall below the confidence intervals in **Figure 6**.

[−] Features ranking lower than expected: Features that rank lower than expected, based on feature-feature correlation, are denoted as the features that fall below the confidence intervals in **Figure 6**.

Accuracy, sensitivity, and specificity rates of the Random Forest and Naive Bayes classifiers on distinct subsets of neuropathological features in the CFAS and ADNI datasets were investigated and compared. **Figure 7A** shows that the dementia classifications obtained from the Random Forest algorithm using ADNI in all group subsets were superior to those derived from CFAS, except for the sensitivity rate calculated by the Random Forest algorithm on the CFAS dataset. The same pattern was observed with the Naive Bayes algorithm based on the same group subsets **Figure 7B**.

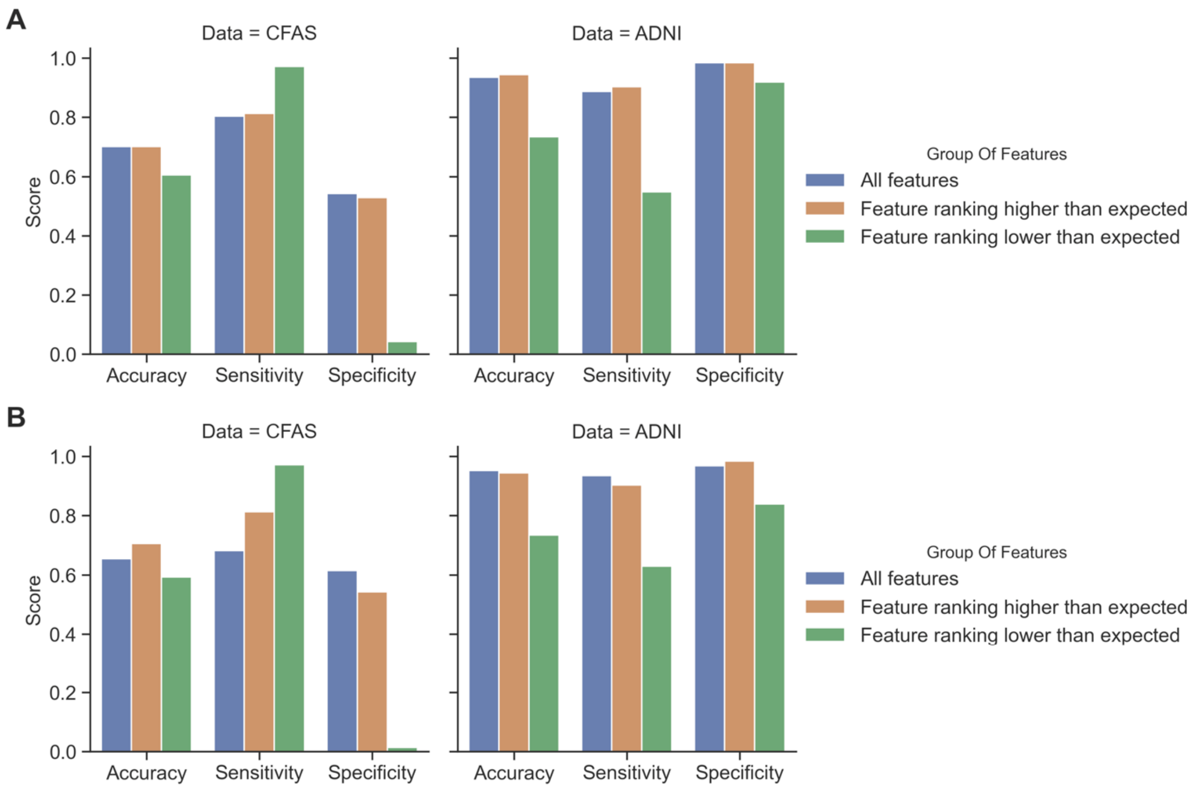


Figure 7: Classification performance using subsets of ranked features. (A) Performance results obtained by the Random Forest algorithm on the different subsets of features. (B) Performance results obtained by the Naive Bayes algorithm on the different subsets of features. For both classification algorithms, the accuracy, sensitivity, and specificity measures were used for subsets of features for CFAS and ADNI datasets. The feature names for the subsets were shown in **Table 2**.

The classification algorithms derived from distinct sets of neuropathological features demonstrate higher sensitivity rates. Higher sensitivity rates are desirable to minimise false negatives and ensure that individuals with dementia are accurately identified. Notably, CFAS exhibits remarkable sensitivity rates for RHE. For instance, the Random Forest classifiers derived from such features in CFAS have a sensitivity rate of 81.3%. A similar trend was observed with Naive Bayes classifiers derived from RHE subset in CFAS with a sensitivity rate of 81.3%.

However, the specificity rate of the CFAS features RHE was low 54.3%. This subset's low specificity implies that the classifiers cannot distinguish patients without dementia from those with the condition. Overall, the classification algorithms and all neuropathological subsets generated low specificity rates, at least for the CFAS dataset.

Conversely, the classification algorithms performed exceptionally well for distinct subsets of the neuropathological features in ADNI. According to the ADNI results, a Random Forest algorithm produced the best classifier for features RHE, with a 94.40% accuracy, 90.0% sensitivity, and 98.4% specificity, demonstrating high predictive power. While in CFAS, Naive Bayes performed best for a subset of ranking higher than expected at 70.6% accuracy, 81.3% sensitivity, and 54.3% specificity.

The feature selection analysis results align with the classification algorithms' findings. Based on these results, clinicians can leverage significant neuropathological features during the clinical assessment of dementia, including features RHE by feature-feature correlation from ADNI. Furthermore, although the sensitivity results obtained from the distinct feature subsets of CFAS by the classification algorithm were remarkable, the ADNI feature subset results were more convincing. This was because the performance measures results were balanced, making ADNI more suitable for dementia analysis, at least when neuropathological features were considered.

Discussions & Conclusions

According to the initial analysis presented in the distribution of neuropathology feature scores across dementia cases, several findings were observed, some of which have been previously published. Both the CFAS and ADNI studies showed that the percentage of individuals with dementia increased as the Braak stage increased, with a peak at stage IV for CFAS and stage V for ADNI [20,49]. Similarly, both studies observed an increase in the Thal phase [20,49]. The CFAS data showed higher cerebral amyloid angiopathy across individuals, while ADNI revealed a higher rate of dementia was associated with a higher number of brain areas with cerebral amyloid angiopathy and its severity. Additionally, both studies observed brain atrophy in dementia patients.

The ranking of neuropathological features was consistent with the results obtained by different filter methods. Using Kendall's tau as a measure of comparison between filter methods, the rank order of features was found to be correlated between ADNI and CFAS for those generated by the same methods. Some filter methods, including reliefF, Chi-Square, and least loss, generated similar rankings for specific neuropathological features in both datasets. However, other filter methods, including gain ratios and symmetrical uncertainty, produced different rankings between the ADNI and CFAS datasets. We also compared feature rankings for ADNI and CFAS in relation to the correlation between features, as measured by multivariate regression R^2. ReliefF filter method had the strongest association with feature-feature correlations in both datasets. Compared to the CFAS dataset, the ADNI dataset showed stronger relationships between filter methods and feature-feature correlations. Feature-feature correlations had the most influence on the rankings from the gain ratio filter method in both datasets.

