UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY
& BIOINFORMATICS

ANA MARTA RODRIGUES PEREIRA DA COSTA
BSc in Cellular and Molecular Biology

Leveraging Machine Learning for Predictive Modelling in Alzheimer's Disease: Integrating Trained Immunity, Infectious Burden, and Serum Cytokine Data

Sep, 2024

ITQB NOVA
NOVA SCHOOL OF SCIENCE & TECHNOLOGY
NOVA MEDICAL SCHOOL
NOVA IMS Information Management School
DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL UNIVERSIDADE NOVA DE LISBOA

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade NOVA de Lisboa

LEVERAGING MACHINE LEARNING FOR PREDICTIVE MODELLING IN ALZHEIMER'S DISEASE

by

Ana Marta Rodrigues Pereira da Costa

Dissertation presented as partial requirement for obtaining the Master's degree in Computational Biology and Bioinformatics

Adviser: Prof. Dr. Leonardo Vanneschi

Co-adviser: Liah Rosenfeld

September, 2024

Leveraging Machine Learning for Predictive Modelling in Alzheimer's Disease

Integrating Trained Immunity, Infectious Burden, and Serum Cytokine Data

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.14) [1].

To grandma.

Acknowledgements

During the course of this thesis, I have received a great deal of support and assistance, whithout which this work wouldn't have been possible, and for this I would like to express my deepest gratitude.

Firstly, I'm extremely grateful to my esteemed professor and supervisor, Leonardo Vanneschi, for his mentorship, constructive feedback and reassuring words. He has been an inspiration in my academic endeavours, and it was through his guidance and encouragement that I was able to delve into the world of data science.

I would also like to express my sincere appreciation to my co-supervisor, Liah Rosenfeld, who provided me with her expertise and kind advisement, along with patience and understanding througout this learning process.

This work would also not have been possible without the continuous support, advice and proficiency of Dr. Paola Bossú and Iliana Piccolino at Santa Lucia Foundation, as well as Professor Francesco Fontanella and Professor Claudio De Stefano.

In addition, I would like to extend my heartfelt gratitude for everyone at the PhD hall, especially Davide Farinati, Berfin Sakallioglu, Emanuele Nardone, Giovanni Pinna and Lena Dewaele. For all the scientific support, friendly incentives and sharing of knowledge, but mostly for the sunny lunch breaks and cheerful discussions in english, portuguese, italian or dutch. I bask in the friendship you have surrounded me with, this year.

Lastly, my thanks to all the wonderful people in my life. My dear family, especially my mom who taught me all about the difference women can make in the academic field and whose journey I always carry with me. To my sweet and bright best friend, Beatriz Xavier, to whom I pass the torch. And to my boyfriend, Rafael Borralho, as the saying goes "to be loved is to be known", what a privilege it is to be seen for who we truly are.

”

"We all woke up this morning and we had with it the amazing return of our conscious mind. We recovered minds with a complete sense of self and a complete sense of our own existence — yet we hardly ever pause to consider this wonder."

— **António Damásio**, Self Comes to Mind: Constructing the Conscious Brain
(Professor of Neuroscience)

Abstract

Alzheimer's disease (AD)'s complex aetheology results in a lack of effective therapies. Studies have suggested the possible involvement of infectious agents and dysregulated inflammation in its pathogenesis. Advances in Machine Learning (ML) permitted the integration and analysis of high-dimensional heterogenous data, uncovering intricate relationships and new disease biomarkers. In this dissertation ML models- including linear and logistic regression, Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Extreme Gradient Boosting (XGB), were deployed to predict AD, using Trained Immunity (TI), Infectious Burden (IB), and serum cytokine data, as well as to model TI data itself, predicting levels of TNFα, IL-6, IL-10, IL-1β and IL-1RA in response to different infectious stimuli.

Best AD-predicting models achieved 87.5% accuracy, an AUC of 0.88, and 100% recall for AD cases. SHAP analysis highlighted elevated pro-inflammatory cytokines (TNFα, IL-6, IL-1β) under primary stimulation conditions, and reduced anti-inflammatory IL-10 levels under infectious challenge, as significant predictors of AD, pointing to a dysregulated immune response. Including IB data revealed a strong impact of Herpes Simplex Virus (HSV) antibody levels in AD prediction, supporting the role of herpes in AD development. Serum-based models achieved 75-80% accuracy, with AUCs of 0.74 and 0.86. IL-18, IL-10 and IL-8 were among most impactful features in optimal models, although discrepancies in literature and our findings, suggest potential stage-specific or inflammatory subtypes in AD, highlighting immune response heterogeneity.

Models predicting TI cytokine levels showed moderate success, with models for TNFα, IL-10 and and IL-1β, being amidst the most effective. These findings suggest existence of modulating relationships between IB, age, sex, and distinct inflammatory responses in AD, although our models do not fully capture the variability of immunological response. Nevertheless, our findings demonstrate the potential of ML in understanding underlying pathways in AD.

Keywords: Alzheimer's disease, Machine learning, Inflammation, Infection hypothesis, Trained immunity, Cytokine profile

Resumo

A complexa etiologia da Doença de Alzheimer (DA) resulta na escassez de terapias eficazes. Estudos sugerem o possível envolvimento de agentes infeciosos e inflamação desregulada na sua patogénese. Avanços em Aprendizagem Automática (AA) permitiram a integração e análise de dados multi-dimensionais heterogéneos, descobrindo relações intrincadas e novos biomarcadores. Nesta dissertação, modelos de AA, como regressão linear e logística, árvores de decisão, *Random Forest*, *Support Vector Machines*, *K-Nearest Neighbors* e *Extreme Gradient Boosting* foram usados para prever a DA, utilizando dados de imunidade treinada (TI), carga infecciosa (IB) e níveis séricos de citocinas. Também foi modelada a TI, prevendo níveis de TNFα, IL-6, IL-10, IL-1β e IL-1RA em resposta a estímulos infecciosos.

Os melhores modelos para a DA atingiram 87,5% de exatidão, AUC de 0,88 e 100% de recuperação para casos de DA. A análise SHAP destacou citocinas pró-inflamatórias elevadas (TNFα, IL-6, IL-1β) em condições de estimulação primária, e níveis reduzidos de IL-10 sob desafio infecioso, como preditores significativos de DA, sugerindo uma resposta imune desregulada. A inclusão da carga infecciosa revelou o forte impacto dos anticorpos contra Herpes Simplex (HSV) na previsão da DA, apoiando o papel do herpes no seu desenvolvimento. Os modelos com dados de soro alcançaram 75-80% de exatidão, com AUCs de 0,74 e 0,86. IL-18, IL-10 e IL-8 destacaram-se como variáveis-chave, embora discrepâncias com a literatura sugiram potenciais estados específicos ou subtipos inflamatórios na DA, refletindo a heterogeneidade da resposta imune.

Modelos para os níveis de citocinas em TI demonstraram um sucesso moderado, com modelos para TNFα, IL-10 e IL-1β entre os mais eficazes. Estes resultados sugerem a existência de relações moduladoras entre IB, idade, sexo e respostas inflamatórias distintas na DA, embora os nossos modelos não capturem totalmente a variabilidade da resposta imunológica. No entanto, os nossos resultados demonstram o potencial da AA na compreensão dos mecanismos subjacentes à DA.

Palavras-chave: Doença de Alzheimer, Aprendizagem automática, Inflamação, Hipótese infecciosa, Imunidade treinada, Perfil de citocinas

Contents

List of Figures

List of Tables

Acronyms

AD Alzheimer's disease *(pp. ix, 1–11, 33–38, 47–49, 55, 57, 59–61, 63–66, 68, 82–84, 88, 90, 91, 94, 97, 99–107, 109, 110)*

AI Artificial Intelligence *(pp. 13, 28, 30, 35)*

ANN Artificial Neural Networks *(pp. 33, 34)*

AUC Area Under the Curve *(pp. 17, 34–37, 48, 49, 83, 88, 90, 101, 102, 104–106, 109, 110, 152–155)*

BDNF Brain-Derived Neurotrophic Factor *(pp. 11, 37, 65)*

Ca LPS Primed with *C. albicans* and then challenged with LPS *(pp. 44–46, 48, 68, 70, 71, 74, 82, 84, 95, 99, 109, 131, 133, 136, 139)*

CMV Cytomegalovirus *(pp. 2, 5, 6, 41, 44, 47, 48, 63, 83, 109)*

CNN Convolutional Neural Network *(p. 34)*

CSF Cerebrospinal Fluid *(pp. 8–11, 103)*

CV Cross-validation *(pp. 43, 45, 47, 49, 50, 55, 56, 109)*

DT Decision Trees *(pp. ix, 25, 26, 34, 35, 44, 51, 53, 74, 75, 78, 81, 82, 88, 97, 100, 105, 109, 129–140, 152–158)*

EN Elastic-Net *(pp. 25, 44, 51, 83, 84, 97, 101, 109, 129–140, 152–158)*

FN False Negatives *(pp. 15, 16)*

FP False Positives *(pp. 15–17)*

G-CSF Granulocyte Colony-Stimulating factor *(p. 65)*

GM-CSF Granulocyte-Macrophage Colony-Stimulating Factor *(pp. 49, 65)*

GS Grid Search *(pp. 50, 51)*

HC Healthy Controls *(pp. 34, 35, 47–49, 59–61, 63–65, 68, 83, 84, 90, 99–104, 106)*

HSV Herpes Simplex Virus *(pp. ix, 1, 2, 4, 5, 38, 41, 44, 47, 48, 61, 63, 83, 88, 105, 109, 110)*

IB Infectious Burden *(pp. ix, 1, 2, 41, 42, 44, 47, 55, 59, 63, 68, 83, 88, 97, 99, 104, 105, 109, 110)*

IgG Immunoglobulin G *(pp. 43, 61, 63, 97)*

KNN K-Nearest Neighbors *(pp. ix, 27, 34, 38, 44, 51, 53, 69, 73, 75, 109, 129–140, 152–158)*

LOO Leave-one-out *(pp. 47, 49, 56)*

LOOCV Leave-one-out Cross Validation *(pp. 45, 47, 48, 50, 55, 56, 83)*

LPS Lipopolysaccharide S *(pp. 8, 9, 41, 84, 100, 101, 105, 106, 109)*

LPS LPS Primed with LPS and then challenged with LPS *(pp. 44, 45, 48, 61, 68, 70, 71, 74, 75, 80–82, 84, 88, 94, 95, 99–101, 105–107, 109, 110, 131, 134, 136, 140)*

MAE Mean Absolute Error *(pp. 18, 45)*

MCI Mild Cognitive Impairment *(pp. 1, 9, 11, 34, 36, 37, 110)*

MdAE Median Absolute Error *(pp. 18, 45, 68–71, 73–75, 77, 78, 80, 81)*

ML Machine Learning *(pp. ix, 1, 2, 13–15, 17, 19–28, 33–36, 38, 39, 43, 46, 48, 100, 104, 105, 109, 110)*

MMSE Mini-Mental State Examination *(p. 35)*

MRI Magnetic Resonance Imaging *(pp. 10, 33, 34, 36, 37)*

NFT Neurofibrillary Tangles *(pp. 3, 103)*

NT Non Treated *(pp. 41, 44–46, 48, 68, 69, 71, 73, 74, 77, 78, 80, 81, 88, 94, 95, 99, 100, 105, 106, 109, 110, 132, 134, 137, 138)*

PET Positron Emission Tomography *(pp. 4, 10, 33, 34)*

Pr Ca Primed with *Candida albicans* pathogen *(pp. 41, 44–46, 48, 68, 70, 71, 73, 74, 78, 82, 95, 99, 106, 130, 133, 135, 138)*

Pr LPS Primed with LPS molecule *(pp. 41, 44, 45, 48, 60, 61, 68–71, 73–75, 77, 78, 80, 81, 84, 88, 95, 99, 101, 105, 106, 109, 130, 132, 133, 135, 137, 139)*

RF Random Forest *(pp. ix, 26, 27, 34, 35, 38, 39, 44, 51, 53, 71–75, 77, 78, 81, 83, 97, 99, 100, 104, 109, 129–140, 152–158)*

RMSE Root Mean squared Error *(p. 18)*

ROC Receiver Operating Characteristic *(pp. 17, 48, 49, 83, 94)*

SHAP SHapley Additive exPlanations *(pp. 28–30, 35, 43, 48–50, 59, 82, 84, 88, 94, 101, 102, 105, 109, 110)*

SMOGN Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise *(pp. 23, 43, 45, 56, 57, 75, 77, 80)*

SMOTE Synthetic Minority Over-Sampling Technique *(pp. 23, 43, 48–50, 56, 57, 68, 78, 81)*

SVC Support Vector Classifier *(pp. 53, 74, 75, 81, 82, 90, 94, 97, 100, 102, 129–140, 152–158)*

SVM Support Vector Machine *(pp. ix, 26, 27, 34, 35, 37, 38, 44, 109)*

SVR Support Vector Regressor *(pp. 26, 51, 69, 70)*

TI Trained Immunity *(pp. ix, 2, 41, 43, 44, 46–49, 55, 59, 83, 88, 94, 97, 101, 102, 105, 107, 109, 110)*

TN True Negatives *(pp. 15, 16)*

TP True Positives *(pp. 15–17)*

XGB Extreme Gradient Boosting *(pp. ix, 27, 44, 51, 53, 69, 70, 80, 99, 109, 129–140, 152–158)*

1 Introduction

AD is a devastating neurodegenerative disease, estimated to affect over 50 million people worldwide [2]. It is the most prevalent form of dementia, being progressive and irreversible, impairing function, cognition, and behavior [3]. With no current effective therapies available, AD is projected to roughly triple in its prevalence by 2050, reaching over 131.5 million people [4]. This jump is partially due to a progressively aging population, but also to the complex and multifactorial nature of this type of dementia, which has severely challenged investigator's efforts to tackle this disease [5]. In this scenario, the need for new and diverse approaches for the studying of AD becomes evident. This is where recent technological advances, such as ML techniques enter. From predicting Alzheimer's progression from Mild Cognitive Impairment (MCI) (a common early form of dementia) through medical image analysis [6, 7] to finding biomarkers of this disease and personalized treatment strategies [8–10], in the past decade, ML has proven to be a very useful and powerful tool in the fight against this debilitating disease [11].

As mentioned, a hefty obstacle to the development of effective therapies against AD is its multifaceted aetheology. As this disease is characterized by the accumulation of amyloid-β plaques and tau protein tangles in the brain [12]. Several studies have focused on these hallmarks in order to explain the underlying cause and mechanisms of AD [13–15] and as avenues of treatment for this form of dementia, however, with little efficacy [16]. In this way, a growing body of research has shifted its attention to the potential role of microbial agents and immune system dysregulation in AD pathogenesis. This mechanism is known as the infection hypothesis, which gained traction with Itzhaki's work in 1997 [17], when researchers found evidence of HSV type 1 viral DNA in the brain of AD patients. More recently, multiple papers have explored the hypothesis of a high IB, involving herpesviruses, *Chlamydia pneumoniae*, *Helicobacter pylori* and other pathogens, being associated with an increased risk of AD, possibly as a result of the chronic inflammation that these infections provoke [18, 19].

Moreover, the concept of innate immune memory, where previous inflammatory stimuli induce long-term epigenetic reprogramming of innate immune cells [20], has

emerged as a crucial factor in modifying the progression of AD. Paradoxically, this phenomenon, while possibly protective against infections, might also exacerbate neurodegeneration in AD by inducing maladaptive immune responses [21, 22]. These processes have been linked to the IL-1 family of cytokines, specifically IL-1β and IL-18, which display a dual function of promoting neuroinflammation and driving disease progression [23].

The work developed in this thesis is inserted in this context and its goal is to develop predictive and interpretable ML models for AD by using TI, IB and serum cytokine data. Our IB data encompasses 5 different pathogens: HSV type 1 and type 2 (HSV-1/2), *Helicobacter pylori*, Cytomegalovirus (CMV), *Chlamydia pneumoniae* and *Borrelia burgdorferi*; and our TI data mainly focuses on 5 different cytokines: TNF-α, IL-6, IL-10, IL-1β and IL-1RA, under varying stimulation conditions. We primarily aim to explore relationships within the data between IB, cytokine levels and AD. We also intend to identify distinct TI cytokine profiles, including pro-inflammatory and anti-inflammatory markers, in both baseline and post-stimulation conditions, to assess their predictive value for distinguishing between AD patients and controls. Additionally, we also strive to investigate the relationship between IB and cytokine profiles in the context of this disease, and, lastly, to build robust models with serum cytokine levels and evaluate potential differences between healthy individuals and patients.

Our expectations for these tasks, in line with our hypotheses, are to find higher levels of different pathogens to be predictive of AD, to encounter, through our models, a TI cytokine profile for AD patients where we expect there to be enhanced production of pro-inflammatory cytokines in priming conditions and decreased production of anti-inflammatory cytokines in challenge conditions, and to find potential serum biomarkers for AD. We also expect to be able to moderately predict TI cytokine levels based on disease status, age and IB data.

In light of this, the document is structured as follows. Chapter 2, Biological Theoretical Background, gives the reader the biological context for this thesis, including notions of the pathophysiology of Alzheimer's, the emergence of the infection hypothesis and the various roles that immunological processes and proteins play in this disease. Chapter 3, Machine Learning Theoretical Background, provides the knowledge necessary to comprehend the elaborated work, presenting a general overview of what is ML, what constitutes a traditional ML pipeline, the theory behind the algorithms deployed in this project and lastly, a definition of Shapley values, our chosen tool for model interpretability. Chapter 4, Literature Review, summarises earlier research that is pertinent to this study, including ML models for predicting AD built with various types of data with a focus on protein and other serum biomarkers, as well as, infectious data approaches. Chapter 5, Methodology, detailedly describes the framework of this project, from data provenance to parameter settings. Chapter 6 presents the results of the work and Chapter 7 its critical discussion. Finally, Chapter 8 concludes the work with a review of its findings, constraints, and suggestions for future research.

2

Biological Theoretical Background

2.1 Overview of Alzheimer's Disease and Pathophysiology

Alzheimer's disease is a condition of progressive nature. Patients are often first diagnosed with mild cognitive and/or behavioural impairment which eventually progresses onto AD dementia. Ongoing research has encouraged clinicians to diagnose AD earlier, before patients progress to the dementia stage. This early and accurate detection of symptoms and underlying pathology is crucial for effective screening, diagnosis, and management of the disease for both patients and caregivers. However, early-stage detection is challenging due to diagnostic difficulties, time constraints, and the tendency to confound symptoms with normal aging [3]. In fact, the most important risk factor in Alzheimer's is age, with the majority cases of AD onset being after 65 years of age [24].

Pathologically, AD is typically defined by memory and learning impairment and executive dysfunction which interferes with the daily life activities of the patient [25]. As far as neuropsychiatric symptoms, the most common ones in AD are depression and apathy, with aggression and psychosis being core symptoms as well [26, 27]. As the disease progresses, the patients may manifest further neuropsychiatric symptoms, such as periods of disorientation and confusion and eventually, in later stages, delusion and hallucination [28]. Whilst these symptoms are of primal concern, currently the diagnosis of AD also heavily relies on molecular biomarkers. This form of dementia is characterized by the accumulation of extracellular amyloid β ($A\beta$) plaques, as well as the accumulation of intracellular Neurofibrillary Tangles (NFT) of hyperphosphorylated tau proteins leading to progressive neuronal loss and cerebral atrophy [12].

While several studies support and describe the central role of the accumulation of $A\beta$ plaques in the brain [13, 29], and of the accumulation of NFTs [30, 31], to date the drugs that have targeted the inhibition of the aggregation/fibrillation of these compounds have faced many challenges as potential treatments for AD. Despite their molecular efficacies demonstrated both *in vivo* and *in vitro*, they often fail during clinical trials, whether that is due to inadequate result or to adverse effects [16].

These challenges might be owed to the complex and multifactorial nature of AD.

Whereas the amyloid cascade hypothesis is the most popular theory to explain AD pathogenesis, and the cascade of tau toxicity has been proved to lead to neuron damage, neuroinflammation and oxidative stress in brain, there are many other pathways strongly associated with cognitive decline in this disease. Deficiencies in various sorts of neurotransmissions are responsible for a plethora of neurodegenerative symptoms, for example cholinergic and glutamatergic deficits for cognitive decline, the excitatory and inhibitory neurotransmission dyshomeostasis for deficits in synaptic plasticity and epileptiform symptoms, and the monoaminergic neurotransmission for neuropsychiatric symptoms [5].

2.2 Neuroinflammation and the Infection Hypothesis

One can not discuss the multifactorial aetiology of AD, without referring to the central role neuroinflammation plays in its development. Neuroinflammation is defined as the brain's activation of the innate immune system, as a defence mechanism which aims to protect the central nervous system (CNS) against infectious insults, injury, or disease. This process is usually controlled and beneficial, with homeostasis being restored once the threat has been eliminated [32]. However, in many neurodegenerative disorders, such as AD, sustained neuroinflammatory processes are involved, potentially contributing to progressive neuronal damage. Research has found neuroinflammation to exacerbate $A\beta$ and tau pathologies, as evidenced by increased microglia activation observed in Positron Emission Tomography (PET) studies of AD patients. Elevated levels of pro-inflammatory cytokines are also found in both serum and brain tissues of AD patients *post mortem*. This response involves various immune cells and molecules, such as glial cells, cytokines, and chemokines, as well as complement, which collectively play an integral role in the onset and progression of AD [33].

Recently, an ongoing body of research has shifted its attention to infection as a potential trigger for the neuroinflammatory processes observed in AD. The infection hypothesis was firstly presented by Alois Alzheimer himself as a hypothetical causative explanation for the disease, and it has resurfaced in the last decades when investigators found that AD amyloid plaques contain remnants of HSV-1 viral DNA [17, 34–36]. This hypothesis posits that microbial pathogens, such as viruses and bacteria, might initiate or exacerbate inflammation in the brain, thereby promoting the development and progression of AD. Evidence supporting this hypothesis includes the previously mentioned findings of microbial DNA in the brains of AD patients and the ability of certain pathogens to induce $A\beta$ plaque formation and tau pathology, mimicking the hallmarks of AD. Exploring the infection hypothesis offers a compelling avenue for understanding the multifactorial nature of AD as well as identifying a potential new line of therapeutics to mitigate its progression [37].

2.3 Infectious Burden and Alzheimer's Disease

AD's infection hypothesis has been linked to several pathogens. Amongst them are several viruses such as HSV type 1, 2 and 6A/B, human CMV, Epstein-Barr virus, hepatitis C virus, influenza virus, and severe acute respiratory syndrome coronavirus 2, SARS-CoV-2. Various bacteria have been implicated as well, such as *Borrelia burgdorferi, Chlamydia pneumoniae, Porphyromonas gingivalis, Prevotella intermedia, Treponema pallidum, Eikenella corrodens, Treponema denticola*, and *Helicobacter pylori*, and even eukaryotic unicellular parasites as is the case of *Toxoplasma gondii* [38].

For this project, 5 of these pathogens were selected, namely HSV (type 1 and type 2) (HSV-1/2), *Helicobacter pylori*, CMV, *Chlamydia pneumoniae* and *Borrelia burgdorferi*.

2.3.1 HSV and AD

The relationship between HSV and Alzheimer's disease has been gaining track in the latest decades, ever since, in 2007, researchers found DNA of the virus allocated inside amyloid plaques of AD patients [35]. The Herpes Simplex Virus is a virus that usually causes oral and lip infections (HSV-1) and genital infections (HSV-2), having a productive phase where clinical symptoms may be expressed, followed by a latent stage where the virus is allocated within sensory neurons. Reactivation from latency can occur and result in recurrent infections [39]. Several studies indicate that infection with this virus might be associated to the development of AD, especially in individuals presenting the type 4 alleles of the apolipoprotein E gene ($APOE - \epsilon 4$) [40–42]. The hypothesis states that as HSV is reactivated from its latent form in the brain, due to events such as stress and inflammation, its productive infection leads to consequent damage which is likely greater in people with the $APOE - \epsilon 4$ gene. This cumulative damage due to the recurrent nature of the reactivation of HSV is suggested to eventually lead to AD [40].

2.3.2 *Helicobacter pylori* and AD

Helicobacter pylori is a gram-negative bacterium that inhabits the gastric environment of over 60% of today's world population [43]. Although this pathogen is mainly associated with gastritis and gastric adenocarcinoma, recent research has established an association between gut dysbiosis and several health issues, comprehending neurological abnormalities, such as AD. However, the precise mechanisms and pathways involved in these neuropathological processes are not yet fully understood [44]. There is emerging scientific evidence supporting that the 'gut-brain axis' interaction between the brain and the immune system is fundamental for the balance between homeostasis and disease, specifically at neurological level. The "Trojan Horse" hypothesis is a mechanism that has been described for several neurological disorders (including AD),

where bacteria, such as *Helicobacter pylori* cross the blood brain barrier, reaching the CNS, where they cause damage leading to neurodegeneration [45].

2.3.3 CMV and AD

Cytomegalovirus (CMV) is a virus from the Herpesviridae family, being known also as human herpesvirus 5. Despite not usually yielding symptoms in healthy individuals, post infection the virus remains in the host's body and several factors can lead to its reactivation. This pathogen has been linked to numerous neurological diseases, amongst them Alzheimer's disease [46]. Again, while the mechanisms of action behind this association are not yet fully understood, it is suggested that the presence of herpesvirus in the brain triggers the formation of extracellular amyloid plaques and intraneuronal hyperphosphorylated peptide aggregates. On the other hand the recurring nature of this infection leading to continuous immune response with a consequent chronic inflammation, could also constitute the underlying mechanisms leading to AD [47].

2.3.4 *Chlamydia pneumoniae* and AD

Chlamydia pneumoniae is an intracellular bacterium member of the genus Chlamydia, responsible for acute respiratory disease [48]. It is one of the most consistent bacterial infections detected in AD brains [49]. Research shows there are likely 3 major routes for chlamydia to reach the brain and affect its biology. First, the systemic effect, when chlamydia infection in specific organs triggers a body-wide response. Inflammatory molecules and potentially even the pathogen itself might travel through the bloodstream and reach the brain. Secondly, the aforementioned Trojan horse, in which Chlamydia is carried to the brain by hiding inside immune cells. Lastly, the direct nasal infection, since *C. pneumoniae* can directly infect the cells in the nasal cavity and a nerve connecting the nose to the brain (the trigeminal nerve) might serve as a direct pathway for *C. pneumoniae* to invade the CNS. Once the pathogen has reached the brain its potential mechanistic pathways affecting AD development and progression are not unlike the ones described before. Studies show it can lead to increasing $A\beta$ production, the presence of the $APOE - \epsilon 4$ gene might also facilitate its growth and attachment to neuronal cells, it may downregulate levels of Sirtuin 1 (a protein with protective effects against AD), it can increase MMP-9, an enzyme involved in triggering inflammatory cascades in AD, and lastly chlamydia infection can activate the NLRP3 inflammasome, a complex that promotes inflammation linked to AD [50].

2.3.5 *Borrelia burgdorferi* and AD

The spirochaete *Borrelia burgdorferi* is the pathogen responsible for Lyme disease, a tick-borne disease transmitted enzootically between ticks and their hosts [51]. This pathogen

is known to be neurotropic and can exist in the CNS for extensive periods eventually resulting in brain atrophy, amyloid deposition, and slow progressive dementia. The association between Borrelia and AD has been investigated for decades, with several studies finding *B. burgdorferi* in AD brain tissues [52] and co-localized expression of amyloid markers [53]. It is found that the chronic inflammation induced by this bacterium may lead to abnormal phosphorylation of the tau protein, resulting in microtubule dysfunction and formation of neurofibrillary tangles, all major hallmarks of AD [38].

While the role that each of these microbial agents may play in the development and progression of AD is of key interest, as well as their various mechanisms of action, from chronic neuroinflammation, to amyloid-β peptide production, and neurodegeneration, it is critical to consider the potential compounded effects of these pathogens when they have infected the same host. Emerging research suggests that $A\beta$ plaque formation might be an immune response to microbial infection. It is posed that the interplay between these pathogens can exacerbate the disease's progression, while AD patients are comprised of a weakened immune system and compromised blood-brain barrier, being particularly susceptible to infections and potentially entering a perpetuate cycle of neuroinflammation and neurodegeneration. This multifaceted pathogen hypothesis could explain the complex and heterogeneous nature of Alzheimer's disease, in which the amyloid-centric approaches so far have yielded disappointing treatments [19].

2.4 Trained Immunity and Alzheimer's Disease

Since we have explored the infection hypothesis for AD it is important to analyse the body's immune response observed in AD. The human immune system consists of two principal stages of response: the innate immune response and the adaptive immune response. On the first line of defence we have the innate immune response. Fast and non-specific, it is mediated by cells such as monocytes, macrophages, dendritic cells, natural-killer cells, among others. These cells detect molecular patterns associated to pathogens and to damage, inducing the release of pro-inflammatory cytokines, such as IL-6, IL-1β and Tumor Necrosis Factor-α (TNF-α). These subsequently lead to the activation of the adaptive immune response, a slower response mediated by memory T and B lymphocytes that yields long-term immune memory [54]. The concept of trained immunity challenges the idea that only the adaptive response possesses memory. It describes functional long-term reprogramming of innate immune cells, giving rise to altered, faster responses towards a second challenge by a pathogen, after the return to a non-activated state. Despite this type of immune response being crucial for the protection against infections, it has also been found to lead to aberrant inflammatory activity in immune-mediated conditions and chronic inflammatory diseases [20].

In Alzheimer's disease, it is hypothesized that innate immune cells (both peripherally and in the brain) may retain memory of past stimulations, altering brain immune

responses to $A\beta$. This, in turn, might lead $A\beta$ to accumulate and the disease to progress. However, since patients are exposed to a multitude of pathogens during their lifetime, innate immune memory responses are expected to be heterogeneous, aligning with clinical differences in disease progression. In pre-symptomatic and early stages of AD, it is suggested that this trained immunity response may enhance pro-inflammatory cytokine release and $A\beta$ production, worsening brain damage. In contrast, at later stages, persistent stimuli, such as $A\beta$, may lead to trained tolerance with decreased production of inflammatory cytokines, with the immune response shifting towards maintenance and repair, though often ineffectively due to the maladaptive nature of the response [22]. In fact a study, with a mouse model of AD pathology demonstrated that peripherally applied inflammatory stimuli of Lipopolysaccharide S (LPS) induced acute immune training and tolerance in the brain, leading to differential epigenetic reprogramming of brain-resident macrophages (microglia) and found that immune training exacerbates cerebral β-amyloidosis while immune tolerance alleviates it [21].

Since neuroinflammation is a central contributor to several aspects of AD pathology, in order to unravel the complex mechanisms of trained immunity behind this disease, it is crucial to look into cytokine signalling. Cytokines are soluble extracellular proteins, or glycoproteins which play an important role as intercellular regulators and mobilizers of cells involved in both innate and adaptive inflammatory host defences, angiogenesis, cell growth, differentiation, and death, as well as development and repair processes meant to restore homeostasis. While some cytokines are sometimes expressed constitutively, most nucleated cell types normally produce them in response to harmful stimuli [55]. In terms of biological function, cytokines mainly divide into two groups, proinflammatory and anti-inflammatory. Pro-inflammatory cytokines such as TNF-α, Interferon-γ (IFN-γ) or IL-1β, mainly promote inflammation, whereas anti-inflammatory cytokines have the role of suppressing the activity of said proinflammatory cytokines, as is the case of, for example, IL-4, IL-10, and IL-13 [56]. In certain cases and infections, some cytokines also exert regulatory functions, being involved in the control of immune responses elicited by pro- and anti-inflammatory cytokines, ensuring the immune system doesn't overreact or underreact [57].

In the context of AD, multiple cytokines have been described to play an important role in its pathogenesis. For this project, the focus was on 5 main cytokines, TNF-α, IL-6, IL-10, IL-1β and IL-1RA, while some others were looked at in the context of serum biomarkers as will be discussed ahead.

2.4.1 Role of TNF-α in AD

The role of Tumor Necrosis Factor-α in AD remains rather puzzling, as some studies focus on its inhibition as a potential therapeutic, whilst others demonstrate beneficial roles for this protein. On one hand, higher Cerebrospinal Fluid (CSF) levels of TNF-α have shown correlation with reduced functional connectivity in the brain. Additionally,

mice deficient in the receptor to this cytokine show decreased levels of $A\beta$, lower expression of inflammatory factors, CSF-blood barrier integrity, and preserved memory compared to their control counterparts. On the other hand, a study showed an injection of murine TNF-α into the hippocampus of mice, to reduce $A\beta$ accumulation and stimulate more microglial responses, suggesting TNF-α may play a role in plaque clearance under specific temporal and spatial conditions [58].

2.4.2 Role of IL-6 in AD

Interleukin-6 is a proinflammatory cytokine, which is involved in the regulation of haematopoiesis and the coordination of the innate and acquired immune responses. It is fundamental in the regulation of metabolism, in neural development and survival, while also participating in several cancerous processes [59]. In the context of AD, studies have shown both serum and CSF IL-6 levels to be increased in AD patients, correlating also with disease severity [60, 61]. IL-6 participates in early-stage $A\beta$ plaque formation in AD brains and has been implicated in tau phosphorylation, synapse loss, and learning deficits in mice, with one study in particular reporting higher IL-6 levels to be associated with smaller brain volumes in patients and lower cognitive scores. In the same study, neutralizing IL-6 in AD mouse models enhanced metabolic issues, memory, and decreased IL-6 levels [62]. An additional paper, found genetic variations in the IL-10 and IL-6 genes to be associated with AD [63].

2.4.3 Role of IL-10 in AD

Interleukin-10 is one of the most important anti-inflammatory cytokines. It is produced by various immune cells, and serves primarily to regulate and suppress inflammatory responses [64]. In AD, lower serum IL-10 levels have been found to correlate with CSF $A\beta$ deposition. [65] One study concluded that loss of IL-10 activates microglia, enhances IL-6, and leads to hyperphosphorylation of tau on AD-relevant epitopes in response to acute systemic inflammation, induced by LPS stimulation [66]. Thus, the role of IL-10 in AD remains complex and appears to have dual nature effects. On one hand, low levels of IL-10 are associated with increased susceptibility to AD, and its overexpression in the hippocampus of AD transgenic mice has been shown to boost neurogenesis and cognition, indicating a possible neuroprotective role. On the other hand, some contradictory studies suggested that IL-10 may worsen cognitive decline in mice models. For instance, IL-10 genetic ablation in transgenic mice led to a reduction in $A\beta$ plaques and cerebral amyloid angiopathy. Additionally, genetic variations in the IL-10 promoter region have been linked to the development of amnestic MCI, a previously referred precursor to AD [67]. Therefore, while IL-10 has potential neuroprotective effects, its role in AD pathology is not straightforward and may vary depending on specific genetic and molecular contexts.

2.4.4 Role of IL-1β and IL-1RA in AD

Interleukin-1β is a proinflammatory cytokine from the IL-1 family. Interleukin-1 receptor antagonist (IL-1RA), from the same family, is an anti-inflammatory cytokine, that serves as an inhibitor of IL-1 proteins. In AD, both these cytokines are dysregulated and display complex and multifaceted roles. IL-1β is important in normal brain functions such as learning and memory, but at aberrant levels, it contributes to infection- and inflammation-induced cognitive dysfunction through pathways resulting in neuroinflammation. This cytokine targets and activates a range of different cells, resulting in a multitude of responses that contribute to neuroinflammation. In AD, IL-1β has mediating roles in neuroinflammation that exacerbate the pathology of the disease, but also acts as a protective factor, influencing the balance between beneficial and detrimental outcomes. The importance of the modulating activity IL-1RA exerts over IL-1β is illustrated by the fact that mice lacking IL-1RA presented worsened AD-like pathologies. Furthermore, variations in the levels of IL-1RA and different IL-1 receptors in the brains and blood of AD patients have been observed. Overall, whether and when these cytokines have a beneficial or detrimental role in neuronal function in the course of AD, and how their pathways function, is still a matter of intense and important research [23].

2.5 Serum Biomarkers in Alzheimer's Disease

Still in the context of trained immunity, it is rather important to look at these cytokines as possible serum biomarkers. As previously mentioned, early diagnosis of AD can significantly enhance the prognosis of the disease and research is actively striving to discover and develop new techniques to reach this target. However, the current gold-standard for AD diagnosis is the pathological analysis of brain tissues, a highly invasive technique. Other alternatives for diagnosis constitute brain-imaging tests/devices, such as Magnetic Resonance Imaging (MRI) devices and PET scans, as well as CSF biomarkers, namely amyloid-beta isoform 42 ($A\beta 42$) and phosphorylated tau. While these techniques are effective in assessing patient status, they remain expensive, highly invasive and uncomfortable for the patient [68]. Thus, serum biomarkers could offer a form of diagnosis much less invasive, more accessible, and affordable, allowing for wider and more repetitive screening, eventually improving patient outcomes through earlier intervention.

Several studies have been conducted in this direction, focusing on several types of molecules as potential biomarkers. Various research papers have shown the potential of measuring serum tau protein levels as biomarkers for AD progression [69–71]. One team of researchers interestingly found serum D-serine (an amino acid with cognitive functions) levels to be altered in early phases of AD, presenting it as a prospective precocious biomarker [72]. Micro RNAS (MiRNAs) have also been investigated as

potential biomarkers, as is the case of serum miR-128 [73] and of miR-106b [74].

Nonetheless, cytokines remain an interesting group of molecules to consider as candidates for serum biomarkers in AD, since, as has been explored before, inflammatory cytokines have been shown to be dysregulated in the brain and CSF of AD patients [75]. On this topic, one study found the Brain-Derived Neurotrophic Factor (BDNF) to be decreased in the serum of AD patients, making this cytokine a potential candidate to be used as a biomarker for early AD detection [76]. Blood IL-6 has also been established as a risk factor in MCI patients for cognitive decline, with high levels of this cytokine being linked to increased risk of AD diagnosis [75]. One meta-analysis comprising 44 studies, reported elevated levels of pro-inflammatory cytokines, such as IL-6, TNF-α, IL-1β, Transforming Growth Factor-β (TGF-β), IL-12 and IL-18 in the blood of patients with AD, when compared to healthy controls [61]. A more recent review, encompassing 175 studies found (as well as the previous markers) IL-2, IFN-γ, C-reactive protein and C-X-C motif chemokine ligand 10 (CXCL10) levels to be increased in AD patients, adding that levels of IL-6 were inversely correlated with cognitive function [77].

To conclude, there is a fair body of research being conducted in this area with promising results, however more research is still required to validate reproducibility, sensitivity specificity and cost-effectiveness of serum AD biomarkers, with the goal of improving diagnosis and early treatment of patients. Eventually, given the intricate role of cytokines in orchestrating immune cell homeostasis and coordinating signal-dependent immune responses, the use of cytokines as blood-derived diagnostic tools is challenging, but the correlation of AD with an abnormal cytokine profile and/or innate immune status could provide valuable insights in this field.

3

Machine Learning Theoretical Background

Having established the biological theoretical background of Alzheimer's disease, its relationship to infection, and the role of trained immunity in its pathogenesis, we turn now to the theoretical framework of Machine Learning (ML). ML holds out powerful tools for analysing complex biological data, including uncovering patterns and making predictions that cannot easily be done by traditional statistical methods. ML methods have been applied to large data sets with the purpose of finding biomarkers, predicting disease course, and trying to understand mechanisms of disease in the context of Alzheimer's research. To that end, this shift towards ML theory will offer insight into how advanced computational methods might be harnessed for the advancement of knowledge and the general frameworks behind them.

3.1 Definition of Machine Learning and Basic Concepts

Machine learning is a specific branch of Artificial Intelligence (AI), in which algorithms improve automatically by means of experience, making it particularly useful for solving problems so complex that are beyond our human capacities. Typical scenarios for this are tasks that are consistently carried out by living beings, but our reflection on how we accomplish them is not detailed enough to enable us to derive a clear algorithm, such as driving or speech recognition, and tasks that require vast amounts of data [78]. AI, a broader concept encompassing ML, refers to computers that mimic cognitive functions we associate to the human mind, such as perception, reasoning, problem-solving and, in fact, learning [79]. This brings us to a crucial question in today's era: "Can machines think?" This question, which centers on the ability of computers to perform complex tasks that typically require human intelligence, has become a key focus in the exploration and development of AI. The purpose of ML, in its essence, however, is to learn an objective function, simply put, ML is about modelling data [80].

With this in mind, a key step of ML is data collection, this is electronic information that will be made available for analysis, and it is usually represented in a structure

known as a dataset [81]. A dataset usually assumes a tabular form. One observation from a dataset is referred to as a data point or instance. An instance is a single experiment, corresponding to each row in tabular data. It is composed of different values for given attributes. These attributes are called features. A vector represents the values of the various attributes of a certain instance. Below is a representation of a dataset (D) with n instances and m number of features.

$$D = \begin{matrix} x_{1_1} & x_{1_2} & \cdots & x_{1_m} & y_1 \\ x_{2_1} & x_{2_2} & \cdots & x_{2_m} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n_1} & x_{n_2} & \cdots & x_{n_m} & y_n \end{matrix} \tag{3.1}$$

A vector of observations (input to a model) can include features and labels (yi) as well, also known as targets, which represent the model's output. Whether the model is or is not trained with this sort of labelled data defines its type of learning scenario, that is, if we are presented with supervised or unsupervised learning.

3.1.1 Unsupervised Learning

Unsupervised learning involves scenarios where the target values are not known for each observation, strictly speaking data is unlabelled, and the ML algorithm must figure out some criteria to group similar inputs together. It can be viewed as finding patterns in the data. Two classic examples of this type of learning are dimensionality reduction and clustering [82].

3.1.2 Supervised Learning

In contrast, When a ML model is trained with labelled data, that is, the learner is given specific input-output pairs, we are facing a supervised learning scenario [79]. In this case, the purpose of the algorithm is to learn to yield the correct output given a new input. The work presented in this thesis made use of supervised learning techniques.

Before we get ahead of ourselves, it has been mentioned that the model is *trained* with some data. It is therefore important to clarify the notions of a training, validation and test set when it comes to ML.

Training set: The portion of the dataset that is used to train the model. In supervised learning, it consists of input-output pairs leading the model to learn the underlying patterns and relationships in the data.

Test set: The portion of the dataset which is left out during the learning phase. It is used to evaluate the model's performance on unseen data.

When a trained model performs exceedingly well on instances used for training but severely underperforms on new unseen samples we face a phenomenon called overfitting, that is to say the model does not generalize well [83]. To avoid this problem and

find a model with an appropriate equilibrium between complexity and generalization ability, we add another split to the data creating the validation set.

Validation set: A subset of the data used to select the appropriate values for the hyperparameters of the model. It helps assess the model's performance on an additional layer of unseen data before the final test set, enabling the fine-tuning of hyperparameters and reducing the risk of overfitting.

Overall, a model is considered a good model if it generalizes well, meaning it can make good predictions on unseen samples, thus effectively learning the patterns and relationships behind the data. This is the true value of a good ML model [78].

3.1.3 Classification vs Regression Problems

ML algorithms seek to solve two different types of problems:

Classification, in which the target values have a discrete or categorical codomain, and are called labels or classes [78]. For instance, predicting a species of a flower based on its features, or classifying tumors as malignant or benign. Classification tasks can be further divided into:

1. **Binary Classification**: Where observations can only be categorized into two possible classes (for example, if a tumor is benign or malignant) [84].

2. **Multi-class Classification**: Where there are more than 2 possible classes for the target (such as predicting flower species) [84].

Regression, in which the target values are continuous and are called expected outputs [78]. For instance, predicting the level of expression of a specific gene based on distinct biological and environmental factors.

Since these problems are very disparate in their nature, the metrics to evaluate models that fall in either of these categories are themselves also very distinct. Let us start by expanding on the metrics for classification problems.

3.1.3.1 Estimating the Performance of a Classifier

The behavior of a classifier is often represented in a confusion matrix as the one shown in Figure 3.1. This visualization displays the number of true predictions which are divided into True Positives (TP) and True Negatives (TN), and false predictions, namely False Positives (FP) and False Negatives (FN), made using labelled data. In ML context, positive cases refer to instances that belong to the target class of interest, while negative cases are those that do not belong to this class.

TP represent instances correctly predicted as belonging to the target class, while TN refer to cases that are correctly identified as not belonging to the target class. In contrast, FP occur when the model incorrectly predicts a negative instance as positive, that is when the instance does not belong to the target class however the algorithm

categorizes it as so, and FN arise when the model fails to identify target class instance, predicting it as negative instead.

On the left side of the matrix of Figure 3.1 are represented the cases that the model predicted to be positive (TP + FP), for example the cases where a model predicted a patient to have a certain disease. On the right side are the ones predicted to be negative (FN + TN), in the previous example, subjects the model predicted to be healthy. Whereas, the upper row of the matrix encodes the cases that are known to be positive (TP + FN) (cases where the subjects actually have the disease) and the bottom row encodes the ones known to be negative (FP + TN) (cases where the subjects don't present the targeted disease). In this way, correct predictions are represented in green and incorrect in red.

The performance/ quality of a classifier can be numerically quantified by a variety of metrics which make use of the aforementioned TP, TN, FP and FN. These metrics are accuracy, precision, recall, f1 score, specificity and negative predictive value and are graphically encoded on Figure 3.1.

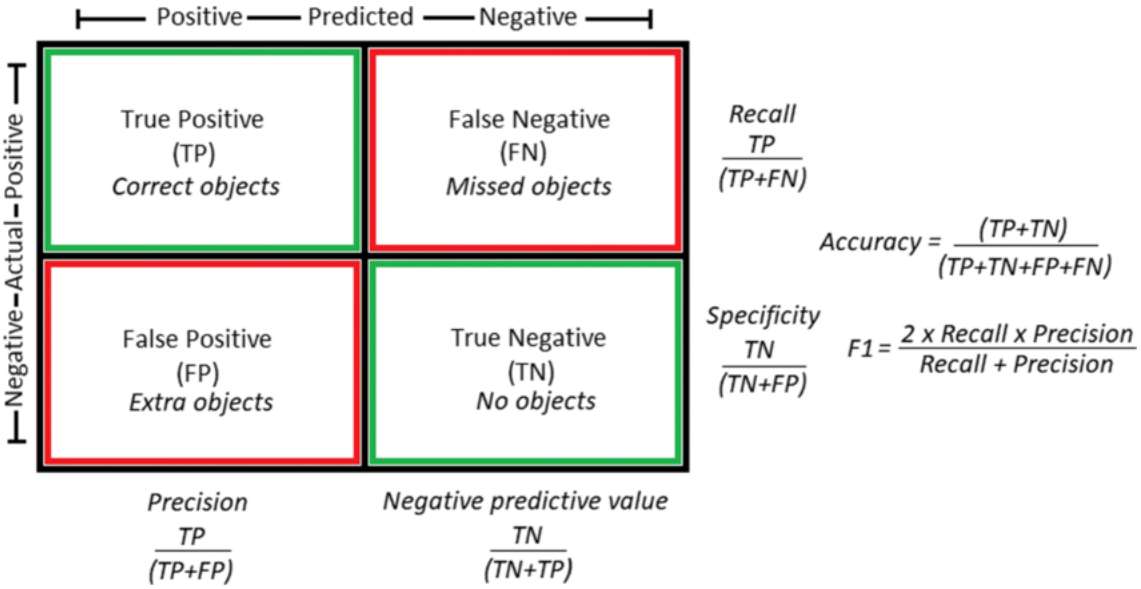


Figure 3.1: General representation of a confusion matrix and classifier error metrics. (Reprinted from [85]).

The most popular and simple form to estimate the predictive ability of a classification model is accuracy, that is, the number of correctly classified instances [78], as shown in the formula presented in Figure 3.1.

While this a powerful measure to assess the performance of the model, it might not be sufficient to understand its quality and behavior. For example, in a dataset where one class dominates, a high accuracy might simply reflect a model who is only predicting majority class, completely disregarding the minority class. Precision and recall, on the other hand, are two of the most widely used metrics to assess a classifier's effectiveness [86]. Precision expresses the number of correct positive predictions (TP)

belonging to a class, divided by the total number of positive predictions (TP + FP), it indicates how often a ML model is correct when predicting the target class [87]. Recall (also called sensitivity) is the number of accurate positive predictions (TP) divided by the total number of positive (P), it shows whether a ML model can find all objects of the target class [87]. The F1 score combines both precision and recall into a single metric, providing a balanced measure of a model's performance [88]. Its formula can be found on Figure 3.1. All these metrics vary between 0 and 1, the closest to 1 the better the classifier.

Lastly, the Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curve are also widely used metrics in binary classification. The ROC curve, as represented on Figure 3.2, plots the true positive rate (TPR) against the false positive rate (FPR) of the classifier at different threshold levels of a model's hyperparameter. AUC represents the area under this curve, ranging from 0 to 1. The diagonal line in the figure represents the ROC curve of a random predictor, with an AUC equal to 0.5. AUC measures the discernibility of the model, a higher AUC indicates better performance, with 1 representing perfect discrimination (the best possible predictor would yield a point in the upper left corner, with coordinates (0,1)) and 0.5 indicating performance similar to random guessing. It is valuable for evaluating classifier performance across various threshold settings and crucial for understanding a model's discriminatory power [78].

