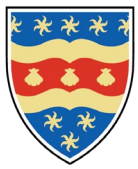
2022

MACHINE LEARNING-BASED EXPLORATION OF BLOOD-BASED BIOMARKERS FOR ALZHEIMER'S DISEASE DIAGNOSIS

Eke, Chima Stanley



UNIVERSITY OF
PLYMOUTH

MACHINE LEARNING-BASED EXPLORATION OF BLOOD-BASED BIOMARKERS FOR ALZHEIMER'S DISEASE DIAGNOSIS

by

CHIMA STANLEY EKE

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

January, 2022

Acknowledgements

My PhD research has been a journey with many feel-good experiences, but not without some moments of doubt. It would not have been possible to complete without the support from several people.

First, I would like to express my deepest appreciation to my Director of Studies Prof. Emmanuel Ifeachor for his academic mentorship, motivation, unwavering support, and generous advice from the outset of my research. I count it as an enormous privilege to learn from his expertise, wealth of knowledge and experience.

I am also extremely grateful to the rest of my supervisory team: Dr Xinzhong Li, Dr Camille Carroll and Dr Stephen Pearson for the constructive discussions and suggestions on our co-authored publications.

Special thank you to the Signal Processing and Multimedia Communications (SPMC) research group: Dr Emmanuel Jammeh for the memorable moments spent sharing ideas, providing support or constructive critique; Dr Is-Haka Mkwawa and Prof. Lingfen Sun for their goodwill and useful suggestions; and student members of the group who made my work easier with their friendship.

I thankfully acknowledge the funding from the EU Horizon 2020 Research and Innovation Programme under Marie Sklodowska-Curie Innovative Training Networks (MSCA-ITN-2016-ETN), Grant Agreement No. 721281, the BBDiag Consortium. I cherish the special moments shared with the Early Career Researchers, supervisors, and program managers at different academic, leadership training and social events organised within and outside the Consortium.

Finally, I am immensely grateful to my family: Chinyere my wife, my mum and my siblings for their unflinching sacrifices, goodwill, and prayers. Posthumous thank you to my dad who would have been most proud of me.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

Word count for the main body of this thesis: 22,526.

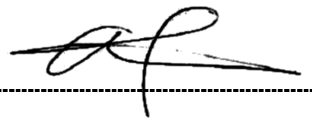
Publications

C. S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor, "Identification of Optimum Panel of Blood-based Biomarkers for Alzheimer's Disease Diagnosis Using Machine Learning," In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018: IEEE, pp. 3991-3994. DOI: 10.1109/EMBC.2018.8513293.

C. S. Eke, F. Sakr, E. Jammeh, P. Zhao, and E. Ifeachor, "A Robust Blood-based Signature of Cerebrospinal Fluid Aβ42 Status," In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020: IEEE, pp. 5523-5526. DOI: 10.1109/EMBC44109.2020.9175158.

C. S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor, "Early Detection of Alzheimer's Disease with Blood Plasma Proteins Using Support Vector Machines," *IEEE journal of biomedical and health informatics (JBHI),* vol. 25, no. 1, pp. 218-226, 2020. DOI: 10.1109/JBHI.2020.2984355.

A. H. Al-Nuaimi, M. Blūma, S. S. Al-Juboori, C. S. Eke, E. Jammeh, L. Sun, E. Ifeachor, "Robust EEG-Based Biomarkers to Detect Alzheimer's Disease," *Brain Sciences*, vol. 11, no. 8, p. 1026, 2021. DO1: 10.3390/brainsci11081026.

Signed 

Date 30.01.2022

Machine Learning-based Exploration of Blood-based Biomarkers for Alzheimer's Disease Diagnosis

Chima Stanley Eke

Abstract

Alzheimer's disease (AD) is a neurodegenerative disease with typical clinical symptoms in the form of progressive cognitive impairment and memory loss. To facilitate early diagnosis of AD and a greater understanding of the mechanisms underlying its clinical expression, the use of biomarkers is necessary. Furthermore, it is believed that biomarkers provide a more objective and accessible means of diagnosis. Currently, established biomarkers include neuroimaging markers, such as those based on positron emission tomography (PET), and biochemical markers such as cerebrospinal fluid (CSF) markers. However, neuroimaging is expensive and may not be widely available and CSF testing is invasive. Blood-based biomarkers offer the potential for the development of minimally invasive, low-cost and time-efficient methods for AD detection to complement CSF and neuroimaging. In this work, a data-driven approach, machine learning in particular, was exploited to identify blood-based biomarker panels consisting of a few markers (as no single marker provides sufficient performance) that may serve as screening tools in a multi-stage diagnostic procedure. Novel contributions were made in biomarker discovery, including identification of novel panels as well as panel selection procedures that emphasize performance and robustness. Identified biomarker panels have remarkable classification performance at discriminating between Alzheimer's dementia as well as mild cognitive impairment subjects and normal controls. Another set of identified blood-based biomarkers could classify individuals with abnormal/normal levels of CSF amyloid β42, which is one of the key early markers of AD. Furthermore, a novel software prototype was developed to demonstrate the possible clinical use of identified biomarker panels. A significance

of this work is its potential contribution to the development of rapid testing and cost-effective point of care devices to facilitate AD diagnosis.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A1M	Alpha-1 microglobulin
A2M	Alpha-2 macroglobulin
AD	Alzheimer's disease
ADD	Alzheimer's disease at dementia stage
ADIP	Adiponectin
ADNI	Alzheimer's disease neuroimaging initiative
ANN	Artificial neural network
APOA2	Apolipoprotein A2
APOE	Apolipoprotein E
APOE4	Apolipoprotein ε4
AUC	Area under receiver operating curve
AUC_{oo}	AUC evaluated on the out-of-bag sample
Aβ	Beta amyloid or amyloid-beta
B2M	Beta-2 microglobulin
BBDiag	Blood Biomarker-based Diagnostic Tools for Early-Stage Alzheimer's Disease
BNP	Brain natriuretic peptide
BTC	Betacellulin
CC3	Complement C3
CFS	Cerebrospinal fluid
CGA	Chromogranin-A
CLA	Complete linear aggregation
CP	Candidate panel

CRP	C-reactive protein
CSF	Cerebrospinal fluid
CSV	Comma separated value
CTL	Normal control
CV	Cross-validation
CWA	Complete weighted aggregation
EOAD	Early-onset Alzheimer's disease
EOT3	Eotaxin-3
FABP	Fatty acid binding protein
FN	False negative
FP	False positive
FVII	Factor VII
GCSF	Granulocyte-colony stimulating factor
GUI	Graphical user interface
HBEGF	Heparin-binding EGF-like growth factor
IGM	Immunoglobulin M
IL18	Interleukin-18
IL3	Interleukin-3
KI	Kuncheva index
KI_{tot}	Kuncheva index - Total
KNN	K-nearest neighbours
LDA	Least discriminant analysis
LOAD	Late-onset Alzheimer's disease

MCC	Matthew's correlation coefficient
MCI	Mild cognitive impairment
MCP1	Monocyte chemotactic protein 1α
MCSF1	Monocyte-colony stimulating factor 1
MPO	Myeloperoxidase
MRI	Magnetic resonance imaging
MSK	Most stable kernel
NFT	Neurofibrillary tangles
NPV	Negative predictive value
PAPPA	Pregnancy-associated plasma protein a
PET	Positron emission tomography
PLGF	Placenta growth factor
PPP	Pancreatic polypeptide
PPV	Positive predictive value
PYY	Peptide YY
RAD	Rapid application development
RAGE	Receptor for advanced glycosylation end
RBF	Radial basis function
RF	Random forests
RFE	Recursive feature elimination
RF-RFE	Random forests with recursive feature elimination
ROC	Receiver operating curve
SD	Standard deviation

SGOT	Serum glutamic oxaloacetic transaminase
SN	Sensitivity
SP	Specificity
SU	Symmetrical uncertainty
SUVR	Standardised uptake value ratio
SVM	Support vector machine
SVM-RFE	Support vector machine with recursive feature elimination
TLSP	T-lymphocyte secreted protein 1.309
TNC	Tenascin C
TP	True positive
TTR	Transthyretin
VCAM	Vascular cell adhesion molecule-1
VIT	Vitronectin
XAI	Explainable artificial intelligence

List of Symbols

$\hat{\alpha}_i$	Lagrange multipliers
b	Bias
Ъ	Number of bootstrap samples
C	Cost parameter for misclassification
d	Data dimensionality
D	Dataset
D\Dt	D less Dt samples at time t
$D(.)$	Decision function
ξ	Slack variable
f	Feature subsets
$h(.)$	Hyperplane function
$H(.)$	Entropy
k	Number of folds of a cross-validation
K	Kernel function
κ	Subsamples
M	Decision margin
m	One-half of decision margin
ɱ	Number of features common to a pair of signatures
N	Sample size
p	Probability
ρ	Proportion of a subsample from the original dataset
ŗ	Ranking

r_{fc}	Feature-class correlation
r_{fc}	Feature-feature correlation
$\mathbf{R}$	Aggregate ranking
s	Signature size
t	Time
w	Weight vector
$\dot{w}$	Bootstrap-dependent weight
x	Predictor variables
y	Class labels
$	American dollar
£	British pound

Chapter 1 Introduction

1.1 Research challenges and motivations

Dementia is the leading cause of disability and dependency among older people. There are over 50 million people living with dementia worldwide and this figure is projected to increase to 152 million by 2050 [1]. The social and economic burden of dementia is enormous with an annual global cost estimated at $1 trillion and projected to double by 2030 [1, 2]. In the United Kingdom, the annual cost is estimated at £26 billion [3]. As a result, addressing the challenge of dementia is now a national and global priority [2, 4]. Alzheimer's disease (AD) is an age-related neurodegenerative disease clinically characterised by a progressive loss of memory and cognition. It is the most prevalent cause of dementia, accounting for 60-80% of cases [5].

There is no cure for AD at present but there is intense research effort to develop interventions that slow, halt, or prevent the disease [6-10]. AD has a long preclinical phase and clinical symptoms may only become apparent decades after disease onset. One of the typical early symptoms of AD is loss of recent memory, followed by mild cognitive impairment (MCI), and then dementia [30]. About 32% of people who develop MCI progress to dementia within 5 years [5].

Emerging treatment and preventive strategies emphasise early diagnosis and intervention before the onset of clinical symptoms or before significant brain cell damage as key to successful treatment and preventive intervention [9-12]. In addition, huge economic savings are foreseen in healthcare through early diagnosis [13]. To facilitate early diagnosis of the disease, and for greater understanding of the mechanisms underlying its clinical expression, the use of biomarkers is necessary [14-16]. Furthermore, it is believed that biomarkers provide a more objective and accessible means of diagnosis. In 2011, it was estimated that up to 50% of people living with the

disease in high-income countriesmay not have received a formal diagnosis, and up to 77% globally [17].

Therefore, given the prevalence of AD, there is a need for non-invasive, low-cost, and reliable biomarkers that can be applied in clinical practice for early diagnosis. Recent diagnostic guidelines recommend the use of two main categories of disease-defining biomarkers of AD, namely cerebral spinal fluid (CSF) and positron emission tomography (PET) neuroimaging biomarkers [14, 16, 18]. However, CSF analysis is not readily used in clinical practice due to the relative invasiveness of sample collection [19]. PET imaging, on the other hand, is expensive and available only in specialist centres [20, 21].

Blood-based biomarkers have shown promising results in early diagnosis of AD and present a less invasive and potentially less expensive (and more accessible) approach compared to CSF and PET biomarkers. At the minimum, even if they are less specific and accurate relative to CSF and PET markers, they can serve as a first-line screening tool to complement the more established biomarkers. This can be particularly beneficial in reducing the current screening failures in clinical trials [21]. Furthermore, blood-based biomarkers may provide insights into yet undiscovered mechanisms of the disease's development.

Blood-based biomarker search usually involves high-dimensional complex data, generated by biosensors capable of producing large arrays of measurements from blood. This complexity, resulting from the inherent nature of blood, and dimensionality of the data present a major challenge in identifying suitable blood biomarkers of the disease. Moreover, no single blood biomarker can accurately detect AD, hence the need to combine many biomarkers. These necessitate the application of advanced

data analysis methods, machine learning techniques in particular, to conduct the biomarker search.

In summary, there is a pressing need to develop minimally invasive, low-cost, and easy to use methods to facilitate early detection of AD and monitoring of response to therapeutic interventions. Exploration of blood-based biomarkers presents a potential to meet this need.

1.2 Aim and objectives

This project deals with an aspect of the blood biomarker-based diagnostics for early-stage Alzheimer's disease (BBDiag) project. BBDiag is an EU H2020 Marie Curie project initiated to address some of the challenges facing the development of clinically useful AD blood-based biomarker diagnostic tools. It aims to develop novel low-cost biosensors to detect multiple biomarkers in blood and point-of-care devices to assess early-stage AD (see Figure 1.1). The outputs of the biosensors are then analysed by an intelligent decision-making algorithm to detect AD.

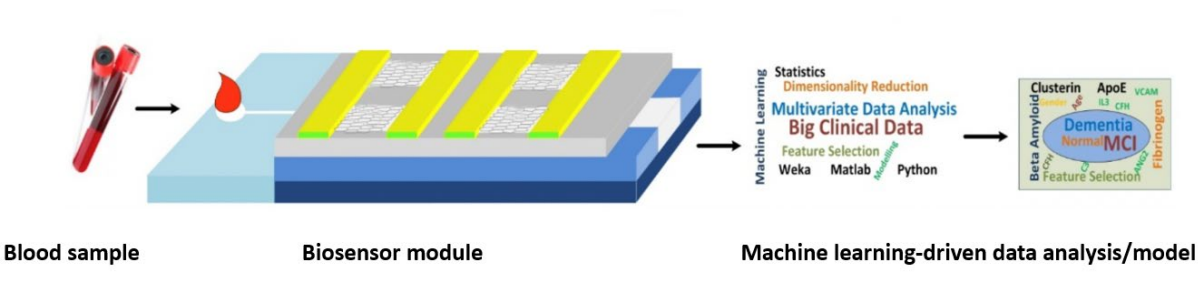


Figure 1.1. A framework of BBDiag project objective.

The aim of this project is to apply intelligent data-driven approaches (machine learning in particular) to identify potentially useful AD blood-based biomarkers for use in clinical practice. The core objectives are to:

1. conduct a literature review to identify existing potential blood-based biomarkers of AD;
2. identify and collect relevant data;
3. conduct machine learning driven data analyses to identify potential clinically useful blood-based biomarkers for the diagnosis of AD;
4. demonstrate potential utility of the identified biomarkers in a real-life clinical setting.

1.3 Contributions of the thesis

This thesis makes the following contributions to knowledge. It:

1. provides a detailed understanding of blood-based biomarkers for AD diagnosis;
2. identifies robust novel blood-based biomarkers for AD detection at later stages;
3. identifies robust blood-based biomarkers for detection of AD at MCI and dementia stages, and potentially at earlier stages;
4. identifies a robust novel biomarker panel for detection of CSF β amyloid status, which is one of the earliest pathological indicators of AD;
5. develops novel methodological frameworks for biomarker search;
6. develops and demonstrates a novel prototype software to illustrate the potential utility of blood-based biomarkers in real-life clinical practice.

1.4 Thesis outline

This Thesis consists of eight chapters. Chapter 1 introduces the project highlighting the motivations, aims and objectives, and contributions. Chapter 2 discusses key background concepts including AD, its diagnosis and biomarkers, diagnostic performance metrics, machine learning, model evaluation techniques, and feature selection methods. In Chapter 3, the materials and methods are described, including the study dataset as well as software tools utilised such as MATLAB, WEKA and

Python. In Chapter 4, the investigation on the identification of a sparse panel of blood-based biomarkers for the detection of AD is discussed. Chapter 5 expands the scope of the preceding chapter, providing a more robust approach, including consideration of individuals at earlier stages of the disease (MCI in particular) to be taken into account. Chapter 6 presents an exploratory study to identify a robust biomarker signature indicative of one of the pathological hallmarks of AD (amyloid abnormality), making stratification of individuals at risk of developing the disease before any clinical symptoms possible. Chapter 7 discusses the implementation of a prototype application to facilitate AD detection based on blood biomarkers, to demonstrate the potential use case of the biomarkers in real-life clinical settings. Finally, Chapter 8 discusses the contributions to knowledge, limitations and future directions, and conclusion.

Chapter 2 Background

2.1 Alzheimer's disease

Alzheimer's disease is a type of neurodegenerative disease with typical clinical symptoms manifesting in the form of progressive cognitive impairment and memory loss. The onset of these symptoms is prevalent in older people above the age of 65 years, usually referred to as late-onset AD (LOAD). There is also the other variant of the disease that affects younger people with a certain genetic predisposition. The disease process begins 10-20 years before the onset of symptoms. This stage is referred to as the preclinical stage of the disease and is clinically indistinguishable from normal aging as illustrated in Figure 2.1. Afterwards, clinical symptoms in the form of mild cognitive impairment (MCI) – a degree of cognitive impairment that is abnormal for age [22] appears, and then finally dementia.

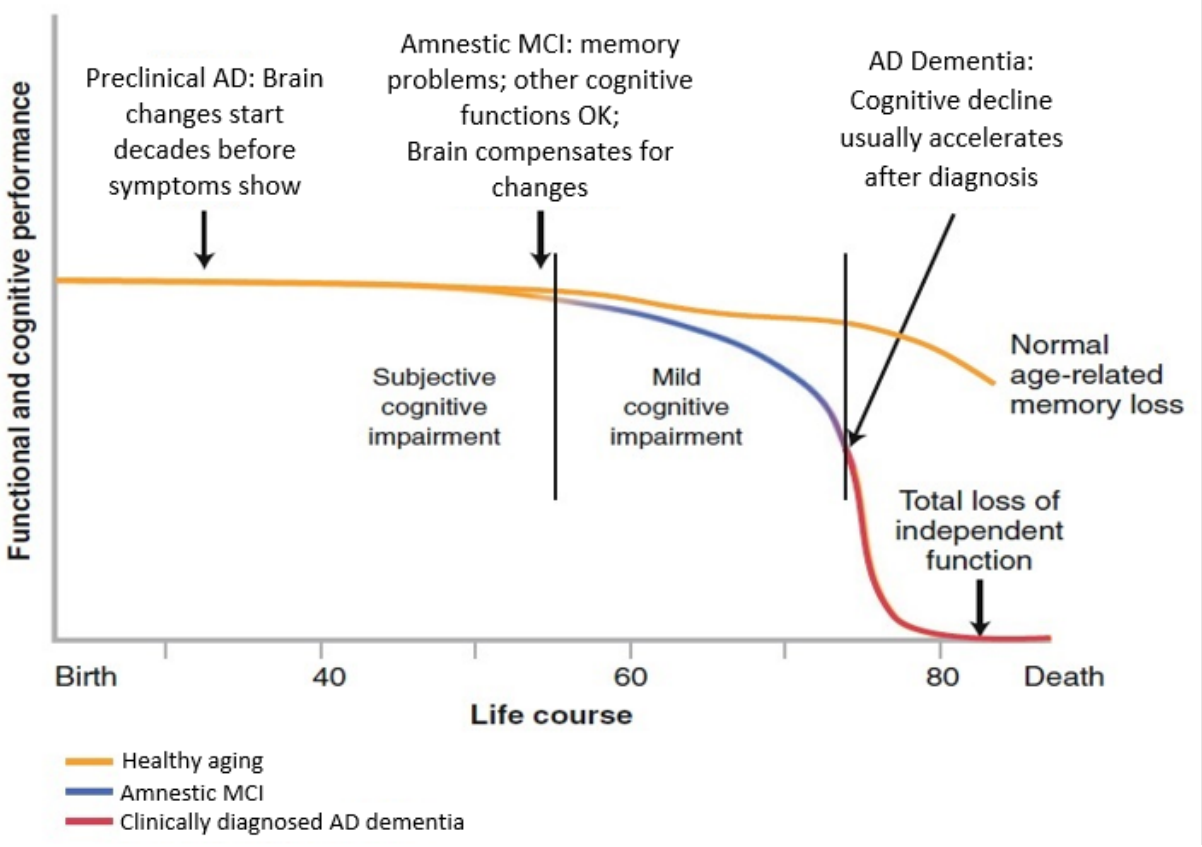


Figure 2.1. Stages of AD development from normal – MCI – ADD (modified from [23]).

Symptoms of dementia include cognitive or behavioural impairment just as MCI but to a degree that interferes significantly with daily activities [24]. It is however worth noting that these clinical symptoms are not specific to AD and therefore can be due to other or multiple causes, one of which is AD [14]. For instance, cognitive impairments may be due to head trauma, substance abuse or metabolic disturbance [25]. Other common causes of dementia include hippocampal sclerosis, frontotemporal lobar degeneration, cerebrovascular, Lewy body and Parkinson's disease [26]. Dementia is ultimately fatal.

2.2 Diagnosis of Alzheimer's disease

The biological definition of AD provides a means to diagnose the disease in terms of its underlying biological expression rather than as a syndrome consisting of signs and symptoms. This approach to defining AD is more sensitive and specific to the disease and can identify individuals even at the preclinical stage [14]. Disease-defining biological hallmarks of AD include the accumulation of Aβ protein fragments (known as amyloid plaques) external to brain neurons and aggregation of an abnormal form of tau protein (known as tau tangles) inside neurons as illustrated in Figure 2.2.

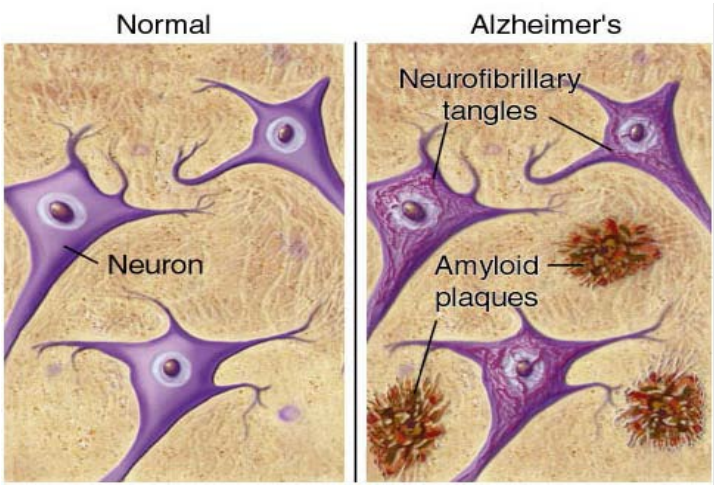


Figure 2.2. An illustration of extracellular amyloid plaques and intracellular neurofibrillary tangles in normal people and AD subjects [27].

The formation of amyloid plaques and smaller soluble aggregates of Aβ called oligomers is believed to contribute to neuronal damage and death (i.e., neurodegeneration) by inhibiting inter-neuronal communication at synaptic junctions. Within neurons, tau neurofibrillary tangles (NFTs) block the transport of nutrients. Accumulation of Aβ may precede the formation of tau tangles and increasing amyloid plaques are associated with subsequent increases in tau tangles [28, 29], albeit the complete sequence of events is unclear.

Other brain changes that accompany AD include inflammation and atrophy. Brain atrophy occurs owing to cell loss resulting from the death of cells. Chronic neuroinflammation is believed to set in due to the inability of the brain to adequately clear toxic beta-amyloid and tau protein accumulations as well as debris from dying cells.

Current diagnostic guidelines advocate diagnosis of AD in living persons based on the two main neuropathologic events of the disease (Aβ and NFT abnormalities) using biomarkers.

2.3 Biomarkers of Alzheimer's disease

A biomarker is a biological feature that can be measured in vivo as indicative of a specific biological state. They may be used to indicate a normal or abnormal process, a condition or disease, progression of disease, or response to treatment [16, 30-32]. The ideal diagnostic biomarker of AD should possess the following characteristics [30, 33].

1. The biomarker should detect a fundamental neuropathologic feature of AD.

2. It should be validated in post-mortem confirmed AD cases.

3. It should be precise, i.e., able to detect AD in its early stages and differentiate it from other causes of dementia.
4. Its measurement should be reliable, and the process should be non-invasive, easy to perform, and inexpensive.

Current recommended diagnostic biomarkers of AD are CSF and PET neuroimaging measurements of amyloid-beta and CSF tau protein abnormalities. Genetic biomarkers are yet another category of AD biomarkers of key importance. There are also other emerging biomarkers of AD at the early stages of development, one of which is blood biomarkers. Besides the diagnostic value of these biomarkers, they are also crucial to understanding the disease mechanism and developing effective pharmacological interventions.

2.3.1 Genetic markers

Studies have shown that genetic factors have a significant impact on the risk of developing Alzheimer's disease. An estimated 1% of AD cases develop because of mutations to one of the genes for amyloid precursor protein (APP), presenilin 1 (PS1) and presenilin 2 (PS2) proteins [34]. Hundreds of distinct mutations have been discovered across these genes [35]. Individuals inheriting a mutation to the APP or PS1 gene are guaranteed to develop AD while those inheriting a mutation to the PS2 gene have a 95% chance of developing the disease [36]. Individuals that have mutations in any of these three genes tend to develop symptoms of AD before age 65, sometimes as early as age 30. This is usually referred to as early-onset AD (EOAD).

Apolipoprotein E ε4 (APOE4) gene is the strongest genetic risk factor for LOAD. Individuals with one or two copies of the e4 allele of the gene have an increased risk of developing LOAD [37, 38]. In contrast to EOAD, LOAD exhibits a more sophisticated

pattern of relationship between genetic and non-genetic factors. Inheriting the APOE4 gene does not guarantee that an individual will have AD.

2.3.2 Neuroimaging biomarkers

Imaging modalities provide several means of observing brain changes due to AD such as deposition of amyloid plaques and tau tangles as well as functional and structural changes that occur due to the disease. PET imaging techniques combined with radiolabeled tracers specific to the target are used to scan Aβ accumulation and brain NFTs. PET scans operate on the principle that positron-emitting radiolabeled tracers accumulate in a region of interest which are then detected by scintillation detectors [39]. The radioligands are injected through a bolus injection, followed by a waiting period to allow for uptake by brain tissue. The degree of uptake is measured in terms of standardised uptake value ratio (SUVR), reflecting the amount of target present. Imaging of Aβ deposition is conducted with amyloid PET [40]. Amyloid PET is one of the recommended diagnostic biomarkers of AD. It was initially developed with carbon-based tracers [11C] such as Pittsburgh Compound B (PiB). However, fluorine-based tracers [18F] such as florbetapir, florbetaben, and flutmetamol have become more widely used due to their extended half-life of nearly 110 minutes compared to 20 minutes for [11C] tracers. Despite being an enormously informative AD biomarker tool, amyloid PET imaging still suffers several technical limitations. Some of these include poor understanding of its relationship with cognition, issues with the choice of reference region as well as harmonization across studies and tracers [40].