We further assessed the impact of selecting subsets of ranked features on the classification of dementia. Classification algorithms developed from distinct sets of neuropathological features had a high accuracy of up to 94.4%, 90.3% sensitivity, and 98.4% specificity using the Random Forest classifier in ADNI for ranked features impacted by feature-feature correlation. While in CFAS, the Naive Bayes classifier achieved the highest performance in classifying dementia status with 70.6% accuracy, 81.3% sensitivity, and 54.3% specificity for the subset of highly ranked features. This classification performance is consistent with the previous classification models using neuropathology features in CFAS [20], and using imaging features from ADNI in deep neural networks [50,51].

In conclusion, this research demonstrated the association between feature-feature correlation and the feature ranking scores obtained by filter methods in medical applications such as dementia diagnosis. The study found that the ReliefF filter method is less sensitive to feature-feature correlations and that these correlations significantly impact the ranking of features and the performance of diagnosis models developed from the two dementia cohorts. The findings of this study indicate that filter methods for selecting neuropathology features associated with dementia are impacted by the feature-feature correlation and may differ between cohort studies. It's important to note that these results are based on the analysis of just two datasets, and further study may be required for broader applicability. By investigating the potential bias in filter methods, it is possible to minimise discrepancies in feature rankings and determine a reliable set of significant features for the purpose of classification algorithms.

Code Availability and Requirements

Links for python script codes in GitHub for the process and producing all results and figures (https://github.com/mdrajab/CFAS-and-ADNI-Neuropathology.git). The machine was used in this study: macOS Monterey version 12.6.2, MacBook Pro (13-inch), and Processor: 2.3 GHz Dual-Core Intel Core i5. Anaconda Navigator 1.9.12 was used to launch Jupyter Notebook version 6.1.4.

Availability of data and materials

Data from the CFAS study is accessible via application to the CFAS (http://www.cfas.ac.uk/cfas-i/data/#cfasi-data-request), under the custodianship of FM and CB. Data from the ADNI study is accessible via application to the ADNI (https://adni.loni.usc.edu/about/), contingent on adherence to the ADNI Data Use Agreement.

Declarations

List of abbreviations

CFAS: The Cognitive Function and Aging Studies; ADNI: The Alzheimer's Disease Neuroimaging Initiative; AD: Alzheimer's disease; MCI: mild cognitive impairment; CN: cognitively normal; NFT: Braak neurofibrillary tangle; CAA: cerebral amyloid angiopathy; CHI: chi-square; R^2: R-squared; SMOTE-NC: synthetic minority oversampling techniques for numerical and categorical features; RF: random forest; GNB: gaussian Naive Bayes.

Ethics approval and consent to participate

For the CFAS dataset, fully written informed consents were obtained from all participants or their authorised representatives, and the study was conducted in accordance with the ethical standards of the Declaration of Helsinki. The study was undertaken with ethical approval from a UK Multicentre Research Ethics Committee (10/H0304/61).

Consent for publication

Not Applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Medical Research Council (MRC/G9901400, U.1052.00.0013, G0900582). SBW is also supported by the Alzheimer's Society (AS-PG-17-007 and AS-PG-14-015). Work in the individual CFAS centres is supported by the UK NIHR Biomedical Research Centre for Ageing and Age – awarded to Newcastle-upon-Tyne Hospitals Foundation Trust; Cambridge Brain Bank supported by the NIHR Cambridge Biomedical Research Centre; Nottingham University Hospitals NHS Trust; University of

Sheffield, Sheffield Teaching Hospitals NHS Foundation Trust and NIHR Sheffield Biomedical Research Centre; The Thomas Willis Oxford Brain Collection, supported by the Oxford Biomedical Research Centre; The Walton Centre NHS Foundation Trust, Liverpool. DW received support from the Academy of Medical Sciences Springboard (SBF004/1052). MR is supported by the Saudi Arabia Ministry of Education. We would like to acknowledge the essential contribution of the liaison officers, the general practitioners, their staff, and nursing and residential home staff. We are grateful to our respondents and their families for their generous gift to medical research, which has made this study possible.

Authors' contributions

Data analysis; MR, DW. Writing of first draft MR, DW. Common Features between CFAS and ADNI; MR, TT, DW. Data oversight and analysis results interpretation; MR, DW. Contribution to interpretation and to the final manuscript; MR, DW. All authors read and approved the final manuscript.

Acknowledgements

We extend our sincere appreciation and gratitude to Stephen B Wharton, Fiona E Matthews, and Carol Brayne for their invaluable contributions in preparing the data and affording us the opportunity to utilise it in this study. Additionally, we would like to acknowledge the crucial role played by the liaison officers, general practitioners, their staff, and the nursing and residential home staff. We express our profound appreciation to the respondents and their families for their remarkable generosity towards medical research, without which this study would not have been possible.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and

DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References (Manuscript 3)

1. Tampi RR, Jeste DV. Dementia Is More Than Memory Loss: Neuropsychiatric Symptoms of Dementia and Their Nonpharmacological and Pharmacological Management. *Am J Psychiatry*. Am Psychiatric Assoc; 179:528–432022;

3. Kivipelto M, Mangialasche F, Ngandu T. Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. *Nat Rev Neurol*. nature.com; 14:653–662018;

4. Hardy J. Alzheimer's disease: the amyloid cascade hypothesis: an update and reappraisal. *J Alzheimers Dis*. 9:151–32006;

5. Hawkins DM. The Problem of Overfitting. ChemInform.

6. Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature Extraction: Foundations and Applications. Springer;

7. Hinrichs A, Prochno J, Ullrich M. The curse of dimensionality for numerical integration on general domains. Journal of Complexity.

8. Khaire UM, Dhanalakshmi R. Stability Investigation of Improved Whale Optimization Algorithm in the Process of Feature Selection. IETE Technical Review.

9. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. *Comput Biol Med*. 112:1033752019;

10. Thabtah, Ong, Peebles. Detection of dementia progression from functional activities data using machine learning techniques. *Intell Decis Technol*. content.iospress.com; 2022;

11. Ceyhan M, Okyay S, Kartal Y, Adar N. The Prediction of Student Grades Using Collaborative Filtering in a Course Recommender System. *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. ieeexplore.ieee.org; p. 177–81.

12. Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Sci Rep*. nature.com; 10:206302020;

13. Breiman L. Random Forests. *Mach Learn*. Springer; 45:5–322001;

14. Quinlan JR. Induction of decision trees. Machine Learning.

15. Pearson K, Henrici OMFE. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*. Royal Society; 187:253–3181896;

16. Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. *J Mach Learn Res*. 5:1205–242004;

17. Pfeffer RI, Kurosaki TT, Chance JM, Filos S, Bates D. USE OF THE MENTAL FUNCTION INDEX IN OLDER ADULTS: RELIABILITY, VALIDITY, AND MEASUREMENT OF CHANGE OVER TIME. American Journal of Epidemiology.

18. Pfeffer RI, Kurosaki TT, Harrah CH, Chance JM, Filos S. Measurement of Functional Activities in Older Adults in the Community. Journal of Gerontology.