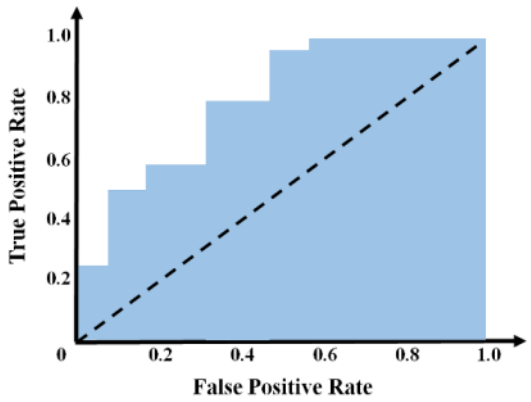


Figure 3.2: ROC curve example for a binary classifier. AUROC is represented in light blue. (Reprinted from [78]).

3.1.3.2 Estimating the Performance of a Regressor

While classifier metrics focus on counting how many instances were correctly predicted to assess performance, regression tasks don't require such counts since the target values are continuous. Instead, performance is evaluated by measuring the difference between predicted and actual values, providing a clear understanding of how far predictions deviate from the desired outcome. Although the general preference for these metrics

is ever-changing, some of the most commonly used metrics for regression problems are the (root) mean squared error (MSE and RMSE), Mean Absolute Error (MAE), the Median Absolute Error (MdAE) and the Pearson correlation coefficient (R) [89]. The study conducted in this thesis employed MdAE as the error measure for regression problems, along with its standard deviation. Pearson's correlation coefficient is also one of the key metrics considered for evaluation of the presented models, MAE is also presented.

The MAE is a metric used to measure the average magnitude of errors in a set of predictions. Being absolute, it does not consider the direction of the errors, but it gives an idea of how close, on average, the predictions of the model are to the actual values. In this project, however, we mainly consider the MdAE, which despite having similar formulation, considers the median of the errors and not the mean, resulting in a metric more robust to outliers [90]. This metric also does not assume any specific distribution of errors, unlike metrics such Root Mean squared Error (RMSE), which assumes a Gaussian distribution [91]. The formulation of the MdAE is hereby represented as:

$$\text{MedAE} = \text{median}\left(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \ldots, |y_n - \hat{y}_n|\right) \tag{3.2}$$

where:

- y_i are the actual values.
- $\hat{y}_i$ are the predicted values.
- n is the number of samples.

The standard deviation of the MdAE is also presented. Lastly, the Pearson correlation coefficient (R) is a measure of the linear correlation between two sets of data points, in this case one set is composed of the predictions output by the model and the other is composed of the real values of the target variable. The formulation of this metric is represented in equation 3.3, but, in sum, it is a metric that varies between 0 and 1 and represents how well our model correlates to real values. The closer R is to 1, the better its predictive performance [89].

$$R = \frac{\text{Cov}(r, p)}{\text{Std}(r) \cdot \text{Std}(p)} \tag{3.3}$$

where:

- $\text{Cov}(r, p)$ denotes the covariance between the actual values r and predicted values p.
- $\text{Std}(r)$ and $\text{Std}(p)$ represent the standard deviations of the actual values r and predicted values p, respectively.

3.2 Data Preprocessing in Machine Learning

Having established the basic concepts of ML required for this project, one ought to refer the basic steps a typical ML pipeline consists of, which are divided into 5 main categories [78].

- 1. Preprocessing the data
- 2. Choosing the algorithm and respective parameters
- 3. Estimating the predictive error on the validation set
- 4. Training the final model
- 5. Evaluating the final model

While these steps are not independent, nor even necessarily linearly sequential, a good starting point is with preprocessing the data, which will be explored in the next sections. This pipeline is essential in the outcome of ML algorithms following the "garbage in, garbage out" (GIGO) principle, which establishes that the quality of the output of a model is directly dependent on the quality of its input [92].

3.2.1 Data Cleaning and Normalization

The importance of data preprocessing is widely accepted as a major step in affecting ML outcomes [93], as evidenced by the aforementioned GIGO principle. Firstly, it is crucial to select the appropriate features for the desired task, then it is essential to convert the instances on the raw data into appropriate data types (as many ML algorithms expect numerical inputs), furthermore it is important to remove unwanted (duplicate or empty) rows or columns of the dataset [94]. These primary steps describe the data cleaning process. After these, data normalization or standardization usually follows.

Data normalization is a pre-processing approach that consists of scaling or transforming the data in order to make an equal contribution of each feature. It does this by transforming the data to comparable dynamic ranges [95]. The two most common data scaling techniques are the standard scaler (which centers the data around the mean 0, with a standard deviation of 1) and the Min-Max scaler (which scales each feature to a given range, usually [0, 1]) [96]. There is no general consensus on which of these methods is better [97]. In the context of this project Min-Max scaler was used where, for each feature, the smallest value becomes 0, the largest value becomes 1, and all other values are scaled to be between this interval [98], as shown in equation 3.4.

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.4}$$

where:

- X is the original value,

- X' is the scaled value,
- X_{min} is the minimum value of the feature,
- X_{max} is the maximum value of the feature.

Along with the general advantages of scaling the data, such as enhancing the model's performance and interpretability, attributing equal importance to features and improving a dataset's resilience to outliers, Min-max scaler offers the primary advantage of keeping the original distribution's shape of the data while adapting the values to a predefined range [99]. This proves useful when working with datasets that contain outlier values or show a wide range of scales in its features [100].

3.2.2 Handling Missing Values and Imputation Strategies

Real-world data is often messy and incomplete, with missing values being a common challenge in any type of data-centric work. Missing data can arise for various reasons and present significant issues as it may lead to biased results, reduce overall quality of analysis and moreover, some algorithms are not equipped to deal with these missing values [101]. One can address this issue from two perspectives. One is the feature selection perspective, when dealing with datasets with several features, it is common to eliminate features that exceed a certain threshold of missing values, typically applying a 50% cut-off [102]. The other is the row elimination perspective, wherein after removing features with high missing value rates, we may still encounter rows with missing data. While some approaches eliminate these rows, this can be problematic, especially with small datasets. To address missing values without losing valuable data, common imputation methods include mean, median and mode imputation. In this approach, the missing values are replaced with the chosen measure for each feature. Overall, while there are more advanced and complex methods being developed, this simple imputation technique, proves effective and much less computationally expensive, whilst enabling further analysis [103].

3.3 Dealing with Small Datasets

One scenario where the management of missing data is particularly critical, is when working with small datasets, specifically within the medical field. One important factor in a ML model's performance is its dataset size. Bigger datasets often perform better, especially in classification, while smaller datasets may lead to over-fitting. This is usually the case since a larger dataset offers more comprehensive information, allowing the underlying model to discern intricate patterns, thereby improving its generalization abilities [104]. However, in reality, gathering medical data is fraught with difficulties because of patient privacy, a lack of instances due to the uncommon nature of certain illnesses, as well as organizational and legal issues [105]. This is a problem encountered

during the course of this thesis. Nonetheless, working with available medical data remains of huge scientific interest and can still provide valuable insights and novel findings. Thus encountering the most appropriate ways to handle these limited size datasets is of utmost importance.

3.3.1 Cross-Validation

Before advancing on to the theoretical background behind the algorithms chosen for this project, let us discuss point 3 of the ML pipeline, estimating the predictive error, which makes use of the validation set.

In ML, data is traditionally partitioned at the outset to create training, validation, and test sets with sufficient instances for each [96]. However, in the case of small datasets, the situation grows nuanced. In the established train/validation split, the idea is using a single set of data (the training data) to develop a predictive ML model, next to employ a second set of data (the validation data), the labels of which are known but not disclosed to the predictor, in order to test different hyperparameters of the model and estimate its error rate. It should be noted that since the ML model is not trained using the entire sample, there is a loss of efficiency [106]. On the other hand, cross-validation works by dividing the dataset into multiple subsets or "folds." The model is trained on a combination of these folds and tested (or validated) on the remaining "left-out" fold. This process is iterated multiple times, with each fold serving as the validation set once. The results from each iteration are then averaged to provide an overall performance estimate. This method helps ensure that the model generalizes well to new, unseen samples and that is not just directly overfitting to the data [107].

3.3.1.1 Nested Cross-Validation Grid-Search for Hyperparameter Tuning

A common feature among nearly all ML methods is the presence of hyperparameters (external model settings or configurations that affect the model's performance and training process). In order to operate, hyperparameters for modern supervised ML algorithms must be configured. To do so, there are three strategies. The user can either apply the software package's default values, manually configure the settings, or execute a tuning process to achieve optimal predictive performance [108]. The latter option is often the most advised, and it is usually achieved using a nested cycle of cross-validation, where the data is split into multiple partitions, ensuring that the final selected hyperparameters perform well across different subsets of the data [78].

An emerging strategy when working with small datasets, where it is not possible to leave out a significant amount of samples for the test set, is to perform hyperparameter tuning by means of a nested cross-validation [109]. A 2 loop strategy which performs a grid-search of optimal hyperparameters in an inner loop and estimates model performance on an outer loop [110], as can be seen in Figure 3.3. As illustrated, in the outer loop of this method the original dataset is split into multiple folds where, per

iteration, each fold serves as the test set (represented in blue) while the others compose the training set. Within each training set, the inner loop further splits the data into training and validation folds (represented in orange). This inner loop is responsible for the hyperparameter tuning, as a grid-search is performed over different combinations of said parameters in the training folds, following the selection of the combination hat minimizes validation error. This optimized hyperparameters are then applied to the outer loop's training set, and the model's performance is assessed on the test fold. The process is iterated across all folds, and the results are averaged to provide an estimate of the model's generalization performance.

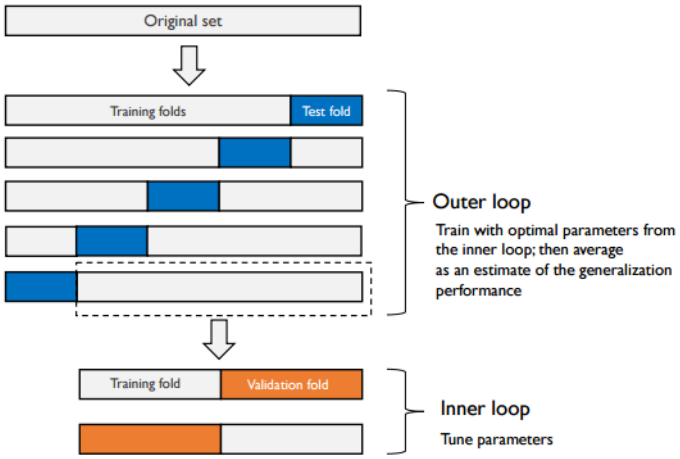


Figure 3.3: Schematic representation of nested cross-validation. (Reprinted from [110]).

One of the major drawbacks of this algorithm is that it does not result in one single final model, but rather various models, due to the different splits of the data. However, when assessing how well the data can be modelled by a given method, researchers have found this procedure to provide a nearly unbiased estimate of the true error [109]. Furthermore, if it is possible to initially set aside a test set, even of a relatively small size, it can be used later to evaluate and finalize the model, with the already optimized combination of hyperparameters.

3.3.2 Oversampling Techniques

On the topic of imbalanced and small datasets, a crucial technique to address class imbalance and data scarcity is oversampling. An imbalanced dataset, in the context of ML, is one where the distribution of target outcomes is unequal. This is primarily a concern in classification tasks, where certain classes may have significantly more instances than others. However, it can also occur in regression when the target values are skewed, leading to a concentration of observations in specific ranges while leaving others sparsely populated [111]. This issue is troublesome because it is associated with misclassification, where the minority (less represented) class (or less represented range

of values in the case of regression) tends to be misclassified or less captured by the algorithm, as compared to the majority (more represented) class/ range [112]. Two primary sampling techniques to address this question are undersampling and oversampling, where samples are either added to the minority class or decreased from the majority class. The case of oversampling is beneficial to small datasets as it consists in adding samples to the minority class. Oversampling techniques fall into two categories, synthetic oversampling and random oversampling. To expand the size of a minority class, the random oversampling approach replicates existing minority samples. The synthetic oversampling approach creates new synthetic samples for the underrepresented class [113]. The Synthetic Minority Over-Sampling Technique (SMOTE) and Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) algorithms fit within this latter category, and will further explained in the methodology part of this thesis (Chapter 5).

3.4 ML Algorithms Deployed in This Project

If the goal of this project is to build ML models that form accurate predictions, a crucial part of this process is selecting which algorithms to deploy. This brings us to the No Free Lunch Theorem posed by David Wolpert in 1995 [114]. Extrapolating to the purpose of ML, this theorem states that if averaged on all possible problems, the generalization ability of all ML algorithms is the same. Therefore, there can not be a "super algorithm" which outperforms all other algorithms on all the problems and thus, selecting the best algorithm is a challenge every time we are presented with a new problem or system to model. Hence, the most appropriate way of choosing the ML algorithm for the task at hand is by means of heuristic/informal processes, based on our knowledge of the algorithms and the exploration of several strategies [78].

The No Free Lunch theorem is central to the methodology of this thesis, which is why several algorithms were evaluated on each of the tasks, before selecting the most effective one. This process included applying linear regression (with and without penalties), logistic regression (for classification tasks), decision trees, random forest, support vector machines, K-nearest neighbors, and extreme gradient boost to the data. These algorithms are particularly advantageous due to their adaptability to both regression and classification tasks.

3.4.1 Linear/ Logistic Regression

Linear and Logistic Regression are fundamental techniques in statistical modelling and ML. Linear regression, is used for predicting a continuous target variable based on one or more independent variables. Providing the simplest and also one of the most popular forms to model regression data, it assumes a linear relationship between the input variables and the target [115]. Its goal is to find the coefficients of the linear

relationship that minimize an error measure (also known as loss function), between the observed and the predicted values [116]. This method is frequently introduced in statistics courses, but is equally prevalent in ML contexts due to its effectiveness and ease of interpretability. The general formulation for this technique is presented in equation 3.5.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{3.5}$$

Where:

- y is the dependent variable (response),
- $x_1, x_2, \ldots, x_p$ are the independent variables (predictors),
- β_0 is the intercept term,
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (regression coefficients),
- ϵ is the error term (residuals), that is the difference between the observed and the predicted value..

In contrast, logistic regression is employed when predicting a categorical target variable. Instead of forecasting a continuous value, logistic regression uses the logistic function, which maps any real-valued number into the [0, 1] interval, to represent the likelihood that a given input belongs to a specific class. To achieve this, it makes use of log odds (logit), which is the natural logarithm of the odds of an event occurring [78]. Hence the model can be formulated as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.6}$$

Where:

- $\ln\left(\frac{p}{1-p}\right)$ represents the log odds of the event probability p,
- $p = p(y = 1 \mid \mathbf{x})$ is the probability of the event given the input vector $\mathbf{x}$,
- β_0 is the intercept term,
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients corresponding to the input variables $x_1, x_2, \ldots, x_p$.

3.4.1.1 Lasso Regression or Logistic with L1 Penalty

Lasso Regression and Logistic Regression with L1 penalty are powerful techniques, that build upon the previously mentioned algorithms, to improve model performance and reduce overfitting. Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a type of linear regression that includes a L1 regularization term in its cost function [117]. This penalty imposes a constraint on the model's parameters, shrinking

them towards zero, by forcing the absolute sum of the model's coefficient to not be greater than a given predefined value (λ). This makes the model less complex and easier to interpret [118]. Logistic regression with an L1 penalty, also known as Lasso Logistic Regression, serves a similar purpose for classification tasks, with the same L1 penalty being applied. Overall, as this penalty leads to the creation of a sparse model, Lasso tends to perform better when only a few variables are relevant for the predictions of the model [78].

3.4.1.2 Ridge Regression or Logistic with L2 Penalty

Ridge Regression and Logistic Regression with L2 penalty constitute the other type of regularization techniques applied to linear/logistic models. Ridge Regression, alike Lasso regression also includes a regularization term in its cost function which penalizes large coefficients, leading effectively to their shrinkage [119]. However, this penalty which is proportional to the sum of the squared coefficients of the model, does not encourage sparsity in the coefficients, instead, the L2 penalty aims to minimize the overall size of the coefficients while keeping them in the model, making the values small, but not forcing them to equal zero. In fact, in Ridge, the coefficients of correlated variables are usually rather similar to each other, consequently this technique usually works best when most variables have a significant impact on the target [78].

3.4.1.3 Elastic-Net Regression or Logistic with L1 and L2 Penalties

Elastic-Net (EN) Regression and Logistic Regression with combined L1 and L2 penalties compose a hybrid technique that blends the strengths of both Lasso and Ridge regularization methods. In this technique, a combination of the L1 and L2 penalties is introduced, as a hybrid penalty that allows for both coefficient shrinkage (as in Ridge) and variable selection (as in Lasso), surpassing the limitations of applying just one of these strategies. Elastic-net performs particularly well in high-dimensional datasets, when the number of predictors is much larger than the number of samples in the data [120].

3.4.2 Decision Trees

Decision Trees (DT) are versatile and intuitive predictive models that can be applied for both classification and regression tasks in ML. Given a training set, these models are represented by a tree-like structure where each internal node constitutes a feature, and each edge represents one possible value for that feature. Leaves (terminal nodes) contain target values, and a prediction is made by following the path from the root node to a leaf [78]. This process results in a highly interpretable model since the decision-making process can be well visualized and understood (particularly in classification scenarios). The algorithm's adaptability to both categorical and numerical data is one

of its strengths, as well its need for little data preprocessing, however it is prone to overfitting, especially when the data leads to the formation of complex trees [121]. This has led to the development of multiple trees algorithms, namely random forest.

3.4.3 Random Forest

Random Forest (RF) is a learning technique that combines several DTs, gaining in performance and stability and forming a robust predictive model. This algorithm works by creating numerous DTs during training, whose output is aggregated into a single prediction using voting (mode) in classification tasks or averaging when is the case of regression [122]. To provide diversity among the trees and reduce overfitting, a random subset of the training data and a random subset of characteristics are used to build each tree in the forest. This randomness combined with the aggregation of multiple trees result in a model that is accurate and resilient to data noise [123]. One of the great advantages of RF is its ability to bring to light meaningful interactions and non-linear relationships in the data, and they are particularly effective when dealing with high dimension datasets [124]. However, this comes at the cost of being very computationally expensive, especially as the number of trees in the forest increases.

3.4.4 Support Vector Machines

Support Vector Machines (SVM) are a classic type of ML algorithm used for both classification and regression tasks. In the case of classification, this algorithm works by mapping data points (observations) to a high-dimensional space and then finding the optimal hyperplane that maximally separates these data points into different classes. This hyperplane is chosen by the largest margin rule, meaning it is the hyperplane with the greatest distance between itself and the closest data points from either class. This rule helps to improve the model's generalization power. For non-linearly separable data, non-linear SVMs employ kernel functions, such as polynomial or Radial Basis Function (RBF) kernels, which transform the input space into a higher-dimensional space allowing for linear separation [125].

In regression, SVM works mildly differently. It is a generalization of the classification problem, in which the model finds not a hyperplane that separates classes, but a function that approximates the relationship between the input features and the target variable. Instead of maximizing the margin between classes, Support Vector Regressor (SVR) finds a line (or hyperplane) that best fits the data points, allowing for a margin of tolerance (ϵ) within which the model's predictions are considered acceptable. The algorithm does this while also keeping the model's coefficients small to prevent overfitting, balancing good fitting of the data with model simplicity [126]. Overall this technique is rather computationally expensive and one of its key limitations is its sensitivity to noise or outliers in the dataset, but it is also growingly popular and has proven successful in various fields and tasks [127].

3.4.5 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple yet effective ML algorithm applicable to both classification and regression tasks. In the case of classification, it classifies a data point based on its κ nearest neighbors' majority label. A neighbor is one of the κ closest training examples, used to make predictions about a new data point. In the case of regression, the continuous value output by the algorithm is based on the average values of said κ closest neighbors. κ is a crucial parameter of the algorithm as is the metric used to calculate the distance between the data points [128]. Despite its simplicity, KNN has proven successful in a large number of domains, competing with more sophisticated methods in their generalization ability. Nonetheless, it presents some drawbacks, such as its performance's heavy reliance on the choice of parameters (namely κ and distance metric), and its hampered reliability in classification when class labels are balanced among neighbors and in regression when the variance of the selected neighbors' outputs is high [78].

3.4.6 Extreme Gradient Boost

Extreme Gradient Boost (XGB) is a sophisticated combined learning model, alike random forest, which combines the strengths of various weaker models, such as DTs, to create one single and powerful predictor. It is an implementation of gradient boosting trees known for its speed and high performance [129]. To make matters simple, XGBoost starts by creating a single simple model (one decision tree) and it follows a sequential approach, building subsequent trees whose focus is on correcting the errors of the previous trees. It repeats this process iteratively with each new tree progressively improving the overall accuracy or diminishing the error of the ensemble. It improves on traditional gradient boosting techniques by incorporating regularization techniques which prevent overfitting and lead to higher interpretability [130]. It is currently one of the most widely used ML algorithms both in industry and academia, due to its robustness, speed and high performance, producing high-quality models even when applied to a range of difficult ML tasks [131].

3.5 Interpretable AI

Lastly, even after successfully training and evaluating an accurate and robust model, we are still faced with a hefty challenge: why is the model good? This is the notorious black-box problem. A black-box model, is one that does not reveal its internal mechanisms, that is to say it can not be interpreted by simply looking at its parameters/coefficients and thus, we do not know how it made its predictions [132]. Of the models described previously, the black-box models are: RF, SVM, and XGB, while the others can be fairly easily interpreted.

Interpretability, the ability for something to be explained or presented in understandable terms to a human, is often disregarded in ML. Nevertheless, there are dire reasons for its importance, namely scientific understanding, safety, ethics and case-by-case decision clarifications when necessary [133]. If these can be of use in any field, in the field of biomedicine, interpretability becomes essential. Where adding to the evident need for safe and ethical use of these systems, there is the promising opportunity to conduct knowledge discovery by using the predictive model's learnings to produce new theories regarding unidentified biological pathways [134].

3.5.1 SHAP values

This consideration brings us to the concept of SHapley Additive exPlanations (SHAP) or SHapley Additive exPlanations [135]. SHAP values are a game theory approach to interpretable ML. They provide explanations to the output of any ML model by assigning an importance value to each of its features representing how relevant said feature is to the model's prediction.

The theory behind this tool stems, as mentioned, from game theory. In 1951, Lloyd S. Shapley devised a theory that allowed for a fair distribution of a game's pay-out among players with different skills who worked together [136]. Translating this idea to ML terms, the "game" is the prediction task at hands, the "pay-out" is the model's output and the "players" are the model's features. SHAP values compute the marginal contribution of each feature by averaging its impact across all possible ways the features can be combined. This is done by evaluating the model's prediction with and without each feature across different coalitions of features, which are weighted based on their representativeness, giving priority to those that showcase independent behavior or interaction effects. Subsequently, these marginal contributions are averaged giving us the feature's SHAP value [132]. This process can be done for each prediction of the model, allowing for a case by case interpretation of feature importance, but SHAP can also provide a global explanation for the model by aggregating several individual predictions (for instances, all predictions made on the test set).

SHAP has a few interesting properties which makes it such a powerful tool for explainable AI [137].

1. **Model Independence**: SHAP values are model-agnostic, meaning they can be used to interpret any ML model.

2. **Efficiency**: The total sum of Shapley values or the marginal contribution of each feature should be equal to the value of the total coalition.

3. **Symmetry**: If two features contribute equally to a prediction, they are assigned the same SHAP value.

4. **Dummy**: SHAP values are zero for missing or irrelevant features for a prediction. This means that the Shapley value of a feature is zero if, independent of the coalition group, it does not alter the prediction.

5. **Additivity**: SHAP values are additive, which means that the contribution of each feature to the final model output can be computed independently and then summed up.

The SHAP framework includes several model-specific explainability methods such as *LinearSHAP* for linear models, *TreeSHAP* for tree-based algorithms and *DeepSHAP* for deep learning algorithms. However, there is one method that is designed to provide explainability for any type of model, which is appropriate when we are exploring several different algorithms in the aims to find the best model. This method is Kernel SHAP [135], which approximates Shapley values by solving a weighted linear regression problem. The idea is to fit a linear model to the predictions of the original model, where the assigned weights represent the contribution of each feature. This permits a flexible interpretation of any kind of model. Importantly, the use of weighted linear regression does not limit Kernel SHAP to either regression or classification problems, it is applicable to both. Regardless of the nature of the task, the weights assigned reflect how much each feature contributes to the prediction, whether that prediction is a continuous value in regression or a probability in classification [137].

While computing SHAP values is highly beneficial, their utility is much enhanced when translated into visualizations. The SHAP Python package [135] is particularly valuable in this regard, as it offers a range of intuitive and accessible visualizations that support both local and global interpretability. In the case of global interpretability, the most common plots are the feature importance bar plot (Figure 3.4) and the summary plot (Figure 3.5). The first one displays the features in descending order of importance and the x-axis represents the mean absolute Shapley values therefore not showcasing whether the feature impacts the prediction positively or negatively. The summary plot is more detailed. In the case of the beeswarm plot as shown in Figure 3.5, every dot on the plot represents a single sample of the data. The features are also ordered by importance and the horizontal axis represents the SHAP value, yet there is an added layer of information, as the color of the dots encodes the magnitude of the feature for that observation. Usually higher values are represented in red and lower values in blue. Hence, we can assess how higher and lower values of the features impact the result and in which direction. In the example of Figure 3.5 higher latitudes and longitudes have a negative impact on the prediction, while lower values have a positive impact.

Lastly, it is important to highlight the main advantages and disadvantages of this method. As far as disadvantages, computing Shapley values can be quite time-consuming, however, this drawback is not significant when working with rather small datasets. Additionally, sometimes Shapley values are misinterpreted. One must keep in mind that the Shapley value of a feature is not the difference of the predicted

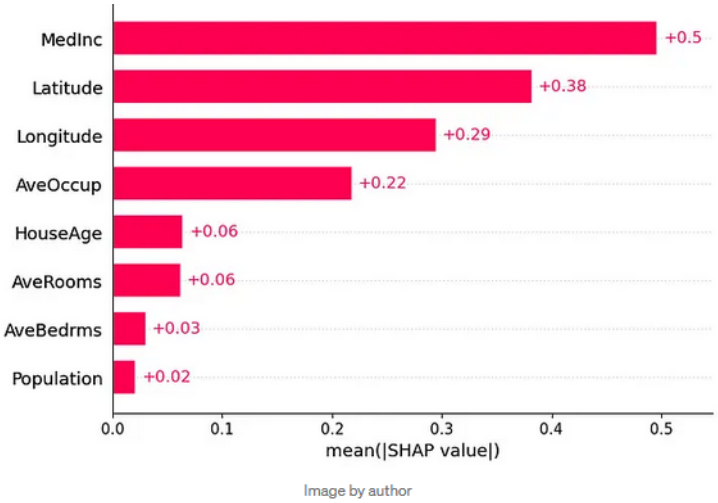


Figure 3.4: Example of a SHAP feature importance bar plot for global interpretability. Reprinted from [138].

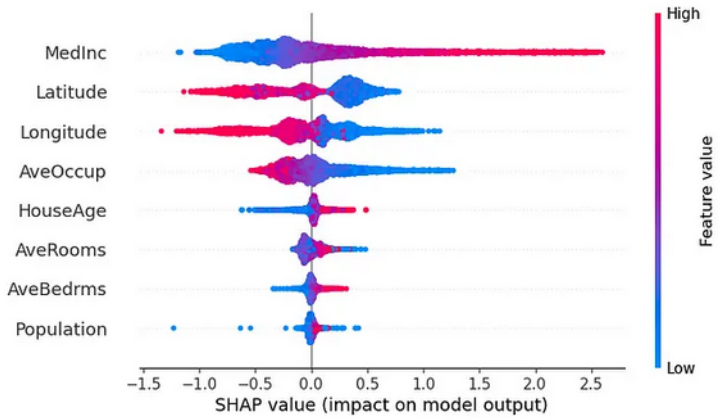


Figure 3.5: Example of SHAP beeswarm summary plot for global interpretability. Reprinted from [138].

value when the feature is removed from the model training, but rather represents the contribution of a feature to the difference between the actual prediction and the mean model's prediction. Lastly, SHAP only provides explanations for the current model and its observations, it does not have prediction power if you change the input. For example in a model predicting credit score, one could not use SHAP to extrapolate a conclusion such as "If I were to earn €500 more a year, my credit score would increase by 5 points" [132].

On the topic of its advantages, there are all the properties already listed, particularly the fact that SHAP can be applied to any kind of model. Furthermore, SHAP allows for contrastive explanations, meaning a prediction does not have to be exclusively compared to the average prediction of the complete dataset, but can be compared to a subset or a single observation. Finally, SHAP is the only theory-based explanation technique for AI. There are several axioms (efficiency, symmetry, dummy, additivity)

supporting it, which gives it a robust foundation and makes it an increasingly popular method in current research [132].

4

Literature review

This chapter aims at reviewing the current works that inspired this thesis, namely what has been done on the field of machine learning and Alzheimer's disease. Section 4.1 summarizes research presenting predictive models for AD based on MRI imagery and PET data. Section 4.2 explores alternative data approaches for building this models, focusing on work conducted with cytokine biomarkers and other proteins, and infectious data. Finally, Section 4.3 presents an overview of papers related to predicting cytokine levels.

4.1 MRI and PET-Based Models

In recent years, the advent of ML has revolutionized the field of medical diagnostics, offering novel methods for predicting the onset and progression of several diseases with increasing accuracy and reliability. Inevitably, with AD being one of the most prevalent diseases of the 21st century [25], various works have focused on this neurological condition.

While there is a plethora of data that can be used to train these intelligent systems, the most common approach is through medical imaging. with two primary types being utilized: Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). MRI is a popular non-invasive imaging method that helps detect physical changes linked to AD by creating detailed images of the structure of the brain by means of magnetic fields and radio waves [139]. Conversely, PET is a nuclear medicine imaging technique that uses the detection of pairs of gamma rays indirectly released by a tracer to visualize functional processes in the body, including the brain [140]. Given the importance of these imaging techniques, naturally a vast body of research has been conducted, exploring their use in cohort with ML models.

Firstly, on the topic of MRI-based models, many algorithms have been deployed to this extent, with Artificial Neural Networks (ANN)s (one of the most complex and popular ML methods) being a leading choice for modelling this type of data. Castelazzi et al., [141] designed an ANN model capable of distinguishing between AD and another

type of dementia (vascular dementia), based on this type of imagery, with over 84% accuracy. Another study, by Salehi et al. [142] developed a Convolutional Neural Network (CNN), a type of ANN algorithm specifically designed to process and analyse visual data, for the classification of AD using MRI images. This study comprised the data of over 5000 subjects and achieved an outstanding accuracy of 99%. However, as mentioned, ANNs aren't the only models being exploited for this kind of task. A team of researchers led by Long et al. [6] has been able to create a SVM model that discriminates patients with AD or MCI from healthy elderly, and predicts AD conversion in MCI patients with accuracies surpassing 90% for the first task and rounding 88% for the latter. All these studies show promising results for the advancement of more accurate AD diagnostics by means of this type of imagery data.

Shifting into ML models constructed with PET data, in this field, several algorithms have been leveraged. Tuan et al. [143] utilized an autoencoder (a type of neural network used for unsupervised learning) to determine, from these images, regions of interest in the brain, and then fed these findings to a SVM classifier, achieving accuracies of around 90% in distinguishing AD patients from Healthy Controls (HC)s. An interesting approach, in line with the vision of this project, is to explore multiple algorithms before deciding on a final model. Peng et al. [144], explored 4 different classifiers, SVM, Naïve Bayes, RF, and KNN, before landing on a final SVM model which achieved an AUC of 0.865, as well as a sensitivity and specificity of over 80%. On a similar note, the paper by R. S. Nancy Noella & J. Priyadarshini [145] analysed the behaviour of multiple classifiers such as Bagged Ensemble, DTs, Naïve Bayes and Multiclass SVM. The final Bagged Ensemble model was trained on over 700 PET images and distinguished between AD brains, Parkinson's patient brains and healthy brains with an accuracy of 90.3%. As a last additional example, one paper by Kumari et al. [7] combined different types of data sources, including MRI, PET images and cognitive assessments of patients, resulting in a high dimension dataset of 100 patients. This data collection was used to develop a new RF-based classifier, which yielded remarkable accuracies in distinguishing controls from AD patients (100% accuracy), controls from MCI patients (91% accuracy) and AD from MCI patients (95% accuracy). Again these findings, reveal great potential for more accurate and standardized diagnostics in AD, by means of ML models constructed on imaging data.

Nonetheless, while it is clear that using MRI and PET data is an interesting and widely adopted approach with promising results, these techniques are often costly, time-consuming and not easily accessible to everyone. Therefore, it is important to explore alternative methods that might be more affordable and accessible, while at the same time allowing for further insights into this disease.

4.2 Alternative Data Approaches

An alternative approach to imagery data involves utilizing demographic and psychological assessments as data, as demonstrated in the paper by J. Neelaveni & M. G. Devasana [146]. This paper uses factors such as age, number of clinical visits, Mini-Mental State Examination (MMSE) score and level of education to build a SVM classifier, able to distinguish between AD patients and HC with 85% accuracy. The research conducted by Antor et al. [147] also utilizes such metrics as well as brain measures to construct several models: logistic regression, DTs, RF and SVM. The SVM classifier ultimately achieved 92% accuracy and F1 score in predicting AD. An additional study conducted this year by Wang et al. [148], created a RF model from sociodemographic data as well (age, sex, marital status), and lipoprotein and metabolite variables. This model achieved an accuracy of 71.01%, a sensitivity of 79.59% and a specificity of 65.28%, which despite not being particularly high performance metrics, permitted the establishment and investigation of the association between certain proteins and metabolites and the onset of AD. These papers illustrate an important trend in AD research: the integration of multi-domain datasets in an attempt to enhance predictive accuracy. A different but growingly popular technique is to make use of genetic data. One team of investigators led by Alatrany in 2021 [149] developed a stacked ML model, more specifically based on single nucleotide polymorphisms of AD patients and controls. This model distinguished between the two groups with 93.7% accuracy, underscoring the potential of genetic profiling in early AD detection. More recently, a team led by Gao X. R. [150] using the UK Biobank dataset, considered genetic and non-genetic factors (such as codes from electronic health records of the biobank participants), to design a XGBoost model with an AUC for AD of 0.88. Furthermore, the researchers utilized SHAP to explain this model, which allowed them to identify important predictors for the development of AD, as, for example, records of urinary tract infection, syncope and collapse or chest pain. An additional study published in the same year, also made use of explainable AI, again using SHAP, to predict AD [151]. This study included information about the participant's sociodemographics, levels of self-reported health as well as blood biomarkers. This latter portion of data significantly increased the model's performance with logistic regression achieving the highest AUC value of 0.818. The researchers were able to identify levels of ptau protein, plasma neurofilament light, blood tau protein, taurine, inosine, xanthine and L-Glutamine as key predictors of AD, along with age, education level and marital status, regarding demographic factors. Again this study encompasses several domains of data in order to create a more comprehensive model of the disease. But a critical shared factor between these latter 2 papers is the introduction of interpretable AI through SHAP, which not only improves model transparency, but on top of that, aids in identifying key predictors, crucial for clinical applicability and for our understanding of the disease.

4.2.1 Protein and Blood/Serum Biomarkers

Bringing to focus a promising approach, closely aligned with the one taken in this thesis, several studies have explored the use of protein and blood/serum biomarkers for predicting AD. While the previous subsection highlighted various alternative approaches for the modelling of this disease, including psychological assessments, sociodemographic factors, and genetic data, the current focus on protein and blood/serum biomarkers represents a more direct investigation into the biological pathways of the disease, as these provide measurable insights into the physiological changes associated with Alzheimer's. This shift emphasizes the importance of biological markers in enhancing diagnostic accuracy and early disease detection and and it leads to studies that are more directly related to our hypothesis and methodology.

One research effort by Gaetani et al. [9] aimed to find protein biomarkers that indicate neuroinflammation in AD. To this end, they analyzed CSF samples from patients with mild cognitive impairment due to AD (AD-MCI) and compared them to samples from patients with other neurological diseases. The team subsequently built a penalized logistic regression model which yielded an AUC of 0.906. This led to the identification of 4 proteins, all linked to neuroinflammatory processes (SIRT2, HGF, MMP-10, and CXCL5) as effective markers, displaying higher levels in AD-MCI patients. Similarly, Khononikhin et al. [152] analysed mass spectrometry data from plasma samples of patients with AD, MCI, vascular dementia, frontotemporal dementia, and an elderly control group. The study highlighted significant decreases in specific proteins associated with AD. ML algorithms were then applied to identify important protein panels and build classifiers for predicting AD. The best classifiers achieved 80% accuracy, 79.4% sensitivity, and 83.6% specificity, in predicting the risk of developing AD within three years for patients with MCI. Proteins found to be candidate biomarkers included afamin, APOE, APOA4, fibronectin, vitronectin, FGG, FGA and beta-2-glycoprotein. One other study by Araujo et al. [10] that envisioned finding a panel of plasma proteins to predict MCI progression to AD, focused on a high-throughput modelling approach. The researchers created over one billion models by exploring different interactions among 146 plasma proteins and randomly selecting up to 30 proteins for each, before choosing the best-performing one. From this, they developed a ML-based panel composed of 12 plasma proteins (ApoB, Calcitonin, C-peptide, CRP, IGFBP-2, Interleukin-3, Interleukin-8, PARC, Serotransferrin, THP, TLSP 1-309, and TN-C). The best model yielded an AUC of 0.91 and accuracy of 91% for predicting the risk of MCI patients converting to AD dementia in a horizon of up to four years. Interestingly, a study aiming to predict conversion from MCI to AD based on protein data had already been conducted in 2011 [153]. However, these investigators combined plasma cytokine and chemokine levels with MRI data and compared these to measures of APOE genotype and clinical evaluation to assess which best predict progression. The study found biochemical markers of inflammation to

be better predictors of conversion than APOE genotype or clinical measures, with the combination of serum inflammation markers and MRI imaging providing the best predictor of conversion. The SVM model created had an AUC of 0.78 and was built with several cytokines including: BDNF , TNF-α, IL-6, IL-1β, IL-1RA, IL-10, IL-12, IL-2, IL-8, VEGF, IFN-γ, IL-4, IL-17, GM-CSF, G-CSF, and MCP-1. Closely related is another study from 2007 by Ray et al. [8], that found 18 blood plasma signalling proteins to be strong predictors of AD. Namely, these included G-CSF, IL-1α, IL-3, IL-11 and TNF-α, which were used to build a classifier that distinguished with 90% accuracy patients with MCI that progressed to AD within 2 to 6 years. One more recent study by Galgani et al. [154], aimed to understand the role of neuroinflammation in AD by examining blood circulating cytokines as well as analysing the effects of age, sex, and the APOE genotype on these biomarkers. Their dataset comprised a cohort of cognitively healthy individuals, patients with MCI, and patients with AD-like dementia. Their findings revealed a robust sex effect on IL-12 and an APOE-related difference in IL-10, with the latter being also related to the presence of advanced cognitive decline. Overall, IL-1β was the most strongly associated variable with the progression from MCI to dementia and the researchers concluded by highlighting the role of plasma cytokines as useful non-invasive tools for studying neuroinflammation in AD. Lastly, a 2019 research article used serum cytokine data to predict several marks of, not Alzheimer's but Parkinson's disease, another age-related neurodegenerative disorder [155]. In fact, cytokine levels have been used to build predictive models for several conditions, namely coronary artery disease [156], lung cancer [157], multiple sclerosis [158], and even COVID-19 [159].

All this literature is summarized in Table 4.1 and collectively gives emphasis to the significant potential of protein and blood/serum biomarkers in improving the early diagnosis and prediction of AD, providing additional support to the approach taken in this thesis.

Table 4.1: Summary of ML studies on blood biomarkers related to AD.

Authors	Year	Reported Blood Biomarkers	Results
Gaetani et al.	2021	SIRT2, HGF, MMP-10, CXCL5	AUC: 0.906
Khononikhin et al.	2022	Afamin, APOE, APOA4, Fibronectin, Vitronectin, FGG, FGA, β2-Glycoprotein	Accuracy: 80%
Araujo et al.	2022	ApoB, Calcitonin, C-peptide, CRP, IGFBP-2, IL-3, IL-8, PARC, Serotransferrin, THP	AUC: 0.91, Accuracy: 91%
Furney et al.	2011	Serum cytokines (various, including BDNF, TNF-α, IL-6, etc.)	AUC: 0.78
Ray et al.	2007	G-CSF, IL-1α, IL-3, IL-11, TNF-α	Accuracy: 90%
Galgani et al.	2022	IL-12, IL-10, IL-1β	Accuracy: 65%

4.2.2 Infectious Data Approaches

Despite the strong hypothesis, previously explored in chapter 2, that pathogens which cause neuroinflammation may lead to the development of AD [160], there is not an

extensive body of research focusing on using infectious data in order to predict AD through ML, which makes our approach rather novel. However, a 2024 study by Tejeda M. et al. [161] used presence and quantification of viral DNA to build several types of predictive ML modes for AD (from the more simple generalized linear models to ensemble methods). The best model found was logistic Lasso regression with 67.2% predictive accuracy for AD status in the test set. The researchers also found HSV-1 and HPV to be the strongest predictors. Another very recent research project aimed to study the hypothesis that herpes virus infection increases the risk of AD, using ML [162]. For this purpose, the investigators developed a RF model to identify 22 key regulatory genes, associated with the occurrence and development of AD and which are genetically regulated by herpes virus infection. These findings highlight the novelty and significance of adding pathogen-related factors into ML models for AD prediction, as well as open the door to a new strategy which could offer novel therapeutic approaches and diagnostic instruments that focus on the inflammatory and infectious aspects of AD.

4.3 Studies on Predicting Cytokine Levels

Lastly, to further explore relationships between pathogens and cytokines, and the pivotal role these proteins play in the neuroinflammatory processes associated with AD, in this project we have also attempted to predict cytokine levels. Therefore, while this type of research is still in its early stages, it is relevant to analyse what has been done in this emerging area.

Starting with TNF-α, the research conducted has mainly been on the field of TNF-α inhibition response, as TNF-α inhibitors are important drugs in treating patients with certain autoimmune diseases [163]. A team of researchers led by Prabha [164] have developed a ML model, *TNFipred*, for classifying TNF-α inhibitors. For this purpose, they explored Naïve Bayes, RF, KNN, and SVM models. The best-performing model was RF, achieving an accuracy of 87.96% and a sensitivity of 86.17%. This study represents the first ML model specifically designed for TNF-α inhibitor prediction. Additionally, several studies have focused on using ML models to predict non-response to anti-TNF treatment, specifically on the case of rheumatoid arthritis [165–167]. Lastly, while there seems to be a lack of papers that directly predict TNF-α levels, one study has aimed at developing models for predicting TNF-α inducing peptides [168], as enhanced expression of this cytokine is associated with the progression of several diseases. Their model achieved an AUROC of 0.83 and interestingly, the researchers also identified potential TNF-α inducing peptides in different proteins of HIV-1, HIV-2 and SARS-CoV-2.

Considering Interleukin-6, again ongoing research seems to be directed towards the prediction of IL-6 inducing peptides and not direct levels of IL-6 in patients [169, 170]. For IL-10 this is also the case, with computational approaches being undertaken

for predicting IL-10 inducing peptides [171]. Even so, one study utilized quantitative morphology data from macrophages to predict these cells' content of IL-10, achieving a 95% accuracy in this task via a RF model [172]. As for IL-1RA and IL-1β, it appears that these cytokines have so far only been utilized as features rather than targets in ML studies, which constitutes an innovative part of our study.

From looking at the current available research, one can visualize the efforts of the scientific community in studying the debilitating disease that is Alzheimer's, from describing its intricate pathways and mechanisms to finding better and more accessible diagnosis techniques. Furthermore, it is a paramount example of the need to integrate technological advances and our novel intelligent tools with medical research, with ML emerging as a powerful new ally and achieving outcomes that traditional techniques could not easily achieve.

5

Methodology

This chapter describes the methodology used in this project. Section 5.1 briefly describes the data collection process, Section 5.2 defines the general *in silico* experimental approach, and Sections 5.3 and 5.4 expand upon the specifics of the experimental procedures and respective settings.

5.1 Data collection

The data used in this study was collected and provided by the Experimental Neuropsychobiology Laboratory, Clinical and Behavioural Neurology Unit at IRCCS Fondazione Santa Lucia. The dataset, which includes data for 51 healthy subjects and 48 AD patients, is composed of 3 main domains: Infectious Burden (IB) data (with measures of antibodies for HSV-1/2, *Helicobacter pylori*, CMV, *Chlamydia pneumoniae* and *Borrelia burgdorferi*), Trained Immunity (TI) data (composed of measures for different cytokines under different stimulation conditions) and serum data (levels of serum circulating cytokines). It also includes information on patients' sex and age. The cytokines included in this study are described in Table 5.1.

TI data was collected according to the protocol by Domínguez-Andrés et al[173]. In this way, TI was inducted *in vitro* in adherent monocytes obtained from the blood of AD patients and control subjects. This process is described in Figure 5.1. After isolation (step 1 in Figure 5.1), these cells were subjected to a trained immunity-inducing stimulus, namely incubation with LPS (a molecule of bacterial origin), these cells are referred to as Primed with LPS molecule (Pr LPS), or incubation with the yeast *Candida albicans*, these are referred to as Primed with *Candida albicans* pathogen (Pr Ca), as seen in step 2 of Figure 5.1. Some cells, identified as Non Treated (NT), were not subjected to any stimulus, serving as controls. After approximately one week, a portion of the cells was rechallenged with one of the previously mentioned stimuli (step 3 of Figure 5.1). Cells that had been primed with LPS and were challenged with LPS are termed as LPS LPS, whereas the ones that were primed with *C. albicans* and then challenged with LPS are referred to as Ca LPS. Lastly, cytokine levels found in the supernatant of the cellular

suspensions were assessed by ELISA (step 4 of Figure 5.1).

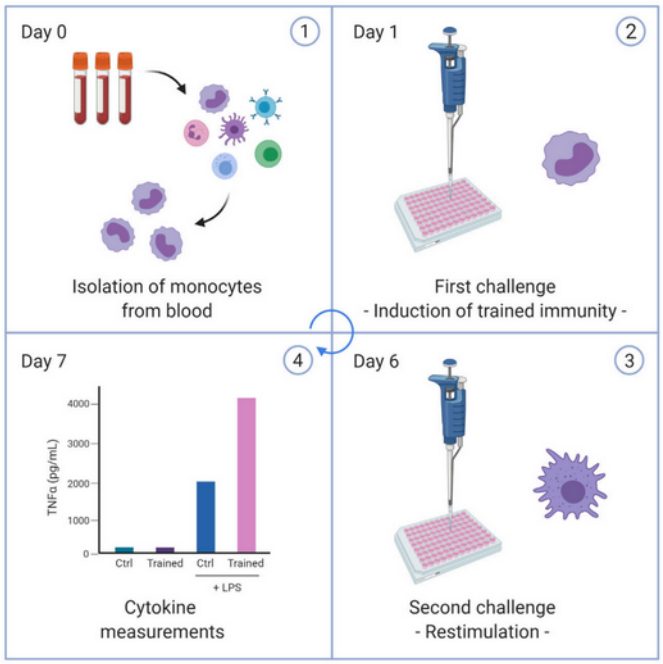


Figure 5.1: Illustration of experimental procedure for trained immunity data collection. Reprinted from [173].

Table 5.1: Cytokines included in the dataset, grouped by function.

Function	Cytokine	Complete Name
Pro-inflammatory	TNFα	Tumor necrosis factor-alpha
	IFNγ	Interferon gamma
	G-CSF	Granulocyte Colony-Stimulating Factor
	IL-1β	Interleukin-1beta
	IL-6	Interleukin-6
	IL-8	Interleukin-8
	IL-17 A	Interleukin-17 A
	IL-18	Interleukin-18
	IL-23	Interleukin-23
Regulatory	CX3CL1	Fractalkine
	GM-CSF	Granulocyte-Macrophage Colony-Stimulating Factor
	MCP-1	Monocyte Chemoattractant Protein-1
	VEGF-A	Vascular Endothelial Growth Factor
	IL-2	Interleukin-2
	IL-4	Interleukin-4
	IL-12p70	Interleukin-12
Regulatory/ Anti-inflammatory	IL-1RA	Interleukin-1 Receptor Antagonist
	IL-33	Interleukin-33
Anti-inflammatory	IL-10	Interleukin-10
	BDNF	Brain-Derived Neurotrophic Factor

For the assessment of IB, an ELISA diagnostic test was employed on serum samples

from individuals. The recorded values represent absorbance measurements that are directly proportional to the levels of specific Immunoglobulin G (IgG) antibodies for the respective pathogen. Values under 0.8 are considered negative for past infection with that microbe, positive values are equal or greater than 1.1 and values in between these thresholds are considered borderline cases.

