MACHINE LEARNING-BASED CLASSIFICATION & FEATURE IDENTIFICATION FOR PSORIASIS

A study submitted in partial fulfilment
of the requirements for the degree of
MSc Data Science

at

THE UNIVERSITY OF SHEFFIELD

by

ROBERT ANDREW JACQUES

240175267

Word Count: 13,993 August 2025

Acknowledgements

I would like to express my sincere gratitude to Dr Zeyneb Kurt, my supervisor, for her guidance, constructive feedback and encouragement throughout the course of this research. Her expertise and patience have been invaluable to both the direction and completion of this dissertation.

I would like to thank the academic staff of the School of Information, Journalism and Communication for their teaching and support throughout my studies, which have provided a strong foundation for this work.

Special thanks go to my fellow *MSc Data Science* students, particularly Brandon Rawson and Peter Hounsell, for their helpful discussions and moral support during challenging stages of the project, as well as for their camaraderie and encouragement throughout the course.

I am deeply grateful to my family for their unwavering encouragement and understanding. Most of all, I wish to thank my partner, Emma, whose patience, kindness and steadfast belief in me have been my greatest source of strength throughout this work.

Abstract

Background: Psoriasis is a chronic, immune-mediated inflammatory skin disease characterised by the dysregulation of keratinocyte proliferation and the immune system. The advent of high-throughput transcriptomic technologies, particularly RNA-sequencing (RNA-seq), has provided large-scale expression data crucial for understanding its pathogenesis. Machine Learning (ML) is increasingly utilised to analyse these complex datasets for disease classification and biomarker discovery.

Aims: This study addresses the knowledge gap in comparative ML studies by systematically assessing feature importance across different classification models. The primary goal is to identify a strong, consensus set of genes that are consistently predictive of psoriasis, thereby enhancing the biological interpretability of ML findings.

Methods: We analysed a large, public RNA-seq dataset comprising 92 psoriatic and 82 'Healthy' skin samples. Following data cleaning and pre-processing, three supervised ML models were developed: Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Machine (SVM). Models were optimised using stratified *k*-fold cross-validation. To identify key features, we consolidated rankings from five sources: LR coefficients, RFC native feature importance and SHapley Additive exPlanations (SHAP) values from all three models.

Results: All three models achieved perfect classification performance on the test set (Area Under the Receiver Operating Characteristic Curve [ROC AUC] = 1.0000). The comparative analysis successfully addressed the *multiplicity problem*, powerfully illustrated by the SVM's reliance on a completely divergent feature set despite its perfect accuracy. A quantitative overlap analysis revealed a robust consensus signature of five genes – *BTC*, *CHI3L2*, *LCE3A*, *S100A9* and *SPRR2G* – consistently identified as top features by four different methods.

Conclusion: This study demonstrates that a multi-model, consensus-based approach can effectively identify highly predictive and biologically relevant biomarkers from complex transcriptomic data. The identified gene signature was strongly validated by existing literature, confirming the methodology's power to provide clear, interpretable insights into the molecular mechanisms of psoriasis and serving as a strong proof of concept for future applications in biomarker discovery.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AEβ7	alpha E beta 7
AKR1B10	Aldo-Keto Reductase family 1 member B10
AMP	Antimicrobial Peptide
APA	Alternative Polyadenylation
AS	Alternative Splicing
BET	Bromodomain and Extra-Terminal
BMI	Body Mass Index
BSA	Body Surface Area
BTC	Betacellulin
CHI3L2	Chitinase 3-Like 2
circRNA	circular RNA
CKD	Chronic Kidney Disease
DC	Dendritic Cell
DEG	Differentially Expressed Gene
DNN	Deep Neural Network
DZ	Dizygotic
EDA	Exploratory Data Analysis
ERV	Endogenous Retrovirus
FC	Fold Change
FDR	First-Degree Relative

FPKM	Fragments Per Kilobase of transcript per Million mapped reads
FPR	False Positive Rate
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Study
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
IFN	Interferon
IFN-α	Interferon-alpha
IFN-γ	Interferon-gamma
ILC	Innate Lymphoid Cell
JAK	Janus Kinase
KIR	Killer-cell Immunoglobulin-like Receptor
LASSO	Least Absolute Shrinkage and Selection Operator
LCE	Late Cornified Envelope
LCE3A	Late Cornified Envelope protein 3A
LIME	Local Interpretable Model-agnostic Explanations
LINE	Long Interspersed Nuclear Element
lncRNA	long non-coding RNA
LR	Logistic Regression
mDC	myeloid Dendritic Cell
MHC	Major Histocompatibility Complex

MI	Myocardial Infarction
miRNA	microRNA
ML	Machine Learning
MR	Mendelian Randomisation
MZ	Monozygotic
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NF-κB	Nuclear Factor kappa-light-chain-enhancer of activated B cells
NK	Natural Killer
NN	Neural Network
OR	Odds Ratio
OVA	One-vs-All
OVO	One-vs-One
PASI	Psoriasis Area and Severity Index
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Component Analysis
pDC	plasmacytoid Dendritic Cell
PE	Paired-End
PsA	Psoriatic Arthritis
RAPTOR	Regulatory-Associated Protein of mTOR
RE	Repetitive Element

RFC	Random Forest Classifier
RLR	RIG-I-like Receptor
RNA-seq	RNA-sequencing
ROC AUC	Receiver Operating Characteristic Area Under the Curve
RPKM	Reads Per Kilobase of transcript per Million mapped reads
RR	Risk Ratio
scRNA-seq	single-cell RNA-sequencing
SE	Single-End
SHAP	SHapley Additive exPlanations
siRNA	small interfering RNA
SNP	Single Nucleotide Polymorphism
SPRR	Small Proline-Rich
SPRR2G	Small Proline-Rich protein 2G
SVM	Support Vector Machine
SVM–RFE	Support Vector Machine–Recursive Feature Elimination
S100A9	S100 calcium-binding protein A9
TLR	Toll-Like Receptor
TNF-α	Tumour Necrosis Factor-alpha
TPM	Transcripts Per Million
TPR	True Positive Rate
Treg	regulatory T-cell

tRNA	transfer RNA
TSS	Transcription Start Site
TWAS	Transcriptome-Wide Association Study
TWEAK	Tumour Necrosis Factor-like Weak inducer of apoptosis
VEGF	Vascular Endothelial Growth Factor
WGCNA	Weighted Gene Co-expression Network Analysis
XAI	Explainable Artificial Intelligence

Chapter 1: Introduction

1.1 Background to Psoriasis

Psoriasis, a chronic, immune-mediated inflammatory skin disease, is influenced by a complex interplay of genetic and environmental factors. This condition affects an estimated 60 million people worldwide, with a notable prevalence of 1.52% in the United Kingdom (Raharja et al., 2021). Psoriasis exhibits a clinically heterogenous presentation and is frequently associated with numerous systemic comorbidities. These extracutaneous comorbid diseases typically develop due to the pathogenesis of systemic inflammation. Despite advancements, a comprehensive understanding of its molecular basis remains an ongoing challenge, necessitating sophisticated analytical approaches.

1.2 Advancements in Psoriasis Research and the role of Omics

Recent years have witnessed significant progress in elucidating the molecular mechanisms underlying psoriasis. Genome-wide studies have been instrumental in identifying numerous genes contributing to the disease's development and revealing potential targets for therapeutic intervention. Specifically, Genome-Wide Association Studies (GWASs), conducted across diverse populations, have identified over 80 genetic loci associated with psoriasis susceptibility. Large-scale gene expression studies, including microarray analyses, have further identified numerous Differentially Expressed Genes (DEGs) in lesional and non-lesional psoriatic skin, providing crucial insights into the molecular basis of the disease. Transcriptomics, a powerful approach facilitated by high-throughput RNA-sequencing (RNA-seq), has become essential for comprehensive transcriptome-wide analysis. This technology enables a deeper understanding of molecular biology in inflammatory skin diseases such as psoriasis, and emerging transcriptomic sequencing techniques are further advancing the ability to resolve intercellular transcriptomic heterogeneity, thereby providing granular insights into specific cellular processes.

1.3 Precision Medicine and Machine Learning in Psoriasis

Precision medicine represents a transformative approach that combines data generated by genomics with other omics disciplines, leveraging big data analytics and Machine Learning (ML).

This integration allows for the analysis of an individual's molecular and clinical data against background population-level data. The goal of this individualised approach is to refine the design of precise therapies by accounting for variations in genes, proteins, environment and lifestyle, ultimately improving patient outcomes in psoriatic disease. Furthermore, insights gleaned from large-scale genetic analyses have been a driving force behind the development of transformative novel treatments. ML plays a central role in precision medicine by enabling the identification of complex, potentially non-linear patterns within high-dimensional biological datasets, crucial for making accurate predictions on new, unseen samples.

1.4 Research Aim and Objectives

The overarching aim of this research is to develop and compare supervised ML models for the accurate classification of psoriatic versus 'Healthy' skin samples using publicly available gene expression data (RNA-seq) from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository (NCBI, n.d.); under accession GSE54456 (Li et al., 2014b), and to subsequently explore the key gene features identified by these classifications. Given the clinical heterogeneity and systemic implications of psoriasis, there is a clear need for advanced analytical techniques to better understand the disease's molecular signature. The primary objective of this study, therefore, is to move beyond simple classification and to conduct a robust, comparative analysis of feature importance across different models and interpretability methods. This approach aims to identify a consensus set of genes that are consistently predictive of psoriasis, thereby enhancing the biological interpretability of the ML findings. The advancement of such gene analysis and ML techniques holds significant clinical relevance, offering a means to overcome challenges associated with traditional diagnoses. The methodology employed in this study specifically facilitates this gene discovery process, which is crucial for improving our understanding of psoriasis and developing more reliable disease signatures.

1.5 Research Questions

This research addresses the following key questions:

- **RQ1: How accurately can optimised Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Machine (SVM) models distinguish psoriatic from 'Healthy' skin samples using the GSE54456 RNA-seq gene expression profile?**

- **RQ2: How do the sets of key gene features identified by different models and interpretability methods compare, and what does this reveal about the *multiplicity problem* in a high-dimensional dataset?**

- **RQ3: What is the robust, consensus-based gene signature identified by a multi-model, multi-method comparative analysis, and how is this signature validated by the existing biological and clinical literature on psoriasis?**

1.6 Dissertation Structure

This dissertation is structured as follows: **Chapter 2: Literature Review**, surveys the genetic underpinnings of psoriasis, the application of transcriptomic data analysis and the potential of ML to improve diagnostic and predictive accuracy. **Chapter 3: Methodology**, details the experimental design; data handling and pre-processing; feature selection; ML algorithms; and evaluation strategies. **Chapter 4: Results and Discussion**, presents and interprets the findings from the ML models and feature analyses, directly addressing the research questions. **Chapter 5: Conclusion**, summarises the main findings, outlines the contributions and discusses implications for future research. **References** provides a comprehensive list of all cited sources. **Appendices** include: **Appendix A: Ethics Approval Certificate**; **Appendix B: Supplementary Results – Consolidated Feature Importance Table**; **Appendix C: Supplementary Visualisations – Per-sample Gene Expression Distributions**; **Appendix D: Dissertation Research Diary**; **Appendix E: Reflection on Research** and **Appendix F: Code and Computational Environment**.

Chapter 2: Literature Review

This chapter provides a comprehensive review of the existing literature directly relevant to the study of psoriasis and the application of Machine Learning (ML) in gene expression analysis. Expanding on the foundational background established in Chapter 1, this review aims to contextualise the current research landscape, highlight key methodological advancements and identify the specific knowledge gap addressed by this dissertation. The chapter begins by further exploring the genetic underpinnings and molecular pathophysiology of psoriasis. It then delves into the evolution and application of transcriptomic data analysis, specifically discussing microarray technology and the significant advancements facilitated by RNA-sequencing (RNA-seq), including the importance of the GSE54456 dataset. Subsequently, the chapter examines the growing role of supervised ML in biomedical research, focusing on its application in psoriasis, and reviews core concepts including feature selection and classification algorithms. This comprehensive review thus establishes the theoretical and empirical basis for the study's methodology and the interpretation of its findings.

2.1 Genetic and Molecular Pathophysiology of Psoriasis

Psoriasis, as introduced in Chapter 1, is a chronic, immune-mediated inflammatory disease, well-established as affecting the skin and potentially other organ systems, driven by a complex interplay between genetic predisposition and environmental factors (Bowcock et al., 2001; Raharja et al., 2021). The condition affects an estimated 60 million people worldwide with a prevalence of 1.52% in the United Kingdom (Raharja et al., 2021) and 3.2% in US adults (Armstrong & Read, 2020). Global prevalence varies significantly, with reported rates ranging from as low as 0.05% in Taiwan to 1.88% in Australia, and is often higher in high-income areas and older populations (Raharja et al., 2021). It presents with significant clinical heterogeneity (Svedbom et al., 2021) and is associated with many systemic comorbidities, which develop due to the pathogenesis of systemic inflammation (Bu et al., 2022; Yamazaki, 2021). These comorbidities include:

- **Psoriatic Arthritis (PsA):** Affecting approximately 33% of patients (Bu et al., 2022);
- **Cardiometabolic diseases:** Including Myocardial Infarction (MI), stroke and cardiovascular mortality, with increased risks even after accounting for traditional risk

factors (for severe psoriasis, the Risk Ratio [RR] for MI is 1.70, for stroke is 1.56 and for cardiovascular mortality is 1.39) (Bu et al., 2022);

- **Mental health disorders:** Such as depression (Odds Ratio [OR] 1.57), anxiety (OR 2.91) and suicidal ideation (OR 2.05) (Bu et al., 2022);
- **Inflammatory bowel diseases:** With a four-fold greater prevalence (Bu et al., 2022);
- **Other frequently associated conditions:** Including chronic kidney disease (CKD), malignancy and infections (Bu et al., 2022).

Environmental factors such as skin trauma (the Koebner phenomenon), infections (e.g., *Streptococcus* infection), smoking, certain medications (e.g., lithium, interferon [IFN]) and stress can exacerbate psoriasis (Armstrong & Read, 2020; Henseler, 1997; Raharja et al., 2021; Liu et al., 2007).

Advancements in recent years have profoundly improved the understanding of psoriasis's genetic and molecular mechanisms, with Genome-Wide Association Studies (GWASs) identifying numerous genes contributing to its development and revealing potential targets for therapeutic intervention (Mateu-Arrom & Puig, 2023). Specifically, GWASs have identified numerous susceptibility loci (Ogawa & Okada, 2020). A recent large-sample-size GWAS meta-analysis by Dand et al. (2025) further identified 109 distinct psoriasis susceptibility loci, including 46 newly reported loci, unequivocally emphasising the complex genetic architecture driving the disease and motivating deeper investigation into how these factors translate into altered cellular functions and tissue pathology (Dand et al., 2025).

Drawing from these broader genomic findings, a more detailed understanding of psoriasis's genetic basis reveals it to be a polygenic disease, with multiple identified alleles and loci (Armstrong & Read, 2020; Babaie et al., 2022; Liu et al., 2007). Family studies consistently highlight a strong inherited risk: approximately 40% of patients with psoriasis or PsA have a family history of these conditions in First-Degree Relatives (FDRs) (Solmaz et al., 2020; Babaie et al., 2022). For patients with PsA, the RR of recurrence for PsA in FDRs can be as high as 30–55, and for psoriasis, 4–10 (Solmaz et al., 2020). This inherited risk also shapes disease phenotype, with earlier onset and more frequent nail disease (Solmaz et al., 2020). In addition, patterns differ by family history: a family history of psoriasis (versus PsA) is associated with enthesitis, whereas a

family history of PsA is associated with deformities and a lower risk of plaque psoriasis (Solmaz et al., 2020).

Twin studies further support a strong genetic component, with concordance rates in monozygotic (MZ) twins estimated between 60–90% (Babaie et al., 2022; Roberson & Bowcock, 2010) and heritability estimated at 68% (Ogawa & Okada, 2020). Specifically, one study found a concordance rate of 72% in MZ twins versus just 15–23% in dizygotic (DZ) twins (Liu et al., 2007). Historically, Type I psoriasis (early-onset, inheritable and strongly *HLA-Cw6* associated) is distinguished from Type II (late-onset and less inherited) (Henseler, 1997).

While the *HLA-C* gene (specifically *HLA-Cw*0602*) remains the strongest genetic risk factor (Krueger & Bowcock, 2005; Dand et al., 2020; Liu et al., 2007; Henseler, 1997), a multitude of non-*HLA* genes also contribute to its polygenic nature (Armstrong & Read, 2020; Babaie et al., 2022; Dand et al., 2020). Recent large-scale GWAS meta-analyses have identified numerous distinct susceptibility loci for psoriasis with common Single Nucleotide Polymorphisms (SNPs) now estimated to explain 46.5% of the variance in liability outside the Major Histocompatibility Complex (MHC) region, increasing to 59.2% when MHC is included (Dand et al., 2025). These loci are associated with genes involved in key biological pathways – such as immune-cell activation, keratinocyte proliferation and cytokine production – that contribute to psoriasis pathogenesis (Dand et al., 2025; Krueger & Bowcock, 2005; Roberson & Bowcock, 2010). Genes within these loci and related pathways include:

- **Skin Barrier Function:** Genes of the Epidermal Differentiation Complex on chromosome 1q21 are essential, including *GJB2* (connexin 26), *DEFB4* (defensin), *KLF4* and the Late Cornified Envelope (*LCE*) genes (Babaie et al., 2022; Roberson & Bowcock, 2010). Psoriatic skin exhibits hyperkeratosis and parakeratosis, with a defective water-vapour barrier due to incomplete differentiation and a dramatically shortened keratinocyte transit time (Henseler, 1997; Roberson & Bowcock, 2010). Key genetic and functional insights include:
    - The widely replicated deletion of *LCE3B* and *LCE3C* as a risk factor for psoriasis (de Cid et al., 2009).
    - The specific upregulation of the *LCE3* gene group (including *LCE3A*) in psoriatic lesions which is thought to be involved in barrier repair, in contrast to other *LCE*

groups which are downregulated (Bergboer et al., 2011). The *LCE3* cluster also shows evidence of epistatic interaction with the *HLA-Cw6* locus and has pleiotropic effects, being associated with other autoimmune diseases like rheumatoid arthritis (Shen et al., 2015).

  - The formation of a functional and physical interaction module in psoriasis between the Small Proline-Rich (*SPRR*) and *LCE* gene families, with direct protein–protein interactions demonstrated between *SPRR2* and *LCE3D* proteins (Tian et al., 2021).
  - The critical role of signalling molecules – such as Betacellulin (*BTC*), which is downregulated in psoriatic skin – in maintaining skin barrier integrity by promoting the expression of key barrier-related genes such as filaggrin and claudin-1 (Peng et al., 2022).

- **Antigen Presentation:** Psoriasis susceptibility has also been linked to numerous genes involved in the antigen presentation pathway, including *ERAP1*, *MBD2*, *RUNX3* (*RUNX1* site variants), *TNFRSF9*, *IRF4*, *ETS1*, *TAGAP*, *B3GNT2* and *HLA-B*/*C* (Babaie et al., 2022; Krueger & Bowcock, 2005).

- **Innate Immunity:** Key pathways and gene families involved include:

  - Nuclear Factor kappa-light-chain-enhancer of activated B cells (NF-κB) pathway, with associated genes including *TNIP1*, *TNFAIP3* (A20) and *CARD14* (Babaie et al., 2022; Roberson & Bowcock, 2010; Zhou et al., 2022).
  - IFN signalling pathways, with genes like *IFIH1* (Babaie et al., 2022; Roberson & Bowcock, 2010; Zhou et al., 2022).
  - Killer-cell Immunoglobulin-like Receptor (KIR) family and related variants (e.g., *MICA*002* and *KIR2DS2* on Natural Killer [NK] cells), associated with PsA risk (Babaie et al., 2022).
  - RIG-I-like Receptor (RLR) family members *DHX58* (LGP2) and *DDX58* (RIG-I), linked to psoriasis risk (Dand et al., 2025).

- **Adaptive Immunity:** Key genes and pathways involved include:

- Th1 pathway genes, such as *IL12B* and *TYK2* (Babaie et al., 2022; Roberson & Bowcock, 2010; Liu et al., 2007).
- Th17 pathway genes, such as *IL-23R*, *TRAF3IP2* (ACT1) and *STAT3* (Babaie et al., 2022; Roberson & Bowcock, 2010; Liu et al., 2007).
- Other key cytokines, including IL-22 and IL-21 (Babaie et al., 2022; Roberson & Bowcock, 2010; Liu et al., 2007).
- Regulatory-Associated Protein of mTOR (RAPTOR), a regulator of mTOR – a key T-cell regulator (Krueger & Bowcock, 2005).

- **Overlap with Other Autoimmune Diseases:** Psoriasis loci overlap with those for inflammatory bowel disease, Crohn's disease, diabetes, multiple sclerosis, rheumatoid arthritis and systemic lupus erythematosus (Krueger & Bowcock, 2005; Roberson & Bowcock, 2010; Liu et al., 2007, Dand et al., 2020, Dand et al., 2025). Building on this correlational evidence, Mendelian Randomisation (MR) studies have confirmed a causal link between Body Mass Index (BMI) and an increased risk of developing psoriasis (Ogawa & Okada, 2020; Dand et al., 2025). Furthermore, recent analyses suggest that genetic risk for psoriasis may have causal effects on traits such as back and generalised pain, increased fracture risk, diabetes and periodontitis (Dand et al., 2025).

Historically, the understanding of psoriasis has evolved from a primary disorder of keratinocytes to a disorder of the immune system, particularly involving T-cells (Campanati et al., 2021; Roberson & Bowcock, 2010). Understanding the complex genetic architecture is crucial for appreciating the molecular heterogeneity of psoriasis and for identifying potential targets for therapeutic intervention.