19. Wharton SB, Brayne C, Savva GM, Matthews FE, Forster G, Simpson J, et al.. Epidemiological neuropathology: the MRC Cognitive Function and Aging Study experience. *J Alzheimers Dis*. content.iospress.com; 25:359–722011;

20. Wharton SB, Minett T, Drew D, Forster G, Matthews F, Brayne C, et al.. Epidemiological pathology of Tau in the ageing brain: application of staging for neuropil threads (BrainNet Europe protocol) to the MRC cognitive function and ageing brain study. *Acta Neuropathologica Communications*. actaneurocomms.biomedcentral …; 4:112016;

21. Wharton SB, Wang D, Parikh C, Matthews FE, Brayne C, Ince PG, et al..

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 Neuropathol Commun*. 7:1982019;

22. Franklin EE, Perrin RJ, Vincent B, Baxter M, Morris JC, Cairns NJ, et al.. Brain collection, standardized neuropathologic assessment, and comorbidity in Alzheimer's Disease Neuroimaging Initiative 2 participants. *Alzheimers Dement*. 11:815–222015;

23. Rajab MD, Jammeh E, Taketa T, Brayne C, Matthews FE, Su L, et al.. Assessment of Alzheimer-related pathologies of dementia using machine learning feature selection. *Alzheimers Res Ther*. 15:472023;

24. Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, et al.. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol*. Wiley Online Library; 65:403–132009;

25. Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N, et al.. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurol*. bmcneurol.biomedcentral.com; 12:462012;

26. Toledo JB, Cairns NJ, Da X, Chen K, Carter D, Fleisher A, et al.. Clinical and multimodal biomarker correlates of ADNI neuropathological findings. *Acta Neuropathol Commun*. actaneurocomms.biomedcentral …; 1:652013;

27. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol*. Springer; 82:239–591991;

28. Braak H, Alafuzoff I, Arzberger T, Kretzschmar H, Del Tredici K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol*. Springer; 112:389–4042006;

29. Alafuzoff I, Thal DR, Arzberger T, Bogdanovic N, Al-Sarraj S, Bodi I, et al.. Assessment of β-amyloid deposits in human brain: a study of the BrainNet Europe Consortium. Acta Neuropathologica.

30. Thal DR, Rüb U, Orantes M, Braak H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD. Neurology.

31. Huan Liu, Setiono R. Chi2: feature selection and discretization of numeric attributes. *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*. p. 388–91.

32. Kononenko I. On biases in estimating multi-valued attributes. *Ijcai*. Citeseer; p. 1034–40.

33. Robnik-Šikonja M, Kononenko I. Machine Learning.

34. Novakovic J, Strbac P, Bulatovic D. Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research.

35. Thabtah F, Kamalov F, Hammoud S, Shahamiri SR. Least Loss: A simplified filter method for feature selection. *Inf Sci* . Elsevier; 534:1–152020;

36. Rajab KD. New Hybrid Features Selection Method: A Case Study on Websites Phishing.

Security and Communication Networks. Hindawi; 2017; doi: 10.1155/2017/9838169.

37. Kamalov F, Thabtah F. A Feature Selection Method Based on Ranked Vector Scores of Features for Classification. Annals of Data Science.

38. Rajab M, Wang D. Practical Challenges and Recommendations of Filter Methods for Feature Selection. *J Info Know Mgmt*. World Scientific Publishing Co.; :20400192020;

39. Hall MA. Correlation-based Feature Selection for Machine Learning.

40. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. ACM; 11:10–82009;

41. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. JMLR. org; 12:2825–302011;

42. Breiman L. Bagging predictors. *Mach Learn*. Springer; 24:123–401996;

43. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering.

44. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al.. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 17:261–722020;

45. Shakeel F, Sabhitha AS, Sharma S. Exploratory review on class imbalance problem: An overview. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. ieeexplore.ieee.org; p. 1–8.

46. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. jair.org; 16:321–572002;

47. Lemaître, Nogueira, Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. jmlr.org; 2017;

48. Chan TF, Golub GH, LeVeque RJ. Updating Formulae and a Pairwise Algorithm for Computing Sample Variances. *COMPSTAT 1982 5th Symposium held at Toulouse 1982*. Physica-Verlag HD; p. 30–41.

49. Wharton SB, Wang D, Parikh C, Matthews FE, Brayne C, Ince PG. 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*. BioMed Central; 7:1–122019;

50. Bae JB, Lee S, Jung W, Park S, Kim W, Oh H, et al.. Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging. *Sci Rep*. 10:222522020;

51. Song Y-H, Yi J-Y, Noh Y, Jang H, Seo SW, Na DL, et al.. On the reliability of deep learning-based classification for Alzheimer's disease: Multi-cohorts, multi-vendors, multi-protocols, and head-to-head validation. *Front Neurosci*. 16:8518712022;

Chapter 5 - Conclusion and Future Directions

This thesis investigated applying filter methods to neuropathology data to address a number of challenges, and to achieve the main aims that are briefly summarised hereunder:

1. Assessing AD-related pathologies in a large cohort of elderly individuals (CAFS) to seek influential features.
2. Investigating neuropathological feature ranking in CFAS and ADNI data to evaluate feature ranking consistency among different dementia data cohorts, and to develop a global score per feature.
3. Evaluating the impact of feature-feature correlation on the ranking of features to determine a sensitive filter method for feature ranking.
4. Testing ML algorithms to determine whether they can better classify dementia conditions.

To achieve the first aim, and answer the research questions on how to rank features in an unbiased way, and determine influential neuropathological features, the study investigated using CFAS that Alzheimer-related and other dementia-related pathologies were measured, and reported the limits of ML classification of dementia utilising neuropathology indicators. Although different feature ranking methods yielded slightly different orders of association with dementia status, the top-ranked features were consistent across methods. Specifically, the Braak NFT and BrainNet tau stages were the top two selected features, in line with previous studies [10–12,124,125]. These indicators, which were detected, are passed to a classification algorithm to develop AD detection models that would help pathologists and clinicians understand dementia pathology. Interestingly, the results also indicated that subpial thorn-shaped astrocytes (TSA) in the mesial temporal lobe were highly ranked, contrasting with prior studies [12]. Additionally, three clusters were found of highly correlated measures in the dataset: CAA, TSA, and microinfarct-related, indicating redundancy. Eliminating redundant features may reduce collinearity and improve the performance of feature selection and classification accuracy [126–130].

To achieve the second aim of the thesis, and answer the research question on which filter methods are less sensitive to feature-feature correlations, the study empirically developed a process to rank scores for neuropathological features, and modelled the correlation between these scores and among the features themselves using ML techniques and rigorous empirical analysis. The approach was evaluated using classification algorithms to ascertain the usability and medical performance of the developed diagnosis models, particularly in terms of evaluation metrics such as detection rate. The empirical evaluation was conducted using AD-related neuropathological indicators and their correlations on two datasets, CFAS and ADNI.

The proposed approach was evaluated to achieve the third aim of the thesis and to answer the research question of whether there is a difference between two cohorts of ageing individuals (ADNI and CFAS) in terms of the association between feature-feature scores and feature rankings. Answering this question will reduce the discrepancy in feature ranking, and possibly generate a concise set of neuropathological indicators that can be passed to the ML algorithms to achieve the fourth aim of the thesis. An in-depth analysis of the neuropathological features of dementia was conducted utilising CFAS and ADNI and a multiple regression method. In this approach, the correlation between each feature and diagnostic class was modelled. Following, a test of whether feature-feature correlations significantly influence feature ranking, and the filter methods most sensitive to such associations are identified.