Development of Tau PET imaging for measuring tau burden has also seen considerable progress, although its reliability is still under investigation [40]. Structural imaging techniques such as MRI provide visualisation of structural brain changes (e.g., cortical thinning, hippocampal atrophy) considered as markers of neurodegeneration

likely to be detected in the later stages of the disease [14]. Functional imaging such as the FDG-PET is used as a marker of neuronal injury and neurodegeneration by reflecting brain neural activity which is usually impaired before changes in brain structure are detectable [41].

In summary, neuroimaging biomarkers have shown immense resourcefulness in AD diagnosis, staging progression and predicting its likely course. However, one of their limitations in application is that they are expensive to obtain.

2.3.3 Cerebrospinal fluid biomarkers

CSF is considered an ideal milieu for evaluation of AD biomarkers given its direct interaction with the interstitial fluid enveloping the brain, making it possible to reflect pathophysiological changes in AD [42, 43]. CSF samples are obtained through a lumber puncture procedure. The three most studied and validated CSF biomarkers of AD are amyloid-beta 42 (Aβ42), phosphorylated tau (p-tau) and total tau protein (t-tau) [30]. The level of amyloid CSF Aβ42 or Aβ42/Aβ40 ratio is used to indicate amyloid pathology and p-tau level is used to indicate tau abnormality. There is a significant reduction of CSF Aβ42 levels in AD, reflecting its accumulation in the brain, and a notable increase in p-tau levels indicating accumulation of NFTs [42, 44]. Increased levels of CSF t-tau are used as a biomarker of neurodegeneration or neuronal injury [44]. Emerging evidence has shown that in the earliest stages of AD, Aβ42 abnormality is evident in the CSF first, before it is detectable on the amyloid PET and before neurodegeneration appears [45].

In summary, CSF biomarkers are highly informative and recommended AD biomarkers, albeit cut-off values may vary between laboratories and the procedure of sample collection is invasive.

2.3.4 Blood-based biomarkers

Blood is a commonly used biological sample in research and clinically. The process of obtaining blood samples is safe and minimally invasive. Blood tests are ubiquitous in clinical practice and can be conducted in a variety of settings, including primary care, community-based medicine centres as well as patients' homes. Consequently, blood-based biomarkers for AD are highly attractive, as they have the potential to provide a simple, low-cost, minimally invasive as well as widely available method of AD diagnosis compared to imaging and CSF-based markers [46, 47]. Diagnosis of AD based on biomarkers in blood is promising owing to evidence suggesting that the disease presence may be reflected in blood [21, 30, 48]. This may be made possible as a result of the normal absorption of CSF and sufficiently small size fragments of proteins across the blood-brain barrier into blood [48]. This may be further enhanced by the compromise of the blood-brain barrier integrity in AD [49-52], making it possible for biochemical changes in the brain due to the disease to reflect in circulating blood [53]. However, development of blood-based AD biomarkers presents several challenges alongside opportunities. One of the major difficulties is that blood is a highly complex biofluid, hence several events can cause a change in its biochemical composition [47]. Fortunately, the complexity also provides opportunity for exploration of further biomarkers of the disease beyond the conventional amyloid and tau markers. Another major challenge is that although biomarkers of AD may be present in blood, their concentrations may be ultra-low, hence the need for highly sensitive biosensors [46] . The difficulty is being overcome with advancements in ultrasensitive biosensing technologies with which biochemical features in blood can now be simultaneously sampled, providing the availability of rich high-dimensional array of biochemical measures [54, 55]. With the availability of these highly complex multidimensional data, another challenge is to discover useful patterns (e.g., biomarkers) from such complex

data. Advanced data analysis methods such as machine learning undertaken in this project, as discussed subsequently, are aiding to overcome this difficulty.

The challenges notwithstanding, the short-term and long-term benefits of AD diagnosis based on blood biomarkers are apparent. Emerging consensus is that blood-based biomarkers could serve as a tremendously useful first-line screening tool in a multistage diagnostic framework for AD prior to conducting a PET scan or CSF-based analysis [46, 47, 55]. This can meet the immediate need in clinical trials recruitment to limit high negative screening failure rates. In the long term, it can also meet the scalability needs required for primary care settings as well as population-based screening that may ensue when treatment becomes available. Furthermore, since AD is a complex polygenic disease, amyloid and/or tau aggregation do not occur in isolation of other relevant molecular or cellular pathophysiological mechanisms. Blood-based biomarker analysis may assist to elucidate these interactions to enable a more comprehensive understanding of the disease to aid the development of suitable interventions [56, 57].

2.4 Diagnostic performance metrics

To measure the performance of a diagnostic test, including with biomarkers, use of relevant metrics is required. Diagnostic performance metrics provide a means of clinically grading the quality of a diagnostic test. An ideal diagnostic test correctly identifies all individuals with disease (referred to as ‘positive’ as determined by the reference standard test) and all disease-free individuals (referred to as ‘negative’ as determined by the reference standard test). In other words, a perfect test is never positive in a subject who is disease free and is never negative in a subject who is diseased in reality. Most diagnostic tests fall short of this ideal in practice. Thus, diagnostic performance measures are normally used to compare alternative tests to select one over the other, depending on the use case or to aid interpretation of the test

outcome. Some of such metrics include sensitivity, specificity, positive and negative predictive values. Other commonly used measures especially in machine learning related domain include F-score, area under receiver operating characteristics curve, accuracy, and Matthew's correlation coefficient. In practice, multiple metrics are usually determined to provide a more robust view of performance [58]. However, depending on the context of use, certain metrics may sometimes be considered more valuable.

Fundamental to the computation of diagnostic performance metrics are some terms such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) illustrated in Table 2.1. True positives refer to subjects who have the disease according to the reference standard also known as the 'gold standard' [59] and correctly diagnosed as positive according to the test being evaluated also referred to as index test. False positives denote disease free individuals misdiagnosed as positive by the index test. True negatives denote individuals who are disease free and correctly diagnosed as negative by the index test. False negatives refer to disease subjects misdiagnosed as negative by the index test. It is worth noting that the gold standard (the best single preferred test or a combination of tests for diagnosing a condition [60]) for AD diagnosis is a direct assessment of brain tissue at autopsy [14].

Another key term is prevalence. Prevalence denotes the proportion of the total population under consideration that has the disease.

It is also worth mentioning that performance metrics discussed here are in the context of their application to a binary state classification (e.g., diagnosis of disease versus no disease).

Table 2.1. Confusion matrix.

Index test	Reference standard	
	Positive	Negative
Predicted positive	TP	FP
Predicted negative	FN	TN

2.4.1 Sensitivity and specificity

The sensitivity (also known as recall or true positive rate) of a test is its probability of correctly identifying positive cases solely from among those who are known to be positive [58]. Similarly, specificity (or true negative rate) is the probability of the test to correctly identify negative cases from among those who are known to be negative. For instance, a test with 90% sensitivity detects 90% of disease cases (i.e., true positives) but 10% of the cases go undetected (i.e., false negatives). Similarly, a test with 90% specificity correctly identifies 90% of disease-free cases as negative (i.e., true negatives) but 10% of the cases are incorrectly identified as positive (i.e., false positives). The ideal situation is a sensitivity and specificity of 100%.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.2}$$

Sensitivity and specificity are an intrinsic property of a test and hence independent of the population of interest (e.g., prevalence of disease). However, they are dependent on the cut-off threshold for a test [61]. As a result, there is a trade-off between the sensitivity and specificity of a given test. A highly sensitive but low specificity test is typically useful as a screening test to identity individuals who may have disease [61].

The identified individuals may then be further subjected to a less sensitive but highly specific test to eliminate the false positives.

2.4.2 Positive and negative predictive values

Positive predictive value (PPV), also referred to as precision, is the probability that an individual identified as positive by a test is in fact positive. It is computed as the proportion of subjects with a positive test outcome that is actually positive as determined by the reference standard. Similarly, negative predictive value (NPV) is the probability that an individual diagnosed as negative by a test is indeed negative.

$$PPV = \frac{TP}{TP + FP} \tag{2.3}$$

$$NPV = \frac{TN}{TN + FN} \tag{2.4}$$

While sensitivity and specificity are useful in evaluating the intrinsic performance of a test (i.e., independent of disease prevalence) during development, the utility of PPV and NPV is in evaluating the value of the test in clinical practice [60, 61]. For example, in real life practice, the index test is conducted first, and result of the reference standard is unknown. To evaluate this index test during use by a clinician, PPV tells how much of the test positives and test negatives are true positives and true negatives, respectively, after testing the cases based on the reference standard. It is pertinent to note that PPV and NPV are dependent on disease prevalence in the population of interest.

2.4.3 Accuracy

Accuracy is defined as the probability of correctly identifying a random case. It is estimated as the ratio between the number of correctly identified cases and the total number of examined cases as shown in (2.5). Although accuracy is a frequently used

metric, it can be unreliable (overoptimistic) especially when the positive and negative cases are highly unbalanced [62].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

2.4.4 F-score

F-score or F-measure is computed from the precision and recall of a test. It is intended to provide a single metric that combines those measures. F_1 score (F_1), the most used form of F-measure, is defined as the harmonic mean of precision and recall.

$$F_1 = \frac{2}{precision^{-1} + recall^{-1}} = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.6}$$

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.7}$$

F_1 ranges in the interval [0,1], with 0 as the worst value and 1 as the best value. The value is minimum when TP=0 (i.e., when all the positive samples are misclassified) and maximum when FN=FP=0. In the situation where TP=FP=FN=0, F_1 is undefined, however, it can be set to 1 [63]. In the event that TP=0, FP>0, and FN>0 the value of F_1 from (2.6) remains undefined, but using (2.7), its value is zero. One of the drawbacks of F1 score is that it does not account for the true negatives [64] and is prone to bias due to class imbalance [63].

2.4.5 Matthew's correlation coefficient

Matthew's correlation coefficient (MCC) is an alternative global performance metric. It is interestingly a measure that is unaffected by the issue of imbalance between the positive and negative cases in the examined population [63], unlike accuracy and F-measure, as well as area under receiver operating characteristics curve. It generates a high score only if the test was able to correctly detect the majority of positive cases

and the majority of negative cases [65, 66]. Furthermore, it is invariant if the positive and negative instances are swapped. MCC can be computed as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.8)$$

It ranges in the interval [−1, +1], with values -1 and +1 corresponding to perfect misidentification and perfect detection, respectively, while MCC=0 is the expected value for a random occurrence.

2.4.6 Area under receiver operating characteristics curve

Accuracy, F score and MCC can be computed when a specific cutoff threshold for the confusion matrix is set. Area under the curve (AUC) of receiver operating characteristics (ROC) is a popular metric that is applicable when a single cutoff threshold is unavailable [63]. ROC curve is obtained from a plot of sensitivity on the vertical axis and (1-specificity) on the horizontal axis for all possible thresholds [67, 68] (see Figure 2.3). AUC value lies in the interval [0,1], with the value 1 denoting a perfect test and value 0.5 denoting a random (worthless) diagnosis. One of the advantages of the AUC is that it is objective, requiring no choices of parameter values, hence the same result is obtained from the same population of interest [68]. One of the flaws of the measure however is that it is sensitive to class imbalance and difficulty in comparing performance when ROC curves cross [69]. Consideration of both area under the precision-recall curve, an alternative measure more informative than AUC in imbalanced class scenarios, and AUC is recommended when a specific confusion matrix threshold is not available [66, 70].

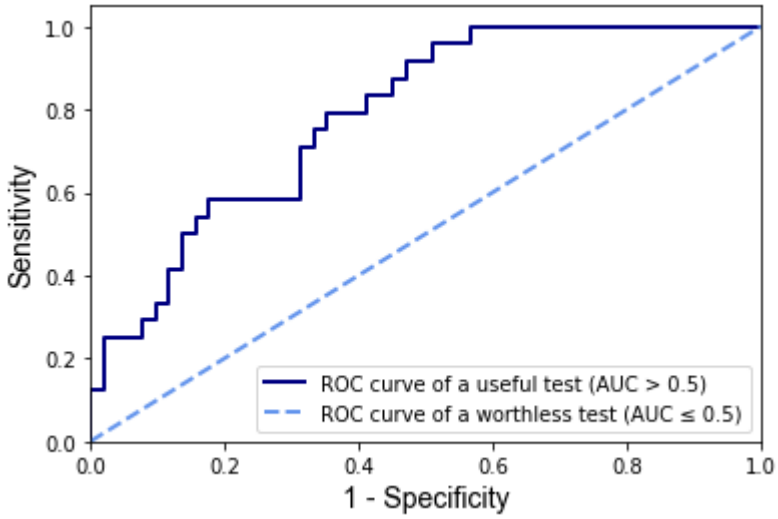


Figure 2.3. A sample ROC curve.

2.5 Machine learning methods

Due to the technological advancements in the processing power of computers and growing availability of large amount of data often referred to as big data, use of machine learning methods to solve computational problems have become a commonplace. Machine learning is a branch of artificial intelligence (AI) that enables computers to automatically detect patterns (learn) from available data (training examples), to gain descriptive knowledge and/or make predictions on new data based on the learned pattern [71, 72]. Machine learning is used to tackle a variety of complex computational tasks by learning from data, rather than following pre-programmed rules, which may be inefficient or too difficult to achieve. In other words, machine learning algorithms can learn from data without being explicitly programmed [73]. Learning in this context can be defined in many ways. One of the popular and foremost definitions is that a computer programme is said to learn from experience (i.e., data), codified with respect to some task and performance measure, if its performance improves with experience [74]. The data could be in the form of digitised human-labelled training sets or other kinds of information obtained through machine-environment interaction [75]. Some of the application areas of machine learning

include data mining (e.g., bioinformatics), email filtering, and speech and image recognition. Machine learning is usually categorised into two main types: supervised learning and unsupervised learning. Another but less common type of machine learning is reinforcement learning.

2.5.1 Unsupervised learning

Unsupervised learning is a form of machine learning that involves learning from unlabelled data. Unlike labelled data that consists of a set of observations tagged with at least one label, unlabelled data do not contain informative or desirable tags. Unsupervised learning algorithms target to uncover hidden patterns or structure in unlabeled data. Popular application areas include clustering (e.g., K-Means, Hierarchical Cluster Analysis), data visualization and dimensionality reduction (e.g., PCA and T-SNE), association rule learning (e.g., Apriori and Eclat).

Clustering

Clustering is a machine learning technique that categorises unlabeled data into groups (i.e., clusters) based on some inherent similarity among them [76]. Therefore, the goal of clustering algorithms is to find distinct clusters that exist within a given dataset such that samples assigned to the same cluster have more similarity according to specific metrics than those in different clusters. Clustering has wide applications in practice, including bioinformatics and computational biology. Because the notion of “cluster” has no precise definition [77], there are several clustering methods, each of which uses a different induction principle. Clustering approaches may be categorised into hierarchical, partitional, density, grid and model-based methods [78, 79].

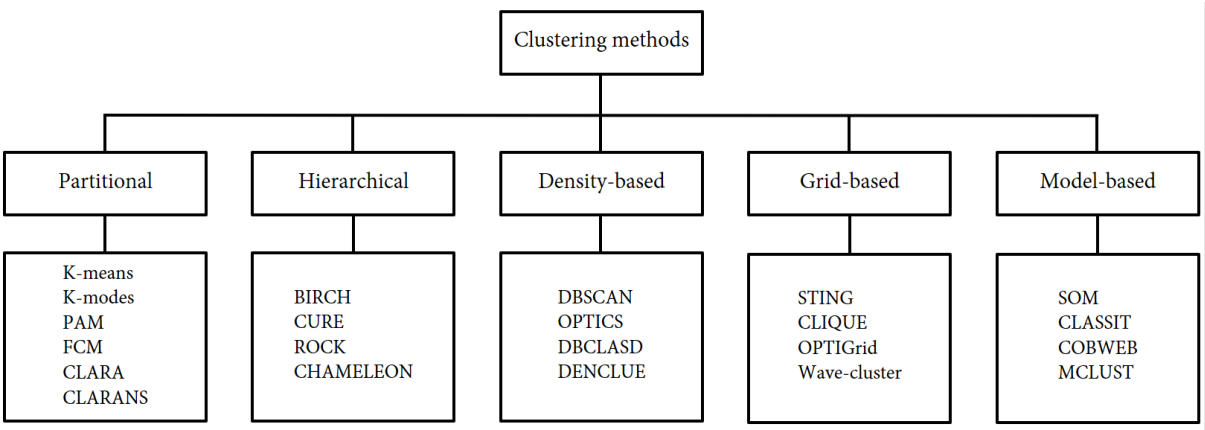


Figure 2.4. An overview of clustering taxonomy of clustering approaches [80].

Hierarchical clustering methods

In hierarchical clustering methods, clusters are constructed by iteratively partitioning the data points in a hierarchical manner, following a top-down or bottom-up approach also known as a divisive or agglomerative method, respectively. An agglomerative hierarchical clustering begins with each instance representing a cluster of its own. Then clusters are successively merged until a stopping criterion is achieved (usually the desired number of clusters, k). In contrast, a divisive hierarchical clustering starts with all instances belonging to one cluster and then successively splits the cluster until the stopping criterion is met. The hierarchical methods usually lead to formation of dendrograms as shown in Figure 2.5. The merging or splitting of clusters is performed according to some similarity measure. Hierarchical clustering can be further grouped into categories such as single-link clustering [81], complete-link [82] and average-link clustering [83], depending on the method of similarity measure. The main drawback of hierarchical clustering is that once a merge or split is performed, it cannot be reversed. Some examples of hierarchical clustering algorithms include BIRCH [84], CURE [85], ROCK [86] and CHAMELEON [87].

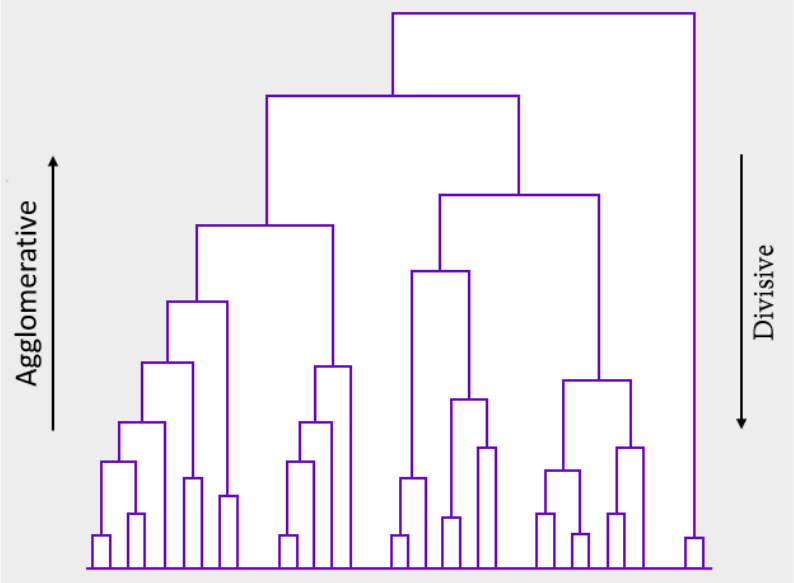


Figure 2.5. Hierarchical clustering dendogram [76].

Partitional clustering methods

Contrary to hierarchical clustering, partitional clustering methods, illustrated in Figure 2.6, allocate data instances to k clusters based on a similarity criterion, without constructing any hierarchical structure. Initial partitions are created and then objects are recursively relocated from one partition to another to form a predefined k number of clusters. Examples of partitional algorithms include K-Means [88] which is the most widely used in practice due to its simplicity, speed and ease of interpretation, K-Modes [89], PAM [90], FCM [91], CLARA [90] and CLARANS [92]. For instance, K-Means algorithm partitions a dataset into predefined k number of clusters by identifying k centroids and assigns each data point to the nearest centroid, such that the mean squared distance between each instance and its closest centroid is minimised. The process consists of the following steps, given k:

i. Select k random data points as the initial centroids
ii. Assign each instance to the closest centroid
iii. Re-compute the centroid of each cluster
iv. Repeat (ii) and (iii) until the centroids become stable

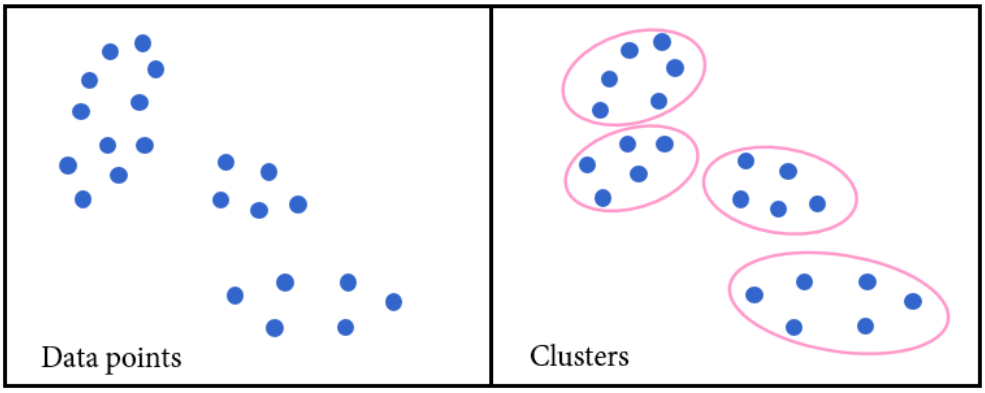


Figure 2.6. Partitional clustering approach [76].

The k-means algorithm is susceptible to local optimum solution as it is sensitive to the initialisation of the centroids as well as noise and outliers. Like other partitional clustering algorithms, k means is efficient on datasets that have isotropic clusters and not applicable when mean is unknown [77], such as non-numeric attributes. However, the algorithm is attractive due to its many benefits [77, 93, 94], including: linear complexity compared to the nonlinear complexity of hierarchical clustering; speed of convergence, offering no limitation on the size and dimensionality of data sets; adaptability to sparse data; and ease of interpretation.

Density-based clustering methods

These clustering methods work with the assumption that the data points belonging to each cluster are drawn from a particular probability distribution and the overall distribution of the data is a mixture of diverse distributions [76]. Therefore, the objective of these clustering methods is to identify the clusters and their distribution parameters, such that the probability of the data points generated by the cluster and parameters is maximised. Under this approach, a given cluster continues to be grown as long as the density (i.e., number of data points) in the neighbourhood exceeds a certain threshold. In other words, the density within a given boundary must contain a minimum number of instances. These methods can discover clusters of arbitrary shapes and provide

protection against the influence of outliers. Example algorithms include DBSCAN [95], DBCLASD [96], DENCLUE [97] and OPTICS [98].

Grid-based clustering methods

Grid-based methods split the data instance space into a finite number of cells to form a grid structure. The approach consists of the following steps:

1. First, the user specifies k number of grid cells, which is usually far less than the dataset size.
2. Then the algorithm assigns each data object to the appropriate grid cell and computes the density of each cell.
3. Low density cells whose number of data points is below a certain threshold is then eliminated.
4. Finally, clusters are formed by merging adjacent high-density cells.

The main strength of this approach is its fast-processing time. Some of the algorithms that use grid-based methods include CLIQUE [99], OPTIGrid [100], STING [101], and WaveCluster [102].

Model-based clustering methods

Model-based methods seek to optimise the fit between the data and a predefined mathematical model. The approach assumes that the data is generated by a mixture of underlying probability distributions. Whilst traditional clustering methods simply identify groups of data instances, model-based clustering methods in addition identify characteristic descriptions for each group, thus each group represents a concept or class. The most widely used induction algorithms include decision trees and neural networks such as COBWEB [103], CLASSIT [104] and SOM [105].

2.5.2 Supervised learning

Supervised learning is the type of machine learning that infers a set of rules from labelled training examples with the goal of creating a model that generalises to other instances. In supervised learning, each training instance comprises the explanatory variables and response variable, otherwise known as the predictors (or attributes) and labels, respectively. The supervised learning algorithm learns the mapping function between the attributes and labels during training. During prediction, the attributes of unseen data are supplied as input to the trained model and then the model predicts the labels. There are two categories of supervised learning algorithms: regression and classification algorithms. Whilst regression algorithms work with continuous labels, classification algorithms work with discrete labels.

State-of-the-art supervised learning algorithms include, but not limited to, support vector machine (SVM), artificial neural network (ANN), decision trees, logistic regression, k nearest neighbours (KNN), variants of discriminant analysis (e.g., LDA), Naive Bayes and ensemble algorithms (e.g., random forests and XGBoost).

Decision tree

Decision tree [106] is a nonparametric machine learning model that may be applied to both classification and regression tasks. The model is constructed using two elements; nodes and branches to form a tree-like structure as shown in Figure 2.7. The structure is formed during training by recursively evaluating all attributes in the training data in order to select nodes (i.e., features) that best split the data according to the sample labels. Thus, an instance is exclusively assigned to each branch based on some condition (like threshold or category) from the attribute's values.