The molecular pathophysiology of psoriasis involves a profound dysregulation of both the innate and adaptive immune systems, forming an IL-17-mediated Type 3 dominant immunological disorder (Yamanaka et al., 2021; Liu et al., 2007). The initiation and perpetuation of psoriatic lesions are typically characterised by an inflammatory cascade involving various immune cells and their secreted cytokines. In addition to canonical cytokines, novel members of the IL-1 family (IL-1F5, -F6, -F8 and -F9) are also highly active in psoriasis, promoting Antimicrobial Peptide (AMP) expression and downregulating genes such as *BTC* (Johnston et al., 2011). Specific immune cells implicated include Th1, Th17 and Th22 lymphocytes, regulatory T-cells (Tregs; Foxp3-positive),

Dendritic Cells (DCs), neutrophils and Innate Lymphoid Cells (ILCs) (Campanati et al., 2021; Yamanaka et al., 2021; Liu et al., 2007; Roberson & Bowcock, 2010). Notably, a specific subset of these, ILC3s, has been identified as having a direct role in inducing the characteristic psoriatic rash (Yamanaka et al., 2021).

This process often initiates when keratinocytes, in response to damage, secrete AMPs such as LL37 and S100 proteins (Campanati et al., 2021; Yamanaka et al., 2021; Zhou et al., 2022, Roberson & Bowcock, 2010). These AMPs activate Toll-Like Receptors (TLRs) on plasmacytoid Dendritic Cells (pDCs), leading to Interferon-alpha (IFN-α) production, which in turn promotes myeloid Dendritic Cell (mDC) maturation (Campanati et al., 2021; Zhou et al., 2022; Liu et al., 2007). Tumour Necrosis Factor-alpha (TNF-α) is a pleiotropic molecule that stimulates T-cell proliferation and differentiation and promotes the adaptive immune effects of the IL-23/IL-17 axis (Campanati et al., 2021).

The IL-23/IL-17 axis is central to disease perpetuation. IL-23 signalling occurs via Janus Kinases (JAKs), primarily TYK2 and JAK2 (Yamanaka et al., 2021). IL-23 is central to Th17 and Th22 cell survival and proliferation, activating ILC3s and converting ILC2s to ILC3s (Yamanaka et al., 2021) and is produced at elevated levels by activated mDCs (Liu et al., 2007). IL-17A (the most important effector cytokine) and IL-17F are key players, with IL-17A acting on keratinocytes to induce inflammatory mediators (IL-1β, IL-6, IL-8, TNF-α, CCL20, IL-19 and IL36) and facilitating abnormal proliferation (Campanati et al., 2021; Yamanaka et al., 2021; Zhou et al., 2022; Liu et al., 2007). This sustained inflammatory cascade, characterised by Interferon-gamma (IFN-γ) leading to TNF-α, IL-23 and ultimately IL-17, drives the characteristic epidermal hyperplasia and inflammation (Campanati et al., 2021; Krueger & Bowcock, 2005; Liu et al., 2007). Further molecular insights highlight:

- **Keratinocyte-specific mechanisms:** Keratinocytes are critical initiators and amplifiers of inflammation (Zhou et al., 2022) with their transit time dramatically shortened in psoriasis (Henseler, 1997). Key mechanisms include:
    - The essential role of *TRAF3IP2* (ACT1) in IL-17A receptor signalling in keratinocytes (Zhou et al., 2022), which is further highlighted by the discovery of independent coding and regulatory variants that contribute to psoriasis susceptibility (Dand et al., 2025).

- The finding that IL-17RA deletion in keratinocytes is protective against psoriasis, which confirms their role as crucial IL-17 responders (Zhou et al., 2022).
- The involvement of the IL-36 family (where IL-36R deletion in keratinocytes is protective) and the cytokine IL-22 (which inhibits differentiation via *IL-22BP*) in regulating keratinocyte function (Yamanaka et al., 2021; Zhou et al., 2022).
- The contribution of NF-κB pathway genes such as *CARD14* and *TNFAIP3* (A20) in keratinocytes (Zhou et al., 2022; Babaie et al., 2022).
- The promotion of inflammation by Tumour Necrosis Factor-like Weak inducer of apoptosis (TWEAK) via its receptor (*Fn14*) on keratinocytes (Zhou et al., 2022).
- The promotion of angiogenesis by *VEGFA* in keratinocytes (Zhou et al., 2022).
- The induction of genes encoding keratins 6, 16 and 17 (Henseler, 1997; Roberson & Bowcock, 2010).
- Dysregulation of Foxp3-positive Treg function, which can lead to Tregs transforming into IL-17–producing cells (Yamanaka et al., 2021).

- **Other mediators:** The inflammatory environment is shaped by a variety of other mediators:

  - Specific autoantigens like LL-37 and keratin 17, which can activate IL-17A–producing T-cells (Yamanaka et al., 2021; Liu et al., 2007).
  - A host of chemokines (e.g., *CCL20*, *CXCL9*, *CXCL10*) that recruit various immune cells and organise lymphoid structures (Liu et al., 2007; Roberson & Bowcock, 2010).
  - The S100 protein family, also encoded within the Epidermal Differentiation Complex, which links regenerative hyperplasia and inflammation (Liu et al., 2007; Roberson & Bowcock, 2010). The S100A8–S100A9 complex, in particular, is not merely a marker but a functional mediator of the disease; it is the most upregulated protein in psoriatic epidermis and drives inflammation by regulating other inflammatory pathways including the complement system (Schonthaler et al., 2013). The massive upregulation of *S100A9* within psoriatic keratinocytes leads to its secretion into the bloodstream, making its serum levels a systemic biomarker

that directly correlates with clinical disease severity as measured by the Psoriasis Area and Severity Index (PASI) (Benoit et al., 2006).

- Other crucial proteins such as Chitinase 3-Like 2 (*CHI3L2*), a member of the chitinase-like protein family, which has been identified as a potential key gene in psoriasis with roles in autoimmune responses and tissue remodelling (J. Zhang et al., 2024).
- The contribution of memory cells, alpha E beta 7 (αEβ7) and *CXCR3* to T-cell trafficking (Liu et al., 2007).

- **Emerging Areas:** While the core genetic and immune pathways provide a solid framework, current research is expanding to explore a broader set of molecular and systemic factors. These emerging areas, encompassing epigenetic modifications, metabolic reprogramming and the influence of the microbiome, are crucial for developing a more holistic understanding of psoriasis and identifying novel targets for diagnosis and therapy:

  - **Metabolic reprogramming:** Including glucose (*Glut1*, *PKM2*), glutamate (*GLS1*) and lipid (*S1P*, *PCSK9*) metabolism in keratinocytes (Zhou et al., 2022).

  - **Oxidative stress and proteostasis:** Enhanced oxidative stress and decreased 20S proteasome activity in blood cells (Kanda, 2021).

  - **Fibroblast dysfunction:** Affecting intercellular signalling in dermal fibroblasts (Kanda, 2021).

  - **Non-coding RNAs (ncRNAs):** The role of microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) as mediators in keratinocyte function is increasingly recognised (Yamanaka et al., 2021; Zhou et al., 2022). This includes:

    - The lncRNA *SPRR2C*, which has been shown to be highly upregulated in psoriasis and is involved in the skin's response to cytokines like IL-22 (Shefler et al., 2022).
    - Specific miRNAs (e.g., *miR-31* upregulated; *miR-146a*/*b* downregulated) and other lncRNAs (e.g., *PRINS* upregulated; *MEG3* downregulated)

(Mateu-Arrom & Puig, 2023). Circular RNAs (circRNAs) also play a role (Mateu-Arrom & Puig, 2023).

- **DNA Methylation:** This is a key epigenetic mechanism altering chromatin structure (Mateu-Arrom & Puig, 2023). Specific insights include:
  - The identification of key psoriasis-related genes, such as *S100A9*, as being regulated by their methylation status, which is predictive of psoriasis risk (Wang et al., 2021).
  - The discovery of differentially methylated genes and pathways in psoriatic lesions, non-lesional skin and Peripheral Blood Mononuclear Cells (PBMCs), which have been linked to disease severity, histopathology and local disease memory (Mateu-Arrom & Puig, 2023).

- **Histone Modification:** Changes in histone methylation (e.g., H3K9, H3K4) and acetylation (e.g., H4) patterns, involving key protein families like Histone Acetyltransferases (HATs), Histone Deacetylases (HDACs) and Bromodomain and Extra-Terminal (BET) proteins, which are known to influence the expression of crucial factors such as RORC, IL-17 and IL-22 (Mateu-Arrom & Puig, 2023).

- **Gut microbiome:** Specific bacterial imbalances (dysbiosis) have been linked to psoriasis. These include a reduction in beneficial genera, such as *Bacteroides* and an increase in *Faecalibacterium*, an imbalance which is thought to influence systemic skin inflammation (Yamanaka et al., 2021; Kanda, 2021).

- **Biomarkers:** Molecules such as elafin, clusterin and selenoprotein P are potential indicators of inflammation and metabolic complications (Kanda, 2021).

- **Pruritus:** Specific mechanisms involving neuropeptides, opioid receptors, IL-31 and Vascular Endothelial Growth Factor (VEGF) (Kanda, 2021).

- **Novel Cytokines:** Other members of the IL-17 family derived from keratinocytes, such as IL-17C and IL-17E (IL-25), also contribute to inflammation (Zhou et al., 2022).

Systemic inflammation not only drives skin lesions but also underlies extracutaneous comorbidities and is modulated by factors such as gut microbiome dysbiosis and circulating biomarkers (e.g., elafin, clusterin, selenoprotein P) (Yamanaka et al., 2021; Kanda, 2021). To resolve these mechanisms, transcriptomic analysis – the large-scale study of gene expression – complements genomic work by revealing the functional consequences of genetic variation in psoriasis.

2.2 Transcriptomic Data Analysis in Psoriasis

Large-scale transcriptomic studies of psoriasis, including microarray analyses, have identified numerous Differentially Expressed Genes (DEGs) in lesional and non-lesional psoriatic skin, providing crucial insights into the molecular basis of the disease (Krishnan & Kõks, 2022; Swindell et al., 2013). Historically, the advent of microarray technology revolutionised the study of gene expression, allowing for the simultaneous measurement of thousands of genes and contributing significantly to the understanding of inflammation and keratinocyte dysregulation in the disease (Bowcock et al., 2001; Swindell et al., 2013). However, this technology harboured limitations including a restricted dynamic range, high background noise and a reliance on prior knowledge of gene sequences, potentially missing important genes expressed at lower levels (Jabbari et al., 2011; McGettigan, 2013; Lowe et al., 2017). Microarrays also suffered from low reproducibility between laboratories due to variations in fluorescence measurements (Rioux et al., 2020). Despite these challenges, early microarray studies laid the groundwork for identifying key molecular signatures; for example, a later, large-scale analysis by Gudjonsson et al. (2010) successfully identified over 900 unique DEGs and highlighted dysregulation in S100 family members and key genes, including *BTC*.

The emergence of high-throughput RNA-seq has since transformed transcriptomic research, largely superseding microarray technology for comprehensive gene expression analysis. RNA-seq provides a digital measure of transcript presence and prevalence by directly sequencing and counting RNA/cDNA species, offering advantages over microarray's analogue-style signals (Jabbari et al., 2011; Mortazavi et al., 2008; Wang et al., 2009). RNA-seq offers significant advancements, providing improved accuracy, greater sensitivity and a more comprehensive characterisation and quantification of the transcriptome. This technology enables the detection of novel transcripts, isoforms and fusion genes and provides a broader dynamic range of more than

9,000-fold for quantifying gene expression levels, which is essential for deeply understanding the molecular biology of inflammatory skin diseases like psoriasis (Mortazavi et al., 2008). In a direct comparison using psoriasis skin samples, RNA-seq detected over four times as many DEGs (2,629 versus 569) as microarrays, while also better capturing those with higher Fold Changes (FC) at lower mRNA expression levels (Jabbari et al., 2011). Importantly, several of these additional genes were independently validated in larger patient cohorts at both the mRNA and protein levels, supporting the robustness of the RNA-seq findings (Jabbari et al., 2011). RNA-seq additionally provides high reproducibility and reliable quantification across a broad range of RNA concentrations (Mortazavi et al., 2008).

The analysis of RNA-seq data involves a sophisticated computational pipeline. Best practices dictate a robust experimental design, including careful RNA extraction protocols and the use of paired-end (PE) reads, which are particularly valuable for *de novo* transcript discovery and isoform analysis (Lowe et al., 2017; Conesa et al., 2016). Sequencing depth remains essential for detecting low-abundance genes and a minimum of three biological replicates is recommended for robust statistical inference (Conesa et al., 2016). Pre-processing typically includes quality control, alignment to a reference genome and quantification of expression levels, followed by normalisation to account for library size and gene length (McGettigan, 2013; Conesa et al., 2016; Wu et al., 2021). Various normalisation metrics, including Reads Per Kilobase of transcript per Million mapped reads (RPKM), Fragments Per Kilobase of transcript per Million mapped reads (FPKM) and Transcripts Per Million (TPM) are employed to facilitate comparisons across samples (Conesa et al., 2016). In this study, expression values were provided as RPKM. RPKM adjusts for gene length and sequencing depth but has recognised limitations for between-sample comparisons (McGettigan, 2013). Differential expression is typically assessed using statistical models such as the negative binomial distribution, implemented in tools like `DESeq2` and `edgeR` (Conesa et al., 2016). Beyond standard DEG identification, RNA-seq enables more advanced molecular insights:

- **Alternative Splicing (AS) Variants:** RNA-seq detects AS events by mapping splice-crossing reads (Mortazavi et al., 2008), providing insights into RNA isoforms relevant for designing precise therapies (Krishnan & Kõks, 2022). Kõks et al. (2016) analysed 173,446 transcripts from psoriatic skin, identifying thousands of differentially expressed RNA isoforms. Key examples include:

- *CARMA2*/*CARD14*, which undergo AS to generate protein variants with differential effects on NF-κB activation (Krishnan & Kõks, 2022).
- *KLK10* and *S100A7A*/*A15*, which have disease-specific alternative isoforms, with the long isoform for *S100A7A/A15* specifically found in lesional psoriatic skin (Krishnan & Kõks, 2022; Kõks et al., 2016).
- *ETV3*, where a specific isoform is downregulated, suggesting suppression of anti-inflammatory signals in psoriatic skin (Kõks et al., 2016).

- **Non-coding RNAs (ncRNAs):** RNA-seq allows for the analysis of ncRNAs, including miRNAs, lncRNAs and circRNAs, which modulate gene expression and play pivotal roles in keratinocyte function (Yamanaka et al., 2021; Zhou et al., 2022).

- **RNA Editing and Alternative Polyadenylation (APA):** RNA-seq facilitates the identification of RNA editing sites (modifications to RNA nucleotides post-transcription) and APA events (alternative processing of mRNA 3' ends), both of which contribute to proteomic diversity (Wu et al., 2021).

- **Transcription Start Site (TSS) Mapping:** RNA-seq allows high-resolution mapping of TSSs (the initiation points of transcription) and their associated promoter regions (Ozsolak & Milos, 2011; Wang et al., 2009).

- **Gene Fusion Detection:** RNA-seq can identify gene fusion events, where parts of two different genes combine, which are relevant in diseases like cancer (Ozsolak & Milos, 2011; Conesa et al., 2016).

- **Transcriptome-Wide Association Studies (TWASs):** This advanced technique integrates GWAS data with gene expression to identify how genetic variants impact gene expression. TWAS analyses have revealed previous non-candidate susceptibility genes and have uncovered crucial mechanistic insights, such as transcriptional regulation of haematopoietic cell development and epigenetic modulation of IFN signalling (Dand et al., 2025).

- **Single-cell RNA-sequencing (scRNA-seq):** scRNA-seq is a powerful technique for resolving intercellular transcriptomic heterogeneity at single-cell resolution and has revealed distinct keratinocytic subpopulations as well as novel T-cell subsets in psoriatic skin (Wu et al., 2021; Lowe et al., 2017; Conesa et al., 2016).

- **Spatial Transcriptomics:** Unlike scRNA-seq, spatial transcriptomics preserves the tissue architecture of psoriatic lesions, enabling gene expression analysis *in situ*. This approach provides insights into immune–epidermal cell interactions and the spatial organisation of pathogenic pathways in psoriasis (Leung et al., 2022).

- **Multi-omics Integration:** To increase confidence in transcriptomic findings, modern studies often integrate data across multi-omic layers. For example, Wang et al. (2021) analysed the GSE54456 dataset, also used in this study, to integrate gene expression and DNA methylation data, identifying genes dysregulated at both the transcriptional and the epigenetic levels.

- **Repetitive Elements (REs):** Transcriptomic analyses now investigate REs, which constitute approximately 45% of the total genome sequence. A study by Krishnan & Kõks (2023) demonstrated that certain REs, including endogenous retroviruses (ERVs) and long and short interspersed nuclear elements (LINEs and SINEs) are differentially expressed in psoriatic skin.

- **3D Psoriatic Skin Models:** To bridge the gap between *in vitro* cultures and *in vivo* biopsies, 3D psoriatic skin models are utilised for transcriptomic studies, allowing for a more representative gene expression profile by incorporating various cell types and supplementing with cytokines (Rioux et al., 2020). These models are also used for personalised medicine approaches (Rioux et al., 2020).

A pivotal dataset in this field, and central to this study, is the GSE54456 RNA-seq dataset, publicly available through the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository (NCBI, n.d.). This dataset originates from a large case–control study of psoriasis, published by Li et al. (2014a). It comprises RNA-seq gene expression profiles from 174 skin samples: 92 psoriatic and 82 healthy samples. This yields a near-balanced class

distribution (psoriatic 52.9%; healthy 47.1%). Expression values are provided as RPKM, which normalises for gene length and library size but is limited for between-sample comparisons. In GSE54456, Li et al. (2014a) reported an average of ~38 million 80-bp single-end (SE) reads per sample generated on the Illumina Genome Analyzer IIx platform and identified 3,577 DEGs between lesional and normal skin. Further analysis using Weighted Gene Co-expression Network Analysis (WGCNA) on this dataset revealed specific gene modules linked to epidermal differentiation, lipid biosynthesis and inflammatory interplay between myeloid cells, T-cells and keratinocytes. The study also provided a crucial architectural insight, suggesting that the observed downregulation of dermally expressed genes in psoriatic biopsies could be explained by the significant expansion of the epidermal compartment. The GSE54456 dataset remains a significant resource due to its substantial sample size for an RNA-seq study of psoriasis, its clear case–control design and its suitability for ML-based classification and feature identification. Its value as a public resource is demonstrated by its use in numerous subsequent bioinformatics studies, including combined transcriptomic analyses that highlighted dysregulated lipid metabolism and identified key biomarkers, such as *AKR1B10* (Gao et al., 2017), and others that used network analysis and ML to investigate DEGs, methylation markers and potential disease pathways (Wang et al., 2021; Zhou et al., 2023a). For example, Federico et al. (2020) compiled and harmonised a collection of 13 RNA-seq datasets for psoriasis, including GSE54456, creating a valuable ready-to-use resource for the scientific community. This demonstrates a growing focus on leveraging and validating findings across multiple datasets to build more robust molecular signatures. As increasingly comprehensive transcriptomic resources like GSE54456 become available, the challenge shifts to extracting meaningful biological insights from these high-dimensional datasets – an area where supervised ML approaches are particularly powerful.

2.3 Supervised Machine Learning in Biomedical Research and Psoriasis

The field of supervised ML has become an increasingly indispensable tool in biomedical research, particularly for making sense of complex, high-dimensional datasets such as gene expression profiles (Camacho et al., 2018; Ghorbian et al., 2024; Cheng et al., 2024). By learning intricate patterns from labelled data (e.g., ‘Healthy’ versus ‘Disease’ Samples), supervised ML models are uniquely positioned to build predictive classifiers that can accurately distinguish between different biological conditions. This approach not only provides a powerful means for biomarker discovery

but also offers a pathway to uncover underlying molecular mechanisms that drive disease pathogenesis. The complexity of psoriasis, influenced by a multitude of genetic and environmental factors, makes it an ideal candidate for ML applications, as traditional statistical methods often struggle to capture the non-linear relationships present in such high-dimensional data (Guo et al., 2014; Tapak et al., 2021). The ML field is an exciting frontier that requires statistical models that can extract accurate and explainable predictions from data that is exponentially growing (Feldner-Busztin et al., 2023). The most popular ML models for multi-omics data integration – such as autoencoders, Random Forest Classifiers (RFCs) and Support Vector Machines (SVMs) – address the challenges of datasets with few samples and many features (Feldner-Busztin et al., 2023). Recent meta-analyses of ML applications in psoriasis provide powerful statistics on their impact, reporting that these techniques have increased the accuracy of diagnosis by 35%, improved treatment efficiency by 29% and boosted the success rate of patient survival predictions by 15% (Ghorbian et al., 2024). In the broader field of dermatology, ML applications are being developed for disease classification, mobile diagnostics, large-scale epidemiology and precision medicine, providing a guide for a wide range of applications that can be developed and used (Chan et al., 2020).