Reported results indicated consistent rankings of indicators across filter methods, with some differences noted between the ADNI and CFAS datasets. The ReliefF filter method was found to be less sensitive to feature-feature correlations, while the gain ratio filter method was the most sensitive to such correlations in both datasets. For the CFAS dataset, ReliefF and information gain showed the strongest positive correlation with feature-feature correlations, while the Chi-square and symmetrical uncertainty filter methods also showed moderate positive correlations. The least loss and gain ratio filter methods exhibited weak positive correlations. For the ADNI dataset, ReliefF and Chi-square showed the strongest positive correlation with feature-feature correlations, while information gain and variable analysis had moderate positive correlations. Gain ratio and the least loss filter methods had weak positive correlations. Overall, results suggested that ReliefF is less sensitive to similarities in data features, showing the strongest positive correlation with feature-feature correlations in both datasets. Chi-square also showed relatively strong positive correlations in both datasets. However, other filter methods showed varying degrees of sensitivity to similarities in data features, with some showing only weak positive correlations (e.g., gain ratio and the least loss) and others showing moderate to strong positive correlations (e.g., information gain and variable analysis).

The study demonstrated that feature-feature correlations significantly impact the ranking of features and the performance of diagnosis models for dementia and can vary between cohort studies. The proposed approach can reduce the discrepancy in feature ranking and generate a robust set of important features for classification algorithms. Thus, future studies on neuropathology in the context of dementia research can employ various filter methods to identify more reliable biomarkers and improve the early detection of diseases.

To achieve the forth aim, the study evaluated several classification algorithms to determine whether the diagnosis models developed are competitive in terms of evaluation metrics. The study assessed the effectiveness of selected neuropathological indicators in classifying dementia and found that dementia classification models constructed from the ADNI data subjects' performance were superior in terms of sensitivity rates of the dementia predictive models to those constructed from the CFAS dataset. Neuropathological indicators, particularly the Braak stage, were found to have the highest correlation with dementia in the CFAS dataset, while neocortical neuritic plaques had the highest correlation in the ADNI dataset.

5.1. Limitations and Ethical Implications

Utilising data–driven approaches based on ML with feature selection techniques have shown superior performance in classification accuracy, recall, and precision rates of the dementia predictive models, in addition to identifying informative dementia neuropathological indicators that can help pathologists during dementia conditions evaluation. However, machine learning with small datasets can be considered one of the limitations of this thesis due to the limited number of subjects with neuropathological indicators. In addition, there were fewer samples after data pre-processing and quality control, which ensured that the input data were adequate for learning classification models by considering class balancing normalisation, smoothing, and missing values, among others. Without considering these pre-processing operations on the original data subsets from ADNI and CFAS may result in biassed models with inaccurate measurements due to the generally proportional relationship between dataset size and the recognition of patterns by ML algorithms [131–133].

Additionally, variations in the measurement of features across different datasets, such as the CFAS and ADNI datasets, may introduce bias when ranking features or training models, thereby limiting the performance in terms of the predictability of classification models. Addressing these discrepancies requires careful consideration of feature selection and data standardisation techniques to ensure consistent interpretation of the same feature across different datasets [134]. Furthermore, the interpretation of common features between datasets may be limited by the high number of neuropathological features and the need for expert interpretation. Also, the time lapse between dementia assessment and post-mortem brain assessment may pose a challenge in relating neuropathology assessments to clinical diagnoses of dementia.

Another limitation of this research is the disproportionate focus on one form of dementia, namely AD. For instance, all data subjects in ADNI are associated with AD, and most of the data subjects in CFAS are also associated with AD although no condition type is provided for this study. This limitation is attributed to the limited accessible data available for scholars for dementia conditions other than AD.

Not considering clinical evaluation of dementia using neuropsychological assessments such as MMSE, ADAS-13, MoCA, can limit the scope of the work to post-mortem investigations. Whereas studying dementia in its preliminary stage is imperative for quick intervention and healthcare accessibility. For example, screening for a dementia precursor such as MCI or light dementia is more challenging since in these stages there is multiple overlapping between dementia and other cognitive conditions. More crucially, patients and caregivers can take advantage of early screening to manage the process of progression, and thus follow more optimised therapy and disease management plans. Analysing cognitive elements could help the proposed model expand the scope of the research work into dementia screening. This could be critical for dementia research, especially when innovative, accurate, and cost-effective technology such as ML is used. We consider this limitation a potential research opportunity for others to pursue in which they include additional cognitive items related to cognitive tests from dementia data studies such as ADNI.

Other possible ethical implications of this study are the model’s accuracy and fairness. Since this research entails the use of ML techniques, the accuracy of the models developed is subject to the features and the data observations used. In applications such as dementia diagnosis, false positives can increase the cost of detecting dementia, including AD, by asking for assessments, while false negatives may postpone effective interventions. It is imperative to reduce the number of false positives and false negatives. For model and process fairness, from one angle the research study involved not only experts in computational theory and ML, but also medical professionals such as pathologists to ensure that the data driven process followed, and the outcomes, are fair and not biased. Consequently, decisions made by the ML models have been contested and evaluated by medical professionals.

Dementia conditions' diagnosis is a sensitive medical application due to the involvement of elderly people who are inherently vulnerable. Therefore, this type of research poses challenging ethical matters requiring an ethical framework to be established for dementia-related research. Examples of ethical matters related to dementia research include consent capacity, consent approval, data collection, patient support, caregiver support, data security, etc.

Since this research study has dealt with dementia, specifically AD, using ML to build models to help pathologists understand certain dementia conditions, and important neuropathological indicators, there are some ethical issues that may arise in relation to data analysis, and decision making. For example, the process of evaluating neuropathological features used for constructing the decision-making models, is automated. In this context, and according to General Data Protection Regulation (GDPR), particularly the section that entails 'automated decision making' for subjects' data processing, there will be some ethical issues such as that the ML models (results generated) are normally biased toward the type of learning involved during the training step. The models generated and tested will possibly hold a certain bias toward their learning scheme.

Moreover, the analysis conducted using feature selection also holds a certain bias to the mathematical models used to model the feature-diagnostic and feature-feature associations, which have been primarily employed to develop the global score and to measure filter sensitivity. Another possible ethical implication of this study is the way the diagnostic class was allocated to the dementia conditions. For instance, in the ADNI study, the diagnostic class was allocated by clinicians using clinical experience as well as clinical neuropsychological tests such as MMSE and CDR-SB. However, the way the diagnostic class was assigned in the CFAS study was primarily based on brain tissue from deceased individuals and by using neuropathological features. Despite these factors, this research investigated common neuropathological features and in each data study the class allocation procedure is different which can be seen also as a limitation.

In terms of ethical implication, the difference in diagnostic class allocation between ADNI and CFAS studies poses an ethical challenge, especially in that invasive procedures have been used, even after the subject's death. These procedures are often costly and require the availability of specialised medical professionals and laboratories. More importantly, in future studies related to the diagnosis of dementia, there is a possibility of using fewer invasive procedures to determine early signs of MCI or mild dementia and using fewer cognitive tests (one or more) or neuroimaging. The former is more cost effective.