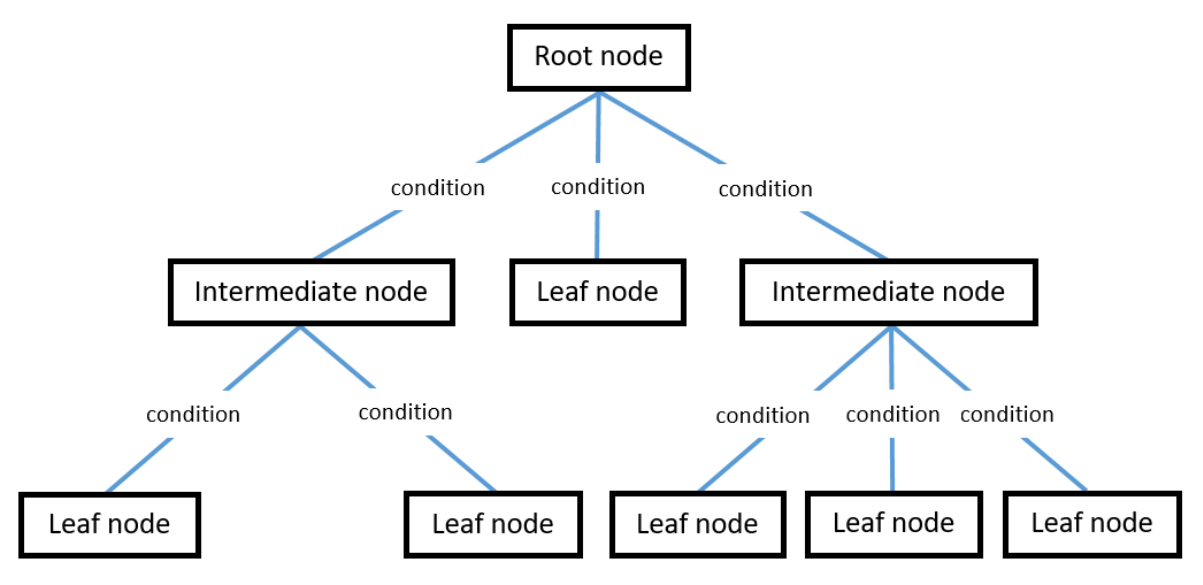


Figure 2.7. A simple decision tree.

Decision tree nodes may be categorised into root, intermediate (branch) and leaf nodes. Root node is the feature that begins the graph. Normally, the feature should best split the data into their respective labels. Intermediate nodes are located between the root and leaf nodes, whereas leaf nodes are the final nodes of the tree, where the prediction of a label is made. Performance of decision tree is dependent on how well the selected nodes split the data accordingly. Power of a node to split the data is usually determined using different proposed metrics such as residual or mean squared error in the case of regression, and Gini index or entropy in the case of classification. However, no single universally superior metric is known. Therefore, metric selection remains an important part of building a decision tree model.

Random Forest

Random forest [107] is one of the most popular ensemble learning algorithms for classification and regression. Ensemble techniques combine several base models to construct a single optimal model. Thus, random forest algorithm consists of an ensemble of decision trees, where each tree is constructed with a bootstrap sample of the training data, and at each split in the learning process, the candidate feature is a

selection from a random subset of the features. Therefore, random forest algorithm applies the general technique of bagging [108] and random selection of feature subset sometimes referred to as feature bagging. The combination of these two techniques help to achieve a model with superior performance compared to a single tree model. This is because the combined techniques can create a model with low bias and low variance, by aggregating predictions over a large ensemble of low bias and high variance trees with low inter-tree correlations.

Artificial Neural Network

Artificial neural networks (ANN) are one of the most popular machine learning algorithms applied in several disciplines. ANN consists of a collection of connected simple processing units called neurons, that can perform parallel computations for data processing and knowledge representation [109, 110]. Conceptually, ANN is loosely modelled after the structure of the brain, as a collection of neurons – each of which is connected to several others, from which it receives stimuli or to which it sends stimuli. The attractiveness of ANN comes from its remarkable information processing properties, including high parallelism, nonlinearity, fault and failure tolerance, and ability to handle imprecise information [111]. ANN models can be constructed as single layer or multilayer. Figure 2.8 shows a simple multilayer ANN. Neurons in a typical ANN architecture are usually arranged in three layers: input, hidden and output. The input layer consists of neurons that receive the original input data into the network for processing by subsequent layers of neurons. Every neural network must have an input layer. Hidden layer, otherwise known as intermediate layer, is any layer of neurons in between the input and output layers. At this layer, the neurons apply weightings to the information received from the input layer and directs them to the output layer after applying an activation function to achieve nonlinear transformations. Although hidden layers are optional, some networks contain many hidden layers. Such networks are

usually referred to as deep neural networks. The non-optional output layer collects inputs from the preceding layer(s) and produces the final output of the network, after applying weights and an activation function to the inputs.

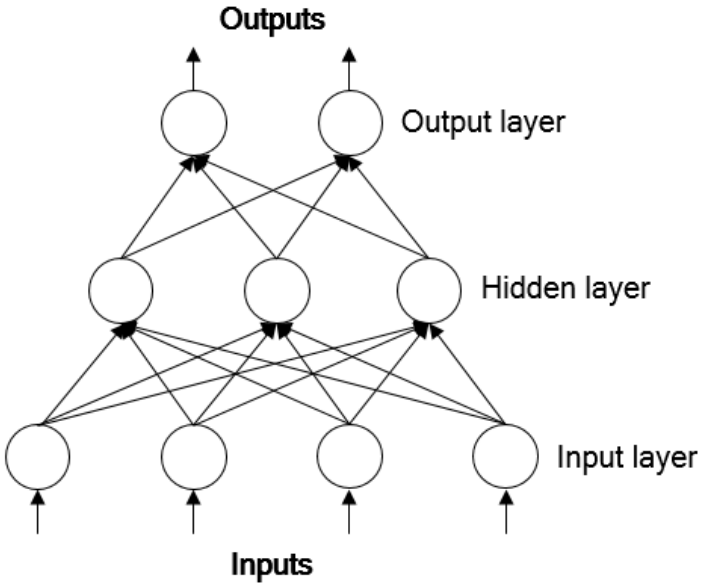


Figure 2.8. Three-layer feedforward neural network.

Support Vector Machine

Support vector machine (SVM) [112] is one of the most popular and powerful traditional machine learning algorithms, and can be applied to classification as well as regression tasks, albeit mostly used for classification. It is a powerful tool widely used in biomedical fields [113, 114]. Its popularity stems from several characteristics, including robustness to outliers, ability to handle high dimensional, small sample size as well as noisy data [115]. During training, SVM constructs a hyperplane or a set of hyperplanes in a high dimensional space to create class separation boundaries, such that the separation margins are maximised. The class of a new instance is determined by the side of the hyperplane(s) to which it is assigned by the trained model. Figure 2.9 illustrates a 2-class SVM classifier.

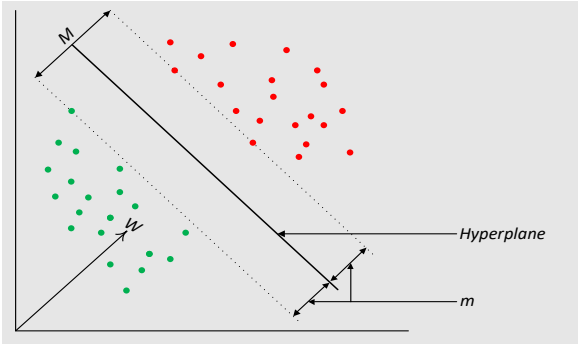


Figure 2.9. Mechanism of classification by SVM [116].

Consider a 2-category classification task with training instances consisting of N samples $(x_1, y_1), (x_2, y_2), \ldots, (x_{N-1}, y_{N-1}), (x_N, y_N)$, with input features $x_i \in \mathbb{R}^d$ and class $y_i \in \{-1,1\}$. During training, assuming the two classes are linearly separable, the goal of SVM is to define a hyperplane *h(x)* given by,

$$h(x) = x^T w + b = 0, \tag{2.9}$$

so as to induce a classification decision rule $D(x)$ that perfectly separates the samples into respective classes and maximises the margin $M(= 2m)$.

$$D(x) = sign(x^T w + b). \tag{2.10}$$

Finding such a hyperplane involves optimizing Mas,

$$\max_{w,b} M \equiv \min_{w,b} \frac{1}{2}\|w\|^2 \tag{2.11}$$

$$\text{subject to } y_i({x_i}^T w + b) \geq 1,$$

where b is a constant, d is the dimension of the data, w is a unit vector normal to the margin, and m is shown to be equal to $\frac{1}{\|w\|}$.

However, because it is often impracticable for the training instances to be perfectly separable, a concept known as soft margin [112] that permits misclassification error is implemented in practical SVM algorithms. In this case, the hyperplane is obtained by,

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \tag{2.12}$$

$$\text{subject to } \xi_i \geq 0, y_i({x_i}^T w + b) \geq 1 - \xi_i \; \forall i,$$

where ξ_i is a slack variable proportional to the amount by which the overlapping instance is on the wrong side of the hyperplane, C is the cost parameter for misclassification that can be tuned as a hyperparameter during model training.

The problem is a convex optimisation problem (in particular quadratic criterion subject to linear inequality constraints) that may be solved using the Lagrange function defined as

$$L_P = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i({x_i}^T w + b) - (1 - \xi_i)] - \sum_{i=1}^{N} m_i\xi_i \qquad (2.13)$$

The resulting w from the optimization is,

$$\widehat{w} = \sum_{i=1}^{N}\hat{\alpha}_i\, y_i x_i \qquad (2.14)$$

$\hat{\alpha}_i$ (Lagrange multipliers) being nonzero for instances i (known as support vectors) where the constraint $y_i({x_i}^T w + b) \geq 1 - \xi_i$ is met. Having found w, b may be determined from (2.9). Thus, the decision rule can be expressed as,

$$\widehat{D}(x) = sign\big(x^T\widehat{w} + \hat{b}\big). \qquad (2.15)$$

When the training data are not linearly separable by a hyperplane, SVM can transform the features to a new feature space where they become linearly separable, using a kernel function K. Several kernel functions exist, some of which include the polynomial, radial basis function and neural network kernels.

The kernel function simply computes dot products of features in the transformed space. One of such kernels is the polynomial kernel [117]. For example, given feature vectors v and z, a polynomial kernel is formulated as,

$$K(v, z) = (1 + v^T z)^r, \qquad (2.16)$$

where r is the degree of the polynomial.

Thus, for a SVM classifier with a kernel function, the solution for the hyperplane and decision rule are respectively modified as,

$$\hat{h}(x) = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i K(x, x_i) + \hat{b} \tag{2.17}$$

$$\widehat{D}(x) = sign\big[\hat{h}(x)\big]. \tag{2.18}$$

2.5.3 Model evaluation

The ultimate goal of data modelling with machine learning is to build models that can generalise to unseen data. A good model makes good predictions on future data, while keeping the problem of overfitting and underfitting minimal. (A good model is one that provides good predictions on unseen data, while keeping the problem of overfitting and underfitting minimal - can be deleted). Estimating the accuracy of a classifier is crucial for model selection and predicting future performance of a model [118]. Therefore, the need to apply the correct estimation method cannot be overemphasized. Several estimation methods are available in literature, such as bootstrapping and cross-validation.

Overfitting and underfitting problems

One of two problems, overfitting or underfitting, can arise during model fitting. Model overfitting is the creation of a model that fits too tightly or perfectly to a particular dataset, adding unnecessary complexity such that the models fail to generalise well to independent data. Such models are said to have high variance and may be remedied using techniques such as regularization, ensembling, dimensionality reduction, cross-validation, and increasing training data size where possible. Underfitting on the other hand occurs when a model fails to capture the underlying structure in the given dataset, thereby resulting in poor predictive performance on the training data as well as poor generalization [119]. Underfitted models are sometimes referred to as high

bias models. Model underfitting can be reduced using techniques including hyperparameter optimization, data denoising, and feature engineering. Preferred machine learning models are ones that strike a balance between variance and bias, often referred to as bias-variance tradeoff to achieve low bias and low variance. From the foregoing, desirable model evaluation techniques such as those discussed below can be used to identify models with low bias and variance.

Given a dataset to conduct a supervised learning task, the dataset is typically partitioned into at least two sets: a training set and test set. Predictive capability of the model on future data is estimated using the test set. However, due to the potential for bias and variance in estimation of the predictive performance, other data partitioning techniques (e.g., resampling methods such as cross-validation and bootstrapping) are often employed to obtain a more accurate estimate, especially when the sample size of available dataset is limited.

Cross-validation

Cross-validation is a technique that randomly partitions a given dataset into a number of complementary subsets, iteratively holding out one subset for validation or testing, and training the model on the remaining subsets. It ensures that the model is trained and tested with every instance in the data. (Its applications include model selection and parameter estimation). The predictive performance estimate is obtained as the aggregation of results from all the holdout sets [120]. In the case of regression, the aggregation is in the form of average value, whilst for classification, it is in the form of classification performance from the combined predictions across the entire subsets. Different variants of cross-validation exist, including k-fold cross-validation and leave-n-out cross-validation as well as their stratified and repeated variants. The difference between the k-fold and leave-n-out cross-validation is in how they conduct the

partitioning of the data into subsets. Given N sample size data, k-fold cross-validation partitions the dataset into k<N subsets, whereas leave-n-out cross-validation iteratively holds out n samples from the dataset as the test set and uses the remaining samples as training set. Figure 2.10 illustrates a k-fold cross-validation operation, with k = 10. Some of the most popular k-fold and leave-n-out cross-validation techniques include 10-fold cross-validation and the more computationally expensive leave-one-out cross-validation.

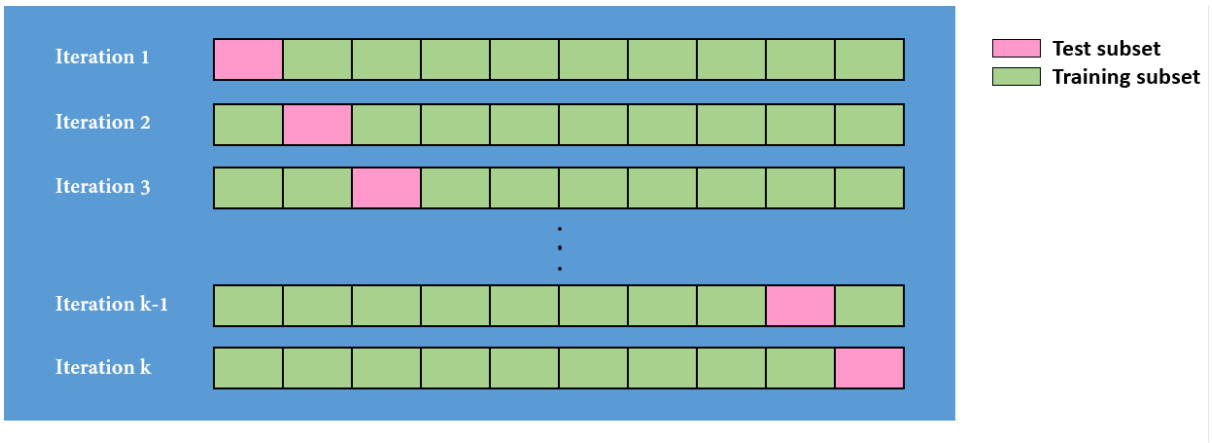


Figure 2.10. An illustration of k-fold cross-validation mechanism with k=10.

Albeit more computationally involving, leave-one-out cross-validation is not necessarily more beneficial compared to k-fold cross-validation. For example, although leave-one-out cross-validation is nearly unbiased, it produces high variance estimates [121].

Cross-validation methods can be further modified through the application of some additional techniques including stratification and repetition, to achieve improved robustness of performance estimation. These techniques give rise to stratified, repeated, and nested k-fold and leave-n-out cross-validation. In a classification task, the stratified technique ensures that an equal proportion of class categories are represented in the respective training and test sets during partitioning of data in the

cross-validation process. Just as the name implies, in repeated cross-validation, the cross-validation process is repeated a number of times, each time the data is randomly shuffled before being split. Nested cross-validation carries out additional cross-validation within a single round of cross-validation. This technique is usually applied to perform hyperparameter tuning during model evaluation. As a precaution, it is important that data processing before model fitting occurs on the training data subset assigned during cross-validation, rather than the full dataset, otherwise it leads to optimistic estimation of performance due to data leakage [122]. It is also good practice to provide a measure of the variance of the estimate, in the form of standard deviation, for example [123].

Bootstrapping

Bootstrapping [124, 125] is a statistical technique for estimating a quantity of a population by randomly resampling with replacement from the original dataset. This technique is usually applied in machine learning model evaluation, especially when there is limited availability of data, with the aim of deriving robust estimates of standard errors and confidence intervals of the predictive performance. However, accuracy of the bootstrap method depends on whether the underlying assumptions are met, such as independence of samples or whether sample size is large enough [126]. The bootstrap scheme involves the following procedure. Given original dataset of sample size N, randomly resample with replacement from the dataset to obtain N sample, then replicate the process B times to obtain B bootstrap samples. Fit the model with each bootstrap sample (as the training set) and estimate predictive performance using the out-of-bag, OOB, sample (i.e., observations in the original data are not present in the bootstrap sample). Final estimate is determined as the aggregate of the B estimates realised. Figure 2.11 shows a simple bootstrap mechanism illustrating pairs of bootstraps and OOB samples, realised from N=14 original data points with a two-class

balanced representation. Notice that due to sampling with replacement in the bootstrapping techniques, there are chances that a bootstrap may contain more than one instance of a single data point. However, an OOB sample can only contain a single instance of the constituent data points.

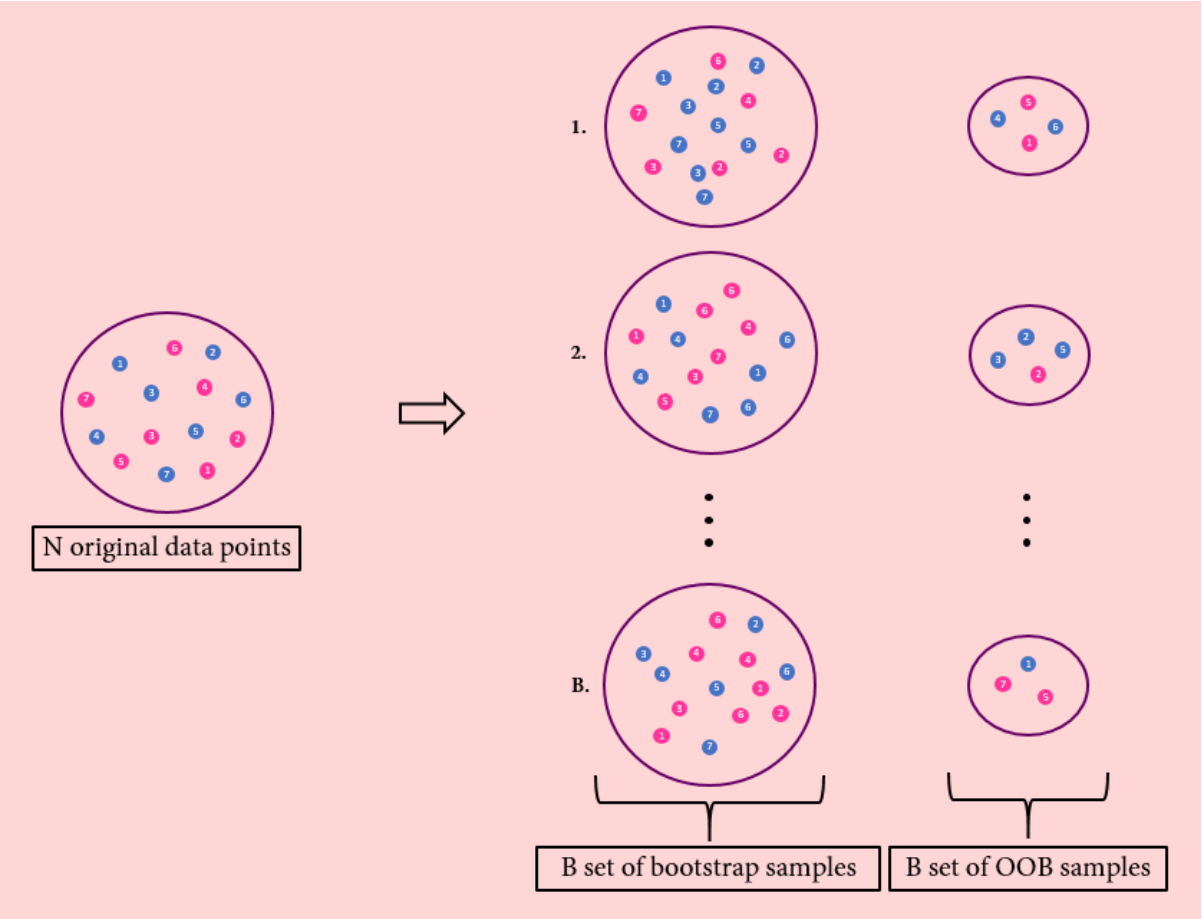


Figure 2.11. A simple bootstrap method.

2.5.4 Data preprocessing

Feature selection

Feature selection is usually an important data pre-processing step, particularly in traditional machine learning. The quality of the modelling data is one of the many factors that determine the success of machine learning on a given task. Knowledge discovery during model training is increasingly difficult if the input data consists of irrelevant or redundant features. Feature selection is the process of identifying and

removing as many of the irrelevant and redundant attributes as possible to retain only a subset of the features that ideally is necessary and sufficient to describe the target concept [127, 128]. However, it is important to highlight that there are various notions of relevance versus usefulness [129-131]. Moreover, features that are useless on their own or together can provide performance improvement when combined with others [132]. Other benefits of feature subset selection include minimising the curse of dimensionality and overfitting problems, reducing measurement and storage requirements as well as model training time, and facilitation of data visualisation [132]. For instance, reducing the dimensionality of the input data minimises the size of the hypothesis space. This causes the learning algorithm to operate faster, more effectively, and produce more compact results and interpretable representation of the underlying concept in the data.

Several methods of feature selection exist, and there is no particular best method. Instead, the effectiveness of a method depends on the specific problem setting. Within the context of supervised learning and in terms of the relationship between a feature selection algorithm and the inductive learning method used to infer a model, feature selection approaches can be broadly categorised into three: filter, wrapper, and embedded methods [133], as shown in Figure 2.12.

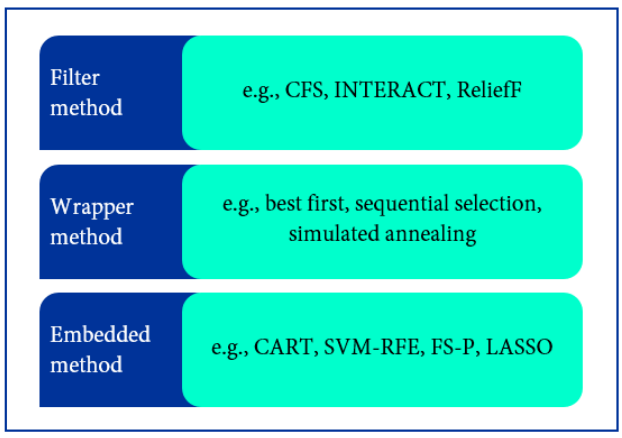


Figure 2.12. Main methods of feature selection.

Filters

Filters operate independent of any learning algorithms, unlike the wrappers and embedded method. The goal of the filter method is to filter out the undesirable features prior to induction. One of the major strengths of filters is that they provide a generic selection of features that are agnostic of the learning algorithm, as they are not tuned for or by any learning algorithm. Consequently, filters can be applied as a true data-preprocessing step to overcome curse of dimensionality and overfitting. Another major strength is that they are typically faster than the other methods, thereby making them easier to apply to high dimensional data. However, they sometimes fail to select the best features in terms of predictive performance, compared to the other two methods. Diverse types of filters approach feature selection problem differently, including selection criteria and evaluation metrics. Furthermore, filter method can be univariate or multivariate. Univariate filters such as InfoGain [134] are very fast and highly scalable but ignore feature dependencies. Multivariate filters such as CFS (Correlation-based Feature Selection) [135], consistency-based filter [136], INTERACT [137], ReliefF [138] and mRMR [139], on the other hand, take feature dependencies into account, but are slower and less scalable than the univariate approach.

Wrappers

In the wrapper method [130, 131], the feature subset selection algorithm wraps around the learning algorithm. It involves the use of the prediction performance of a learning algorithm applied as a black box (i.e., no knowledge of the algorithm is needed, just the interface) to assess the relative usefulness of feature subsets, as shown in Figure 2.13. The learning algorithm is run on the training dataset, usually divided into internal training and validation sets, with a different subset of the features applied at each run.

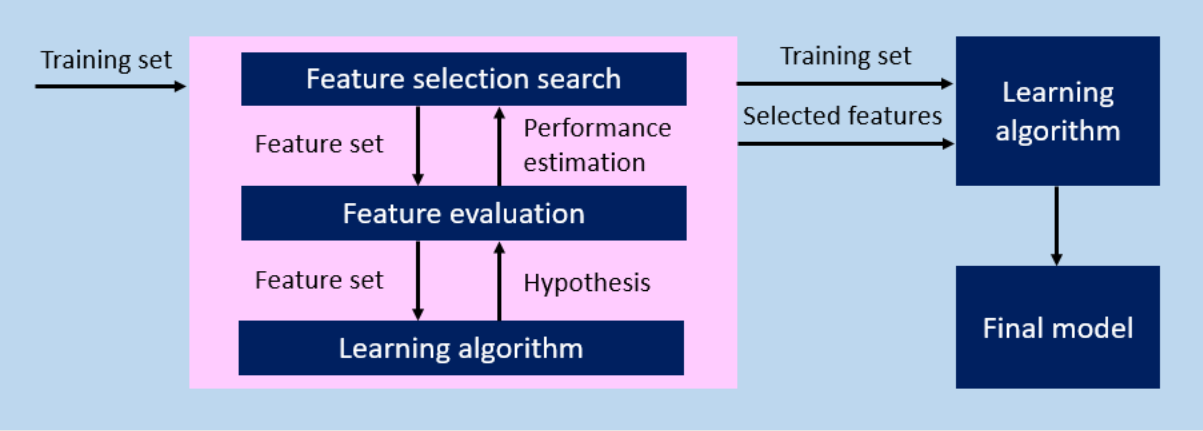


Figure 2.13. Wrapper method of feature subset selection (modified from [130]).

Feature subset with the highest prediction performance on the internal validation set is selected as the final set with which to train the learning algorithm on the full training dataset, to obtain the final prediction model. The wrapper method includes a search within the possible feature subset space and requires the use of a search strategy. For n features, the size of the search space is $O(2^n)$, thus it is computationally intractable to exhaustively search the entire space, except n is small. Thus, the main limitation of wrappers is the amount of computation needed to obtain a feature subset. Search strategies employed with the wrapper method can be grouped into three broad categories: exponential, sequential, and randomized [140]. Examples of these search strategies include the best-first [141], genetic algorithms [142, 143], simulated annealing [144], branch-and-bound [145], sequential forward and backward elimination [146], beam and bidirectional search [147]. Popular learning algorithms used in wrapper methods include decision trees, naive Bayes, least-square linear predictors, and support vector machines [132].