To achieve robust and accurate predictions from high-dimensional datasets like those derived from RNA-seq, several key ML concepts and techniques are commonly employed. The primary challenge is the *curse of dimensionality*, where a high number of features, P, relative to the number of samples, n (the $n \ll P$ problem), leads to data sparsity and complicates inference (Feldner-Busztin et al., 2023). A fundamental first step to overcome this is dimensionality reduction, which has two approaches: feature extraction and feature selection. Feature extraction transforms data by projecting the original features into a new, lower-dimensional space. In contrast, feature selection – the focus of this study – directly selects a subset of the most relevant original features for model construction (Li et al., 2017). This addresses the $n \ll P$ problem without altering the original feature space, which is crucial because it maintains the physical meaning of the selected genes, enhancing model interpretability. Beyond this interpretability, by focusing on a manageable subset of the most relevant genes, feature selection can also improve model accuracy and reduce computational complexity (Díaz-Uriarte & Alvarez de Andrés, 2006; Feldner-Busztin et al., 2023; Saeys et al., 2007; Mitić et al., 2025). However, this process can lead to the *multiplicity problem*, where different feature selection methods identify distinct sets of genes that perform with

comparable predictive accuracy, highlighting the need for robust validation (Díaz-Uriarte & Alvarez de Andrés, 2006). The importance of these techniques is particularly evident in the context of single-cell transcriptomics, where a high feature-to-sample ratio is a fundamental challenge (Mitić et al., 2025).

The most common ML goals for multi-omics data are classification, dimensionality reduction and survival prediction, with classification being the most popular (Feldner-Busztin et al., 2023). To this end, powerful strategies often involve multi-model approaches; for instance, researchers have combined methods like Least Absolute Shrinkage and Selection Operator (LASSO) and Support Vector Machine–Recursive Feature Elimination (SVM–RFE) to identify crucial genes shared across related disease (J. Zhang et al., 2024), while others have used combinations of RFC, Neural Networks (NNs) and SVM to distil a small set of characteristic genes from thousands of candidates (Dan et al., 2025). Similarly, Zhou et al. (2023a) used a combination of LASSO and RFC to screen for genes common to both psoriasis and its comorbidities. For this purpose, various supervised ML classification algorithms have been successfully applied and compared. Common approaches include:

- **Logistic Regression (LR):** As a foundational linear model, LR is frequently used in bioinformatics to establish a powerful and highly interpretable performance baseline against which more complex algorithms can be compared (Saeys et al., 2007). Shen et al. (2024) demonstrated its effective use in predicting treatment efficacy, while Yu et al. (2020) highlighted its role in identifying predictors of disease risk and comorbidities, underscoring its versatility across clinical and epidemiological applications. However, LR is inherently a binary classification technique. Its application to multiclass problems requires adaptation through decomposition strategies, such as one-vs-all (OVA), where a separate classifier is trained for each class against all others (Galar et al., 2011).

- **Random Forest Classifier (RFC):** This powerful tree-based ensemble method is particularly valued in bioinformatics for its ability to handle high-dimensional data where the number of variables exceeds the number of observations and its capacity to capture complex, non-linear interactions between features (Díaz-Uriarte & Alvarez de Andrés, 2006; Feldner-Busztin et al., 2023; Dan et al., 2025). Its robustness has led to its successful application in psoriasis research, with studies using the method to identify novel

biomarkers such as angiogenesis-related genes (M. J. Zhang et al., 2024), to stratify patients into distinct molecular subtypes, revealing underlying biological heterogeneity not visible clinically (Ainali et al., 2012) and to establish evaluation models whose scores were positively correlated with the PASI, demonstrating utility not just for classification but also for assessing clinical severity (Hu et al., 2022).

- **Support Vector Machine (SVM):** This kernel-based model is highly effective in high-dimensional spaces where features outnumber samples, making it a strong and popular choice for gene expression classification (Díaz-Uriarte & Alvarez de Andrés, 2006; Feldner-Busztin et al., 2023; Vanitha et al., 2015). Its application to psoriasis classification has yielded highly stable prediction accuracies (Guo et al., 2014), with some hybrid approaches achieving perfect classification on test sets (Tapak et al., 2021). Like LR, the standard SVM is a binary classifier and its extension to multiclass problems also necessitates the use of binarization techniques. Approaches like one-vs-one (OVO), which trains a classifier for each pair of classes and one-vs-all (OVA) are common methods to achieve this adaptation (Galar et al., 2011).

To ensure the models' generalisability and to avoid overfitting, rigorous evaluation processes are crucial. These typically involve using stratified *k*-fold cross-validation during hyperparameter tuning and assessing performance on an independent, held-out test set (Feldner-Busztin et al., 2023). The primary evaluation metric for binary classification tasks is often the Area Under the Receiver Operating Characteristic Curve (ROC AUC), a robust, threshold-independent measure of classifier performance. Unlike accuracy, which can be biased by class imbalance, ROC AUC provides a more reliable assessment by capturing the trade-off between sensitivity and specificity (Feldner-Busztin et al., 2023; Ghorbian et al., 2025; J. Zhang et al., 2024; Saito & Rehmsmeier, 2015). Finally, to address the black-box nature of some ML models, which can limit the interpretability of a model's predictions and hinder its utility for yielding insights on underlying biological mechanisms, methods like SHapley Additive exPlanations (SHAP) have been developed. SHAP is a unified approach to explain the output of any ML model by assigning an importance value to each feature, thereby providing insights into which genes contribute most significantly to a model's output for a given prediction (Camacho et. al., 2018; Feldner-Busztin et al., 2023; Yagin et al., 2023). A pivotal aspect of this modelling approach is a move towards

Explainable Artificial Intelligence (XAI), which seeks to provide both predictive power and transparency in model recommendations (Carrieri et al., 2021). While SHAP is a powerful tool for interpreting complex models, some researchers argue that for high-stakes decisions, inherently interpretable models should be used from the outset, rather than creating post hoc explanations for black boxes (Rudin, 2019). The value of such models is increasingly being demonstrated through their use in clinical settings to provide a clear interpretation of the impact of key genomic features on disease risk and prognosis (Yagin et al., 2023). Together, these advances in transcriptomics and ML underscore how far the field has progressed while also highlighting the need to consolidate these insights and identify remaining gaps in knowledge.

2.4 Knowledge Gap and Rationale

Building on this foundation, large-scale transcriptomic analyses, particularly those using modern technologies like RNA-seq, have proven powerful for identifying DEGs, molecular pathways and novel RNA species that are implicated in psoriasis pathogenesis (Krishnan & Kõks, 2022; Wu et al., 2021; Swindell et al., 2013; Jabbari et al., 2011). Furthermore, supervised ML models, including LR, RFC and SVM have been effectively applied to these datasets, achieving high-accuracy classification and robust feature selection (Camacho et al., 2018; Ghorbian et al., 2024; Cheng et al., 2024; Tapak et al., 2021).

However, a notable gap remains in current methodology. While many papers build and validate individual ML models for psoriasis classification, few systematically compare feature importance across model types within a unified framework. Many studies rely on a single model's native metric (e.g., RFC native feature importance [mean decrease in impurity; Gini importance] or LR coefficients), which are not directly comparable across algorithms. This lack of a standardised, comparative approach makes it challenging to identify a consensus set of biomarkers that are robustly predictive regardless of the underlying model's assumptions.

The primary rationale for this study, therefore, is to address this knowledge gap by providing a comprehensive, comparative analysis of feature importance across different models and interpretation methods. This study will utilise the well-established GSE54456 RNA-seq dataset and will apply three distinct ML models (LR, RFC and SVM).

Crucially, this study will generate feature importance rankings from multiple perspectives: the inherent methods of LR coefficients and RFC native feature importance, alongside the advanced,

model-agnostic SHAP framework applied to all three classifiers. By performing a consolidated feature analysis and quantitative overlap analysis across all of these generated lists, this study aims to identify a consensus set of genes that are consistently highlighted as the most influential for classification, regardless of the model or interpretation technique. This multi-faceted approach is expected to provide a more robust and biologically meaningful set of potential biomarkers than a single-model analysis, thereby enhancing the interpretability of ML in this domain and furthering our understanding of the core molecular mechanisms that drive psoriasis pathology. The successful implementation and validation of this methodology will serve as a strong proof of concept for future applications of a comparative ML and interpretability framework in biomedical research.

Chapter 3: Methodology

3.1 Overview

This study aimed to develop and compare supervised Machine Learning (ML) models for the accurate classification of psoriatic versus 'Healthy' skin samples using gene expression data, with a subsequent exploration of key gene features. The methodology encompassed data pre-processing, feature selection, model training and rigorous evaluation. All computational analyses were performed in Python 3.13.5, primarily leveraging libraries including `pandas` (data manipulation), `scikit-learn` (ML), `matplotlib` and `seaborn` (visualisation), and `shap` (model interpretability; see Appendix F for code and environment details).

3.2 Data Acquisition

The data used in this study were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) GSE54456 RNA-sequencing (RNA-seq) dataset. This dataset originates from a large case–control study of psoriasis, published by Li et al. (2014a). It comprises RNA-seq gene expression profiles from 174 skin samples, specifically 92 psoriatic ('Disease') and 82 normal ('Healthy') samples. The expression matrix comprised 174 samples × 21,510 genes (features). The gene expression levels within this dataset were originally quantified using Reads Per Kilobase of transcript per Million mapped reads (RPKM) normalisation, a method designed to account for variations in both gene length and total sequencing depth. The GSE54456 dataset is publicly available through the NCBI GEO repository (NCBI, n.d.) under accession GSE54456 (Li et al., 2014b).

3.3 Ethical Considerations

The GSE54456 dataset used in this study is secondary data originally published by Li et al. (2014a). All subjects involved in the original study provided written informed consent and the dataset is fully anonymised, containing no personally identifiable information.

In adherence to institutional requirements, a University Research Ethics Committee-approved self-declaration (Application Reference Number: 068508) was completed for this research. This declaration confirmed that the study involved only existing, robustly anonymised research data

and was judged to be unlikely to cause offence to the original participants. Formal approval for the study to proceed was granted on 23/05/2025. No further specific ethical considerations were identified for the scope of this study, consistent with the assessment that a Secondary Data Questionnaire was appropriate. The Ethics Approval Certificate is provided in Appendix A (Figure A.1).

3.4 Data Pre-processing

Prior to ML model application, the raw expression data underwent a series of pre-processing steps to ensure data integrity, appropriate distribution and suitability for downstream analysis.

3.4.1 Initial Data Inspection and Cleaning

Upon loading the dataset, an initial overview was conducted to understand its structure, identify data types and inspect descriptive statistics. A check for missing values across all columns revealed no missing data, confirming the completeness of the dataset for the intended analysis.

3.4.2 $\mathbf{Log_2}$ Transformation (RPKM + 1)

The raw RPKM gene expression values exhibited a right-skewed distribution, as evidenced by skewness statistics and by both the cohort-level pre-transformation density plot (Figure 3.1, left) and representative per-sample density plots (Appendix C, Figure C.1). To mitigate this skewness, stabilise variance and render the data more suitable for linear models and distance-based algorithms, a $\log_2(\text{RPKM} + 1)$ transformation was applied element-wise to the entire expression matrix (all genes, all samples). Because this transformation is parameter-free, it was applied uniformly to all samples, avoiding data leakage. Post-transformation, the cohort-level and per-sample distributions were effectively normalised (Figure 3.1, right; Appendix C, Figure C.2). Representative per-sample pre- and post-transformation density plots for four samples (two 'Healthy', two 'Disease') are provided in Appendix C (Figures C.1 and C.2).

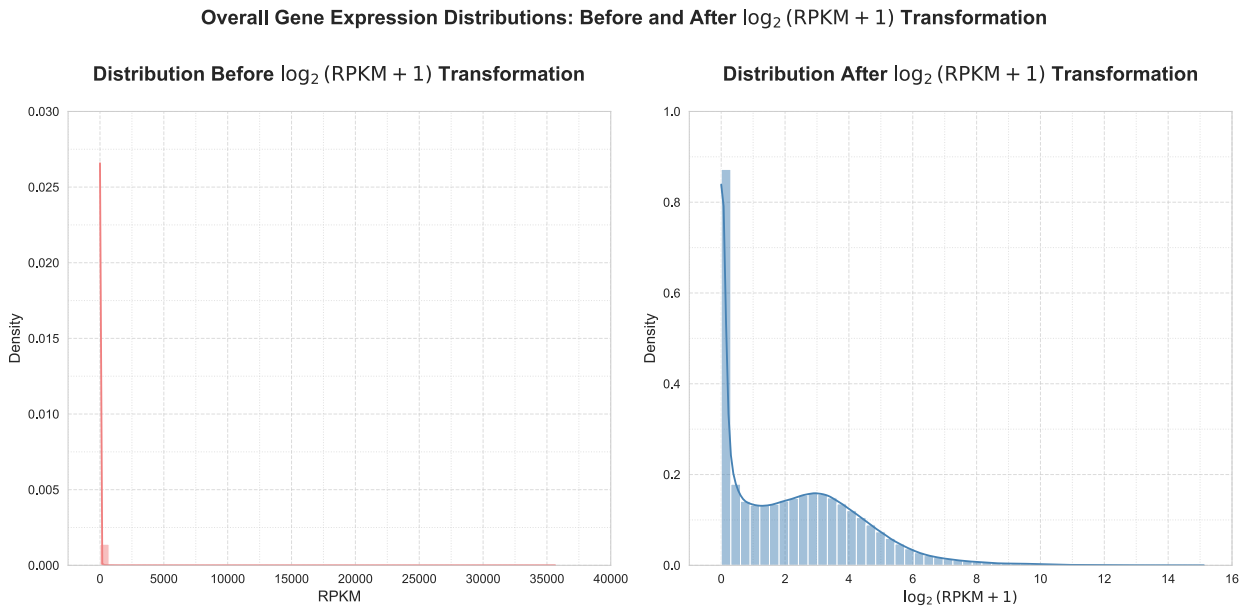


Figure 3.1 Overall Gene Expression Distributions: Before and After $log_2(RPKM + 1)$ Transformation

3.4.3 Data Definition and Encoding

The pre-processed dataset was then separated into features (`X_full`) and target labels (`y_full`). The 'labels' column contained the string categories 'Healthy' and 'Disease'. For modelling, the target was encoded to binary numeric values, mapping 'Disease' → 1 and 'Healthy' → 0. This convention designates 'Disease' as the positive class for all reported metrics and figures (see Section 3.5.1, 'Hyperparameter Tuning: Randomised Search with Stratified *k*-Fold Cross-Validation').

3.4.4 Train–Test Split

To ensure rigorous model evaluation and generalisability to unseen data, the dataset was partitioned using a stratified 80:20 train–test split (train n = 139; test n = 35), allocating 80% for model fitting and cross-validation and holding out 20% as an independent test set. The full cohort is approximately balanced ('Disease' 92/174, 52.9%; 'Healthy' 82/174, 47.1%) and the stratified split preserved this balance (train 52.5%/47.5%; test 54.3%/45.7%), mitigating class-imbalance concerns.

3.4.5 Feature Selection: Variance-Based Selection

Given the high-dimensional nature of RNA-seq data (21,510 genes versus 174 samples), dimensionality reduction was necessary to mitigate overfitting and improve interpretability. A preliminary variance-based selection was applied to identify and retain genes exhibiting substantial variation across samples in the training set. Specifically, the top 5,000 most variable genes (by variance in the training data) were selected and this fixed subset was then applied unchanged to the test set. This choice is a pragmatic compromise that captures most of the signal while keeping the feature space tractable for model training and SHAP analyses. Feature ranking and selection were performed on the training set only to avoid data leakage. Overall, this approach mitigates the $n \ll P$ problem (5,000 genes versus 174 samples) while avoiding overly aggressive filtering that might discard disease-relevant variance.

3.4.6 Feature Scaling (Standardisation)

All selected gene features underwent *z*-score normalisation before dimensionality reduction. Each feature was scaled to mean = 0 and standard deviation = 1. The `StandardScaler` was fitted exclusively on the training partition (`X_train_selected_variance`) to avoid data leakage, and then used to transform both the training and test sets (`X_train_scaled`, `X_test_scaled`). Standardisation is essential for distance-based algorithms like Support Vector Machines (SVMs) and regularisation models like Logistic Regression (LR), ensuring no single feature dominates the learning process due to scale.

3.4.7 Exploratory Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed on the scaled training data (`X_train_scaled`) to visualise separability in a reduced feature space. The first two principal components were retained for exploration, revealing clear separation between ‘Healthy’ and ‘Disease’ samples and providing an early indication of the data’s inherent discriminative power. To avoid data leakage, the PCA model was fitted on the training set only. PCA-derived components were not used as inputs to the classifiers.

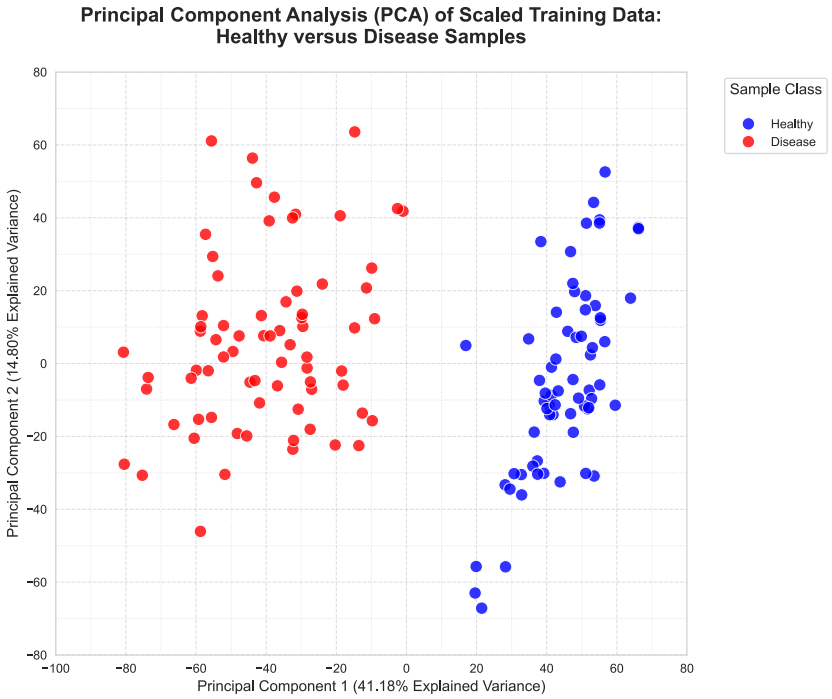


Figure 3.2 Principal Component Analysis (PCA) of Scaled Training Data: 'Healthy' versus 'Disease' Samples

3.5 Machine Learning Models

This study evaluated three supervised classification algorithms, each representing different approaches to learning from complex, high-dimensional data: Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Machine (SVM).

3.5.1 Hyperparameter Tuning: Randomised Search with Stratified *k*-Fold Cross-Validation

For each model, hyperparameter tuning was conducted using `RandomizedSearchCV`, which efficiently searches a defined distribution of hyperparameter values and is particularly useful for large search spaces. `StratifiedKFold` cross-validation with five folds (`N_SPLITS_CV`) was employed to obtain reliable performance estimates and mitigate variability. Stratification ensured that class proportions ('Healthy' versus 'Disease') were maintained within each fold, which is critical under class imbalance and still good practice in balanced cohorts. The primary scoring metric for hyperparameter tuning was the Area Under the Receiver Operating Characteristic Curve (ROC AUC), a robust threshold-independent measure of classifier performance. A fixed

`RANDOM_STATE = 42` was maintained to ensure reproducibility of the search process. For SVM, the search spanned `C` and kernel candidates (`'linear'`, `'rbf'`); `gamma` was tuned only when the `'rbf'` kernel was considered and is not applicable to the `'linear'` kernel. The following tables provide an overview of the fixed settings used for each model (Table 3.1), as well as the hyperparameter ranges explored during the tuning process (Table 3.2).

Table 3.1 Fixed Model Settings for LR, RFC and SVM

Model	Fixed Settings
LR	`solver='liblinear', max_iter=1000, random_state=RANDOM_STATE`
RFC	`random_state=RANDOM_STATE`
SVM	`probability=True, random_state=RANDOM_STATE`

Randomised search iterations were set to 50 for each model to balance computational efficiency with adequate exploration of the hyperparameter space.

Table 3.2 Hyperparameter Search Ranges Explored for LR, RFC and SVM Models

Model	Hyperparameter	Range/Options	Description
LR	`C`	0.02–10 (log-uniform)	Inverse regularisation strength; larger values = weaker regularisation
	`penalty`	`'l1'`, `'l2'`	Type of regularisation (LASSO, Ridge)
RFC	`n_estimators`	50–500 (integers)	Number of trees in the forest
	`max_features`	`'sqrt'`, `'log2'`, 0.5, 0.8, 1.0	Number of features considered at each split
	`max_depth`	10–100 (integers)	Maximum depth of each tree
	`min_samples_split`	2–20 (integers)	Minimum samples required to split an internal node
	`min_samples_leaf`	1–10 (integers)	Minimum samples required at a leaf node
	`bootstrap`	`True`, `False`	Whether to use bootstrap samples when building trees
SVM	`C`	0.1–100 (log-uniform)	Inverse regularisation strength; larger values = weaker regularisation
	`kernel`	`'linear'`, `'rbf'`	Kernel type for SVM
	`gamma`	0.0001–1 (log-uniform) – applies to `'rbf'` kernel only; not applicable to `'linear'` kernel	Kernel coefficient for `'rbf'`

These ranges were selected to cover plausible values reported in the literature and allow sufficient exploration of the hyperparameter space without excessive computational cost. This approach ensures a thorough search while keeping model training feasible.

3.6 Model Evaluation

The performance of each optimised model was rigorously evaluated on the held-out test set, which had not been used during training or hyperparameter tuning.

3.6.1 Performance Metrics

A comprehensive suite of classification metrics was employed to assess model performance beyond simple accuracy:

- **ROC AUC (Receiver Operating Characteristic Area Under the Curve):** A threshold-independent measure of the model's ability to discriminate between positive and negative classes. This was the primary metric for comparison across models.
- **Accuracy:** The proportion of correctly classified samples.
- **Precision:** The proportion of positive identifications that were actually correct.
- **Recall (sensitivity; true positive rate [TPR]):** The proportion of actual positives that were correctly identified.
- **F1-score:** The harmonic mean of precision and recall, providing a balance between the two.