Serum cytokine levels were also obtained via ELISA on serum samples of the subjects.

5.2 General Approach

The present project was divided in 3 main ML tasks: Predicting Cytokine Levels - Regression (described in Section 5.3.1), Predicting Cytokine Levels - Classification (found in Section 5.3.2) and Predicting Alzheimer's Disease - Binary Classification (Section 5.3.3). Additionally, for comparison, age group (being over or under the age of 65) was also predicted based on TI data (Section 5.3.4).

All this multifaceted analysis sought to explore the issue from all angles while still relying on the same core methodology, which can be visualized in Figure 5.2.

Firstly, the data underwent cleaning and preprocessing (including data scaling) before analysis. As will be explained later on in Section 5.3.3, when dataset size allowed it, a test set was set aside. Subsequently, a nested Cross-validation (CV) grid-search was implemented to identify the optimal hyperparameters for each given algorithm, and provide a general assessment of model performance, guiding the decision to proceed to the test set evaluation and construction of final models. Details on hyperparameters explored for the regressors and classifiers deployed are provided in Sections 5.4.1 and 5.4.2, respectively. Section 5.4.3 provides the specifics of the implementation of the nested CV strategy. The algorithms explored for both regression and classification are presented in Table 5.2. All the models were implemented using the scikit-learn library[174].

In scenarios where the dataset size permitted the division of a separate test set, models were constructed using the entire dataset utilized in the nested CV procedure with the optimized hyperparameters. These models were then evaluated on the separate test set, with respective performance metrics recorded and plotted for visualization. Furthermore, the best models were interpreted using SHAP.

Additionally, to address class imbalance issues, an attempt was made to oversample the training data in every instance it was used (both in the nested CV procedure and training of final models). For regression tasks, the SMOGN technique was employed, while for classification tasks, the SMOTE method was utilized, (details for these algorithms can be found in Section 5.4.4). To enable comparative analysis, results for models trained with and without oversampling were separately documented.

Table 5.2: Table of the algorithms deployed for both regression and classification.

Regression Algorithms	**Classification Algorithms**
Linear Regression	Logistic Regression
Lasso Regression	Logistic Lasso Regression
Ridge Regression	Logistic Ridge Regression
Elastic-Net (EN) Regression	Logistic Elastic-Net (EN) Regression
Decision Trees (DT)	
Random Forest (RF)	
Support Vector Machine (SVM)	
K-Nearest Neighbors (KNN)	
Extreme Gradient Boost (XGB)	

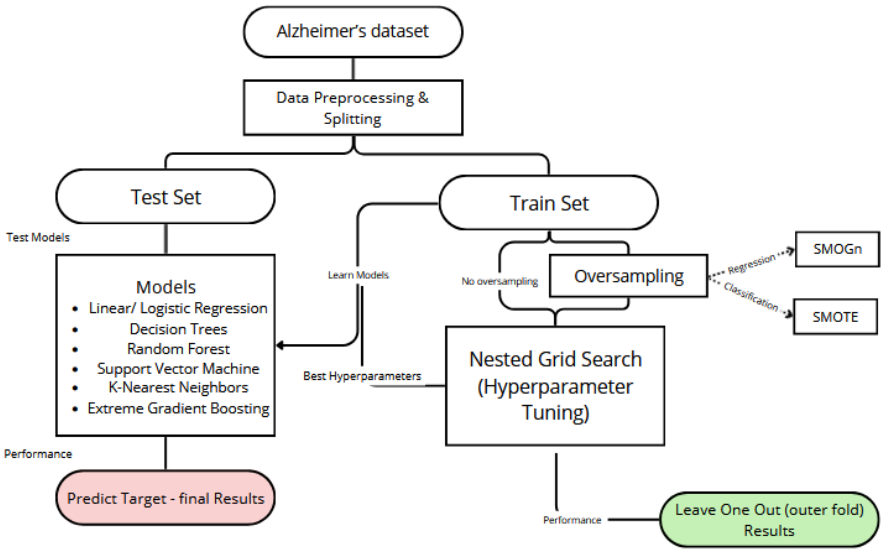


Figure 5.2: Schematic overview of methodology pipeline.

5.3 Experimental Procedures

5.3.1 Predicting TI Cytokine Levels - Regression

The task of predicting TI cytokines consisted of multiple targets, effectively representing several independent tasks. However, all tasks shared the same set of features. Training data consisted of IB data (antibody levels for HSV, CMV, *H. pylory*, *B. burgdorferi* and *C. pneumoniae*), sex, age and whether subjects had Alzheimer's disease. Target variables were TNFα, IL-6, IL-10, IL-1β and IL-1RA in each of the 5 considered stimulation conditions: NT, Pr LPS, Pr Ca, Primed with LPS and then challenged with LPS (LPS LPS) and Primed with *C. albicans* and then challenged with LPS (Ca LPS), with the exception of IL-1β where there was no data available for the challenge conditions, as displayed on Table 5.3.

Table 5.3: Table of targets for regression task.

Cytokine	Condition	Dataset size	Oversampling possible	Samples added by SMOGN
TNFα	NT	32	yes	18
	Pr LPS	26	yes	16
	Pr Ca	27	no	-
	Ca LPS	16	no	-
	LPS LPS	34	yes	17
IL-6	NT	28	yes	14
	Pr LPS	19	yes	9
	Pr Ca	27	no	-
	Ca LPS	13	no	-
	LPS LPS	29	yes	16
IL-10	NT	22	yes	14
	Pr LPS	16	no	-
	Pr Ca	22	yes	14
	Ca LPS	16	yes	12
	LPS LPS	20	no	-
IL-1β	NT	32	yes	19
	Pr LPS	32	yes	22
	Pr Ca	32	yes	19
IL-1RA	NT	29	no	-
	Pr LPS	29	no	-
	Pr Ca	29	no	-
	Ca LPS	17	no	-
	LPS LPS	20	no	-

Data preprocessing involved identifying the desired target column and eliminating rows where the label was missing, resulting in variations in dataset size for each target, as shown on Table 5.3. Data was scaled using MinMax Scaler.

Following this, hyperparameter optimization via nested Leave-one-out Cross Validation (LOOCV) for each of the algorithms mentioned in the regression column of Table 5.2 was executed and performance assessed. Hyperparameters grid-searched and details for CV are presented in Sections 5.4.1 and 5.4.3, respectively. During this procedure, oversampling of the training data was attempted by means of the SMOGN algorithm. As shown on Table 5.3, this algorithm was not able to perform oversampling on all cases, for some target variables' distribution did not contain box plot extremes which the algorithm requires in order to oversample the data.

Finally, the actual and predicted values for the validation instance in each outer fold of the nested CV were stored, allowing, at the end of each execution, for the computation of the MdAE, the MAE, the absolute error's standard deviation and the Pearson's Correlation Coefficient between real and output values. Visualizations with this gathered data were built, namely the actual vs. predicted scatter plot and the distribution of residuals.

5.3.2 Predicting Cytokine Levels - Multi-Class Classification

The previously described ML problem was converted to a multi-class classification task, in an attempt to simplify the problem, increase robustness to outliers as well as achieve meaningful insights into the model's performance by means of different metrics (such as accuracy and F1 score).

For the purpose of this conversion, the tertiles approach was deployed, based on the distribution of all considered protein values combined. The continuous protein levels were divided into three categories based on their distribution: Low (first tertile, values in the bottom 33%), Medium (second tertile, values between 34% and 66%), and High (third tertile, values in the top 33%). The final distribution of values for each target is represented on Figure 5.3.

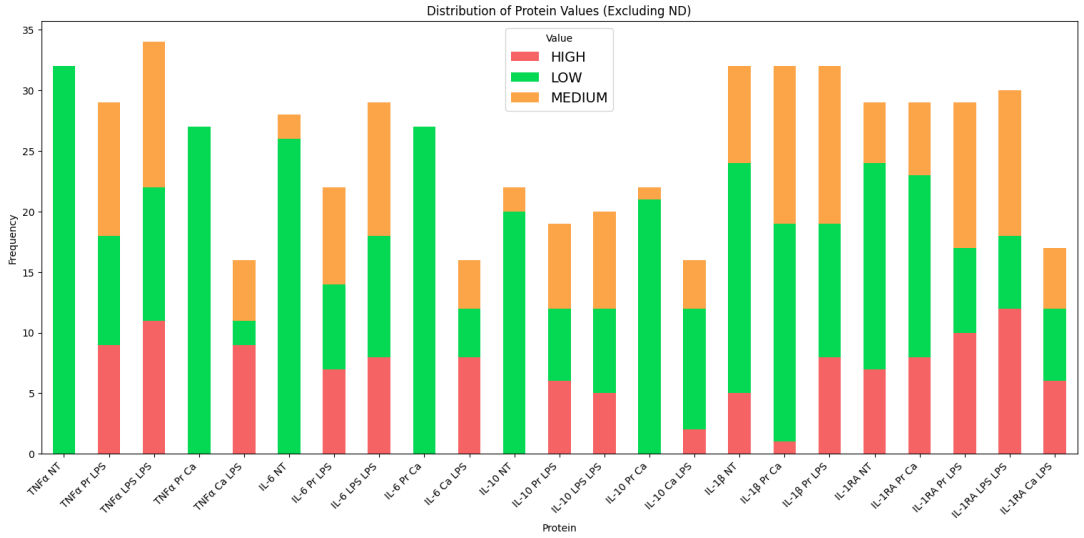


Figure 5.3: Distribution of discrete trained immunity protein values.

As can be seen in Figure 5.3, not all TI proteins considered in regression exhibited a distribution encompassing all three classes, which led to the exclusion of certain targets. A lack of sufficient samples in each class—specifically, having fewer than three instances per class— means that the model cannot reliably learn the characteristics of those classes due to insufficient representative data. This situation increases the risk of overfitting, renders the performance metrics unreliable, and results in such significant class imbalance that there aren't enough instances of the minority class to enable effective oversampling. Accordingly, TNFα NT, TNFα Pr Ca and IL-6 Pr Ca had to be disregarded as targets due to only presenting "Low" values. IL-6 NT, IL-10 NT and IL-10 Pr Ca were also disregarded for not presenting any instances of class "High" and presenting less than 3 instances of class "Medium". Lastly, TNFα Ca LPS presented only 2 "Low" instances, and IL-10 Ca LPS and IL-β Pr Ca presented less than 3 instances for class "High", thus these variables were also not accounted for.

After establishing the possible targets for this task, an approach very similar to

the one described in Section 5.3.1 was followed. Features considered were once more antibody levels for HSV, CMV, *H. pylory, B. burgdorferi* and *C. pneumoniae*, sex, age and whether subjects had AD. Data was cleaned according to specific target column, rows where the label was missing were eliminated and data was scaled. For every method listed in the classification column of Table 5.2, hyperparameter optimization using nested LOOCV was carried out, and performance was evaluated. Grid-searched hyperparameters and CV details are provided in Sections 5.4.2 and 5.4.3, accordingly.

Once again, the predicted and actual values for the validation instance in each outer fold of the nested CV were saved, enabling the computation of the general accuracy, precision, recall and F1 score at the conclusion of each execution. With the use of this data, the confusion matrix for each method and each target variable was plotted to facilitate analysis of model performance.

5.3.3 Predicting Alzheimer's Disease - Binary Classification

Predicting AD was a central task to this thesis. Due to differences in objectives and variations in dataset size and available features in line with those objectives, the task of predicting AD was divided into 4 different subtasks. Methodology for predicting AD from infectious burden data is explored in Section 5.3.3.1, prediction from trained immunity data in Section 5.3.3.2, prediction from both IB and TI data simultaneously is presented in Section 5.3.3.3 and, lastly, Section 5.3.3.4 refers to modelling AD from serum cytokine data. Main differences between these subtasks can be consulted on Table 5.4, from dataset size, separation of an independent test set and its number of samples, existence of missing data and nested CV outer fold strategy. In all these tasks age was discarded as a feature due to high correlation with AD.

Table 5.4: Summary of major differences between subtasks of predicting AD.

Predict AD	Total sample size	independent test set	test set size	Nones in data	CV outer fold
From IB	35	no	-	no	Leave-one-out (LOO)
From TI	38	yes	8	yes	LOO
From IB + TI	38	yes	8	yes	LOO
From serum	99	yes	20	yes	8 fold

5.3.3.1 Predicting AD from Infectious Burden Data

Firstly, with the intuition of exploring the infection hypothesis and thus the possibility of leveraging IB data in order to predict AD, antibody levels for HSV, CMV, *H. pylory, B. burgdorferi* and *C. pneumoniae* were selected as features, along with sex of the subjects.

This selection resulted in a dataset composed of 35 instances (15 HCs and 20 AD patients). Owing to this small sample size, it was not viable to separate an independent test set, thus, only LOOCV results are available. There was no missing data in the training set, so data needed only to be scaled, which was executed using MinMax scaler.

SMOTE algorithm was deployed for oversampling, creating 5 additional artificial samples for HCs. Nested LOOCV was performed for each technique indicated in the classification column of Table 5.2, for hyperparameter optimization and performance assessment. General accuracy, precision, recall and F1 score, based on the validation fold of this procedure was computed, as well as the confusion matrix for each ML model.

5.3.3.2 Predicting AD from Trained Immunity Data

This predictive modelling task had the objective of exploring relationships between AD and selected TI proteins. Features selected were the ones used as targets for predicting cytokine levels, expressly, TNFα, IL-6, IL-10, IL-1β and IL-1RA in different stimulation conditions (as present on Table 5.3). However, IL-6 in conditions Ca LPS was dropped because of the high proportion of missing values (over 50%). Subjects' sex was also included.

This dataset comprised 38 instances (18 HCs and 20 AD patients). A small test set of 8 samples (4 HCs and 4 ADs) was set aside for the settlement of final models. Data was scaled and missing data entries were filled with mean, median or mode of column (independent results for these 3 strategies were assessed). Training data (composed of 16 patients and 14 controls) was oversampled using SMOTE, creating 2 additional artificial controls (details for SMOTE found in Section 5.4.4).

Nested LOOCV was performed, information on optimal hyperparameters and performance was stored. For building the final models, the classification algorithms were trained on the 30 samples used in the nested procedure with the optimal hyperparameters found. These were chosen based on most frequent combination of hyperparameters yielded via LOOCV. In cases where 2 or more combinations were found with the same frequency, a random one of them was chosen (*seed* was set to 12 for reproducibility).

At last, model's performance was assessed using the test set. This evaluation included metrics such as overall accuracy, precision, recall, and F1 score, as well as these metrics specific to AD. Additionally, the evaluation involved calculating the AUC, plotting the test set's confusion matrix, and generating the ROC curve. Best models were interpreted using SHAP kernel explainer, specifically via importance bar and beeswarm summary plots.

5.3.3.3 Predicting AD from Infectious Burden and Trained Immunity Data

This task, aiming to explore relationships between both infection and TI cytokines with Alzheimer's, combined features from infectious burden and trained immunity. The resulting features included: sex, antibody levels for HSV, CMV, *H. pylory, B. burgdorferi* and *C. pneumoniae*, TNFα, IL-10, and IL-1RA under all five considered stimulation conditions, IL-6 under NT, Pr LPS, Pr Ca, and LPS LPS conditions, and IL-1β under NT, Pr LPS, and Pr Ca conditions.

Notwithstanding, the methodology for this task is nearly identical to the previously described approach. The dataset consisted of 38 samples, with 8 samples allocated to the test set and the remaining 30 used for hyperparameter optimization and final model training. Missing data was filled with mean, median or mode. Oversampling via SMOTE led to the addition of 2 artificial controls. Final models created with optimal hyperparameters were interpreted using SHAP.

5.3.3.4 Predicting AD from Serum Cytokine Levels

Lastly, on the topic of constructing models to predict AD, serum cytokine levels made up a substantially larger dataset. These serum records, were obtained for 99 patients (51 HCs and 48 AD patients) and included levels for each cytokine described in Table 5.1.

Accordingly, we utilized these serum cytokine levels as features for building predictive models, along with the sex of the patients. The larger size of this dataset, allowed for the allocation of a test set comprised of 10 patients and 10 controls. IL-4, IL-33, IL-23, Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF), IL-2, and IL-12p70 were excluded from the dataset due to a high proportion of missing values (>50%). The training set with the remaining features was then preprocessed by imputing missing values with the mean, median, or mode, and subsequently scaled. To keep in line with the chosen approach, an attempt to oversample the data via SMOTE was still deployed, however resulting in the addition of just 3 artificial patients to the dataset.

Initial performance was estimated and hyperparameters optimized for each of the classifiers via nested CV. However, due to the larger dataset size, LOO strategy was not deployed, instead a 20-fold CV composed the outer layer as specified in Section 5.4.3. This resulted in more robust results for the validation layer, which led to a mildly different strategy for the choice of optimal hyperparameters. These were once again primarily chosen by most frequent combination of hyperparameters found, however when ties arose, the combination which yielded the highest AUC was selected.

Afterwards, final models were set up and performance evaluated on test set (by means of train and test accuracies, precision, recall and F1 score, general and specific to AD, and AUC). Confusion matrices and ROC curves were plotted for each model. The SHAP kernel explainer was utilized to interpret the best models, through the use of importance bar and beeswarm summary charts.

5.3.4 Predicting Age (Over/Under 65) for Comparison

One concluding task was to attempt to predict whether subjects were over the age of 65 from TI data, for comparison with the task of predicting AD from TI as well.

For this purpose, target variable age was binarized (1 for subjects over 65 and 0 otherwise). Features considered were: sex (since this was also included in prediction of AD), and the monocyte expressed cytokines in different stimulation conditions described in Sections 5.3.3.2 and 5.3.3.3. Thus, the dataset resulted again in 38 samples,

with 8 being separated for the constitution of the test set. In this case, due to a larger proportion of elderly individuals in the data, a stratified train test split was utilized, and the test set was composed of 5 subjects over 65 and 3 under this threshold. The train set consisted in 19 samples belonging to the elder group and 11 of the alternative class.

The approach used is as previously described, with the filling of missing data by means of median, mean or mode, hyperparameter optimization via nested LOOCV and final models being set up and evaluated on test set. SMOTE was deployed for oversampling, in this case creating 8 artificial samples (labelled under 65). Final models with ideal hyperparameters were interpreted via SHAP.

With the experimental setups for each task having been outlined, and to provide a clearer overview of the data characteristics and treatments used, a summary table is presented below (Table 5.5). This table encapsulates key information regarding dataset sizes, excluded instances, data treatment strategies, and oversampling techniques across all tasks.

Table 5.5: Overview of dataset characteristics and treatment for each task.

Task	Dataset Size	Number of excluded training instances	Data Treatment	Oversampling Technique
Predicting TI cytokine levels- regression	Varying from 13 to 34 - see Table 5.3	0	Cleaning and normalization	SMOTE
Praedicting TI cytokine levels- classification	Varying from 13 to 34 - see Figure 5.3	0	Cleaning and normalization	SMOGN
Predict AD from IB	35	0	Cleaning and normalization	SMOTE
Predict AD from TI	38	8 - for test set	Cleaning, normalization, imputation of nones	SMOTE
Predict AD from IB and TI	38	8 - for test set	Cleaning, normalization, imputation of nones	SMOTE
Predict AD from serum	99	20 - for test set	Cleaning, normalization, imputation of nones	SMOTE
Predict age group from TI	38	8 - for test set	Cleaning, normalization, imputation of nones	SMOTE

5.4 Experimental Settings

This section describes with more detail the experimental settings adopted. Specifically, Subsections 5.4.1 and 5.4.2 entail the hyperparameter grids for each algorithm used in this project. Subsection 5.4.3 elaborates on CV specifics and lastly, Subsection 5.4.4 defines oversampling settings.

5.4.1 Hyperparameters for Each Regressor

On the topic of model hyperparameters, we aimed for a comprehensive Grid Search (GS), striving for thoroughness while balancing the necessity to control computational

time effectively.

The hyperparameters searched for in each of the regression algorithms are present in Table 5.6. For basic linear regression, only 1 hyperparameter was optimizable which was whether to calculate the intercept for this model. For Ridge, Lasso and EN regressions, similar and comprehensive grids were searched, with alpha (a common parameter among them) being tested across values ranging from 1×10^{-5} to 100, spanning several orders of magnitude. The chosen range aims to explore a broad spectrum of regularization strengths, ensuring performance of the models is evaluated across very weak to very strong regularization, thereby increasing the likelihood of finding the optimal alpha value for each model. As for EN, an additional parameter was tuneable, the L1 ratio, that is the balance between both penalties. As *l1_ratio = 0* equals applying an L2 penalty and *l1_ratio = 1* signifies an L1 penalty, the selected range allows for a thorough exploration in order to identify the optimal combination of L1 and L2 regularization.

In the case of DTs and RFs, the searching criteria was to ensure a comprehensive search over key parameters, while minimizing computational time. Thus, variations in tree depth, splitting criteria, and minimum samples for splits and leaves were included, concerning finding a balance between underfitting and overfitting. Values covered in these intervals were rather small, adapted to our small size dataset. For RF, number of estimators was optimized between 50, 100 and 200, striking a balance between ensemble stability and computational efficiency.

Regarding SVR, all kernels were considered, permitting the exploration of different relationships in the data. C parameter (which controls regularization strength) and epsilon (which determines the margin of tolerance within which no penalty is given to errors), were searched across small ranges that comprehended slight increases of default values, allowing to adjust model flexibility.

Concerning KNN, many parameters were explored within all their possible values (weight function, algorithm to compute nearest neighbors and power parameter). Number of neighbors ranged from 3 to 7, balancing between having too few neighbors, which might lead to overfitting and high variance, and too many neighbors (as 7 is rather a high value for our number of available samples), which might lead to underfitting and high bias.

Lastly, XGB's grid was designed similarly to RF for number of estimators and maximum depth (albeit including smaller values in this last interval, as this model works with smaller trees by default). Gamma parameter tested varying regularization strengths and the learning rate covered a broad spectrum of step sizes for updating the model weights, with smaller values requiring more iterations but potentially improving generalization. Here, (as with RF), the goal was to conduct GS over critical hyperparameters within the limited computing time limitations, being cautious not to overextend the search.

The range of values for all hyperparameters always included default values, as these

are a reasonable starting point and have been found as the optimal value in a number of tasks [108]. Therefore, including them ensures a balanced and effective exploration of model settings, leveraging well-known optimal combinations while looking for possible enhancements. This is also true for hyperparameter grids for classifiers.

Table 5.6: Table for each regressor with detailed hyperparameter grids. Any hyperparameter not mentioned was left as default by scikit learn.

Model	Hyperparameter	Values searched	Description
Linear Regression	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
Lasso Regression	alpha	[0.00001, 0.0001, 0.001, 0.01, 0.01, 1, 10, 100]	Constant that controls regularization strength.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	selection	[cyclic, random]	Controls wheter coefficients are updated sequentially or at random.
Ridge Regression	alpha	[0.00001, 0.0001, 0.001, 0.01, 0.01, 1, 10, 100]	Constant that controls regularization strength.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	solver	[auto, svd, cholesky, lsgr, sparse_cg, sag, saga]	Algorithm to optimize loss function.
Elastic-Net Regression	alpha	[0.00001, 0.0001, 0.001, 0.01, 0.01, 1, 10, 100]	Constant that controls regularization strength.
	l1_ratio	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	Balance between L1 and L2 regularization.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	selection	[cyclic, random]	Controls whether coefficients are updated sequentially or at random.
Decision Trees	criterion	[squarred_error, friedman_mse]	Function to measure the quality of a split.
	max_depth	[None, 5, 10, 15]	The maximum depth of a tree.
	min_samples_split	[2, 5, 10]	The minimum number of samples required to split an internal node.
	min_samples_leaf	[1, 2, 4]	The minimum number of samples required to be at a leaf node.
	max_features	[None, sqrt, log2]	The number of features to consider when looking for best split.
Random Forest	n_estimators	[50, 100, 200]	The number of trees in the forest.
	max_depth	[None, 10, 20, 30]	The maximum depth of the trees.
	min_samples_split	[2, 5, 10]	The minimum number of samples required to split an internal node.
	min_samples_leaf	[1, 2, 4]	The minimum number of samples required to be at a leaf node.
Support Vector Regressor	kernel	[linear, rbf, poly, sigmoid]	The kernel type to be used in the algorithm.
	C	[1, 10, 100]	Strength of regularization (inversely proportional to C).
	epsilon	[0.1, 0.2, 0.5]	Defines the margin within which no penalty is given to prediction errors.
K-Nearest Neighbors	n_neighbors	[3, 4, 5, 6, 7]	Number of neighbors to use.
	weights	[uniform, distance]	Weight function to be used (distribute weights uniformly, or by inverse of distance)
	algorithm	[ball_tree, kd_tree, brute]	Algorithm to compute nearest neighbors.
	p	[1, 2]	Power parameter (1 uses Manhattan distance, 2 uses Euclidian distance)
Extreme Gradient Boost	n_estimators	[100, 200, 500]	Number of gradient boosted trees.
	max_depth	[3, 6, 9]	The maximum depth of a tree.
	gamma	[0.01, 0.1]	Minimum loss reduction needed to split a leaf node in a tree.
	learning_rate	[0.001, 0.01, 0.1, 1]	Step size shrinkage used in update to prevent overfitting.

5.4.2 Hyperparameters for Each Classifier

The hyperparameters searched for each classifier can be consulted on Table 5.7.

The strategy for basic logistic regression and logistic regression with L1, L2 and elastic-net penalties was in every way identical to the one described previously in

Section 5.4.1 for the linear regressors, with the solvers chosen for each of the algorithms being in accordance with the penalty applied.

Hyperparameter grids for DTs and RF are also the same as described formerly for the regressors (with the criterion parameter for DT being adapted to classification). The same is also true for the KNN algorithm. For Support Vector Classifier (SVC), probability estimates were enabled and as computation time allowed for it, gamma parameter (which certain kernels make use of) was also included in the search (between the 2 possible built-in values: *auto* or *scale*). Finally, due to computational efficiency on our high-dimensional datasets in several classification tasks, a linear booster model for XGB was implemented, a method which uses a linear model as the base learner (or booster) [175]. In this way, the chosen hyperparameters to optimize were alpha and lambda (corresponding to L1 and L2 regularization terms, respectively), the learning rate as before and the choice of algorithm to fit the linear model (either shotgun or ordinary coordinate descent algorithm).

Table 5.7: Table for each classifier with detailed hyperparameter grids. Any hyperparameter not mentioned was left as default by scikit learn.

Model	Hyperparameter	Values searched	Description
Basic Logistic Regression	penalty	None	Specifies the norm of the penalty.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	solver	[newton-cg, lbfgs, sag, saga]	Algorithm to optimize loss function.
Logistic Lasso Regression	penalty	L1	Specifies the norm of the penalty.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	C	[0.001, 0.01, 1, 10, 100]	Inverse of regularization strength.
	solver	[liblinear, saga]	Algorithm to optimize loss function.
Logistic Ridge Regression	penalty	L2	Specifies the norm of the penalty.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	C	[0.001, 0.01, 1, 10, 100]	Inverse of regularization strength.
	solver	[liblinear, newton-cg, lbfgs, sag, saga]	Algorithm to optimize loss function.
Logistic Elastic-Net Regression	penalty	elasticnet	Specifies the norm of the penalty.
	fit_intercept	[True, False]	Whether to calculate the intercept for the model.
	C	[0.001, 0.01, 1, 10, 100]	Inverse of regularization strength.
	solver	[liblinear, newton-cg, lbfgs, sag, saga]	Algorithm to optimize loss function.
	l1_ratio	[0.3, 0.4, 0.5, 0.6, 0.7]	Balance between L1 and L2 regularization.
Decision Trees	criterion	[gini, entropy]	Function to measure the quality of a split.
	max_depth	[None, 5, 10, 15]	The maximum depth of a tree.
	min_samples_split	[2, 5, 10]	The minimum number of samples required to split an internal node.
	min_samples_leaf	[1, 2, 4]	The minimum number of samples required to be at a leaf node.
	max_features	[None, sqrt, log2]	The number of features to consider when looking for best split.
Random Forest	n_estimators	[50, 100, 200]	The number of trees in the forest.
	max_depth	[None, 10, 20, 30]	The maximum depth of the trees.
	min_samples_split	[2, 5, 10]	The minimum number of samples required to split an internal node.
	min_samples_leaf	[1, 2, 4]	The minimum number of samples required to be at a leaf node.
Support Vector Classifier	probability	True	Enables probability estimates.
	C	[0.1, 1, 10, 100]	Strength of regularization (inversely proportional to C).
	kernel	[linear, rbf, poly]	The kernel type to be used in the algorithm.
	degree	[2, 3]	Degree for "poly" kernel.
	gamma	[scale, auto]	Defines how far the influence of a single training instance reaches.
K-Nearest Neighbors	n_neighbors	[3, 4, 5, 6, 7]	Number of neighbors to use.
	weights	[uniform, distance]	Weight function to be used (distribute weights uniformly, or by inverse of distance).
	algorithm	[ball_tree, kd_tree, brute]	Algorithm to compute nearest neighbors.
	p	[1, 2]	Power parameter (1 uses Manhattan distance,2 uses Euclidian distance)
Extreme Gradient Boost	alpha	[0, 0.1, 1]	L1 regularization term on weights. The higher the more conservative the model.
	lambda	[0, 0.1, 1]	L2 regularization term on weights. The higher the more conservative the model.
	learning_rate	[0.001, 0.01, 0.1, 1]	Step size shrinkage used in update to prevent overfitting.
	updater	[shotgun, coord_descent]	Choice of algorithm to fit linear model.

5.4.3 Cross-Validation Details

For the tasks of predicting cytokine levels (both in regression and classification), predicting AD from TI, from IB, from TI and IB and of predicting age group, due to the limited size of the datasets (varying between 13 and 38, without oversampling), the nested CV was performed with a leave-one-out approach.

LOOCV is a CV technique that is particularly useful when the size of the dataset is limited. In LOOCV, the model is trained on all observations except one (naturally, test set observations are excluded as well). That one observation is used to esteem the predictive power of the model and the procedure is repeated the same number of times as the number of samples in the dataset [176]. Overall, this method provides a good estimate for the model's performance, but it is very computationally expensive (since it requires training n models on an n large dataset), hence it is ideal for small or imbalanced datasets [177].

In most tasks of this study this technique was viable for the outer loop of the nested approach, in order to achieve a robust measure of algorithm performance. The inner loop of the nested CV used a 5-fold cross-validation for hyperparameter tuning. The outer loop was then used to assess the general performance of the model with the optimal hyperparameters identified in the inner loop. To better illustrate this procedure the pseudocode outlining the steps involved in both the inner and outer loops is presented below.

Algorithm 1 Nested LOOCV and 5-Fold Cross-Validation.

Input: Dataset D with N samples, Model M, Hyperparameter set H
Output: Average performance of M across all N samples (LOOCV results)
for each sample i in dataset D **do**
    Split D into:
        Training set $D_{train} = D \setminus \{i\}$
        Test set $D_{test} = \{i\}$
    Inner Loop: Perform 5-Fold Cross-Validation on D_{train} for hyperparameter tuning
    for each fold f_j in D_{train} (with 5 folds) **do**
        Train M on $D_{train} \setminus f_j$
        Validate M on fold f_j
    end for
    Identify best hyperparameters H_{opt} from the 5-Fold CV
    Train final model M on full D_{train} using H_{opt}
    Test model M on D_{test} and record performance
end for
Aggregate performance metrics from all N iterations
Optional: For larger datasets, replace LOOCV with K-Fold CV in the outer loop:
    Set K (e.g., $K = 8$)
    Perform steps (1-13) but divide D into K folds for outer validation

In the case of predicting AD utilizing serum data, by virtue of the larger dataset size,

an 8-fold stratified CV was utilized instead of the LOOCV (represented in the optional part of the pseudocode provided). This decision was made for two fundamental reasons: firstly, the LOO approach would have been computationally prohibitive; secondly, with a sufficient number of samples (79 in this train set), the K-fold CV provided a more robust and reliable validation process. This 8-fold CV, allowed for the iterative division of approximately 70 samples to be used in the inner loop for hyperparameter tuning and around 10 samples to be left out for outer validation layer, providing a balanced division of the data.

5.4.4 Oversampling Techniques

The Synthetic Minority Over-Sampling Technique (SMOTE) algorithm is a common data resampling technique which functions by identifying the nearest k neighbors to any instance of the minority class and then generating new data points along the lines connecting these neighbors. This is achieved by selecting random positions on these lines, calculated as a combination of the original data point and a neighbor [178]. The formal representation is presented in equation 5.1.

$$X_{\text{new}} = X_i + (X_j - X_i) \times \text{rand}(0, 1) \tag{5.1}$$

Where:

- X_{new} is synthetic data created by SMOTE,
- $X_i \in T$ is the selected instance from the minority class,
- $X_j \in T$ is one of the K nearest neighbors of X_i,
- $\text{rand}(0, 1)$ is a random number between 0 and 1 that leads to the selection of a random point along the "line segment" between the instances.

SMOTE has become the benchmark for dealing with imbalanced data, proving, despite its simplicity, to be successful and robust when applied to several problems from various domains [179]. Nonetheless, regardless of its versatile nature, this technique, as described, functions only for classification problems [178]. In actuality, the continuous nature of the target variable, in regression problems, makes the oversampling task more complex, for, in theory, there could be an endless number of values to consider. Adding to that, there is also the issue of determining which values of the target are more or less relevant [180]. Some variations to the SMOTE algorithm, such as SMOTE for Regression (SMOTE-R) [181] or Geometric SMOTE (G-SMOTE) [182] have been presented to allow the use of this oversampling strategy for regression. An algorithm of particular relevance to this project was SMOGN [180], owing to its free availability and ease of implementation as a Python package. SMOGN combines two oversampling strategies: SMOTE-R and introduction of gaussian noise, with

random under-sampling, however the algorithm provides the flexibility to disable this undersampling feature. SMOGN generates new synthetic data points with SMOTE-R when the data point selected and its selected k-nearest neighbor are “close enough”, and uses the introduction of gaussian noise when instances are "more distant" [180]. This algorithm has shown to improve the performance of regression models in several different fields [183–185].

Referring to the details of the oversampling procedure used in this project, in the case of regression, SMOGN resampling algorithm was utilized on the training data. For its execution, undersampling was turned off, for it was not of interest to eliminate samples from the dataset. The *Seed* parameter was set to 1 for reproducibility and *samp_method* parameter was defined to "extreme" so a higher level of oversampling was performed, compared to the default setting. Other hyperparameters were set with their default values. Regarding classification, the algorithm for oversampling, SMOTE, only possesses 3 tuneable parameters: *sampling_strategy*, *random_state* and *k_neighbors*. *Sampling_strategy*, specifies the class targeted by the resampling, it was set to ’auto’ so the minority class is oversampled. *Random_state* was set to 42 for reproducibility. Finally, *k_neighbors* was set to 2 when predicting cytokine levels, as this was the highest number which enabled oversampling for all targets. When the task was to predict AD or age group, this parameter was set to its default value of 5 as this has been frequently found to be an optimal value for this parameter [186].

6

Results

This chapter presents the findings of the implementation of the approaches presented in the previous chapter. Section 6.1 introduces the general analysis of the dataset utilized. Section 6.2 entails the results for the models predicting various TI cytokine levels under different stimulation conditions, focusing on the most successful regression and classification models in each task. At last, results for models predicting AD are entailed in Section 6.3, with respective SHAP plots, for facilitation of interpretability. More extensive details of model results can be found in annex I.

6.1 General Data Analysis

For the purpose of general data analysis, the primary dataset was partitioned into two distinct subsets in order to enable a more focused investigation. The first subset, referred to as the serum dataset, encompasses all 99 participants from the study, with corresponding measurements of serum circulating cytokine levels. The second subset, termed the TI dataset, consists of 38 participants from the original 99, for whom TI measurements are available (refer to Table 5.5 in Chapter 5 for clarification on data available for each predictive task). Notably, within this TI subset, additional IB data was collected, however, due to constraints of data availability, only 35 participants have these corresponding measurements.

For the sake of clarity and methodological rigor, the analysis was conducted in two parts: the first one (presented on Subsection 6.1.1) focuses on the 38 samples in the TI subset, while the second section (presented on Subsection 6.1.2) focuses on the whole cohort of 99 samples in the serum dataset.

6.1.1 Analysis on TI and IB part of Dataset

Firstly, we looked at the distribution of sex for the TI part of the dataset. As depicted in Figure 6.1 there is overall a higher proportion of females in the dataset, however with sex being balanced across HCs and AD patients. Subsequently, distribution of age across the dataset was analysed. Figure 6.2 highlights that AD patients are generally

older than the HC subjects in this dataset (median age of 55 years for HCs versus median of around 78 years for AD patients). This is not unexpected as age is highly correlated with the development of AD, with, as has previously been mentioned, most patients' age of onset being over 65 years [24].

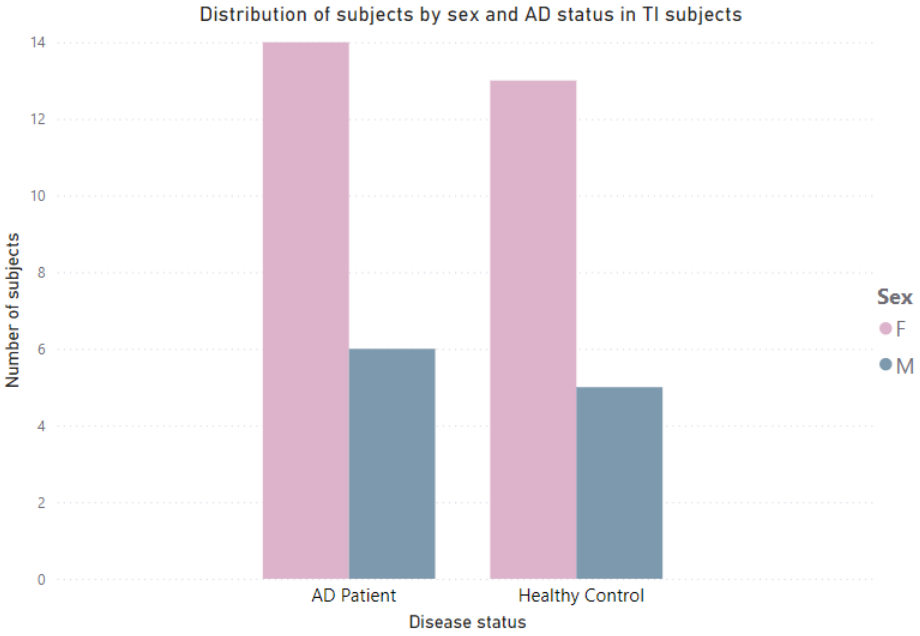


Figure 6.1: Box plot of TI subjects' sex per disease status

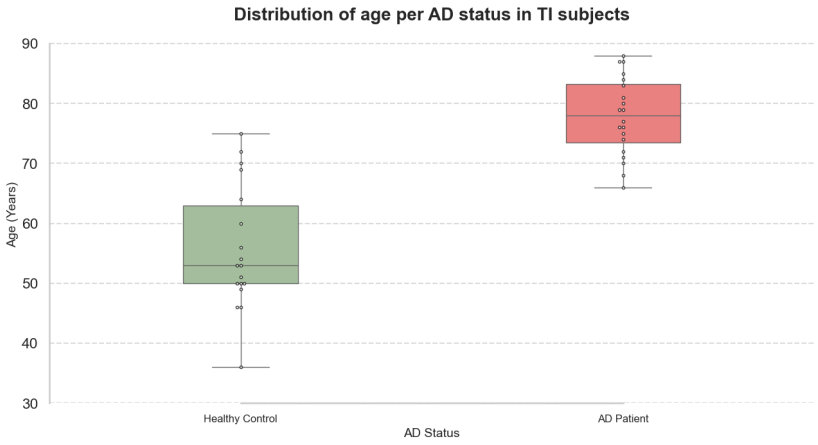


Figure 6.2: Box plot of TI subjects' age per disease status

Afterwards, we analysed the distribution of macrophage expressed TNFα, IL-6, IL-10, IL-1β and IL-1RA under the different stimulation conditions applied in the study and respective differences between HCs and subjects with AD. For TNFα (Figure 6.3), this cytokine appears to express the most variation between the 2 groups in conditions Pr LPS, with its expression in AD patients being enhanced. The same is true for IL-6 as evidenced in the boxplot of Figure 6.4. For IL-10, Figure 6.5 depicts apparently

close distributions for the 2 groups under all considered stimuli. For IL-1β, again this cytokine displays similar distributions within the 2 groups, however with a more widespread variation encompassing higher values in conditions Pr LPS (Figure 6.6). Lastly, IL-1RA seemingly presents similar distributions across AD patients and HCs, with slightly decreased median values for AD patients except for LPS LPS conditions (as observable in Figure 6.7).

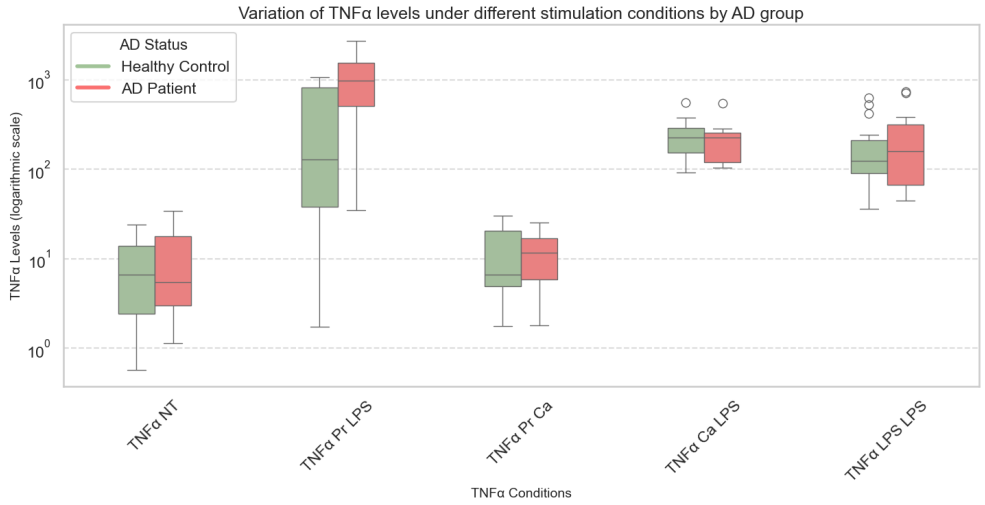


Figure 6.3: Box plot of macrophage expressed TNFα levels under different stimulation conditions per disease status. Logarithmic scale was applied.

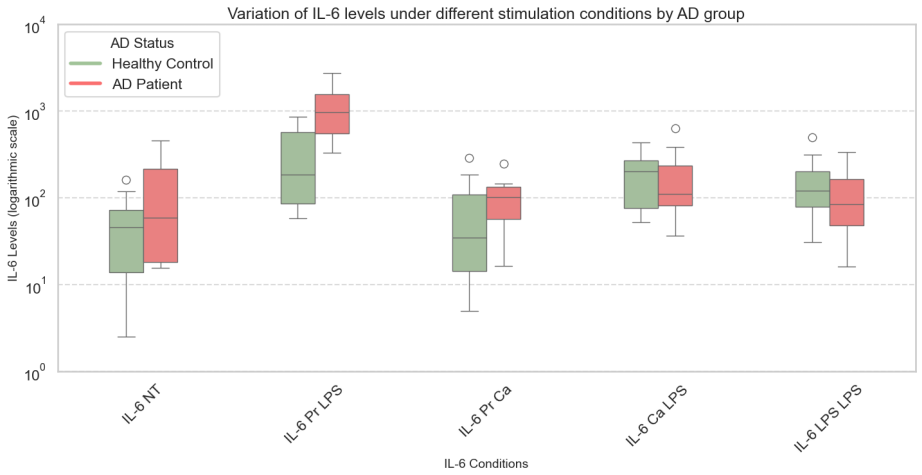


Figure 6.4: Box plot of macrophage expressed IL-6 levels under different stimulation conditions per disease status. Logarithmic scale was applied.

Following, we plotted the levels of IgG antibodies against the pathogens considered in this study, for comparison between HCs and AD patients as presented in Figure 6.8. Concerning, HSV-1/2, HCs show a wider distribution of IgG levels compared to AD patients, with median antibody levels being slightly increased in AD patients. On the other hand, the AD cohort exhibits a wider spread of IgG levels for *H. pylori*

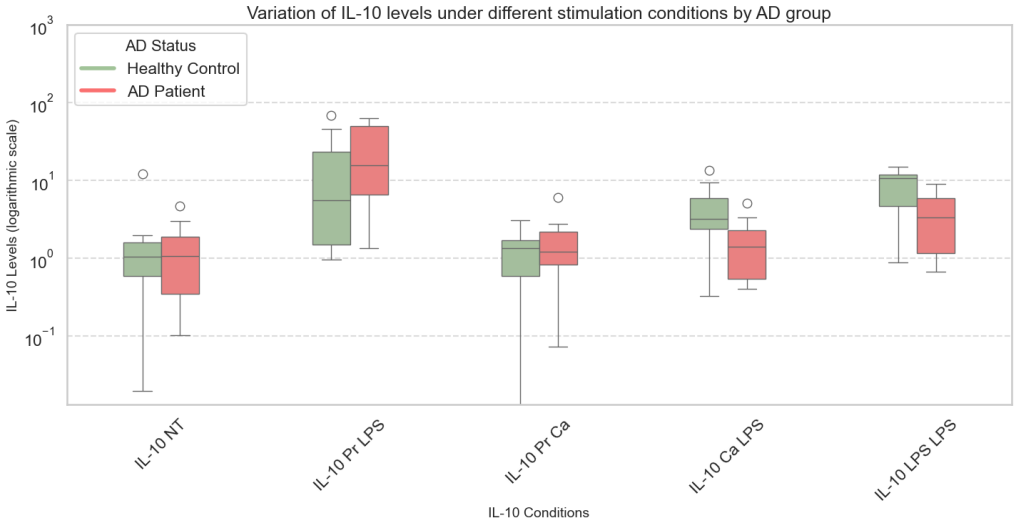


Figure 6.5: Box plot of macrophage expressed IL-10 levels under different stimulation conditions per disease status. Logarithmic scale was applied.

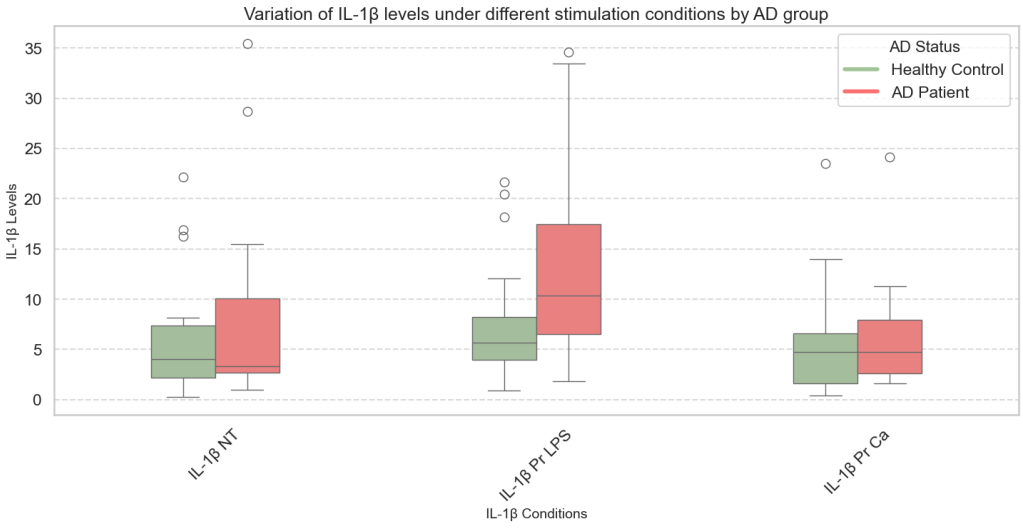


Figure 6.6: Box plot of macrophage expressed IL-1β levels under different stimulation conditions per disease status.