5.2. Future Directions

Further investigations of the approach can be carried out by exploring alternative ML techniques, including embedded feature selection, and evaluating additional cohorts with similar pathology features and clinical outcomes, such as the Rush Memory and Ageing Project [135] or UK Biobank, to validate the findings from the CFAS and ADNI datasets. As the number of available samples is limited, it may be necessary to employ simulated or resampled data to evaluate the approach on larger datasets. Despite that there are datasets that captures the cognitive status of patients and cognitively normal subjects, there is limited research on how the cognitive status associates with neuropathological indicators during the disease progression, and how such association differs between dementia severity, e.g. mild dementia to severe dementia, pre-dementia to mild dementia.

It is important to note that different types of dementia may exhibit distinct pathological features that must be quantified to link with dementia symptoms [136–138]. To address this issue, follow-up reports on the cognitive status of participants could be collected from individuals who knew the person until the time of death. To improve the conciseness and standardisation of neuropathological features across different datasets, it is recommended that pathologists or other experts contribute to the interpretation and inclusion of features. Furthermore, to ensure generalizability, the approach should be applied to datasets beyond dementia to other complex disorders. Finally, create a tool that can independently select relevant features in ML. This tool will have multiple filter methods that can be selected, allowing users to rank features based on their importance. Additionally, the tool will consider the association between the rank scores of features and feature-feature correlations to produce global feature ranking results.

ML algorithms, through their profound ability to analyse vast medical datasets, are paving the way for transformative real-world applications in healthcare. Furthermore, ML-driven decision-support systems are now becoming part of many hospital infrastructures, aiding physicians in making evidence-based decisions that directly impact patient outcomes. The fusion of ML with real-world medical applications ensures that the path from diagnosis to treatment becomes more streamlined, precise, and patient-centric.

Non-Manuscript References

1. Chertkow H, Feldman HH, Jacova C, Massoud F. Definitions of dementia and predementia states in Alzheimer's disease and vascular cognitive impairment: consensus from the Canadian conference on diagnosis of dementia [Internet]. Alzheimer's Research & Therapy. 2013. p. S2. Available from: http://dx.doi.org/10.1186/alzrt198

2. WHO. Dementia [Internet]. World Health Organization. 2022 [cited 2023 Jan 15]. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia

3. Duong S, Patel T, Chang F. Dementia: What pharmacists need to know. Can Pharm J . 2017;150:118–29.

4. Smith M, Buckwalter K. BEHAVIORS ASSOCIATED WITH DEMENTIA [Internet]. AJN, American Journal of Nursing. 2005. p. 40–52. Available from: http://dx.doi.org/10.1097/00000446-200507000-00028

5. Wittenberg R, Hu B, Jagger C, Kingston A, Knapp M, Comas-Herrera A, et al. Projections of care for older people with dementia in England: 2015 to 2040. Age Ageing. 2020;49:264–9.

6. Weidner WS, Barbarino P. P4-443: THE STATE OF THE ART OF DEMENTIA RESEARCH: NEW FRONTIERS [Internet]. Alzheimer's & Dementia. 2019. p. P1473–P1473. Available from: http://dx.doi.org/10.1016/j.jalz.2019.06.4115

7. Global status report on the public health response to dementia. WHO; 2021.

8. Rostamzadeh A, Bohr L, Wagner M, Baethge C, Jessen F. Progression of Subjective Cognitive Decline to MCI or Dementia in Relation to Biomarkers for Alzheimer Disease: A Meta-Analysis [Internet]. Neurology. 2022. p. 10.1212/WNL.0000000000201072. Available from: http://dx.doi.org/10.1212/wnl.0000000000201072

9. Hall A, Pekkala T, Polvikoski T, van Gils M, Kivipelto M, Lötjönen J, et al. Prediction models for dementia and neuropathology in the oldest old: the Vantaa 85+ cohort study. Alzheimers Res Ther. 2019;11:11.

10. Wharton SB, Brayne C, Savva GM, Matthews FE, Forster G, Simpson J, et al. Epidemiological neuropathology: the MRC Cognitive Function and Aging Study experience. J Alzheimers Dis. 2011;25:359–72.

11. Wharton SB, Minett T, Drew D, Forster G, Matthews F, Brayne C, et al. Epidemiological pathology of Tau in the ageing brain: application of staging for neuropil threads (BrainNet Europe protocol) to the MRC cognitive function and ageing brain study. Acta Neuropathologica Communications. 2016;4:11.

12. Wharton SB, Wang D, Parikh C, Matthews FE, Brayne C, Ince PG, et al. 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 Neuropathol Commun. 2019;7:198.

13. Weiner MW, Aisen PS, Jack CR, Jagust WJ, Trojanowski JQ, Shaw L, et al. The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans [Internet].

Alzheimer’s & Dementia. 2010. p. 202. Available from: http://dx.doi.org/10.1016/j.jalz.2010.03.007

14. Cahill S. WHO’s global action plan on the public health response to dementia: some challenges and opportunities [Internet]. Aging & Mental Health. 2020. p. 197–9. Available from: http://dx.doi.org/10.1080/13607863.2018.1544213

15. Kraepelin E. Das senile und prasenilar Irrsein. Psychiatrie : Ein Lehrbuch fur Studierende und Alzte. 1910;533:593–632.

16. Hippius H, Neundörfer G. The discovery of Alzheimer’s disease. Dialogues Clin Neurosci. 2003;5:101–8.

17. Tampi RR, Jeste DV. Dementia Is More Than Memory Loss: Neuropsychiatric Symptoms of Dementia and Their Nonpharmacological and Pharmacological Management. Am J Psychiatry. 2022;179:528–43.

18. Costa PT, Widiger TA. Introduction: Personality disorders and the five-factor model of personality [Internet]. Personality disorders and the five-factor model of personality (2nd ed.). p. 3–14. Available from: http://dx.doi.org/10.1037/10423-001

19. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020;396:413–46.

20. 2022 Alzheimer’s disease facts and figures. Alzheimers Dement. 2022;18:700–89.

21. WHO. Risk Reduction of Cognitive Decline and Dementia [Internet]. World Health Organization. 2019 [cited 2023 Jan 10]. Available from: https://www.who.int/publications/i/item/9789241550543

22. Alzheimer’s Society. Drug treatments and medication for Alzheimer’s disease [Internet]. Alzheimer’s Society. [cited 2023 Jan 25]. Available from: https://www.alzheimers.org.uk/about-dementia/treatments/dementia-drugs/drug-treatments-and-medication-alzheimers-disease

23. van Loenhoud AC, van der Flier WM, Wink AM, Dicks E, Groot C, Twisk J, et al. Cognitive reserve and clinical progression in Alzheimer disease: A paradoxical relationship. Neurology. 2019;93:e334–46.

24. Lee ATC, Richards M, Chan WC, Chiu HFK, Lee RSY, Lam LCW. Higher Dementia Incidence in Older Adults with Poor Visual Acuity. J Gerontol A Biol Sci Med Sci. 2020;75:2162–8.

25. Cullen B, Smith DJ, Deary IJ, Pell JP, Keyes KM, Evans JJ. Understanding cognitive impairment in mood disorders: mediation analyses in the UK Biobank cohort. Br J Psychiatry. 2019;215:683–90.