A detailed classification report was also generated for each model, summarising these metrics for both classes ('Healthy' and 'Disease').

3.6.2 Confusion Matrix

Confusion matrices were generated for each model to visualise counts of true positives, true negatives, false positives and false negatives on the test set. This allowed for a clear understanding of the error types made by each classifier.

3.6.3 ROC Curve Analysis

ROC curves were plotted for each model, visually representing the trade-off between the true positive rate (TPR; recall) and the false positive rate (FPR) at various classification thresholds. The ROC AUC score, already used in hyperparameter tuning, provided a quantitative summary of these curves.

3.7 Feature Importance and Model Interpretability

Beyond predictive performance, understanding which gene features were most influential for classification was a key objective. Various methods were used to identify and interpret feature importance:

- **Logistic Regression Coefficients:** The coefficients of the trained LR model were analysed for feature importance. The magnitude and sign of these coefficients indicate the strength and direction of a gene's association with the 'Disease' (positive coefficients) or 'Healthy' (negative coefficients) class;
- **Random Forest Classifier Native Feature Importance:** RFC models inherently provide a measure of feature importance (the mean decrease in impurity; sometimes called Gini importance);
- **SHapley Additive exPlanations (SHAP):** SHAP values were computed for all three classifiers to provide a unified and theoretically sound measure of feature importance and interpretability:
    - For LR, `shap.LinearExplainer` was used;
    - For RFC, `shap.TreeExplainer` was used;
    - For SVM (a non-tree, non-linear model), `shap.KernelExplainer` was employed, utilising a *k*-means–summarised background dataset for computational efficiency;
    - SHAP summary plots were generated to visualise the global impact of features, showing how each feature contributes to pushing the model output from the base value to the final prediction;

- Individual SHAP force plots were also generated for select ‘Disease’ and ‘Healthy’ samples to provide explanations for specific predictions.

3.8 Consolidated Feature Analysis and Overlap

To provide a comprehensive view of important genes, a consolidated feature importance table was compiled by taking the union of all genes that appeared in the top 20 of any of the five methods (LR coefficients; RFC native feature importance; SHAP values for LR, RFC and SVM). For any given method, genes not present in its top 20 were assigned a value of zero. Each method’s raw importance values were then independently scaled to [0, 1] using min–max normalisation and the five scaled scores were summed to give a per-gene Consensus Score in the range 0–5. The complete, normalised per-gene values are provided in Appendix B, Table B.1. **Feature overlap analysis:** A quantitative analysis of feature overlap was conducted:

- **Pairwise Overlap:** The number of common genes between any two feature importance methods was calculated and presented.
- **Multi-method Overlap:** Overlaps across combinations of three or more methods and genes common to all methods, were identified.
- **Visualisations:** A heatmap of pairwise overlaps was generated to provide a quick overview of the agreement between methods. Furthermore, two bar charts were created to summarise the overlap counts: one showing the total summed overlaps and another showing the count of unique genes at each level of consensus.

Chapter 4: Results and Discussion

4.1 Logistic Regression Classification

4.1.1 Logistic Regression – Model Training and Hyperparameter Tuning

A Logistic Regression (LR) model was trained on the standardised and variance-selected training data. Hyperparameter tuning was conducted using `RandomizedSearchCV` with 50 iterations and `StratifiedKFold` cross-validation (five splits), optimising for ROC AUC. The hyperparameters were selected from predefined ranges, as detailed in Table 3.2.

The best hyperparameters found for the LR model were:

- `C`: 0.1329
- `penalty`: `'l1'`

The best cross-validation ROC AUC score achieved was 1.0000.

4.1.2 Logistic Regression – Model Evaluation on Test Set

The optimised LR model's performance on the independent **test set** was assessed using various metrics. The results were as follows:

- ROC AUC: 1.0000
- Accuracy: 1.0000
- Precision: 1.0000
- Recall: 1.0000
- F1-score: 1.0000

Table 4.1 provides the detailed classification report for the LR model on the test set. It shows perfect precision, recall and F1-score of 1.0000 for both the 'Healthy' (Class 0) and 'Disease' (Class 1) categories, with overall accuracy also 1.0000, confirming the model's ideal classification performance across class-specific and aggregate metrics.

Table 4.1 Classification Report for Logistic Regression (LR) on the Test Set (Class 0 = Healthy, Class 1 = Disease)

	Precision	Recall	F1-score	Support
0	1.0000	1.0000	1.0000	16
1	1.0000	1.0000	1.0000	19
Accuracy			1.0000	35
Macro Avg.	1.0000	1.0000	1.0000	35
Weighted Avg.	1.0000	1.0000	1.0000	35

4.1.3 Logistic Regression – Confusion Matrix

As illustrated in Figure 4.1, the LR model achieved perfect classification on the test set, correctly identifying 16 ‘Healthy’ samples (true negatives) and 19 ‘Disease’ samples (true positives). The absence of any false positives or false negatives further underscores the model’s precise performance for this dataset.

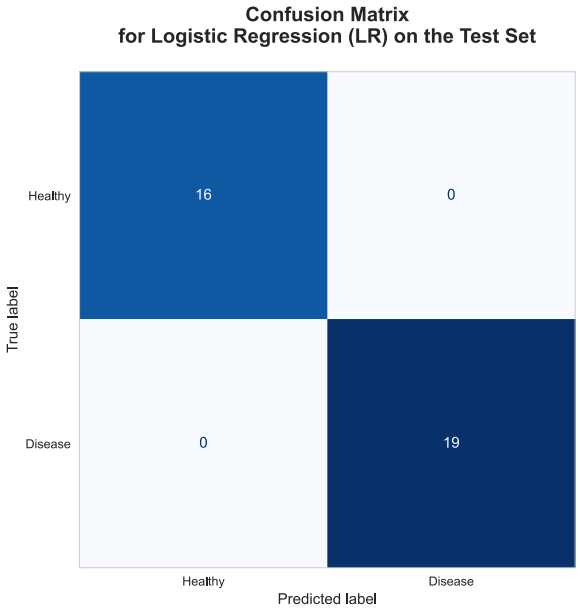


Figure 4.1 Confusion Matrix for Logistic Regression (LR) on the Test Set

4.1.4 Logistic Regression – ROC Curve

Figure 4.2 displays the Receiver Operating Characteristic (ROC) curve for the LR model. The curve hugs the top-left corner of the plot with an Area Under the Curve (AUC) of 1.0000, which

visually confirms the model's perfect discriminative ability between the two classes across all possible classification thresholds.

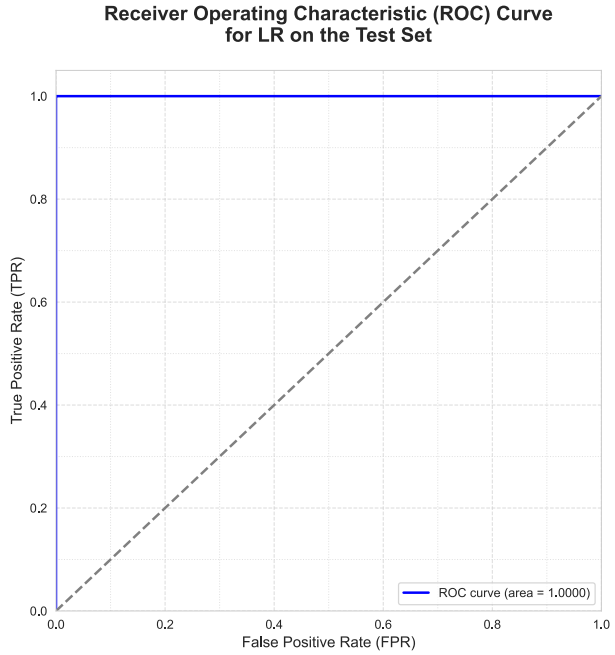


Figure 4.2 Receiver Operating Characteristic (ROC) Curve for LR on the Test Set

4.1.5 Logistic Regression – Coefficients

The coefficients derived from the LR model indicated the influence and direction of each gene's expression on the classification outcome. Table 4.2 lists the top 20 most influential genes as determined by their absolute LR coefficient values. Genes like *AKR1B10* (0.8653), *S100A7A* (0.2387) and *TPBG* (0.2300) are identified as having the largest overall impact on the model's prediction of psoriasis status.

Table 4.2 Top 20 Most Influential Genes Ranked by Absolute LR Coefficient

Gene	Coefficient	Absolute Coefficient
AKR1B10	0.8653	0.8653
S100A7A	0.2387	0.2387
TPBG	0.2300	0.2300
BTC	-0.1448	0.1448
FABP5	0.1421	0.1421

PI3	0.1220	0.1220
SPRR2G	0.1130	0.1130
F3	-0.1044	0.1044
INA	0.0973	0.0973
ABCG4	0.0953	0.0953
S100A7	0.0806	0.0806
S100A8	0.0799	0.0799
LCE3A	0.0771	0.0771
FUT3	0.0684	0.0684
S100A9	0.0643	0.0643
CHRM4	-0.0643	0.0643
TAP2	0.0455	0.0455
DEFB4A	0.0432	0.0432
CHI3L2	0.0399	0.0399
ANKFN1	-0.0364	0.0364

Table 4.3 specifically presents the top 20 genes whose positive LR coefficients indicate a strong association with the 'Disease' (Class 1) phenotype. *AKR1B10* (0.8653), *S100A7A* (0.2387) and *TPBG* (0.2300) are among the genes whose increased expression is most strongly linked to the presence of psoriasis.

Table 4.3 Top 20 Genes Positively Associated with the 'Disease' Class by LR Coefficient

Gene	**Coefficient**	**Absolute Coefficient**
AKR1B10	0.8653	0.8653
S100A7A	0.2387	0.2387
TPBG	0.2300	0.2300
FABP5	0.1421	0.1421
PI3	0.1220	0.1220
SPRR2G	0.1130	0.1130
INA	0.0973	0.0973
ABCG4	0.0953	0.0953
S100A7	0.0806	0.0806
S100A8	0.0799	0.0799
LCE3A	0.0771	0.0771
FUT3	0.0684	0.0684
S100A9	0.0643	0.0643

TAP2	0.0455	0.0455
DEFB4A	0.0432	0.0432
CHI3L2	0.0399	0.0399
S100A12	0.0353	0.0353
SERPINB4	0.0298	0.0298
C10orf99	0.0252	0.0252
SPRR2A	0.0239	0.0239

Conversely, Table 4.4 enumerates the top 20 genes whose negative LR coefficient suggested a positive association with the 'Healthy' (Class 0) phenotype. *BTC* (-0.1448), *F3* (-0.1044) and *CHRM4* (-0.0643) are highlighted as genes whose expression patterns are most indicative of 'Healthy' skin.

Table 4.4 Top 20 Genes Positively Associated with the 'Healthy' Class by LR Coefficient

Gene	Coefficient	Absolute Coefficient
BTC	-0.1448	0.1448
F3	-0.1044	0.1044
CHRM4	-0.0643	0.0643
ANKFN1	-0.0364	0.0364
C6orf26	-0.0160	0.0160
COBL	-0.0028	0.0028
ST3GAL6	0.0000	0.0000
COPZ2	0.0000	0.0000
CD46	0.0000	0.0000
SEMA4G	0.0000	0.0000
ITGB5	0.0000	0.0000
ABHD14A	0.0000	0.0000
ST3GAL3	0.0000	0.0000
ORC1L	0.0000	0.0000
HOMER1	0.0000	0.0000
TMSB15B	0.0000	0.0000
KIAA0802	0.0000	0.0000
FGGY	0.0000	0.0000
CCNG1	0.0000	0.0000
NES	0.0000	0.0000

Figure 4.3 provides a visual representation of the signed coefficients for the top 20 genes identified by the LR model. Notably, *AKR1B10* exhibited the largest positive coefficient (0.8653), while *BTC* displayed the largest negative coefficient, indicating their strong respective associations with the 'Disease' and 'Healthy' classes. The distribution of coefficients across positive and negative values is also evident.

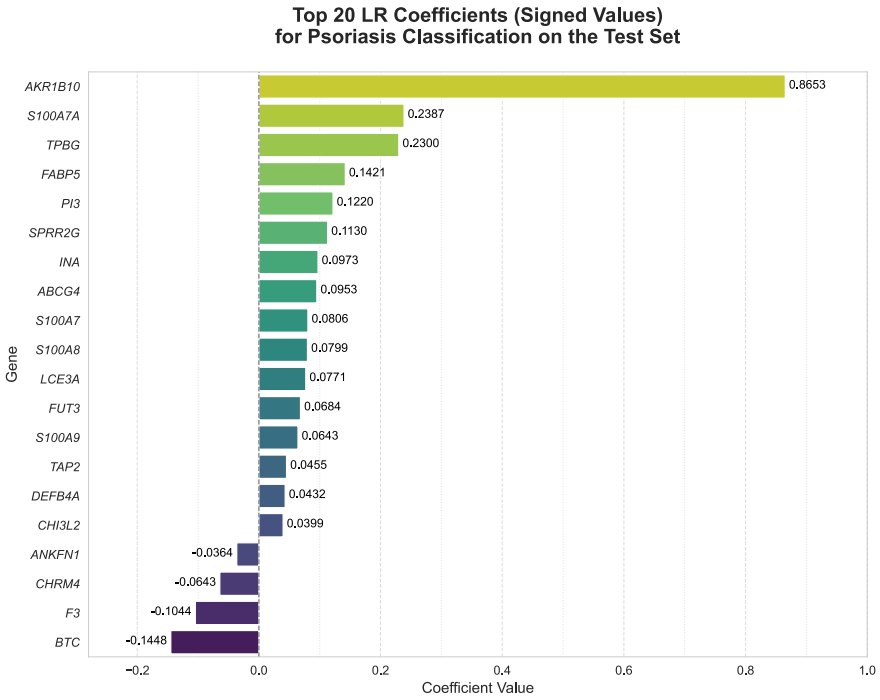


Figure 4.3 Top 20 LR Coefficients (Signed Values) for Psoriasis Classification on the Test Set

4.1.6 Logistic Regression – SHAP Global Feature Importance

Mean absolute SHapley Additive exPlanations (SHAP) values were used to further interpret the LR model's predictions, providing a model-agnostic measure of each gene's average impact on the prediction. These are presented for the top 20 features in Table 4.5.

Table 4.5 Top 20 Global Features Ranked by Mean Absolute SHAP Value for LR

Gene	Mean Absolute SHAP Value
AKR1B10	0.8351
TPBG	0.2350
S100A7A	0.2339

FABP5	0.1548
BTC	0.1381
PI3	0.1180
SPRR2G	0.1072
INA	0.0958
ABCG4	0.0937
F3	0.0926
LCE3A	0.0800
S100A7	0.0765
S100A8	0.0751
FUT3	0.0634
S100A9	0.0616
CHRM4	0.0534
DEFB4A	0.0431
TAP2	0.0404
S100A12	0.0349
CHI3L2	0.0325

Table 4.5 identifies *AKR1B10* (0.8351), *TPBG* (0.2350) and *S100A7A* (0.2339) as the most globally important features for this model, based on their mean absolute SHAP values.

Figure 4.4 presents the SHAP summary plot for the LR model. This visualisation indicates that *AKR1B10* is the most influential feature, followed by *TPBG* and *S100A7A*, based on their mean absolute SHAP values. The plot also illustrates the directional impact of each feature, showing how high (red) or low (blue) expression values contribute to the model's output (positive SHAP values pushing towards 'Disease', negative pushing towards 'Healthy').

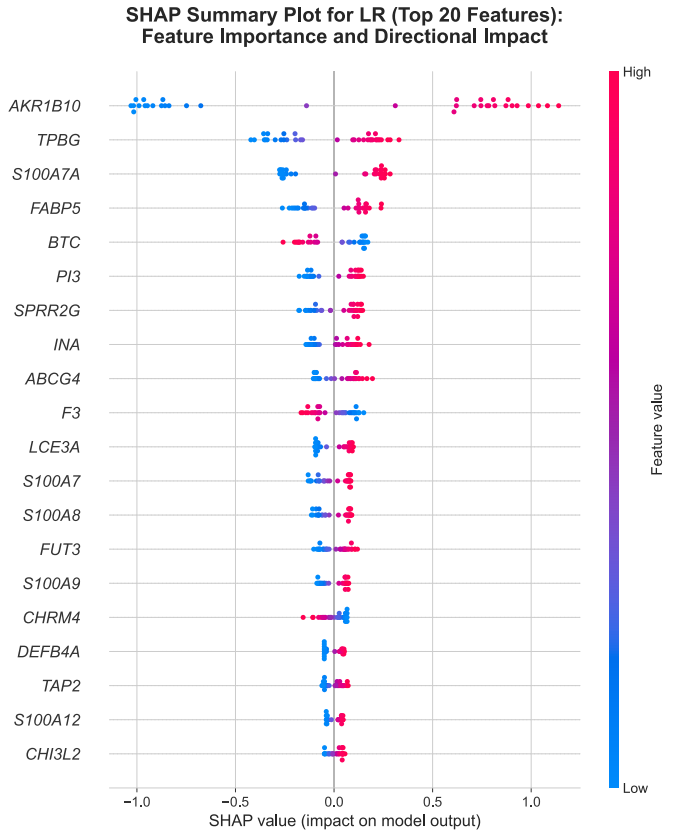


Figure 4.4 SHAP Summary Plot for LR (Top 20 Features) showing Feature Importance and Directional Impact on Model Output

To assess agreement between native coefficients and model-agnostic explanations, we compared LR coefficients with LR SHAP values. **Concordance with LR coefficients:** The SHAP rankings closely mirror the native LR coefficients: 19/20 (95%) of the genes in Table 4.5 also appear in Table 4.2 (*S100A12* replaces *ANKFN1*). Rank shifts are typically ≤2 places. See Section 4.4.3, 'Quantitative Feature Overlap Analysis' and Table 4.13 for the quantified overlap.

4.1.7 Logistic Regression – SHAP Force Plots (Representative Samples)

To demonstrate individual prediction explanations, SHAP force plots were generated for a representative 'Disease' sample and a 'Healthy' sample from the test set. Figure 4.5 illustrates the individual feature contributions for a representative 'Disease' sample (ID: M8481). The plot demonstrates how specific gene expression values combine to push the model's output from the baseline expected value towards the final prediction probability of 'Disease'.

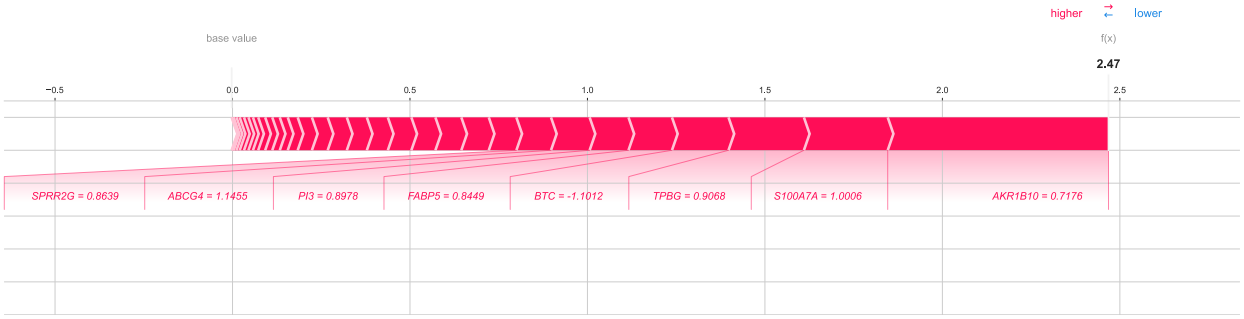


Figure 4.5 SHAP Force Plot for LR Explaining the Prediction of a Representative 'Disease' Sample (ID: M8481)

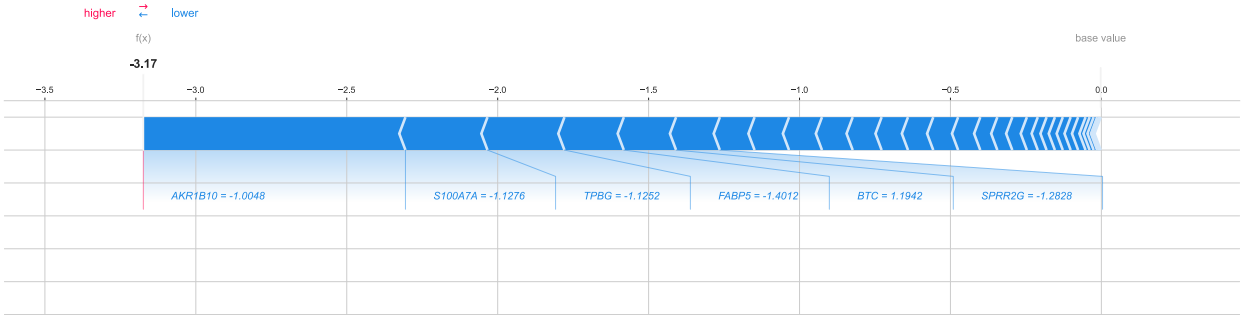


Figure 4.6 SHAP Force Plot for LR Explaining the Prediction of a Representative 'Healthy' Sample (ID: M8702)

Conversely, Figure 4.6 provides insight into the prediction for a representative ‘Healthy’ sample (ID: M8702). Here, the influence of genes collectively drives the model’s output from the baseline expected value towards a low probability of ‘Disease’, consistent with a ‘Healthy’ classification.

4.2 Random Forest Classification

4.2.1 Random Forest Classifier – Model Training and Hyperparameter Tuning

A Random Forest Classifier (RFC) was trained and optimised using `RandomizedSearchCV` (50 iterations, five-fold `StratifiedKFold` cross-validation), with ROC AUC as the scoring metric. Candidate hyperparameters were sampled from the predefined ranges shown in Table 3.2.