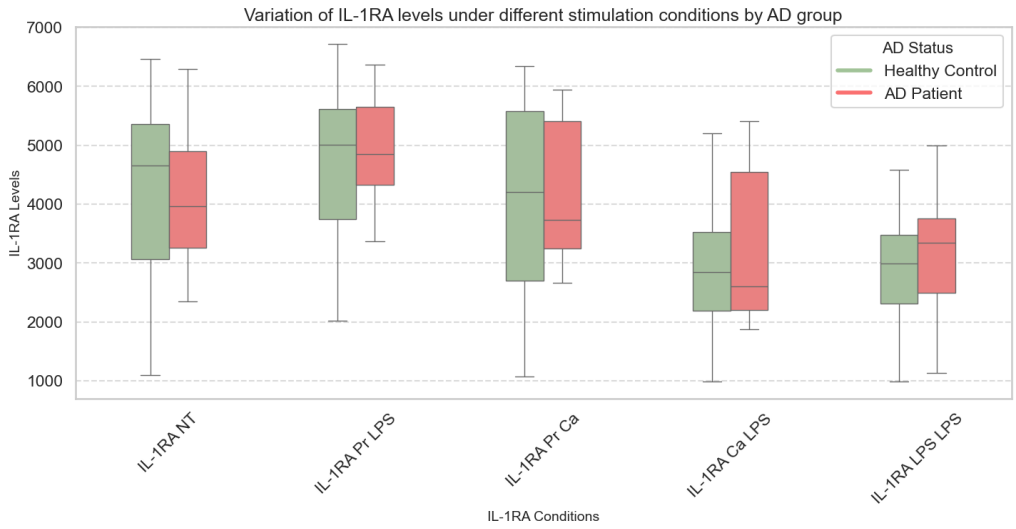


Figure 6.7: Box plot of macrophage expressed IL-1RA levels under different stimulation conditions per disease status.

with a higher median compared to HCs, indicating that some AD patients present significantly higher antibody levels for this pathogen. Regarding CMV, both groups display similar IgG distributions, with overlapping medians and similar interquartile ranges. Pertaining to *C. pneumoniae*, AD patients show a significantly higher median compared to HCs and also present a notable increase in the upper range of antibody levels for this bacterium. Lastly, both groups present relatively low IgG levels for *B. burgdorferi* with similar distributions and a small number of outliers.

Furthermore, the correlation matrix for these IB features was also plotted and is presented in Figure 6.9, allowing for the detection of moderate positive correlation between HSV-1/2 and CMV IgG levels, as well as a moderate negative correlation between CMV and *B. burgdorferi* serum antibody concentrations.

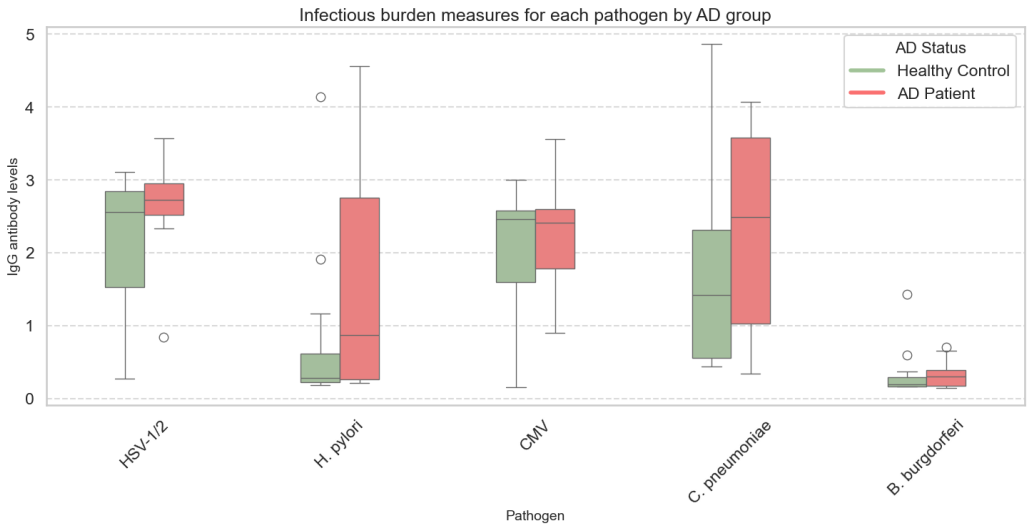


Figure 6.8: Box plot of IB levels for different pathogens per disease status.

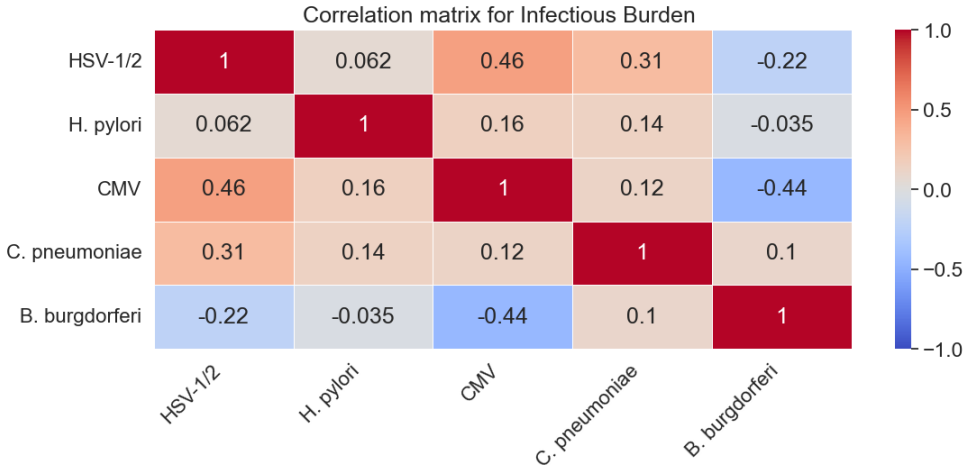


Figure 6.9: Correlation matrix for IB levels for different pathogens.

6.1.2 Analysis on Serum Part of Dataset

For the analysis of the broader serum dataset, again we looked into the distribution of sex across AD patients and HCs (Figure 6.10), with females representing once more a higher proportion of either class, with a slightly higher number of female controls compared to female AD patients. The boxplots for age (Figure 6.11), depict once more a higher proportion of elder individuals in the AD cohort, although this difference is less pronounced, with the median age for HCs being around 71 years and the median for AD patients being around 76 years.

In order to visualize variations in serum cytokine levels between AD patients and HCs, boxplots for the proteins were created and organized into separate plots based on their biological functions related to the disease as can be see in Figure 6.12. Figure 6.12a displays boxplots for serum concentration of pro-inflammatory cytokines. In general,

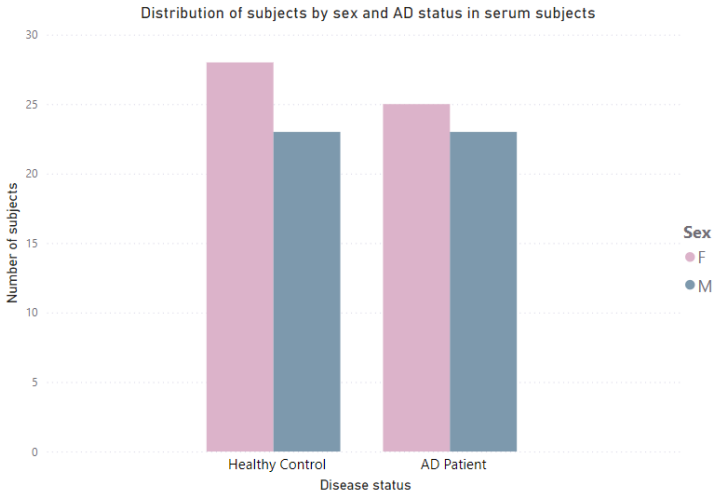


Figure 6.10: Box plot of serum subjects' sex per disease status.

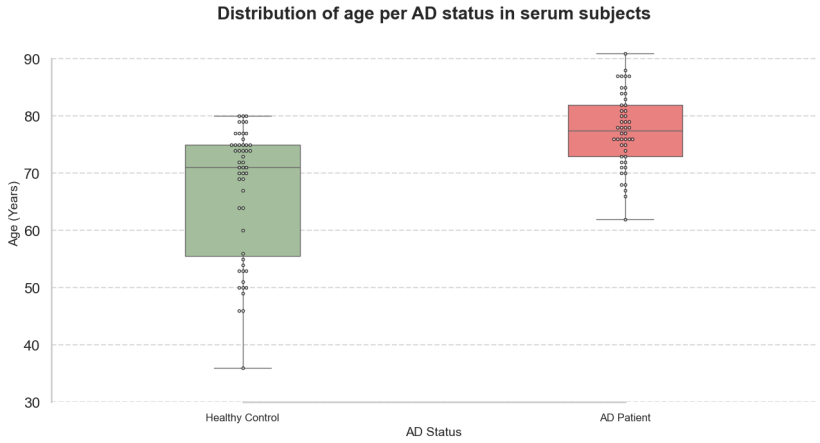


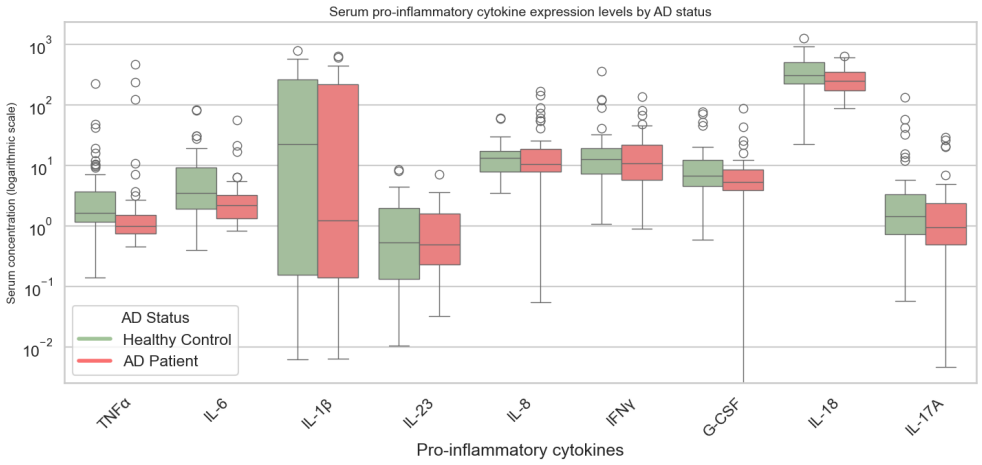
Figure 6.11: Box plot of serum subjects' age per disease status.

across most cytokines (e.g., TNFα, IL-6, IL-1β, Granulocyte Colony-Stimulating factor (G-CSF)), AD patients tend to have lower median concentrations than HCs, and also often display less variability, with the exceptions of IL-8, IFNγ, G-CSF and IL-17A which have broader or similar distributions with lower concentrations overall.

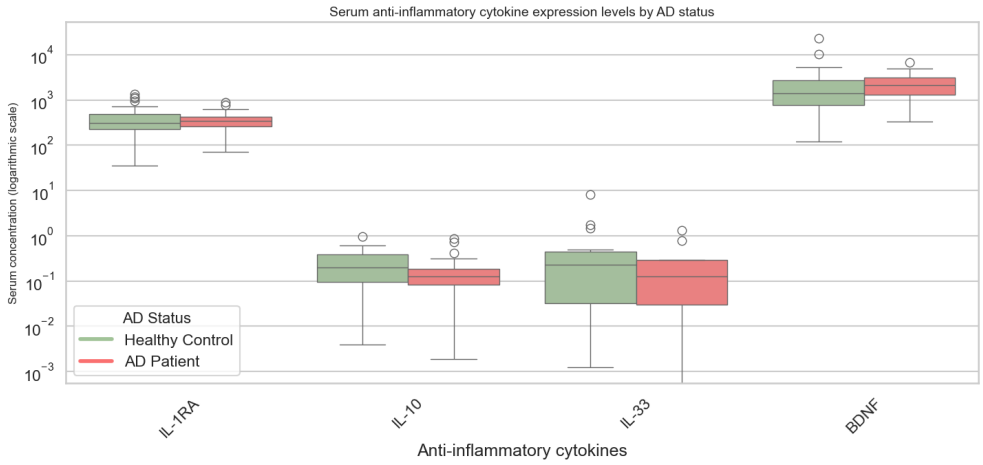
Figure 6.12b presents boxplots for serum concentration of anti-inflammatory cytokines. AD patients show narrower distributions for most cytokines, with the exception of IL-33 and in general also tend to have lower median values compared to HCs, BDNF being the only anti-inflammatory cytokine with a higher median level.

Finally, boxplots for serum concentration of regulatory cytokines are depicted in Figure 6.12c. Again the distributions of the majority of these cytokines are narrower in AD patients than in HCs. GM-CSF presents the most significant difference between the 2 groups with AD patients exhibiting higher levels of this cytokine. Concerning

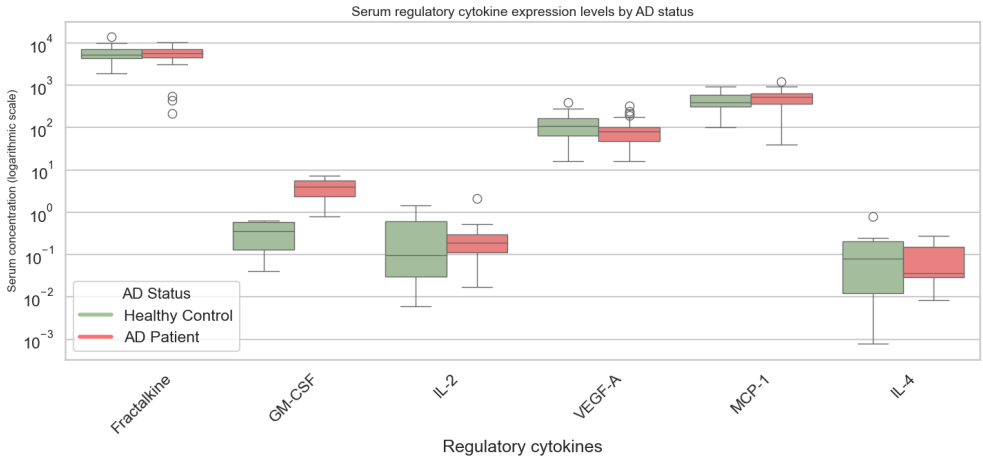
fractalkine, VEGF-A and MCP-1, both groups show high concentrations of these proteins, with minimal differences between the two. IL-2 and IL-4 present overall low concentrations, however AD patients have a higher median value for IL-2 and present a lower median value for IL-4 in comparison to the control group.



(a) Pro-inflammatory cytokine levels.



(b) Anti-inflammatory cytokine levels.



(c) Regulatory cytokine levels.

Figure 6.12: Box plots of serum circulating cytokine levels per disease status for (a) pro-inflammatory, (b) anti-inflammatory, and (c) regulatory cytokines. Logarithmic scale was applied.

6.2 Predicting TI Cytokine Levels

The objective of this part of the work was to assess the predictability of several macrophage- expressed cytokines under different stimulation conditions in AD patients and HCs based mainly on IB features (making use also of subjects' sex and age).

Table 6.1 summarizes the results of the most successful regression models employed, with and without oversampling the data, highlighting the best-performing algorithms for each cytokine and condition, along with their corresponding most illustrative metrics, the MdAE and Pearson's correlation coefficient (R value). Best results (in which the R value is higher than 0.3, indicating moderate correlation between actual and predicted values) were achieved for TNFα in NT and Pr LPS conditions, IL-6 in NT, Pr LPS and LPS LPS conditions, IL-10 Pr Ca, Ca LPS, and LPS LPS, IL-1β NT, Pr LPS and Pr Ca and IL-1RA in conditions NT, Pr LPS and LPS LPS. Notably, oversampling the data improved model performance in nearly every instance where it was feasible.

Table 6.1: Table of best results for predicting cytokine levels - regression.

			Only original data			With oversampled data		
Protein	**Treatment**	**Range of values**	**Best Algorithm**	**MdAE**	**Pearson's coefficient (R)**	**Best Algorithm**	**MdAE**	**Pearson's coefficient (R)**
TNFα	NT	0.65 - 27.93	Non predictive	-	-	XGB	5.792	0.509
	Pr LPS	1.75-2705.91	SVR	354.53	0.490	KNN	381.148	0.593
	Pr Ca	1.75 - 30.00	Non predictive	-	-	No oversample	-	-
	Ca LPS	92.12 - 556.73	Non predictive	-	-	No oversample	-	-
	LPS LPS	44.5 - 741.82	Non predictive	-	-	Non predictive	-	-
IL-6	NT	2.51 - 461.71	Non predictive	-	-	RF	61.976	0.320
	Pr LPS	76.04 - 2739.48	RF	520.27	0.316	RF	445.829	0.436
	Pr Ca	5.02 - 250.31	Non predictive	-	-	No oversample	-	-
	Ca LPS	36.93 - 640.01	Non predictive	-	-	No oversample	-	-
	LPS LPS	16.17 - 498.24	Linear Regression	88.88	0.338	RF	71.456	0.365
IL-10	NT	0.10 - 12.15	Non predictive	-	-	Non predictive	-	-
	Pr LPS	0.98 - 68.47	Non predictive	-	-	No oversample	-	-
	Pr Ca	0 - 6.05	Non predictive	-	-	Decision Trees	0.7012	0.359
	Ca LPS	0.32 - 13.64	DT	1.784	0.306	RF	2.133	0.657
	LPS LPS	0.67 - 11.71	KNN	4.05	0.331	No oversample	-	-
IL-1β	NT	0.25 - 35.43	Non predictive	-	-	RF	3.814	0.400
	Pr LPS	0.90 - 34.61	Ridge Regression	3.289	0.345	RF	4.197	0.451
	Pr Ca	0.42 - 24.12	Non predictive	-	-	RF	2.286	0.474
IL-1RA	NT	1095.66 - 6293.75	XGB	1075.939	0.313	No oversample	-	-
	Pr LPS	2013.46 - 6365.57	XGB	488.807	0.573	No oversample	-	-
	Pr Ca	1066.64 - 5947.04	Non predictive	-	-	No oversample	-	-
	Ca LPS	984.26 - 5406.42	Non predictive	-	-	No oversample	-	-
	LPS LPS	987.22 - 4996.25	Linear Regression	777.045	0.376	No oversample	-	-

Table 6.2 displays the results for the top-scoring classifiers when values for each cytokine were converted to categorical labels, as explained in Section 5.3.2 of Chapter 5, with and without the application of SMOTE for oversampling the data, focusing as well on the optimal algorithm for the task and presenting accuracy and general F1 score. Not all variables were suitable for classification, as explained in the previous chapter, due to complete lack or insufficient representation within certain classes. Among the variables that were classifiable, only IL-10 Pr LPS, IL-10 LPS LPS, and IL-1RA Ca LPS reached an accuracy of 50% or greater without oversampling. Upon applying SMOTE algorithm for oversampling, several additional variables achieved or surpassed the 50% accuracy threshold, including IL-6 in Pr LPS and LPS LPS conditions, IL-10 Pr LPS and LPS LPS, IL-1β NT, and IL-1RA in conditions NT, Pr LPS, Pr Ca, and Ca LPS. Among these, a model for IL-1β NT yielded the highest performance, achieving an accuracy of

65% and an F1 score of 61%.

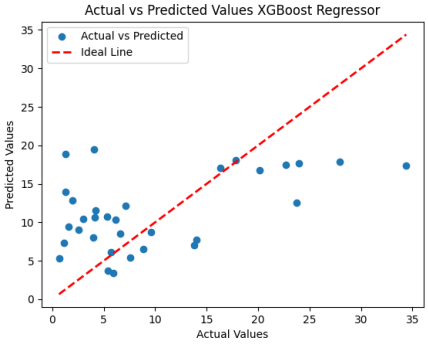
Table 6.2: Table of best results for predicting cytokine levels - classification.

		Only original data			With oversampled data		
Protein	**Treatment**	**Best Algorithm**	**Accuracy**	**F1 Score**	**Best Algorithm**	**Accuracy**	**F1 Score**
TNFα	NT	Non classifiable					
	Pr LPS	DT	42.31%	42.94%	KNN	42.3%	37.0%
	Pr Ca	Non classifiable					
	Ca LPS	Non classifiable					
	LPS LPS	DT	48.39%	48.82%	DT	48.4%	48.4%
IL-6	NT	Non classifiable					
	Pr LPS	KNN	47.37%	45.25%	KNN	57.9%	47.2%
	Pr Ca	Non classifiable					
	Ca LPS	DT	38.46%	30.00%	SVC	46.2%	52.1%
	LPS LPS	Logistic Ridge	46.15%	47.88%	Logistic Regression	50.0%	51.1%
IL-10	NT	Non classifiable					
	Pr LPS	DT	50.00%	49.46%	RF	56.3%	56.7%
	Pr Ca	Non classifiable					
	Ca LPS	Non classifiable					
	LPS LPS	Logistic Regression	52.94%	53.33%	SVC	58.8%	58.6%
IL-1β	NT	DT	48.28%	34.66%	RF	65.5%	61.0%
	Pr LPS	Logistic Lasso	34.48%	31.04%	RF	44.8%	44.2%
	Pr Ca	Non classifiable					
IL-1RA	NT	DT	42.31%	36.85%	RF	53.8%	54.1%
	Pr LPS	DT	46.15%	43.21%	SVC	50.0%	50.6%
	Pr Ca	Logistic Regression	30.77%	29.63%	SVC	50.0%	50.7%
	Ca LPS	SVC	50.00%	50.59%	SVC	57.1%	53.3%
	LPS LPS	DT	44.44%	33.10%	Logistic Ridge	44.4%	43.9%

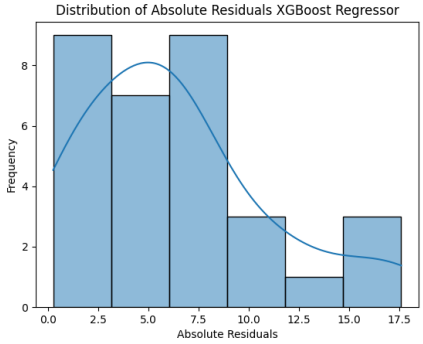
Predicting TNFα

The deployed regression algorithms were not able to reach performance levels that would allow the models to be considered reliably predictive using only the original data for TNFα in NT conditions, (with the best-performing algorithm yielding an R value of 0.169 between real and predicted values). However, with oversampled data, the XGB algorithm achieved a significantly higher R value of 0.509, indicating a more accurate prediction, and a MdAE of 5.792. Figure 6.13a presents the scatter plot for the actual versus predicted values by the model and Figure 6.13b the model's distribution of residues. These together, indicate that the model predictions are generally close to the actual values for many observations, nevertheless there are some instances where the predictions deviate more significantly, notably for higher values of this variable.

For the Pr LPS condition, SVR performed best on the original data, yielding an R value of 0.49 and MdAE of 354.53 (as this variable presents a wide range of values, varying between 1.75 and 2705.91). Notwithstanding, the KNN algorithm outperformed it



(a) Scatter plot of actual vs predicted values



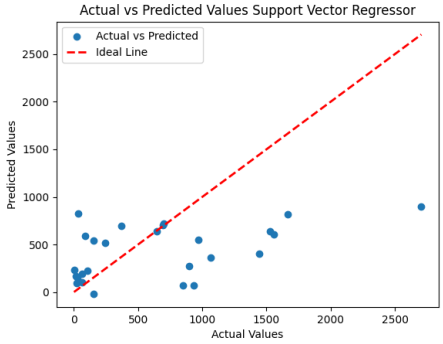
(b) Distribution of residues

Figure 6.13: KNN model LOOCV results for predicting TNFα NT with oversampled data.

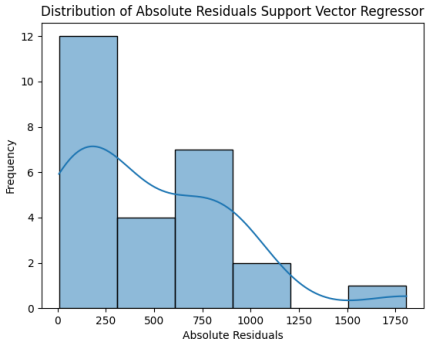
when using oversampled data, resulting in a higher correlation of 0.593 but with an increased MdAE of 381.148. As can be observed in Figure 6.14, the SVR model performs well on lower values of the target variable, which significantly improves in the XGB model when data is oversampled, indicating this model captures best the variation in the data, presenting also a smoother distribution of residues.

Interestingly, for the Pr Ca, Ca LPS and LPS LPS conditions, the regression models were non-predictive, with R values lower than 0.3.

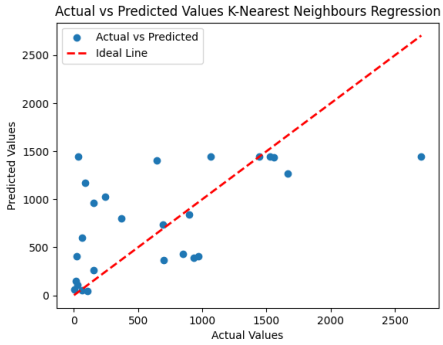
In classification, only TNFα Pr LPS and TNFα LPS LPS were classifiable, with the best classifier for the first condition achieving an accuracy of only 42% even when data was oversampled and the top-performing algorithm for the LPS LPS condition achieving only around 48% accuracy, even when trained with oversampled data.



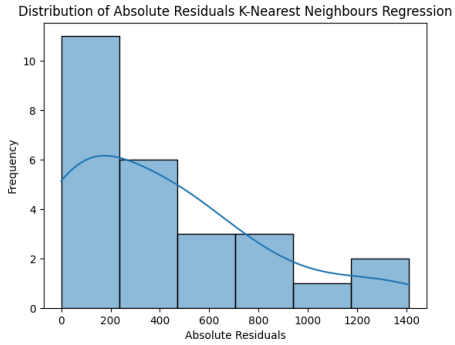
(a) SVR: Actual vs predicted values scatter plot



(b) SVR: Distribution of residues



(c) KNN (oversampled): Actual vs predicted values scatter plot



(d) KNN (oversampled): Distribution of residues

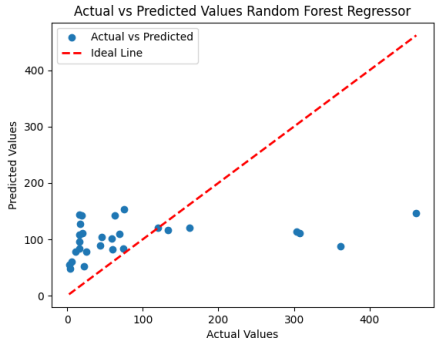
Figure 6.14: LOOCV results for predicting TNFα Pr LPS: (a-b) SVR model without oversampling; (c-d) KNN model with oversampling.

Predicting IL-6

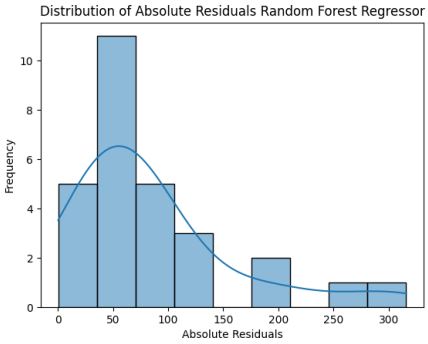
For IL-6, the original data yielded non-predictive regression models for the NT, Pr Ca and Ca LPS conditions. However, when data was oversampled the RF model for IL-6 NT yielded an R value of 0.32 and a MdAE of 61.976. Figure 6.15a illustrates this moderate correlation between the real values and the model's outputs, which once again shows more significant deviation (higher residues displayed in Figure 6.15b) for the few instances corresponding to higher values of the cytokine in this condition.

For the Pr LPS condition, RF also performed best in regression, with an R value of 0.316, which improved to 0.436 with oversampling. The scatter plots presented in Figures 6.16a and 6.16c demonstrate the mild improvement in the second model as the points in this plot are closer to the ideal line (represented in red) and as the distribution of residues in Figure 6.16d for the oversampled data model displays overall a lower range of absolute residues, indicating that the model yields improved accuracy and tighter predictions compared to the non-oversampled data model.

Finally, regarding IL-6, the performance in predicting the LPS LPS condition also saw

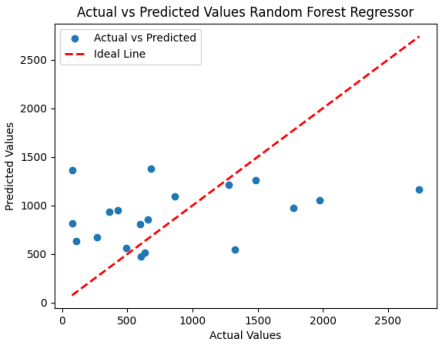


(a) Scatter plot of actual vs predicted values

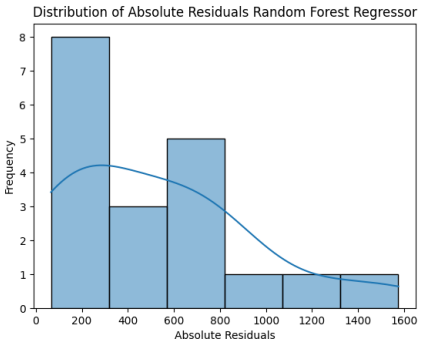


(b) Distribution of residues

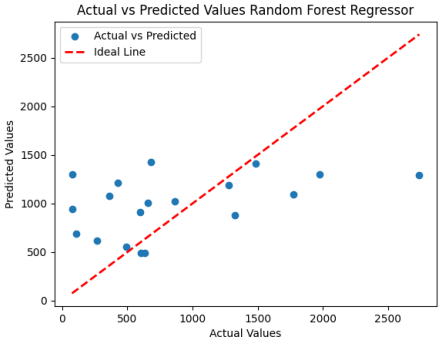
Figure 6.15: RF model LOOCV results for predicting IL-6 NT with oversampled data.



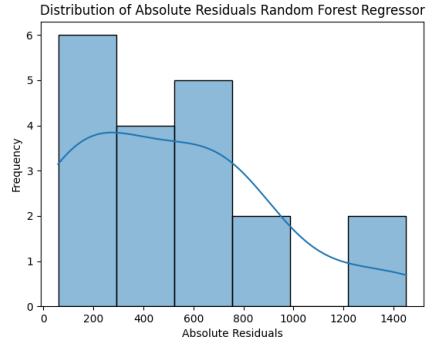
(a) RF: Actual vs predicted values scatter plot



(b) RF: Distribution of residues



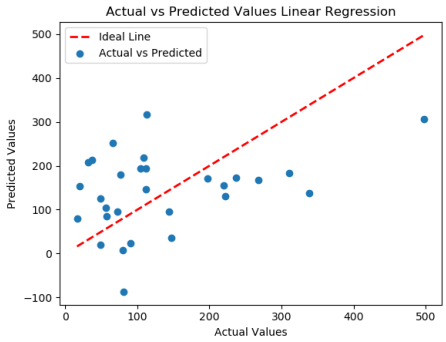
(c) RF (oversampled): Actual vs predicted values scatter plot



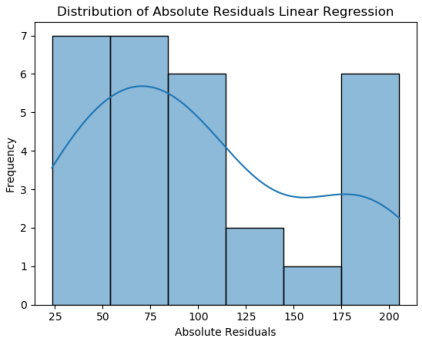
(d) RF (oversampled): Distribution of residues

Figure 6.16: LOOCV results for predicting IL-6 Pr LPS: (a-b) RF model without oversampling; (c-d) RF model with oversampling.

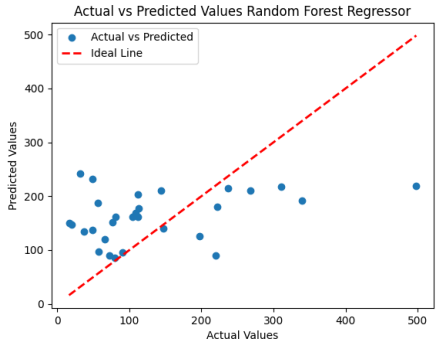
improvement with oversampling, with RF regressor outperforming linear regression by reducing the MdAE (from 88.88 to 71.46) and increasing the correlation coefficient (from 0.338 to 0.365). Even so, as the improvement is only mild the scatter plots for these models are rather similar as can be observed in Figures 6.17a and 6.17c, however major differences can be visualized in the plots for the distribution of residues, as the one for the RF model (Figure 6.17d) shows a higher concentration of lower residuals and fewer large errors, despite having a wider overall range of residuals when compared to the linear model (Figure 6.17b).



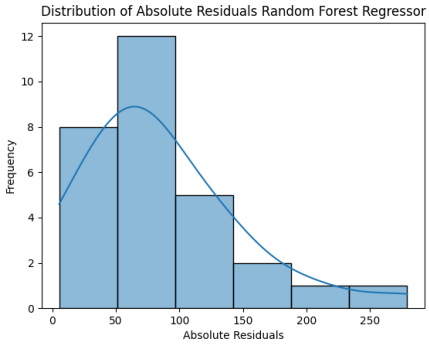
(a) Linear Regression: Actual vs predicted values scatter plot



(b) Linear Regression: Distribution of Residues



(c) RF (oversampled): Actual vs predicted values scatter plot



(d) RF (oversampled): Distribution of Residues

Figure 6.17: LOOCV results for predicting IL-6 LPS LPS: (a-b) Linear Regression model; (c-d) Random Forest model with oversampled data.

With regards to classification results for this cytokine, neither IL-6 NT nor IL-6 Pr Ca enabled classification. In the case of IL-6 Pr LPS, KNN was the best classifier, achieving 47.4% accuracy and an F1 score of 45.3% on the original data. With oversampled data, KNN improved, being again the best-performing algorithm and reaching an accuracy of 57.9%, although the F1 score only modestly improved to 47.2%. This performance is illustrated in the confusion matrix of the model in Figure 6.18 as this was the best performance achieved for this protein. However, as evidenced by the absence of instances in the central vertical column of the matrix, this model fails to

capture and predict the instances representing medium levels of this variable.

For IL-6 Ca LPS, the best results were yielded by the DT algorithm, which, even so, only achieved an accuracy of 38.5% and an F1 score of 30% on original data. Oversampling helped improve performance, with SVC reaching an accuracy of 46.2% and an F1 score of 52.1%.

Lastly, in LPS LPS conditions, logistic Ridge regression and logistic regression both performed similarly with original and oversampled data, attaining accuracies of 46.2% and 50.0%, and F1 scores of 47.9% and 51.1%, respectively.

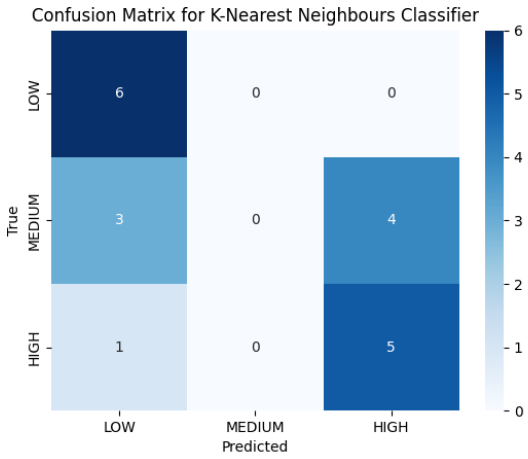
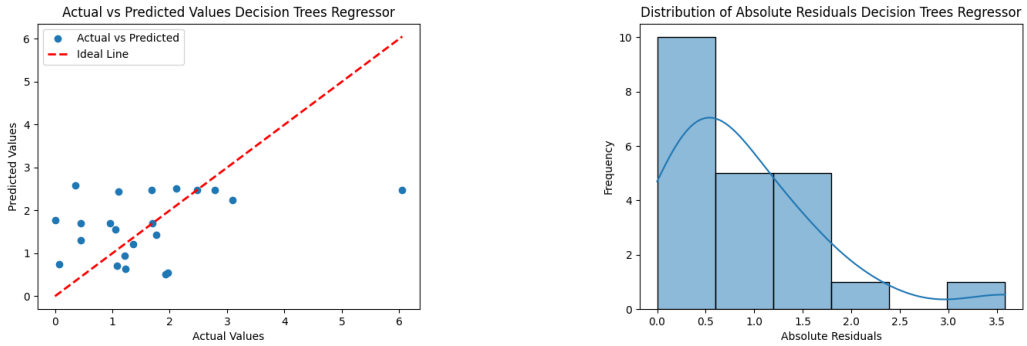


Figure 6.18: Confusion matrix for KNN model predicting IL-6 Pr LPS with oversampled data.

Predicting IL-10

In the case of IL-10, conditions where regression models did not turn out predictive included IL-10 NT and Pr LPS. On the other hand, the Pr Ca condition showed notable improvement when data was augmented, with the DT algorithm achieving an R value of 0.359 and a MdAE of 0.701. The scatter plot for actual versus predicted values in Figure 6.19a displays data points fairly closely aligned with the ideal line, indicating a general good fit of the model. Nonetheless, the presence of one significant outlier suggests that the model struggles to accurately predict certain cases, which is reflected in the higher absolute residue value depicted in Figure 6.19b.

Concerning the Ca LPS condition, the DT regressor algorithm performed best on original data with an R value of 0.306. This mild correlation between the model's output and the real values is patent in Figure 6.20a where data points are rather scattered across the plot and not so distributed along the ideal line, in conformity with the somewhat increased proportion of higher residue values displayed in Figure 6.20b. Howbeit, when data was oversampled, RF emerged as the best algorithm with a significant increase in correlation, in fact achieving the best results overall, with an R value of 0.657, despite the slight increase in MdAE compared to the previous model (from 1.784



(a) Scatter plot of actual vs predicted values

(b) Distribution of residues

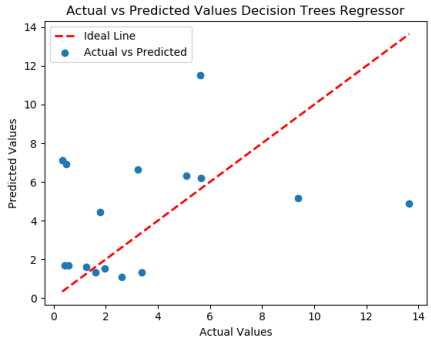
Figure 6.19: DT model LOOCV results for predicting IL-10 Pr Ca with oversampled data.

to 2.133). Figure 6.20c displays a scatter plot with data points much more aligned with the ideal line and in Figure 6.20d the distribution of residues presents a lower frequency for larger residues.

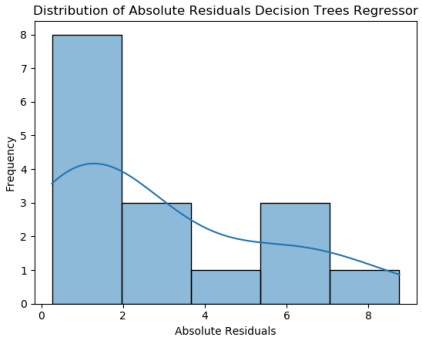
For IL-10 LPS LPS, although it was not possible to oversample the data via SMOGN algorithm, the KNN model still reached an R value above the threshold of 0.3 solely on original data, and a MdAE of 4.05. The scatter plot for the model is displayed in Figure 6.21a, with mid-range values being better captured by the model, and Figure 6.21b displays the model's residue distribution, which is rather spread out, with a significant number of residuals in the higher range and the frequencies being nearly evenly distributed across the residual range, with several peaks.

In respects to classification results, only Pr LPS and LPS LPS conditions showed sufficient label variation (significant instances of each target class) in order to train predictive models. In the case of IL-6 Pr LPS, the DT algorithm gave an accuracy of 50.00% and an F1 score of 49.46% on original data. After oversampling, RF improved these metrics, achieving 56.3% accuracy and an F1 score of 56.7%, this performance is shown in Figure 6.22. The analysis of this confusion matrix of the model reveals the classifier's limitation in its ability to discern between high and medium instances.

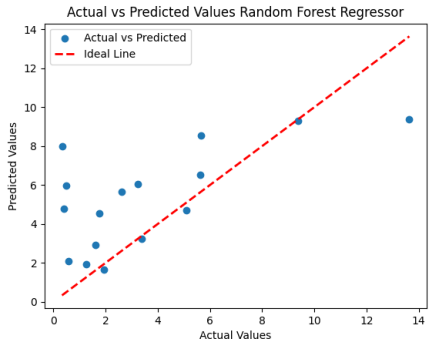
As for IL-10 with LPS LPS treatment, the logistic regression model achieved 52.94% accuracy and an F1 score of 53.33% on original data, while after oversampling, SVC performed better with 58.8% accuracy and an F1 score of 58.6% (see Figure 6.23).



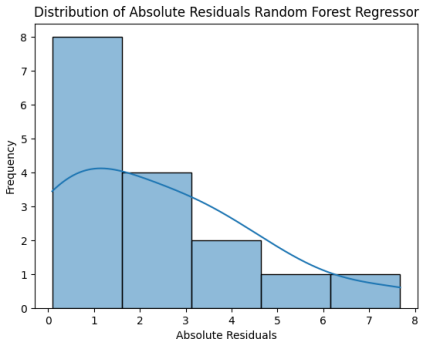
(a) DT: Actual vs predicted values scatter plot



(b) DT: Distribution of Residues

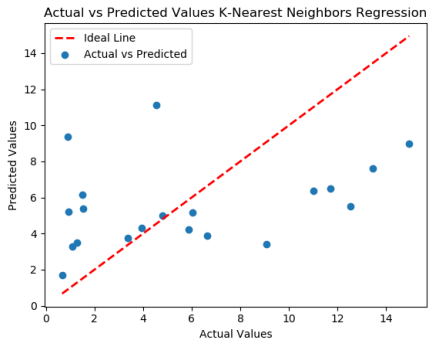


(c) RF (oversampled): Actual vs predicted values scatter plot

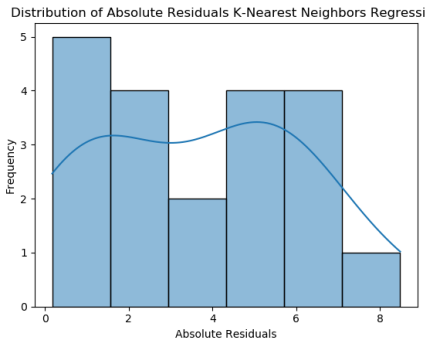


(d) RF (oversampled): Distribution of Residues

Figure 6.20: LOOCV results for predicting IL-10 Ca LPS: (a-b) Decision Trees model; (c-d) Random Forest model with oversampling.



(a) Scatter plot of actual vs predicted values



(b) Distribution of residues

Figure 6.21: KNN model LOOCV results for predicting IL-10 LPS LPS.

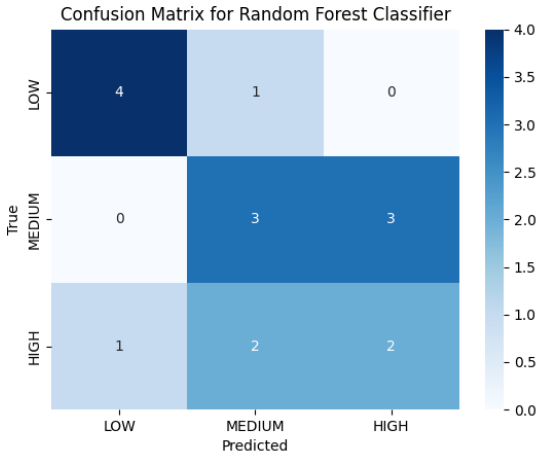


Figure 6.22: Confusion matrix for RF model predicting IL-10 Pr LPS levels with oversampled data.

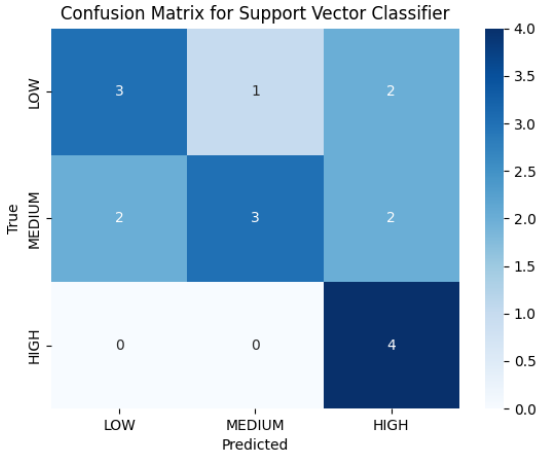


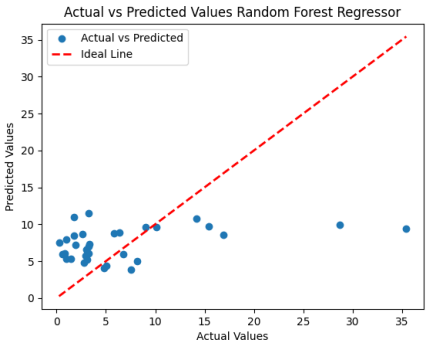
Figure 6.23: Confusion matrix for SVC model predicting IL-10 LPS LPS levels with oversampled data.

Predicting IL-1β

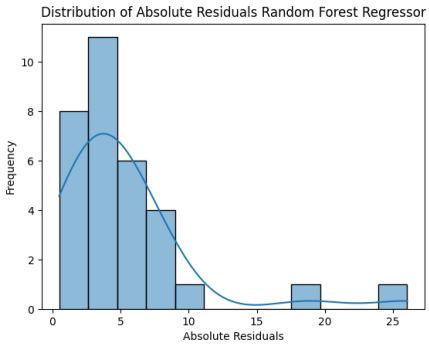
IL-1β predictions were most successful for the Pr LPS condition when data was not oversampled, however all three conditions considered for IL-1β yielded an R value equal or superior to 0.4 when oversampling was deployed.

IL-1β NT was most successfully modelled by RF when data was oversampled, with an R value of 0.4 and MdAE of 3.814. Nonetheless, the scatter plot of Figure 6.24a shows the model again does not fit well to the instances that display significantly higher values and in this way the distribution of residues in Figure 6.24b presents a few larger residuals, extending the tail of the distribution.

Without oversampling, IL-1β Pr LPS was most successfully modelled by Ridge regression, with an R value of 0.345 and MdAE of 3.29. Application of SMOGN once again improved performance with the correlation coefficient of a RF model increasing



(a) Scatter plot of actual vs predicted values



(b) Distribution of residues

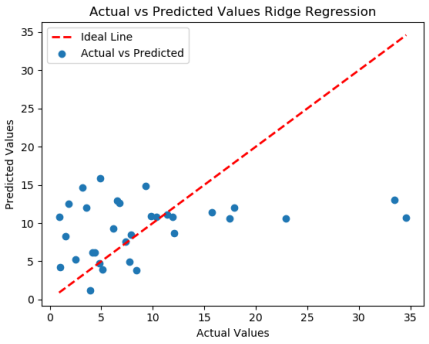
Figure 6.24: RF model LOOCV results for predicting IL-1β NT with oversampled data.

to 0.451. Comparing these metrics and the scatter plots depicted on Figure 6.25a for the Ridge regressor and Figure 6.25c for the RF model with oversampled data, it can be assessed that the latter is better suited to handle the variability of this variable. Respective residue distributions (Figures 6.25b and 6.25d) also highlight the lower frequency of residuals in the second model as well as overall smaller range for the errors.

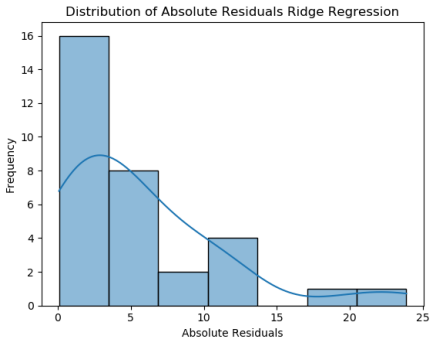
IL-1β Pr Ca displayed mildly predictive results when data was oversampled, as a RF model yielded an R value of 0.474 and MdAE of 2.286. This model exhibits a moderate level of accuracy, as illustrated by the scatter plot in Figure 6.26a, however with the presence of outliers, particularly one data point far from the ideal line. The distribution of residuals in Figure 6.26b also shows that while many predictions are rather close to the actual values, there are some notable errors. In this way, the model captures the variability of a major part of the dataset but is off in certain cases, particularly when the variable presents significantly higher values than typically observed.

Concerning these variables' conversion to categorical values, classification was possible for NT and Pr LPS conditions. Results for IL-1β NT show DT yielded 48.28% accuracy and a relatively low F1 score of 34.66%. After oversampling via SMOTE, RF improved performance to 65.5% accuracy and an F1 score of 61.0%, representing a substantial improvement, and the highest performance among the cytokines in the dataset. The model's performance is depicted in Figure 6.27.

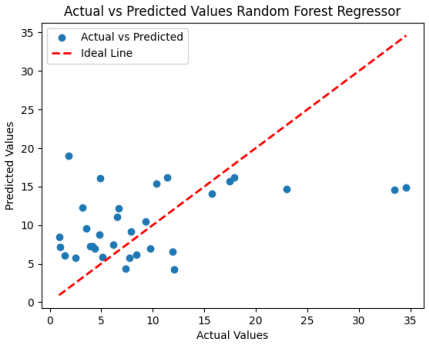
In respects to IL-1β Pr LPS, logistic Lasso regression only achieved 34.48% accuracy and an F1 score of 31.04% on original data. After oversampling, the best-performing algorithm, RF, improved performance to 44.8% accuracy and an F1 score of 44.2%.