26. National Academies of Sciences Engineering and Medicine, Division of Behavioral and Social Sciences and Education, Board on Behavioral Cognitive and Sensory Sciences, Committee on the Decadal Survey of Behavioral and Social Science Research on Alzheimer’s Disease and Alzheimer's Disease-Related Dementias. Reducing the Impact of Dementia in America: A Decadal Survey of the Behavioral and Social Sciences. 2022.

27. ALZHEIMER'S ASSOCIATION [Internet]. Alzheimer's & Dementia. 2005. Available from: http://dx.doi.org/10.1016/s1552-5260(05)00414-0

28. United States. Congress. House. Committee on Foreign Affairs. Subcommittee on Africa, Global Health, Global Human Rights, International Organizations. The Global Challenge of Alzheimer's: The G-8 Dementia Summit and Beyond : Hearing Before the Subcommittee on Africa, Global Health, Global Human Rights, and International Organizations of the Committee on Foreign Affairs, House of Representatives, One Hundred Thirteenth Congress, First Session, November 21, 2013. 2014.

29. World Health Organization. World Report on Ageing and Health. World Health Organization; 2015.

30. Connell J, Page S. Tourism, ageing and the demographic time bomb – the implications of dementia for the visitor economy: a perspective paper [Internet]. Tourism Review. 2019. p. 81–5. Available from: http://dx.doi.org/10.1108/tr-02-2019-0070

31. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state" [Internet]. Journal of Psychiatric Research. 1975. p. 189–98. Available from: http://dx.doi.org/10.1016/0022-3956(75)90026-6

32. Jin R, Pilozzi A, Huang X. Current Cognition Tests, Potential Virtual Reality Applications, and Serious Games in Cognitive Assessment and Non-Pharmacological Therapy for Neurocognitive Disorders. J Clin Med Res [Internet]. 2020;9. Available from: http://dx.doi.org/10.3390/jcm9103287

33. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7:263–9.

34. Nasreddine ZS, Phillips NA, BÃ©dirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment [Internet]. Journal of the American Geriatrics Society. 2005. p. 695–9. Available from: http://dx.doi.org/10.1111/j.1532-5415.2005.53221.x

35. Freedman M, Leach L, Kaplan E, Winocur G, Shulman K, Delis DC. Clock Drawing: A Neuropsychological Analysis. Oxford University Press; 1994.

36. Reisberg B, Ferris SH, de Leon MJ, Crook T. The Global Deterioration Scale for assessment of primary degenerative dementia. Am J Psychiatry. 1982;139:1136–9.

37. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry. 1984;141:1356–64.

38. Breijyeh Z, Karaman R. Comprehensive Review on Alzheimer's Disease: Causes and Treatment [Internet]. Molecules. 2020. p. 5789. Available from: http://dx.doi.org/10.3390/molecules25245789

39. Querfurth HW, LaFerla FM. Alzheimer's Disease [Internet]. New England Journal of Medicine. 2010. p. 329–44. Available from: http://dx.doi.org/10.1056/nejmra0909142

40. Mayeux R, Stern Y. Epidemiology of Alzheimer Disease [Internet]. Cold Spring Harbor

Perspectives in Medicine. 2012. p. a006239–a006239. Available from: http://dx.doi.org/10.1101/cshperspect.a006239

41. Selkoe DJ. Alzheimer's disease: genes, proteins, and therapy. Physiol Rev. 2001;81:741–66.

42. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82:239–59.

43. Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science. 2002;297:353–6.

44. Savva GM, Wharton SB, Ince PG, Forster G, Matthews FE, Brayne C. Age, Neuropathology, and Dementia [Internet]. New England Journal of Medicine. 2009. p. 2302–9. Available from: http://dx.doi.org/10.1056/nejmoa0806142

45. Ellison D, Love S, Chimelli LMC, Harding B, Lowe JS, Vinters HV, et al. Neuropathology E-Book: A Reference Text of CNS Pathology. Elsevier Health Sciences; 2012.

46. Hyman BT, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Carrillo MC, et al. National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. Alzheimers Dement. 2012;8:1–13.

47. Wong K-P. Medical Image Segmentation: Methods and Applications in Functional Imaging. In: Suri JS, Wilson DL, Laxminarayan S, editors. Handbook of Biomedical Image Analysis: Volume II: Segmentation Models Part B. Boston, MA: Springer US; 2005. p. 111–82.

48. Olsson B, Lautner R, Andreasson U, Öhrfelt A, Portelius E, Bjerke M, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. Lancet Neurol. 2016;15:673–84.

49. Braak H, Alafuzoff I, Arzberger T, Kretzschmar H, Del Tredici K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol. 2006;112:389–404.

50. Alafuzoff I, Thal DR, Arzberger T, Bogdanovic N, Al-Sarraj S, Bodi I, et al. Assessment of β-amyloid deposits in human brain: a study of the BrainNet Europe Consortium [Internet]. Acta Neuropathologica. 2009. p. 309–20. Available from: http://dx.doi.org/10.1007/s00401-009-0485-4

51. Thal DR, Rüb U, Orantes M, Braak H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD [Internet]. Neurology. 2002. p. 1791–800. Available from: http://dx.doi.org/10.1212/wnl.58.12.1791

52. Ince PG, Minett T, Forster G, Brayne C, Wharton SB, Function MRCC, et al. Microinfarcts in an older population-representative brain donor cohort (MRC CFAS): Prevalence, relation to dementia and mobility, and implications for the evaluation of cerebral Small Vessel Disease. Neuropathol Appl Neurobiol. 2017;43:409–18.

53. Ikeda K. Glial fibrillary tangles and argyrophilic threads: Classification and disease specificity [Internet]. Neuropathology. 1996. p. 71–7. Available from:

http://dx.doi.org/10.1111/j.1440-1789.1996.tb00158.x

54. Ikeda K, Akiyama H, Arai T, Nishimura T. Glial Tau Pathology in Neurodegenerative Diseases: Their Nature and Comparison with Neuronal Tangles [Internet]. Neurobiology of Aging. 1998. p. S85–91. Available from: http://dx.doi.org/10.1016/s0197-4580(98)00034-7

55. Ikeda K, Akiyama H, Kondo H, Haga C, Tanno E, Tokuda T, et al. Thorn-shaped astrocytes: possibly secondarily induced tau-positive glial fibrillary tangles [Internet]. Acta Neuropathologica. 1995. p. 620–5. Available from: http://dx.doi.org/10.1007/bf00318575

56. Nishimura M, Namba Y, Ikeda K, Oda M. Glial fibrillary tangles with straight tubules in the brains of patients with progressive supranuclear palsy [Internet]. Neuroscience Letters. 1992. p. 35–8. Available from: http://dx.doi.org/10.1016/0304-3940(92)90227-x

57. Crary JF, Trojanowski JQ, Schneider JA, Abisambra JF, Abner EL, Alafuzoff I, et al. Primary age-related tauopathy (PART): a common pathology associated with human aging. Acta Neuropathol. 2014;128:755–66.

58. Dugger BN, Dickson DW. Pathology of Neurodegenerative Diseases. Cold Spring Harb Perspect Biol [Internet]. 2017;9. Available from: http://dx.doi.org/10.1101/cshperspect.a028035

59. Terry RD, Masliah E, Hansen LA. The neuropathology of Alzheimer disease and the structural basis of its cognitive alterations. Alzheimer Dis Assoc Disord. 1999;2:187–206.