Embedded method

The embedded method of feature selection serves as a trade-off in speed and performance between the filter and wrapper methods. In the embedded method, feature selection is incorporated in the training of the learning algorithm. Thus, it is

faster compared to wrappers that must evaluate feature subsets iteratively and provides comparable performance. Examples of embedded methods include CART (classification and regression trees) [106], SVM-RFE (recursive feature elimination for SVM) [113], FS-P (feature selection–perceptron) [148], and regularised techniques such as LASSO [149] and elastic net [150]. Despite the advantages, one of the drawbacks of the embedded method is that, like wrappers, feature selection is usually biased toward the particular learning algorithm used. Therefore, the selected features may not be optimal for other learning machines.

Hybrid and ensemble methods

Some literatures also define additional feature selection methods, such as the hybrid and ensemble, derived from the earlier discussed methods [151-153]. Hybrid method can be developed as a combination of different methods, such as a filter and wrapper [116, 154], or techniques within the same method. The purpose is to harness the complementary strengths of the combined approaches. It uses distinct evaluation criteria at various stages of the selection process to improve prediction performance and computational cost. Ensemble method on the other hand aims to create a bucket of feature subsets and then produce an aggregate subset from it [155, 156]. It is developed to combat the issue of instability associated with feature selection algorithms in the face of data perturbations, especially in the case of high-dimensional datasets. The method is based on subsampling techniques where a specific feature selection method is applied to different subsamples and the resulting feature subsets are merged to create a more stable subset.

2.6 Summary

In this chapter, important relevant background concepts and techniques were discussed in depth. Some of these include the definition of AD, diagnosis of AD, use

of biomarkers and challenges, blood-based biomarkers, machine learning concepts and techniques. Some of the highlights of the discussion include the following. Typically, AD has a long preclinical phase before clinical symptoms become apparent. Biomarkers can be used as surrogate measures to indicate development or progression of the disease. A number of such biomarkers exist, and more are being studied. Blood-based biomarkers are some of the potential biomarkers attracting a growing interest due to some special benefits of using blood. Machine learning is providing several techniques to conduct advanced data analysis on complex data to drive research and innovation, including search for and potential clinical utility of blood-based biomarkers of AD discussed in subsequent chapters.

Chapter 3 Materials and Methods

3.1 General overview of data

Proteomics-based data used in this project were obtained from the publicly available Alzheimer's disease neuroimaging initiative (ADNI – http://adni.loni.ucla.edu) database, after securing approval through online application. ADNI study, which began in 2004 and organized in phases, is a longitudinal multicentre study funded by the National Institute on Aging, some pharmaceutical companies, and foundations. Among its objectives is the development of biochemical markers for early detection and monitoring of Alzheimer's disease, as well as enabling sharing of relevant scientific research data globally. ADNI phase one (ADNI-1) cohort consists of Alzheimer's dementia patients, mild cognitive impairment subjects, and elderly controls aged between 55 and 90. Alzheimer's dementia was diagnosed based on expert opinions according to the NINCDS-ADRDA criteria for probable ADD. A detailed description of the protocol may be found in the ADNI. The subjects were age matched and with about 16 years of education. Data obtained for this project included measurements of blood and CSF proteins, as well as demographic and diagnostic information of participants from ADNI-1. The blood plasma measurements comprised 190 proteins analysed on a Rule-Based Medicine platform. Forty-four (44) of the proteins were later excluded due to quality control issues and missingness, leaving 146 proteins. Data for three (3) other proteins including homocysteine, Aβ40 and Aβ42 were also obtained. CSF Aβ42 levels were measured using the Luminex Xmap platform. Clinical phenotypes such as apolipoprotein E epsilon 4 (APOE4) genotype of the participants were also obtained, and the demographic information included age and level of education. Table 3.1 shows the baseline characteristics of the 258

participants involved in the data, whilst a list of the 146 proteins is as shown in Table 3.2. CSF Aβ42 status for the individuals was obtained by dichotomizing their CSF Aβ42 levels as normal (high) or abnormal (low) according to clinically recognized threshold of 192pg/ml for the Luminex platform.

Table 3.1. Demographic characteristics of study subjects.

Demographics	Clinical Diagnosis		
	CTL	MCI	ADD
No. of participants at baseline (n)	58	198	102(+6 with no CSF record)
Age (mean, (SD))	75.11(5.77)	74.37(7.49)	74.86(7.88)
Gender, female (n, (%))	28(48.28)	65(32.83)	43(42.16)
Years of education (mean, (SD))	15.67(2.78)	15.80(2.99)	15.16(3.30)
APOE4 carriers (n, (%))	5(8.62)	106(53.56)	71(69.61)
Low CSF Aβ42 status (n, (%))	1(1.72)	147(74.24)	93(91.18)

CTL: Healthy control; MCI: Mild cognitive impairment; ADD: Alzheimer's dementia; n: Number of subjects; SD: Standard deviation. n-CTL, n-MCI at month 12 from baseline: 54, 136.

3.2 Software tools

Initial machine learning analyses at the beginning of this project were conducted using MATLAB programming language and WEKA workbench, as they were the legacy tools used by the research group. Subsequently, Python programming was used to conduct analyses due to its versatility for machine learning and data science, open-source license and large community base.

3.2.1 MATLAB

MATLAB is a programming platform developed for analysis and design of systems, enabled by the matrix-based MATLAB language that allows a natural expression of computational mathematics. It combines a desktop environment designed for iterative analysis and design processes with the programming language together with a live editor for creating scripts. MATLAB is used by engineers and scientist in a wide variety of sectors ranging from academia to industry to solve scientific and engineering problems, including machine learning, computational biology, control systems, signal

processing and communications. MATLAB statistical and machine learning toolbox provides functions and apps for analysing data, developing algorithms and models, and creating applications. With the toolbox, machine learning tasks including classification, regression and clustering can be performed. The toolbox also provides for other functionalities such as visualisation, dimensionality reduction and feature selection.

3.2.2 WEKA workbench

WEKA workbench (https://www.cs.waikato.ac.nz/~ml/weka/index.html) provides a collection of machine learning algorithm implementations and data pre-processing tools with an interactive interface to enable a quick exploration of existing machine learning methods, without writing any programming code. The data pre-processing tools provides support for data preparation including data visualisation as well as algorithms for discretization, sampling and feature selection. With the workbench, it is easy to pre-process a dataset, feed it to a learning algorithm, analyse results and compare the performance from different methods to identify potential solutions.

3.2.3 Python

Python is an open-source high-level general-purpose programming language. It uses a simple and easy to read syntax, and supports multiple programming paradigms, including procedural, object-oriented and functional programming. It is one of the most popular programming languages. The first version of the language was released in 1991, and version 2.0 that introduced a number of features was released in 2000. In 2008, Python 3.0, a major revision of the language that is not completely backward compatible was released. Python was designed to be highly extensible with the use of modules. The compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications. In the machine learning and

data science world, Python has become the go-to language. Python's Scikit-learn module (https://scikit-learn.org/stable/) provides several simple and efficient tools for predictive data analysis including data pre-processing, classification, regression, clustering, dimensionality reduction and model selection. Python also provides several other supportive libraries such as Numpy, Pandas and SciPy to facilitate data manipulation. Anaconda distribution of python and code editor was used for writing and running codes.

3.3 Methods

Figure 3.1 illustrates a typical machine learning-based methodological framework for the identification of biomarkers applied in this research. This general framework is modified accordingly at distinct stages of the research to suit the specific research question at hand. The pipeline begins with the collection of raw data, followed by feature extraction and pre-processing. Feature extraction and some levels of pre-processing were already conducted on the raw data to produce the study data obtained from ADNI. At the feature extraction stage, protein measurements were extracted from blood samples using biosensors. The extracted protein measurements were then subjected to quality control protocols to validate the measurements. The obtained data were sometimes further pre-processed, for example, normalised or standardised to facilitate a faster convergence when fed to a machine learning algorithm. At the feature selection stage, a subset of the proteins is dropped due to redundancy or a perceived lack of usefulness. The selection is based on the outcome of an evaluation process that is usually based on a combination of filter and wrapper methods (i.e., hybrid method) or ensemble approach. Model development stage involved the selection of certain machine learning algorithms over the others and tuning of the selected model's hyperparameter(s) to obtain an optimised model. Finally, estimation for generalisation

performance of the model (i.e., learning algorithm + identified biomarkers) is conducted at the testing phase.

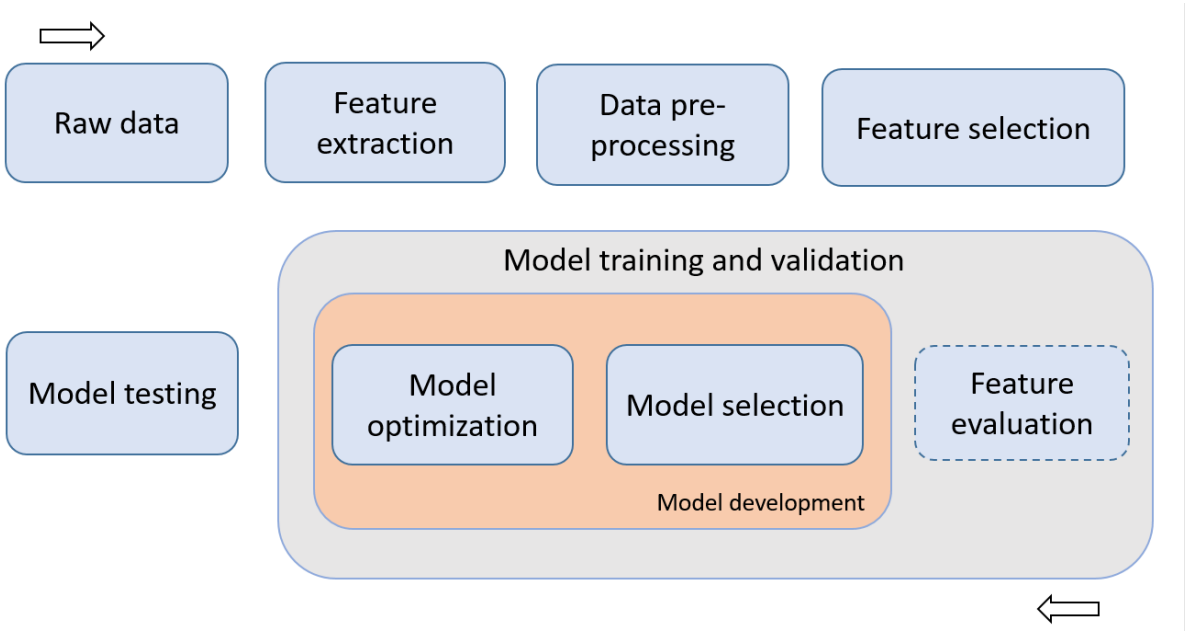


Figure 3.1. Typical methodological framework.

3.4 Summary

In this chapter, the materials and methods used in this research work were described. These include description of the data provider (ADNI), dataset, software tools (such as MATLAB, WEKA and Python) and general methodology applied in conducting machine learning-based analysis on the obtained dataset using the tools, as discussed in more detail in the following chapters.

Table 3.2. List of 146 plasma proteins obtained from ADNI.

Adiponectin	Interleukin-6 receptor
Agouti-Related Protein	Interleukin-8
Alpha-1-Antichymotrypsin	Interleukin-13
Alpha-1-Antitrypsin	Interleukin-16
Alpha-1-Microglobulin	Interleukin-18
Alpha-2-Macroglobulin	Kidney Injury Molecule-1
Alpha-Fetoprotein	Leptin
Angiopoietin-2	Luteinizing Hormone
Angiotensin-Converting Enzyme	Macrophage Colony-Stimulating Factor 1
Angiotensinogen	Macrophage Inflammatory Protein-1 alpha
Apolipoprotein A-I	Macrophage Inflammatory Protein-1 beta

Table 3.2. List of 146 plasma proteins obtained from ADNI (Cont.).

Apolipoprotein A-II	Macrophage Inflammatory Protein-3 alpha
Apolipoprotein A-IV	Macrophage Migration Inhibitory Factor
Apolipoprotein B	Macrophage-Derived Chemokine
Apolipoprotein C-I	Matrix Metalloproteinase-1
Apolipoprotein C-III	Matrix Metalloproteinase-10
Apolipoprotein D	Matrix Metalloproteinase-2
Apolipoprotein E	Matrix Metalloproteinase-7
Apolipoprotein H	Matrix Metalloproteinase-9
Apolipoprotein(a)	Matrix Metalloproteinase-9- total
AXL Receptor Tyrosine Kinase	Monocyte Chemotactic Protein 1
B Lymphocyte Chemoattractant	Monocyte Chemotactic Protein 2
Beta-2-Microglobulin	Monocyte Chemotactic Protein 3
Betacellulin	Monocyte Chemotactic Protein 4
Bone Morphogenetic Protein 6	Monokine Induced by Gamma Interferon
Brain Natriuretic Peptide	Myeloid Progenitor Inhibitory Factor 1
Brain-Derived Neurotrophic Factor	Myeloperoxidase
Calcitonin	Myoglobin
Cancer Antigen 19-9	Neuronal Cell Adhesion Molecule
Carcinoembryonic Antigen	Neutrophil Gelatinase-Associated Lipocal
CD 40 antigen	Osteopontin
CD40 Ligand	Pancreatic Polypeptide
CD5-Antigen-like Precursor	Peptide YY
Chemokine CC-4	Placenta Growth Factor
Chromogranin-A	Plasminogen Activator Inhibitor 1
Ciliary Neurotrophic Factor	Platelet-Derived Growth Factor BB
Clusterin	Pregnancy-Associated Plasma Protein A
Complement C3	Proinsulin- Intact
Complement Factor H	Proinsulin-Total
Cortisol	Prolactin
C-peptide	Prostatic Acid Phosphatase
C-Reactive Protein	Pulmonary and Activation-Regulated Chemo
Creatine Kinase-MB	Receptor for advanced glycosylation end
Cystatin-C	Resistin
Eotaxin-1	Serotransferrin
Eotaxin-3	Serum Amyloid P-Component
Epidermal Growth Factor	Serum Glutamic Oxaloacetic Transaminase
Epidermal Growth Factor Receptor	Sex Hormone-Binding Globulin
Epithelial-Derived Neutrophil-Activating	Sortilin
E-Selectin	Stem Cell Factor
Factor VII	Superoxide Dismutase 1- Soluble
Fas Ligand	T Lymphocyte-Secreted Protein I-309
FASLG Receptor	Tamm-Horsfall Urinary Glycoprotein
Fatty Acid-Binding Protein	T-Cell-Specific Protein RANTES
Ferritin	Tenascin-C

Table 3.2. List of 146 plasma proteins obtained from ADNI (Cont.).

Fetuin-A	Testosterone-Total
Fibrinogen	Thrombomodulin
Fibroblast Growth Factor 4	Thrombopoietin
Follicle-Stimulating Hormone	Thrombospondin-1
Glutathione S-Transferase alpha	Thymus-Expressed Chemokine
Growth Hormone	Thyroid-Stimulating Hormone
Growth-Regulated alpha protein	Thyroxine-Binding Globulin
Haptoglobin	Tissue Inhibitor of Metalloproteinases 1
Heparin-Binding EGF-Like Growth Factor	TNF-Related Apoptosis-Inducing Ligand
Hepatocyte Growth Factor	Transthyretin
Immunoglobulin A	Trefoil Factor 3
Immunoglobulin E	Tumor Necrosis Factor alpha
Immunoglobulin M	Tumor Necrosis Factor Receptor-Like 2
Insulin	Vascular Cell Adhesion Molecule-1
Insulin-like Growth Factor-Binding Protein	Vascular Endothelial Growth Factor
Intercellular Adhesion Molecule 1	Vitamin K-Dependent Protein
Interferon gamma Induced Protein 10	Vitronectin
Interleukin-3	von Willebrand Factor

Chapter 4 Identification of a Sparse Panel of Blood-based Biomarkers for Alzheimer's Disease Detection Using Machine Learning

4.1 Introduction

A considerable proportion of dementia patients remain undiagnosed because of inadequate access to diagnosis. Of those that receive a diagnosis, a considerable percentage may have received it late, when extensive cell damage would have occurred and when treatments are less effective. In view of this, it is thought that providing accessible diagnosis may decrease the burden of dementia, facilitate access to evidence-based pathway to treatment. It may also facilitate planning and timely receipt of suitable health and social care services [157].

Being that AD accounts for most dementia cases, research studies are investigating several putative AD biomarkers, including ones found in peripheral blood. Huge research efforts are being made to identify and validate AD biomarkers that are minimally invasive, simple to use, cost-effective and able to reliably discriminate target population in the light of the disease [33, 158]. Blood-based biomarkers may be more cost and time-efficient to assess AD, compared to the more established biomarkers from CSF and amyloid PET. Therefore, blood-based biomarkers can serve to complement CSF and PET markers. Although blood-based biomarkers have shown the potential to meet these targets, no single marker is reliable to provide sufficient diagnostic performance, in terms of sensitivity and specificity. Consequently, a number of research studies have investigated AD diagnostic performance of some blood biomarker panels using machine learning. This is due to the multivariate modelling efficiency of machine learning.

Among these studies, Ray et al. [159] identified an 18-biomarker panel that attained sensitivity and specificity values of 90% and 88% respectively, while a 30-biomarker panel was identified by O'Bryant et al. [160], which achieved sensitivity and specificity

values of 94% and 84%, respectively. Daniel et al. [161] identified 5 to 15 biomarker panels that detected AD with 74% sensitivity and 85% specificity. Using ADNI dataset, Doecke et al. [162] identified an 18-marker panel that identified AD with sensitivity and specificity values of 80%. A study by Guo et al. [163] obtained sensitivity and specificity values of 89.36% and 79.17%. Furthermore, Jammeh et al. [164] identified a panel of six blood biomarkers that was able to detect AD with sensitivity and specificity values of 85.4% and 78.6%, respectively.

Despite the progress, some of the identified panels of blood biomarkers consist of a large number of biomarkers or do not meet the recommended performance specification. Furthermore, there are difficulties with replicating results, due to many factors such as overfitting in model development [165]. In addition, some of the studies cannot be replicated because panels were identified using datasets that are difficult to access or are based on biomarkers that are not found in accessible databases. These challenges impede continued investigation of the utility of blood biomarkers in AD diagnosis and progress in identifying blood biomarker panels with clinical utility.

The main objectives of this research are to:

i. identify a sparse panel of adequately cross-validated blood biomarkers of AD that can discriminate between Alzheimer's patients and healthy controls with acceptable diagnostic performance of at least 80% sensitivity and specificity values [33], using a widely accessible and well characterized blood proteomic dataset;
ii. demonstrate a technique to evaluate robustness of the identified biomarker panel to facilitate future replication of results.

Realising a minimum number of biomarkers that provide high and reliable diagnostic performance may result in reduced complexity and cost of implementation of point of care diagnostic devices for AD detection.

4.2 Methods

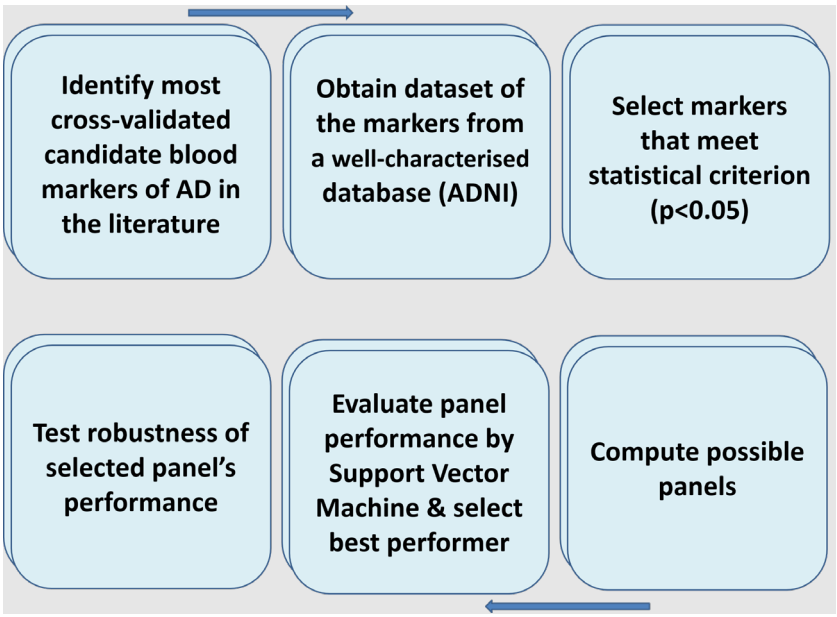


Figure 4.1. Description of methodology

The overall analytical pipeline adopted in this study is illustrated in Figure 4.1, and discussed in detail in the following sections. Briefly, it involves preselection of features from the study dataset based on literature and filter method of feature selection (use of p-value in particular), formation of possible panels, wrapper-based evaluation of the panels, selection of a final panel and investigating robustness of the selected panel.

4.2.1 Study data

Data used for this study were obtained from ADNI as described in Section 3.1. The subset of the data used for this analysis include measurements of 146 blood plasma proteins derived from a cohort of 112 Alzheimer's dementia (ADD) patients and 58 healthy controls (CTL) taken at ADNI-1 baseline. Data from four of the ADD subjects, including three that were diagnosed possible Alzheimer's disease and one diagnosed with mild level of confidence, were removed.

4.2.2 Feature preselection

First, a review of the literature was conducted to identify blood biomarkers with the most association with AD. A literature search was carried out electronically on the MEDLINE and Embase databases using the PubMed and Ovid interfaces. The top-ranking ones were preselected for analysis with the study data, whilst the rest were discarded. The probability distribution of each of the markers was then examined and normalised where necessary. As a filter-based feature selection approach, differential abundance between the two clinical groups (ADD and CTL) was analysed using Student's t-test. Then markers with statistically significant differences (p-value < 0.05) were selected as candidates for the identification of potential optimum biomarker panels from subsequent analyses.

4.2.3 Panel search

A brute-force search strategy was applied to generate and evaluate biomarker panels: beginning with single markers and using 5-element panels as the stopping criterion. That is, individual markers as well as all their possible combinations consisting of 2, 3, 4, and 5 markers were generated. Each of the panels was used for the classification procedure described in the following section. This method of panel generation is different from the usual methods seen in blood biomarker studies, where some sort of reductionist approach is often implemented. The drawback of such methods is that some potentially useful biomarker panels might be missed.

4.2.4 Classification and biomarker panel selection

A wrapper-based feature selection using a supervised linear kernel SVM classifier implemented in MATLAB was applied to identify an optimum panel of biomarkers that met the desired performance. SVM has been extensively applied in Alzheimer's research [166, 167]. Its popularity stems from a number of desirable characteristics,

including robustness to outliers as well as the ability to handle high dimensional, small sample size and noisy data [115] as is the case with the study data. The classifier algorithm was trained and tested with each of the panels generated as described in the preceding section using 10-fold cross-validation technique. This technique randomly partitions the applied dataset into 10 subsets and ensures that each subset is used for both training and testing. Cross-validation was applied to overcome model overfitting and to obtain a more realistic estimate of the model's classification performance. Clinical status of the individuals in the form of binary values was used as the class labels in the data (i.e., 0 and 1 represented CTL and ADD, respectively). APOE4 genotype was used as a covariate to each panel, since it has been established as one of the major clinical AD risk factors [15]. The training and testing of a model with each panel was repeated five times and the performance metrics were recorded per time. Performance was measured in terms of sensitivity, specificity, accuracy and area under the operating curve (AUC). The panel that showed high consistency in performance with sensitivity and specificity values greater than 80%, was selected for further evaluation of robustness.

4.2.5 Evaluation of robustness

In this evaluation phase, the 10-fold cross-validation was iterated one thousand times (with the training and testing subsets internally randomized each time) and the average performance recorded. The purpose was to rigorously investigate the robustness of the selected panel in detecting AD across the dataset. The percentage of times that the panel achieved sensitivity and specificity of at least 80% (i.e., loosely referred to as success rate in this analysis) was calculated. The success rate demonstrates an estimation of the robustness of the panel's performance. The panel that met the pre-specified performance threshold with a high success rate was selected as final. This

method was implemented to improve the feasibility of replicating the observed performance and facilitate further refinements of the existing panels.

4.3 Results

From the review of literature, 173 blood proteins associated with AD (excluding groups of microRNAs) were identified from 54 studies, from which 40 markers were most acknowledged. However, only 31 of these markers were available in the study dataset. Of the 31 proteins, 14 that are listed in Table 4.1 showed statistically significant difference between the ADD and CTL subjects. There were 3,458 candidate 2-5 marker panels generated from the 14 markers. The wrapper-based evaluation of the candidate panels with SVM classifier identified a panel of five markers including Alpha-1 microglobulin (A1M), Alpha-2 macroglobulin (A2M), Complement C3 (CC3), Immunoglobulin M (IGM), and Tenascin C (TNC) achieving high performance and success rate. This panel detected AD with average sensitivity, specificity, accuracy and AUC of 86.5%, 82.1%, 85% and 0.89, respectively. It also achieved a success rate of 77.8% in the robustness evaluation.

Table 4.1. List of candidate and selected blood biomarkers.