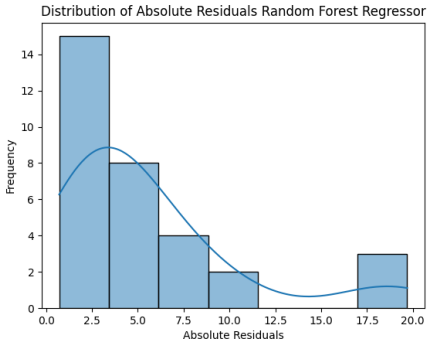
(a) Ridge Regression: Actual vs predicted values scatter plot



(b) Ridge Regression: Distribution of Residues

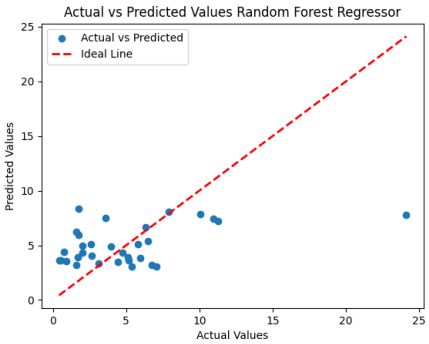


(c) RF (oversampled): Actual vs predicted values scatter plot

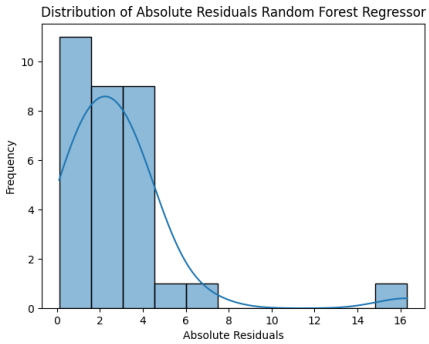


(d) RF (oversampled): Distribution of Residues

Figure 6.25: LOOCV results for predicting IL-1β Pr LPS: (a-b) Ridge Regression model; (c-d) Random Forest model with oversampling.



(a) Scatter plot of actual vs predicted values



(b) Distribution of residues

Figure 6.26: RF model LOOCV results for predicting IL-1β Pr Ca with oversampled data.

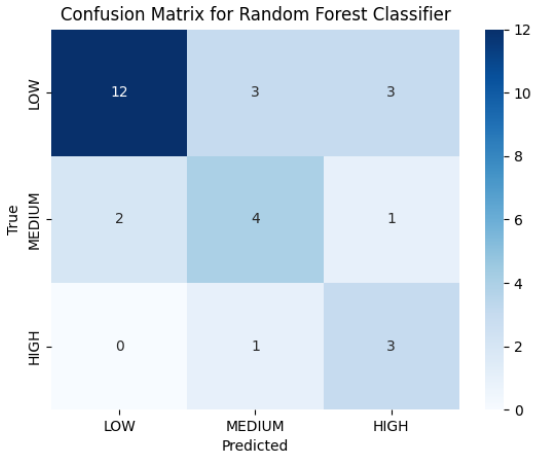
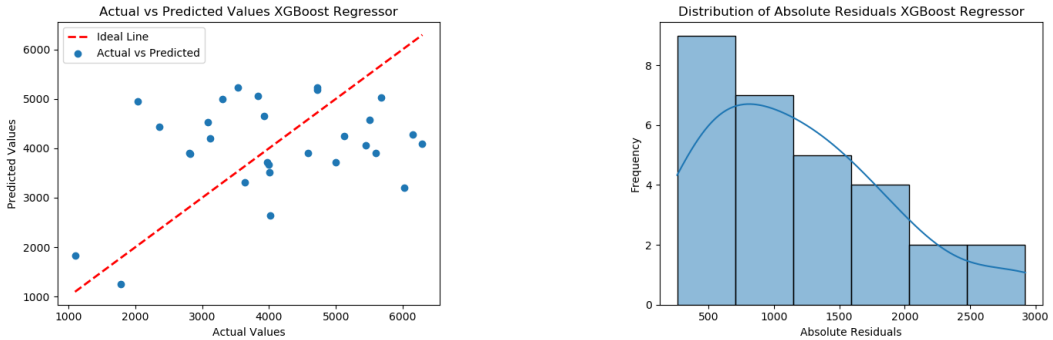


Figure 6.27: Confusion matrix for RF model predicting IL-1β NT levels with oversampled data.

Predicting IL-1RA

For IL-1RA regression results, SMOGN algorithm did not allow for oversampling of the data. Even so, moderately predictive models were still found for NT, Pr LPS and LPS LPS conditions.

XGB was the best performer for the NT condition with an R value of 0.313 and MdAE of 1075.939 (having in account that the values for these variable range between 1095.66 and 6293.75). This model captures the general trend of the data but exhibits variability, as indicated by the scatter of the data points in Figure 6.28a and the distribution of residuals in Figure 6.28b which shows that larger prediction errors are less common, however displaying the presence of larger residuals (of over 2000).



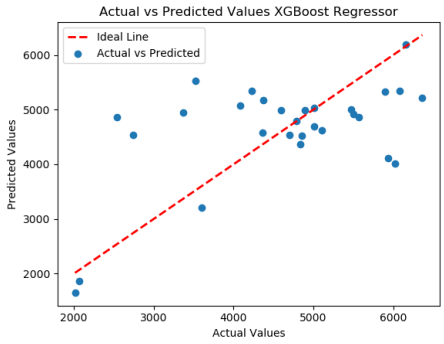
(a) Scatter plot of actual vs predicted values

(b) Distribution of residues

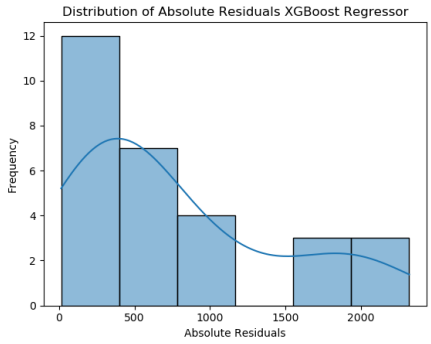
Figure 6.28: XGB model LOOCV results for predicting IL-1RA NT.

XGB was once again the best-performing algorithm for the prediction of IL-1RA Pr LPS levels, with the highest R value on original data of 0.573 and MdAE of 488.81. Once again, the plot of actual versus predicted values displayed in Figure 6.29a shows the model captures the general trend in the data, without significant outliers, however data

points are still considerably scattered and the distribution of residues (Figure 6.29b) displays a substantial number of larger residuals despite the majority of residuals being clustered around the range of 0 to 500.



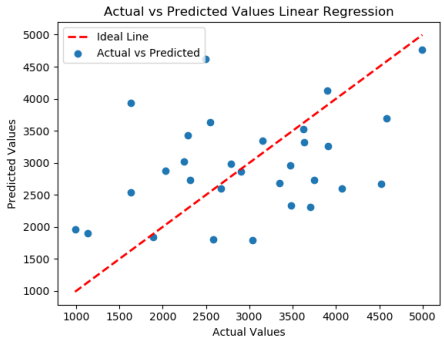
(a) Scatter plot of actual vs predicted values



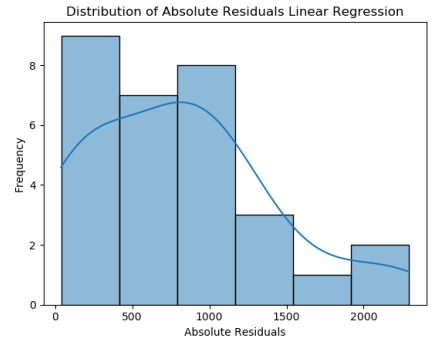
(b) Distribution of residues

Figure 6.29: XGB model LOOCV results for predicting IL-1RA Pr LPS.

Finally, IL-1RA LPS LPS was best modelled by simple linear regression, which achieved a MdAE of 777.05 and R value of 0.376. Again this model shows modest performance. The data points in the plot of Figure 6.30a spread around the ideal line, however with significant scatter and no visible outliers. The relatively broad spread of the residuals and relatively high frequency of moderate to substantial prediction errors depicted in Figure 6.30b demonstrate the model's difficulty in capturing more nuanced variations in the data.



(a) Scatter plot of actual vs predicted values



(b) Distribution of residues

Figure 6.30: Linear regression model LOOCV results for predicting IL-1RA LPS LPS.

In the matter of classification results for this cytokine, models were built for each condition and nearly every model substantially improved with the deployment of SMOTE. For IL-1RA NT, the DT model yielded 42.31% accuracy and an F1 score of 36.85% on original data, but oversampling improved these metrics to 53.8% accuracy and F1 score of 54.1% with RF.

For Pr LPS condition, again DT was the best algorithm when trained only on original data, achieving 46.15% accuracy and an F1 score of 43.21%. After oversampling, SVC

slightly improved the best performance achieved, resulting in 50.0% accuracy and a 50.6% F1 score.

In Pr Ca condition, models for IL-1RA, were not very successful when trained solely on original data with logistic regression achieving a top performance of just 30.77% accuracy and 29.63% F1 score. Oversampling improved performance, enabling SVC to reach an accuracy of 50.0% and an F1 score of 50.7%.

The best performing model on both original and oversampled data for IL-1RA Ca LPS, was SVC with an accuracy of 50.00% and an F1 score of 50.59% in the first case, and improvement to 57.1% accuracy and 53.3% F1 score in the latter. This was the best result achieved for this cytokine, and the confusion matrix for the oversampled model is presented in Figure 6.31. The analysis of the matrix reveals that the classifier's primary limitation is its difficulty in accurately predicting instances categorized as "high".

To conclude, IL-1RA LPS LPS was one of the few cases where oversampling did not significantly improve the classifier's performance. When training only with original data, the DT algorithm performed with 44.4% accuracy and an F1 score of 33.1%. Following oversampling of the train set, the best metrics were achieved by logistic Ridge regression, with an accuracy of 44.4% and a mildly improved F1 score of 43.9%

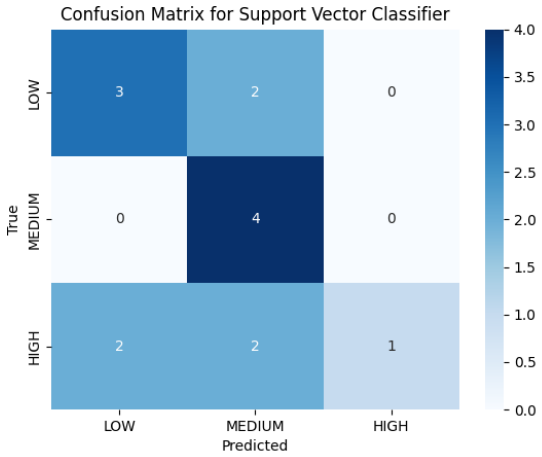


Figure 6.31: Confusion matrix for SVC model predicting IL-1RA Ca LPS levels with oversampled data.

6.3 Predicting Alzheimer's Disease

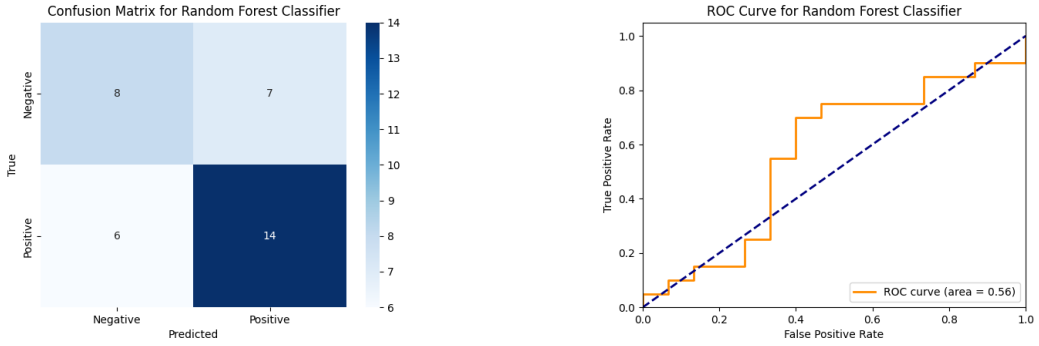
The purpose of this portion of the work is to assess AD predictability based on different features of our dataset and thus exploring different relationships within the data and the disease. When sample size allows it, final models have also been developed and interpreted and, thus, SHAP plots are presented.

6.3.1 Predicting AD from Infectious Burden

Firstly, AD was predicted based on IB data, with measures of subjects' antibody levels for HSV-1/2, *Helicobacter pylori*, CMV, *Chlamydia pneumoniae* and *Borrelia burgdorferi*. As for this task the sample size is only of 35, LOOCV results are presented. As displayed on Table 6.3 the best model achieved was RF without oversampling. This model achieved an accuracy of around 63% and a general F1 score of around 62%. As illustrated in the confusion matrix of the model (Figure 6.32a), the model was particularly successful in identifying AD cases, successfully classifying 14 out of the 20 cases, however with a lower recall for the negative class, only correctly identifying 8 out of the 15 HCs. The AUC achieved, as shown in the ROC curve plot of Figure 6.32b was of 0.56.

Table 6.3: Performance metrics of top-scoring predictive model (RF) for AD using IB data.

RF	**Precision**	**Recall**	**F1 Score**
General	0.619	0.617	0.617
AD	0.667	0.700	0.683
Accuracy		0.629	
AUC		0.560	



(a) (b)

Figure 6.32: RF model LOOCV results for predicting AD from IB data: (a) Confusion matrix; (b) ROC curve.

6.3.2 Predicting AD from Trained Immunity Cytokines

For this task, models predicting AD were trained on TI data. Final models were deployed and performance assessed on a test set of 8 samples. Table 6.4 displays performance metrics for the two best models achieved, median-filled logistic regression with EN penalty and mean-filled logistic regression with L1 penalty and oversampling enabled. Both models achieve similar results with identical confusion matrices (Figures 6.33a and 6.34a), an accuracy of 87.5% in the test set and recall of 100% for AD class. Both models reach moderately high AUCs (Figures 6.33b and 6.34b) of 0.88 and 0.81 respectively.

Table 6.4: Performance metrics of top-scoring predictive models for AD Using TI cytokine data.

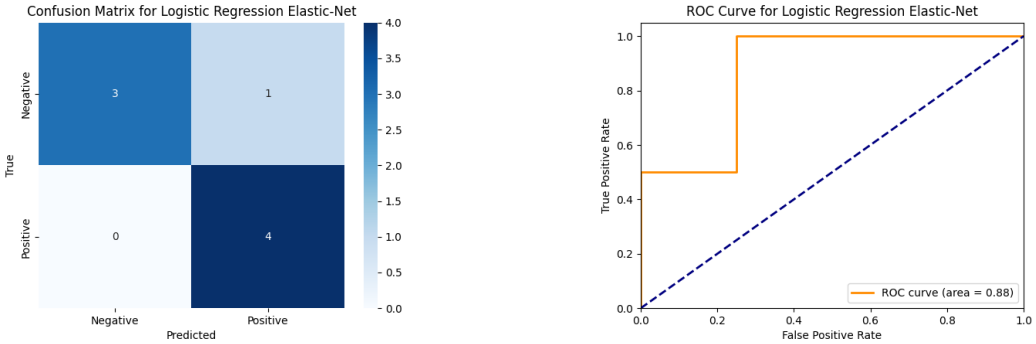
Logistic EN	**Precision**	**Recall**	**F1 Score**
General	0.900	0.875	0.873
AD	0.800	1.00	0.889
Accuracy		0.875	
AUC		0.88	
Logistic Lasso	**Precision**	**Recall**	**F1 Score**
General	0.900	0.875	0.873
AD	0.800	1.00	0.889
Accuracy		0.875	
AUC		0.81	

Having established the performance metrics of the top-performing models in this task, it was essential to us to gain deeper insights into the decision-making processes of these models. To achieve this, SHAP plots were employed to interpret and explain each feature's contribution to the predictions output by the models.

The findings for median-filled logistic regression with EN penalty are presented in Figure 6.35. Figure 6.35a depicts that, by order of importance, the top contributing features for predicting AD are: IL-10 LPS LPS, IL-1β Pr LPS, TNFα Pr LPS, IL-6 Pr LPS and with less impact but still present IL-10 Ca LPS. The test and train set beeswarm plots presented in Figures 6.35b and 6.35c respectively display that higher values of IL-1β Pr LPS, TNFα Pr LPS and IL-6 Pr LPS lead to the prediction of AD, whereas the opposite happens with IL-10 LPS LPS, where lower levels of this feature seemingly lead to prediction of the disease.

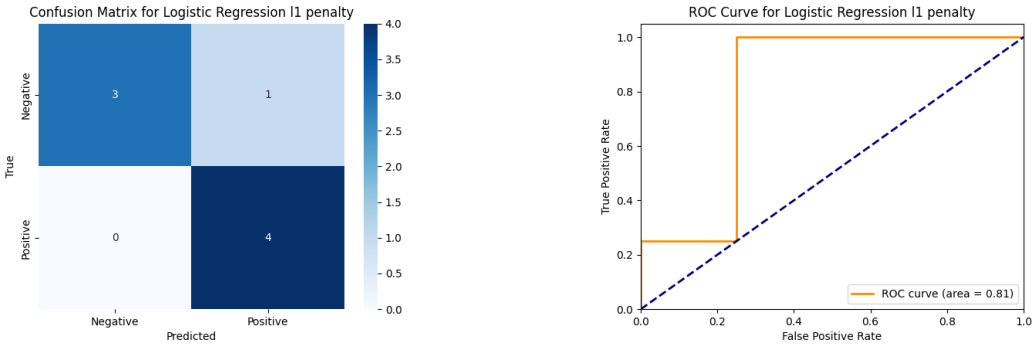
Similarly, the SHAP analysis for mean-filled logistic regression with l1 penalty and data oversampled, displayed in Figure 6.36, reveals a comparable ranking of top features influencing AD predictions. The primary difference between both models lies in the order of importance, where TNFα Pr LPS and IL-6 Pr LPS have swapped places, indicating a slight shift in their relative impact on the model's predictions, however maintaining that higher values lead to predictions of the positive class as happens with IL-1β Pr LPS. Additionally, it is noteworthy that IL-10 Ca LPS, which mildly contributed to the previous model, no longer plays a significant role in this alternative one.

In summary, both of our top-scoring models are in large agreement with each other. Features that are important for both models' prediction of AD vs HC are IL-10 in conditions primed with LPS challenged with LPS, IL-1β in Pr LPS conditions, IL-6 Pr LPS and TNFα Pr LPS. Interestingly, the top-contributing feature in both models is IL-10 LPS LPS, where lower values of the anti-inflammatory cytokine in such conditions lead to prediction of AD.



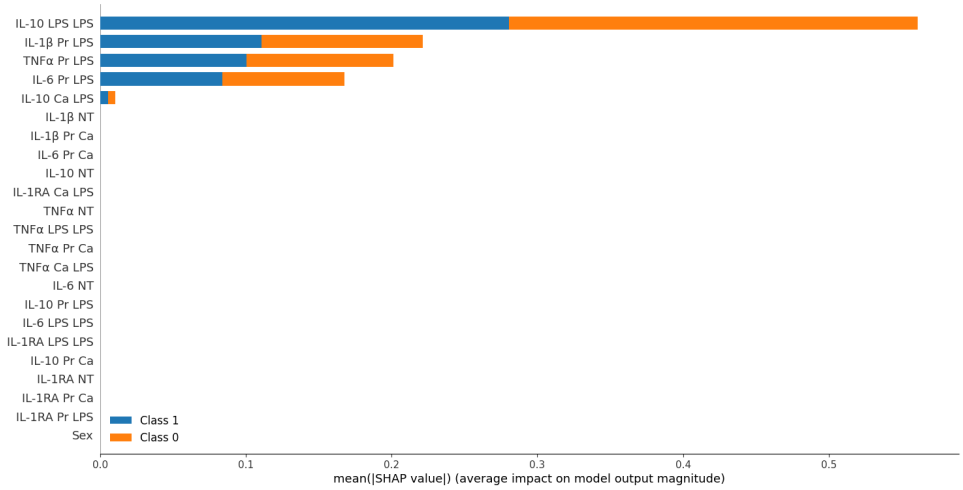
(a) (b)

Figure 6.33: Logistic EN regression median-filled model test set results for predicting AD from TI data: (a) Confusion matrix; (b) ROC curve.

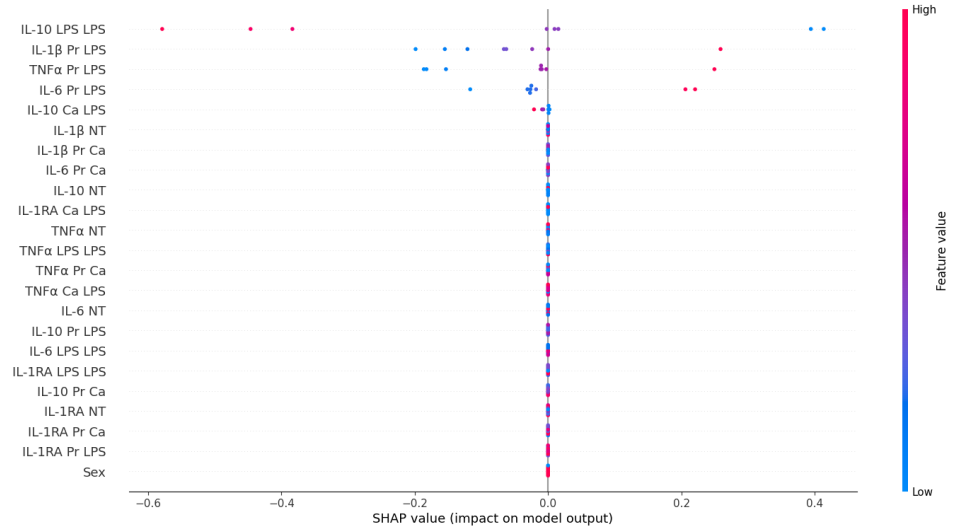


(a) (b)

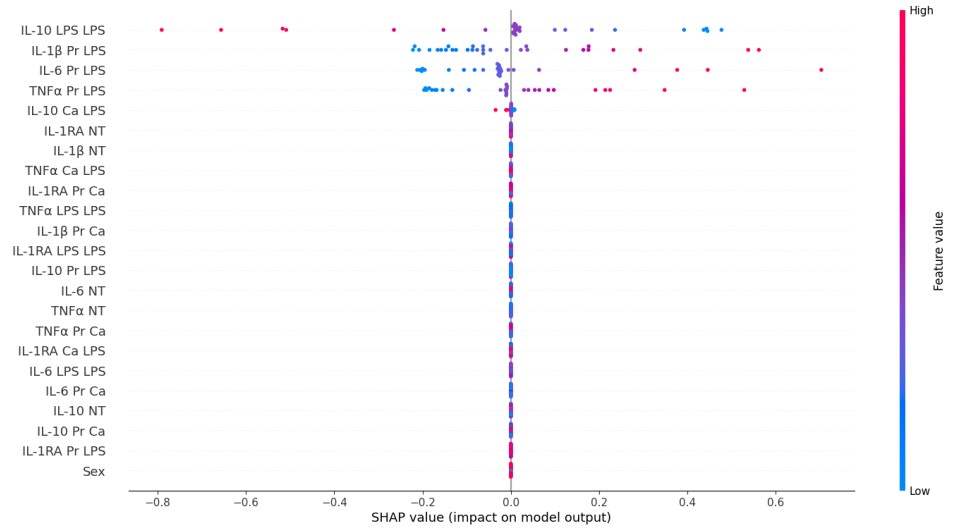
Figure 6.34: Logistic Lasso regression mean-filled model test set results for predicting AD from TI data, with oversampling enabled: (a) Confusion matrix; (b) ROC curve.



(a)

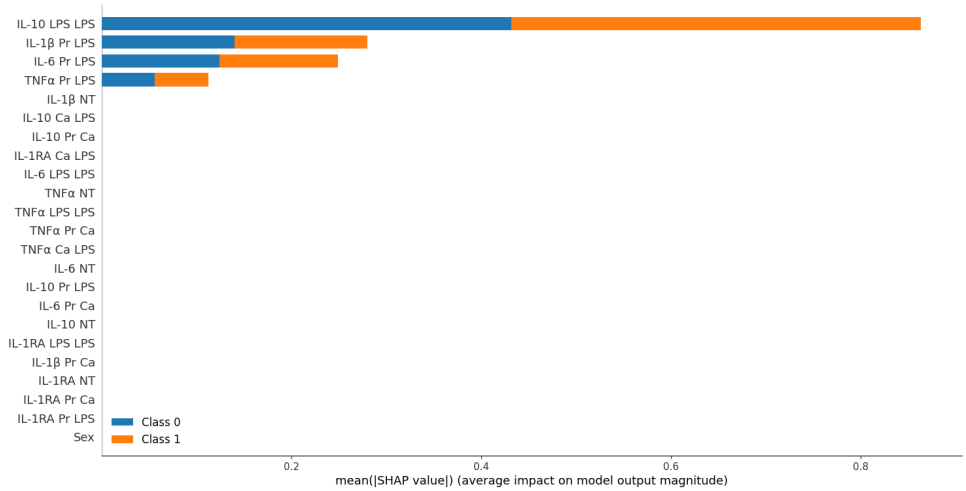


(b)

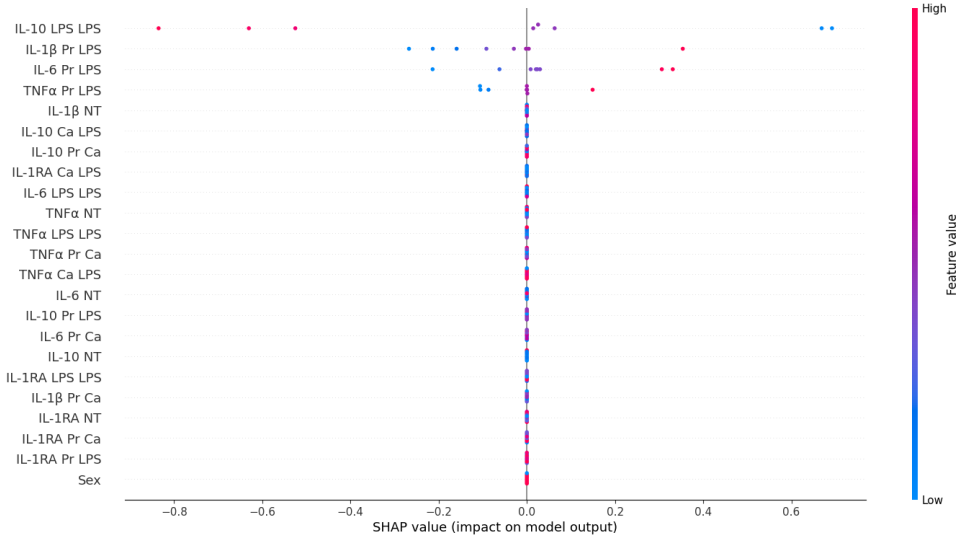


(c)

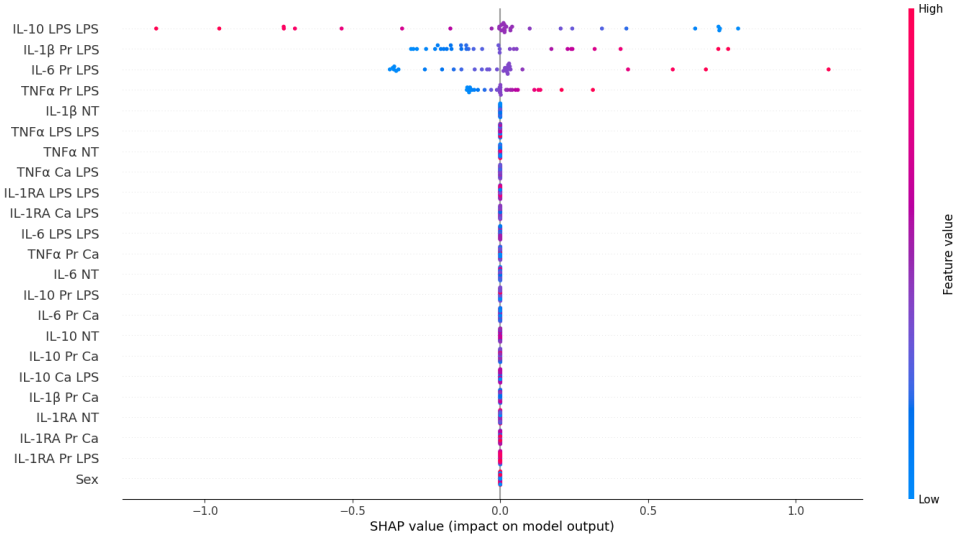
Figure 6.35: SHAP plots for logistic EN regression median-filled model predicting AD from TI data: (a) Feature importance plot; (b) Test set beeswarm plot; (c) Train set beeswarm plot.



(a)



(b)



(c)

Figure 6.36: SHAP plots for logistic Lasso regression mean-filled model predicting AD from TI data, with oversampling enabled: (a) Feature importance plot; (b) Test set beeswarm plot; (c) Train set beeswarm plot.

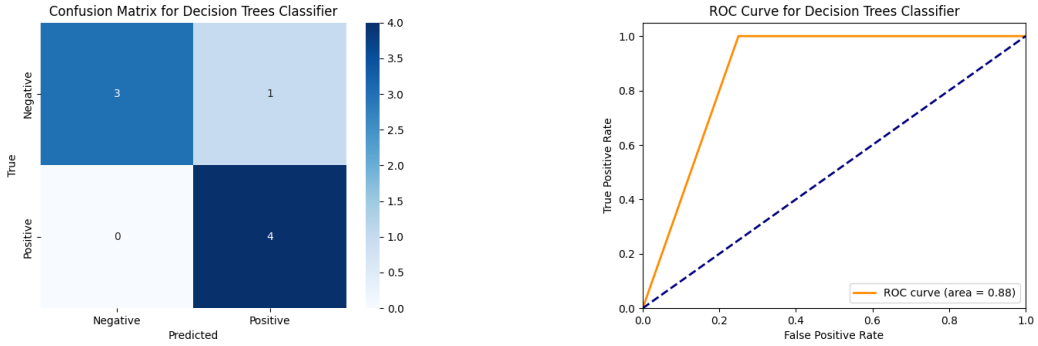
6.3.3 Predicting AD from IB and TI

This task combined the data of the two previous tasks (both IB and TI features) for modelling AD. Once more, a test set consisting of eight samples was used to evaluate the performance of the final models. The optimal model was a DT classifier obtained by utilizing this combined dataset in which missing values were imputed using the median. The performance achieved was in every way similar to when utilizing just the TI data, with 100% recall for AD class, a general 87.5% accuracy and AUC of 0.88. Table 6.5 along with Figure 6.37 provide further metrics and insights into this model.

Table 6.5: Performance metrics of top-scoring predictive model for AD using IB and TI data.

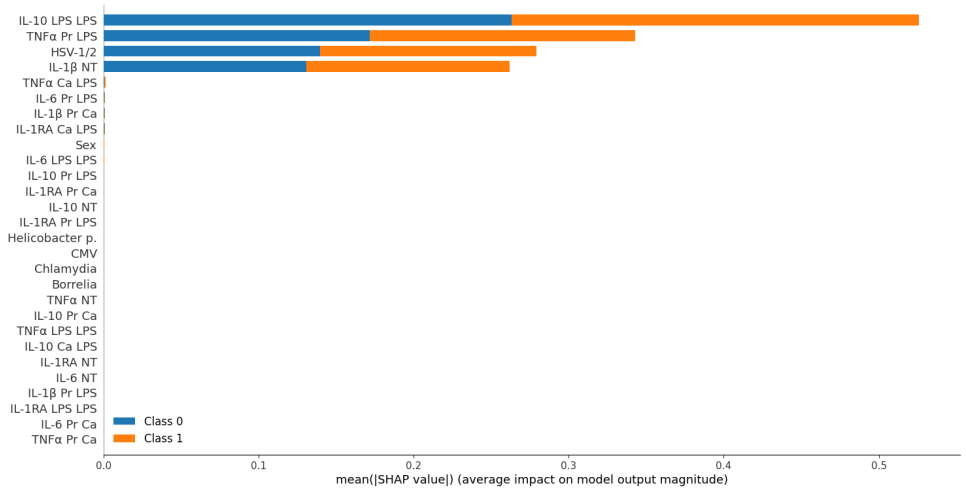
median-filled DT	**Precision**	**Recall**	**F1 Score**
General	0.900	0.875	0.873
AD	0.800	1.00	0.889
Accuracy		0.875	
AUC		0.88	

Looking into the SHAP explanations provided in Figure 6.38 for the median-filled DT model, results are very similar to the ones described in the preceding section for the models trained with solely TI data. Examining Figures 6.38b and 6.38c reveals, once again, decreased values of IL-10 LPS LPS have strong positive impacts on prediction of AD, as have increased values of TNFα Pr LPS. This model did not quite capture the predictive nature of other pro-inflammatory cytokines when primed with LPS (IL-1β Pr LPS and IL-6 Pr LPS). However, intriguingly, it picked up on the predictive power of IL-1β NT, where lower values of this cytokine appear to be linked to the prediction of a healthy subject. As this model was also trained with IB, one feature is of key importance, HSV-1/2, where increased values of this IB feature are positively linked to the prediction of AD.

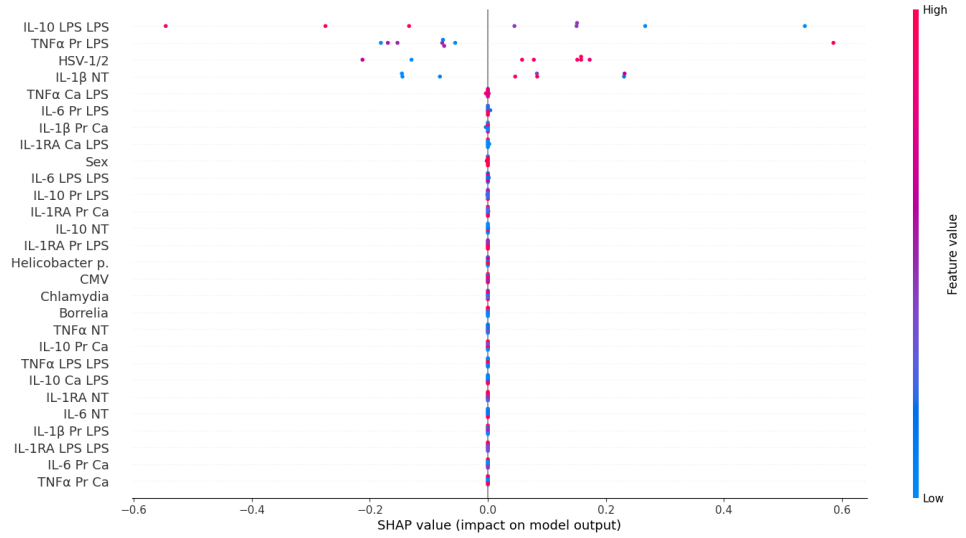


(a) (b)

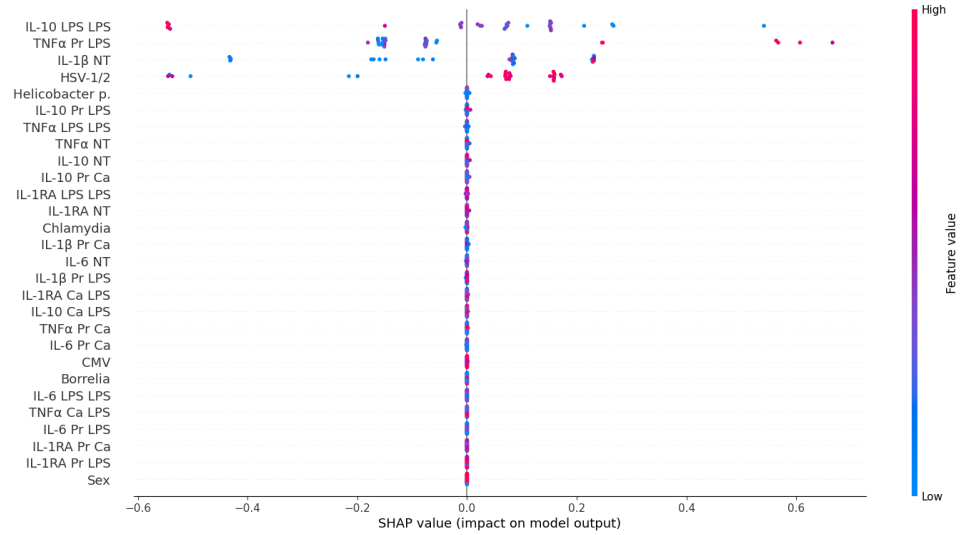
Figure 6.37: DT median-filled model test set results for predicting AD from IB and TI data: (a) Confusion matrix; (b) ROC curve.



(a)



(b)



(c)

Figure 6.38: SHAP plots for DT median-filled model predicting AD from IB and TI data: (a) Feature importance plot; (b) Test set beeswarm plot; (c) Train set beeswarm plot.

6.3.4 Predicting AD from Serum Cytokines

For the last task of predicting AD, serum circulating cytokine data was utilized, with a training set of 79 samples and a test set consisting of 10 HCs and 10 AD patients. Best models achieved were mean-filled and mode-filled SVCs with accuracies of 75% and 80% respectively. As shown in Table 6.6 and the confusion matrices of Figures 6.39a and 6.40a the models differ slightly in their successful predictions. The mode-filled model has a higher overall accuracy and successfully recalls all instances of the negative class, however only recalling 60% of the samples belonging to the AD class, assigning 4 false negatives. The mean-filled model despite having a lower accuracy, shows superior recall for class AD and misclassifies only 2 false negatives. The AUC score as shown in Figures 6.39b and 6.40b is, however, higher in the mode-filled model.

Table 6.6: Performance metrics of top-scoring predictive models for AD using serum cytokine data.

Mean-filled SVC	**Precision**	**Recall**	**F1 Score**
General	0.753	0.750	0.749
AD	0.727	0.80	0.762
Accuracy		0.75	
AUC		0.74	
Mode-filled SVC	**Precision**	**Recall**	**F1 Score**
General	0.857	0.800	0.792
AD	1.00	0.600	0.750
Accuracy		0.80	
AUC		0.86	

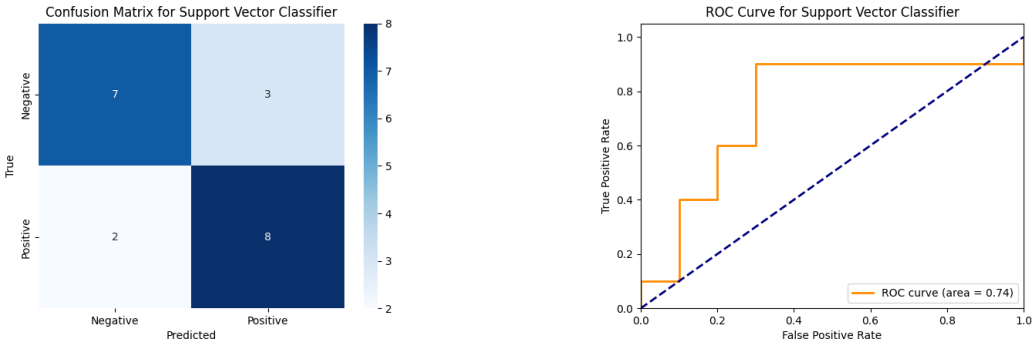
Analysing Figures 6.41 and 6.42 provides further insights into the classifiers. In both these top-scoring models, sex and IL-18 are leading features, being most apparent in the second model that being female leads to positive predictions, and in either model (but more evidently in the latter) lower values of IL-18 push the predictions towards class AD. IL-6, VEGF-A and IL-1β are of somewhat importance in both models and, once again, lower levels of these proteins seem to lead to prediction of AD.

A key difference between these models is the importance of IL-10, wherein the first model it is only of moderate relevance, it becomes a major player in the second model, being the leading feature. Only in this second model does it have a clear relationship of lower values significantly pushing towards positive predictions. This might indicate a better capture of its predictive power in the mean-filled model.

Another protein whose relevance varies between the two models is IL-8 which has a rather more important role in the mean-filled SVC. Nevertheless, both models seem to take higher values of this cytokine as a predictor for positive class.

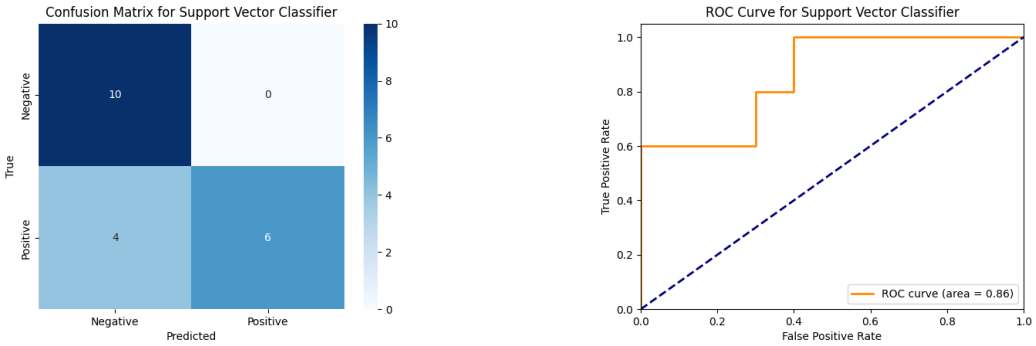
IL-1RA, of moderate importance in the two classifiers, shows contradicting impacts. In the mode-filled model, higher levels of this protein in the serum seem to be associated

with predicting AD, whereas in the mean-filled model the opposite appears to be happening.



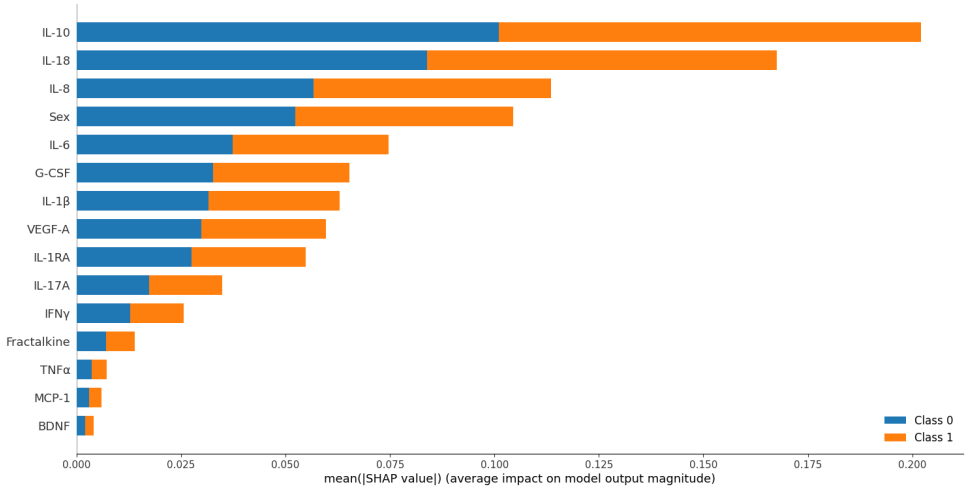
(a) (b)

Figure 6.39: SVC mean-filled model test set results for predicting AD from serum data: (a) Confusion matrix; (b) ROC curve.

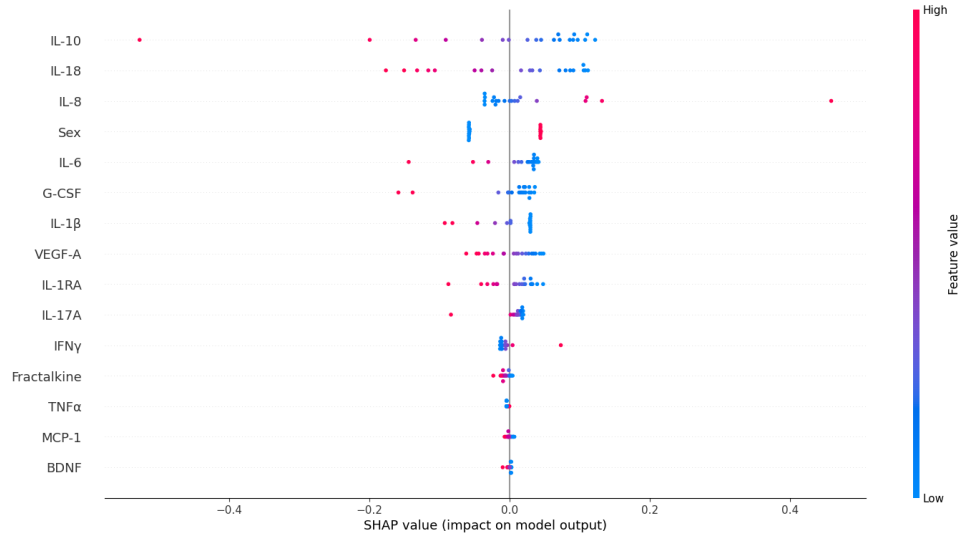


(a) (b)

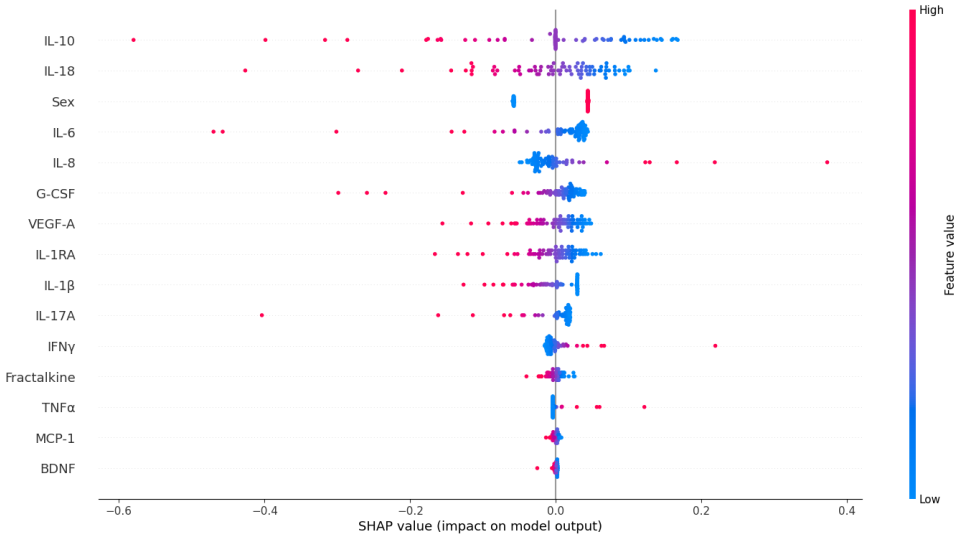
Figure 6.40: SVC mode-filled model test set results for predicting AD from serum data: (a) Confusion matrix; (b) ROC curve.