60. American Psychiatric Association D. Diagnostic and statistical manual of mental disorders: DSM-5. 2013; Available from: https://www.academia.edu/download/38718268/csl6820_21.pdf

61. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer’s disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984;34:939–44.

62. Hastie T, Tibshirani R, Friedman J. The elements of statistical learnin. Cited on [Internet]. 2009;33. Available from: http://sutlib2.sut.ac.th/sut_contents/H128492.pdf

63. Abu-Mostafa YS, Magdon-Ismail M, Lin HT. Learning From Data Yaser Abu-Mostafa, Caltech http://work.caltech.edu/telecourse Self-paced version [Internet]. work.caltech.edu; 2012 [cited 2023 Jan 11]. Available from: https://home.work.caltech.edu/homework/final.pdf

64. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380:1347–58.

65. Nti IK, Quarcoo JA, Aning J, Fosu GK. A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics. 2022;5:81–97.

66. M.a. A, Thomas PA. Comparative Review of Feature Selection and Classification modeling. 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). ieeexplore.ieee.org; 2019. p. 1–9.

67. Abiodun EO, Alabdulatif A, Abiodun OI, Alawida M, Alabdulatif A, Alkhawaldeh RS. A

systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Comput Appl. 2021;33:15091–118.

68. Kamalov F, Thabtah F. A Feature Selection Method Based on Ranked Vector Scores of Features for Classification [Internet]. Annals of Data Science. 2017. p. 483–502. Available from: http://dx.doi.org/10.1007/s40745-017-0116-1

69. Alelyani S. On Feature Selection Stability: A Data Perspective [Internet]. search.proquest.com; 2013. Available from: https://search.proquest.com/openview/bb0db40c3aa975f4d598d812c532b9bf/1?pq-origsite=gscholar&cbl=18750

70. Dong G, Liu H. Feature Engineering for Machine Learning and Data Analytics. CRC Press; 2018.

71. Zawbaa HM, Emary E, Grosan C, Snasel V. Large-dimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach [Internet]. Swarm and Evolutionary Computation. 2018. p. 29–42. Available from: http://dx.doi.org/10.1016/j.swevo.2018.02.021

72. Abdel-Basset M, El-Shahat D, El-henawy I, de Albuquerque VHC, Mirjalili S. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection [Internet]. Expert Systems with Applications. 2020. p. 112824. Available from: http://dx.doi.org/10.1016/j.eswa.2019.112824

73. Karegowda AG, Manjunath AS, Jayaram MA. Feature Subset Selection Problem using Wrapper Approach in Supervised Learning [Internet]. International Journal of Computer Applications. 2010. p. 13–7. Available from: http://dx.doi.org/10.5120/169-295

74. Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature Extraction: Foundations and Applications. Springer; 2008.

75. Hinrichs A, Prochno J, Ullrich M. The curse of dimensionality for numerical integration on general domains [Internet]. Journal of Complexity. 2019. p. 25–42. Available from: http://dx.doi.org/10.1016/j.jco.2018.08.003

76. Khaire UM, Dhanalakshmi R. Stability Investigation of Improved Whale Optimization Algorithm in the Process of Feature Selection [Internet]. IETE Technical Review. 2022. p. 286–300. Available from: http://dx.doi.org/10.1080/02564602.2020.1843554

77. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review [Internet]. Journal of King Saud University - Computer and Information Sciences. 2022. p. 1060–73. Available from: http://dx.doi.org/10.1016/j.jksuci.2019.06.012

78. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications [Internet]. Computers in Biology and Medicine. 2019. p. 103375. Available from: http://dx.doi.org/10.1016/j.compbiomed.2019.103375

79. Hawkins DM. The Problem of Overfitting [Internet]. ChemInform. 2004. Available from: http://dx.doi.org/10.1002/chin.200419274

80. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for

biologists. Nat Rev Mol Cell Biol. 2022;23:40–55.

81. Lo Vercio L, Amador K, Bannister JJ, Crites S, Gutierrez A, MacDonald ME, et al. Supervised machine learning tools: a tutorial for clinicians. J Neural Eng. 2020;17:062001.

82. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling [Internet]. Family Medicine and Community Health. 2020. p. e000262. Available from: http://dx.doi.org/10.1136/fmch-2019-000262

83. Venkatesh B, Anuradha J. A Review of Feature Selection and Its Methods. Cybern Inf Technol. 2019;19:3–26.

84. Rajab M, Wang D. Practical Challenges and Recommendations of Filter Methods for Feature Selection. J Info Know Mgmt. 2020;2040019.

85. Abd-Alsabour N. On the role of dimensionality reduction. J Comput. 2018;571–9.

86. Jindal P, Kumar D. A Review on Dimensionality Reduction Techniques. Int J Comput Appl. 2017;173:42–6.

87. Kambayashi Y, Prague CRD 200, Mohania M. Data Warehousing and Knowledge Discovery: 5th International Conference, DaWaK 2003, Prague, Czech Republic, September 3-5,2003, Proceedings. Springer Science & Business Media; 2003.

88. Zeebaree DQ, Haron H, Abdulazeez AM, Zebari DA. Machine learning and Region Growing for Breast Cancer Segmentation. 2019 International Conference on Advanced Science and Engineering (ICOASE). ieeexplore.ieee.org; 2019. p. 88–93.

89. Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal. 2018;19:179–89.

90. Dash M, Liu H. Feature selection for classification [Internet]. Intelligent Data Analysis. 1997. p. 131–56. Available from: http://dx.doi.org/10.1016/s1088-467x(97)00008-5

91. Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends. 2020;1:56–70.

92. Bommert A, Welchowski T, Schmid M, Rahnenführer J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. Brief Bioinform [Internet]. 2022;23. Available from: http://dx.doi.org/10.1093/bib/bbab354

93. Braak H, Del Tredici K, Rüb U, de Vos RAI, Jansen Steur EN, Braak E. Staging of brain pathology related to sporadic Parkinson's disease [Internet]. Neurobiology of Aging. 2003. p. 197–211. Available from: http://dx.doi.org/10.1016/s0197-4580(02)00065-9

94. Price JL, Morris JC. Tangles and plaques in nondemented aging and ?preclinical? Alzheimer's disease [Internet]. Annals of Neurology. 1999. p. 358–68. Available from: http://dx.doi.org/10.1002/1531-8249(199903)45:3<358::aid-ana12>3.0.co;2-x

95. Thabtah, Ong, Peebles. Detection of dementia progression from functional activities data using machine learning techniques. Intell Decis Technol [Internet]. 2022; Available from: https://content.iospress.com/articles/intelligent-decision-technologies/idt220054

96. Rajab MD, Jammeh E, Taketa T, Brayne C, Matthews FE, Su L, et al. Assessment of Alzheimer-related pathologies of dementia using machine learning feature selection [Internet]. bioRxiv. 2022. Available from: https://www.medrxiv.org/content/10.1101/2022.04.28.22274107.abstract

97. Bharati S, Podder P, Thanh DNH, Surya Prasath VB. Dementia classification using MR imaging and clinical data with voting based machine learning models [Internet]. Multimedia Tools and Applications. 2022. p. 25971–92. Available from: http://dx.doi.org/10.1007/s11042-022-12754-x

98. Mahendran N, Durai Raj Vincent P. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease [Internet]. Computers in Biology and Medicine. 2022. p. 105056. Available from: http://dx.doi.org/10.1016/j.compbiomed.2021.105056

99. Fathi S, Ahmadi M, Dehnad A. Early diagnosis of Alzheimer's disease based on deep learning: A systematic review. Comput Biol Med. 2022;146:105634.