Candidate markers	Selected 5-marker panel
Alpha-1 microglobulin Alpha-2 macroglobulin Alpha-1 antitrypsin Apolipoprotein E Beta-2 microglobulin Brain natriuretic peptide Complement C3 Eotaxin-3 Immunoglobulin M Interleukin-3 Macrophage inflammatory protein-1 alpha Pancreatic polypeptide Tenascin C Vascular cell adhesion molecule-1	Alpha-1 microglobulin Alpha-2 macroglobulin Complement C3 Immunoglobulin M Tenascin C

4.4 Discussion

In this study, 5-biomarker panel (consisting of A1M, A2M, CC3, IGM and TNC) was identified for the discrimination between AD patients and control subjects from the ADNI cohort, whilst using APOE4 genotype as an additional feature. The panel was further evaluated rigorously for robustness to improve the replicability of results. Size of the panel was deliberately limited in consideration of the feasible complexity of the mutli-marker biosensing platform being developed in the BBDiag project to demonstrate a point care device. Each of the five markers is well associated with AD in the literature. A1M is a protein involved in inflammatory response [168] that has also been identified as a plasma marker of brain atrophy in AD [169]. The role of A2M in AD has been extensively researched; Bauer et al. [170] showed that A2M was present in amyloid plaques. Since then, it has further been linked to blood-brain barrier damage [171], hippocampal metabolism in early AD [172] and neuronal injury [173]. Complement C3 has been identified as a marker of brain atrophy in AD [169] and cerebral amyloid in non-demented elderly [174]. IGM has been identified as blood protein marker of neocortical amyloid-beta burden [175, 176]. TNC is an extracellular glycoprotein that has been linked to different biological processes, including inflammation and angiogenesis, which have an association with AD [177]. Both IGM and TNC have been linked to APOE4 genotype, a well-established risk marker of late-onset AD [178].

4.5 Summary

Notwithstanding the prospects of blood-based biomarkers to provide a low-cost and non-invasive method of AD detection to complement CSF and neuroimaging markers, no single blood biomarker is yet able to detect AD with a reliable performance. As a result, several studies have considered a combination (or panel) of markers using machine learning approach as it provides advanced methods of analysing complex

data. However, a large number of biomarkers are often needed to achieve a satisfactory detection performance. In addition, it is often difficult to reproduce reported results within and across different study cohorts due to overfitting of data and lack of access to the datasets used in the studies. In this chapter, an optimum panel (in terms of the least number to meet a clinically recognised diagnostic performance of 80% sensitivity and specificity) of blood biomarkers based on a widely accessible data set was identified. Key contributions of the study include the identified panel with reduced feature size and high performance, as well as the novel feature selection and model evaluation process implemented to reinforce replicability of findings. Despite the findings and benefits of the applied methodology, it is also necessary to: (1) involve a more exploratory approach to the initial selection of markers that includes all the available features in the study dataset and (2) put into consideration the performance of candidate panels in detecting earlier phase of the disease such as MCI. The next chapter attempts to address these amongst other perspectives.

[1]

[1] This chapter is a slightly modified version of “Identification of Optimum Panel of Blood-based Biomarkers for Alzheimer’s Disease Diagnosis Using Machine Learning” published and presented at the 40th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2018 and has been reproduced here with the permission of the copyright holder.

Chapter 5 Early Detection of Alzheimer's Disease with Blood-based Biomarkers Using Machine Learning

5.1 Introduction

As there is currently no cure for AD, clinical interventions being developed are aimed at individuals in the early (including preclinical and prodromal [179]) stages of the disease, when it is thought that treatment is more likely to be effective. AD is characterised by deposition of amyloid plaques in the brain, which are observable in vivo using amyloid PET or CSF biomarkers. However, despite progress with the development of amyloid based biomarkers for early AD diagnosis, they face some limitations [180-182]. Amyloid-based biomarkers provide limited information about the disease pathological aetiology and pathways [183-185]. In addition, tests based on these biomarkers are unable to identify individuals at risk of AD prior to a significant amyloid-beta deposition in the brain. Therefore, there is a need for biomarkers that have the potential to detect biological processes that precede brain amyloid-beta accumulation (amyloid pathology) during the disease's development. Such biomarkers may advance understanding of the disease, aid identification of individuals at the early disease stages and the development of new interventions.

Emerging findings suggest that AD is characterised by metabolic alterations [22] that may precede amyloid pathology [185]. Signatures of such metabolic abnormalities may therefore serve as biomarkers of early disease. Such biomarkers may be obtained from blood since blood has a rich metabolic information content. The use of blood is also attractive because blood biomarker-based test is relatively non-invasive compared to CSF and may be more cost-effective than PET imaging.

A number of studies have attempted to identify blood-based (non-amyloid) biomarkers of disease by profiling a large array of proteins in blood and examining their association with the disease [172, 178, 186], but this approach is difficult to apply in practice. One

of the most promising approaches is the use of machine learning techniques to find appropriate combinations of blood proteins that can achieve a reliable detection as machine learning makes it possible to fit multivariable data to a model by learning complex patterns from the data.

Several studies [160, 161, 163, 164, 187-194] have applied machine learning to develop classifiers to differentiate between AD subjects and healthy controls as discussed in the preceding chapter. Despite the promising results from these studies, nearly all the models (including those proposed in Chapter 4) were developed and evaluated using data from only cognitively healthy controls and AD dementia patients (i.e., subjects at the later stages of the disease). The models were not evaluated in individuals at earlier stages of the disease. Therefore, the panels underlying such models may not be suitable as biomarker signatures of early AD. This study extends the scope of the work presented in Chapter 4, in terms of approach to feature selection and model development. The main objective in this chapter is to develop a machine learning-driven method to identify potential blood biomarker panels of early AD based on non-amyloid proteins that have the potential to identify the disease prior to the accumulation of brain amyloid burden. In addition, the potential of existing machine learning-based methods to achieve early disease detection is assessed.

5.2 Methods

5.2.1 Study data

The study data consist of ADNI-1 baseline measurement of 146 plasma proteins derived from 58 CTLs and 108 ADD patients just like in the preceding chapter. However, they additionally contain month 12 records from 54 CTLs and 136 individuals with MCI who were later diagnosed with AD dementia within 10 years of follow-up. Data description has been previously provided in Section 3.1.

5.2.2 Data partitioning

To make optimal use of the available data while minimizing susceptibility of the proposed approach to overfitting problems, the study data were partitioned into two non-overlapping datasets; Datasets 1 and 2. Dataset 1 consists of baseline data from the ADD and CTL subjects. All existing methods evaluated in this study except [160] were originally developed based on Dataset 1. In the proposed approach, Dataset 1 was used to conduct a robust feature preselection and model development. The resulting models were further evaluated with Dataset 2. Dataset 2 consists of month-12 data from MCI and CTL subjects. It was used to assess the performance of the developed models (trained on the entirety of Dataset 1) for MCI vs. CTL classification. Models were trained with only Dataset 1 during model development using the entirety of it or its subsamples.

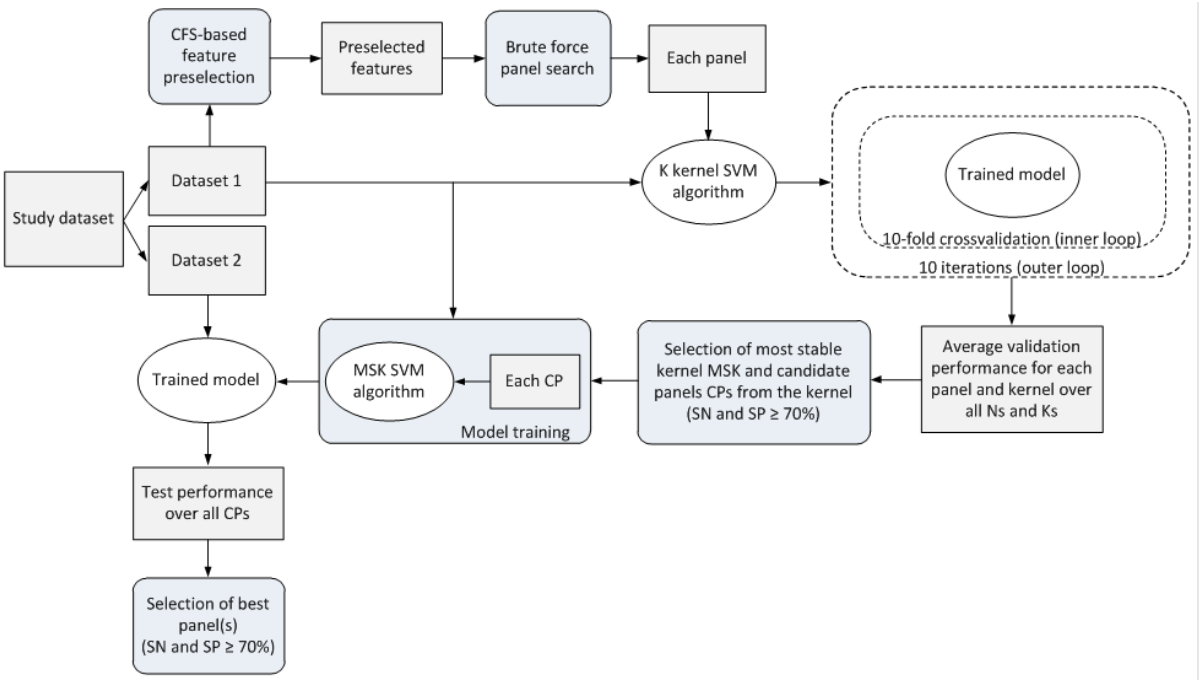


Figure 5.1. Overall framework for identification of novel putative biomarker panels and model development for early AD detection. K: Different kernels of SVM including linear, second- and third-degree polynomials, and radial basis function (RBF), respectively. MSK: Most stable kernel. A stable kernel is one that showed most moderate to high performance for most panels. CV: Cross-validation (CV). CP: Candidate panel. A candidate panel is one that meets the pre-specified performance criteria (SN and SP of at least 70%) in the model training and CV step.

5.2.3 Replication and evaluation of existing methods

Machine learning models reported in previous studies for the classification of ADD and CTL subjects (Dataset 1) were replicated. The models were evaluated using 10-fold cross-validation with the average performance of the models taken after 10 repetitions. In 10-fold cross-validation, the dataset D is randomly split into 10 mutually exclusive subsets (the folds) D_1, D_2, ..., D_{10} of approximately equal size. The classifier is trained and tested 10 times; each time $t \in \{1, 2, \ldots, 10\}$, it is trained on $D \backslash D_t$ and tested on D_t [195]. The cross-validation estimate of the classifier performance is the overall performance over all the folds. Repeated cross-validation was implemented to ensure robust estimation of performance [195]. The ability of the models to classify MCI and CTL was then tested with Dataset 2 to assess their potential and hence the underlying protein panels to detect early AD.

5.2.4 Novel panel identification and model development

Figure 5.1 shows the methodological framework that was applied to identify novel blood biomarker panels and to develop the new models for early detection of AD. The framework is described in detail subsequently. Briefly, the framework consists of three major procedures, including feature subset preselection, biomarker panel formation, and machine learning-based model development and evaluation. A feature subset preselection process was performed to identify marker subsets that may have strong discriminatory power between disease subjects (ADD) and CTLs. A brute force search was applied to the preselected feature subset to form several panels. Each of the panels was then used to develop and cross-validate SVM classifiers of different kernels (K) using Dataset 1. Data from ADD subjects were used in these initial procedures on the basis that dementia subjects are most likely to exhibit the metabolic alterations that are associated with the disease. The most stable kernel and candidate

panels (i.e., promising models) trained on Dataset 1 were further evaluated for the classification of individuals with MCI and CTLs using Dataset 2. The promising models with the best performance at this stage were selected. Finally, the marker panels that underlie the selected models are reported as potential blood-based non-amyloid biomarker signatures of early disease.

Feature subset preselection

A feature subset preselection procedure was implemented with Dataset 1 using CFS method [135]. The goal of this task was to make an initial selection of the most relevant and non-redundant features for the classification of ADD and CTL subjects and consequently reduce the dimension of the study data prior to model development. Reduction of the dimension of the study data was necessary because it would otherwise be computationally expensive to implement an exhaustive search to evaluate the classification performance of all possible feature subsets with machine learning algorithms. For d-dimensional data (where d is 146 in this case) there are 2^d possible feature subsets.

The CFS approach comes under the broad category of filter-based feature subset evaluation methods that attempt to remove irrelevant and redundant features from data by using correlation-based heuristic to determine the worth (merit) of a feature subset. This technique has been shown to compare favourably with wrapper-based approaches in selecting the best feature subsets that achieve high classification accuracy while incurring far less computational cost [127]. It is based on a heuristic that evaluates the merit of feature subsets following the hypothesis that a good feature subset consists of features highly correlated with the class, yet uncorrelated with each other. Correlation in this sense refers to the predictability of one variable by another. Equation (5.1) shows the mathematical formulation of the CFS heuristics, a concept

borrowed from test theory [196].

$$Merit = \frac{f \times r_{fc}}{\sqrt{f + f(f-1)r_{ff}}} \tag{5.1}$$

$Merit$ is the heuristic merit of a feature subset consisting of f features, r_{fc} is the mean feature-class correlation and r_{ff} is the mean feature-feature inter-correlation. The parameters, r_{fc} and r_{ff} are measures of feature relevance and redundancy, respectively, based on the proposition that a feature is relevant if it is correlated with the class, otherwise it is irrelevant. Redundant features are correlated with one or more other features.

To determine the correlations, continuous features were firstly discretized using the discretization method proposed in [197] to ensure that all features were uniformly handled. The correlations were calculated in terms of modified information gain known as symmetrical uncertainty (SU) [198] to cater for the bias of information gain in favour of features with more values. Values were normalised to the range [0, 1] to ensure that they were comparable and had a similar effect.

$$\mathrm{SU} = 2 \times \left[\frac{gain}{H(Y) + H(X)}\right], \tag{5.2}$$

where $gain$ is the information gain [199] for nominal features X and Y, $H(X)$ and $H(Y)$ are the entropy [200] of X and Y, respectively. The gain is formulated as,

$$gain = H(Y) - H(Y|X) = H(X) - H(X|Y), \tag{5.3}$$

where,

$$H(Y) = -\sum_{y \in Y} p(y)\, log_2\, p(y)\,; \tag{5.4}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x)\, log_2\, p(y|x). \tag{5.5}$$

Novel panel formation and SVM-based evaluation

First, feature panels were formed from the CFS-preselected proteins based on a brute force approach. Each panel was then evaluated using a wrapper-based method to identify the ML algorithm and panels with the best performance for the classification of ADD and CTL subjects. Using each panel, several SVM [112] classification models were constructed with different kernels including linear, second- and third-degree polynomials, and radial basis function (RBF) using Dataset 1. The average performance of each model to classify ADD and CTL subjects was obtained using a 10-fold cross-validation [195] scheme repeated 10 times. Second, the performance of most stable models (SVM algorithm and feature panels) that met the performance criteria of average SN and SP ≥ 70% for classification of ADD and CTL subjects was tested with Dataset 2 for discrimination of MCI and CTL groups. Finally, the models and underlying panels with best performance in classifying MCI and CTL groups were selected as putative models and non-amyloid biomarker panels for early detection of AD.

5.2.5 Implementation and performance evaluation

Feature selection using CFS as discussed earlier was conducted with attribute selection toolbox in WEKA software package [201]. All classification tasks were conducted with MATLAB and WEKA software packages accordingly. MATLAB codes are available on https://github.com/chimastan/earlydetectionofAD. In evaluating the models from previous studies, WEKA was used where previous studies had used it for model development. Training of learning models and validation of performance for ADD vs. CTL discrimination was based on 10-fold cross-validation scheme repeated 10 times. The data (Dataset 1) were randomly re-partitioned after each run to ensure that data subsets used for training and validation varied from the ones used in the preceding run. This way, a more robust average performance was obtained.

Classification performance metrics of primary consideration were measures of SN and SP in accordance with international recommendations for clinically usable AD biomarkers [33]. A performance threshold of 70% for SN and SP was adopted in the model development task. This is on the grounds that the diagnostic accuracy of human experts reaches 77% [202] with sensitivity and specificity reaching 81% and 70% [24], respectively. Moreover, sensitivity and specificity greater than 80% is the target performance for ideal AD biomarkers [33]. No class imbalance handling procedure was applied to the training dataset (Dataset 1) in model development as minority to majority class distribution was 35:65% which is acceptable in machine learning-based classification problems [203, 204].

5.3 Results

5.3.1 Replication and evaluation of existing models

Seven existing models for classification of ADD patients and CTL subjects were successfully replicated. The model proposed by [160] could not be replicated because it was originally trained on a dataset obtained from a private source. Nevertheless, a similar model was constructed with Dataset 1 based on the learning algorithm and blood biomarker panel proposed by the ([160]) study. Existing models investigated in this study were ones constructed with blood biomarkers available in this study's dataset. Table 5.1 shows the average cross-validated performance of the models repeated over 10 runs for classification of ADD and CTL subjects. Nearly all the models achieved SN, SP, and AUC greater than 80%, 60%, and 0.70, respectively. However, when evaluated for possible detection of early AD by classifying MCI and CTL with Dataset 2, the SN values of the models remained moderately high while their SP values drastically dropped (with only one model achieving up to 50%). This implies that the models may have undesirably high levels of false positives when applied for

early disease detection. Consequently, the underlying protein panels may not serve as good biomarker signatures of early disease.

Table 5.1. Performance of existing blood biomarker panels for ad detection.

Study	Panel size	Panel	ML model	ADD vs. CTL (Dataset 1)			MCI vs. CTL (Dataset 2)		
				SN	SP	AUC	SN	SP	AUC
[160]	11	ADIP, B2M, CRP, FABP, FVII, IL18, MCP1, PPP, TLSP, TNC, VCAM	Random forest	85.2	25.9	0.62	81.6	46.3	0.72
[163]	5	A1M, APOE, BNP, IL16, SGOT	Logistic regression	85.2	74.1	0.90	79.0	50.0	0.70
[161]	8	A1M, APOA2, APOE, BNP, EOT3, IGM, PLGF, SGOT	Random forest	88.0	72.4	0.87	80.9	46.3	0.69
	5	A1M, APOA2, APOE, BNP, SGOT		87.0	62.1	0.83	83.1	38.9	0.67
	13	APOA2, APOE, BNP, EOT3, HBEGF, IGM, IL16, PLGF, PYY, SGOT, TNC, TTR, VIT		92.6	60.3	0.87	85.3	42.6	0.72
	14	A1M, A2M, APOA2 APOE, BNP, BTC, CRP, EOT3, IGM, IL16, MPO, PLGF, RAGE, SGOT		92.6	67.2	0.91	83.1	44.4	0.70
[164]	6	A1M, A2M, AAT, APOE, CC3, PPP	Naive Bayes	86.1	63.8	0.82	78.3	37.0	0.62
[191]*	5	A1M, A2M, CC3, IGM, TNC	SVM	81.1	60.5	0.77	75.7	35.2	0.65

* Use of apolipoprotein ε4 (APOE4) genotype as covariate in original model proposed in [191] was excluded as distribution of APOE4 status is highly uneven in CTL group (less than 9% of CTLs are positive). A1M: Alpha-1 microglobulin; A2M: Alpha-2 macroglobulin; ADIP: Adiponectin; APOA2: Apolipoprotein A2; APOE: Apolipoprotein E; B2M: Beta-2 microglobulin; BNP: Brain natriuretic peptide; BTC: Betacellulin; CC3: Complement C3; CRP: C-reactive protein; EOT3: Eotaxin-3; FABP: Fatty acid binding protein; FVII: Factor VII; GCSF: Granulocyte-colony stimulating factor; HBEGF: Heparin-binding EGF-like growth factor; IGM: Immunoglobulin M; IL: Interleukin; MCP1: Monocyte chemotactic protein 1 α; MPO: myeloperoxidase; PLGF: placenta growth factor; PPP: Pancreatic Polypeptide; PYY: -Peptide YY; RAGE: Receptor for advanced glycosylation end; SGOT: Serum glutamic oxaloacetic transaminase; TLSP: T-lymphocyte secreted protein 1.309; TNC: Tenascin C; TTR: Transthyretin; VCAM: Vascular cell adhesion molecule-1; VIT: Vitronectin.

5.3.2 Feature subset preselection

Using the proposed methodological approach, sixteen proteins with a merit ($Merit$) of 0.36 were preselected with the CFS technique from the 146 proteins in the original study data. The 16 proteins are shown in Table 5.2 together with their statistical significance P as calculated with z-test. The z-test was used to estimate the statistical significance of the difference between the pair of clinical groups being considered

together (AD vs. CTL) and (MCI vs. CTL) for the pre-selected features. All except a few features were statistically significant (p-value < 0.05) in the ADD vs. CTL pair (Dataset 1). Most of the features were not statistically significant in the MCI vs. CTL pair (Dataset 2). This may be due to the high imbalance between the sample sizes of MCI and the CTL in the dataset.

Table 5.2. CFS-based preselected proteins.

Protein	p-value	
	ADD vs. CTL (Dataset 1)	MCI vs. CTL (Dataset 2)
A1M	2.9E-6	3.3E-1
A2M	2.5E-3	3.2E-1
APOA2	3.2E-8	1.1E-1
APOE	1.1E-7	3.8E-4
BNP	7.7E-7	5.2E-2
BTC	4.4E-2	2.4E-1
CD5L	1.0 E-1	8.6E-1
EOT3	5.5E-5	6.2E-3
IGM	9.7E-7	3.9E-5
IL3	8.1E-3	6.9E-15
MCSF1	4.0E-1	8.4E-2
PAPPA	7.7E-4	1.6E-1
PLGF	1.3E-5	3.2E-1
PYY	2.7E-6	5.9E-1
RAGE	6.5E-3	6.3E-1
SGOT	9.2E-6	2.2E-6

A1M: Alpha-1 microglobulin; A2M: Alpha-2 macroglobulin; APOA2: Apolipoprotein A2; APOE: Apolipoprotein E; BNP: Brain natriuretic peptide; BTC: Betacellulin; CD5L: CD5; EOT3: Eotaxin-3; IGM: Immunoglobulin M; IL3: Interleukin-3; MCSF1: Monocyte-colony stimulating factor 1; PAPPA: Pregnancy-Associated Plasma Protein A; PLGF: Placenta growth factor; PPP: Pancreatic Polypeptide; PYY: peptide YY; RAGE: Receptor for advanced glycosylation end; SGOT: Serum glutamic oxaloacetic transaminase.

5.3.3 Novel panel formation and SVM-based evaluation

From the 16 CFS-preselected protein subset, 2^{16} different panels were formed. Results from wrapper-based evaluation of all the panels for classification of ADD and CTL groups using Dataset 1 showed that models constructed with 2-degree polynomial kernel had a better and more stable performance. Consequently, SVM with 2-degree polynomial kernel was selected as the algorithm of choice. Only (10,699) 2-degree polynomial kernelised SVM models that met the pre-specified performance benchmark (SN and SP $\geq 70\%$) for ADD vs. CTL classification were further evaluated

for their potential to detect early disease with Dataset 2. Two models constructed with six and eight -marker panels (A1M, A2M, ApoA2, CD5L, IL3, SGOT and A1M, A2M, ApoA2, BNP, BTC, CD5L, IL3, SGOT, respectively) achieved a remarkable cross-validated performance (SN of 92% and 93%, SP of 81% and 83%, AUC of 0.90 and 0.94 respectively) in classifying ADD and CTL subjects. This perhaps highlights a performance benefit of the CFS-based feature preselection technique. Nevertheless, the two models performed poorly when evaluated for classification of MCI and CTL subjects. The implication is that, in line with the hypothesis of this study, an excellent model at later stages of the disease does not necessarily imply a good disease detection model at the early disease stages. This may be attributed to subtle differences in the underlying patterns as well as noise in the data among other factors, thus highlighting the need for further evaluations. Five models constructed with panels shown in Table 5.3 realised best performance for classification of MCI and CTL groups. All but one of the models detected AD subjects with SN and SP above 80% and 70% respectively at dementia as well as MCI stage. A larger panel formed by combining all five panels in Table 5.3 achieved a cross-validated SN, SP, and AUC of 85%, 70%, and 0.88, respectively in classifying ADD vs. CTL. However, its specificity dropped drastically to 52% with 82% SN and 0.73 AUC when tested for MCI vs. CTL classification. The introduction of well-known risk factors of AD [205] such as age and level of education as covariates to the models did not improve performance significantly. APOE4 genotype was not used as a covariate to avoid bias since less than 9% of CTL group have positive status.

Table 5.3. Performance of identified novel blood-based biomarker panels.

Panel size	Panel	ADD vs. CTL (Dataset 1)			MCI vs. CTL (Dataset 2)		
		SN	SP	AUC	SN	SP	AUC
7	A2M, APOE, BNP, Eot3, PLGF, RAGE, SGOT	88.5	70.4	0.87	80.1	70.4	0.80
7	A2M, APOE, BNP, EOT3, PYY, RAGE, SGOT	88.9	73.8	0.89	77.9	74.1	0.80
8	A2M, APOE, EOT3, IGM, MCSF1, PYY, RAGE, SGOT	85.3	71.6	0.86	83.8	70.4	0.83
9	A2M, APOA2, APOE, BNP, BTC, EOT3, PYY, RAGE, SGOT	85.0	75.0	0.89	80.1	72.2	0.80
10	A1M, A2M, APOE, BNP, BTC, EOT3, IGM, MCSF1, PAPPA, SGOT	88.1	72.9	0.89	83.1	70.4	0.80

A1M: Alpha-1 microglobulin ; A2M: Alpha-2 macroglobulin; APOA2: Apolipoprotein A2; APOE: Apolipoprotein E; BNP: Brain natriuretic peptide; BTC: Betacellulin; CD5L: CD5; EOT3: Eotaxin 3; IGM: Immunoglobulin M; IL3: Interleukin 3; MCSF1: Monocyte-colony stimulating factor 1; PAPPA: Pregnancy-Associated Plasma Protein A; PLGF: Placenta growth factor; PPP: Pancreatic Polypeptide; PYY: peptide YY; RAGE: Receptor for advanced glycosylation end; SGOT: Serum glutamic oxaloacetic transaminase.

5.4 Discussion

In this study, potentially useful novel blood-based biomarker panels and the corresponding machine learning models for early detection of AD were produced using a fresh approach, having demonstrated that existing biomarkers panels may not be suitable for early detection. The models and panels were selected based on their performance at both the prodromal and dementia stages of the disease, thus improving the chance that signals about the disease were captured rather than noise resulting from individual variations between study participants. Ideally, the smaller the size of a panel, the better in terms of interpretability and cost of implementation in practical applications such as point of care technology. However, because this study was exploratory, it was important to flag all the panels that achieved reasonably good performance since it is unclear which panel or markers are the most important. Gaining such clarification may require further investigation such as analysis of protein-protein interaction for the proposed panels (see later). The performance of a larger panel derived by combining all five identified panels was also shown, although it has a lower

performance relative to the individual panels, perhaps due to curse of dimensionality. Comparing the realised results Table 5.3 with those of existing models investigated Table 5.1; the best existing model identified AD subjects at MCI stage with high sensitivity and fairly good specificity (79% SN and 50% SP) while the model with the least panel size developed in this study achieved better performance with 80% SN and 70% SP. At the dementia stage, the proposed models achieved a performance that is comparable to the best model from the investigated studies.