(a)

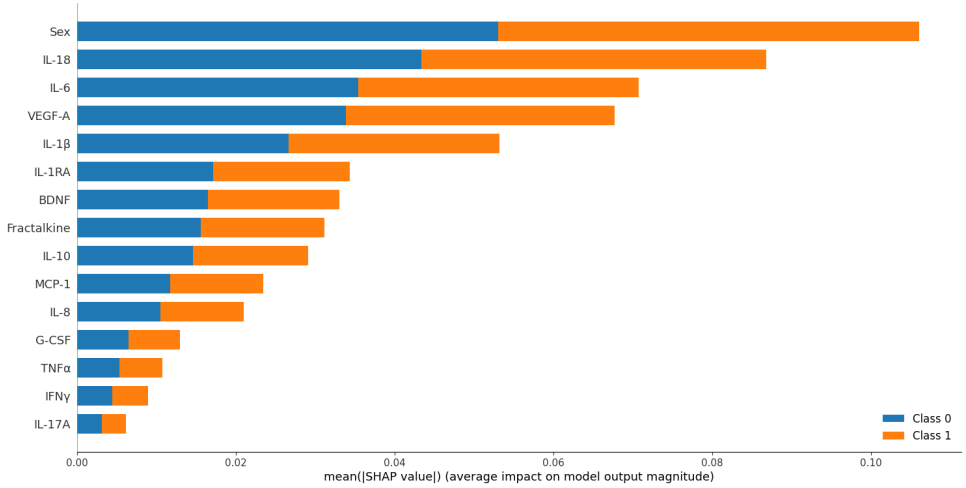


(b)

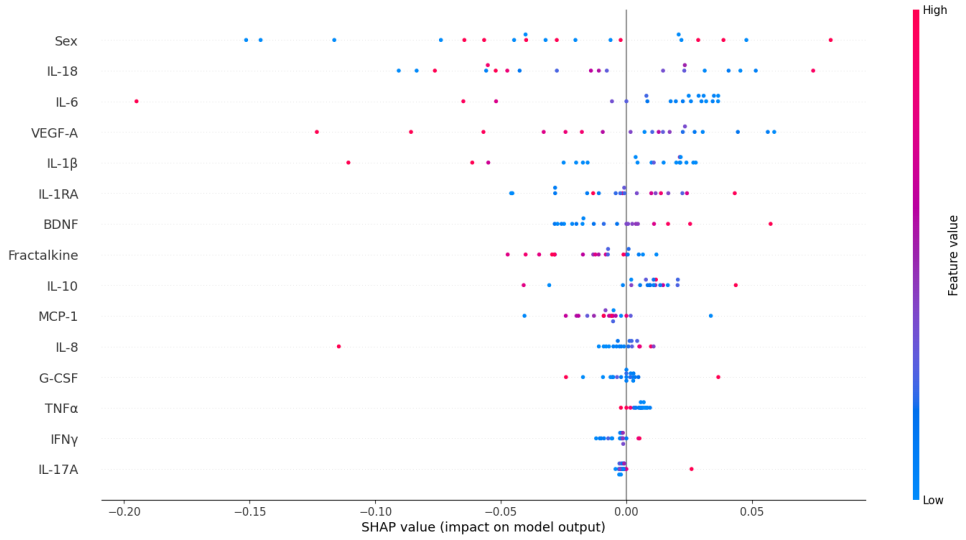


(c)

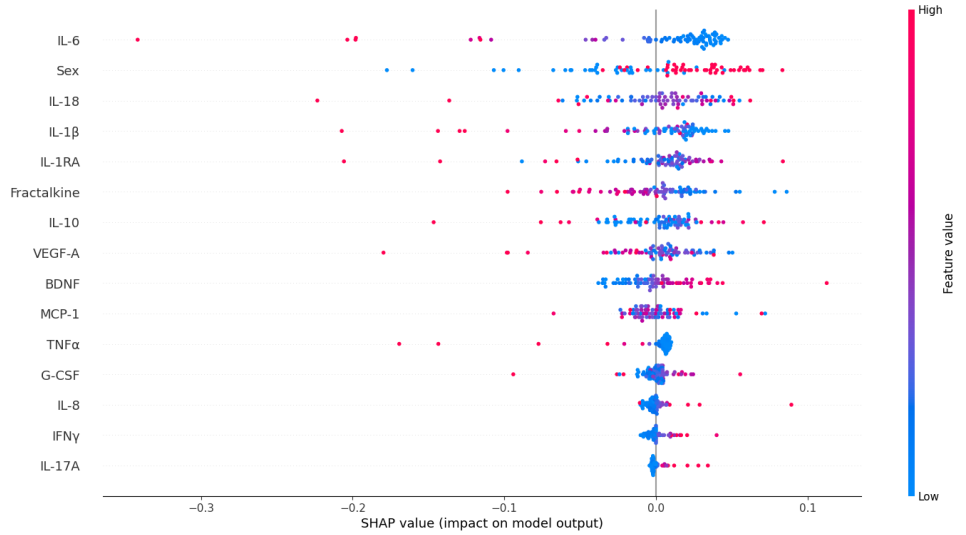
Figure 6.41: SHAP plots for SVC mean-filled model predicting AD from serum data: (a) Feature importance plot; (b) Test set beeswarm plot; (c) Train set beeswarm plot.



(a)



(b)



(c)

Figure 6.42: SHAP plots for SVC mode-filled model predicting AD from serum data: (a) Feature importance plot; (b) Test set beeswarm plot; (c) Train set beeswarm plot.

6.4 Predicting Age Group (Over/ Under 65)

Lastly, for comparative purposes we predicted age group (being over or under 65 years) based on TI data. In this dataset this task has vast similarities with predicting AD, since there are no AD patients with age under 65, as shown in Figure 6.43.

The best models achieved 100% on all metrics as they successfully classified every instance in the test set. Results are shown for median-filled SVC in Table 6.7 and in the confusion matrix and ROC curve presented in Figure 6.44.

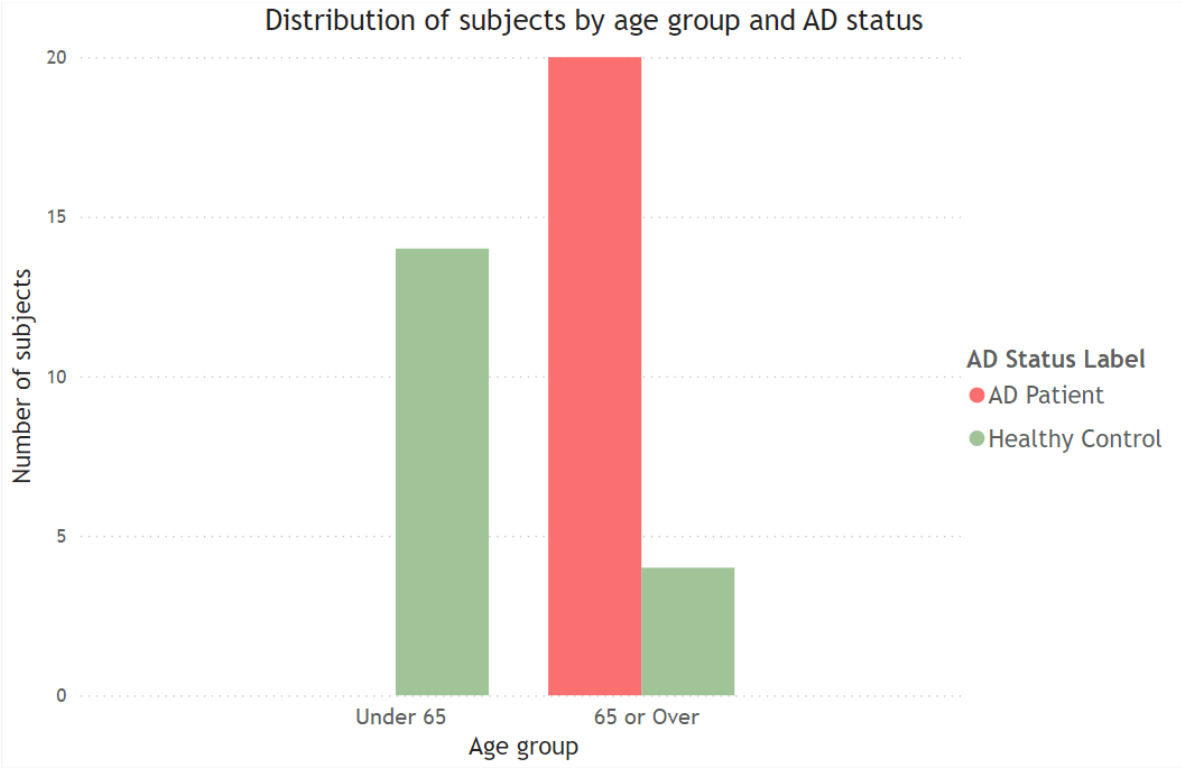


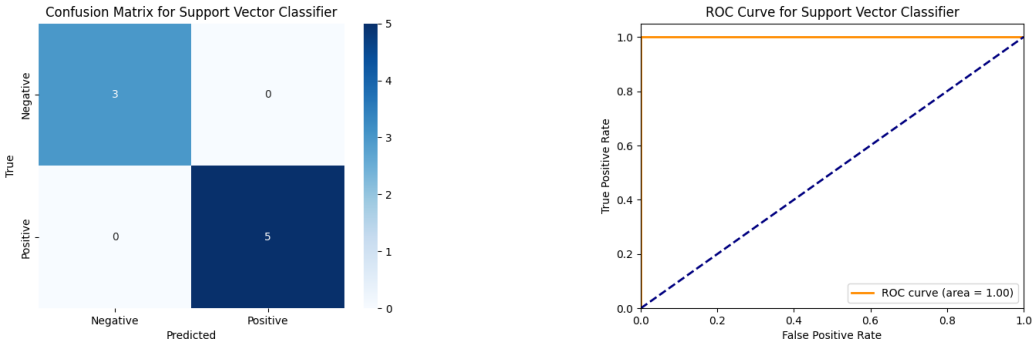
Figure 6.43: Bar plot of TI subjects' distribution per age group (over or under 65) and disease status.

Table 6.7: Performance metrics of top-scoring predictive model for age group using TI data.

Median-filled SVC	**Precision**	**Recall**	**F1 Score**
General	1.00	1.00	1.00
AD	1.00	1.00	1.00
Accuracy		1.00	
AUC		1.00	

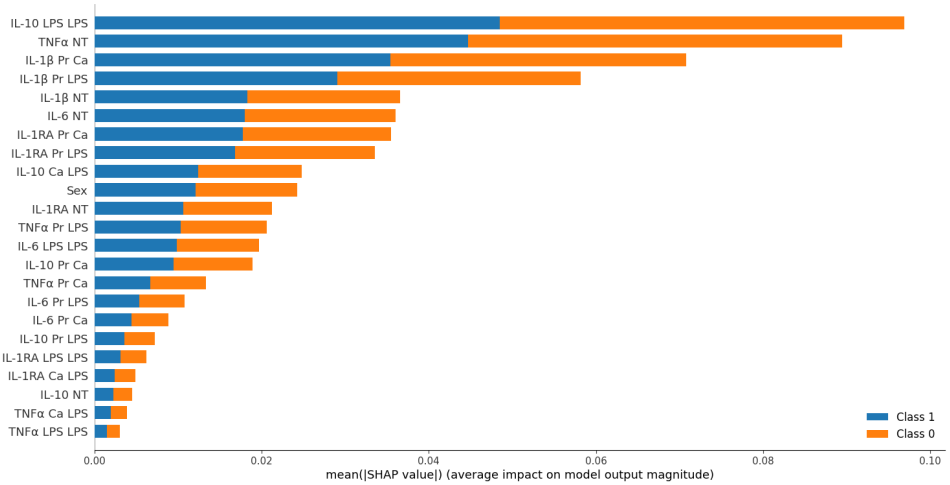
Analysing the SHAP plots for this optimal model highlights the impact of TI features on the model's performance for this task. Firstly, IL-10 LPS LPS is the leading feature on test set, and its lower values lead to the prediction of being over 65. Additionally, TNFα NT is also a principal feature of the model, with again lower values predicting

belonging to the elderly group. In priority order, IL-1β increased values in all conditions included (Pr Ca, Pr LPS and NT) appear to lead to the prediction of an individual being over the age of 65. Regarding IL-6 NT, higher values lead to positive predictions and IL-1RA displays different interactions in different conditions, with higher levels of IL-1RA Pr Ca driving to a positive prediction, in opposition to lower levels of IL-1RA Pr LPS which lead towards to this elder group prediction. Lastly, for features with mild relative contributions, IL-10 Ca LPS decreased expression is a contributor to the model's prediction of an individual being over the age of 65, as happened with IL-10 LPS LPS.

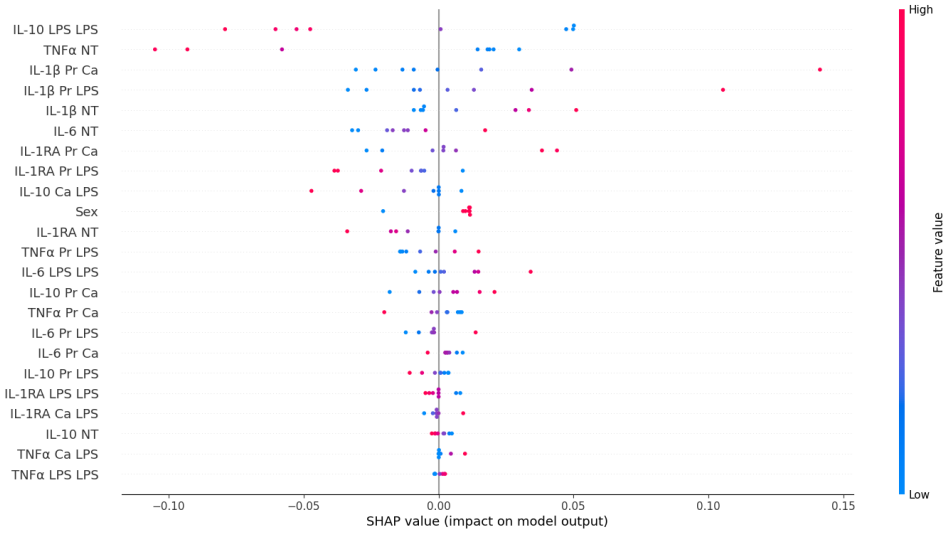


(a) (b)

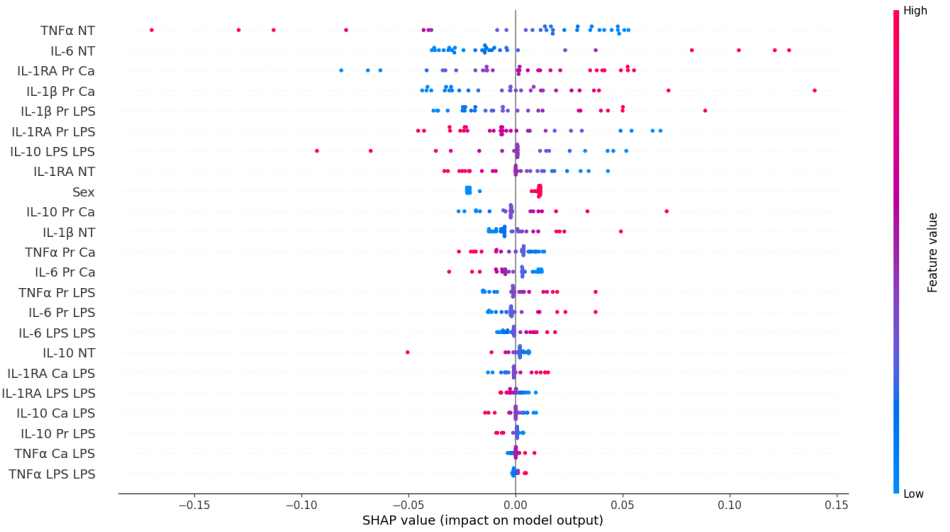
Figure 6.44: SVC median-filled model test set results for predicting age group from TI data: (a) Confusion matrix; (b) ROC curve.



(a)



(b)



(c)

Figure 6.45: SHAP plots for SVC mode-filled model predicting AD from serum data: (a) Feature importance plot; (b) Test set beeswarm plot; (c) Train set beeswarm plot.

To conclude this chapter, Table 6.8 presents an overview of the best results for every AD-predicting task, as well as the age-predicting task. The corresponding overviews of results for tasks predicting cytokine levels are shown in Tables 6.1 and 6.2 of Section 6.2. In Table 6.8, the data employed in each task is summarized (with sample size and brief description) along with the optimal model identified. The performance of these models is indicated by accuracy and F1 score. These findings provide a solid foundation for the subsequent analysis and discussion of the predictive models, which will be explored in the following chapter.

Table 6.8: Summary of best results for each task predicting AD or age group.

Task	**Data**	**Best Model**	**Accuracy**	**F1 Score**
Predict AD from IB	35 instances (IgG measures)	RF	62.9%	68.3%
Predict AD from TI	38 instances (cytokine levels expressed by isolated monocytes)	Logistic EN Regression	87.5%	87.3%
		Logistic Lasso Regression	87.5%	87.3%
Predict AD from IB and TI	38 instances (TI and IB data)	Median-filled DT	87.5%	87.3%
Predict AD from serum	99 instances (serum cytokine levels)	Mean-filled SVC	75%	74.9%
		Mode-filled SVC	80%	79.2%
Predict age group (over/ under 65)	38 instances (TI data)	Median-filled SVC	100%	100%

7

Discussion

This section analyses the findings and contributions made by the present study. Section 7.1 discusses the results from the cytokine-predicting models. Section 7.2 comprises all findings and implications from the AD predicting models. Namely, Section 7.2.1 discusses what can be concluded in terms of identifying a cytokine profile in primed and/or challenged cell response for the prediction of AD. Section 7.2.2 explores any serum cytokine profile found by our predictive models in AD and HC as well as discussing how these serum models compare to existing ones. Section 7.2.3 reports and delves into relationships found between IB, AD and cytokine profile. Lastly, Section 7.3 presents how the findings of our AD models compare to the results from the age predicting task and highlights this study's main limitations.

7.1 Modelling TI Cytokine Levels in AD

In the present study, we built moderately predictive regression models for *ex vivo* monocyte expressed TNFα in NT and Pr LPS conditions, IL-6 in NT, Pr LPS and LPS LPS conditions, IL-10 in Pr Ca, and both challenge conditions, IL-1β in NT and both priming conditions, and IL-1RA in NT, Pr LPS and LPS LPS conditions. Overall, our results display high variability performance across different cytokines and treatments which highlights the complex and heterogenous immune response of different individuals. Despite the challenging nature of this task and the possible influence of other regulatory pathways, external factors, or immune modulators not considered in this study, the results still demonstrate potential for predictability. This is particularly evident in the cases of IL-1RA Pr LPS and TNFα NT, TNFα Pr LPS, and IL-10 Ca LPS when oversampling was applied. The improvements in performance enabled by the oversampling of the data in many cases, also indicate the need for more samples for the training of models and suggest that data imbalance might influence predictions, with oversampling algorithms being a promising strategy for overcoming said obstacle. Furthermore, RF and XGB were the most effective algorithms across proteins and treatments, consistently outperforming others, potentially due the non-linear relationships

present in the data better captured by these complex algorithms. It is also interesting to explore why the response to LPS (whether in priming or challenge conditions) seems to be more predictive rather than non-treated or priming with *C. albicans* treatment. This effect might be due to the nature of LPS as a potent activator of the innate immune system. LPS robustly triggers inflammatory pathways, leading to a heightened cytokine response [187]. Perhaps, this response is both more intense and involved in AD specific pathways than the one elicited by *C. albicans* or the one reflected in homeostasis-regulated non-treated conditions, making them less predictive in this context. Additional studies could aid in clarifying how the distinct immune responses elicited by these pathogens are involved in AD mechanisms and pathways. One way to explore that could be through comparative transcriptomic or proteomic analyses, focusing on the signalling pathways activated by LPS and *C. albicans* in immune cells from AD patients and HCs, to determine whether and which inflammatory responses are specifically linked to AD pathology.

Attempting to explore the problem further and solve it as a classification problem by converting the targets to categorical variables, did not lead to significant improvement in performance. This conversion resulted in many targets having insufficient instances in each class, limiting the ability to effectively build classifiers. Even so, results across the cytokines and respective treatments which were classifiable, generally show moderate to low classification accuracy and F1 scores, with a mild improvement in most cases, when data was oversampled, as happened in regression. DT, RF and SVC were consistently among the top-performing algorithms, reinforcing that algorithms that capture more intricate non-linear relationships are better equipped for handling the dataset's complexities. Overall, cytokines like IL-10 and IL-1RA demonstrate relatively higher classification potential compared to others, especially after oversampling. Particularly, models for IL-10 LPS LPS and IL-1β NT showed best performance, suggesting a good starting point for the modelling of inflammatory response to infections in aging processes and in AD context.

In conclusion, our cytokine models, both in regression and classification, did not yield particularly high performance metrics which highlights the challenging nature of this task. Nevertheless, given that this approach to studying inflammatory pathways and responses to infectious burden in both healthy and diseased individuals, through the application of ML models, has not yet been explored extensively in the literature, it presents a promising and innovative avenue for future research.

7.2 Modelling Alzheimer's Disease

7.2.1 Identifying a Different Cytokine Profile in Primed and/or Challenged Cell Inflammatory Response in Predicting AD

In order to identify a cytokine profile in primed or challenged cell-response for AD patients, the results of the models predicting AD from TI data are analysed. Both of the top-performing models, median-filled logistic regression with EN penalty and mean-filled logistic regression with L1 penalty, achieved strong predictive performance on the test set, recalling every instance of the AD class. The high AUCs of 0.88 and 0.81, respectively, also suggest that these models offer a reliable framework for differentiating between AD patients and HCs, as models in recognized literature for this field usually present similar AUC values [144, 151, 153]. These metrics allow the conclusion that the models perform well, giving good results and thus allow for insights into the decision-making process, permitted by SHAP.

The SHAP analysis provides critical insights into the cytokine profiles that significantly influence the predictions of the models. In the two models, IL-10 LPS LPS, IL-1β Pr LPS, TNFα Pr LPS, and IL-6 Pr LPS are consistently identified as top predictors for AD. Notably, IL-10 LPS LPS is identified as the most significant feature in both models, where decreased levels of this anti-inflammatory cytokine correlate with predictions of AD, as opposite to the effect of the pro-inflammatory cytokines (IL-1β, IL-6 and TNFα) in Pr LPS conditions, in which lower levels lead to the prediction of the disease.

In this way, the obtained results suggest that higher values of the pro-inflammatory cytokines after the cells have been primed with LPS are indicators of AD, with the regulatory cytokine, IL-1RA, having no impact in these conditions nor the anti-inflammatory IL-10. This result ties well with previous studies wherein a potentiated trained immunity status is observed in AD, and the blood cells of patients display an amplified inflammatory response to a stimulus, along with increased expression of pro-inflammatory cytokines [22]. On the other hand, the negative impact of IL-10 when cells have been primed and challenged with LPS seems to suggest an impaired anti-inflammatory response after successive inoculations in AD patients. In this model, it is the response after the secondary challenge occurs, several days after the primary inoculation with LPS, that is characteristic of AD, translated into lower levels of IL-10 expression as compared to controls. These results are consistent with the proposed role of exacerbated inflammation in AD, as one study has reported patients to show a significant decrease in IL-10 production after stimulation in peripheral blood mononuclear cells [188] and another found AD patients to have a higher prevalence of the β 1082A allele, linked to lower IL-10 production [189]. The fact that this effect has been picked up in the "challenged" conditions as opposed to the "primed" conditions, indicates that the impaired anti-inflammatory response in AD patients may become more apparent after repeated exposure to inflammatory stimuli, rather than from the initial priming.

This finding aligns with the idea that chronic inflammation plays a pivotal role in the progression of AD, with patients' immune systems becoming progressively less capable of regulating inflammatory responses [190].

7.2.2 Evaluating Serum Cytokine Profile to Identify Differences Between AD Patients and Controls

The best models found for the task of predicting AD from serum-circulating cytokines were mode-filled and mean-filled SVC models with respective accuracies of 80% and 75%. Additionally, the mode-filled SVC model demonstrated superior performance compared to the mean-filled model, achieving a higher AUC of 0.86 in contrast to 0.74, thereby indicating enhanced overall efficacy. However, the superior recall for the AD class, showed by the mean-filled model, identifying 80% of AD patients correctly, leads us to the consideration and interpretation of both models. Although, not achieving the metrics yielded by the models trained on TI data, the performance yielded in this task is still significant as the models were built on a larger dataset, making for arguably more robust results. Furthermore, while there are other studies which achieve higher accuracies on serum data for the prediction of AD [8–10], the presented models still perform comparatively well, with the AUCs of our models (of 0.74 and 0.86), being consistent with those reported in studies using similar datasets for AD prediction [150, 153], while making use of a different plethora of proteins.

The SHAP analysis of our predictive models unveils a complex interplay between pro and anti-inflammatory serum cytokines in AD. Firstly, IL-18, a pro-inflammatory cytokine, is a key factor in both models, however with varying impacts. In the mean-filled model it is clear that reduced levels of this cytokine correlate with the prediction of the disease, wherein the mode-filled model this relationship is more complex, with some high level instances leading to prediction of HCs but others being linked to the prediction of AD. This is an interesting and controversial finding as serum studies have often found no differences in circulating IL-18 levels between AD patients and controls [191, 192], whilst a few studies, in opposition, found AD patients to have elevated peripheral IL-18 levels compared to HCs [77, 193].

The anti-inflammatory cytokine, IL-10, on the other hand, was found to be of moderate importance in the mean-filled SVC but emerged as the most significant predictor in the more accurate mode-filled classifier, with lower values considerably pushing predictions towards AD. This in line with current research, as has been explored in the previous section, with studies finding decreased IL-10 levels in serum of patients with AD [188, 194], which in turn reinforces the hypothesis that impaired anti-inflammatory responses, particularly lower IL-10 levels, play an important role in the pathogenesis of this disease.

Interestingly, in the two models, the pro-inflammatory cytokines such as IL-6, VEGF-A, and IL-1β, consistently show moderate to strong importance. In both classifiers,

lower levels of these cytokines contributed to the prediction of AD. This result is coherent with existing literature in the case of VEGF-A, since low serum levels of this protein have consistently been associated with AD [195, 196]. However, these are intriguing findings for IL-6 and IL-1β, where numerous papers have reported the opposite relationship. Increased levels of IL-6 have consistently been found in the serum of AD patients [60, 61, 197], with the same applying to serum levels of IL-1β [198–201]. Important questions about the intricate function of inflammation in AD are raised by this apparent discrepancy between our serum models' conclusions and the body of recognized literature. It can be hypothesized that while elevated levels of IL-6 and IL-1β are often reported in AD patients, lower serum concentrations of these cytokines may also hold significance in certain disease states or patient subgroups. This discrepancy could be further explored by conducting longitudinal studies that track cytokine levels across different stages of AD progression and different age groups or by stratifying patients based on specific clinical features, such as disease severity or comorbidities, to uncover potential subgroups where lower cytokine levels play a more prominent role.

A key area of divergence between the two models is the role of IL-8 and IL-1RA. IL-8, another pro-inflammatory cytokine, appears to have a stronger impact in the mean-filled model, where higher levels push towards prediction of AD. Although there is a lack of research with serum levels of IL-8 in AD patients, it has been observed that this cytokine may increase Tau phosphorylation and subsequent NFT formation (a hallmark of AD) with IL-8 being found to be increased in the CSF of AD patients [202]. An additional research paper found higher concentrations of serum IL-8 to be associated withAD, but only in the presence of significant cerebrovascular disease [203]. In this way, the findings of our models reinforce the potential relevance of IL-8 as a biomarker in AD, as captured by the models' ability to leverage higher serum levels for AD prediction.

Conversely, IL-1RA, whilst of only mild importance in either model, showed varying impacts. While in the mean-filled model, the correlation of lower levels with AD predictions is clear (and vice-versa), in the mode-filled model, it is higher levels of IL-1RA that are most often associated with prediction of the disease, but with some elevated instances also seemingly pushing towards the prediction of HCs. Current research does not systematically find variations in serum levels of IL-1RA in AD patients in comparison to controls [204], as there is only one study that reports IL-1RA serum levels to be increased in AD [205]. This inconsistency may point to the nuanced role that IL-1RA plays in AD, formerly analysed in Chapter 2, where its regulatory function could be context-dependent, balancing between beneficial and detrimental outcomes [23].

Lastly, it is important to discuss the role of sex in these models, as this variable also consistently appears as a feature with strong impact, as female patients apparently lead the model towards the prediction of AD. This finding is interesting, as sex is

moderately balanced in both classes, and as there is, in fact, a higher proportion of female healthy controls in the dataset as displayed in Figure 6.10. In this manner, the models' bias toward predicting AD in female subjects could reflect underlying biological or epidemiological trends rather than sample imbalance. On this note, current research has well documented that women are disproportionately affected by AD, in terms of both prevalence and progression of the disease. Studies suggest that higher life expectancy, postmenopausal hormonal changes, genetic predispositions such as the APOE $\epsilon 4$ allele, and variations in cognitive reserve may contribute to this increased risk in women [206]. This proven link may be captured by the models' reliance on sex as a predictive feature, highlighting the need of taking sex-specific factors into account in AD research and clinical diagnosis.

In conclusion, we have developed well performing models for the prediction of AD, envisioning the exploration of cytokine levels as accessible serum biomarkers, as well as shedding light into inflammatory processes behind this form of dementia, following in line with current research trends. The larger dataset used in these models provides a more robust foundation for these findings, however with only approximately 100 samples in total, the results still face limitations. Many studies and ML models in this domain typically utilize datasets comprising several hundred of samples, which offers greater statistical power and generalizability. Nevertheless, obtaining such datasets is known for being challenging, since typically clinical samples are scarce and data collection is resource-intensive.

Overall, the evaluation of serum cytokine profiles from the models achieved, highlights significant differences between inflammatory markers in AD prediction. Anti-inflammatory IL-10 and pro-inflammatory IL-18, IL-6, and IL-1β emerge as key contributors in both models, with a clear trend indicating that lower anti-inflammatory and dysregulated pro-inflammatory cytokines are important markers that lead to the prediction of AD. The differences between the two models, particularly in the roles of IL-18 and IL-1RA, reveal the need for further investigation into the specific inflammatory pathways involved in this disease. Moreover, discrepancies between our model results (particularly for the role of IL-18, IL-6, and IL-1β) make for intriguing findings and may reflect a specific stage or subtype of AD pathology, as well as variations in population, potentially underscoring the heterogeneous nature of immune responses in AD patients.

7.2.3 Reporting any Relationship Between IB, Cytokine Profile and AD

When predicting AD based solely on IB data, the RF classifier described in Section 6.3.1 performed with moderate success, as evidenced by its accuracy of approximately 63% and general F1 score of 62%. Nonetheless, while demonstrating some predictive power, particularly in its recall for AD cases, it's far from an ideal model as the AUC value of 0.56 underscores a limited ability to truly discriminate between AD and HC. This

model was limited by its sample size and thus, there was insufficient data to build a robust classifier which prevented the development of SHAP explainability plots for deeper insights into the model. Even so, the model's performance, which relies solely on six features (sex and antibody levels for the five pathogens assessed), aligns with our hypothesis of a potential relationship between certain infections and AD. In this way, it points towards further exploration of infection-related features in the prediction of AD with larger datasets, which could aid in uncovering more definitive relationships between IB and disease progression.

To this effect, when the IB data was fused with TI data, the increase in features and slight increase in sample size, allowed for the development of more robust models and for the final model setup. The optimal model, a median-filled DT classifier, performed comparably to models using only TI data, likewise achieving an accuracy of 87.5%, an AUC of 0.88, and 100% recall for the AD class. SHAP explanations revealed a similar pattern to the TI-based models, with decreased IL-10 LPS LPS and increased TNFα Pr LPS levels being major contributors to the prediction of AD. Intriguingly, as this model relies only on a small subset of the available features to make predictions due to its tree nature, it did not quite capture the predictive nature of other pro-inflammatory cytokines when primed with LPS. However, it picked up on the predictive power of IL-1β NT, where increased values of this cytokine appear to be linked to the prediction of AD, which in opposition to our findings in serum, is in accordance with recent studies [200, 201].

More to the point of IB relationships with AD, HSV-1/2 emerged as a feature of strong impact in the model, with increased values of this infectious feature being positively correlated with the prediction of AD. This finding not only aligns with the hypothesis that HSV is linked to the development of this disease [17, 36, 40], but also introduces the novel approach of using ML to integrate infection-related features with immune response data, providing valuable insights into AD classification. In addition, it opens an avenue for explaining the impact of IL-1β NT in this model, as HSV infection can trigger increased production of this cytokine [207]. The presence of HSV as a feature may enhance the model's ability to detect patterns involving both this pathogen and IL-1β production, contributing to a more comprehensive understanding of their roles in AD and undercovering pathways that could be further explored.

7.3 Comparative Analysis of Predicting Age Group and Predicting AD, from TI Data

Finally, we analyse the task of predicting age group from TI data and its comparison to the feature-like models predicting AD. This comparison is fundamental because, in our TI dataset, AD patients have a significantly higher median age than healthy individuals. Given that inflammation also changes with age, this comparison helps

ensure that our findings aren't solely attributable to the age of the subjects. However, since in our dataset this results in a high proportion of elder individuals being AD patients and all younger patients (under the age of 65) being HCs, as illustrated in Figure 6.2, this task also presents itself as rather challenging.

The optimal model for predicting age group performed exceptionally well, achieving perfect scores across all metrics, including 100% accuracy and an AUC of 1.00. In contrast, the models predicting AD, while still performing strongly, recorded slightly lower accuracies and AUC values. This disparity in performance may be partially attributed to the clearer cytokine profile distinctions for age group classification, whereas predicting AD requires the identification of more nuanced patterns associated with immune dysregulation specific to the disease.

Despite the difference in performance, both tasks share some similarities in feature impact, with IL-10 LPS LPS being the top feature in both models. In both cases, lower levels of anti-inflammatory IL-10, in primed and challenged with LPS conditions predicted AD and older age, reflecting the link between reduced anti-inflammatory responses and both aging and AD. With that in mind, the behaviour of pro-inflammatory cytokines such as IL-1β, TNFα, and IL-6 diverged between the tasks. In predicting AD, these cytokines were significant in primed with LPS conditions, while in predicting age, they were more relevant in non-treated conditions (with the exception of IL-1β Pr LPS which was impactful in both tasks), indicating potentially different immune responses for aging versus disease. In aging, as with AD, studies show increase in pro-inflammatory cytokine release [208], this phenomenon is not likely a coincidence, but arguably part of the pathophysiological mechanism of AD through age-linked dysregulation of inflammatory pathways [209]. In this regard, the difference found by our models is both interesting and intriguing, as it opens up new grounds for research. Particularly, it raises questions about why age is more accurately predicted by pro-inflammatory cytokine release in NT conditions, while AD is more strongly predicted after stimulation, in primed with LPS conditions. This distinction suggests a divergence in how immune responses are activated in the context of aging in opposition to AD progression, and warrants further research.

The regulatory and anti-inflammatory cytokine, IL-1RA played a limited role in AD prediction, not really displaying any impact in optimal models, but was more complex in age prediction, with its levels influencing predictions differently depending on the priming conditions. IL-1RA Pr LPS (as for NT) lead to prediction of being over 65 when the feature presents lower values, however in Pr Ca conditions, higher levels positively impact the model's prediction of an individual belonging to the older class. This highlights a more intricate role for inflammation regulation in aging and goes beyond previous studies which only report an increase of IL-1RA production with age [210, 211].

One major limitation of our work is the overlap between age and AD in the used dataset, as all AD patients are over 65 and the majority of HCs are under this threshold.

This makes it difficult to fully separate the effects of aging from AD. Hence, future work should focus on including age-matched controls to better isolate the cytokine signatures of AD and remove the confounding effect of age. Even so, our results still show distinction in TI cytokine profile, namely they indicate a much more clear pattern of increased pro-inflammatory cytokine levels in primed cell response, whilst the impact of IL-10 in LPS LPS remains to be attributed to AD, age or potentially both. Furthermore, this last task presents intriguing findings for the complex relationships between aging and inflammation, while potentially providing a basis for more research aimed at understanding how baseline immune activity and immune responses to external stimuli vary and evolve with age and in the development of AD.

8

Conclusion

In this study, we have explored the ability of modelling immune response to different stimuli in AD patients and healthy individuals. For this we leveraged ML algorithms, including linear and logistic regression with L1, L2 and EN penalties, DTs, RFs, SVMs, KNN and XGB, recurred to oversampling for data augmentation and adopted a nested CV approach for hyperparameter optimization. The regressors and classifiers deployed (the latter were used when target variables were converted to categorical labels) were trained on subjects' age, sex, antibody levels for HSV types 1 and 2, *H. pylori*, CMV, *C. pneumoniae* and *B. burgdorferi*, and AD status. With this approach, we achieved moderately performing models for predicting TNFα, IL-6, IL-10, IL-1β and IL-1RA in various responses to stimulation with LPS and *C. albicans*. Regression models were most successful for TNFα Pr LPS, IL-10 Ca LPS and IL-1RA Pr LPS, whereas classifiers for IL-10 LPS LPS and IL-1β NT displayed the most efficacy. Our results show the challenging nature of predicting immune response in individuals, particularly in the context of AD, nonetheless they hint at the existence of relationships between infectious burden, age, sex, AD and distinct inflammatory response in individuals. Furthermore, they show promise for the application of this novel approach in order to better understand mechanistic pathways behind inflammation and this form of dementia.

Another fundamental part of this work was the the development of ML models for the prediction of AD, following the same core approach. Models trained solely on IB performed moderately well, but vast improvements were achieved when TI data was utilized. Top performing models (trained solely on TI data and on both TI and IB data) achieved 87.5% accuracy, an AUC of 0.88 and and 100% recall for instances of AD class, on test set.

Through the deployment and interpretation of our models via SHAP, we found increased levels of pro-inflammatory cytokines (TNFα, IL-6 and IL-1β) in Pr LPS to lead to the prediction of AD, in line with our expectations and existent research [22]. We also found decreased levels of anti-inflammatory IL-10 in LPS LPS conditions to be correlated with prediction of the disease, reflecting an impaired anti-inflammatory

response in AD patients in reaction to successive infections [188, 190]. In models trained with IB data as well, higher antibody levels for HSV and higher levels of IL-1β NT were also major contributors to AD prediction. The impact of HSV is significant as it aligns with the infection hypothesis, fundament of our study, supporting that herpes may play a critical role in the development of AD [40].

However, strong age correlation with AD present in our dataset, elicited comparison with models trained on the same data, monocyte expressed cytokine levels in response to stimuli, for predicting age group (under or over 65). The optimal model found achieved perfect metric scores, correctly identifying every instance in the train and test sets, and its interpretation shared some similarities with the AD predicting models, namely the impact for IL-10 LPS LPS. In this way, our results reflect the link between dysregulated inflammatory responses in both aging and AD, and highlight the importance of thoroughly investigating how age-related changes in immune function, particularly in cytokine regulation, might contribute to AD pathogenesis.

We also made use serum cytokine data (which encompassed a larger portion of the dataset and various pro-inflammatory and anti-inflammatory cytokines) in order to build predictive models for AD and find potential biomarkers. The optimal models in this task achieved accuracies of 75% and 80% with AUC scores of 0.74 and 0.86, respectively. SHAP explanations revealed lower levels of IL-10 to be consistently linked to AD, reinforcing its role in impaired anti-inflammatory responses [188, 194]. VEGF-A, IL-6, and IL-1β also contributed, although lower levels predicted AD, which contrasts with existing literature for the latter two proteins [60, 197, 200, 201]. Some variables showed contrasting impacts in how they influenced AD prediction, namely IL-18 and IL-1RA suggesting nuanced roles of these cytokines in the pathology of this disease. Overall, these models underscore the importance of inflammatory processes in AD and the need for further investigation into the roles these cytokines play in disease progression, while the discrepancies found suggest potential stage or subtype-specific immune responses in AD which require further exploration.

Main limitations of the present work lie in sample size (as ML require vast amounts of data for better performance and generalizability) and confounding effect of age, namely in models predicting AD from TI data. Future research should include validation of our findings on larger datasets, including age-matched controls for discarding the confounding effects of natural aging processes. Further research could also extend to MCI patients in order to better illustrate disease progression, and explore other potential biomarkers or infectious agents reportedly linked to AD, such as SARS-CoV-2 [212].

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[2] M. Guerchet, M. Prince, M. Prina, et al. "Numbers of people with dementia worldwide: An update to the estimates in the World Alzheimer Report 2015". In: (2020) (cit. on p. 1).

[3] A. P. Porsteinsson et al. "Diagnosis of early Alzheimer's disease: clinical practice in 2021". In: *The journal of prevention of Alzheimer's disease* 8 (2021), pp. 371–386 (cit. on pp. 1, 3).

[4] B. J. Grabher. "Effects of Alzheimer disease on patients and their family". In: *Journal of nuclear medicine technology* 46.4 (2018), pp. 335–340 (cit. on p. 1).

[5] Y. Ju and K. Y. Tam. "Pathological mechanisms and therapeutic strategies for Alzheimer's disease". In: *Neural Regeneration Research* 17.3 (2022), pp. 543–549 (cit. on pp. 1, 4).

[6] X. Long et al. "Prediction and classification of Alzheimer disease based on quantification of MRI deformation". In: *PloS one* 12.3 (2017), e0173372 (cit. on pp. 1, 34).

[7] R. Kumari, A. Nigam, and S. Pushkar. "An efficient combination of quadruple biomarkers in binary classification using ensemble machine learning technique for early onset of Alzheimer disease". In: *Neural Computing and Applications* 34.14 (2022), pp. 11865–11884 (cit. on pp. 1, 34).

[8] S. Ray et al. "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins". In: *Nature medicine* 13.11 (2007), pp. 1359–1362 (cit. on pp. 1, 37, 102).

[9] L. Gaetani et al. "Neuroinflammation and Alzheimer's disease: a machine learning approach to CSF proteomics". In: *Cells* 10.8 (2021), p. 1930 (cit. on pp. 1, 36, 102).

[10] D. C. Araújo et al. "A novel panel of plasma proteins predicts progression in prodromal Alzheimer's disease". In: *Journal of Alzheimer's Disease* 88.2 (2022), pp. 549–561 (cit. on pp. 1, 36, 102).

[11] M. Tanveer et al. "Machine learning techniques for the diagnosis of Alzheimer's disease: A review". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16.1s (2020), pp. 1–35 (cit. on p. 1).

[12] E. Joe and J. M. Ringman. "Cognitive symptoms of Alzheimer's disease: clinical management and prevention". In: *bmj* 367 (2019) (cit. on pp. 1, 3).

[13] S. Sadigh-Eteghad et al. "Amyloid-beta: a crucial factor in Alzheimer's disease". In: *Medical principles and practice* 24.1 (2015), pp. 1–10 (cit. on pp. 1, 3).

[14] R. Roychaudhuri et al. "Amyloid β-protein assembly and Alzheimer disease". In: *Journal of Biological Chemistry* 284.8 (2009), pp. 4749–4753 (cit. on p. 1).

[15] C. L. Masters and D. J. Selkoe. "Biochemistry of amyloid β-protein and amyloid deposits in Alzheimer disease". In: *Cold Spring Harbor perspectives in medicine* 2.6 (2012), a006262 (cit. on p. 1).

[16] H. Ashrafian, E. H. Zadeh, and R. H. Khan. "Review on Alzheimer's disease: inhibition of amyloid beta and tau tangle formation". In: *International journal of biological macromolecules* 167 (2021), pp. 382–394 (cit. on pp. 1, 3).

[17] R. F. Itzhaki et al. "Herpes simplex virus type 1 in brain and risk of Alzheimer's disease". In: *The Lancet* 349.9047 (1997), pp. 241–244 (cit. on pp. 1, 4, 105).

[18] X.-L. Bu et al. "A study on the association between infectious burden and A lzheimer's disease". In: *European journal of neurology* 22.12 (2015), pp. 1519–1525 (cit. on p. 1).

[19] D. Vigasova et al. "Multi-pathogen infections and Alzheimer's disease". In: *Microbial cell factories* 20 (2021), pp. 1–13 (cit. on pp. 1, 7).

[20] M. G. Netea et al. "Defining trained immunity and its role in health and disease". In: *Nature Reviews Immunology* 20.6 (2020), pp. 375–388 (cit. on pp. 1, 7).

[21] A.-C. Wendeln et al. "Innate immune memory in the brain shapes neurological disease hallmarks". In: *Nature* 556.7701 (2018), pp. 332–338 (cit. on pp. 2, 8).

[22] F. Salani et al. "Is innate memory a double-edge sword in Alzheimer's disease? a reappraisal of new concepts and old data". In: *Frontiers in Immunology* 10 (2019), p. 466833 (cit. on pp. 2, 8, 101, 109).

[23] D. Boraschi et al. "Cause or consequence? The role of IL-1 family cytokines and receptors in neuroinflammatory and neurodegenerative diseases". In: *Frontiers in Immunology* 14 (2023), p. 1128190 (cit. on pp. 2, 10, 103).

[24] R. Guerreiro and J. Bras. "The age factor in Alzheimer's disease". In: *Genome medicine* 7 (2015), pp. 1–3 (cit. on pp. 3, 60).

[25] P. Scheltens et al. "Alzheimer's disease". In: *The Lancet* 397.10284 (2021), pp. 1577–1590 (cit. on pp. 3, 33).

[26] X.-L. Li et al. "Behavioral and psychological symptoms in Alzheimer's disease". In: *BioMed research international* 2014 (2014) (cit. on p. 3).

[27] C. G. Lyketsos et al. *Neuropsychiatric symptoms in Alzheimer's disease*. 2011 (cit. on p. 3).

[28] A. A. Tahami Monfared et al. "Alzheimer's disease: epidemiology and clinical progression". In: *Neurology and therapy* 11.2 (2022), pp. 553–569 (cit. on p. 3).

[29] H. Hampel et al. "The amyloid-β pathway in Alzheimer's disease". In: *Molecular psychiatry* 26.10 (2021), pp. 5481–5503 (cit. on p. 3).

[30] S. Muralidar et al. "Role of tau protein in Alzheimer's disease: The prime pathological player". In: *International journal of biological macromolecules* 163 (2020), pp. 1599–1617 (cit. on p. 3).

[31] S. Wegmann, J. Biernat, and E. Mandelkow. "A current view on Tau protein phosphorylation in Alzheimer's disease". In: *Current opinion in neurobiology* 69 (2021), pp. 131–138 (cit. on p. 3).

[32] J. P. Spencer et al. "Neuroinflammation: modulation by flavonoids and mechanisms of action". In: *Molecular aspects of medicine* 33.1 (2012), pp. 83–97 (cit. on p. 4).

[33] I. G. Onyango et al. "Neuroinflammation in Alzheimer's disease". In: *Biomedicines* 9.5 (2021), p. 524 (cit. on p. 4).

[34] T. Fulop et al. "Can an infection hypothesis explain the beta amyloid hypothesis of Alzheimer's disease?" In: *Frontiers in aging neuroscience* 10 (2018), p. 224 (cit. on p. 4).

[35] M. Wozniak, A. Mee, and R. Itzhaki. "Herpes simplex virus type 1 DNA is located within Alzheimer's disease amyloid plaques". In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 217.1 (2009), pp. 131–138 (cit. on pp. 4, 5).

[36] R. F. Itzhaki. "Herpes and Alzheimer's disease: subversion in the central nervous system and how it might be halted". In: *Journal of Alzheimer's Disease* 54.4 (2016), pp. 1273–1281 (cit. on pp. 4, 105).

[37] I. Vojtechova et al. "Infectious origin of Alzheimer's disease: Amyloid beta as a component of brain antimicrobial immunity". In: *PLoS Pathogens* 18.11 (2022), e1010929 (cit. on p. 4).

[38] T. Piekut et al. "Infectious agents and Alzheimer's disease". In: *Journal of Integrative Neuroscience* 21.2 (2022), p. 73 (cit. on pp. 5, 7).

[39] T. J. Taylor et al. "Herpes simplex virus". In: *Front Biosci* 7.1-3 (2002), pp. d752–64 (cit. on p. 5).

[40] R. F. Itzhaki. "Overwhelming evidence for a major role for herpes simplex virus type 1 (HSV1) in Alzheimer's disease (AD); underwhelming evidence against". In: *Vaccines* 9.6 (2021), p. 679 (cit. on pp. 5, 105, 110).

[41] S. Feng et al. "Mechanistic insights into the role of herpes simplex virus 1 in Alzheimer's disease". In: *Frontiers in Aging Neuroscience* 15 (2023) (cit. on p. 5).

[42] M. Linard et al. "Interaction between APOE4 and herpes simplex virus type 1 in Alzheimer's disease". In: *Alzheimer's & Dementia* 16.1 (2020), pp. 200–208 (cit. on p. 5).

[43] M. L. C. Santos et al. "Helicobacter pylori infection: Beyond gastric manifestations". In: *World journal of gastroenterology* 26.28 (2020), p. 4076 (cit. on p. 5).

[44] M. Kandpal et al. "Gut-brain axis interplay via STAT3 pathway: Implications of Helicobacter pylori derived secretome on inflammation and Alzheimer's disease". In: *Virulence* 15.1 (2024), p. 2303853 (cit. on p. 5).

[45] M. Doulberis et al. "Alzheimer's disease and gastrointestinal microbiota; impact of Helicobacter pylori infection involvement". In: *International Journal of Neuroscience* 131.3 (2021), pp. 289–301 (cit. on p. 6).

[46] S. Sanami et al. "Association between cytomegalovirus infection and neurological disorders: A systematic review". In: *Reviews in Medical Virology* 34.3 (2024), e2532 (cit. on p. 6).

[47] P. Biagio et al. "Alzheimer's disease and herpes viruses: Current events and perspectives". In: *Reviews in Medical Virology* 34.3 (2024), e2550 (cit. on p. 6).

[48] M. R. Hammerschlag, S. A. Kohlhoff, and C. A. Gaydos. "Chlamydia pneumoniae". In: *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases* (2015), p. 2174 (cit. on p. 6).

[49] H. C. Gérard et al. "Chlamydophila (Chlamydia) pneumoniae in the Alzheimer's brain". In: *FEMS Immunology & Medical Microbiology* 48.3 (2006), pp. 355–366 (cit. on p. 6).

[50] L. Subedi et al. "Chlamydia pneumoniae in Alzheimer's disease pathology". In: *Frontiers in Neuroscience* 18 (2024), p. 1393293 (cit. on p. 6).

[51] C. Kurokawa et al. "Interactions between Borrelia burgdorferi and ticks". In: *Nature Reviews Microbiology* 18.10 (2020), pp. 587–600 (cit. on p. 6).

[52] A. B. MacDonald. "Borrelia invasion of brain pyramidal neurons and biofilm Borrelia plaques in neuroborreliosis dementia with Alzheimer's phenotype". In: *Microbiol Infect Dis* 5.1 (2021), pp. 1–11 (cit. on p. 7).

[53] A. G. Senejani et al. "Borrelia burgdorferi co-localizing with amyloid markers in Alzheimer's disease brain tissues". In: *Journal of Alzheimer's Disease* 85.2 (2022), pp. 889–903 (cit. on p. 7).