The best hyperparameters identified for the RFC model were:

- `n_estimators`: 171
- `max_features`: 1.0

- `max_depth`: 61
- `min_samples_split`: 8
- `min_samples_leaf`: 8
- `bootstrap`: `True`

The best cross-validation ROC AUC score achieved was 1.0000.

4.2.2 Random Forest Classifier – Model Evaluation on Test Set

The RFC model's performance on the independent **test set** was as follows:

- ROC AUC: 1.0000
- Accuracy: 1.0000
- Precision: 1.0000
- Recall: 1.0000
- F1-score: 1.0000

Table 4.6 provides the detailed classification report for the RFC on the test set. It confirms perfect precision, recall and F1-score of 1.0000 across both 'Healthy' and 'Disease' classes, with overall accuracy also 1.0000, demonstrating ideal performance across class-specific and aggregate metrics.

Table 4.6 Classification Report for Random Forest Classifier (RFC) on the Test Set (Class 0 = Healthy, Class 1 = Disease)

	Precision	Recall	F1-score	Support
0	1.0000	1.0000	1.0000	16
1	1.0000	1.0000	1.0000	19
Accuracy			1.0000	35
Macro Avg.	1.0000	1.0000	1.0000	35
Weighted Avg.	1.0000	1.0000	1.0000	35

4.2.3 Random Forest Classifier – Confusion Matrix

Figure 4.7 displays the confusion matrix for the RFC on the test set. Similar to LR, this model also achieved perfect classification, demonstrating no misclassifications of either 'Healthy' or 'Disease' samples.

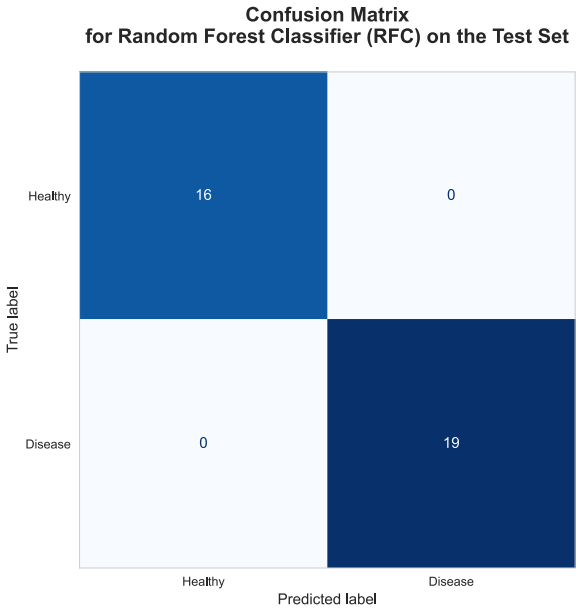


Figure 4.7 Confusion Matrix for Random Forest Classifier (RFC) on the Test Set

4.2.4 Random Forest Classifier – ROC Curve

The ROC curve for the RFC is presented in Figure 4.8. The perfect Area Under the Curve (AUC) of 1.0000 visually confirms the model's ideal performance in discriminating between the two classes, mirroring the results obtained from LR.

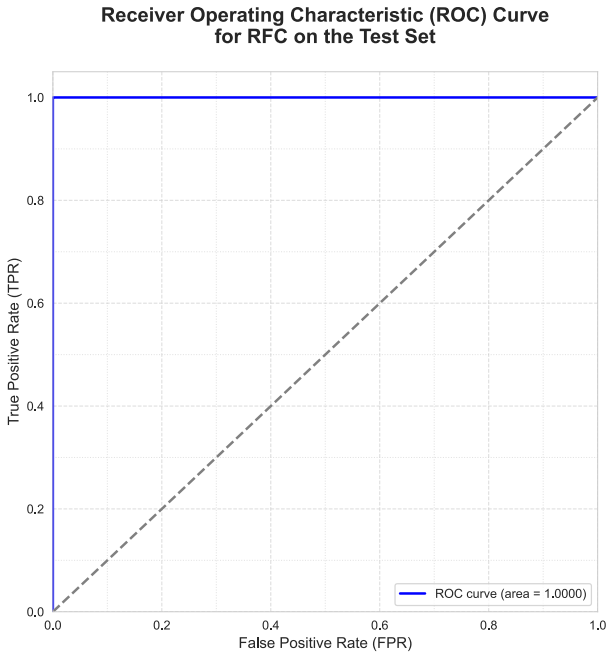


Figure 4.8 Receiver Operating Characteristic (ROC) Curve for RFC on the Test Set

4.2.5 Random Forest Classifier – Native Feature Importance

RFC native feature importance (mean decrease in impurity) identified the top 20 most important genes as shown in Table 4.7.

Table 4.7 Most Important Genes Identified by RFC Native Feature Importance

Gene	Native Feature Importance
COBL	0.0292
CHI3L2	0.0234
ANKRD33B	0.0234
PLA2G4D	0.0234
SPRR2C	0.0234
SNTB1	0.0234
CCR7	0.0234
LCE3A	0.0175
SPRR2A	0.0175
SPRR2G	0.0175
KYNU	0.0175
SPRR2D	0.0175
S100A9	0.0175

RND1	0.0175
C10orf99	0.0175
BTC	0.0175
FOXE1	0.0175
CD24	0.0117
WLS	0.0117
KRT6A	0.0117

Figure 4.9 visualises the top 20 genes identified as important by the RFC. Genes such as *COBL* (0.0292), *CHI3L2* (0.0234) and *ANKRD33B* (0.0234) exhibit among the highest RFC native feature importance scores (i.e., greater mean decrease in impurity across trees).

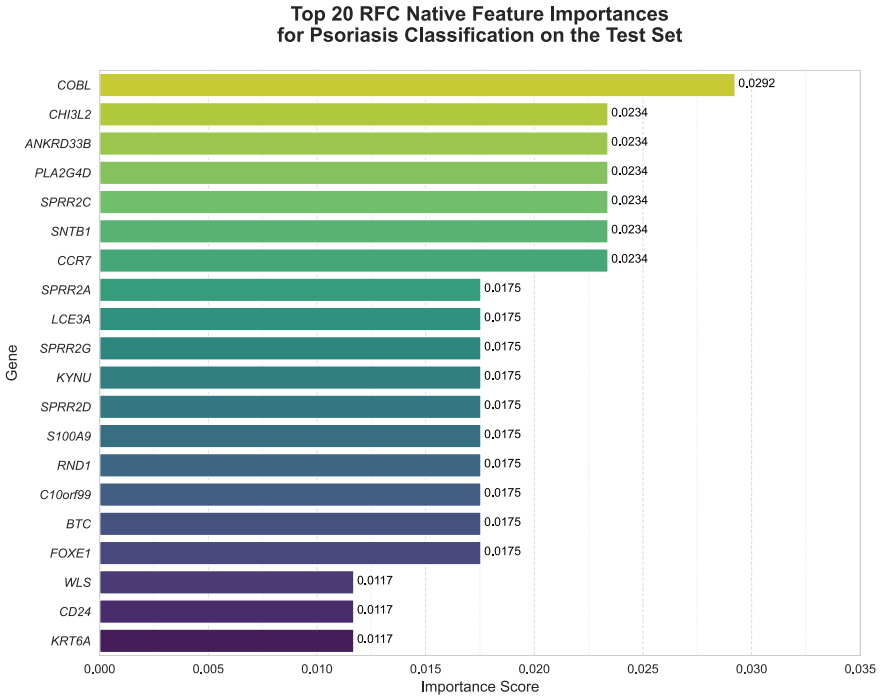


Figure 4.9 Top 20 RFC Native Feature Importances for Psoriasis Classification on the Test Set

4.2.6 Random Forest Classifier – SHAP Global Feature Importance

SHAP values were also computed for the RFC to provide model-agnostic insights into feature importance. Table 4.8 presents the top 20 global features based on their mean absolute SHAP values.

Table 4.8 Top 20 Global Features Ranked by Mean Absolute SHAP Value for RFC

Gene	Mean Absolute SHAP Value
COBL	0.0146
ANKRD33B	0.0117
PLA2G4D	0.0116
SPRR2C	0.0116
SNTB1	0.0115
CCR7	0.0112
CHI3L2	0.0112
C10orf99	0.0087
KYNU	0.0087
LCE3A	0.0087
SPRR2D	0.0087
BTC	0.0087
S100A9	0.0087
SPRR2A	0.0087
FOXE1	0.0087
SPRR2G	0.0083
RND1	0.0081
EPN2	0.0058
RAB3B	0.0058
FABP5L3	0.0058

Figure 4.10 displays the SHAP summary plot for the RFC. The plot highlights *COBL* (0.0146), *ANKRD33B* (0.0117) and *PLA2G4D* (0.0116) as particularly influential features based on their mean absolute SHAP values, with the colour indicating the original gene expression level's contribution to the prediction.

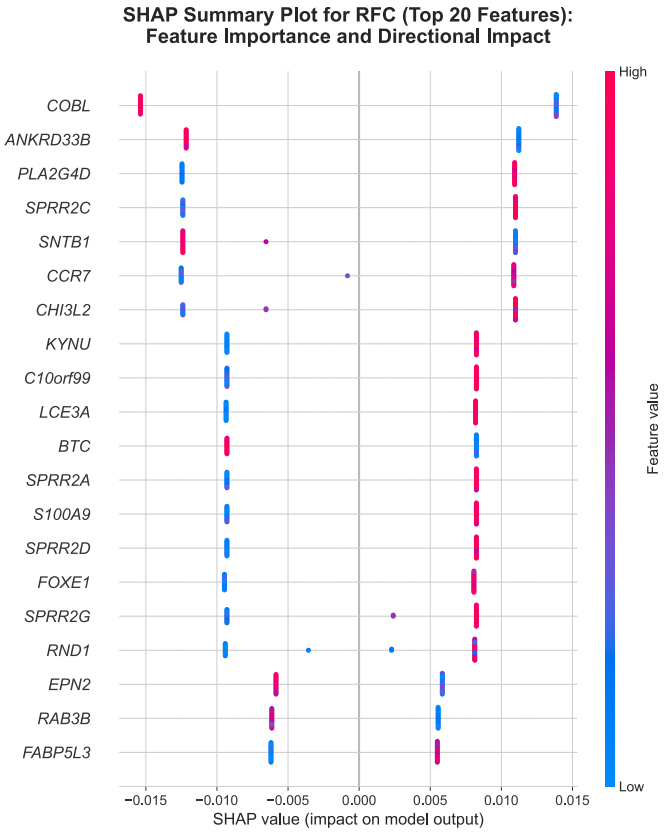


Figure 4.10 SHAP Summary Plot for RFC (Top 20 Features) showing Feature Importance and Directional Impact on Model Output

We next compared the RFC's native impurity-based importances with its SHAP rankings. **Concordance with RFC native feature importance:** The SHAP rankings closely match the RFC native feature importance (mean decrease in impurity): 17/20 (85%) of the genes are shared between Tables 4.7 and 4.8. Native-only genes are *CD24*, *WLS*, *KRT6A*; SHAP-only genes are *EPN2*, *RAB3B*, *FABP5L3* (see Section 4.4.3, 'Quantitative Feature Overlap Analysis' and Table 4.13 for the quantified overlap).

4.3 Support Vector Machine Classification

4.3.1 Support Vector Machine – Model Training and Hyperparameter Tuning

A Support Vector Machine (SVM) model was trained and tuned using `RandomizedSearchCV` (50 iterations, five-fold `StratifiedKFold` cross-validation), with ROC AUC as the

optimisation metric. Candidate hyperparameters were sampled from the predefined ranges shown in Table 3.2; for the selected `'linear'` kernel, `gamma` is not applicable.

The best hyperparameters for the SVM model were:

- `C`: 1.3293
- `kernel`: `'linear'`
- `gamma`: not applicable to `'linear'` kernel

The best cross-validation ROC AUC score achieved was 1.0000.

4.3.2 Support Vector Machine – Model Evaluation on Test Set

The SVM model's performance on the independent **test set** was measured as follows:

- ROC AUC: 1.0000
- Accuracy: 1.0000
- Precision: 1.0000
- Recall: 1.0000
- F1-score: 1.0000

Table 4.9 provides the detailed classification report for the SVM on the test set. It shows perfect precision, recall and F1-score of 1.0000 for both class categories, with overall accuracy also 1.0000, affirming the SVM's ideal classification performance.

Table 4.9 Classification Report for Support Vector Machine (SVM) on the Test Set (Class 0 = Healthy, Class 1 = Disease)

	Precision	Recall	F1-score	Support
0	1.0000	1.0000	1.0000	16
1	1.0000	1.0000	1.0000	19
Accuracy			1.0000	35
Macro Avg.	1.0000	1.0000	1.0000	35
Weighted Avg.	1.0000	1.0000	1.0000	35

4.3.3 Support Vector Machine – Confusion Matrix

Figure 4.11 presents the confusion matrix for SVM on the test set. Consistent with the other models, the SVM classifier also achieved perfect discrimination, showing no false positives or false negatives.

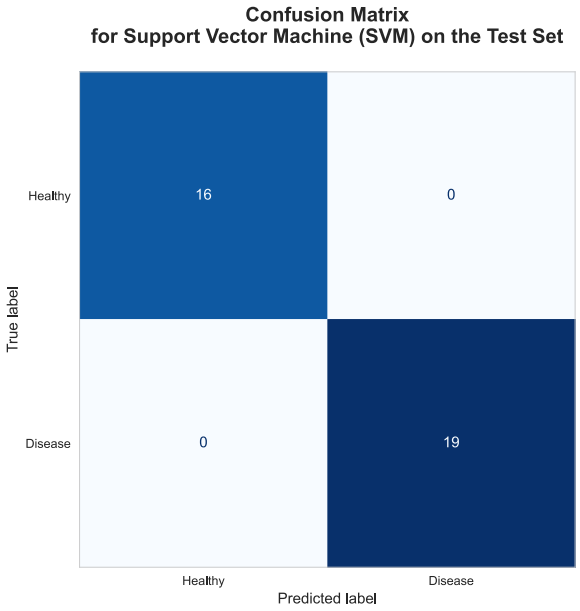


Figure 4.11 Confusion Matrix for Support Vector Machine (SVM) on the Test Set

4.3.4 Support Vector Machine – ROC Curve

The ROC curve for the SVM is shown in Figure 4.12, displaying a perfect AUC of 1.0000. The result further enforces the strong signal present in the dataset, allowing all tested models to achieve optimal classification performance.

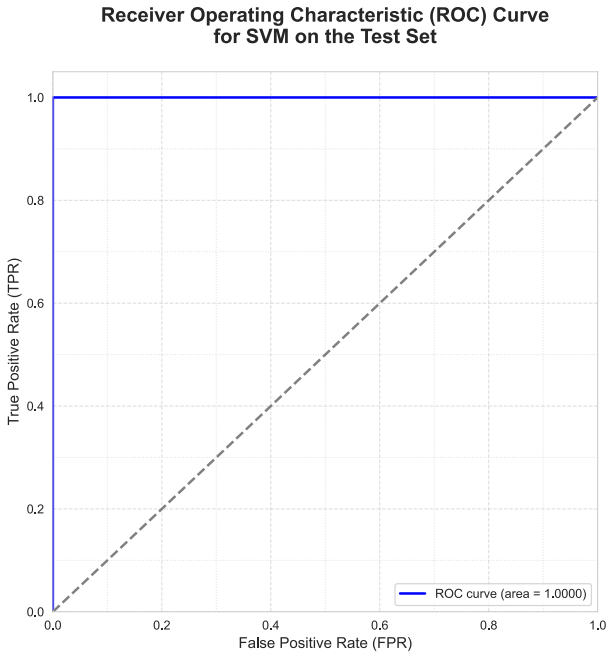


Figure 4.12 Receiver Operating Characteristic (ROC) Curve for SVM on the Test Set

4.3.5 Support Vector Machine – SHAP Global Feature Importance

SHAP values were calculated for the SVM model using `shap.KernelExplainer` with a *k*-means–summarised background dataset. Table 4.10 displays the top 20 global features based on their mean absolute SHAP values.

Table 4.10 Top 20 Global Features Ranked by Mean Absolute SHAP Value for SVM

Gene	Mean Absolute SHAP Value
SENP7	0.0030
IGFL1	0.0029
STAT1	0.0028
HPGDS	0.0028
PMCH	0.0028
EPS8L1	0.0027
AMY2B	0.0027
GTSE1	0.0027
APOC2	0.0027
C1orf31	0.0027
GPT	0.0027

KDELC2	0.0026
MS4A4A	0.0026
MIR221	0.0025
SNORD34	0.0025
MDM4	0.0017
DHX58	0.0017
PPARD	0.0017
C6orf48	0.0017
SPINK7	0.0016

Figure 4.13 illustrates the SHAP summary plot for the SVM. Features like *SENP7* (0.0030), *IGFL1* (0.0029) and *STAT1* (0.0028) are most prominent by mean absolute SHAP value, indicating their contribution to the SVM's predictions.

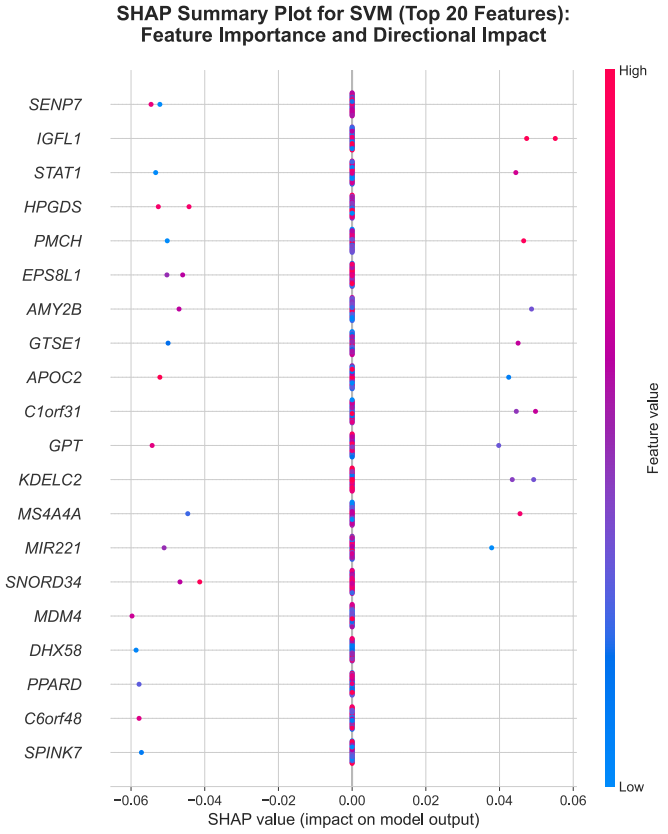


Figure 4.13 SHAP Summary Plot for SVM (Top 20 Features) showing Feature Importance and Directional Impact on Model Output

The distribution is tightly clustered near zero – consistent with a regularised linear model that spreads importance thinly across many genes. We interpret these magnitudes only in the context of the SVM, as SHAP scales are not comparable across different explainers.

Because SVMs do not expose a reliable native importance metric, we report SHAP-based importances only and summarise cross-model agreement. **Cross-model concordance among SHAP rankings:** SVM parameters are not directly interpretable as feature importances, nor are they comparable to coefficient- or impurity-based measures. Across SHAP top 20 lists, LR versus RFC share 5/20 genes (25%; *BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*), whereas LR versus SVM and RFC versus SVM share 0/20 (0%) (See Section 4.4.3, 'Quantitative Feature Overlap Analysis', Table 4.13 and Figure 4.15).

4.4 Overall Model Comparison and Consolidated Feature Analysis

4.4.1 Comparative Summary of Model Performance Metrics

A consolidated summary of the performance metrics obtained from the test set for all three classification models is provided in Table 4.11. It consistently shows a perfect score of 1.0000 across all metrics (ROC AUC, accuracy, precision, recall and F1-score) for LR, RFC and SVM, highlighting their uniform optimal performance on this dataset.

Table 4.11 Comparative Summary of Classification Model Performance Metrics on the Test Set

Model	ROC AUC	Accuracy	Precision	Recall	F1-score
Logistic Regression (LR)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Random Forest Classifier (RFC)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Support Vector Machine (SVM)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

4.4.2 Consolidated Top Gene Features

To identify robustly important gene features across different models and interpretability methods, the top 20 genes from LR coefficients, RFC native feature importance and SHAP values for LR, RFC and SVM were merged. Table 4.12 presents these top 20 consolidated gene features, ranked by a final Consensus Score. This score, which summarises a gene's overall importance, was

calculated by summing, for each gene, the five values after min–max normalisation to [0, 1]. Genes not present in a method's top 20 received a value of zero before normalisation. Full, normalised per-gene values for all five methods are provided in Appendix B, Table B.1.

Table 4.12 Consolidated Top 20 Gene Features Across LR, RFC and SVM Models, Ranked by Consensus Score

Gene	Consensus Score
AKR1B10	2.0000
COBL	2.0000
CHI3L2	1.6533
ANKRD33B	1.6008
PLA2G4D	1.5991
SPRR2C	1.5967
SNTB1	1.5935
CCR7	1.5726
BTC	1.5302
SPRR2G	1.4294
LCE3A	1.3841
S100A9	1.3456
C10orf99	1.1996
KYNU	1.1996
SPRR2D	1.1975
SPRR2A	1.1975
FOXE1	1.1958
RND1	1.1569
SENP7	1.0000
IGFL1	0.9602

Figure 4.14 visually compares the top 20 consolidated gene features in a stacked bar chart, ranked by their final Consensus Score. This visualisation demonstrates both the overall ranking and the composition of each gene's score. The total length of each bar represents the Consensus Score, while the coloured segments illustrate the min–max normalised [0, 1] contribution of each of the five analytical methods.