Comparing the results from this analysis with the three recent relevant studies (see Table 5.4), it can be observed that the panels identified in [192] and [193] classified ADD and CTL with high performance, but the markers were reported by the authors to be poor at distinguishing between MCI and CTL. Furthermore, while study [194] achieved a high AUC of 0.88 with XGBoost model for classification of ADD and CTL, the model's performance has not been evaluated for disease detection at MCI stage. Due to the unavailability of biomarkers used in the study at the time of this analysis, the performance of the models for MCI and CTL classification was not investigated herein.

Table 5.4. Comparison of realised results with recent relevant studies.

Study	ML model	ADD vs. CTL			MCI vs. CTL		
		SN	SP	AUC	SN	SP	AUC
[192]	Logistic regression	84.0	70.0	0.79		Poor	
[193]	Random forest	90.0	67.0	0.77		Poor	
[194]	XGBoost	-	-	0.88	-	-	-
	Random forest	-	-	0.85	-	-	-
	Deep learning	-	-	0.85	-	-	-
Current study	SVM	85.0	75.0	0.89	80.1	72.2	0.80

In contrast to the recent studies, the proposed models from this study achieved high performance for disease detection at ADD stage (with one of the models shown in the table realising best AUC, with high sensitivity and specificity) as well as the MCI stage. The identified panels differ significantly from those of existing methods. This may be due to significant differences in the approaches including feature preselection and evaluation modalities which were deliberately applied in this study. To the best of my knowledge, no relevant study has previously applied CFS for feature preselection. Details of the learning algorithm used, including kernel information as well as their selection process were provided to ensure transparency of approach and reproducibility. It is noteworthy that no existing AD model based on non-amyloid proteins has hitherto been evaluated for early disease detection using ADNI data. Regarding the proteins evaluated in this study, besides PAPPA, which is rather highly associated with depressive symptoms in older adults [206] other proteins preselected by CFS have been previously identified in several studies [159-161, 163, 187, 189-191, 207] to have classification value in discriminating between ADD and CTL groups. From the five selected panels shown in Table 5.3, six proteins (i.e., A2M, APOE, BNP, EOT3, RAGE, and SGOT) appear as most prominent, featuring in nearly all the panels. A combination of the six proteins therefore appears to play a significant role in the discrimination of disease (prodromal and dementia) subjects and healthy controls. The panel classified both groups with sensitivity and specificity > 80% and 65%, respectively and AUC of at least 0.80. Several of these proteins are found in nearly all the previously reported models investigated in this study. Studies show that blood plasma levels of A2M are linked to mechanisms related to blood-brain barrier damage and neuronal injury as well as hippocampus metabolism in early AD [172, 173]. ApoE in blood is speculated as a dementia risk marker in preclinical AD [208]. BNP levels in

plasma is associated with a decline in cognitive function [209]. Plasma levels of RAGE are altered in AD [210]. RAGE has been reported to play a critical role in AD and is considered a potential therapeutic target [211]. SGOT is a biomarker of peripheral inflammation and an essential metabolic enzyme. It is often used as a clinical measure of liver function [212]. Interestingly, a recent finding has implicated liver function as a potential significant confounding factor in the onset of AD (https://www.alz.org/aaic/releases_2018/AAIC18-Tues-gut-liver-brain-axis.asp).

5.5 Summary

As emerging findings suggest that AD is characterised by metabolic changes possibly detectable in blood and may precede amyloid pathology, one of the hallmarks of AD, signatures of such metabolic abnormalities may therefore serve as biomarkers of early disease. In this chapter, peculiar feature selection and evaluation modalities were applied to identify potential blood-based (non-amyloid) biomarkers for early detection of AD as existing machine learning-based solutions are optimised for detection of the disease at later stages. The main contributions of this study include the potential biomarker panels identified and the innovative methodological approach developed for the search to bridge this important research gap. The developed machine learning models based on these panels classified prodromal AD as well as AD dementia and normal controls with sensitivity above 80%, specificity higher than 70%, and AUC of at least 0.80. Existing models performed poorly in comparison at this stage of the disease, suggesting that the underlying marker panels may not be suitable for early disease detection. A combination of A2M, APOE, BNP, EOT3, RAGE and SGOT was identified as a key biomarker profile with significant contribution to the classification performance. Overall, the results suggest that it may be feasible to detect AD at early stages using a profile of non-amyloid proteins in blood that may indicate metabolic processes that accompany or precede the disease. However, this requires further

studies.

2

[2] This chapter is a slightly modified version of “Early Detection of Alzheimer’s Disease with Blood Plasma Proteins Using Support Vector Machines” published in the IEEE Journal of Biomedical and Health Informatics (JBHI) and has been reproduced here with the permission of the copyright holder.

Chapter 6 Robust Blood Biomarker Signature of Cerebrospinal Fluid Amyloid-beta 42 Status

6.1 Introduction

In the preceding chapters, it has been highlighted that current disease-modifying clinical trials target individuals at the earliest stages of AD, where intervention is thought to be most likely successful, following the high failure rates of previous trials [4]. Accumulation of amyloid-beta (Aβ42) plaques in the brain, also known as amyloid pathology, has been also discussed as one of the key biochemical events that characterise AD and that it is present long before clinical symptoms are apparent [2, 3]. Amyloid screening is being used in these trials to identify individuals with amyloid pathology and is envisaged to be beneficial in the future for population-based screening [5, 6]. The amyloid screening tests, which aim to detect abnormal amyloid accumulation, are conducted with the aid of amyloid PET scan and Aβ42 measurement in CSF [7]. However, as previously pointed out, PET scan is expensive and available only at specialised centres while lumbar puncture required for CSF testing is invasive, thereby posing an economic burden and challenges in recruitment of participants.

There is a growing body of evidence that CSF Aβ42 may be an earlier indicator of AD pathology compared to amyloid PET [9-11] and thus may be a more suitable biomarker for disease detection at the earliest stages. To mitigate the limitation of invasiveness posed by CSF-based amyloid testing, there is evolving interest in identifying blood-based biomarkers reflective of amyloid status, as would CSF. Such biomarkers may be used as a reliable initial step in a multistage diagnostic procedure.

A few studies [12, 13] have demonstrated the potential of blood-based markers predictive of amyloid status as measured by CSF Aβ42 with area under receiver operating curve (AUC) reaching 0.88 (in 46 samples) and 0.81 (in 358 samples),

respectively. However, the novel method employed by [12] in measuring the blood-based markers remains to be established and the results from [13] are yet to be validated in independent cohorts.

In this study, the utility of blood-based proteins to predict CSF Aβ42 status using support vector machines with recursive feature elimination (SVM-RFE) that has shown effectiveness in similar research domains [14] was explored. Particular consideration was given to the robustness of identified markers, to enhance the likelihood of reproducing results since reproducibility of results is one of the lingering challenges in blood biomarker discovery for AD [15].

6.2 Methods

6.2.1 Study data

Baseline data from 358 individuals including CTL, MCI and ADD subjects were obtained from ADNI-1 cohort. The data comprised of APOE4 genotype and blood-based measurement of 146 proteins as earlier described (see Section 3.1). Particular to this study, 3 additional blood-based proteins including homocysteine, Aβ40, and Aβ42 were obtained alongside CSF Aβ42 measurements of the individuals. CSF Aβ42 status for the individuals was derived by dichotomizing their CSF Aβ42 levels as normal (high) or abnormal (low) according to clinically recognized threshold of 192pg/ml for the Luminex biosensing platform.

6.2.2 Robust biomarker selection

The objective here was to identify potential blood biomarker signatures predictive of CSF Aβ42 status, from which a signature can be selected based on robustness and performance. The measure of robustness was intended to be transparent and simple to evaluate. The method used is based on the approach proposed by Abeel et al. [213], with some modifications.

Similar to [213], SVM-RFE [113] combined with ensemble technique illustrated in Figure 6.1 was used to select features for signatures formation, while Kuncheva index (KI) [214] was used to evaluate the robustness of signatures. SVM-RFE combines the embedded feature selection capability of linear SVM with the backward feature elimination strategy of RFE. Absolute values of the weights (coefficients) the linear SVM provides is the contribution of each feature to the SVM hyperplane and may be used as a means of ranking the importance of individual features. A feature with a larger weight is regarded as one of higher importance, and one with a lower weight is considered less important.

RFE implements a backward feature elimination procedure that iteratively removes the least important features in the training data samples. The algorithm starts out by fitting the training data with all the available features to a linear SVM, then ranks the features according to their weights and eliminates the least important one(s). The training data is subsequently refitted to the linear SVM but with only the retained features. This process is repeated until all features have been eliminated or a desired number of features to retain is attained.

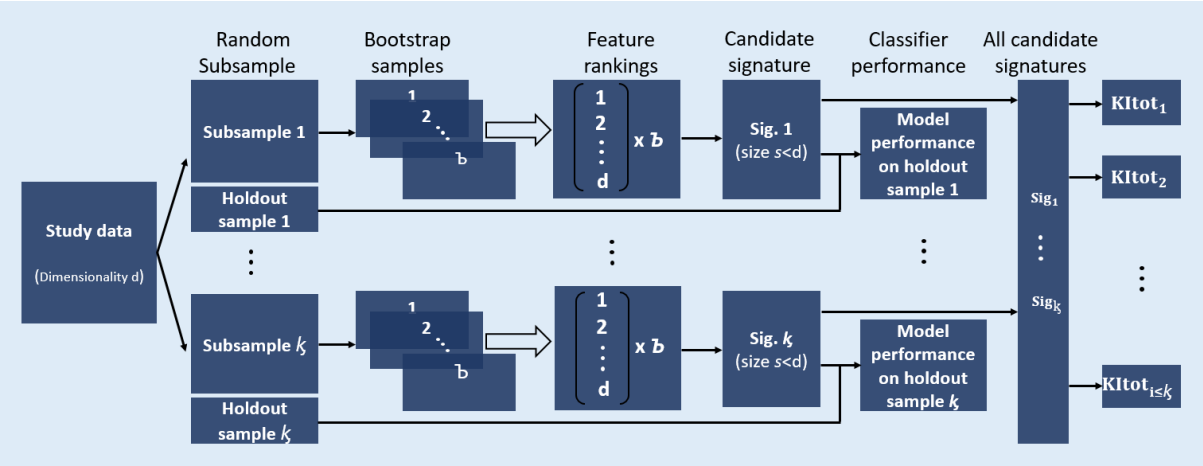


Figure 6.1. Visual overview of the implemented ensemble learning approach.

Finally, each feature in the training data is assigned an overall rank ├ (an integer with 1 as minimum and dimension of training data d as maximum) according to the observed feature contributions, with most significant features assigned lowest ranks. SVM-RFE with ensemble learning is implemented to improve the robustness (stability) of feature subset selection by SVM-RFE. In this approach, ƙ different subsamples of the original dataset (of d dimension) are generated using random sampling without replacement, each subsample containing only a slight variation (ρ samples) of the original dataset. For each subsample (in the ƙ subsamples), Ƃ bootstrap samples are generated. SVM-RFE provided with a specified signature size s as a stopping criterion is then applied to each bootstrap. The rank of each feature in d as well as the AUC performance ($\mathrm{AUC_{OO}}$) of the selected features on the out-of-bag samples is recorded. A candidate signature of size s is subsequently selected according to an ensemble ranking R obtained by aggregating ├ over all Ƃ bootstrap samples as shown in (6.1). An estimate of the generalization performance of the signature is obtained by training the linear SVM on the subsample and its performance evaluated on the $1-\rho$ held out samples. The ensemble method of generating signatures has been shown to improve robustness and classification performance compared to simply applying SVM-RFE directly to subsamples [213]. In addition to the approach proposed in [213], repeated stratified cross-validation of the candidate signature on the corresponding subsample was carried out as a supportive evaluation of the signature's classification performance.

$$\mathrm{R} = \sum_{i=1}^{Ƃ} \omega_i ├_i \tag{6.1}$$

The weight ω_i is bootstrap-dependent. It takes either of two values depending on the chosen aggregation method. In the complete linear aggregation (CLA) method, ω_i is

set to 1, while $w_i = 1 - \mathrm{AUC}_{\mathrm{OO}}$ in the complete weighted aggregation (CWA) strategy. The two methods were explored in this study although CWA was shown to be marginally better than CLA in [213].

To evaluate the robustness of the k candidate signatures, a stability measure defined by the Kuncheva index (KI) [214] shown in (6.2) was applied.

$$\mathrm{KI} = \frac{m - (s^2/k)}{s - (s^2/k)} \tag{6.2}$$

KI with range [-1, 1] measures the similarity between two signatures. The variable m is the number of features common to both signatures. The greater the value of KI, the larger the number of common features. A negative index indicates that feature intersection is mostly due to chance. The overall stability $\mathrm{KI}_{\mathrm{tot}}$ of a signature can be defined as the average of all pairwise similarity comparisons between the signature and the rest of $k - 1$ signatures as in (6.3).

$$\mathrm{KI}_{\mathrm{tot}} = \frac{\sum_{i=1}^{k-1} \mathrm{KI}_i}{(k-1)} \tag{6.3}$$

6.2.3 Implementation

The robust biomarker selection task was implemented in Python programming language. The machine learning subtasks were conducted with the Scikit-learn package. Codes are available at https://github.com/chimastan/robust-blood-based-signature-of-csf-abeta42-status. The values of k, b, and ρ used were 500, 50, and 0.8, respectively, considering the recommendations by [213]. Cross-validation used was 10-fold with 10 repetitions with samples stratified according to the target label distribution. The C parameter for the linear SVM was set to default (C=1). In the RFE, the number of features to eliminate per run was set to 20% of the total available features to improve the speed of processing.

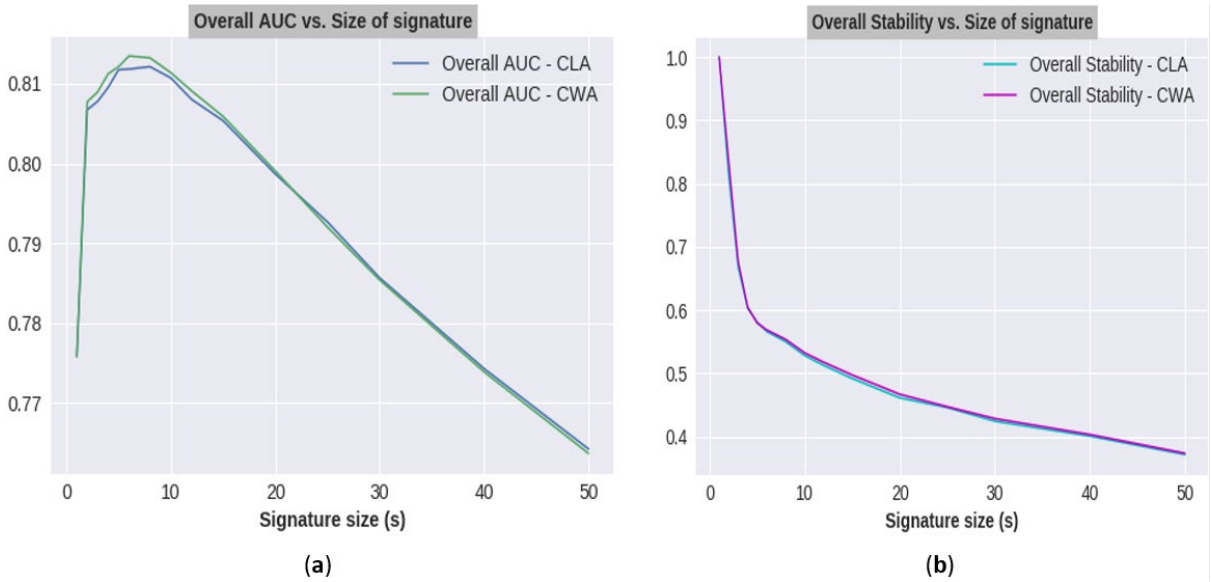


Figure 6.2. Comparison of (a) classification and (b) stability performance of CLA and CWA-based ensemble methods. The overall AUC and stability are the average AUC and KI_{tot} over the ķ (500) subsamples.

6.3 Results

6.3.1 Potential robust signatures

Several potential signatures with various levels of classification and stability performance for prediction of CSF Aβ42 status were realised. Figure 6.2 illustrates the variation between signature size s and the average cross-validated AUC as well as average KI_{tot} over the 500 subsamples. The average AUC gradually increased with increasing s up to a point ($s \approx 8$) and then declined, while stability steadily dropped with increasing value of s. The results of CWA and CLA ensemble methods were largely equivalent as shown in Figure 6.2. Thus, all further reports are based on results of the simpler CLA method. Consideration of potential signatures was also limited to ones consisting of 5 biomarkers, being that stability remained moderate at $s = 5$ while the increase in average AUC beyond that point was minimal. A total of 229 unique candidate signatures were obtained from the 500 subsamples. The top 10 signatures

with the best values of stability KI_{tot} (ranging between 0.67 and 0.61) were preselected. Then further analysis was conducted on them to aid in making a final selection.

6.3.2 Final selection of signature

Additional analyses with similar approach described in Section 6.2.2, but with s limited to 5 and random forests (RF) used as the machine learning algorithm, were conducted. Therefore, in this case, RF-RFE was applied instead of SVM-RFE. The number of trees per forest was set to 2000, each forest containing a maximum of $d^{3/4}$ features as recommended in [215]. The purpose was to obtain candidate signatures with best KI_{tot} values and compare them to the top 10 realized earlier with SVM-RFE. This would allow identifying signatures whose classification and stability performance may be agnostic to the type of machine learning algorithm and thus likely to generalize better. With the RF-RFE, 169 unique potential signatures were realised and the top 10 with the best stability values were identified. A comparison of the signatures with ones obtained with SVM-RFE implicated one signature as common. The signature consists of APOE4 genotype, eotaxin-3 (EOT3), apolipoprotein-C1 (APOC1), chromogranin-A (CGA), and Aβ42. The signature achieved 0.64 stability (KI_{tot}) value. Average AUC, sensitivity, specificity, negative predictive value (PPV) and negative predictive value (NPV) for the repeated 10-fold cross-validation were 0.85, 84%, 63%, 83% and 67%, respectively. The average values on the unseen held-out samples were 0.84 AUC, 82% sensitivity, 62% specificity, 81% PPV, and 64% NPV, respectively. The contribution of individual biomarkers to the classification performance of the signature is as shown in Figure 6.3, with APOE4 unsurprisingly making the most contribution.

6.4 Discussion

In this study, the utility of blood-based signature to predict CSF Aβ42 status with a robust performance was investigated. It was demonstrated that APOE4 genotype and

levels of four proteins predicted CSF Aβ42 status with high AUC. This is the first study to demonstrate a signature with a stable performance beyond a single machine learning algorithm. It is a positive indicator of the potential of the identified signature to generalize to other cohorts. Compared to existing studies, four out of the five predictors (APOE4, CGA, Aβ42 and EOT3) in the signature were implicated in a multi-marker panel from a recent study [216] as predictive of CSF Aβ42 status with random forests. A number of studies have shown evidence of association between some of the identified markers and AD. In line with the observed prominent contribution of APOE4 in the identified signature, it is the strongest and most prevalent genetic risk factor for late-onset AD and is considered as a possible therapeutic target [217]. Serum and CSF but not plasma levels of EOT3 have been shown to be dysregulated in individuals with AD [218]. APOC1 genes, in combination with APOE4, are suggested to play an important risk factor role in AD [219, 220].

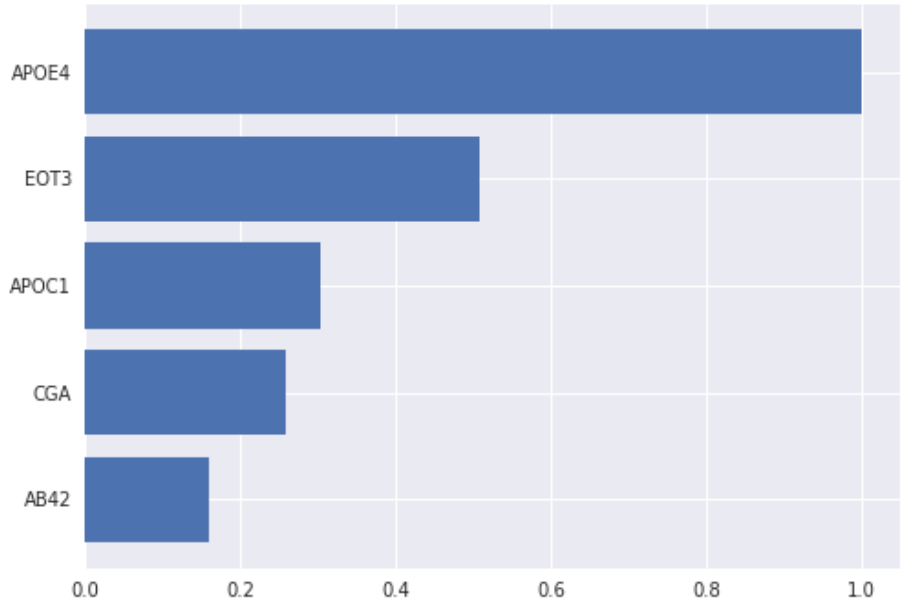


Figure 6.3. Contribution of individual marker to classification performance of the selected signature. The contribution was determined from the feature weights of linear SVM, normalized by the largest weight during training.

However, the association between plasma levels of APOC1 and AD has not been evidenced. CGA on the other hand has an amount of co-localisation with brain amyloid plaques [221], but CSF and blood levels of CGA have not been reported to be correlated. Interestingly, plasma and CSF Aβ42 have shown to be correlated in individuals with AD [222, 223].

This study has several limitations. All analyses were conducted with the ADNI cohort with its peculiarity such as age and level of education of participants. The distribution of individuals with abnormal CSF Aβ42 levels across the clinical groups (CTL, MCI, and ADD) was biased, with nearly all samples belonging to the MCI or ADD group. This might have influenced the analyses, as the individuals are likely to have developed other confounding conditions.

6.5 Summary

In view of evolving evidence suggesting that CSF Aβ42 level may indicate AD risk earlier compared to amyloid PET marker, identification of blood-based biomarkers to serve as a surrogate measure indicative of CSF Aβ42 status has become necessary. This is because blood collection is minimally invasive and inexpensive compared to the procedure for collecting CSF. In this chapter, it was shown that APOE4 genotype and blood markers comprising EOT3, APOC1, CGA, and Aβ42 may be a suitable biomarker profile to predict CSF Aβ42 status. The major contributions of the research include the identified novel biomarker signature and innovative machine learning approach implemented for the identification. Given that early detection is believed to be central to a successful development of disease-modifying intervention, the potential utility of the identified markers is enormous. They may be applied as a minimally invasive and cost-effective first-line screening tool in a multistage diagnostic procedure to facilitate the enrichment of clinical trials.

[3]

[3] This chapter is a slightly modified version of “A Robust Blood-based Signature of Cerebrospinal Fluid Aβ42 Status” published and presented at the 42nd International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2020 and has been reproduced here with the permission of the copyright holder.

Chapter 7 Prototype Software to Facilitate Detection of AD with Blood Biomarkers

7.1 Introduction

The conceptual framework of the proposed BBDiag blood-based AD diagnostic point of care device shown in Figure 1.1 includes an intelligent computational component that should collect multiplex outputs from a biosensor module producing blood biomarker measurements, analyse the outputs, and suggest the diagnostic status of the individual from which the blood sample was obtained. The purpose of this intelligent firmware is not to provide definitive diagnostic information, but to furnish assisting information to a medical expert, based on the individual's blood biomarker profile. In this chapter, a prototype software developed to provide such assistive information using blood biomarker panels identified in this project is demonstrated. The software is demonstrated here as a standalone desktop application (a.k.a. BBDiag App).

7.2 Methods

7.2.1 Requirements analysis and design

As a set of minimum requirements, the software should be able to collect data from measurements of blood biomarkers levels and visually provide a corresponding visually interpretable output. Development of the software was carried out using rapid application building together with an incremental software development approach. Rapid application development (RAD) model focuses on producing a useable prototype that can evolve into a complete product [224]. This approach was adopted due to time constraints and because the software is a proof of concept. An incremental build model was adopted in managing the complexity of the software development as shown in Figure 7.1, making room for it to evolve in future.

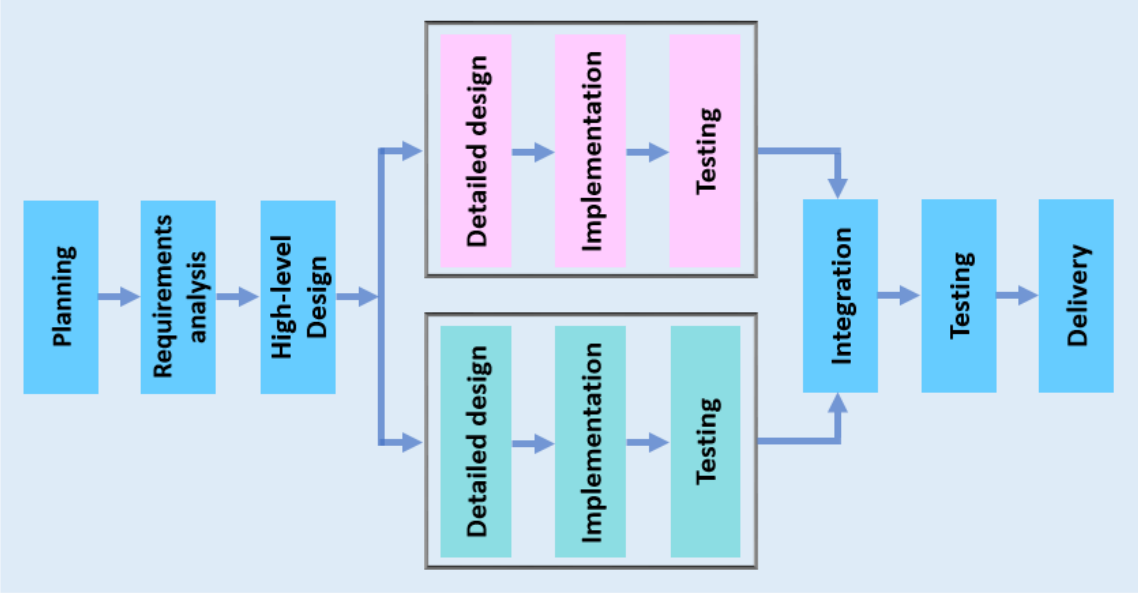


Figure 7.1. Incremental development model for BBDiag App.