[54] J. Ochando et al. "Trained immunity—basic concepts and contributions to immunopathology". In: *Nature Reviews Nephrology* 19.1 (2023), pp. 23–37 (cit. on p. 7).

[55] J. J. Oppenheim. "Cytokines: past, present, and future". In: *International journal of hematology* 74 (2001), pp. 3–8 (cit. on p. 8).

[56] C. A. Dinarello. "Proinflammatory cytokines". In: *Chest* 118.2 (2000), pp. 503–508 (cit. on p. 8).

[57] W. J. Branchett and C. M. Lloyd. "Regulatory cytokine function in the respiratory tract". In: *Mucosal immunology* 12.3 (2019), pp. 589–600 (cit. on p. 8).

[58] H. E. Ennerfelt and J. R. Lukens. "The role of innate immunity in Alzheimer's disease". In: *Immunological reviews* 297.1 (2020), pp. 225–246 (cit. on p. 9).

[59] S. Rose-John. "Interleukin-6 signalling in health and disease". In: *F1000Research* 9 (2020) (cit. on p. 9).

[60] N. Jain et al. "ASSESSMENT OF INTERLEUKIN-6 (IL-6) LEVELS IN ALZHEIMER'S DISEASE AND ITS RELATION WITH SEVERITY OF DISEASE: A CASE CONTROL STUDY'". In: *NeuroQuantology* 21.6 (2023), p. 1436 (cit. on pp. 9, 103, 110).

[61] W. Swardfager et al. "A meta-analysis of cytokines in Alzheimer's disease". In: *Biological psychiatry* 68.10 (2010), pp. 930–941 (cit. on pp. 9, 11, 103).

[62] N. M. Lyra e Silva et al. "Pro-inflammatory interleukin-6 signaling links cognitive impairments and peripheral metabolic alterations in Alzheimer's disease". In: *Translational psychiatry* 11.1 (2021), p. 251 (cit. on p. 9).

[63] B. Arosio et al. "Interleukin-10 and interleukin-6 gene polymorphisms as risk factors for Alzheimer's disease". In: *Neurobiology of Aging* 25.8 (2004), pp. 1009–1015 (cit. on p. 9).

[64] R. Sabat et al. "Biology of interleukin-10". In: *Cytokine & growth factor reviews* 21.5 (2010), pp. 331–344 (cit. on p. 9).

[65] L. D'anna et al. "Serum interleukin-10 levels correlate with cerebrospinal fluid amyloid beta deposition in Alzheimer disease patients". In: *Neurodegenerative Diseases* 17.4-5 (2017), pp. 227–234 (cit. on p. 9).

[66] L. L. Weston et al. "Interleukin-10 deficiency exacerbates inflammation-induced tau pathology". In: *Journal of neuroinflammation* 18 (2021), pp. 1–13 (cit. on p. 9).

[67] C. Porro, A. Cianciulli, and M. A. Panaro. "The regulatory role of IL-10 in neurodegenerative diseases". In: *Biomolecules* 10.7 (2020), p. 1017 (cit. on p. 9).

[68] S. Gunes et al. "Biomarkers for Alzheimer's disease in the current state: a narrative review". In: *International journal of molecular sciences* 23.9 (2022), p. 4962 (cit. on p. 10).

[69] E. Nam et al. "Serum tau proteins as potential biomarkers for the assessment of Alzheimer's disease progression". In: *International journal of molecular sciences* 21.14 (2020), p. 5007 (cit. on p. 10).

[70] P. R. Kac et al. "Diagnostic value of serum versus plasma phospho-tau for Alzheimer's disease". In: *Alzheimer's Research & Therapy* 14.1 (2022), p. 65 (cit. on p. 10).

[71] F. Gonzalez-Ortiz et al. "Brain-derived tau: a novel blood-based biomarker for Alzheimer's disease-type neurodegeneration". In: *Brain* 146.3 (2023), pp. 1152–1165 (cit. on p. 10).

[72] L. Piubelli et al. "Serum D-serine levels are altered in early phases of Alzheimer's disease: towards a precocious biomarker". In: *Translational Psychiatry* 11.1 (2021), p. 77 (cit. on p. 10).

[73] M. Zhang et al. "Serum miR-128 serves as a potential diagnostic biomarker for Alzheimer's disease". In: *Neuropsychiatric Disease and Treatment* (2021), pp. 269–275 (cit. on p. 11).

[74] S. Madadi et al. "Downregulation of serum miR-106b: a potential biomarker for Alzheimer disease". In: *Archives of physiology and biochemistry* 128.4 (2022), pp. 875–879 (cit. on p. 11).

[75] A. Varesi et al. "Blood-based biomarkers for Alzheimer's disease diagnosis and progression: an overview". In: *Cells* 11.8 (2022), p. 1367 (cit. on p. 11).

[76] Y. Mori et al. "Serum BDNF as a potential biomarker of Alzheimer's disease: verification through assessment of serum, cerebrospinal fluid, and medial temporal lobe atrophy". In: *Frontiers in neurology* 12 (2021), p. 653267 (cit. on p. 11).

[77] K. S. P. Lai et al. "Peripheral inflammatory markers in Alzheimer's disease: a systematic review and meta-analysis of 175 studies". In: *Journal of Neurology, Neurosurgery & Psychiatry* 88.10 (2017), pp. 876–882 (cit. on pp. 11, 102).

[78] L. Vanneschi and S. Silva. *Lectures on Intelligent Systems*. Springer, 2023 (cit. on pp. 13, 15–17, 19, 21, 23–25, 27).

[79] H. Jiang. *Machine learning fundamentals: A concise introduction*. Cambridge University Press, 2021 (cit. on pp. 13, 14).

[80] T. P. Trappenberg. *Fundamentals of machine learning*. Oxford University Press, 2019 (cit. on p. 13).

[81] R. E. Schapire and Y. Freund. "Foundations of machine learning". In: (2012) (cit. on p. 14).

[82] Z. Ghahramani. "Unsupervised learning". In: *Summer school on machine learning*. Springer, 2003, pp. 72–112 (cit. on p. 14).

[83] Y. Xu and R. Goodacre. "On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning". In: *Journal of analysis and testing* 2.3 (2018), pp. 249–262 (cit. on p. 14).

[84] S. Raschka and V. Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd, 2019 (cit. on p. 15).

[85] N. S. Punn and S. Agarwal. "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks". In: *Applied Intelligence* 51.5 (2021), pp. 2689–2702 (cit. on p. 16).

[86] Ž. Vujović et al. "Classification model evaluation metrics". In: *International Journal of Advanced Computer Science and Applications* 12.6 (2021), pp. 599–606 (cit. on p. 16).

[87] P. Fränti and R. Mariescu-Istodor. "Soft precision and recall". In: *Pattern Recognition Letters* 167 (2023), pp. 115–121 (cit. on p. 17).

[88] N. W. S. Wardhani et al. "Cross-validation metrics for evaluating classification performance on imbalanced data". In: *2019 international conference on computer, control, informatics and its applications (IC3INA)*. IEEE. 2019, pp. 14–18 (cit. on p. 17).

[89] V. Plevris et al. "Investigation of performance metrics in regression analysis and machine learning-based prediction models". In: *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*. European Community on Computational Methods in Applied Sciences. 2022 (cit. on p. 18).

[90] M. V. Shcherbakov et al. "A survey of forecast error measures". In: *World applied sciences journal* 24.24 (2013), pp. 171–176 (cit. on p. 18).

[91] C. J. Willmott and K. Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". In: *Climate research* 30.1 (2005), pp. 79–82 (cit. on p. 18).

[92] B. Vidgen and L. Derczynski. "Directions in abusive language training data, a systematic review: Garbage in, garbage out". In: *Plos one* 15.12 (2020), e0243300 (cit. on p. 19).

[93] C. V. G. Zelaya. "Towards explaining the effects of data preprocessing on machine learning". In: *2019 IEEE 35th international conference on data engineering (ICDE)*. IEEE. 2019, pp. 2086–2090 (cit. on p. 19).

[94] J. Brownlee. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020 (cit. on p. 19).

[95] D. Singh and B. Singh. "Investigating the impact of data normalization on classification performance". In: *Applied Soft Computing* 97 (2020), p. 105524 (cit. on p. 19).

[96] A. Subasi. *Practical machine learning for data analysis using python*. Academic Press, 2020 (cit. on pp. 19, 21).

[97] M. M. Ahsan et al. "Effect of data scaling methods on machine learning algorithms and model performance". In: *Technologies* 9.3 (2021), p. 52 (cit. on p. 19).

[98] V. G. Raju et al. "Study the influence of normalization/transformation process on the accuracy of supervised classification". In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE. 2020, pp. 729–735 (cit. on p. 19).

[99] S. Patro and K. K. Sahu. "Normalization: A preprocessing stage". In: *arXiv preprint arXiv:1503.06462* (2015) (cit. on p. 20).

[100] C. J. Cruz. "Exercise 3: Implemeting Simple Linear Regression Model using Neural Networks". In: () (cit. on p. 20).

[101] T. Emmanuel et al. "A survey on missing data in machine learning". In: *Journal of Big data* 8 (2021), pp. 1–37 (cit. on p. 20).

[102] S. M. Mostafa et al. "CBRG: a novel algorithm for handling missing data using Bayesian ridge regression and feature selection based on gain ratio". In: *IEEE Access* 8 (2020), pp. 216969–216985 (cit. on p. 20).

[103] N. A. A. Wafaa Mustafa Hameed. "Comparison of seventeen missing value imputation techniques". In: *Journal of Hunan University Natural Sciences* 49.7 (2022) (cit. on p. 20).

[104] S. Uddin and H. Lu. "Dataset meta-level and statistical features affect machine learning performance". In: *Scientific Reports* 14.1 (2024), p. 1670 (cit. on p. 20).

[105] A. Althnian et al. "Impact of dataset size on classification performance: an empirical evaluation in the medical domain". In: *Applied Sciences* 11.2 (2021), p. 796 (cit. on p. 20).

[106] R. D. King, O. I. Orhobor, and C. C. Taylor. "Cross-validation is safe to use". In: *Nature Machine Intelligence* 3.4 (2021), pp. 276–276 (cit. on p. 21).

[107] T. J. Bradshaw et al. "A guide to cross-validation for artificial intelligence in medical imaging". In: *Radiology: Artificial Intelligence* 5.4 (2023), e220232 (cit. on p. 21).

[108] H. J. Weerts, A. C. Mueller, and J. Vanschoren. "Importance of tuning hyperparameters of machine learning algorithms". In: *arXiv preprint arXiv:2007.07588* (2020) (cit. on pp. 21, 52).

[109] S. Varma and R. Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC bioinformatics* 7 (2006), pp. 1–8 (cit. on pp. 21, 22).

[110] S. Raschka. "Model evaluation, model selection, and algorithm selection in machine learning". In: *arXiv preprint arXiv:1811.12808* (2018) (cit. on pp. 21, 22).

[111] R. Mohammed, J. Rawashdeh, and M. Abdullah. "Machine learning with oversampling and undersampling techniques: overview study and experimental results". In: *2020 11th international conference on information and communication systems (ICICS)*. IEEE. 2020, pp. 243–248 (cit. on p. 22).

[112] B. Santoso et al. "Synthetic over sampling methods for handling class imbalanced problems: A review". In: *IOP conference series: earth and environmental science*. Vol. 58. 1. IOP Publishing. 2017, p. 012031 (cit. on p. 23).

[113] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya. "A review on imbalanced data handling using undersampling and oversampling technique". In: *Int. J. Recent Trends Eng. Res* 3.4 (2017), pp. 444–449 (cit. on p. 23).

[114] D. H. Wolpert, W. G. Macready, et al. *No free lunch theorems for search*. Tech. rep. Citeseer, 1995 (cit. on p. 23).

[115] L. E. Eberly. "Multiple linear regression". In: *Topics in biostatistics* (2007), pp. 165–187 (cit. on p. 23).

[116] X. Su, X. Yan, and C.-L. Tsai. "Linear regression". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.3 (2012), pp. 275–294 (cit. on p. 24).

[117] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288 (cit. on p. 24).

[118] J. Ranstam and J. A. Cook. "LASSO regression". In: *Journal of British Surgery* 105.10 (2018), pp. 1348–1348 (cit. on p. 25).

[119] G. C. McDonald. "Ridge regression". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1 (2009), pp. 93–100 (cit. on p. 25).

[120] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320 (cit. on p. 25).

[121] B. De Ville. "Decision trees". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (2013), pp. 448–455 (cit. on p. 26).

[122] S. J. Rigatti. "Random forest". In: *Journal of Insurance Medicine* 47.1 (2017), pp. 31–39 (cit. on p. 26).

[123] Y. Qi. "Random forest for bioinformatics". In: *Ensemble machine learning: Methods and applications* (2012), pp. 307–323 (cit. on p. 26).

[124] G. Biau and E. Scornet. "A random forest guided tour". In: *Test* 25 (2016), pp. 197–227 (cit. on p. 26).

[125] W. S. Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567 (cit. on p. 26).

[126] M. Awad et al. "Support vector regression". In: *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (2015), pp. 67–80 (cit. on p. 26).

[127] M. Singla and K. Shukla. "Robust statistics-based support vector machine and its variants: a survey". In: *Neural Computing and Applications* 32.15 (2020), pp. 11173–11194 (cit. on p. 26).

[128] M. Steinbach and P.-N. Tan. "kNN: k-nearest neighbors". In: *The top ten algorithms in data mining*. Chapman and Hall/CRC, 2009, pp. 165–176 (cit. on p. 27).

[129] R. Odegua. "Predicting bank loan default with extreme gradient boosting". In: *arXiv preprint arXiv:2002.02011* (2020) (cit. on p. 27).

[130] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794 (cit. on p. 27).

[131] J. Brownlee. *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery, 2016 (cit. on p. 27).

[132] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020 (cit. on pp. 27, 28, 30, 31).

[133] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017) (cit. on p. 28).

[134] V. Chen et al. "Best practices for interpretable machine learning in computational biology". In: *Biorxiv* (2022), pp. 2022–10 (cit. on p. 28).

[135] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 28, 29).

[136] L. S. Shapley et al. "A value for n-person games". In: (1953) (cit. on p. 28).

[137] A. Bhattacharya. *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022 (cit. on pp. 28, 29).

[138] T. D. Science. *Using SHAP Values to Explain How Your Machine Learning Model Works*. Accessed: 2024-07-19. 2020. URL: https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137 (cit. on p. 30).

[139] A. S. Fleisher et al. "Identification of Alzheimer disease risk by functional magnetic resonance imaging". In: *Archives of Neurology* 62.12 (2005), pp. 1881–1888 (cit. on p. 33).

[140] G. Muehllehner and J. S. Karp. "Positron emission tomography". In: *Physics in Medicine & Biology* 51.13 (2006), R117 (cit. on p. 33).

[141] G. Castellazzi et al. "A machine learning approach for the differential diagnosis of Alzheimer and vascular dementia fed by MRI selected features". In: *Frontiers in neuroinformatics* 14 (2020), p. 25 (cit. on p. 33).

[142] A. W. Salehi et al. "A CNN model: earlier diagnosis and classification of Alzheimer disease using MRI". In: *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE. 2020, pp. 156–161 (cit. on p. 34).

[143] P. M. Tuan et al. "AutoEncoder-based feature ranking for Alzheimer Disease classification using PET image". In: *Machine Learning with Applications* 6 (2021), p. 100184 (cit. on p. 34).

[144] J. Peng et al. "18F-FDG-PET radiomics based on white matter predicts the progression of mild cognitive impairment to Alzheimer disease: A machine learning study". In: *Academic Radiology* 30.9 (2023), pp. 1874–1884 (cit. on pp. 34, 101).

[145] R. Nancy Noella and J. Priyadarshini. "Machine learning algorithms for the diagnosis of Alzheimer and Parkinson disease". In: *Journal of Medical Engineering & Technology* 47.1 (2023), pp. 35–43 (cit. on p. 34).

[146] J. Neelaveni and M. G. Devasana. "Alzheimer disease prediction using machine learning algorithms". In: *2020 6th international conference on advanced computing and communication systems (ICACCS)*. IEEE. 2020, pp. 101–104 (cit. on p. 35).

[147] M. Bari Antor et al. "A comparative analysis of machine learning algorithms to predict alzheimer's disease". In: *Journal of Healthcare Engineering* 2021.1 (2021), p. 9917919 (cit. on p. 35).

[148] F. Wang et al. "Lipoproteins and metabolites in diagnosing and predicting Alzheimer's disease using machine learning". In: *Lipids in Health and Disease* 23.1 (2024), p. 152 (cit. on p. 35).

[149] A. S. Alatrany et al. "Stacked machine learning model for predicting alzheimer's disease based on genetic data". In: *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE. 2021, pp. 594–598 (cit. on p. 35).

[150] X. R. Gao et al. "Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction". In: *Scientific reports* 13.1 (2023), p. 450 (cit. on pp. 35, 102).

[151] M. Jia et al. "Predicting Alzheimer's Disease with Interpretable Machine Learning". In: *Dementia and Geriatric Cognitive Disorders* 52.4 (2023), pp. 249–257 (cit. on pp. 35, 101).

[152] A. S. Kononikhin et al. "Prognosis of Alzheimer's disease using quantitative mass spectrometry of human blood plasma proteins and machine learning". In: *International journal of molecular sciences* 23.14 (2022), p. 7907 (cit. on p. 36).

[153] S. J. Furney et al. "Combinatorial markers of mild cognitive impairment conversion to Alzheimer's disease-cytokines and MRI measures together predict disease progression". In: *Journal of Alzheimer's Disease* 26.s3 (2011), pp. 395–405 (cit. on pp. 36, 101, 102).

[154] A. Galgani et al. "Biological determinants of blood-based cytokines in the Alzheimer's disease clinical continuum". In: *Journal of Neurochemistry* 163.1 (2022), pp. 40–52 (cit. on p. 37).

[155] D. Ahmadi Rastegar et al. "Parkinson's progression prediction using machine learning and serum cytokines". In: *NPJ Parkinson's disease* 5.1 (2019), p. 14 (cit. on p. 37).

[156] S. S. Saharan et al. "Machine learning and statistical approaches for classification of risk of coronary artery disease using plasma cytokines". In: *BioData Mining* 14 (2021), pp. 1–14 (cit. on p. 37).

[157] F. Wei et al. "Machine learning for prediction of immunotherapeutic outcome in non-small-cell lung cancer based on circulating cytokine signatures". In: *Journal for Immunotherapy of Cancer* 11.7 (2023) (cit. on p. 37).

[158] M. Goyal et al. "Computational intelligence technique for prediction of multiple sclerosis based on serum cytokines". In: *Frontiers in neurology* 10 (2019), p. 781 (cit. on p. 37).

[159] S. Cabaro et al. "Cytokine signature and COVID-19 prediction models in the two waves of pandemics". In: *Scientific Reports* 11.1 (2021), p. 20793 (cit. on p. 37).

[160] E. K. Pauwels and G. J. Boer. "Friends and foes in Alzheimer's disease". In: *Medical Principles and Practice* 32.6 (2023), pp. 313–322 (cit. on p. 37).

[161] M. Tejeda et al. "DNA from multiple viral species is associated with Alzheimer's disease risk". In: *Alzheimer's & Dementia* 20.1 (2024), pp. 253–265 (cit. on p. 38).

[162] Y. Wang et al. "Integrative Multi-omics Analysis to Characterize Herpes Virus Infection Increases the Risk of Alzheimer's Disease". In: *Molecular Neurobiology* (2024), pp. 1–16 (cit. on p. 38).

[163] D.-i. Jang et al. "The role of tumor necrosis factor alpha (TNF-α) in autoimmune disease and current TNF-α inhibitors in therapeutics". In: *International journal of molecular sciences* 22.5 (2021), p. 2719 (cit. on p. 38).

[164] N. K. Prabha et al. "TNFipred: a classification model to predict TNF-α inhibitors". In: *Molecular Diversity* (2023), pp. 1–11 (cit. on p. 38).

[165] N. Yoosuf et al. "Early prediction of clinical response to anti-TNF treatment using multi-omics and machine learning in rheumatoid arthritis". In: *Rheumatology* 61.4 (2022), pp. 1680–1689 (cit. on p. 38).

[166] B. Prasad et al. "ATRPred: A machine learning based tool for clinical decision making of anti-TNF treatment in rheumatoid arthritis patients". In: *PLOS Computational Biology* 18.7 (2022), e1010204 (cit. on p. 38).

[167] Y. Guan et al. "Machine learning to predict anti–tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers". In: *Arthritis & Rheumatology* 71.12 (2019), pp. 1987–1996 (cit. on p. 38).

[168] A. Dhall et al. "TNFepitope: A webserver for the prediction of TNF-α inducing epitopes". In: *Computers in Biology and Medicine* 160 (2023), p. 106929 (cit. on p. 38).

[169] P. Charoenkwan et al. "StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides". In: *Briefings in bioinformatics* 22.6 (2021), bbab172 (cit. on p. 38).

[170] Y.-h. Liao et al. "UsIL-6: An unbalanced learning strategy for identifying IL-6 inducing peptides by undersampling technique". In: *Computer Methods and Programs in Biomedicine* 250 (2024), p. 108176 (cit. on p. 38).

[171] O. Singh, W.-L. Hsu, and E. C.-Y. Su. "ILeukin10Pred: a computational approach for predicting IL-10-inducing immunosuppressive peptides using combinations of amino acid global features". In: *Biology* 11.1 (2021), p. 5 (cit. on p. 39).

[172] M. Selig et al. "Prediction of six macrophage phenotypes and their IL-10 content based on single-cell morphology using artificial intelligence". In: *Frontiers in Immunology* 14 (2024), p. 1336393 (cit. on p. 39).

[173] J. Domínguez-Andrés et al. "In vitro induction of trained immunity in adherent human monocytes". In: *STAR protocols* 2.1 (2021), p. 100365 (cit. on pp. 41, 42).

[174] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on p. 43).

[175] A. Samat et al. "Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles". In: *Remote Sensing* 12.12 (2020), p. 1973 (cit. on p. 53).

[176] N. Del Buono, F. Esposito, and L. Selicato. "Methods for hyperparameters optimization in learning approaches: an overview". In: *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I 6*. Springer. 2020, pp. 100–112 (cit. on p. 55).

[177] F. Maleki et al. "Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment". In: *Neuroimaging Clinics* 30.4 (2020), pp. 433–445 (cit. on p. 55).

[178] G. A. Pradipta et al. "SMOTE for handling imbalanced data problem: A review". In: *2021 sixth international conference on informatics and computing (ICIC)*. IEEE. 2021, pp. 1–8 (cit. on p. 56).

[179] A. Fernández et al. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary". In: *Journal of artificial intelligence research* 61 (2018), pp. 863–905 (cit. on p. 56).

[180] P. Branco, L. Torgo, and R. P. Ribeiro. "SMOGN: a pre-processing approach for imbalanced regression". In: *First international workshop on learning with imbalanced domains: Theory and applications*. PMLR. 2017, pp. 36–50 (cit. on pp. 56, 57).

[181] L. Torgo et al. "Smote for regression". In: *Portuguese conference on artificial intelligence*. Springer. 2013, pp. 378–389 (cit. on p. 56).

[182] L. Camacho, G. Douzas, and F. Bacao. "Geometric SMOTE for regression". In: *Expert Systems with Applications* 193 (2022), p. 116387 (cit. on p. 56).

[183] P.-H. Kuo, Y.-T. Chen, and H.-T. Yau. "SMOGN, MFO, and XGBoost Based Excitation Current Prediction Model for Synchronous Machine." In: *Computer Systems Science & Engineering* 46.3 (2023) (cit. on p. 57).

[184] G. R. Yun and S. H. Bae. "Analysis of incident impact factors and development of SMOGN-DNN model for prediction of incident clearance time". In: *Journal of the Korea Institute of Intelligent Transportation Systems* 20.4 (2021), pp. 46–56 (cit. on p. 57).

[185] D. H. Rudd, H. Huo, and G. Xu. "Predicting Financial Literacy via Semi-supervised Learning". In: *Australasian Joint Conference on Artificial Intelligence*. Springer. 2022, pp. 304–319 (cit. on p. 57).

[186] A. Pertiwi et al. "Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data". In: *Journal of Physics: Conference Series*. Vol. 1524. 1. IOP Publishing. 2020, p. 012048 (cit. on p. 57).

[187] J. Thorn. "The inflammatory response in humans after inhalation of bacterial endotoxin: a review". In: *Inflammation Research* 50 (2001), pp. 254–261 (cit. on p. 100).

[188] L. Speciale et al. "Lymphocyte subset patterns and cytokine production in Alzheimer's disease patients". In: *Neurobiology of aging* 28.8 (2007), pp. 1163–1169 (cit. on pp. 101, 102, 110).

[189] D. Lio et al. "Interleukin-10 promoter polymorphism in sporadic Alzheimer's disease". In: *Genes & Immunity* 4.3 (2003), pp. 234–238 (cit. on p. 101).

[190] R. A. Whittington, E. Planel, and N. Terrando. "Impaired resolution of inflammation in Alzheimer's disease: a review". In: *Frontiers in immunology* 8 (2017), p. 1464 (cit. on pp. 102, 110).

[191] P. Bossu et al. "Interleukin-18 produced by peripheral blood cells is increased in Alzheimer's disease and correlates with cognitive impairment". In: *Brain, behavior, and immunity* 22.4 (2008), pp. 487–492 (cit. on p. 102).

[192] J. Ojala et al. "Expression of interleukin-18 is increased in the brains of Alzheimer's disease patients". In: *Neurobiology of aging* 30.2 (2009), pp. 198–209 (cit. on p. 102).

[193] J.-M. Chen et al. "Increased serum levels of interleukin-18,-23 and-17 in chinese patients with Alzheimer's disease". In: *Dementia and geriatric cognitive disorders* 38.5-6 (2014), pp. 321–329 (cit. on p. 102).

[194] D. Gezen-Ak et al. "BDNF, TNFα, HSP90, CFH, and IL-10 serum levels in patients with early or late onset Alzheimer's disease or mild cognitive impairment". In: *Journal of Alzheimer's Disease* 37.1 (2013), pp. 185–195 (cit. on pp. 102, 110).

[195] I. Mateo et al. "Low serum VEGF levels are associated with Alzheimer's disease". In: *Acta Neurologica Scandinavica* 116.1 (2007), pp. 56–58 (cit. on p. 103).

[196] L. Huang, J. Jia, and R. Liu. "Decreased serum levels of the angiogenic factors VEGF and TGF-β1 in Alzheimer's disease and amnestic mild cognitive impairment". In: *Neuroscience letters* 550 (2013), pp. 60–63 (cit. on p. 103).

[197] X. Li and W. Wang. "Levels of IL-6 in peripheral blood and cerebrospinal fluid of Alzheimer's disease: a meta-analysis". In: *receptor (IL-6R)* 5.2 (2023), pp. 53–60 (cit. on pp. 103, 110).

[198] V. K. Khemka et al. "Raised serum proinflammatory cytokines in Alzheimer's disease with depression". In: *Aging and disease* 5.3 (2014), p. 170 (cit. on p. 103).

[199] O. V. Forlenza et al. "Increased serum IL-1β level in Alzheimer's disease and mild cognitive impairment". In: *Dementia and geriatric cognitive disorders* 28.6 (2010), pp. 507–512 (cit. on p. 103).

[200] D. Scarabino et al. "Relationship between proinflammatory cytokines (Il-1beta, Il-18) and leukocyte telomere length in mild cognitive impairment and Alzheimer's disease". In: *Experimental Gerontology* 136 (2020), p. 110945 (cit. on pp. 103, 105, 110).

[201] E. Mombelli et al. "Effect of a probiotic administration on inflammatory profile and clinical features in patients with Alzheimer's disease: Human/Human trials: Nutraceuticals and non-pharmacological interventions". In: *Alzheimer's & Dementia* 16 (2020), e042737 (cit. on pp. 103, 105, 110).

[202] J. Doroszkiewicz et al. "The cerebrospinal fluid interleukin 8 (IL-8) concentration in Alzheimer's disease (AD)". In: *Alzheimer's & Dementia* 17 (2021), e051317 (cit. on p. 103).

[203] Y. Zhu et al. "Serum IL-8 is a marker of white-matter hyperintensities in patients with Alzheimer's disease". In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 7 (2017), pp. 41–47 (cit. on p. 103).

[204] A.-S. Lanzrein et al. "Longitudinal Study of Inflammatory Factors in Serum, Cerebrospinal Fluid, and Brain Tissue in Alzheimer Disease: Interleukin-1β, Interleukin-6, Interleukin-1 Receptor Antagonist, Tumor Necrosis Factor-a, the Soluble Tumor Necrosis Factor Receptors I and II, and α: 1:-Antichymotrypsin". In: *Alzheimer Disease & Associated Disorders* 12.3 (1998), pp. 215–227 (cit. on p. 103).

[205] P. Italiani et al. "Circulating levels of IL-1 family cytokines and receptors in Alzheimer's disease: new markers of disease progression?" In: *Journal of neuroinflammation* 15 (2018), pp. 1–12 (cit. on p. 103).

[206] M. A. O'Neal. "Women and the risk of Alzheimer's disease". In: *Frontiers in Global Women's Health* 4 (2024), p. 1324522 (cit. on p. 104).

[207] J. N. Gabhann-Dromgoole et al. "Systemic IL-1β production as a consequence of corneal HSV-1 infection-contribution to the development of herpes simplex keratitis". In: *International Journal of Ophthalmology* 12.9 (2019), p. 1493 (cit. on p. 105).

[208] L. Álvarez-Rodríguez et al. "Aging is associated with circulating cytokine dysregulation". In: *Cellular immunology* 273.2 (2012), pp. 124–132 (cit. on p. 106).

[209] M. Michaud et al. "Proinflammatory cytokines, aging, and age-related diseases". In: *Journal of the American Medical Directors Association* 14.12 (2013), pp. 877–882 (cit. on p. 106).

[210] R. Roubenoff et al. "Monocyte cytokine production in an elderly population: effect of age and inflammation". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 53.1 (1998), pp. M20–M26 (cit. on p. 106).

[211] Y. Stephan et al. "The prospective relationship between subjective aging and inflammation: evidence from the health and retirement study". In: *Psychophysiology* 60.2 (2023), e14177 (cit. on p. 106).

[212] M. Ciaccio et al. "COVID-19 and Alzheimer's disease". In: *Brain sciences* 11.3 (2021), p. 305 (cit. on p. 110).

I

Annex 1

I.1 Predicting TI Cytokine Levels (Regression) - Extensive Results

Table I.1: LOOCV results for predicting TNFα NT. Best results (models with Pearson's correlation coefficient over 0.3) are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
TNFα NT	0.65 - 27.93	Original data	32	Linear Regression	8.65	9.06	5.93	-0.238
				Lasso Regression	6.87	7.89	5.09	-0.467
				Ridge Regression	7.25	7.89	5.12	-0.456
				EN Regression	6.95	7.85	5.13	-0.469
				DT	6.22	8.21	6.97	0.169
				RF	7.64	8.05	5.18	-0.106
				SVC	5.23	7.28	6.91	-0.622
				KNN	7.63	8.81	5.48	-0.572
				XGB	6.01	6.87	6	0.156
		Oversampled data	50	Linear Regression	7.610	8.820	5.738	-0.030
				Lasso Regression	7.204	8.126	4.525	0.067
				Ridge Regression	7.528	8.203	4.494	0.032
				EN Regression	7.432	8.356	4.587	-0.062
				DT	5.756	7.853	6.971	0.341
				RF	5.299	7.047	4.547	0.465
				SVC	8.970	9.379	4.757	0.034
				KNN	7.230	7.821	4.223	0.391
				XGB	5.792	6.280	4.599	0.509

Table I.2: LOOCV results for predicting TNFα Pr LPS. Best results (models with Pearson's correlation coefficient over 0.3) are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
TNFα Pr LPS	1.75 - 2705.91	Original data	26	Linear Regression	635.68	735.08	477.60	-0.054
				Lasso Regression	527.75	584.53	356.58	0.220
				Ridge Regression	555.55	561.77	372.69	0.251
				EN Regression	517.03	568.42	357.64	0.245
				DT	663.32	698.02	566.39	-0.078
				RF	593.29	631.95	401.75	-0.062
				SVC	354.53	472.57	426.35	0.490
				KNN	461.12	541.97	417.24	0.220
				XGB	595.39	626.87	419.95	-0.197
		Oversampled data	42	Linear Regression	601.051	612.179	384.221	0.276
				Lasso Regression	558.073	565.385	323.425	0.394
				Ridge Regression	506.609	553.707	356.425	0.384
				EN Regression	559.262	562.384	341.597	0.376
				DT	446.970	539.935	372.489	0.402
				RF	499.975	527.603	323.645	0.470
				SVC	400.095	498.338	404.123	0.499
				KNN	381.148	417.565	390.504	0.593
				XGB	487.131	562.824	368.308	0.388

Table I.3: LOOCV results for predicting TNFα Pr Ca. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
TNFα Pr Ca	1.75 - 30.00	Original data	27	Linear Regression	9.114	9.757	5.477	-0.077
				Lasso Regression	7.984	7.862	4.799	-0.210
				Ridge Regression	6.965	7.573	4.574	-0.710
				EN Regression	6.965	7.559	4.557	-0.684
				DT	10.451	10.557	7.045	-0.536
				RF	8.533	8.706	5.223	-0.467
				SVC	6.793	8.242	5.687	-0.556
				KNN	8.865	8.689	5.338	-0.416
				XGB	8.384	9.393	6.828	-0.430

Table I.4: LOOCV results for predicting TNFα Ca LPS. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
TNFα Ca LPS	92.12 - 556.73	Original data	16	Linear Regression	226.070	265.911	171.354	-0.494
				Lasso Regression	129.311	143.507	112.000	-0.653
				Ridge Regression	162.687	145.182	98.127	-0.477
				EN Regression	146.639	162.826	121.553	-0.516
				DT	144.307	167.361	108.539	-0.519
				RF	127.756	135.875	109.868	-0.612
				SVC	87.085	118.937	108.611	-0.312
				KNN	119.076	129.564	103.262	-0.606
				XGB	118.169	152.076	134.481	-0.447

Table I.5: LOOCV results for predicting TNFα LPS LPS.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
TNFα LPS LPS	44.5 - 741.82	Original data	34	Linear Regression	118.13	168.29	171.8	-0.263
				Lasso Regression	127.85	149.43	127.32	-0.408
				Ridge Regression	103.29	138.16	132.81	-0.129
				EN Regression	103.66	138.74	132.57	-0.148
				DT	176.22	198.19	165.11	-0.39
				RF	129.46	162.21	135.96	-0.342
				SVC	87.57	127.38	150.05	-0.531
				KNN	103.56	149.19	134.04	-0.326
				XGB	89.6	147.2	157.04	-0.002
		Oversampled data	51	Linear Regression	111.194	153.783	136.146	-0.071
				Lasso Regression	120.064	152.542	126.316	-0.251
				Ridge Regression	127.667	154.586	123.999	-0.264
				EN Regression	118.498	153.453	122.742	-0.273
				DT	114.771	148.773	117.245	0.185
				RF	118.964	144.472	118.130	0.154
				SVC	135.381	153.644	115.826	-0.024
				KNN	95.899	138.735	125.749	0.129
				XGB	118.899	142.195	113.989	0.122

Table I.6: LOOCV results for predicting IL-6 NT. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-6 NT	2.51 - 461.71	Original data	28	Linear Regression	1.620	2.753	3.121	-0.390
				Lasso Regression	1.007	1.464	2.266	-0.988
				Ridge Regression	1.009	1.458	2.265	-0.982
				EN Regression	1.009	1.457	2.264	-0.982
				DT	1.024	1.805	2.439	-0.174
				RF	1.209	1.992	2.603	-0.340
				SVC	1.064	1.584	2.214	-0.259
				KNN	1.005	1.673	2.228	-0.314
				XGB	0.786	1.464	2.384	-0.215
		Oversampled data	42	Linear Regression	102.306	114.343	82.445	-0.292
				Lasso Regression	65.805	90.188	77.412	0.102
				Ridge Regression	64.443	89.409	77.339	0.126
				EN Regression	63.850	88.972	77.802	0.124
				DT	55.887	85.374	83.379	0.215
				RF	61.976	85.667	74.293	0.320
				SVC	63.228	88.800	79.752	0.053
				KNN	86.062	104.097	84.105	-0.205
				XGB	78.635	88.206	79.977	0.120

Table I.7: LOOCV results for predicting IL-6 Pr LPS. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-6 Pr LPS	76.04 - 2739.48	Original data	19	Linear Regression	648.76	919.42	606.57	-0.004
				Lasso Regression	525.20	632.48	383.71	0.162
				Ridge Regression	508.42	569.12	412.49	0.188
				EN Regression	530.61	611.93	386.67	0.163
				DT	395.83	565.32	515.00	0.198
				RF	520.27	530.42	410.05	0.316
				SVC	541.51	619.64	457.34	-0.160
				KNN	409.91	587.18	476.08	-0.023
				XGB	574.06	642.20	438.10	-0.288
		Oversampled data	28	Linear Regression	579.981	690.337	468.385	0.235
				Lasso Regression	556.362	619.662	372.411	0.233
				Ridge Regression	552.323	647.175	387.805	0.130
				EN Regression	574.427	561.996	385.963	0.309
				DT	503.159	582.317	440.738	0.252
				RF	445.829	516.217	386.301	0.436
				SVC	579.538	554.884	344.372	0.413
				KNN	457.781	511.057	428.634	0.384
				XGB	474.427	604.114	467.628	0.154

Table I.8: LOOCV results for predicting IL-6 Pr Ca. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-6 Pr Ca	5.02 - 250.31	Original data	27	Linear Regression	72.461	72.442	45.927	-0.354
				Lasso Regression	43.250	52.241	35.879	0.019
				Ridge Regression	43.016	51.097	32.934	0.164
				EN Regression	43.120	51.000	32.823	0.170
				DT	59.705	66.516	44.567	-0.163
				RF	54.303	58.714	33.088	-0.239
				SVC	64.842	65.893	40.384	-0.365
				KNN	47.866	54.654	36.944	-0.071
				XGB	40.706	51.479	45.787	0.043

Table I.9: LOOCV results for predicting IL-6 Ca LPS. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-6 Ca LPS	36.93 - 640.01	Original data	13	Linear Regression	297.691	356.216	252.109	-0.3434
				Lasso Regression	169.701	177.850	148.305	-0.6828
				Ridge Regression	166.885	178.742	132.182	-0.6265
				EN Regression	164.044	177.774	162.845	-0.8623
				DT	177.987	209.170	135.405	-0.2923
				RF	148.900	161.069	129.644	-0.8481
				SVC	91.434	150.937	134.869	-0.3471
				KNN	139.553	166.644	129.202	-0.4486
				XGB	138.322	152.532	116.867	-0.5566

I.2 Predicting TI Cytokine Levels (Classification) - Extensive Results

Table I.24: LOOCV results for predicting TNFα Pr LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
TNFα Pr LPS	Original data	26	Logistic Regression	23.08%	22.22%	23.33%	22.17%
			Logistic Lasso	19.23%	19.44%	20.83%	20.10%
			Logistic Ridge	26.92%	27.78%	26.67%	26.95%
			Logistic EN	11.54%	13.89%	12.50%	13.10%
			DT	42.31%	42.80%	43.33%	42.94%
			RF	26.92%	27.04%	27.50%	26.88%
			SVC	7.69%	11.43%	8.33%	9.57%
			KNN	26.92%	24.44%	29.17%	26.58%
			XGB	11.54%	14.29%	12.50%	13.33%
	Oversampled	133	Logistic Regression	30.8%	27.9%	32.5%	28.9%
			Logistic Lasso	34.6%	25.4%	37.5%	29.9%
			Logistic Ridge	34.6%	28.5%	37.5%	32.4%
			Logistic EN	38.5%	28.7%	41.7%	33.9%

Table I.10: LOOCV results for predicting IL-6 LPS LPS. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-6 LPS LPS	16.17 - 498.24	Original data	29	Linear Regression	88.88	99.86	56.95	0.338
				Lasso Regression	89.85	109.34	74.14	-0.043
				Ridge Regression	91.68	108.57	75.47	0.001
				EN Regression	91.92	108.45	75.22	-0.063
				DT	114.2	120.05	104.97	-0.151
				RF	80.41	96.79	75.87	-0.207
				SVC	57.01	86.42	84.09	-0.122
				KNN	80.69	93.71	78.79	-0.069
				XGB	90.29	103.95	91.94	-0.208
		Oversampled data	45	Linear Regression	82.693	90.457	54.319	0.363
				Lasso Regression	70.663	87.604	61.340	0.296
				Ridge Regression	78.195	88.117	68.236	0.194
				EN Regression	79.650	90.919	60.681	0.270
				DT	50.898	81.674	72.770	0.359
				RF	71.456	85.738	62.148	0.365
				SVC	92.606	88.582	67.892	0.306
				KNN	92.024	96.842	68.274	0.263
				XGB	83.779	92.098	66.233	0.178

Table I.11: LOOCV results for predicting IL-10 NT.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-10 NT	0.10 - 12.15	Original data	22	Linear Regression	116.084	121.946	87.367	-0.198
				Lasso Regression	68.221	91.143	82.143	-0.068
				Ridge Regression	60.640	88.860	81.438	0.012
				EN Regression	59.881	89.320	82.488	-0.029
				DT	70.230	106.372	92.819	0.038
				RF	69.071	103.386	83.317	-0.270
				SVC	41.822	80.889	100.943	-0.061
				KNN	68.756	97.998	90.973	-0.345
				XGB	48.843	86.885	105.034	-0.176
		Oversampled data	36	Linear Regression	0.9508	1.6916	2.2052	-0.1235
				Lasso Regression	1.1555	1.4602	2.1696	-0.3292
				Ridge Regression	1.0420	1.5052	2.0818	-1.0000
				EN Regression	1.0420	1.5052	2.0818	-1.0000
				DT	0.7699	1.5468	2.2542	0.0673
				RF	0.8847	1.4858	2.0239	0.0572
				SVC	0.8739	1.4355	2.0954	0.0571
				KNN	1.4489	1.7643	2.0091	-0.1238
				XGB	0.9679	1.4392	2.0879	-0.4160

Table I.12: LOOCV results for predicting IL-10 Pr LPS. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-10 Pr LPS	0.98 - 68.47	Original data	16	Linear Regression	38.116	44.544	30.325	-0.4472
				Lasso Regression	22.576	23.642	10.575	-0.1254
				Ridge Regression	27.410	25.332	10.723	-0.5153
				EN Regression	27.410	25.377	10.864	-0.5246
				DT	37.337	31.209	16.907	-0.3206
				RF	23.305	25.285	12.092	-0.4956
				SVC	21.517	23.832	14.628	-0.6499
				KNN	27.751	26.482	11.219	-0.5323
				XGB	26.943	24.892	12.702	-0.5958

Table I.13: LOOCV results for predicting IL-10 Pr Ca. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-10 Pr Ca	0 - 6.05	Original data	22	Linear Regression	1.111	1.539	1.319	-0.178
				Lasso Regression	0.882	1.036	0.929	-0.406
				Ridge Regression	0.857	1.031	0.923	-0.361
				EN Regression	0.853	1.019	0.927	-0.331
				DT	0.789	1.314	1.287	-0.050
				RF	0.779	0.997	1.050	-0.350
				SVC	0.698	0.917	0.934	0.173
				KNN	0.559	0.988	1.101	-0.371
				XGB	0.659	1.010	0.989	0.136
		Oversampled data	36	Linear Regression	0.9292	1.1712	1.0085	-0.0764
				Lasso Regression	0.9757	1.0960	0.9818	-0.0428
				Ridge Regression	1.0371	1.1029	1.0190	-0.0804
				EN Regression	1.0008	1.0845	0.9670	-0.0112
				DT	0.7012	0.8993	0.8162	0.3592
				RF	0.7626	0.9222	0.8425	0.3006
				SVC	1.2058	1.2249	0.9562	-0.0724
				KNN	0.4875	0.9012	0.9282	0.3024
				XGB	0.6678	0.9222	0.8878	0.2337

Table I.14: LOOCV results for predicting IL-10 Ca LPS. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-10 Ca LPS	0.32 - 13.64	Original data	16	Linear Regression	4.582	5.042	3.203	-0.037
				Lasso Regression	2.692	3.332	2.390	-0.051
				Ridge Regression	1.999	2.742	2.644	-0.086
				EN Regression	1.864	2.611	2.693	0.012
				DT	1.784	2.936	2.611	0.306
				RF	1.695	2.748	2.660	0.045
				SVC	2.085	2.962	2.990	-0.111
				KNN	1.425	2.541	2.680	0.186
				XGB	0.766	2.383	3.017	0.210
		Oversampled data	28	Linear Regression	5.835	5.293	2.740	0.310
				Lasso Regression	2.140	2.874	1.949	0.483
				Ridge Regression	4.002	3.884	2.556	0.342
				EN Regression	2.804	3.164	2.088	0.449
				DT	1.895	2.330	2.133	0.632
				RF	2.133	2.413	2.117	0.657
				SVC	3.703	3.844	2.522	0.264
				KNN	1.417	2.370	2.668	0.567
				XGB	1.815	2.481	2.140	0.612

Table I.15: LOOCV results for predicting IL-10 LPS LPS. Best results are highlighted in green. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-10 LPS LPS	0.67 - 11.71	Original data	20	Linear Regression	6.24	6.44	2.87	-0.066
				Lasso Regression	4.8	4.39	2.16	0.175
				Ridge Regression	4.56	4.55	2.7	-0.308
				EN Regression	4.39	4.44	2.47	-0.177
				DT	4.05	4.97	3.2	-0.149
				RF	4.22	4.46	2.23	-0.039
				SVC	4.08	4.84	2.88	-0.438
				KNN	4.05	3.69	2.44	0.331
				XGB	5.08	5.07	3.29	-0.075

Table I.16: LOOCV results for predicting IL-1β NT. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1β NT	0.25 - 35.43	Original data	32	Linear Regression	5.986	7.307	6.465	-0.218
				Lasso Regression	4.570	5.547	5.675	0.054
				Ridge Regression	4.449	5.488	5.689	0.063
				EN Regression	4.422	5.476	5.693	0.071
				DT	4.338	6.571	6.168	0.063
				RF	4.645	6.026	5.946	-0.180
				SVC	2.694	5.409	6.983	-0.456
				KNN	5.129	6.316	6.159	-0.042
				XGB	2.283	5.874	7.272	-0.139
		Oversampled data	51	Linear Regression	6.719	6.682	5.847	-0.117
				Lasso Regression	4.832	5.732	5.387	0.140
				Ridge Regression	4.802	5.688	5.364	0.163
				EN Regression	4.815	5.676	5.367	0.164
				DT	4.783	6.309	5.345	0.172
				RF	3.814	5.229	5.086	0.400
				SVC	4.750	5.803	5.417	0.158
				KNN	5.629	6.477	5.690	0.051
				XGB	4.090	5.421	5.366	0.253

Table I.17: LOOCV results for predicting IL-1β Pr LPS. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1β Pr LPS	0.90 - 34.61	Original data	32	Linear Regression	5.689	6.912	5.974	0.156
				Lasso Regression	3.876	5.659	5.563	0.295
				Ridge Regression	3.289	5.364	5.562	0.345
				EN Regression	3.469	5.501	5.545	0.322
				DT	3.721	6.384	6.716	0.214
				RF	4.447	6.115	5.795	0.126
				SVC	3.823	5.370	5.665	0.325
				KNN	4.304	5.944	5.995	0.187
				XGB	4.249	6.691	7.016	0.007
		Oversampled data	51	Linear Regression	5.036	6.383	5.173	0.333
				Lasso Regression	3.690	6.048	5.280	0.319
				Ridge Regression	3.746	5.802	5.007	0.411
				EN Regression	3.803	5.898	5.112	0.375
				DT	4.303	5.687	5.470	0.408
				RF	4.197	5.547	4.892	0.451
				SVC	4.002	6.281	4.920	0.373
				KNN	4.271	5.912	5.639	0.296
				XGB	6.176	6.016	4.707	0.423

Table I.18: LOOCV results for predicting IL-1β Pr Ca. Best results are highlighted in green.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1β Pr Ca	0.42 - 24.12	Original data	32	Linear Regression	3.485	4.261	3.876	-0.107
				Lasso Regression	2.599	3.149	3.143	0.196
				Ridge Regression	2.568	3.120	3.146	0.210
				EN Regression	2.498	3.096	3.162	0.213
				DT	3.252	3.702	3.378	0.126
				RF	3.215	3.651	3.449	-0.117
				SVC	2.790	3.556	3.503	-0.125
				KNN	2.615	3.343	3.278	0.105
				XGB	2.011	3.345	4.176	0.109
		Oversampled data	54	Linear Regression	3.438	3.674	3.323	0.113
				Lasso Regression	2.500	3.253	3.035	0.256
				Ridge Regression	2.448	3.243	3.029	0.260
				EN Regression	2.457	3.238	3.026	0.263
				DT	2.362	3.275	3.327	0.290
				RF	2.286	2.827	2.843	0.474
				SVC	2.910	3.552	2.758	0.265
				KNN	2.761	3.198	2.932	0.333
				XGB	2.962	3.372	3.147	0.193

Table I.19: LOOCV results for predicting IL-1RA NT. Best results are highlighted in green. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1RA NT	1095.66 - 6293.75	Original data	29	Linear Regression	1754.201	1619.610	1126.908	-0.669
				Lasso Regression	1024.752	1143.689	802.774	-0.993
				Ridge Regression	1063.517	1180.766	824.766	-0.900
				EN Regression	1048.523	1171.517	817.439	-0.931
				DT	954.298	1456.023	1046.716	0.122
				RF	1239.353	1278.665	885.787	-0.327
				SVC	1019.006	1105.108	806.398	-0.347
				KNN	983.123	1147.909	948.943	0.046
				XGB	1075.939	1186.211	702.391	0.313

Table I.20: LOOCV results for predicting IL-1RA Pr LPS. Best results are highlighted in green. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1RA Pr LPS	2013.46 - 6365.57	Original data	29	Linear Regression	1132.246	1291.767	834.213	-0.315
				Lasso Regression	1051.477	1097.747	863.824	-0.571
				Ridge Regression	1166.393	1082.069	781.527	-0.345
				EN Regression	1093.629	1057.416	792.259	-0.414
				DT	1050.779	1226.796	849.503	0.071
				RF	882.237	992.138	786.357	0.116
				SVC	762.788	952.740	787.156	-0.675
				KNN	881.976	1078.148	787.296	-0.110
				XGB	488.807	764.151	667.981	0.573

Table I.21: LOOCV results for predicting IL-1RA Pr Ca. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1RA Pr LPS	1066.64 - 5947.04	Original data	29	Linear Regression	1490.768	1578.273	1190.021	-0.583
				Lasso Regression	1056.634	1180.267	768.745	-0.748
				Ridge Regression	1190.101	1256.343	722.669	-0.393
				EN Regression	1110.247	1209.580	740.364	-0.634
				DT	1399.137	1604.703	1131.916	-0.021
				RF	1516.146	1395.430	801.603	-0.489
				SVC	992.084	1297.438	805.311	-0.731
				KNN	662.050	1041.820	1002.278	0.096
				XGB	1142.263	1274.336	838.113	0.197

Table I.22: LOOCV results for predicting IL-1RA Ca LPS. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1RA Ca LPS	984.26 - 5406.42	Original data	17	Linear Regression	1768.325	1780.153	1089.770	-0.334
				Lasso Regression	919.432	1114.834	759.477	-0.830
				Ridge Regression	1347.913	1254.559	750.005	-0.507
				EN Regression	1300.228	1216.615	745.775	-0.492
				DT	1243.899	1519.255	1011.073	-0.338
				RF	1248.145	1226.016	737.501	-0.686
				SVC	770.175	1075.270	809.204	-0.756
				KNN	1100.696	1135.757	747.961	-0.666
				XGB	1619.831	1453.425	992.146	-0.355

Table I.23: LOOCV results for predicting IL-1RA LPS LPS. Best results are highlighted in green. Oversampling via SMOGN was not possible.