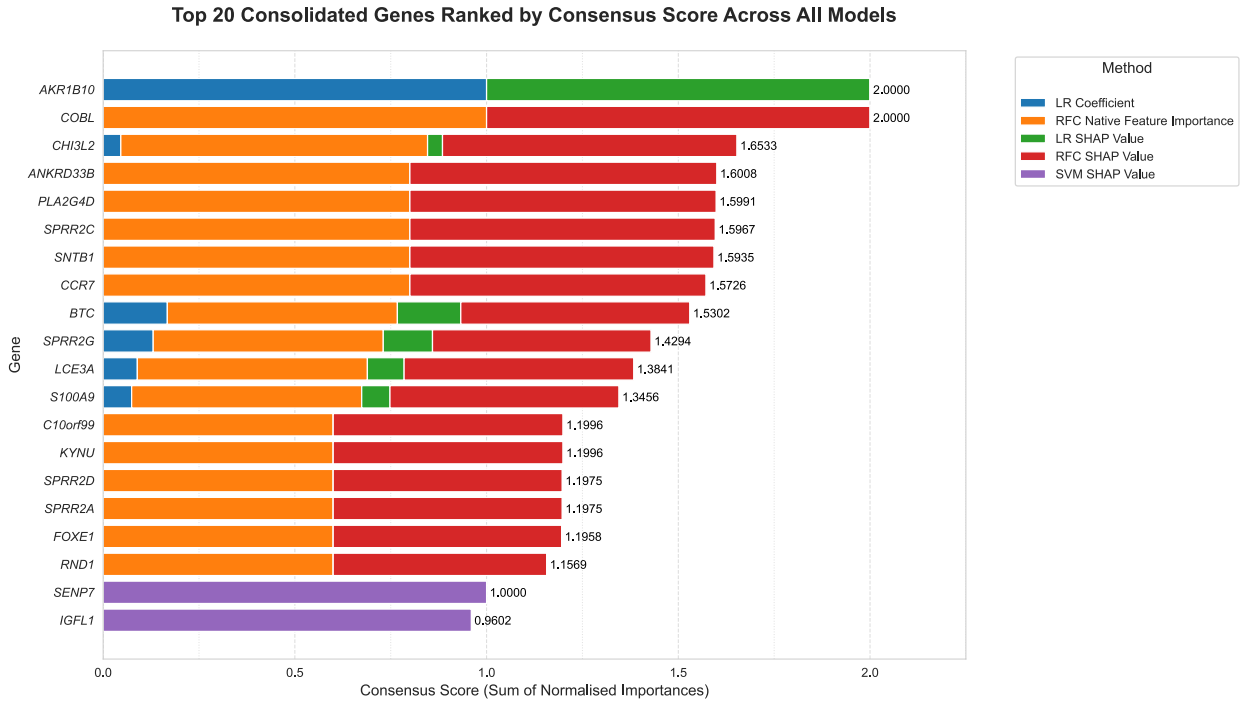


Figure 4.14 Top 20 Consolidated Genes Ranked by Consensus Score Across All Models

4.4.3 Quantitative Feature Overlap Analysis

A quantitative analysis was performed to assess the commonality of identified top genes (based on the top 20 features from each method) across the different models and importance metrics. Table 4.13 details the pairwise overlaps among the top 20 genes identified by each feature importance method, showing the number of genes common to every pair of methods, with the highest overlap being 19 genes between the LR coefficients and LR SHAP values analyses. Figure 4.15 provides a heatmap visualisation of the pairwise overlaps among the top 20 genes identified by each feature importance method, providing a quick overview of the agreement between methods.

The most striking result from this analysis was the unique feature set identified by the SVM. Despite achieving perfect classification accuracy on par with the other models, the SVM's top 20 features (Figure 4.13) had zero overlap with those from the LR and RFC approaches (Table 4.13 and Figure 4.15). The final visualisation (Figure 4.14) and the underlying data (Appendix B, Table B.1) further confirm this divergence, showing that the SVM's predictions were driven by an entirely separate group of genes. A full interpretation of this finding in the context of the

multiplicity problem is provided in Section 4.5, 'Interpretation of Results and Answers to Research Questions'.

Table 4.13 Pairwise Overlaps of Top 20 Genes Across Feature Importance Methods

Methods	Overlap Count	Common Genes
LR Coefficients versus RFC Native Feature Importance	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*
LR Coefficients versus LR SHAP Values	19	*ABCG4*, *AKR1B10*, *BTC*, *CHI3L2*, *CHRM4*, *DEFB4A*, *F3*, *FABP5*, *FUT3*, *INA*, *LCE3A*, *PI3*, *S100A7*, *S100A7A*, *S100A8*, *S100A9*, *SPRR2G*, *TAP2*, *TPBG*
LR Coefficients versus RFC SHAP Values	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*
RFC Native Feature Importance versus LR SHAP Values	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*
RFC Native Feature Importance versus RFC SHAP Values	17	*ANKRD33B*, *BTC*, *C10orf99*, *CCR7*, *CHI3L2*, *COBL*, *FOXE1*, *KYNU*, *LCE3A*, *PLA2G4D*, *RND1*, *S100A9*, *SNTB1*, *SPRR2A*, *SPRR2C*, *SPRR2D*, *SPRR2G*
LR SHAP Values versus RFC SHAP Values	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*

Note: Pairs with zero overlap are omitted for brevity (LR Coefficients versus SVM SHAP Values; RFC Native Feature Importance versus SVM SHAP Values; LR SHAP Values versus SVM SHAP Values; RFC SHAP Values versus SVM SHAP Values). See Figure 4.15 for the full pairwise overlap heatmap.

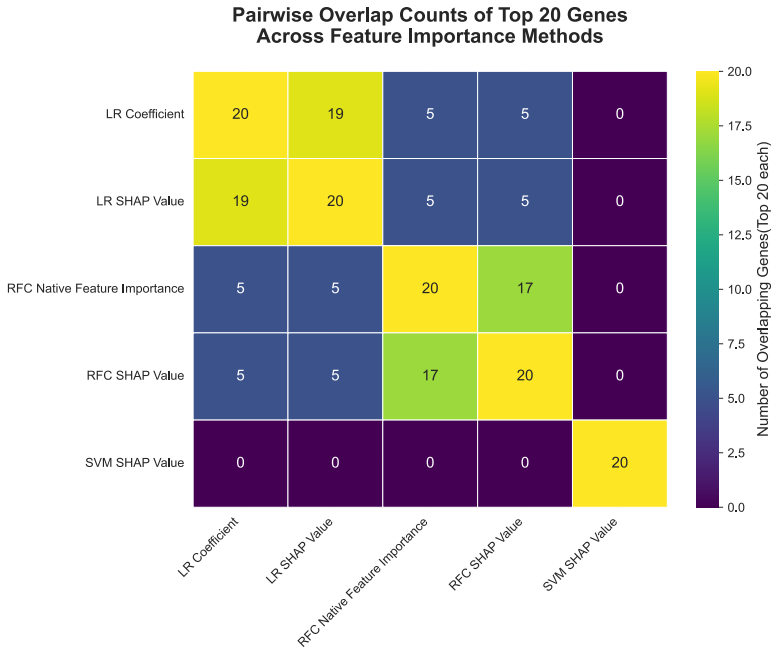


Figure 4.15 Pairwise Overlap Heatmap of Top 20 Genes Across Feature Importance Methods

Extending the pairwise overlap analysis (Table 4.13 and Figure 4.15), this investigation subsequently aimed to identify genes common to combinations of three or more feature importance methods, thereby highlighting genes consistently identified by multiple approaches.

Table 4.14 presents the unique genes found to be common across three feature importance methods. This analysis identified five unique genes (*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*) that demonstrated a moderate level of consensus by appearing in the top 20 features of at least three distinct methods.

Table 4.14 Genes Common to Three Feature Importance Methods (Top 20 Genes from Each Method)

Methods	Overlap Count	Common Genes
LR Coefficient, RFC Native Feature Importance, LR SHAP Value	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*
LR Coefficient, RFC Native Feature Importance, RFC SHAP Value	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*
LR Coefficient, LR SHAP Value, RFC SHAP Value	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*

RFC Native Feature Importance, LR SHAP Value, RFC SHAP Value	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*

Further analysis revealed genes consistently identified by four different methods, as detailed in Table 4.15. This tighter consensus highlighted these same five unique genes (*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*) that were common to three methods, indicating a high degree of robustness for this specific set of features as the consensus threshold increased.

Table 4.15 Genes Common to Four Feature Importance Methods (Top 20 Genes from Each Method)

Method	Overlap Count	Common Genes
LR Coefficient, RFC Native Feature Importance, LR SHAP Value, RFC SHAP Value	5	*BTC*, *CHI3L2*, *LCE3A*, *S100A9*, *SPRR2G*

Despite evaluating various combinations, no individual genes were found to be consistently present across the top 20 lists of all five feature importance methods, suggesting that perfect consensus on the most influential genes is not achieved under these conditions.

Figure 4.16 presents a bar chart summarising the total count of gene overlaps across all combinations for increasing numbers of methods in consensus. It quantitatively illustrates the level of consensus, showing a total of 56 genes for two methods (summed across combinations), 20 for three methods (summed across combinations) and five for four methods (summed across combinations), before the consensus drops to zero for all five methods.

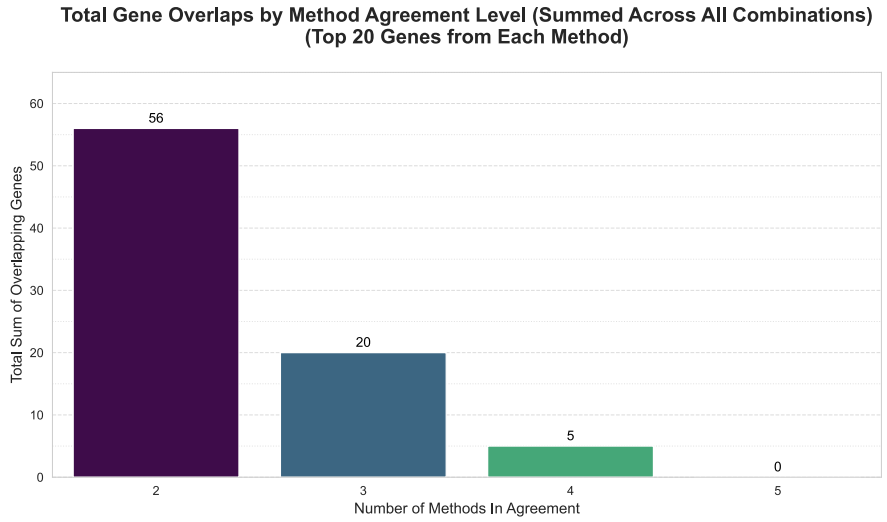


Figure 4.16 Total Gene Overlaps by Method Agreement Level (Summed Across All Combinations) (Top 20 Genes from Each Method)

To further clarify the distinction between total overlaps and unique genes, Figure 4.17 presents the total count of unique genes that reached each level of consensus, providing a direct visualisation of the breadth of the robust gene set. This bar chart shows 31 unique genes common to two methods, five unique genes common to three methods and five unique genes common to four methods, with zero unique genes common to all five methods.

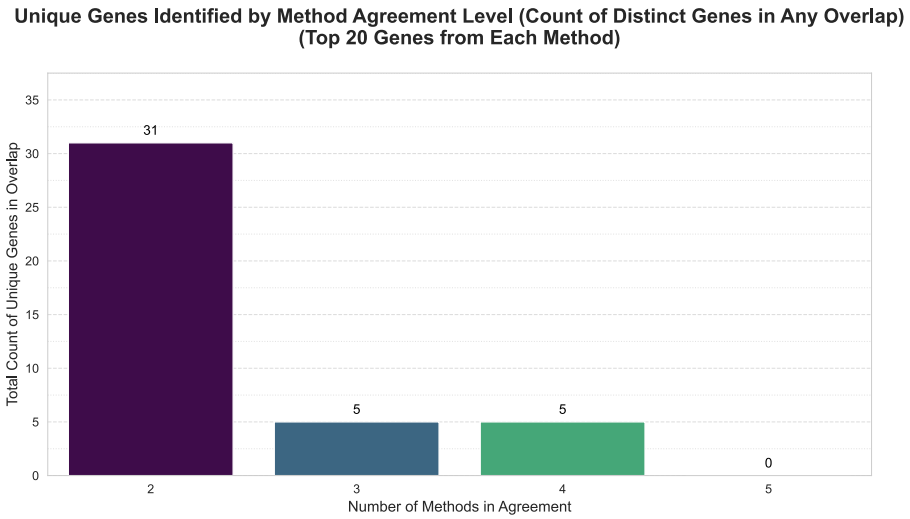


Figure 4.17 Unique Genes Identified by Method Agreement Level (Count of Distinct Genes in Any Overlap) (Top 20 Genes from Each Method)

4.5 Interpretation of Results and Answers to Research Questions

This study aimed to develop and compare supervised Machine Learning (ML) models for the accurate classification of psoriatic versus 'Healthy' skin samples using RNA-seq data, and to subsequently explore the key gene features underlying this classification. The results presented in this chapter provide comprehensive insights that directly address the proposed research questions.

RQ1: How accurately can optimised Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Machine (SVM) models distinguish psoriatic from 'Healthy' skin samples using the GSE54456 RNA-seq gene expression profile?

The evaluation on the independent test set revealed that all three optimised ML models – LR, RFC and SVM – achieved perfect classification performance, resulting in a score of 1.0000 across every metric measured. This included ROC AUC, accuracy, precision, recall and F1-score for all models (Table 4.11). This exceptional discriminative ability was highly consistent, as cross-validation ROC AUC scores during hyperparameter tuning were also uniformly 1.0000 for all models. This perfect performance was consistent with the clear separability visually observed in the initial Principal Component Analysis (PCA) of the scaled, variance-selected training data (Figure 3.2). Although PCA is unsupervised, the visualisation provided an early indication that the GSE54456 gene expression profiles appeared separable into 'Healthy' and 'Disease' groups.

RQ2: How do the sets of key gene features identified by different models and interpretability methods compare, and what does this reveal about the *multiplicity problem* in a high-dimensional dataset?

While all three classifiers demonstrated perfect predictive performance, the comparative analysis of their feature importance lists revealed a crucial finding. The feature sets identified by the LR and RFC models showed a moderate and meaningful degree of overlap, especially between their respective SHAP analyses and native importance metrics. However, a particularly noteworthy outcome was the complete divergence of the features identified by the SVM, which achieved its perfect score by leveraging a set of top 20 genes that had zero overlap with those identified by the LR and RFC approaches (Table 4.13). The final visualisation (Figure 4.14) and the detailed data in the consolidated feature importance table (Appendix B, Table B.1) provide the definitive

quantitative evidence for this conclusion, showing precisely how the SVM achieved its perfect score while ignoring the genes prioritised by all other methods. This finding is a powerful, practical illustration of the *multiplicity problem* in high-dimensional data, where different, equally valid models can find distinct feature sets to achieve the same predictive goal. It strongly validates the core rationale for this study, demonstrating that a multi-model and consensus-based approach is essential for moving beyond model-specific idiosyncrasies to identify a truly robust and model-agnostic gene signature.

RQ3: What is the robust, consensus-based gene signature identified by a multi-model, multi-method comparative analysis, and how is this signature validated by the existing biological and clinical literature on psoriasis?

The quantitative overlap analysis proved essential in filtering through the model-specific preferences to answer this question. The analysis revealed a highly reliable, core consensus signature of five unique genes: *BTC*, *CHI3L2*, *LCE3A*, *S100A9* and *SPRR2G*. This set of genes was consistently identified within the top 20 most influential features by four of the five different feature importance methods (Table 4.15). Crucially, the composition of the final Consensus Score (Table 4.12) provides a powerful internal validation for this five-gene signature, as it clearly shows that these are precisely the genes that derive their high scores from a consensus across the same four feature importance methods. The sharp drop-off in consensus from 31 unique genes shared by any two methods to just these five shared by four (Figure 4.17) strongly indicates that this signature represents the fundamental, model-agnostic signal for psoriasis classification within this dataset.

The ultimate validation of this data-driven signature is its support from the existing biological and clinical literature. As demonstrated in the following section (4.6, 'Comparison to Existing Literature'), each of these five genes is deeply implicated in key aspects of psoriasis pathophysiology, confirming that the methodology successfully identified a set of biologically meaningful and clinically relevant biomarkers.

4.6 Comparison to Existing Literature

The successful development of high-accuracy classifiers is a pivotal first step, but the true value of a ML approach in a biomedical context lies in its ability to provide biologically meaningful insights. The ultimate test of this study's methodology is not just its predictive power, but whether the gene features it identifies as most important can be validated by and understood through the existing body of biological and clinical literature. As this section will demonstrate, the key gene features identified through our multi-model consensus approach not only align with previous large-scale transcriptomic studies but are also validated by mechanistic and multi-omic research, underscoring their biological relevance in the pathogenesis of psoriasis.

A primary outcome of this research was the identification of a highly robust set of five consensus genes – *BTC*, *CHI3L2*, *LCE3A*, *S100A9* and *SPRR2G* – that were consistently ranked within the top 20 features across four different feature importance methods. This finding is strongly corroborated by the literature.

The consistent identification of S100 calcium-binding protein A9 (*S100A9*) as a top-ranked discriminative feature by our models provides a powerful empirical validation of its known role as a key functional mediator of the disease. Proteomic analysis by Schonthaler et al. (2013) identified the S100A8–S100A9 complex as the most upregulated protein in psoriatic lesions, where it drives inflammation through regulation of the complement system. The massive upregulation of *S100A9* within psoriatic keratinocytes is so significant that it leads to elevated serum levels that directly correlate with the Psoriasis Area and Severity Index (PASI) (Benoit et al., 2006), establishing it as a validated clinical biomarker.

Our models' consistent selection of Epidermal Differentiation Complex genes, such as Late Cornified Envelope protein 3A (*LCE3A*) and Small Proline-Rich protein 2G (*SPRR2G*), is also well-supported. Bergboer et al. (2011) demonstrated that the *LCE3* gene group, including *LCE3A*, is specifically upregulated in psoriatic lesions in response to injury, suggesting a role in skin barrier repair. Most compellingly, our ML approach independently identified two key interacting partners of a pathogenic protein complex, as Tian et al. (2021) showed that the *LCE* and *SPRR* gene families form a functional and physical interaction module in psoriasis. Their work not only confirmed the significant upregulation of both *LCE3A* and *SPRR2G* in psoriasis but also demonstrated direct protein–protein interactions. This highlights our methodology's ability to uncover biologically meaningful, co-regulated gene sets and provides powerful external validation for our findings.

Our models' ability to identify Betacellulin (*BTC*) as a top feature, particularly as the gene with the largest negative coefficient, directly validates the findings of earlier large-scale studies. For example, Gudjonsson et al. (2010), in a large microarray study, also highlighted *BTC* as one of the most strongly downregulated genes in psoriasis, confirming that our ML learning approach successfully captured this well-documented downregulation event in the psoriatic transcriptome. Its importance is further reinforced by other ML analyses: Ainali et al. (2012), using an RFC, identified *BTC* as one of the top five most important genes for discriminating between psoriatic and non-lesional skin. Mechanistically, the downregulation of *BTC* is likely a direct consequence of the pro-inflammatory cytokine milieu, as studies have shown that treating skin models with novel IL-1 family cytokines, which are highly active in psoriasis, consistently downregulates *BTC* expression (Johnston et al., 2011).

Completing the consensus signature, the inclusion of Chitinase 3-Like 2 (*CHI3L2*) is strongly supported by recent ML studies investigating the shared molecular pathology between psoriasis and its comorbidities. In an integrated bioinformatic analysis, J. Zhang et al. (2024) identified a set of eight crucial genes common to both psoriasis and ulcerative colitis, with *CHI3L2* being a core member of this shared signature. Their work highlights that *CHI3L2* is known to be involved in inducing autoimmune responses and tissue remodelling and their analysis further demonstrated a close correlation between *CHI3L2* expression and the infiltration of various immune cells in patients with ulcerative colitis. The consistent identification of *CHI3L2* by our models therefore aligns with this independent research, suggesting that this gene is a robust and meaningful biomarker whose role in psoriatic inflammation warrants further investigation.

Beyond the consensus list, the single most dominant feature identified by our LR model was Aldo-Keto Reductase family 1 member B10 (*AKR1B10*). This finding is powerfully validated by the work of Gao et al. (2017) who also analysed the GSE54456 dataset and identified *AKR1B10* as one of the most highly overexpressed genes. This study went further, using small interfering RNA (siRNA) knockdown experiments to prove that *AKR1B10* plays a direct functional role in promoting the overproliferation and migration of keratinocytes. Our model, therefore, successfully identified not just a statistically powerful feature, but one that has been experimentally proven to be a driver of a key pathological process in psoriasis.

Finally, the methodological approach of this study is consistent with modern trends in bioinformatics. Other recent studies have successfully used ML pipelines including RFC and

LASSO, on the GSE54456 and other GEO datasets to identify biomarkers and build evaluation models that correlate with clinical severity (Hu et al., 2022; Zhou et al., 2023b; Wang et al., 2022). While these studies sometimes converge on different gene sets (e.g., Wang et al., 2022, who highlighted *ARG1* and *CXCL2*), this underscores the complexity of the psoriasis transcriptome, where different analytical strategies can reveal distinct but equally valid aspects of the disease's pathology. Our work contributes to this field by demonstrating a robust, multi-model consensus approach that successfully identifies a core set of genes strongly validated by existing biological, clinical and multi-omic evidence.