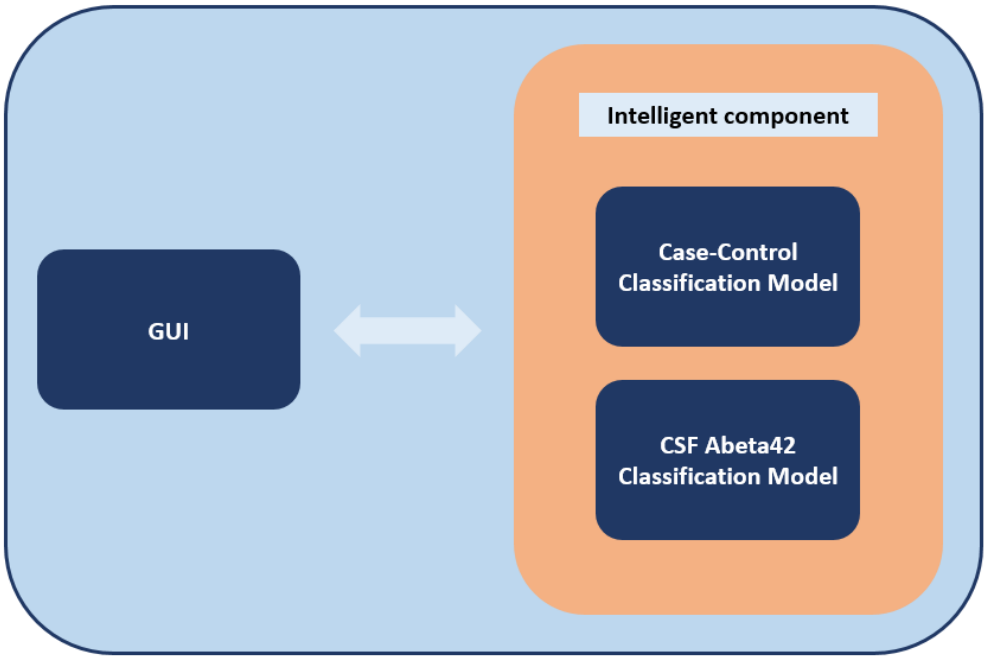


Figure 7.2. High-level design of BBDiag App.

In line with this model, the software was decomposed into two main components (see Figure 7.2), each of which was built separately and then later integrated. The two major components include the intelligent component that is responsible for analysing blood

biomarker data inputted into the system and providing results, and the graphical user interface (GUI) to facilitate input-output interaction between the software and a user.

Predictive models design

The intelligent component was designed to comprise pre-trained classifiers (i.e., predictive models) realised from previously presented analyses. The Case-Control Classifier in Figure 7.2 is one of the models realised in Chapter 5 for having a high classification performance between individuals with ADD or MCI and CTL subjects. It was built with a biomarker panel (A2M, APOE, BNP, EOT3, PYY, RAGE, and SGOT) and second-degree polynomial kernel SVM algorithm (see Table 5.3) to predict whether an individual has a blood biomarker profile consistent with clinical manifestation of AD. The other predictive model - CSF Abeta42 Classification Model - is the model realised in Chapter 6 constructed with a blood biomarker panel (APOE4, EOT3, APOC1, CGA, and Aβ42) identified as indicative of CSF Aβ42 status of an individual, together with a linear SVM algorithm.

GUI design

The user interface was designed as shown in Figure 7.3. It consists of input buttons for sending inputs or commands to trigger an event in the software and output fields for providing visually perceptible outputs or labels, as described below.

Input buttons

i. The “Purpose” dropdown button enables a user to select which classifier to use in making prediction at a given instance. Two options are available: to predict CSF Aβ42 status using the CSF Abeta42 Classification Model or predict clinical phenotype using the Case-Control Classification Model.

ii. The "Load data" button enables a user to select the relevant blood biomarker data of an individual or a group to use for prediction. The data must be in comma separated value (CSV) file format. The data must also contain column headers corresponding to each biomarker acronym listed in the Predictive models design section above, whilst replacing letter “β” with “B” where it appears.

iii. The "Predict" button enables a user to request the selected model to run a prediction and provide results. When the result is ready, a notification "Prediction Complete" is shown on the upper part of the UI. For the case-control prediction, given a single input record, the prediction result is either “LikelyNormal" or "LikelyAD” corresponding to normal cognition or AD-related phenotype (dementia or MCI). For CSF Aβ42 status prediction, given a single input record, the prediction result is either "High" or "Low”.

iv. The "Export result" button enables a user to export the prediction results CSV file to a selected storage location. The file contains the prediction and degree of confidence in the prediction for each corresponding record in the input data.

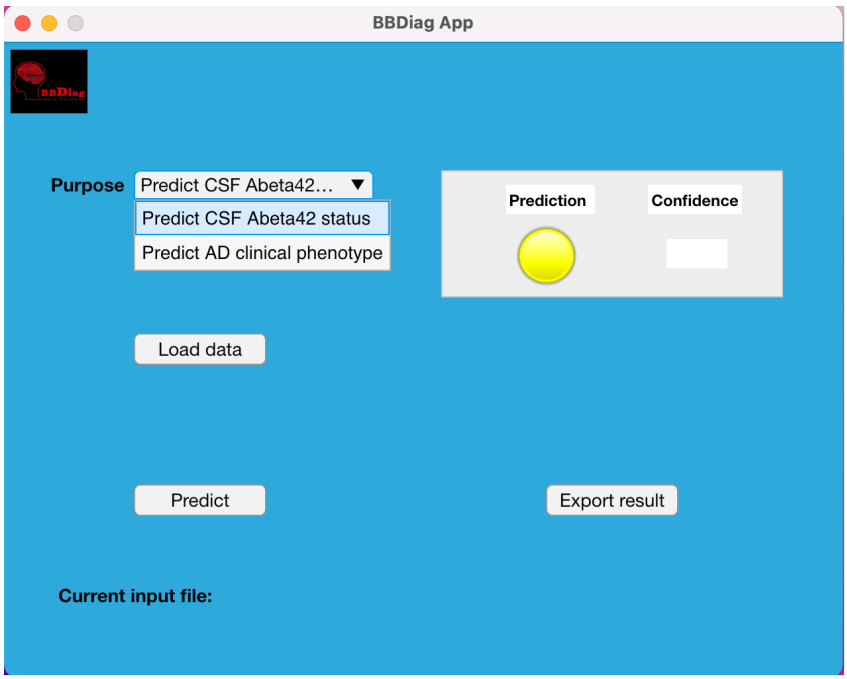


Figure 7.3. GUI design

Labels and output fields

v. The first two fields in the UI denote the name of the software “BBDiag App” and logo of the BBDiag Consortium, respectively.

vi. The "Current input file" label shows the current input file selected by the user in the "Load data" operation if any.

vii. The spherical “Prediction” output field (coloured in yellow) dynamically changes colour according to the result of the prediction for a single record input data. When the input data contains multiple records, it visually indicates only the result of the prediction for the first record in the data. The field is yellow in colour by default, indicating a neutral state. It is green if the prediction is "Low" or "LikelyNormal" but red if "High" or "LikelyAD”.

viii. The "Confidence" label provides the user with the estimated confidence of the prediction for a record. The confidence (measured as a probability, ranging between 0 and 1) is estimated by applying a sigmoid function to the score attribute of the support vector machine, which reflects the distance of the multidimensional data point from the decision boundary. This is because SVM does not naturally provide probability estimate for a prediction. A value of 0 indicates lowest confidence, and a value of 1 maximum confidence. Alternatively, Platt’s scaling may be applied as an advanced method of obtaining the probability.

7.2.2 Implementation and integration

The software designs were implemented using MATLAB. The predictive models were built, trained, and exported as .mat files. The GUI was designed using App Designer. Codes were written with MATLAB editor and compiled using MATLAB compiler. MATLAB publisher was used to package and publish the software into a standalone desktop application for Windows and MAC operating systems, respectively.

Installation and running

File names of the installers are “BBDiagAppInstaller_WindowsOS.exe” for the Windows version and “BBDiagAppInstaller_MacOS.app” for the MAC operating system version. Run the installer to install the program. To use the software in Windows, open from the Start menu or a created shortcut. To use in MAC, start the program from the Applications repository or a created shortcut. Program can also be started from the command line.

Figure 7.4. BBDiag App in operation.

7.2.3 Testing

The application was robustly tested to ensure that all parts are working effectively and efficiently as expected. Figure 7.4 illustrates the software in operation: after a csv file, containing relevant blood biomarker information from one of the ADD patients was loaded onto the system for case-control classification and the Predict button was clicked. The red colour signal indicates that the individual may have AD-kind of clinical phenotype with a probability of 65%.

7.3 Summary

In this chapter, the potential utility of blood biomarker-based predictive models to assist in AD detection within real-life clinical environment was demonstrated using the developed prototype software application. The software was demonstrated here as a standalone desktop application rather than a firmware. Similarly, blood biomarker data applied in developing and testing the software were obtained from external source (the same data obtained from ADNI and used in the preceding chapters) rather than those produced by the proposed BBDiag biosensor module. This was because the development of the required multiplex biosensor module and its packaging is still ongoing. However, the software tool significantly demonstrates the potential of blood-based biomarkers in real-life scenarios; where it could be used as a first-line screening tool to identify individuals to undergo further tests using CSF or PET-based biomarkers.

Chapter 8 Discussion, Future Direction and Conclusion

8.1 Contributions to knowledge

The contributions of this thesis apply significantly to two principal areas of AD research, namely blood biomarker discovery and development of practical use case of blood biomarkers which are discussed as follows.

8.1.1 Blood biomarker discovery

Identified potential biomarker panel of AD at dementia stages

A blood biomarker panel consisting of eight markers (A1M, A2M, APOA2, BNP, BTC, CD5L, IL3 and SGOT) was identified, as highlighted in Section 5.3.3, to discriminate between AD dementia patients and control subjects with a remarkable robustly cross-validated performance of 0.94 AUC with sensitivity and specificity values of 93% and 83%, respectively. The 8-marker panel achieved the best performance compared to any reported similar blood biomarkers in the literature, owing to the innovative biomarker search strategy implemented. Although, the identified panel was poor at discriminating between individuals at earlier stages of AD (in this case, MCI) and control subjects, making it unsuitable for early detection where efforts towards the development of effective treatments are focused, it can still play a key role. Through further investigations, the panels can be useful for gaining a better understanding of the biological dysfunctions that accompany AD at later stages. Such knowledge can be useful for non-curative management of the condition and planning adequate resource provision to support patients. The understanding may also aid to elucidate modifiable predispositions (lifestyle) that may influence the severity of disease.

Identified potential biomarker panel of early AD

A number of biomarker panels were identified as having high classification performance between AD dementia patients as well as MCI subjects and control group (see Table 5.3), contrary to the usual robust performance at only dementia stages from preceding studies using similar features. Of particular interest, six markers (A2M, APOE, BNP, EOT3, RAGE and SGOT) were predominant from the five slightly different panels identified and their combination appeared to be key profile driving the performance of the different panels as discussed in earlier Section 5.4. The potential value of the biomarkers is enormous, considering the current emphasis on early diagnosis as key to developing effective clinical interventions, and the benefits that the use of blood as a medium of biomarker collection affords. More so, the identified biomarker profile is not directly tied to the more established early biomarkers of AD such as amyloid pathology. Understanding the interactions between the markers may deepen understanding of the early processes or pathways of the disease.

Identified potential blood biomarker of amyloid pathology in CSF

Another biomarker panel (EOT3, APOC1, CGA, and Aβ42 together with APOE4 status) with an impressive robust performance was identified (see Section 6.3.2) as being able to distinguish between individuals with abnormal and normal levels of CSF Aβ42. This is of particular importance as CSF Aβ42 is one of the well-established biomarkers of AD and suggested as the earliest preclinical indicator of AD development. Therefore, a reliable blood-based surrogate method of assessing its level would be a huge scientific breakthrough to facilitate enrichment of clinical trials, or population-based screening if an effective treatment becomes available. In fact, a novel graphene-based biosensor for detecting one of the identified markers (plasma

Aβ42) has been developed by one of the collaborators within the BBDiag Consortium [225].

Innovative methodological frameworks for biomarker search

In terms of the analytical approach to biomarker search, contributions were made through effective adaptation of feature selection and evaluation techniques, thereby providing direction on what may be effective in the case of blood biomarkers. Some of these include a carefully designed combination of filter and wrapper-based approaches (i.e., hybrid) illustrated in Section 5.2.4, as well as embedded and ensemble techniques as described in Sections 6.2.2 and 6.3.2. These innovative approaches enabled the successful identification of potential blood biomarkers that may aid AD diagnosis.

8.1.2 Demonstrating a potential practical use case for blood-based biomarkers

A practical demonstration of a potential use case for AD blood biomarkers in real life clinical settings was provided in the form of a prototype application (see Chapter 7). No such relatable tool is currently available to demonstrate the envisioned future of blood biomarkers to stakeholders, e.g., researchers, funders, or the public. The proposed concept may inspire new research and innovation. For instance, the application can be extended and calibrated for different biosensor technologies such as digital ELISA and SOMAscan, to provide an online-based predictive software tool for AD with blood biomarkers.

8.2 Limitations and future directions

Notwithstanding the success of this research, it has some limitations as discussed below.

8.2.1 Data

In this work, sample size of the study data was small. This is because of the limited availability of relevant data, partly due to the high cost of collection of such specialised data. As a result of the limited dataset, data intensive machine learning methods such as deep learning were not profusely pursued. Preliminary analyses conducted during this study with the limited data based on deep learning techniques did not provide superior performance, similar to the results from a related study [194]. Further to this, data augmentation using generative adversarial networks [226] was investigated but proved ineffective at improving performance. This was likely due to the small sample size and nature of the data, as the techniques are usually more effective in problems like computer vision. Typically, data augmentation involves a range of techniques for data synthesis and generation to create either more training data or labelled data to avoid overfitting or minimise cost of labelling. This provides an opportunity for future research to develop suitable augmentation techniques for blood-based biomarker-kind of data. Similarly, with a future increase in original or effectively augmented data, a deep learning approach may prove to provide superior performance and biomarker selection from resulting models may be investigated using explainable artificial intelligence (XAI) techniques [227].

Another limitation is that the study data only consists of older and educated subjects. Thus, findings may be biased and not generalise well to other cohorts such as less educated individuals, given that level of education is a known risk factor for clinical AD. More educated individuals or those involved in intellectually engaging activities tend to have higher brain resilience to AD. Therefore, future research involving collection of new data should consider balancing the demographics of participants.

8.2.2 Biomarker search methods

Notwithstanding the usefulness of data-driven biomarker search techniques applied, some important biomarkers with strong biological links to AD may have been eliminated as most of the analyses were blind to prior knowledge. Furthermore, aspects such as protein-protein interaction were not investigated as these were beyond the scope of the research. Potentially, analysis of the interactions between proteins in the identified panels may facilitate understanding of their joint role in AD process and clarify which panel(s) are more clinically relevant. Therefore, there is a need for a strong collaboration of expertise from advanced data analysis, biology, and mechanistic pathology in this area of research. This opens a whole new window of innovative multidisciplinary research.

Furthermore, besides proteomics-based biomarkers which were used in this research, there are also other types of blood biomarkers such as mRNA [228-230] and autoantibodies [230] where progress is being made in AD detection and improving understanding of the disease. Therefore, investigation of the full range of blood-based biomarkers within a single research theme, rather than in isolation, is a promising research direction.

8.2.3 External validation

There is a need to conduct additional follow-up studies and validation of findings in large and independent cohorts considering that validation of findings is a crucial step for clinical acceptance and translation into clinical practice. Attempts to collect new relevant albeit small size data during this research were unsuccessful. Hopefully, future research will address this challenge.

8.2.4 Other non-invasive low-cost biomarkers

There are also other emerging non-invasive low-cost markers such as electroencephalogram (EEG) markers. Although it was not the focus of this PhD research, preliminary contributions were made in that direction [231], in collaboration with one of the partnering groups within the BBDiag Consortium that was investigating EEG markers. Future research could explore the potential of combining blood biomarkers with other emerging non-invasive low-cost markers to improve performance.

8.3 Conclusion

Given the prevalence of AD, there is a need for non-invasive, low-cost and reliable biomarkers that can be applied in clinical practice for diagnosis. Diagnostic guidelines recommend the use of biomarkers, which may serve the purpose of diagnosis as well as furthering understanding of the disease. Current disease-defining biomarkers of AD include those from cerebrospinal fluid and positron emission tomography neuroimaging, which are either invasive or expensive to collect. Blood-based biomarkers present a complementary alternative, as they are easier and inexpensive to collect. However, identification of suitable blood biomarkers is a huge research challenge. One of the main challenges is that blood contains rich but high dimensional and complex amount of information that may be used for this purpose. Biomarker search involves mining through this complex high dimensional data, which is further made more difficult as no single blood biomarker has shown to provide reliable performance, thereby making traditional statistical approaches unsuitable for the task. Machine learning methods provide analytical tools to confront this challenge. Further to existing relevant research, this thesis identified a number of potentially suitable blood biomarkers to aid AD diagnosis, after extensive machine learning-based analytical exploration of blood proteomics data. In addition, a potential practical use

case for blood-based biomarkers in real-life clinical settings was demonstrated. The identified biomarker panels can be useful in developing a suitable point of care technology for diagnosis, as each panel consists of only a few biomarkers. Besides aiding diagnosis, they may be also useful in deepening understanding of the disease's development mechanism to aid the realisation of suitable treatments. Overall, this research demonstrates the huge prospects of blood-based biomarkers for real-life applications in AD diagnosis.

References

[1] D. Australia, S. Baker, and S. Banerjee, "Alzheimer's Disease International: World Alzheimer Report 2019: Attitudes to Dementia. 2019," *Alzheimer's Dis. Int,* 2019.

[2] W. H. Organization, "Global action plan on the public health response to dementia 2017-2025," 2017.

[3] K. Donegan, Donegan, N. Fox, N. Black, G. Livingston, S. Banerjee, and A. Burns, "Trends in diagnosis and treatment for people with dementia in the UK from 2005 to 2015: a longitudinal retrospective cohort study," *The Lancet Public Health,* vol. 2, no. 3, pp. e149-e156, 2017

[4] Policy Paper (2015). *Prime Minister's challenge on dementia 2020*. [Online] Available: https://www.gov.uk/government/publications/prime-ministers-challenge-on-dementia-2020/prime-ministers-challenge-on-dementia-2020.

[5] A. Association, "2019 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 15, no. 3, pp. 321-387, 2019.

[6] J. Lehtisalo *et al.*, "Dietary changes and cognition over 2 years within a multidomain intervention trial–The Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER)," *Alzheimer's and Dementia,* Article vol. 15, no. 3, pp. 410-417, 2019.

[7] A. Rosenberg *et al.*, "Multidomain lifestyle intervention benefits a large elderly population at risk for cognitive decline and dementia regardless of baseline characteristics: The FINGER trial," *Alzheimer's & Dementia,* vol. 14, no. 3, pp. 263-270, 2018.

[8] A. Solomon, M. Kivipelto, J. L. Molinuevo, B. Tom, and C. W. Ritchie, "European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): study protocol," *BMJ Open,* vol. 8, no. 12, p. e021017, 2018.

[9] G. Livingston *et al.*, "Dementia prevention, intervention, and care," *The Lancet,* vol. 390, no. 10113, pp. 2673-2734, 2017.

[10] A. Sommerlad and G. Livingston, "Preventing Alzheimer's dementia," *BMJ,* vol. 359, 2017.

[11] M. Kivipelto, F. Mangialasche, and T. Ngandu, "Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer's disease," *Nature Reviews Neurology,* Review vol. 14, no. 11, pp. 653-666, 2018.

[12] M. Kivipelto *et al.*, "World Wide Fingers will advance dementia prevention," *The Lancet Neurology,* vol. 17, no. 1, p. 27, 2018.

[13] A. Association, "2018 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 14, no. 3, pp. 367-429, 2018.

[14] C. R. Jack Jr *et al.*, "NIA-AA research framework: toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia,* vol. 14, no. 4, pp. 535-562, 2018.

[15] R. A. Sperling *et al.*, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 7, no. 3, pp. 280-292, 2011.

[16] C. R. Jack *et al.*, "Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 7, no. 3, pp. 257-262, 2011.

[17] M. Prince, R. Bryce, C. Ferri, and I. Alzheimer's Disease, "World Alzheimer's report 2011 : the benefits of early diagnosis and intervention," (in English), 2011.

[18] B. Dubois *et al.*, "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 12, no. 3, pp. 292-323, 2016.

[19] S. M. de Almeida *et al.*, "Incidence of post-dural puncture headache in research volunteers," *Headache: The Journal of Head and Face Pain,* vol. 51, no. 10, pp. 1503-1510, 2011.

[20] S. Lista, F. Faltraco, D. Prvulovic, and H. Hampel, "Blood and plasma-based proteomic biomarker research in Alzheimer's disease," *Progress in neurobiology,* vol. 101, pp. 1-17, 2013.

[21] L. Shi *et al.*, "A decade of blood biomarkers for Alzheimer's disease research: an evolving field, improving study designs, and the challenge of replication," *Journal of Alzheimer's Disease,* no. Preprint, pp. 1-18, 2018.

[22] M. S. Albert *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia,* vol. 7, no. 3, pp. 270-279, 2011.

[23] M. M. González, *Atlas of Biomarkers for Alzheimer's Disease*. Springer, 2014.

[24] G. M. McKhann *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," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 7, no. 3, pp. 263-269, 2011.

[25] C. G. Lyketsos *et al.*, "Position statement of the American Association for Geriatric Psychiatry regarding principles of care for patients with dementia resulting from Alzheimer's disease," *The American journal of geriatric psychiatry,* vol. 14, no. 7, pp. 561-573, 2006.

[26] A. Association, "2020 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 16, no. 3, pp. 391-460, 2020.

[27] L. C. Silbert, "Does statin use decrease the amount of Alzheimer's disease pathology in the brain?," *Neurology,* vol. 69, no. 9, pp. e8-11, 2007.

[28] B. J. Hanseeuw *et al.*, "Association of amyloid and tau with cognition in preclinical Alzheimer's disease: a longitudinal study," *JAMA neurology,* vol. 76, no. 8, pp. 915-924, 2019.

[29] C. Sato *et al.*, "Tau kinetics in neurons and the human central nervous system," *Neuron,* vol. 97, no. 6, pp. 1284-1298. e7, 2018.

[30] R. Khoury and E. Ghossoub, "Diagnostic biomarkers of Alzheimer's disease: A state-of-the-art review," *Biomarkers in Neuropsychiatry,* vol. 1, p. 100005, 2019.

[31] J. A. AJ *et al.*, "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical pharmacology and therapeutics,* vol. 69, no. 3, pp. 89-95, 2001.

[32] C. Humpel, "Identifying and validating biomarkers for Alzheimer's disease," *Trends in biotechnology,* vol. 29, no. 1, pp. 26-32, 2011.

[33] P. Davies *et al.*, "Consensus report of the working group on:'Molecular and biochemical markers of Alzheimer's disease'," *Neurobiology of Aging,* vol. 19, no. 2, pp. 109-116, 1998.

[34] L. M. Bekris, C.-E. Yu, T. D. Bird, and D. W. Tsuang, "Genetics of Alzheimer's disease," *Journal of geriatric psychiatry and neurology,* vol. 23, no. 4, pp. 213-227, 2010.

[35] M. Cruts, J. Theuns, and C. Van Broeckhoven, "Locus-specific mutation databases for neurodegenerative brain diseases," *Human mutation,* vol. 33, no. 9, pp. 1340-1344, 2012.

[36] J. S. Goldman *et al.*, "Genetic counseling and testing for Alzheimer's disease: joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors," *Genetics in medicine,* vol. 13, no. 6, pp. 597-605, 2011.

[37] W. J. Strittmatter *et al.*, "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer's disease," *Proceedings of the National Academy of Sciences,* vol. 90, no. 5, pp. 1977-1981, 1993.

[38] D. M. Holtzman, J. Herz, and G. Bu, "Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer's disease," *Cold Spring Harbor perspectives in medicine,* vol. 2, no. 3, p. a006312, 2012.

[39] N. R. Evans, J. M. Tarkin, J. R. Buscombe, H. S. Markus, J. H. Rudd, and E. A. Warburton, "PET imaging of the neurovascular interface in cerebrovascular disease," *Nature Reviews Neurology,* vol. 13, no. 11, pp. 676-688, 2017.

[40] F. Márquez and M. A. Yassa, "Neuroimaging biomarkers for Alzheimer's disease," *Molecular neurodegeneration,* vol. 14, no. 1, p. 21, 2019.

[41] Y.-N. Ou *et al.*, "FDG-PET as an independent biomarker for Alzheimer's biological diagnosis: a longitudinal study," *Alzheimer's research & therapy,* vol. 11, no. 1, p. 57, 2019.

[42] K. Blennow, B. Dubois, A. M. Fagan, P. Lewczuk, M. J. de Leon, and H. Hampel, "Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease," *Alzheimer's & Dementia,* vol. 11, no. 1, pp. 58-69, 2015.

[43] W. J. Strittmatter, "Bathing the brain," *The Journal of clinical investigation,* vol. 123, no. 3, pp. 1013-1015, 2013.

[44] O. V. Forlenza *et al.*, "Cerebrospinal fluid biomarkers in Alzheimer's disease: diagnostic accuracy and prediction of dementia," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring,* vol. 1, no. 4, pp. 455-463, 2015.

[45] S. Palmqvist, N. Mattsson, O. Hansson, and A. s. D. N. Initiative, "Cerebrospinal fluid analysis detects cerebral amyloid-β accumulation earlier than positron emission tomography," *Brain,* vol. 139, no. 4, pp. 1226-1236, 2016.