100. van der Schaar J, Visser LNC, Bouwman FH, Ket JCF, Scheltens P, Bredenoord AL, et al. Considerations regarding a diagnosis of Alzheimer's disease before dementia: a systematic review. Alzheimers Res Ther. 2022;14:31.

101. Boccardi M, Handels R, Gold M, Grazia A, Lutz MW, Martin M, et al. Clinical research in dementia: A perspective on implementing innovation. Alzheimers Dement. 2022;18:2352–67.

102. Odusami M, Maskeliūnas R, Damaševičius R. An Intelligent System for Early Recognition of Alzheimer's Disease Using Neuroimaging. Sensors [Internet]. 2022;22. Available from: http://dx.doi.org/10.3390/s22030740

103. Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. Alzheimers Dement [Internet]. 2022; Available from: http://dx.doi.org/10.1002/alz.12721

104. Alvi AM, Siuly S, Wang H, Wang K, Whittaker F. A deep learning based framework for diagnosis of mild cognitive impairment. Knowledge-Based Systems. 2022;248:108815.

105. Esiri MM, Matthews F, Brayne C, Ince PG, Matthews FE, Xuereb JH, et al. Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. Lancet [Internet]. 2001;357. Available from: https://ora.ox.ac.uk/objects/uuid:848a4dac-6a33-4fcd-904b-22c30c5b64c9

106. Matthews FE, Brayne C, Lowe J, McKeith I, Wharton SB, Ince P. Epidemiological pathology of dementia: attributable-risks at death in the Medical Research Council Cognitive Function and Ageing Study. PLoS Med. 2009;6:e1000180.

107. Shilaskar S, Ghatol A. Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. Expert Syst Appl. 2013;40:4146–53.

108. Verma AK, Pal S, Kumar S. Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study. Appl Biochem Biotechnol.

2020;190:341–59.

109. Castellazzi G, Cuzzoni MG, Cotta Ramusino M, Martinelli D, Denaro F, Ricciardi A, et al. A Machine Learning Approach for the Differential Diagnosis of Alzheimer and Vascular Dementia Fed by MRI Selected Features. Front Neuroinform. 2020;14:25.

110. Boyle PA, Yu L, Wilson RS, Leurgans SE, Schneider JA, Bennett DA. Person-specific contribution of neuropathologies to cognitive loss in old age. Ann Neurol. 2018;83:74–83.

111. Boyle PA, Yu L, Leurgans SE, Wilson RS, Brookmeyer R, Schneider JA, et al. Attributable risk of Alzheimer's dementia attributed to age-related neuropathologies [Internet]. Annals of Neurology. 2019. p. 114–24. Available from: http://dx.doi.org/10.1002/ana.25380

112. Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. Sci Rep. 2020;10:20630.

113. Haider F, de la Fuente S, Luz S. An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. IEEE J Sel Top Signal Process. 2020;14:272–81.

114. Lombardi A, Diacono D, Amoroso N, Biecek P, Monaco A, Bellantuono L, et al. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. Brain Inform. 2022;9:17.

115. Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. J Mach Learn Res. 2004;5:1205–24.

116. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226–38.

117. Cai Y, Huang T, Hu L, Shi X, Xie L, Li Y. Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids. 2012;42:1387–95.

118. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. 2017;18:9.

119. Lin L, Chen Q, Hirsch JP, Yoo S, Yeung K, Bumgarner RE, et al. Temporal genetic association and temporal genetic causality methods for dissecting complex networks. Nat Commun. 2018;9:3980.

120. Gu Q, Li Z, Han J. Generalized Fisher Score for Feature Selection [Internet]. arXiv [cs.LG]. 2012. Available from: http://arxiv.org/abs/1202.3725

121. Hall MA. Correlation-based Feature Selection for Machine Learning. 1999.

122. Alirezanejad M, Enayatifar R, Motameni H, Nematzadeh H. Heuristic filter feature selection methods for medical datasets. Genomics. 2020;112:1173–81.

123. Khagi B, Kwon G-R, Lama R. Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques. Int J Imaging Syst Technol. 2019;29:297–310.

124. Lace G, Ince PG, Brayne C, Savva GM, Matthews FE, de Silva R, et al. Mesial temporal astrocyte tau pathology in the MRC-CFAS ageing brain cohort. Dement Geriatr Cogn Disord. 2012;34:15–24.

125. Keo A, Mahfouz A, Ingrassia AMT, Meneboo J-P, Villenet C, Mutez E, et al. Transcriptomic signatures of brain regional vulnerability to Parkinson's disease. Commun Biol. 2020;3:101.

126. Jain I, Jain VK, Jain R. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification [Internet]. Applied Soft Computing. 2018. p. 203–15. Available from: http://dx.doi.org/10.1016/j.asoc.2017.09.038

127. Mwadulo MW. A review on feature selection methods for classification tasks [Internet]. Citeseer; 2016 [cited 2021 Apr 6]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1075.7828&rep=rep1&type=pdf

128. Shi H, Li H, Zhang D, Cheng C, Cao X. An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification. Computer Networks. 2018;132:81–98.

129. Gómez Flores W, Pereira WC de A, Infantosi AFC. Improving classification performance of breast lesions on ultrasonography. Pattern Recognit. 2015;48:1125–36.

130. Agarwal B, Mittal N. Prominent feature extraction for review analysis: an empirical study. J Exp Theor Artif Intell. 2016;28:485–98.

131. Malave N, Nimkar AV. A survey on effects of class imbalance in data pre-processing stage of classification problem. International Journal of Computational Systems Engineering. 2020;6:63–75.

132. Sun Y, Fung BCM, Haghighat F. The generalizability of pre-processing techniques on the accuracy and fairness of data-driven building models: A case study [Internet]. Energy and Buildings. 2022. p. 112204. Available from: http://dx.doi.org/10.1016/j.enbuild.2022.112204

133. Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings. 2022;3:91–9.

134. Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, et al. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective [Internet]. arXiv [cs.LG]. 2022. Available from: http://arxiv.org/abs/2202.01602

135. Bennett DA, Schneider JA, Buchman AS, Mendes de Leon C, Bienias JL, Wilson RS. The Rush Memory and Aging Project: study design and baseline characteristics of the study cohort. Neuroepidemiology. 2005;25:163–75.

136. Elahi FM, Miller BL. A clinicopathological approach to the diagnosis of dementia [Internet]. Nature Reviews Neurology. 2017. p. 457–76. Available from: http://dx.doi.org/10.1038/nrneurol.2017.96

137. Barker WW, Luis CA, Kashuba A, Luis M, Harwood DG, Loewenstein D, et al. Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the State of Florida Brain Bank. Alzheimer Dis Assoc Disord. 2002;16:203–12.

138. Geldmacher DS, Whitehouse PJ. Evaluation of dementia. N Engl J Med. 1996;335:330–6.