4.7 Limitations

Despite the robust findings, this study has several limitations that should be acknowledged. Firstly, the analysis is based solely on a single publicly available dataset, GSE54456. While this dataset is large for its type and the signals within it are strong, reliance on a single source may limit the generalisability of the specific gene weightings and rankings to broader patient populations or different experimental conditions. Secondly, although the dataset's RPKM normalisation was accounted for through $\log_2(\text{RPKM} + 1)$ transformation, inherent limitations of RPKM for between-sample comparisons could introduce subtle biases, as acknowledged in the methodology. Further studies could explore the impact of alternative normalisation methods (e.g., TPM or count-based methods). Thirdly, while variance-based selection served as an effective initial feature selection strategy, other sophisticated methods exist – significance-based filter methods (e.g., *t*-tests), wrapper methods and embedded methods – that might identify slightly different sets of optimal features. Finally, as a computational bioinformatics study, direct experimental validation of the interactions between the identified genes was outside the scope of this work. While Section 4.6, 'Comparison to Existing Literature', confirms that many of the individual genes in the consensus signature have established functional roles in psoriasis, their collective action as a multi-gene biomarker panel remains a data-driven hypothesis that awaits direct experimental confirmation.

4.8 Future Work

In light of the findings and addressing the limitations of this study, several avenues for future research emerge. Given the exceptional and uniform classification performance achieved, future work should pivot from improving accuracy to focusing on the generalisability, biological validation and clinical translation of the identified gene signature.

Firstly, the most critical next step is to focus on external validation. This involves applying the developed multi-model consensus methodology to larger, independent RNA-seq datasets of psoriasis. This is crucial to validate the five-gene signature and assess its robustness across more diverse patient cohorts and experimental conditions, thereby confirming that the discovered associations are not specific to the GSE54456 dataset. An excellent resource for this is the manually curated data collection by Federico et al. (2020), which provides several RNA-seq datasets, including GSE47944, a dataset of 84 samples suitable for robust validation.

Secondly, there is significant scope for methodological expansion. Future work could explore more advanced feature selection techniques, particularly those incorporating biological network information, which might refine the set of discriminative genes and reveal other important pathways, especially considering that perfect consensus was not achieved across five feature importance methods. Furthermore, while classification performance was optimal, investigating a wider range of ML algorithms, including deep learning approaches, would be valuable for more complex tasks such as predicting therapeutic response or disease prognosis. For example, models like Deep Neural Networks (DNNs) and autoencoders could be explored for their ability to learn hierarchical features from transcriptomic data. Integrating other omics data (e.g., genomics, proteomics, metabolomics) alongside transcriptomics could also provide a more holistic understanding of psoriasis, potentially enhancing predictive power for these advanced applications.

Finally, the ultimate goal is biological and clinical translation. The identified key gene features warrant focused *in vitro* and *in vivo* investigation to confirm their causal roles in psoriasis pathogenesis. While the literature confirms the individual importance of genes like *S100A9* and *AKR1B10*, functional studies should aim to elucidate the precise interactions between the five consensus genes. A key priority should be exploring the mechanistic role of the less-characterised gene in the signature, *CHI3L2*, to fully understand its potential as a novel diagnostic biomarker or therapeutic target. Successfully transitioning from a high-performing classification model to a

validated multi-gene prognostic tool would represent a significant clinical advancement, building on the strong and biologically relevant signal uncovered in this study.

Chapter 5: Conclusion

This research successfully developed and rigorously evaluated a multi-model Machine Learning (ML) pipeline for the classification of psoriatic versus 'Healthy' skin using the high-dimensional GSE54456 RNA-sequencing (RNA-seq) dataset. The study confirmed that with robust pre-processing and feature selection, supervised learning models can achieve perfect classification accuracy, underscoring the strong and distinct molecular signature present in psoriatic tissue. All three optimised models – Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Machine (SVM) – demonstrated uniformly perfect performance across all evaluation metrics, including an Area Under the Receiver Operating Characteristic Curve (ROC AUC) = 1.0000 on the independent test set.

The primary contribution of this work, however, lies not in the classification performance itself, but in the robust multi-faceted approach to feature identification. This comparative methodology successfully addressed the *multiplicity problem*, which was powerfully illustrated by the complete divergence of the SVM's feature set despite its perfect predictive accuracy. The final Consensus Score analysis not only provided the definitive quantitative evidence of this divergence (Table 4.12) but also internally validated the core consensus signature of five highly robust genes – *BTC*, *CHI3L2*, *LCE3A*, *S100A9* and *SPRR2G* – that were initially identified through overlap counts.

Crucially, this data-driven signature was shown to be strongly validated by the existing biological and clinical literature. Each gene in the consensus set is deeply implicated in key aspects of psoriasis pathophysiology, including the skin barrier function (*LCE3A*, *SPRR2G*, *BTC*), innate immunity and inflammation (*S100A9*) and autoimmune responses (*CHI3L2*). Furthermore, the single most influential gene identified by the LR model, *AKR1B10*, was also confirmed to be a functionally significant driver of keratinocyte hyperproliferation.

While the findings are promising, the study's limitations must be acknowledged. The analysis was based on a single, albeit large, public dataset and as a computational study it awaits direct experimental validation of the identified gene functions. Further work should therefore focus on applying this methodology to larger, independent patient cohorts to confirm the generalisability of the gene signature. However, it is a major strength of this study that many of the identified consensus genes already have established functional roles in psoriasis according to the existing literature. This provides a strong external validation of our data-driven approach. Therefore,

further functional studies should focus on elucidating the precise interactions between these key genes and exploring the mechanistic roles of less-characterised genes in the signature, such as *CHI3L2*, to fully understand their potential as novel diagnostic biomarkers or therapeutic targets.

In conclusion, this dissertation successfully demonstrates the power of a multi-model, consensus-based ML approach not only to build high-accuracy classifiers but also to identify a core set of biologically meaningful and externally validated gene features from a complex transcriptomic dataset. This work validates a robust methodology for biomarker discovery and serves as a strong proof of concept for the future application of comparative ML and interpretability frameworks in biomedical research, contributing to a deeper understanding of the molecular mechanisms that drive psoriasis.

References

Ainali, C., Valeyev, N., Perera, G., Williams, A., Gudjonsson, J. E., Ouzounis, C. A., ... & Tsoka, S. (2012). Transcriptome classification reveals molecular subtypes in psoriasis. *BMC genomics*, *13*(1), 472.

Armstrong, A. W., & Read, C. (2020). Pathophysiology, clinical presentation, and treatment of psoriasis: a review. *Jama*, *323*(19), 1945-1960.

Babaie, F., Omraninava, M., Gorabi, A. M., Khosrojerdi, A., Aslani, S., Yazdchi, A., ... & Sahebkar, A. (2022). Etiopathogenesis of psoriasis from genetic perspective: an updated review. *Current Genomics*, *23*(3), 163-174.

Benoit, S., Toksoy, A., Ahlmann, M., Schmidt, M., Sunderkötter, C., Foell, D., ... & Goebeler, M. (2006). Elevated serum levels of calcium-binding S100 proteins A8 and A9 reflect disease activity and abnormal differentiation of keratinocytes in psoriasis. *British Journal of Dermatology*, *155*(1), 62-66.

Bergboer, J. G., Tjabringa, G. S., Kamsteeg, M., van Vlijmen-Willems, I. M., Rodijk-Olthuis, D., Jansen, P. A., ... & Schalkwijk, J. (2011). Psoriasis risk genes of the late cornified envelope-3 group are distinctly expressed compared with genes of other LCE groups. *The American journal of pathology*, *178*(4), 1470-1477.

Bowcock, A. M., Shannon, W., Du, F., Duncan, J., Cao, K., Aftergut, K., ... & Menter, A. (2001). Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Human molecular genetics, 10*(17), 1793-1805.

Bu, J., Ding, R., Zhou, L., Chen, X., & Shen, E. (2022). Epidemiology of psoriasis and comorbid diseases: a narrative review. *Frontiers in immunology, 13*, 880201.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell, 173*(7), 1581-1592.

Campanati, A., Marani, A., Martina, E., Diotallevi, F., Radi, G., & Offidani, A. (2021). Psoriasis as an immune-mediated and inflammatory systemic disease: from pathophysiology to novel therapeutic approaches. *Biomedicines*, *9*(11), 1511.

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L. J., Murphy, B., Mayes, A. E., ... & Pyzer-Knapp, E. O. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Scientific reports*, *11*(1), 4565.

Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N., & Liao, W. (2020). Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatology and therapy*, *10*(3), 365-386.

Cheng, Y., Xu, S. M., Santucci, K., Lindner, G., & Janitz, M. (2024). Machine learning and related approaches in transcriptomics. *Biochemical and Biophysical Research Communications*, 150225.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology, 17*, 1-19.

Dan, W., Lv, S., Gao, W., Liao, X., Wang, Z., & Zhang, G. (2025). Identification of characteristic genes and herbal compounds for the treatment of psoriasis based on machine learning and molecular dynamics simulation. *Journal of Biomolecular Structure and Dynamics*, *43*(5), 2646-2665.

Dand, N., Mahil, S. K., Capon, F., Smith, C. H., Simpson, M. A., & Barker, J. N. (2020). Psoriasis and genetics. *Acta dermato-venereologica*, *100*(100-year theme: Psoriasis), 54-64.

Dand, N., Stuart, P. E., Bowes, J., Ellinghaus, D., Nititham, J., Saklatvala, J. R., ... & Elder, J. T. (2025). GWAS meta-analysis of psoriasis identifies new susceptibility alleles impacting disease mechanisms and therapeutic targets. *Nature communications, 16*(1), 2051.

De Cid, R., Riveira-Munoz, E., Zeeuwen, P. L., Robarge, J., Liao, W., Dannhauser, E. N., ... & Estivill, X. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature genetics*, *41*(2), 211-215.

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics, 7*, 1-13.

Federico, A., Hautanen, V., Christian, N., Kremer, A., Serra, A., & Greco, D. (2020). Manually curated and harmonised transcriptomics datasets of psoriasis and atopic dermatitis patients. *Scientific data*, *7*(1), 343.

Feldner-Busztin, D., Firbas Nisantzis, P., Edmunds, S. J., Boza, G., Racimo, F., Gopalakrishnan, S., ... & de Polavieja, G. G. (2023). Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics, 39*(2), btad021.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*(8), 1761-1776.

Gao, Y., Yi, X., & Ding, Y. (2017). Combined transcriptomic analysis revealed AKR1B10 played an important role in psoriasis through the dysregulated lipid pathway and overproliferation of keratinocyte. *BioMed research international*, *2017*(1), 8717369.

Ghorbian, M., Ghobaei-Arani, M., & Ghorbian, S. (2024). A comprehensive study on the application of machine learning in psoriasis diagnosis and treatment: taxonomy, challenges and recommendations. *Artificial Intelligence Review, 58*(2), 60.

Gudjonsson, J. E., Ding, J., Johnston, A., Tejasvi, T., Guzman, A. M., Nair, R. P., ... & Elder, J. T. (2010). Assessment of the psoriatic transcriptome in a large sample: additional regulated genes and comparisons with in vitro models. *Journal of Investigative Dermatology*, *130*(7), 1829-1840.

Guo, P., Luo, Y., Mai, G., Zhang, M., Wang, G., Zhao, M., ... & Zhou, F. (2014). Gene expression profile based classification models of psoriasis. *Genomics, 103*(1), 48-55.

Henseler, T. (1997). The genetics of psoriasis. *Journal of the American Academy of Dermatology*, *37*(2), S1-S11.

Hu, Y., Jiang, L., Lei, L., Luo, L., Guo, H., Zhou, Y., ... & Zeng, Q. (2022). Establishment and validation of psoriasis evaluation models. *Fundamental Research*, *2*(1), 166-176.

Jabbari, A., Suárez-Fariñas, M., Dewell, S., & Krueger, J. G. (2011). Transcriptional profiling of psoriasis using RNA-seq reveals previously unidentified differentially expressed genes. *The Journal of investigative dermatology, 132*(1), 246.

Johnston, A., Xing, X., Guzman, A. M., Riblett, M., Loyd, C. M., Ward, N. L., ... & Gudjonsson, J. E. (2011). IL-1F5,-F6,-F8, and-F9: a novel IL-1 family signaling system that is active in psoriasis and promotes keratinocyte antimicrobial peptide expression. *The Journal of Immunology*, *186*(4), 2613-2622.

Kanda, N. (2021). Psoriasis: pathogenesis, comorbidities, and therapy updated. *International journal of molecular sciences*, *22*(6), 2979.

Kõks, S., Keermann, M., Reimann, E., Prans, E., Abram, K., Silm, H., ... & Kingo, K. (2016). Psoriasis-specific RNA isoforms identified by RNA-seq analysis of 173,446 transcripts. *Frontiers in Medicine*, *3*, 46.

Krishnan, V. S., & Kõks, S. (2022). Transcriptional basis of psoriasis from large scale gene expression studies: the importance of moving towards a precision medicine approach. *International Journal of Molecular Sciences, 23*(11), 6130.

Krishnan, V. S., & Kõks, S. (2023). Transcriptional landscape of repetitive elements in psoriatic skin from large cohort studies: relevance to psoriasis pathophysiology. *International journal of molecular sciences*, *24*(23), 16725.

Krueger, J. G., & Bowcock, A. (2005). Psoriasis pathophysiology: current concepts of pathogenesis. *Annals of the rheumatic diseases*, *64*(suppl 2), ii30-ii36.

Leung, A., Marquez-Grap, G., Kranyak, A., & Liao, W. (2025). A review of spatial transcriptomics in psoriasis: new insights into cellular contributions. *Current Opinion in Immunology*, *95*, 102585.

Li, B., Tsoi, L. C., Swindell, W. R., Gudjonsson, J. E., Tejasvi, T., Johnston, A., ... & Elder, J. T. (2014a). Transcriptome analysis of psoriasis in a large case–control sample: RNA-seq provides insights into disease mechanisms. *Journal of Investigative Dermatology, 134*(7), 1828-1838.

Li, B., Tsoi, L. C., Swindell, W. R., Gudjonsson, J. E., Tejasvi, T., Johnston, A., ... & Elder, J. T. (2014b). *Transcriptome analysis of psoriasis in a large case–control sample: RNA-seq provides insights into disease mechanisms* [Data set]. NCBI Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54456

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1-45.

Liu, Y., Krueger, J. G., & Bowcock, A. M. (2007). Psoriasis: genetic associations and immune system changes. *Genes & Immunity*, *8*(1), 1-12.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS computational biology*, *13*(5), e1005457.

Mateu-Arrom, L., & Puig, L. (2023). Genetic and Epigenetic mechanisms of psoriasis. *Genes, 14*(8), 1619.

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current opinion in chemical biology, 17*(1), 4-11.

McMullen, E. P., Al Naser, Y. A., Maazi, M., Grewal, R. S., Abdel Hafeez, D., Folino, T. R., & Vender, R. B. (2025). Predicting psoriasis severity using machine learning: a systematic review. *Clinical and Experimental Dermatology*, *50*(3), 520-528.

Mitić, N. S., Malkov, S. N., Maljković Ružičić, M. M., Veljković, A. N., Čukić, I. L., Lin, X., ... & Brusić, V. (2025). Correlation-based feature selection of single cell transcriptomics data from multiple sources. *Journal of Big Data, 12*(1), 4.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods, 5*(7), 621-628.

National Center for Biotechnology Information. (n.d.). *Gene Expression Omnibus (GEO).* U.S. National Library of Medicine. Retrieved May 24, 2025 from https://www.ncbi.nlm.nih.gov/geo/

Ogawa, K., & Okada, Y. (2020). The current landscape of psoriasis genetics in 2020. *Journal of dermatological science, 99*(1), 2-8.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics, 12*(2), 87-98.

Peng, G., Tsukamoto, S., Umehara, Y., Kishi, R., Tominaga, M., Takamori, K., ... & Niyonsaba, F. (2022). Experimental and clinical evidence suggests that treatment with betacellulin can alleviate Th2-type cytokine-mediated impairment of skin barrier function. *International Journal of Molecular Sciences*, *23*(19), 11520.

Raharja, A., Mahil, S. K., & Barker, J. N. (2021). Psoriasis: a brief overview. *Clinical Medicine, 21*(3), 170-173.

Rioux, G., Ridha, Z., Simard, M., Turgeon, F., Guérin, S. L., & Pouliot, R. (2020). Transcriptome profiling analyses in psoriasis: a dynamic contribution of keratinocytes to the pathogenesis. *Genes*, *11*(10), 1155.

Roberson, E. D., & Bowcock, A. M. (2010). Psoriasis genetics: breaking the barrier. *Trends in Genetics*, *26*(9), 415-423.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507-2517.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Schonthaler, H. B., Guinea-Viniegra, J., Wculek, S. K., Ruppen, I., Ximénez-Embún, P., Guío-Carrión, A., ... & Wagner, E. F. (2013). S100A8-S100A9 protein complex mediates psoriasis by regulating the expression of complement factor C3. *Immunity*, *39*(6), 1171-1181.

Shefler, A., Patrick, M. T., Wasikowski, R., Chen, J., Sarkar, M. K., Gudjonsson, J. E., & Tsoi, L. C. (2022). Skin-expressing lncRNAs in inflammatory responses. *Frontiers in genetics*, *13*, 835740.

Shen, C., Gao, J., Yin, X., Sheng, Y., Sun, L., Cui, Y., & Zhang, X. (2015). Association of the late cornified envelope-3 genes with psoriasis and psoriatic arthritis: a systematic review. *Journal of Genetics and Genomics*, *42*(2), 49-56.

Solmaz, D., Bakirci, S., Kimyon, G., Gunal, E. K., Dogru, A., Bayindir, O., ... & Aydin, S. Z. (2020). Impact of having family history of psoriasis or psoriatic arthritis on psoriatic disease. *Arthritis care & research*, *72*(1), 63-68.

Svedbom, A., Mallbris, L., Larsson, P., Nikamo, P., Wolk, K., Kjellman, P., ... & Ståhle, M. (2021). Long-term outcomes and prognosis in new-onset psoriasis. *JAMA dermatology, 157*(6), 684-690.

Swindell, W. R., Johnston, A., Voorhees, J. J., Elder, J. T., & Gudjonsson, J. E. (2013). Dissecting the psoriasis transcriptome: inflammatory-and cytokine-driven gene expression in lesions from 163 patients. *BMC genomics, 14*, 1-20.

Tapak, L., Afshar, S., Afrasiabi, M., Ghasemi, M. K., & Alirezaei, P. (2021). Application of genetic algorithm-based support vector machine in identification of gene expression signatures for psoriasis classification: a hybrid model. *BioMed Research International, 2021*(1), 5520710.

Tian, S., Chen, S., Feng, Y., & Li, Y. (2021). The interactions of small proline-rich proteins with late cornified envelope proteins are involved in the pathogenesis of psoriasis. *Clinical, Cosmetic and Investigational Dermatology*, 1355-1365.

Vanitha, C. D. A., Devaraj, D., & Venkatesulu, I. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. *procedia computer science*, *47*, 13-21.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.

Wang, H., Chen, W., He, J., Xu, W., & Liu, J. (2021). Network analysis of potential risk genes for psoriasis. *Hereditas*, *158*(1), 21.

Wu, J., Fang, Z., Liu, T., Hu, W., Wu, Y., & Li, S. (2021). Maximizing the utility of transcriptomics data in inflammatory skin diseases. *Frontiers in Immunology, 12*, 761890.

Yagin, F. H., Cicek, I. B., Alkhateeb, A., Yagin, B., Colak, C., Azzeh, M., & Akbulut, S. (2023). Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. *Computers in Biology and Medicine*, *154*, 106619.

Yamanaka, K., Yamamoto, O., & Honda, T. (2021). Pathophysiology of psoriasis: A review. *The Journal of dermatology*, *48*(6), 722-731.

Yamazaki, F. (2021). Psoriasis: comorbidities. *The Journal of dermatology*, *48*(6), 732-740.

Yu, K., Syed, M. N., Bernardis, E., & Gelfand, J. M. (2020). Machine learning applications in the evaluation and management of psoriasis: a systematic review. *Journal of Psoriasis and Psoriatic Arthritis®*, *5*(4), 147-159.

Zhang, J., Feng, S., Chen, M., Zhang, W., Zhang, X., Wang, S., ... & Wang, G. (2024). Identification of potential crucial genes shared in psoriasis and ulcerative colitis by machine learning and integrated bioinformatics. *Skin Research and Technology*, *30*(2), e13574.

Zhang, M. J., Xue, T. T., Fei, X. Y., Zhang, Y., Luo, Y., Ru, Y., ... & Li, B. (2024). Identification of angiogenesis-related genes and molecular subtypes for psoriasis based on random forest algorithm. *Clinical and Experimental Immunology*, *218*(2), 199-212.

Zhou, X., Chen, Y., Cui, L., Shi, Y., & Guo, C. (2022). Advances in the pathogenesis of psoriasis: from keratinocyte perspective. *Cell death & disease*, *13*(1), 81.

Zhou, Y., Han, L., Wang, Z., Fang, R., Wan, Y., Yang, Z., ... & Ni, Q. (2023a). Bioinformatic analysis of the potential common pathogenic mechanisms for psoriasis and metabolic syndrome. *Inflammation*, *46*(4), 1381-1395.

Zhou, Y., Wang, Z., Han, L., Yu, Y., Guan, N., Fang, R., ... & Li, J. (2023b). Machine learning-based screening for biomarkers of psoriasis and immune cell infiltration. *European Journal of Dermatology*, *33*(2), 147-156.

Appendices

Appendix A: Ethics Approval Certificate

This appendix reproduces the University Research Ethics Committee Approval Certificate for this study (Application Reference Number: 068508; approved 23/05/2025). The certificate confirms that the study used existing, anonymised data and required a self-declaration only.

The University Of Sheffield.