[46] K. Henriksen *et al.*, "The future of blood-based biomarkers for Alzheimer's disease," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 10, no. 1, pp. 115-131, 2014.

[47] H. Hampel *et al.*, "Blood-based biomarkers for Alzheimer's disease: mapping the road to the clinic," *Nature Reviews Neurology,* vol. 14, no. 11, pp. 639-652, 2018.

[48] S. Patel, R. J. Shah, P. Coleman, and M. Sabbagh, "Potential peripheral biomarkers for the diagnosis of Alzheimer's disease," *International Journal of Alzheimer's Disease,* vol. 2011, 2011.

[49] E. Zenaro, G. Piacentino, and G. Constantin, "The blood-brain barrier in Alzheimer's disease," *Neurobiology of disease,* vol. 107, pp. 41-56, 2017.

[50] M. A. Erickson and W. A. Banks, "Blood-brain barrier dysfunction as a cause and consequence of Alzheimer's disease," *Journal of Cerebral Blood Flow & Metabolism,* vol. 33, no. 10, pp. 1500-1513, 2013.

[51] A. Carrano, J. J. Hoozemans, S. M. van der Vies, A. J. Rozemuller, J. van Horssen, and H. E. de Vries, "Amyloid beta induces oxidative stress-mediated blood-brain barrier changes in capillary amyloid angiopathy," *Antioxidants & redox signaling,* vol. 15, no. 5, pp. 1167-1178, 2011.

[52] R. Deane and B. V. Zlokovic, "Role of the blood-brain barrier in the pathogenesis of Alzheimer's disease," *Current Alzheimer's Research,* vol. 4, no. 2, pp. 191-197, 2007.

[53] B. Zipser *et al.*, "Microvascular injury and blood-brain barrier leakage in Alzheimer's disease," *Neurobiology of aging,* vol. 28, no. 7, pp. 977-986, 2007.

[54] S.-Y. Yang, M.-J. Chiu, T.-F. Chen, and H.-E. Horng, "Detection of plasma biomarkers using immunomagnetic reduction: a promising method for the early diagnosis of Alzheimer's disease," *Neurology and therapy,* vol. 6, no. 1, pp. 37-56, 2017.

[55] H. Zetterberg, "Blood-based biomarkers for Alzheimer's disease–An update," *Journal of neuroscience methods,* vol. 319, pp. 2-6, 2019.

[56] H. Hampel *et al.*, "A precision medicine initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling," *Climacteric,* vol. 20, no. 2, pp. 107-118, 2017.

[57] H. Hampel *et al.*, "Precision medicine-the golden gate for detection, treatment and prevention of Alzheimer's disease," *The journal of prevention of Alzheimer's disease,* vol. 3, no. 4, p. 243, 2016.

[58] R. Trevethan, "Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice," *Frontiers in public health,* vol. 5, p. 307, 2017.

[59] L. G. Portney and M. P. Watkins, *Foundations of clinical research: applications to practice*. Pearson/Prentice Hall Upper Saddle River, NJ, 2009.

[60] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian journal of ophthalmology,* vol. 56, no. 1, p. 45, 2008.

[61] A. G. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specificity," *Continuing Education in Anaesthesia Critical Care & Pain,* vol. 8, no. 6, pp. 221-223, 2008.

[62] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, Springer, pp. 1015-1021, 2006.

[63] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics,* vol. 21, no. 1, p. 6, 2020.

[64] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061,* 2020.

[65] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of MCC and CEN error measures in multi-class prediction," *PloS one,* vol. 7, no. 8, p. e41882, 2012.

[66] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData mining,* vol. 10, no. 1, p. 35, 2017.

[67] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143, no. 1, pp. 29-36, 1982.

[68] D. J. Hand, "Evaluating diagnostic tests: the area under the ROC curve and the balance of errors," *Statistics in medicine,* vol. 29, no. 14, pp. 1502-1510, 2010.

[69] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. R. Dougherty, "Small-sample precision of ROC-related estimates," *Bioinformatics,* vol. 26, no. 6, pp. 822-830, 2010.

[70] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS one,* vol. 10, no. 3, p. e0118432, 2015.

[71] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[72] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[73] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development,* vol. 3, no. 3, pp. 210-229, 1959.

[74] T. M. Mitchell, *Machine learning*. New York: McGraw Hill. , 1997.

[75] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[76] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing,* vol. 267, pp. 664-681, 2017.

[77] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*: Springer, 2005, pp. 321-352.

[78] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The computer journal,* vol. 41, no. 8, pp. 578-588, 1998.

[79] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[80] A. Fahad *et al.*, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing,* vol. 2, no. 3, pp. 267-279, 2014.

[81] P. H. Sneath and R. R. Sokal, *Numerical taxonomy. The principles and practice of numerical classification*. 1973.

[82] B. King, "Step-wise clustering procedures," *Journal of the American Statistical Association,* vol. 62, no. 317, pp. 86-101, 1967.

[83] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association,* vol. 58, no. 301, pp. 236-244, 1963.

[84] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *ACM sigmod record,* vol. 25, no. 2, pp. 103-114, 1996.

[85] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," *ACM Sigmod record,* vol. 27, no. 2, pp. 73-84, 1998.

[86] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Information systems,* vol. 25, no. 5, pp. 345-366, 2000.

[87] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer,* vol. 32, no. 8, pp. 68-75, 1999.

[88] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14: Oakland, CA, USA, pp. 281-297.

[89] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes clustering," *Journal of classification,* vol. 18, no. 1, pp. 35-55, 2001.

[90] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: an Introduction to Cluster Analysis," 1990.

[91] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & geosciences,* vol. 10, no. 2-3, pp. 191-203, 1984.

[92] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE transactions on knowledge and data engineering,* vol. 14, no. 5, pp. 1003-1016, 2002.

[93] V. Estivill-Castro and J. Yang, "Fast and robust general purpose clustering algorithms," in *Pacific Rim International Conference on Artificial Intelligence*, 2000: Springer, pp. 208-218.

[94] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine learning,* vol. 42, no. 1, pp. 143-175, 2001.

[95] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, 1996, vol. 96, no. 34, pp. 226-231.

[96] X. Xu, M. Ester, H.-P. Kriegel, and J. Sander, "A distribution-based clustering algorithm for mining in large spatial databases," in *Proceedings 14th International Conference on Data Engineering*, IEEE, pp. 324-331, 1998.

[97] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *KDD*, 1998, vol. 98, pp. 58-65.

[98] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM Sigmod record,* vol. 28, no. 2, pp. 49-60, 1999.

[99] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 94-105.

[100] M. Ishida, H. Takakura, and Y. Okabe, "High-performance intrusion detection using optigrid clustering and grid-based labelling," in *2011 IEEE/IPSJ International Symposium on Applications and the Internet*, IEEE, pp. 11-19, 2011.

[101] W. Wang, J. Yang, and R. Muntz, "STING: A statistical information grid approach to spatial data mining," in *VLDB*, 1997, vol. 97: Citeseer, pp. 186-195.

[102] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: a wavelet-based clustering approach for spatial data in very large databases," *The VLDB Journal,* vol. 8, no. 3, pp. 289-304, 2000.

[103] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine learning,* vol. 2, no. 2, pp. 139-172, 1987.

[104] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial intelligence,* vol. 40, no. 1-3, pp. 11-61, 1989.

[105] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on neural networks,* vol. 11, no. 3, pp. 586-600, 2000.

[106] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software, 1984.

[107] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. IEEE, pp. 278-282, 1995.

[108] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[109] R. Hecht-Nielsen, "Neurocomputing: picking the human brain," *IEEE spectrum,* vol. 25, no. 3, pp. 36-41, 1988.

[110] R. J. Schalkoff, *Artificial neural networks*. McGraw-Hill Higher Education, 1997.

[111] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer,* vol. 29, no. 3, pp. 31-44, 1996.

[112] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[113] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning,* vol. 46, no. 1-3, pp. 389-422, 2002.

[114] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer genomics & proteomics,* vol. 15, no. 1, pp. 41-51, 2018.

[115] B. Scholkopf *et al.*, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Transactions on Signal Processing,* vol. 45, no. 11, pp. 2758-2765, 1997.

[116] C. S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor, "Early Detection of Alzheimer's Disease with Blood Plasma Proteins Using Support Vector Machines," *IEEE journal of biomedical and health informatics,* vol. 25, no. 1, pp. 218-226, 2020.

[117] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks,* vol. 12, no. 6, pp. 783-789, 1999.

[118] D. H. Wolpert, "Stacked generalization," *Neural networks,* vol. 5, no. 2, pp. 241-259, 1992.

[119] D. Anderson and K. Burnham, "Model selection and multi-model inference," *Second. NY: Springer-Verlag,* vol. 63, no. 2020, p. 10, 2004.

[120] G. Seni and J. F. Elder, "Ensemble methods in data mining: improving accuracy through combining predictions," *Synthesis lectures on data mining and knowledge discovery,* vol. 2, no. 1, pp. 1-126, 2010.

[121] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association,* vol. 78, no. 382, pp. 316-331, 1983.

[122] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[123] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[124] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[125] B. Efron, "Second thoughts on the bootstrap," *Statistical science,* pp. 135-140, 2003.

[126] G. A. Young, "Bootstrap: More than a Stab in the Dark?," *Statistical Science,* pp. 382-395, 1994.

[127] M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper," in *FLAIRS conference*, 1999, vol. 1999, pp. 235-239.

[128] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*: Elsevier, 1992, pp. 249-256.

[129] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence,* vol. 97, no. 1-2, pp. 245-271, 1997.

[130] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence,* vol. 97, no. 1-2, pp. 273-324, 1997.

[131] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine learning proceedings 1994*, Elsevier, pp. 121-129, 1994,.

[132] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, no. Mar, pp. 1157-1182, 2003.

[133] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems,* vol. 34, no. 3, pp. 483-519, 2013.

[134] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," 1998.

[135] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[136] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial intelligence,* vol. 151, no. 1-2, pp. 155-176, 2003.

[137] Z. Zhao and H. Liu, "Searching for interacting features," in *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1156–1161, 2007.

[138] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *European conference on machine learning*, Springer, pp. 171-182, 1994.

[139] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 27, no. 8, pp. 1226-1238, 2005.

[140] R. Gutierrez-Osuna, "Pattern analysis for machine olfaction: A review," *IEEE Sensors journal,* vol. 2, no. 3, pp. 189-202, 2002.

[141] L. Xu, P. Yan, and T. Chang, "Best first strategy for feature selection," in *9th international conference on pattern recognition*, IEEE Computer Society, pp. 706-708, 1988.

[142] H. Vafaie and K. A. De Jong, "Genetic Algorithms as a Tool for Feature Selection in Machine Learning," in *ICTAI*, 1992, pp. 200-203.

[143] H. Vafaie and K. De Jong, "Robust feature selection algorithms," in *Proceedings of 1993 ieee conference on tools with al (tai-93)*, IEEE, pp. 356-363, 1993.

[144] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science,* vol. 220, no. 4598, pp. 671-680, 1983.

[145] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on computers,* vol. 26, no. 09, pp. 917-922, 1977.

[146] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *IEEE transactions on Information Theory,* vol. 9, no. 1, pp. 11-17, 1963.

[147] W. Siedlecki and J. Sklansky, "On automatic feature selection," in *Handbook of pattern recognition and computer vision*: World Scientific, 1993, pp. 63-87.

[148] M. Mejía-Lavalle, E. Sucar, and G. Arroyo, "Feature selection with a perceptron neural net," in *Proceedings of the international workshop on feature selection for data mining*, 2006, pp. 131-135.

[149] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 58, no. 1, pp. 267-288, 1996.

[150] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology),* vol. 67, no. 2, pp. 301-320, 2005.

[151] J. Li *et al.*, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR),* vol. 50, no. 6, pp. 1-45, 2017.

[152] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007.

[153] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 13, no. 5, pp. 971-989, 2015.

[154] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," *Journal of Biomedical Informatics,* vol. 43, no. 1, pp. 15-23, 2010.

[155] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: homogeneous and heterogeneous approaches," *Knowledge-Based Systems,* vol. 118, pp. 124-139, 2017.

[156] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, 2012: IEEE, pp. 356-363.

[157] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World Alzheimer's report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[158] P. Schneider, H. Hampel, and K. Buerger, "Biological marker candidates of Alzheimer's disease in blood, plasma, and serum," *CNS neuroscience & therapeutics,* vol. 15, no. 4, pp. 358-374, 2009.

[159] S. Ray *et al.*, "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins," *Nature medicine,* vol. 13, no. 11, p. 1359, 2007.

[160] S. E. O'Bryant *et al.*, "A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI," *PloS one,* vol. 6, no. 12, p. e28092, 2011.

[161] D. A. Llano, V. Devanarayan, A. J. Simon, and A. s. D. N. Initiative, "Evaluation of plasma proteomic data for Alzheimer's disease state classification and for the prediction of progression from mild cognitive impairment to Alzheimer's disease," *Alzheimer's Disease & Associated Disorders,* vol. 27, no. 3, pp. 233-243, 2013.

[162] J. D. Doecke *et al.*, "Blood-based protein biomarkers for diagnosis of Alzheimer disease," *Archives of neurology,* vol. 69, no. 10, pp. 1318-1325, 2012.

[163] L.-H. Guo, P. Alexopoulos, S. Wagenpfeil, A. Kurz, R. Perneczky, and A. s. D. N. Initiative, "Plasma proteomics for the identification of Alzheimer's disease," *Alzheimer's disease and associated disorders,* vol. 27, no. 4, 2013.

[164] E. Jammeh, P. Zhao, C. Carroll, S. Pearson, and E. Ifeachor, "Identification of blood biomarkers for use in point of care diagnosis tool for Alzheimer's disease," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, IEEE, pp. 2415-2418, 2016.

[165] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Frontiers in Aging Neuroscience,* vol. 9, p. 329, 2017.

[166] L. K. Ferreira *et al.*, "Support vector machine-based classification of neuroimages in Alzheimer's disease: direct comparison of FDG-PET, rCBF-SPECT and MRI data acquired from the same individuals," *Revista Brasileira de Psiquiatria,* no. AHEAD, pp. 0-0, 2017.

[167] J. Ramírez *et al.*, "Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features," *Information Sciences,* vol. 237, pp. 59-72, 2013.

[168] R. Zhang *et al.*, "Mining biomarkers in human sera using proteomic tools," *Proteomics,* vol. 4, no. 1, pp. 244-256, 2004.

[169] M. Thambisetty *et al.*, "Plasma biomarkers of brain atrophy in Alzheimer's disease," *PloS one,* vol. 6, no. 12, p. e28527, 2011.

[170] J. Bauer *et al.*, "Interleukin-6 and α-2-macroglobulin indicate an acute-phase state in Alzheimer's disease cortices," *FEBS letters,* vol. 285, no. 1, pp. 111-114, 1991.

[171] L. Cucullo, N. Marchi, M. Marroni, V. Fazio, S. Namura, and D. Janigro, "Blood-brain barrier damage induces release of α2-macroglobulin," *Molecular & Cellular Proteomics,* vol. 2, no. 4, pp. 234-241, 2003.

[172] M. Thambisetty *et al.*, "Proteome-based identification of plasma proteins associated with hippocampal metabolism in early Alzheimer's disease," *Journal of neurology,* vol. 255, no. 11, pp. 1712-1720, 2008.

[173] V. R. Varma *et al.*, "Alpha-2 macroglobulin in Alzheimer's disease: a marker of neuronal injury through the RCAN1 pathway," *Molecular psychiatry,* vol. 22, no. 1, p. 13, 2017.

[174] X. Zhao *et al.*, "A candidate plasma protein classifier to identify Alzheimer's disease," *Journal of Alzheimer's Disease,* vol. 43, no. 2, pp. 549-563, 2015.

[175] S. J. Kiddle *et al.*, "Plasma based markers of [11C] PiB-PET brain amyloid burden," *PloS one,* vol. 7, no. 9, p. e44260, 2012.

[176] N. Voyle *et al.*, "Blood protein markers of neocortical amyloid-β burden: a candidate study Using SOMAscan technology," *Journal of Alzheimer's Disease,* vol. 46, no. 4, pp. 947-961, 2015.

[177] K. S. Midwood, T. Hussenet, B. Langlois, and G. Orend, "Advances in tenascin-C biology," *Cellular and molecular life sciences,* vol. 68, no. 19, p. 3175, 2011.

[178] H. D. Soares *et al.*, "Plasma biomarkers associated with the apolipoprotein E genotype and Alzheimer's disease," *Archives of neurology,* vol. 69, no. 10, pp. 1310-1317, 2012.

[179] B. Dubois *et al.*, "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria," *Alzheimer's & Dementia,* vol. 12, no. 3, pp. 292-323, 2016.

[180] G. P. Morris, I. A. Clark, and B. Vissel, "Questions concerning the role of amyloid-β in the definition, aetiology and diagnosis of Alzheimer's disease," *Acta neuropathologica,* vol. 136, no. 5, pp. 663-689, 2018.

[181] K. H. Tse and K. Herrup, "Re-imagining Alzheimer's disease–the diminishing importance of amyloid and a glimpse of what lies ahead," *Journal of neurochemistry,* vol. 143, no. 4, pp. 432-444, 2017.

[182] F. Zhang, J. Wei, X. Li, C. Ma, and Y. Gao, "Early candidate urine biomarkers for detecting Alzheimer's disease before amyloid-β plaque deposition in an APP (swe)/PSEN1 dE9 transgenic mouse model," *Journal of Alzheimer's Disease,* vol. 66, no. 2, pp. 613-637, 2018.

[183] F. Kametani and M. Hasegawa, "Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease," *Frontiers in neuroscience,* vol. 12, p. 25, 2018.

[184] M. Gold, "Phase II clinical trials of anti-amyloid β antibodies: When is enough, enough?," *Alzheimer's & Dementia: Translational Research & Clinical Interventions,* vol. 3, no. 3, pp. 402-409, 2017.

[185] S. Makin, "The amyloid hypothesis on trial," *Nature,* vol. 559, no. 7715, pp. S4-S4, 2018.

[186] H. D. Soares, Y. Chen, M. Sabbagh, A. Rohrer, E. Schrijvers, and M. Breteler, "Identifying early markers of Alzheimer's disease using quantitative multiplex proteomic immunoassay panels," *Annals of the New York Academy of Sciences,* vol. 1180, no. 1, pp. 56-67, 2009.

[187] A. Hye *et al.*, "Proteome-based plasma biomarkers for Alzheimer's disease," *Brain,* vol. 129, no. 11, pp. 3042-3050, 2006.

[188] S. Ray *et al.*, "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins," *Nature medicine,* vol. 13, no. 11, pp. 1359-1362, 2007.

[189] M. G. Ravetti and P. Moscato, "Identification of a 5-protein biomarker molecular signature for predicting Alzheimer's disease," *PloS one,* vol. 3, no. 9, p. e3111, 2008.

[190] S. E. O'Bryant *et al.*, "A blood-based algorithm for the detection of Alzheimer's disease," *Dementia and geriatric cognitive disorders,* vol. 32, no. 1, pp. 55-62, 2011.

[191] C. S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor, "Identification of Optimum Panel of Blood-based Biomarkers for Alzheimer's Disease Diagnosis Using Machine Learning," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, pp. 3991-3994, 2018.

[192] A. R. Morgan *et al.*, "Inflammatory biomarkers in Alzheimer's disease plasma," *Alzheimer's & Dementia,* vol. 15, no. 6, pp. 776-787, 2019.

[193] X. Zhao *et al.*, "A Machine Learning Approach to Identify a Circulating MicroRNA Signature for Alzheimer's Disease," *The Journal of Applied Laboratory Medicine,* vol. 5, no. 1, pp. 15-28, 2020.

[194] D. Stamate *et al.*, "A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer's disease biomarker discovery cohort," *Alzheimer's & Dementia: Translational Research & Clinical Interventions,* vol. 5, no. C, pp. 933-938, 2019.

[195] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," presented at the Proc. Int. Joint Conf. on Artificial Intelligence, Montreal, 1995.

[196] E. E. Ghiselli, *Theory of psychological measurement*. McGraw-Hill, 1964.

[197] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," 1993.

[198] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical recipes in C," *Cambridge University Press,* vol. 1, p. 3, 1988.

[199] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, no. 1, pp. 81-106, 1986.

[200] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal,* vol. 27, no. 3, pp. 379-423, 1948.

[201] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.

[202] M. N. Sabbagh, L.-F. Lue, D. Fayard, and J. Shi, "Increasing precision of clinical diagnosis of Alzheimer's disease using a combined algorithm incorporating clinical and novel biomarker data," *Neurology and therapy,* vol. 6, no. 1, pp. 83-95, 2017.

[203] T. M. Khoshgoftaar, C. Seiffert, J. Van Hulse, A. Napolitano, and A. Folleco, "Learning with limited minority class data," presented at the ICMLA 2007.

[204] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 40, no. 1, pp. 185-197, 2010.

[205] J. Lindsay *et al.*, "Risk factors for Alzheimer's disease: a prospective analysis from the Canadian Study of Health and Aging," *American Journal of Epidemiology,* vol. 156, no. 5, pp. 445-453, 2002.

[206] S. Arnold *et al.*, "Plasma biomarkers of depressive symptoms in older adults," *Translational psychiatry,* vol. 2, no. 1, p. e65, 2012.

[207] E. Jammeh, P. Zhao, C. Carroll, S. Pearson, and E. Ifeachor, "Identification of blood biomarkers for use in point of care diagnosis tool for Alzheimer's disease," presented at the Proc. IEEE Eng Med Biol Soc, Aug, 2016.

[208] M. Thambisetty and S. Lovestone, "Blood-based biomarkers of Alzheimer's disease: challenging but feasible," *Biomarkers in medicine,* vol. 4, no. 1, pp. 65-79, 2010.

[209] E. Begic, S. Hadzidedic, A. Kulaglic, B. Ramic-Brkic, Z. Begic, and M. Causevic, "SOMAscan-based proteomic measurements of plasma brain natriuretic peptide are decreased in mild cognitive impairment and in Alzheimer's dementia patients," *PloS one,* vol. 14, no. 2, 2019.

[210] X.-Y. Xu *et al.*, "Plasma levels of soluble receptor for advanced glycation end products in Alzheimer's disease," *International Journal of Neuroscience,* vol. 127, no. 5, pp. 454-458, 2017.

[211] Z. Cai *et al.*, "Role of RAGE in Alzheimer's disease," *Cellular and molecular neurobiology,* vol. 36, no. 4, pp. 483-495, 2016.

[212] E. G. Giannini, R. Testa, and V. Savarino, "Liver enzyme alteration: a guide for clinicians," (in eng), *Canadian Medical Association Journal,* vol. 172, no. 3, pp. 367-79, 2005.

[213] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics,* vol. 26, no. 3, pp. 392-398, 2010.

[214] L. I. Kuncheva, "A stability index for feature selection," in *Artificial intelligence and applications*, 2007, pp. 421-427.

[215] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn, "Random survival forests for high-dimensional data," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 4, no. 1, pp. 115-132, 2011.

[216] B. Goudey, B. J. Fung, C. Schieber, and N. G. Faux, "A blood-based signature of cerebrospinal fluid A β 1–42 status," *Scientific reports,* vol. 9, no. 1, pp. 1-12, 2019.

[217] M. Safieh, A. D. Korczyn, and D. M. Michaelson, "ApoE4: an emerging therapeutic target for Alzheimer's disease," *BMC medicine,* vol. 17, no. 1, pp. 1-17, 2019.

[218] A. K. Huber, D. A. Giles, B. M. Segal, and D. N. Irani, "An emerging role for eotaxins in neurodegenerative disease," *Clinical Immunology,* vol. 189, pp. 29-33, 2018.

[219] Q. Zhou *et al.*, "Association between APOC1 polymorphism and Alzheimer's disease: a case-control study and meta-analysis," *PloS one,* vol. 9, no. 1, 2014.

[220] M. Prendecki *et al.*, "Biothiols and oxidative stress markers and polymorphisms of TOMM40 and APOC1 genes in Alzheimer's disease patients," *Oncotarget,* vol. 9, no. 81, p. 35207, 2018.

[221] T. Lechner *et al.*, "Chromogranin peptides in Alzheimer's disease," *Experimental gerontology,* vol. 39, no. 1, pp. 101-113, 2004.

[222] C. E. Teunissen *et al.*, "Plasma Amyloid-β (Aβ 42) Correlates with Cerebrospinal Fluid Aβ 42 in Alzheimer's Disease," *Journal of Alzheimer's Disease,* vol. 62, no. 4, pp. 1857-1863, 2018.

[223] O. Hanon *et al.*, "Plasma amyloid levels within the Alzheimer's process and correlations with central biomarkers," *Alzheimer's & Dementia,* vol. 14, no. 7, pp. 858-868, 2018.

[224] J. Martin, *Rapid application development*. Macmillan Publishing Co., Inc., 1991.

[225] J. Sethi, M. Van Bulck, A. Suhail, M. Safarzadeh, A. Perez-Castillo, and G. Pan, "A label-free biosensor based on graphene and reduced graphene oxide dual-layer for electrochemical determination of beta-amyloid biomarkers," *Microchimica Acta,* vol. 187, no. 5, pp. 1-10, 2020.

[226] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM,* vol. 63, no. 11, pp. 139-144, 2020.

[227] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging,* vol. 6, no. 6, p. 52, 2020.

[228] X. Li *et al.*, "Systematic analysis and biomarker study for Alzheimer's disease," *Scientific reports,* vol. 8, no. 1, p. 17394, 2018.

[229] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: from experimental results to computational models," *Briefings in bioinformatics,* vol. 20, no. 2, pp. 515-539, 2019.

[230] C. A. DeMarshall *et al.*, "Detection of Alzheimer's disease at mild cognitive impairment and disease progression using autoantibodies as blood-based biomarkers," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring,* vol. 3, pp. 51-62, 2016.

[231] A. H. Al-Nuaimi *et al.*, "Robust EEG-Based Biomarkers to Detect Alzheimer's Disease," *Brain Sciences,* vol. 11, no. 8, p. 1026, 2021.