	Range	Data utilized	Sample Size	Model	Median absolute error	Mean absolute error	Standard Deviation	Pearson's coefficient
IL-1RA LPS LPS	987.22 - 4996.25	Original data	30	Linear Regression	777.045	810.859	593.764	0.376
				Lasso Regression	777.973	849.390	556.958	0.206
				Ridge Regression	973.505	902.822	592.555	0.139
				EN Regression	790.698	871.784	618.618	0.142
				DT	1142.707	1150.310	683.731	-0.035
				RF	812.266	859.287	527.055	0.216
				SVC	756.525	949.103	536.076	-0.436
				KNN	703.342	793.184	546.157	0.300
				XGB	710.654	878.008	684.186	0.200

Table I.25: LOOCV results for predicting TNFα LPS LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
TNFα LPS LPS	Original data	31	Logistic Regression	12.90%	12.59%	12.73%	12.64%
			Logistic Lasso	3.23%	3.03%	3.03%	3.03%
			Logistic Ridge	6.45%	6.11%	6.36%	6.23%
			Logistic EN	3.23%	2.38%	3.03%	2.67%
			DT	48.39%	51.59%	48.48%	48.82%
			RF	19.35%	17.14%	19.09%	18.00%
			SVC	12.90%	11.43%	12.73%	11.90%
			KNN	35.48%	35.48%	35.76%	35.35%
			XGB	6.45%	6.55%	6.36%	6.37%
	Oversampled data	33	Logistic Regression	16.1%	16.1%	16.1%	16.0%
			Logistic Lasso	9.7%	8.1%	10.0%	8.9%
			Logistic Ridge	16.1%	16.3%	16.1%	16.1%
			Logistic EN	9.7%	8.5%	10.0%	9.1%
			DT	48.4%	49.0%	48.8%	48.4%
			RF	22.6%	18.6%	22.7%	20.2%
			SVC	16.1%	16.5%	16.4%	16.4%
			KNN	32.3%	29.1%	33.0%	30.6%
			XGB	9.7%	8.6%	10.0%	9.2%

Table I.26: LOOCV results for predicting IL-6 Pr LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-6 Pr LPS	Original data	19	Logistic Regression	36.84%	37.78%	36.51%	36.74%
			Logistic Lasso	31.58%	32.54%	32.54%	32.54%
			Logistic Ridge	36.84%	38.33%	37.30%	37.68%
			Logistic EN	31.58%	32.38%	32.54%	32.27%
			DT	42.11%	29.17%	44.44%	35.12%
			RF	21.05%	21.43%	21.43%	21.43%
			SVC	26.32%	23.33%	27.78%	25.11%
			KNN	47.37%	43.45%	49.21%	45.25%
			XGB	36.84%	38.33%	37.30%	37.68%
	Oversampled data	21	Logistic Regression	31.6%	32.8%	31.7%	32.1%
			Logistic Lasso	36.8%	32.4%	38.1%	34.0%
			Logistic Ridge	47.4%	45.3%	48.4%	46.2%
			Logistic EN	36.8%	35.6%	38.1%	36.3%
			DT	26.3%	20.6%	27.8%	23.6%
			RF	36.8%	35.1%	38.1%	35.7%
			SVC	47.4%	43.5%	49.2%	44.9%
			KNN	57.9%	38.5%	61.1%	47.2%
			XGB	36.8%	35.6%	38.1%	36.3%

Table I.27: LOOCV results for predicting IL-6 Ca LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-6 Ca LPS	Original data	31	Logistic Regression	36.84%	37.78%	36.51%	36.74%
			Logistic Lasso	31.58%	32.54%	32.54%	32.54%
			Logistic Ridge	36.84%	38.33%	37.30%	37.68%
			Logistic EN	31.58%	32.38%	32.54%	32.27%
			DT	42.11%	29.17%	44.44%	35.12%
			RF	21.05%	21.43%	21.43%	21.43%
			SVC	26.32%	23.33%	27.78%	25.11%
			KNN	47.37%	43.45%	49.21%	45.25%
			XGB	36.84%	38.33%	37.30%	37.68%
	Oversampled data	33	Logistic Regression	31.6%	32.8%	31.7%	32.1%
			Logistic Lasso	36.8%	32.4%	38.1%	34.0%
			Logistic Ridge	47.4%	45.3%	48.4%	46.2%
			Logistic EN	36.8%	35.6%	38.1%	36.3%
			DT	26.3%	20.6%	27.8%	23.6%
			RF	36.8%	35.1%	38.1%	35.7%
			SVC	47.4%	43.5%	49.2%	44.9%
			KNN	57.9%	38.5%	61.1%	47.2%
			XGB	36.8%	35.6%	38.1%	36.3%

Table I.28: LOOCV results for predicting IL-6 LPS LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-6 LPS LPS	Original data	26	Logistic Regression	42.31%	43.49%	43.49%	43.49%
			Logistic Lasso	46.15%	48.89%	47.20%	47.88%
			Logistic Ridge	46.15%	46.30%	47.20%	46.63%
			Logistic EN	42.31%	45.45%	43.86%	44.37%
			DT	30.77%	32.85%	30.63%	31.02%
			RF	23.08%	22.22%	22.91%	22.12%
			SVC	23.08%	19.44%	25.40%	22.02%
			KNN	26.92%	19.81%	29.10%	23.33%
			XGB	42.31%	42.59%	43.86%	43.12%
	Oversampled data	30	Logistic Regression	50.0%	50.5%	52.0%	51.1%
			Logistic Lasso	46.2%	46.3%	47.2%	46.4%
			Logistic Ridge	50.0%	50.9%	52.0%	50.8%
			Logistic EN	50.0%	50.9%	52.0%	50.8%
			DT	30.8%	28.5%	30.3%	29.1%
			RF	50.0%	51.7%	49.5%	49.9%
			SVC	38.5%	40.6%	40.2%	40.2%
			KNN	42.3%	60.4%	45.7%	38.6%
			XGB	46.2%	46.7%	48.6%	46.3%

Table I.29: LOOCV results for predicting IL-10 Pr LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-10 Pr LPS	Original data	16	Logistic Regression	31.25%	30.56%	30.00%	30.13%
			Logistic Lasso	18.75%	13.33%	18.89%	14.95%
			Logistic Ridge	18.75%	9.09%	16.67%	11.76%
			Logistic EN	6.25%	3.33%	6.67%	4.44%
			DT	50.00%	49.29%	50.00%	49.46%
			RF	37.50%	34.29%	36.67%	35.38%
			SVC	6.25%	4.17%	5.56%	4.76%
			KNN	6.25%	5.56%	5.56%	5.56%
			XGB	12.50%	20.83%	12.22%	14.29%
	Oversampled data	18	Logistic Regression	31.3%	33.3%	31.1%	31.5%
			Logistic Lasso	25.0%	25.0%	25.6%	24.8%
			Logistic Ridge	18.8%	18.9%	18.9%	18.8%
			Logistic EN	18.8%	19.4%	18.9%	18.8%
			DT	56.3%	58.3%	55.6%	56.1%
			RF	56.3%	56.7%	56.7%	56.7%
			SVC	25.0%	24.4%	25.6%	24.8%
			KNN	31.3%	31.2%	32.2%	31.1%
			XGB	31.3%	35.2%	32.2%	31.7%

Table I.30: LOOCV results for predicting IL-10 LPS LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-10 LPS LPS	Original data	17	Logistic Regression	52.94%	56.48%	52.38%	53.33%
			Logistic Lasso	52.94%	52.38%	49.60%	49.40%
			Logistic Ridge	29.41%	26.92%	24.60%	21.67%
			Logistic EN	35.29%	32.78%	32.94%	31.88%
			DT	41.18%	39.29%	39.29%	39.29%
			RF	35.29%	26.67%	30.16%	27.81%
			SVC	47.06%	47.62%	50.40%	46.90%
			KNN	52.94%	55.56%	55.16%	52.59%
			XGB	35.29%	34.34%	32.94%	31.75%
	Oversampled data	21	Logistic Regression	58.8%	61.7%	60.7%	60.0%
			Logistic Lasso	52.9%	55.6%	55.2%	52.6%
			Logistic Ridge	47.1%	48.1%	49.6%	45.0%
			Logistic EN	47.1%	48.1%	49.6%	45.0%
			DT	47.1%	47.6%	50.4%	46.9%
			RF	47.1%	48.1%	54.0%	46.0%
			SVC	58.8%	61.7%	64.3%	58.6%
			KNN	52.9%	55.6%	59.5%	52.2%
			XGB	52.9%	55.6%	55.2%	52.6%

Table I.31: LOOCV results for predicting IL-1β NT through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1β NT	Original data	29	Logistic Regression	31.03%	17.65%	16.67%	17.14%
			Logistic Lasso	62.07%	20.69%	33.33%	25.53%
			Logistic Ridge	62.07%	20.69%	33.33%	25.53%
			Logistic EN	62.07%	20.69%	33.33%	25.53%
			DT	48.28%	34.66%	34.66%	34.66%
			RF	55.17%	19.75%	29.63%	23.70%
			SVC	44.83%	18.06%	24.07%	20.63%
			KNN	48.28%	18.67%	25.93%	21.71%
			XGB	62.07%	20.69%	33.33%	25.53%
	Oversampled data	54	Logistic Regression	31.0%	32.7%	41.9%	31.9%
			Logistic Lasso	31.0%	32.7%	41.9%	31.9%
			Logistic Ridge	31.0%	33.3%	41.9%	32.7%
			Logistic EN	31.0%	33.3%	41.9%	32.7%
			DT	51.7%	52.2%	56.0%	52.7%
			RF	65.5%	59.5%	66.3%	61.0%
			SVC	44.8%	43.4%	49.3%	42.9%
			KNN	44.8%	47.9%	52.2%	45.5%
			XGB	37.9%	39.4%	45.6%	37.6%

Table I.32: LOOCV results for predicting IL-1β Pr LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1β Pr LPS	Original data	29	Logistic Regression	24.14%	22.69%	22.54%	22.57%
			Logistic Lasso	17.24%	16.16%	16.43%	16.27%
			Logistic Ridge	34.48%	30.95%	31.43%	31.04%
			Logistic EN	31.03%	31.09%	30.63%	30.67%
			DT	17.24%	16.89%	16.98%	16.91%
			RF	27.59%	21.76%	23.89%	22.22%
			SVC	20.69%	19.91%	20.32%	20.08%
			KNN	31.03%	31.37%	33.17%	30.77%
			XGB	20.69%	14.33%	17.22%	14.92%
	Oversampled data	36	Logistic Regression	31.0%	32.1%	34.0%	30.6%
			Logistic Lasso	31.0%	31.1%	34.0%	30.7%
			Logistic Ridge	37.9%	38.9%	40.2%	37.9%
			Logistic EN	31.0%	31.0%	34.0%	31.1%
			DT	34.5%	34.4%	35.4%	34.6%
			RF	44.8%	44.0%	45.4%	44.2%
			SVC	31.0%	30.7%	34.6%	31.4%
			KNN	27.6%	26.0%	30.4%	27.3%
			XGB	31.0%	30.4%	35.2%	30.1%

Table I.33: LOOCV results for predicting IL-1RA NT through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1RA NT	Original data	26	Logistic Regression	42.31%	33.10%	32.64%	32.82%
			Logistic Lasso	57.69%	20.00%	31.25%	24.39%
			Logistic Ridge	57.69%	20.00%	31.25%	24.39%
			Logistic EN	61.54%	20.51%	33.33%	25.40%
			DT	42.31%	36.39%	38.89%	36.85%
			RF	53.85%	19.44%	29.17%	23.33%
			SVC	53.85%	19.44%	29.17%	23.33%
			KNN	53.85%	19.44%	29.17%	23.33%
			XGB	61.54%	20.51%	33.33%	25.40%
	Oversampled data	48	Logistic Regression	38.5%	38.2%	40.3%	38.0%
			Logistic Lasso	34.6%	32.4%	35.4%	32.0%
			Logistic Ridge	30.8%	28.5%	29.9%	28.1%
			Logistic EN	30.8%	29.5%	29.9%	28.5%
			DT	50.0%	48.7%	56.3%	49.5%
			RF	53.8%	56.7%	65.3%	54.1%
			SVC	34.6%	38.5%	38.9%	33.2%
			KNN	34.6%	41.7%	38.9%	32.2%
			XGB	38.5%	36.7%	37.5%	34.8%

Table I.34: LOOCV results for predicting IL-1RA Pr LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1RA Pr LPS	Original data	26	Logistic Regression	46.15%	43.43%	43.43%	43.43%
			Logistic Lasso	38.46%	28.54%	32.32%	30.29%
			Logistic Ridge	42.31%	37.78%	37.88%	37.48%
			Logistic EN	23.08%	13.33%	18.18%	15.38%
			DT	46.15%	46.19%	43.43%	43.21%
			RF	30.77%	35.24%	28.11%	28.68%
			SVC	26.92%	26.52%	25.08%	25.45%
			KNN	15.38%	13.10%	13.47%	12.81%
			XGB	30.77%	26.79%	27.44%	26.59%
	Oversampled data	33	Logistic Regression	50.0%	49.5%	50.3%	49.6%
			Logistic Lasso	42.3%	42.4%	43.6%	42.8%
			Logistic Ridge	42.3%	42.4%	43.6%	42.8%
			Logistic EN	42.3%	42.4%	43.6%	42.8%
			DT	38.5%	38.3%	41.8%	39.5%
			RF	42.3%	42.5%	44.8%	43.3%
			SVC	50.0%	50.8%	52.2%	50.6%
			KNN	30.8%	28.8%	36.4%	30.4%
			XGB	46.2%	47.2%	49.2%	47.9%

Table I.35: LOOCV results for predicting IL-1RA Pr Ca through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1RA Pr Ca	Original data	26	Logistic Regression	30.77%	32.05%	27.62%	29.63%
			Logistic Lasso	46.15%	16.67%	28.57%	21.05%
			Logistic Ridge	53.85%	17.95%	33.33%	23.33%
			Logistic EN	46.15%	16.67%	28.57%	21.05%
			DT	26.92%	17.50%	19.05%	18.24%
			RF	46.15%	16.67%	28.57%	21.05%
			SVC	34.62%	15.79%	21.43%	18.18%
			KNN	30.77%	14.04%	19.05%	16.16%
			XGB	50.00%	17.33%	30.95%	22.22%
	Oversampled data	42	Logistic Regression	30.8%	31.0%	43.3%	34.0%
			Logistic Lasso	30.8%	33.3%	41.0%	34.2%
			Logistic Ridge	30.8%	35.3%	41.0%	35.6%
			Logistic EN	30.8%	34.6%	38.6%	35.0%
			DT	46.2%	48.0%	49.0%	46.0%
			RF	38.5%	43.8%	52.9%	39.2%
			SVC	50.0%	56.3%	55.7%	50.7%
			KNN	30.8%	37.5%	41.4%	31.1%
			XGB	30.8%	33.3%	41.0%	34.2%

Table I.36: LOOCV results for predicting IL-1RA Ca LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1RA Ca LPS	Original data	14	Logistic Regression	21.43%	19.44%	21.67%	20.45%
			Logistic Lasso	14.29%	12.04%	13.33%	12.17%
			Logistic Ridge	21.43%	14.29%	20.00%	16.67%
			Logistic EN	14.29%	12.04%	13.33%	12.17%
			DT	21.43%	17.78%	21.67%	19.53%
			RF	0.00%	0.00%	0.00%	0.00%
			SVC	50.00%	50.00%	51.67%	50.59%
			KNN	35.71%	27.78%	38.33%	32.12%
			XGB	0.00%	0.00%	0.00%	0.00%
	Oversampled data	15	Logistic Regression	28.6%	30.0%	30.0%	30.0%
			Logistic Lasso	7.1%	11.1%	8.3%	9.5%
			Logistic Ridge	21.4%	22.2%	21.7%	21.6%
			Logistic EN	0.0%	0.0%	0.0%	0.0%
			DT	35.7%	33.3%	35.0%	33.9%
			RF	7.1%	8.3%	8.3%	8.3%
			SVC	57.1%	70.0%	60.0%	53.3%
			KNN	35.7%	27.6%	38.3%	31.5%
			XGB	7.1%	8.3%	8.3%	8.3%

Table I.37: LOOCV results for predicting IL-1RA LPS LPS through classification. Best results are highlighted in green.

	Data utilized	Sample Size	Model	Accuracy	Precision	Recall	F1 Score
IL-1RA LPS LPS	Original data	27	Logistic Regression	37.04%	33.89%	33.94%	33.86%
			Logistic Lasso	25.93%	17.08%	21.21%	18.69%
			Logistic Ridge	37.04%	26.71%	30.30%	28.38%
			Logistic EN	33.33%	23.08%	27.27%	25.00%
			DT	44.44%	30.71%	36.36%	33.10%
			RF	33.33%	25.24%	27.27%	26.03%
			SVC	25.93%	18.89%	21.21%	19.78%
			KNN	33.33%	21.19%	27.27%	22.46%
			XGB	33.33%	24.15%	27.27%	25.60%
	Oversampled data	33	Logistic Regression	44.4%	42.6%	43.6%	42.8%
			Logistic Lasso	33.3%	33.2%	34.5%	33.8%
			Logistic Ridge	44.4%	44.7%	43.6%	43.9%
			Logistic EN	33.3%	34.3%	34.5%	34.2%
			DT	25.9%	31.2%	28.5%	29.5%
			RF	37.0%	37.5%	41.2%	38.8%
			SVC	29.6%	29.3%	31.5%	29.6%
			KNN	40.7%	40.7%	44.2%	41.7%
			XGB	40.7%	40.7%	47.9%	42.4%

I.3 Predicting AD from IB - Extensive Results

Table I.38: LOOCV results for predicting AD from IB data.

	Model	Accuracy	Precision	Recall	F1 score	AUC	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)
Original data	Logistic Regression	48.6%	45.3%	45.8%	44.9%	0.458	54.2%	65.0%	59.1%
	Logistic Lasso	57.1%	55.4%	55.0%	54.8%	0.550	60.9%	70.0%	65.1%
	Logistic Ridge	57.1%	54.6%	53.3%	51.4%	0.533	59.3%	80.0%	68.1%
	Logistic EN	57.1%	55.0%	54.2%	53.3%	0.542	60.0%	75.0%	66.7%
	DT	45.7%	42.8%	43.3%	42.7%	0.433	52.2%	60.0%	55.8%
	RF	62.9%	61.9%	61.7%	61.7%	0.617	66.7%	70.0%	68.3%
	SVC	48.6%	45.3%	45.8%	44.9%	0.458	54.2%	65.0%	59.1%
	KNN	51.4%	49.1%	49.2%	48.8%	0.492	56.5%	65.0%	60.5%
	XGB	51.4%	50.8%	50.8%	50.8%	0.508	57.9%	55.0%	56.4%
Oversampled data	Logistic Regression	57.1%	58.3%	58.3%	57.1%	0.583	66.7%	50.0%	57.1%
	Logistic Lasso	54.3%	56.0%	55.8%	54.2%	0.558	64.3%	45.0%	52.9%
	Logistic Ridge	51.4%	52.5%	52.5%	51.4%	0.525	60.0%	45.0%	51.4%
	Logistic EN	54.3%	54.9%	55.0%	54.2%	0.550	62.5%	50.0%	55.6%
	DT	57.1%	57.4%	57.5%	57.0%	0.575	64.7%	55.0%	59.5%
	RF	57.1%	57.4%	57.5%	57.0%	0.575	64.7%	55.0%	59.5%
	SVC	45.7%	45.9%	45.8%	45.5%	0.458	52.9%	45.0%	48.6%
	KNN	60.0%	60.7%	60.8%	60.0%	0.608	68.8%	55.0%	61.1%
	XGB	57.1%	57.4%	57.5%	57.0%	0.575	64.7%	55.0%	59.5%

I.4 Predicting AD from TI cytokines - Extensive Results

Table I.39: Test set results of mean filled models for predicting AD from TI data.

	Model	Accuracy	Precision	Recall	F1 score	AUC	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	62.5%	63.3%	62.5%	61.9%	0.44	60.0%	75.0%	66.7%	100.0%
	Logistic Lasso	75.0%	83.3%	75.0%	73.3%	0.69	66.7%	100.0%	80.0%	100.0%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.44	60.0%	75.0%	66.7%	100.0%
	Logistic EN	75.0%	83.3%	75.0%	73.3%	0.81	66.7%	100.0%	80.0%	86.7%
	DT	62.5%	63.3%	62.5%	61.9%	0.44	66.7%	50.0%	57.1%	96.7%
	RF	75.0%	75.0%	75.0%	75.0%	0.59	75.0%	75.0%	75.0%	100.0%
	SVC	75.0%	75.0%	75.0%	75.0%	0.75	75.0%	75.0%	75.0%	100.0%
	KNN	50.0%	50.0%	50.0%	46.7%	0.50	50.0%	75.0%	60.0%	80.0%
	XGB	62.5%	63.3%	62.5%	61.9%	0.75	60.0%	75.0%	66.7%	100.0%
Oversampled data	Logistic Regression	62.5%	63.3%	62.5%	61.9%	0.44	60.0%	75.0%	66.7%	100.0%
	Logistic Lasso	87.5%	90.0%	87.5%	87.3%	0.81	80.0%	100.0%	88.9%	87.5%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.63	60.0%	75.0%	66.7%	96.9%
	Logistic EN	87.5%	90.0%	87.5%	87.3%	0.81	80.0%	100.0%	88.9%	87.5%
	DT	62.5%	63.3%	62.5%	61.9%	0.63	66.7%	50.0%	57.1%	100.0%
	RF	62.5%	63.3%	62.5%	61.9%	0.88	60.0%	75.0%	66.7%	100.0%
	SVC	62.5%	63.3%	62.5%	61.9%	0.63	60.0%	75.0%	66.7%	87.5%
	KNN	62.5%	78.6%	62.5%	56.4%	0.50	100.0%	25.0%	40.0%	81.3%
	XGB	87.5%	90.0%	87.5%	87.3%	0.88	100.0%	75.0%	85.7%	100.0%

Table I.40: Test set results of median filled models for predicting AD from TI data.

	Model	Accuracy	Precision	Recall	F1 score	AUC	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	62.5%	63.3%	62.5%	61.9%	0.50	60.0%	75.0%	66.7%	100.0%
	Logistic Lasso	50.0%	50.0%	50.0%	50.0%	0.75	50.0%	50.0%	50.0%	93.3%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.44	60.0%	75.0%	66.7%	100.0%
	Logistic EN	87.5%	90.0%	87.5%	87.3%	0.88	80.0%	100.0%	88.9%	83.3%
	DT	37.5%	36.7%	37.5%	36.5%	0.31	40.0%	50.0%	44.4%	93.3%
	RF	62.5%	63.3%	62.5%	61.9%	0.63	60.0%	75.0%	66.7%	100.0%
	SVC	37.5%	36.7%	37.5%	36.5%	0.25	40.0%	50.0%	44.4%	100.0%
	KNN	12.5%	10.0%	12.5%	11.1%	0.25	20.0%	25.0%	22.2%	100.0%
	XGB	75.0%	75.0%	75.0%	75.0%	0.69	75.0%	75.0%	75.0%	100.0%
Oversampled data	Logistic Regression	50.0%	50.0%	50.0%	50.0%	0.44	50.0%	50.0%	50.0%	100.0%
	Logistic Lasso	50.0%	50.0%	50.0%	50.0%	0.63	50.0%	50.0%	50.0%	93.8%
	Logistic Ridge	50.0%	50.0%	50.0%	50.0%	0.56	50.0%	50.0%	50.0%	93.8%
	Logistic EN	75.0%	75.0%	75.0%	75.0%	0.88	75.0%	75.0%	75.0%	81.3%
	DT	62.5%	63.3%	62.5%	61.9%	0.63	66.7%	50.0%	57.1%	100.0%
	RF	75.0%	75.0%	75.0%	75.0%	0.63	75.0%	75.0%	75.0%	100.0%
	SVC	50.0%	50.0%	50.0%	46.7%	0.63	50.0%	25.0%	33.3%	93.8%
	KNN	50.0%	50.0%	50.0%	46.7%	0.50	50.0%	25.0%	33.3%	100.0%
	XGB	75.0%	75.0%	75.0%	75.0%	0.81	75.0%	75.0%	75.0%	100.0%

Table I.41: Test set results of mode filled models for predicting AD from TI data.

	Model	Accuracy	Precision	Recall	F1 score	AUC	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	50.0%	50.0%	50.0%	46.7%	0.44	50.0%	25.0%	33.3%	100.0%
	Logistic Lasso	62.5%	63.3%	62.5%	61.9%	0.50	66.7%	50.0%	57.1%	96.7%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	96.7%
	Logistic EN	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	96.7%
	DT	62.5%	78.6%	62.5%	56.4%	0.69	100.0%	25.0%	40.0%	83.3%
	RF	75.0%	75.0%	75.0%	75.0%	0.69	75.0%	75.0%	75.0%	100.0%
	SVC	50.0%	50.0%	50.0%	46.7%	0.56	50.0%	25.0%	33.3%	100.0%
	KNN	50.0%	50.0%	50.0%	50.0%	0.75	50.0%	50.0%	50.0%	100.0%
	XGB	62.5%	63.3%	62.5%	61.9%	0.75	66.7%	50.0%	57.1%	96.7%
Oversampled data	Logistic Regression	50.0%	50.0%	50.0%	46.7%	0.44	50.0%	25.0%	33.3%	100.0%
	Logistic Lasso	62.5%	63.3%	62.5%	61.9%	0.50	66.7%	50.0%	57.1%	96.9%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	96.9%
	Logistic EN	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	96.9%
	DT	62.5%	78.6%	62.5%	56.4%	0.69	100.0%	25.0%	40.0%	84.4%
	RF	75.0%	75.0%	75.0%	75.0%	0.56	75.0%	75.0%	75.0%	100.0%
	SVC	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	96.9%
	KNN	50.0%	50.0%	50.0%	50.0%	0.69	50.0%	50.0%	50.0%	100.0%
	XGB	50.0%	50.0%	50.0%	46.7%	0.63	50.0%	25.0%	33.3%	100.0%

I.5 Predicting AD from IB and TI - Extensive Results

Table I.42: Test set results of mean filled models for predicting AD from IB and TI data.

	Model	Accuracy	Precision	Recall	F1 score	AUC	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	50.0%	50.0%	50.0%	50.0%	0.44	50.0%	50.0%	50.0%	100.0%
	Logistic Lasso	37.5%	36.7%	37.5%	36.5%	0.56	33.3%	25.0%	28.6%	100.0%
	Logistic Ridge	50.0%	50.0%	50.0%	50.0%	0.38	50.0%	50.0%	50.0%	100.0%
	Logistic EN	75.0%	83.3%	75.0%	73.3%	0.81	66.7%	100.0%	80.0%	90.0%
	DT	87.5%	90.0%	87.5%	87.3%	0.88	80.0%	100.0%	88.9%	100.0%
	RF	50.0%	50.0%	50.0%	50.0%	0.47	50.0%	50.0%	50.0%	100.0%
	SVC	50.0%	50.0%	50.0%	50.0%	0.44	50.0%	50.0%	50.0%	100.0%
	KNN	50.0%	50.0%	50.0%	46.7%	0.44	50.0%	25.0%	33.3%	100.0%
	XGB	50.0%	50.0%	50.0%	46.7%	0.69	50.0%	25.0%	33.3%	100.0%
Oversampled data	Logistic Regression	50.0%	50.0%	50.0%	50.0%	0.44	50.0%	50.0%	50.0%	100.0%
	Logistic Lasso	37.5%	36.7%	37.5%	36.5%	0.50	33.3%	25.0%	28.6%	100.0%
	Logistic Ridge	50.0%	50.0%	50.0%	50.0%	0.56	50.0%	50.0%	50.0%	100.0%
	Logistic EN	62.5%	63.3%	62.5%	61.9%	0.69	66.7%	50.0%	57.1%	100.0%
	DT	87.5%	90.0%	87.5%	87.3%	0.88	80.0%	100.0%	88.9%	100.0%
	RF	50.0%	50.0%	50.0%	50.0%	0.69	50.0%	50.0%	50.0%	100.0%
	SVC	50.0%	50.0%	50.0%	50.0%	0.44	50.0%	50.0%	50.0%	100.0%
	KNN	50.0%	50.0%	50.0%	50.0%	0.56	50.0%	50.0%	50.0%	100.0%
	XGB	62.5%	78.6%	62.5%	56.4%	0.75	100.0%	25.0%	40.0%	100.0%

Table I.43: Test set results of median filled models for predicting AD from IB and TI data.

	Model	Accuracy	Precision	Recall	F1 score	AUC	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	62.5%	63.3%	62.5%	61.9%	0.63	66.7%	50.0%	57.1%	100.0%
	Logistic Lasso	62.5%	63.3%	62.5%	61.9%	0.63	66.7%	50.0%	57.1%	100.0%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.69	66.7%	50.0%	57.1%	96.7%
	Logistic EN	75.0%	83.3%	75.0%	73.3%	0.88	100.0%	50.0%	66.7%	90.0%
	DT	87.5%	90.0%	87.5%	87.3%	0.88	80.0%	100.0%	88.9%	100.0%
	RF	75.0%	75.0%	75.0%	75.0%	0.66	75.0%	75.0%	75.0%	100.0%
	SVC	50.0%	50.0%	50.0%	50.0%	0.69	50.0%	50.0%	50.0%	90.0%
	KNN	62.5%	78.6%	62.5%	56.4%	0.56	100.0%	25.0%	40.0%	100.0%
	XGB	75.0%	83.3%	75.0%	73.3%	1.00	100.0%	50.0%	66.7%	100.0%
Oversampled data	Logistic Regression	62.5%	63.3%	62.5%	61.9%	0.63	66.7%	50.0%	57.1%	100.0%
	Logistic Lasso	62.5%	63.3%	62.5%	61.9%	0.75	60.0%	75.0%	66.7%	90.6%
	Logistic Ridge	62.5%	63.3%	62.5%	61.9%	0.63	66.7%	50.0%	57.1%	96.9%
	Logistic EN	75.0%	83.3%	75.0%	73.3%	0.88	100.0%	50.0%	66.7%	90.6%
	DT	87.5%	90.0%	87.5%	87.3%	0.81	80.0%	100.0%	88.9%	90.6%
	RF	50.0%	50.0%	50.0%	46.7%	0.56	50.0%	25.0%	33.3%	100.0%
	SVC	50.0%	50.0%	50.0%	50.0%	0.69	50.0%	50.0%	50.0%	90.6%
	KNN	62.5%	63.3%	62.5%	61.9%	0.59	66.7%	50.0%	57.1%	87.5%
	XGB	75.0%	83.3%	75.0%	73.3%	0.63	100.0%	50.0%	66.7%	100.0%

Table I.44: Test set results of mode filled models for predicting AD from IB and TI data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	50.0%	50.0%	50.0%	46.7%	0.50	50.0%	25.0%	33.3%	100.0%
	Logistic Lasso	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	100.0%
	Logistic Ridge	50.0%	50.0%	50.0%	50.0%	0.69	50.0%	50.0%	50.0%	83.3%
	Logistic EN	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	93.3%
	DT	50.0%	50.0%	50.0%	50.0%	0.50	50.0%	50.0%	50.0%	100.0%
	RF	62.5%	63.3%	62.5%	61.9%	0.69	66.7%	50.0%	57.1%	100.0%
	SVC	62.5%	63.3%	62.5%	61.9%	0.56	66.7%	50.0%	57.1%	90.0%
	KNN	62.5%	63.3%	62.5%	61.9%	0.75	66.7%	50.0%	57.1%	80.0%
	XGB	62.5%	63.3%	62.5%	61.9%	0.69	66.7%	50.0%	57.1%	100.0%
Oversampled data	Logistic Regression	50.0%	50.0%	50.0%	46.7%	0.50	50.0%	25.0%	33.3%	100.0%
	Logistic Lasso	50.0%	50.0%	50.0%	46.7%	0.56	50.0%	25.0%	33.3%	100.0%
	Logistic Ridge	50.0%	50.0%	50.0%	46.7%	0.50	50.0%	25.0%	33.3%	100.0%
	Logistic EN	50.0%	50.0%	50.0%	46.7%	0.56	50.0%	25.0%	33.3%	93.8%
	DT	37.5%	36.7%	37.5%	36.5%	0.50	33.3%	25.0%	28.6%	81.3%
	RF	75.0%	83.3%	75.0%	73.3%	0.75	100.0%	50.0%	66.7%	100.0%
	SVC	62.5%	63.3%	62.5%	61.9%	0.81	66.7%	50.0%	57.1%	100.0%
	KNN	62.5%	78.6%	62.5%	56.4%	0.94	100.0%	25.0%	40.0%	100.0%
	XGB	37.5%	21.4%	37.5%	27.3%	0.50	0.0%	0.0%	0.0%	100.0%

I.6 Predicting AD from Serum Cytokines - Extensive Results

Table I.45: Test set results of mean filled models for predicting AD from serum data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	65.0%	65.2%	65.0%	64.9%	0.730	63.6%	70.0%	66.7%	69.6%
	Logistic Lasso	65.0%	65.2%	65.0%	64.9%	0.710	63.6%	70.0%	66.7%	69.6%
	Logistic Ridge	65.0%	65.2%	65.0%	64.9%	0.760	63.6%	70.0%	66.7%	68.4%
	Logistic EN	65.0%	65.2%	65.0%	64.9%	0.710	63.6%	70.0%	66.7%	69.6%
	DT	50.0%	50.0%	50.0%	49.5%	0.460	50.0%	40.0%	44.4%	88.6%
	RF	60.0%	60.4%	60.0%	59.6%	0.570	62.5%	50.0%	55.6%	100.0%
	SVC	75.0%	75.3%	75.0%	74.9%	0.740	72.7%	80.0%	76.2%	74.7%
	KNN	75.0%	77.5%	75.0%	74.4%	0.795	85.7%	60.0%	70.6%	74.7%
	XGB	35.0%	34.8%	35.0%	34.8%	0.470	33.3%	30.0%	31.6%	100.0%
Oversampled data	Logistic Regression	75.0%	75.3%	75.0%	74.9%	0.790	72.7%	80.0%	76.2%	68.3%
	Logistic Lasso	60.0%	60.4%	60.0%	59.6%	0.710	58.3%	70.0%	63.6%	69.5%
	Logistic Ridge	65.0%	65.2%	65.0%	64.9%	0.770	63.6%	70.0%	66.7%	68.3%
	Logistic EN	70.0%	70.8%	70.0%	69.7%	0.790	66.7%	80.0%	72.7%	67.1%
	DT	55.0%	55.1%	55.0%	54.9%	0.590	55.6%	50.0%	52.6%	89.0%
	RF	65.0%	66.5%	65.0%	64.2%	0.660	71.4%	50.0%	58.8%	100.0%
	SVC	65.0%	65.2%	65.0%	64.9%	0.760	63.6%	70.0%	66.7%	87.8%
	KNN	75.0%	77.5%	75.0%	74.4%	0.795	85.7%	60.0%	70.6%	76.8%
	XGB	40.0%	40.0%	40.0%	40.0%	0.450	40.0%	40.0%	40.0%	98.8%

Table I.46: Test set results of median filled models for predicting AD from serum data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	65.0%	65.2%	65.0%	64.9%	0.720	63.6%	70.0%	66.7%	64.6%
	Logistic Lasso	65.0%	65.2%	65.0%	64.9%	0.740	63.6%	70.0%	66.7%	63.3%
	Logistic Ridge	65.0%	65.2%	65.0%	64.9%	0.710	63.6%	70.0%	66.7%	64.6%
	Logistic EN	70.0%	70.0%	70.0%	70.0%	0.750	70.0%	70.0%	70.0%	65.8%
	DT	50.0%	50.0%	50.0%	47.9%	0.500	50.0%	30.0%	37.5%	97.5%
	RF	60.0%	60.4%	60.0%	59.6%	0.630	62.5%	50.0%	55.6%	100.0%
	SVC	70.0%	70.8%	70.0%	69.7%	0.770	66.7%	80.0%	72.7%	73.4%
	KNN	75.0%	75.3%	75.0%	74.9%	0.855	77.8%	70.0%	73.7%	68.4%
	XGB	45.0%	44.5%	45.0%	43.7%	0.440	42.9%	30.0%	35.3%	100.0%
Oversampled data	Logistic Regression	65.0%	65.2%	65.0%	64.9%	0.720	63.6%	70.0%	66.7%	67.1%
	Logistic Lasso	65.0%	65.2%	65.0%	64.9%	0.690	63.6%	70.0%	66.7%	69.5%
	Logistic Ridge	70.0%	70.8%	70.0%	69.7%	0.820	66.7%	80.0%	72.7%	67.1%
	Logistic EN	75.0%	75.3%	75.0%	74.9%	0.790	72.7%	80.0%	76.2%	67.1%
	DT	60.0%	60.4%	60.0%	59.6%	0.600	62.5%	50.0%	55.6%	100.0%
	RF	65.0%	65.2%	65.0%	64.9%	0.620	66.7%	60.0%	63.2%	98.8%
	SVC	75.0%	77.5%	75.0%	74.4%	0.890	85.7%	60.0%	70.6%	97.6%
	KNN	75.0%	83.3%	75.0%	73.3%	0.830	100.0%	50.0%	66.7%	74.4%
	XGB	55.0%	55.5%	55.0%	54.0%	0.460	57.1%	40.0%	47.1%	100.0%

Table I.47: Test set results of mode filled models for predicting AD from serum data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	65.0%	66.5%	65.0%	64.2%	0.780	61.5%	80.0%	69.6%	63.3%
	Logistic Lasso	65.0%	66.5%	65.0%	64.2%	0.670	61.5%	80.0%	69.6%	62.0%
	Logistic Ridge	65.0%	66.5%	65.0%	64.2%	0.680	61.5%	80.0%	69.6%	62.0%
	Logistic EN	65.0%	66.5%	65.0%	64.2%	0.680	61.5%	80.0%	69.6%	62.0%
	DT	50.0%	50.0%	50.0%	45.1%	0.500	50.0%	20.0%	28.6%	100.0%
	RF	65.0%	65.2%	65.0%	64.9%	0.680	66.7%	60.0%	63.2%	97.5%
	SVC	80.0%	85.7%	80.0%	79.2%	0.860	100.0%	60.0%	75.0%	97.5%
	KNN	60.0%	61.9%	60.0%	58.3%	0.765	66.7%	40.0%	50.0%	67.1%
	XGB	45.0%	44.5%	45.0%	43.7%	0.470	42.9%	30.0%	35.3%	100.0%
Oversampled data	Logistic Regression	60.0%	60.4%	60.0%	59.6%	0.700	58.3%	70.0%	63.6%	65.9%
	Logistic Lasso	45.0%	44.9%	45.0%	44.9%	0.570	45.5%	50.0%	47.6%	59.8%
	Logistic Ridge	60.0%	60.0%	60.0%	60.0%	0.710	60.0%	60.0%	60.0%	63.4%
	Logistic EN	50.0%	50.0%	50.0%	49.5%	0.590	50.0%	60.0%	54.5%	63.4%
	DT	60.0%	60.0%	60.0%	60.0%	0.670	60.0%	60.0%	60.0%	93.9%
	RF	65.0%	65.2%	65.0%	64.9%	0.680	66.7%	60.0%	63.2%	96.3%
	SVC	80.0%	85.7%	80.0%	79.2%	0.860	100.0%	60.0%	75.0%	97.6%
	KNN	70.0%	81.3%	70.0%	67.0%	0.750	100.0%	40.0%	57.1%	70.7%
	XGB	50.0%	50.0%	50.0%	50.0%	0.490	50.0%	50.0%	50.0%	100.0%

I.7 Predicting Age (Over/Under 65) - Extensive results

Table I.48: Test set results of mean filled models for predicting age group from TI data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic Lasso	87.5%	87.5%	90.0%	87.3%	100.0%	100.0%	80.0%	88.9%	100.0%
	Logistic Ridge	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic EN	87.5%	91.7%	83.3%	85.5%	93.3%	83.3%	100.0%	90.9%	93.3%
	DT	62.5%	62.5%	63.3%	61.9%	63.3%	75.0%	60.0%	66.7%	100.0%
	RF	50.0%	46.7%	46.7%	46.7%	63.3%	60.0%	60.0%	60.0%	100.0%
	SVC	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	KNN	50.0%	46.7%	46.7%	46.7%	56.7%	60.0%	60.0%	60.0%	76.7%
	XGB	62.5%	62.5%	63.3%	61.9%	53.3%	75.0%	60.0%	66.7%	100.0%
Oversampled data	Logistic Regression	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic Lasso	87.5%	87.5%	90.0%	87.3%	100.0%	100.0%	80.0%	88.9%	100.0%
	Logistic Ridge	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic EN	75.0%	80.0%	80.0%	75.0%	86.7%	100.0%	60.0%	75.0%	92.1%
	DT	62.5%	62.5%	63.3%	61.9%	76.7%	75.0%	60.0%	66.7%	97.4%
	RF	50.0%	46.7%	46.7%	46.7%	53.3%	60.0%	60.0%	60.0%	100.0%
	SVC	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	97.4%
	KNN	62.5%	62.5%	63.3%	61.9%	66.7%	75.0%	60.0%	66.7%	100.0%
	XGB	62.5%	62.5%	63.3%	61.9%	50.0%	75.0%	60.0%	66.7%	100.0%

Table I.49: Test set results of median filled models for predicting age group from TI data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic Lasso	87.5%	91.7%	83.3%	85.5%	100.0%	83.3%	100.0%	90.9%	100.0%
	Logistic Ridge	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic EN	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	93.3%
	DT	37.5%	25.0%	30.0%	27.3%	40.0%	50.0%	60.0%	54.5%	90.0%
	RF	50.0%	46.7%	46.7%	46.7%	60.0%	60.0%	60.0%	60.0%	100.0%
	SVC	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	KNN	75.0%	85.7%	66.7%	66.7%	56.7%	71.4%	100.0%	83.3%	100.0%
	XGB	37.5%	25.0%	30.0%	27.3%	33.3%	50.0%	60.0%	54.5%	100.0%
Oversampled data	Logistic Regression	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Logistic Lasso	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	89.5%
	Logistic Ridge	75.0%	80.0%	80.0%	75.0%	100.0%	100.0%	60.0%	75.0%	86.8%
	Logistic EN	87.5%	87.5%	90.0%	87.3%	100.0%	100.0%	80.0%	88.9%	89.5%
	DT	50.0%	46.7%	46.7%	46.7%	46.7%	60.0%	60.0%	60.0%	100.0%
	RF	50.0%	46.7%	46.7%	46.7%	53.3%	60.0%	60.0%	60.0%	94.7%
	SVC	50.0%	46.7%	46.7%	46.7%	60.0%	60.0%	60.0%	60.0%	94.7%
	KNN	87.5%	91.7%	83.3%	85.5%	80.0%	83.3%	100.0%	90.9%	100.0%
	XGB	37.5%	25.0%	30.0%	27.3%	46.7%	50.0%	60.0%	54.5%	100.0%

Table I.50: Test set results of mode filled models for predicting age group from TI data.

	Model	Accuracy	Precision	Recall	F1 score	\gls{AUC}	Precision (Class AD)	Recall (Class AD)	F1 score (Class AD)	Training Accuracy
Original data	Logistic Regression	62.5%	75.0%	70.0%	61.9%	93.3%	100.0%	40.0%	57.1%	100.0%
	Logistic Lasso	87.5%	87.5%	90.0%	87.3%	93.3%	100.0%	80.0%	88.9%	100.0%
	Logistic Ridge	62.5%	31.3%	50.0%	38.5%	13.3%	62.5%	100.0%	76.9%	63.3%
	Logistic EN	87.5%	91.7%	83.3%	85.5%	100.0%	83.3%	100.0%	90.9%	86.7%
	DT	37.5%	25.0%	30.0%	27.3%	40.0%	50.0%	60.0%	54.5%	93.3%
	RF	37.5%	25.0%	30.0%	27.3%	53.3%	50.0%	60.0%	54.5%	100.0%
	SVC	62.5%	31.3%	50.0%	38.5%	80.0%	62.5%	100.0%	76.9%	100.0%
	KNN	62.5%	31.3%	50.0%	38.5%	66.7%	62.5%	100.0%	76.9%	100.0%
	XGB	50.0%	46.7%	46.7%	46.7%	56.7%	60.0%	60.0%	60.0%	90.0%
Oversampled data	Logistic Regression	62.5%	75.0%	70.0%	61.9%	93.3%	100.0%	40.0%	57.1%	100.0%
	Logistic Lasso	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	92.1%
	Logistic Ridge	50.0%	46.7%	46.7%	46.7%	46.7%	60.0%	60.0%	60.0%	78.9%
	Logistic EN	50.0%	46.7%	46.7%	46.7%	46.7%	60.0%	60.0%	60.0%	81.6%
	DT	50.0%	46.7%	46.7%	46.7%	50.0%	60.0%	60.0%	60.0%	92.1%
	RF	37.5%	25.0%	30.0%	27.3%	46.7%	50.0%	60.0%	54.5%	100.0%
	SVC	62.5%	31.3%	50.0%	38.5%	60.0%	62.5%	100.0%	76.9%	100.0%
	KNN	62.5%	58.3%	56.7%	56.4%	66.7%	66.7%	80.0%	72.7%	100.0%
	XGB	50.0%	46.7%	46.7%	46.7%	50.0%	60.0%	60.0%	60.0%	92.1%

UNIVERSIDADE NOVA
DE LISBOA

2024

Leveraging Machine Learning for Predictive Modelling in Alzheimer's Disease

ANA MARTA RODRIGUES PEREIRA DA COSTA

MASTER IN **COMPUTATIONAL BIOLOGY & BIOINFORMATICS**

SPECIALIZATION Biosystems Simulation for Life and Health

NOVA UNIVERSIDADE NOVA DE LISBOA