Downloaded: 23/05/2025
Approved: 23/05/2025

Robert Jacques
Registration number: 240175267
Information School [a.k.a iSchool]
Programme: MSc Data Science

Dear Robert

PROJECT TITLE: Machine Learning-Based Classification and Feature Identification for Psoriasis
APPLICATION: Reference Number 068508

This letter confirms that you have signed a University Research Ethics Committee-approved self-declaration to confirm that your research will involve only existing research, clinical or other data that has been robustly anonymised. You have judged it to be unlikely that this project would cause offence to those who originally provided the data, should they become aware of it.

As such, on behalf of the University Research Ethics Committee, I can confirm that your project can go ahead on the basis of this self-declaration.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since full ethical review may be required.

Yours sincerely

Zeyneb Kurt
Departmental Ethics Administrator

Figure A.1 Ethics Approval Certificate (Application Reference Number: 068508; approved 23/05/2025)

Appendix B: Supplementary Results – Consolidated Feature Importance Table

This table reports the full, method-specific importance values (min–max normalised to [0, 1]) for every gene that appeared in the top 20 of at least one method (LR coefficients; RFC native feature importance; SHAP values for LR, RFC and SVM). It provides the source data for the Consensus Scores in Section 4.4.2, 'Consolidated Top Gene Features'.

Table B.1 Consolidated Feature Importance Table

Gene	Consensus Score	LR Coefficient	RFC Native Feature Importance	LR SHAP Value	RFC SHAP Value	SVM SHAP Value
AKR1B10	2.0000	1.0000	0.0000	1.0000	0.0000	0.0000
COBL	2.0000	0.0000	1.0000	0.0000	1.0000	0.0000
CHI3L2	1.6533	0.0461	0.8000	0.0389	0.7682	0.0000
ANKRD33B	1.6008	0.0000	0.8000	0.0000	0.8008	0.0000
PLA2G4D	1.5991	0.0000	0.8000	0.0000	0.7991	0.0000
SPRR2C	1.5967	0.0000	0.8000	0.0000	0.7967	0.0000
SNTB1	1.5935	0.0000	0.8000	0.0000	0.7935	0.0000
CCR7	1.5726	0.0000	0.8000	0.0000	0.7726	0.0000
BTC	1.5302	0.1674	0.6000	0.1654	0.5975	0.0000
SPRR2G	1.4294	0.1305	0.6000	0.1284	0.5704	0.0000
LCE3A	1.3841	0.0891	0.6000	0.0958	0.5992	0.0000
S100A9	1.3456	0.0743	0.6000	0.0737	0.5975	0.0000
C10orf99	1.1996	0.0000	0.6000	0.0000	0.5996	0.0000
KYNU	1.1996	0.0000	0.6000	0.0000	0.5996	0.0000
SPRR2D	1.1975	0.0000	0.6000	0.0000	0.5975	0.0000
SPRR2A	1.1975	0.0000	0.6000	0.0000	0.5975	0.0000
FOXE1	1.1958	0.0000	0.6000	0.0000	0.5958	0.0000
RND1	1.1569	0.0000	0.6000	0.0000	0.5569	0.0000
SENP7	1.0000	0.0000	0.0000	0.0000	0.0000	1.0000
IGFL1	0.9602	0.0000	0.0000	0.0000	0.0000	0.9602

Note: All values are reported to four decimal places. Method-specific columns are min–max normalised to [0, 1]; a value of 0.0000 indicates the gene did not appear in that method's top 20 list. The Consensus Score is the sum of the five normalised values (range 0.0000–5.0000).

Appendix C: Supplementary Visualisations – Per-sample Gene Expression Distributions

This appendix provides per-sample gene expression distributions to complement the cohort-level plot in Figure 3.1. A set of four representative samples – two 'Healthy' and two 'Disease' – were selected at random (`RANDOM_STATE = 42`). Figure C.1 shows the distributions before the $\log_2(\text{RPKM} + 1)$ transform; Figure C.2 shows the same samples after the transform, illustrating reduced right-skew and variance stabilisation.

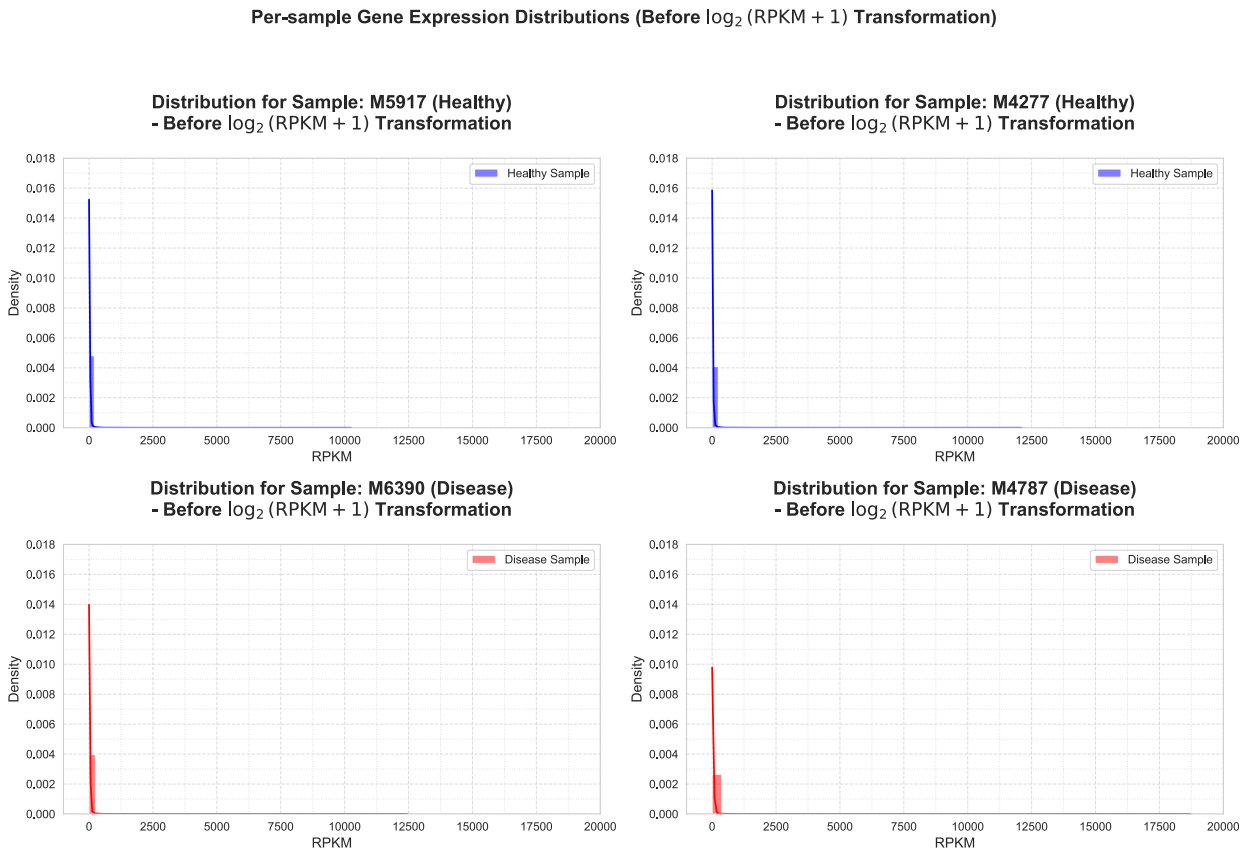


Figure C.1 Per-sample Gene Expression Distributions (Before $log_2(RPKM + 1)$ Transformation)

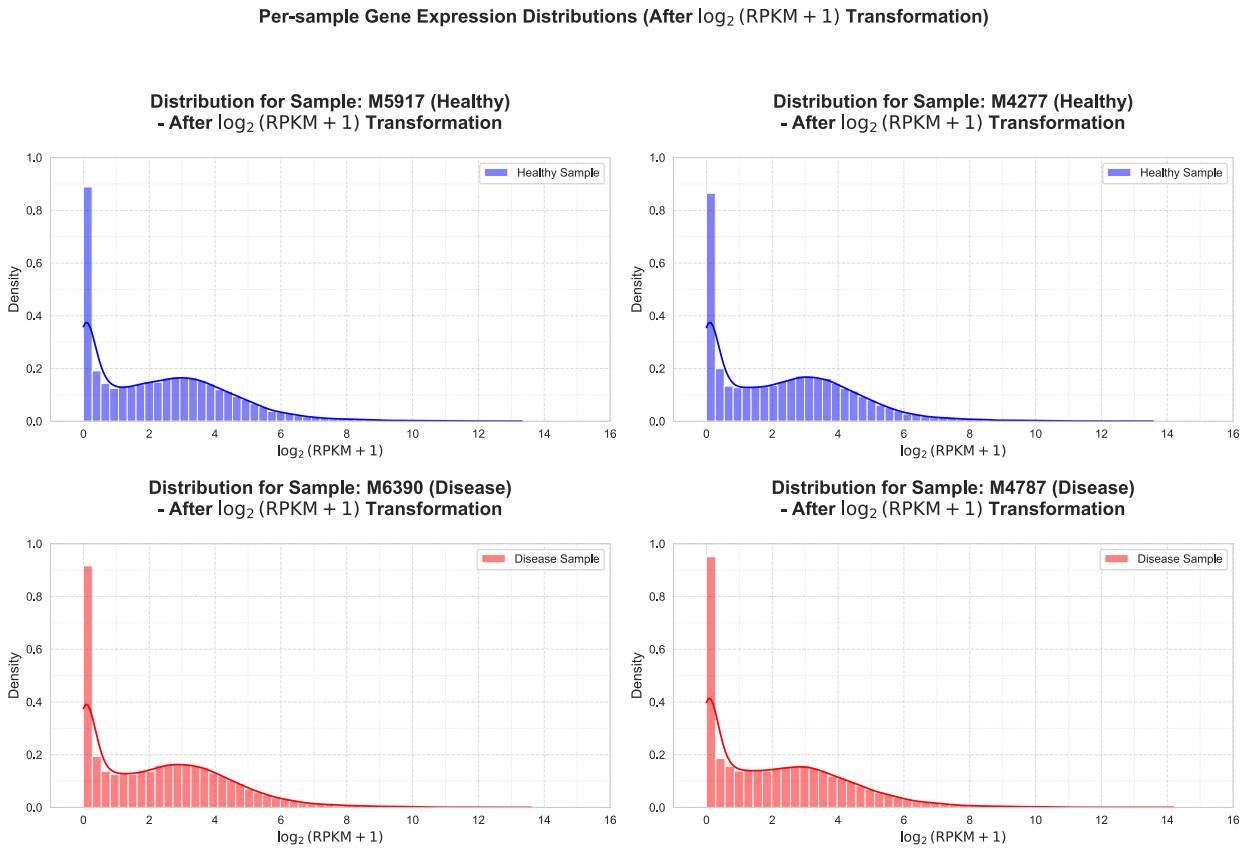


Figure C.2 Per-sample Gene Expression Distributions (After $log_2(RPKM + 1)$ Transformation)

Appendix D: Dissertation Research Diary

A chronological record of the key activities, decisions and milestones encountered during the dissertation project.

Meeting 1 (February) Date: 20/02/2025

Work Completed:

I drafted and submitted the *Dissertation/Project – Initial Idea* document prior to the meeting.

Items Discussed:

An initial meeting was held to establish contact between myself and my supervisor.

The initial project idea, which proposed investigating the link between mental health diagnoses and physical symptoms using a large medical dataset like UK Biobank, was discussed.

My supervisor provided critical feedback, suggesting the initial idea was likely beyond the scope of a Master's dissertation, primarily due to the significant challenges and potential infeasibility of accessing required patient data.

A strategic pivot to a new topic was recommended. My supervisor suggested exploring publicly available, high-quality health datasets from sources like Kaggle or the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). The GSE54456 dataset was specifically highlighted as a promising and interesting alternative for a bioinformatics-focused project.

Key Outcomes/Decisions:

The decision was made to abandon the original research idea in favour of a more feasible and well-scoped project.

I agreed with my supervisor's assessment and expressed strong interest in pursuing a project based on the GSE54456 dataset.

The new project direction was defined: a Machine Learning (ML) analysis of the GSE54456 psoriasis dataset, with the immediate next step being the development of a formal research proposal.

Action Points:

Write a comprehensive dissertation proposal based on the new topic (analysis of the GSE54456 dataset).

The proposal must be completed and submitted by the 17/03/2025 deadline.

Begin an initial literature review focused on psoriasis, transcriptomics and the application of ML to inform the proposal.

Meeting 2 (May) Date: 14/05/2025

Work Completed:

Between late February and mid-May, I conducted an extensive literature review on psoriasis, transcriptomics and the application of ML in this domain.

I drafted the full dissertation proposal based on the new topic.

Items Discussed:

Following the decision to use the GSE54456 dataset, the project's primary goal was established as a binary classification task to distinguish between 'Disease' and 'Healthy' samples.

The initial data preparation steps were discussed. It was anticipated that the gene expression data would be right-skewed and a $\log_2$ transformation was proposed as the likely solution to normalise the distributions.

A provisional plan for the main analysis was discussed, proposing the use of multiple ML models, including Random Forest Classifier (RFC) and Support Vector Machine (SVM). It was decided from the outset that the project would include a strong interpretability focus, using post hoc methods like SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) to identify key features for subsequent biological validation against existing literature.

The process for seeking ethics approval was discussed and it was confirmed that a self-declaration would be the appropriate route.

Key Outcomes/Decisions:

The initial data preparation and analysis plan was approved as a robust strategy for a classification task.

My supervisor advised that the ethics self-declaration should first be drafted in a shared Google Document for collaborative review before official submission.

Action Points:

Draft the ethics self-declaration in a Google Document and share it with my supervisor for review.

Conduct a deeper literature review to finalise the specific pre-processing steps and model hyperparameters.

Prepare the foundational Python script for data loading and pre-processing, ready to be run immediately following the key milestone of ethics approval (received 23/05/2025).

Meeting 3 (June) Date: 11/06/2025

Work completed:

Following ethics approval (23/05/2025), I acquired the GSE54456 dataset and performed the full data pre-processing pipeline between late May and early June.

This included all key steps: Exploratory Data Analysis (EDA), a $\log_2(x + 1)$ transformation to address skewed data, and a pilot Principal Component Analysis (PCA).

Items Discussed:

The results of the completed pre-processing pipeline were presented, including the pilot PCA which showed clear class separability.

The standard data preparation steps were discussed, including the use of a $\log_2$ transformation to address skewed data and a variance-based selection approach for initial feature selection to reduce dimensionality (~1000–5000 genes).

The suitability of several ML models was discussed. It was agreed that using three methodologically distinct models (LR, RFC and SVM) would provide a better basis for comparison than using only two, which might yield contradictory results.

The potential use of Neural Networks (NNs) was discussed but deemed overly complex and time-consuming for the project's scope.

It was confirmed that the ethics self-declaration was fully approved and no further action was required.

Key Outcomes/Decisions:

The data preparation and pre-processing plan was approved as robust.

The final selection of LR, RFC and SVM as the core classification models was confirmed.

The decision was made to incorporate a multi-faceted feature importance analysis, combining native model methods (e.g., LR coefficients) with a model-agnostic framework (SHAP).

Action Points:

Proceed with the full data pre-processing pipeline as discussed.

Develop, train and tune all three ML models.

Generate the five feature importance lists (LR coefficients, RF native feature importance and SHAP from all three models).

Prepare the initial results for the next supervision meeting, scheduled for early July 2025.

Meeting 4 (July) Date: 07/07/2025

Work completed:

Between late June and early July, I developed and trained all models, achieving perfect classification accuracy.

I also generated all five initial feature importance lists and conducted a preliminary overlap analysis.

Items Discussed:

The key result of perfect classification performance (1.0000 ROC AUC) for all models was reviewed.

It was confirmed that the perfect scores were not an error but a result of the strong, separable signal in the dataset, as had been anticipated from the initial PCA.

The pre-processing pipeline was discussed and validated, ensuring all steps correctly prevent data leakage.

Key Outcomes/Decisions:

The project's focus was officially pivoted from finding the best classifier to a comparative analysis of the features identified by the different models and interpretability methods.

The new objective is to identify a consensus gene signature and use the divergent model results to discuss the *multiplicity problem*.

Action Points:

Complete the quantitative feature overlap analysis to identify the final consensus gene set.

Conduct a targeted literature search for the identified consensus genes to gather evidence for their biological relevance (my supervisor specifically recommended using the DisGeNET website as a resource for exploring gene–disease associations).

Draft the 'Comparison to Existing Literature' section.

Prepare a full draft of the dissertation ready for a supervisor check at least two weeks before the final 27/08/2025 submission deadline.

Meeting 5 (August) Date: 13/08/2025

Work completed:

Between mid-July and early August, I completed the quantitative feature overlap analysis, which successfully identified the final five-gene consensus signature.

I conducted a final, targeted literature search for the key genes identified in the analysis. This search yielded excellent results, providing strong external validation for all key genes from high-quality, peer-reviewed papers.

A near-final draft of the entire dissertation, including the 'Comparison to Existing Literature' section, was submitted to my supervisor for review.

Items Discussed:

The near-final draft of the entire dissertation was discussed.

My supervisor gave initial feedback during the meeting, including the observation that the features in the consolidated feature analysis had not been normalised.

It was agreed that this would need to be corrected to ensure methodological consistency.

Key Outcomes/Decisions:

The methodological framework was confirmed as robust overall, with the exception of the missing normalisation step in the consolidated analysis.

The normalisation step will be incorporated to strengthen the rigour of the final results.

My supervisor indicated that a more detailed review of the draft would follow after the meeting.

Action Points:

Normalise the feature scores in the consolidated feature analysis and update the results accordingly.

Await review of the draft and then incorporate my supervisor's detailed written feedback (received later on 14/08/2025).

Complete all appendices, including this Research Diary and the Reflection on Research.

Perform a final proofread of the entire dissertation before the 27/08/2025 submission deadline.

Mid-August: Final amendments and proofreading of the full dissertation document were completed. Preparation of all appendices was finalised.

Late-August: Milestone: The final dissertation was submitted.

Appendix E: Reflection on Research

Feedback on Initial Idea: Pivoting the Research Scope

Reflection on feedback:

Situation: At the outset of the project, my initial idea was to investigate the link between undiagnosed mental health conditions and physical symptoms using a large, restricted-access medical dataset like UK Biobank.

Task: The task, following my first supervision meeting (20/02/2025), was to refine this broad area of interest into a feasible and well-scoped research project that could be successfully completed within the timeframe of a Master's dissertation.

Action (Responding to Feedback): My supervisor provided critical feedback that the initial idea, while interesting, was likely beyond the scope of a Master's project due to the immense practical and ethical challenges of accessing the required patient data. Acting directly on this feedback, I made the decision to pivot the project entirely. Following her suggestion, I researched and ultimately selected the publicly available NCBI GEO GSE54456 dataset as the basis for a new, more focused bioinformatics project.

Result: This resulted in a complete but necessary change in the research topic. The new project had a clearly defined scope, a readily accessible dataset and a feasible methodology. This improved the dissertation fundamentally, transforming it from a logistically challenging and potentially unachievable proposal into a well-defined project with a high probability of success.

Reflection: This was the most important learning moment in the project's management. It taught me the crucial difference between a broad area of interest and a well-defined, achievable research question. The key skill I developed was pragmatism in research planning. I learned to critically assess the feasibility of a project and make a strategic decision to align my research goals with available resources and a realistic timeline. This initial pivot, based directly on supervisor feedback, was the single most important decision that ensured the project's eventual success.

Associated skills:

Research and critical thinking, Applying knowledge, Personal development, Purpose

mySkills Screengrab:

Feedback on Methodological Rigour: Preventing Data Leakage

Reflection on feedback:

Situation: During the initial data analysis phase, my pre-processing pipeline involved performing variance-based selection on the entire dataset before splitting it into training and testing sets.

Task: In my fourth supervision meeting (07/07/2025), the task was to ensure my methodology was robust and adhered to best practices in Machine Learning (ML) to guarantee the validity of my results.

Action (Responding to Feedback): My supervisor highlighted that my initial approach could lead to data leakage, where information from the test set inadvertently influences the training process, potentially leading to an overly optimistic evaluation of model performance. In response to this feedback, I immediately amended my methodology. I restructured my Python code to ensure that the variance-based selection was fitted only on the training data after the split and the learned transformation was then applied separately to the test data.

Result: This significant modification ensured that the test set remained a truly unseen holdout set. This improved the dissertation by ensuring the scientific validity of my findings. The subsequent perfect classification scores could now be reported with high confidence, knowing they were generated from a methodologically sound and rigorous experimental design.

Reflection: This experience was a critical lesson in the subtleties of ML best practice. It taught me that data leakage is a non-obvious but serious risk that must be actively managed. The primary skill I strengthened was critical technical awareness. I learned that the integrity of the evaluation process is paramount and that every pre-processing step must be carefully considered to prevent information leakage from the test set to the training set. This feedback was crucial for ensuring my results were trustworthy and reproducible.

Associated skills:

Digital capability, Research and critical thinking, Ethics and sustainability, Academic skills

mySkills screengrab:

Feedback on Results Interpretation: From Classification to Feature Importance

Reflection on feedback:

Situation: My initial analysis revealed all three of my ML models achieved perfect classification accuracy. My original plan had been to compare them to find the best-performing one.

Task: In my fourth supervision meeting (07/07/2025), the task was to interpret this result and decide on the most meaningful way to frame the contribution of my research.

Action (Responding to Feedback): My supervisor confirmed the perfect performance was a result of the strong biological signal in the data. She advised that since no model was best in terms of accuracy, the most valuable research contribution would be a deep analysis of the features the models were using. Based on this guidance, I pivoted the focus of my entire 'Results and Discussion' section. The project's core contribution became a robust, comparative analysis of the features to identify a consensus gene signature.

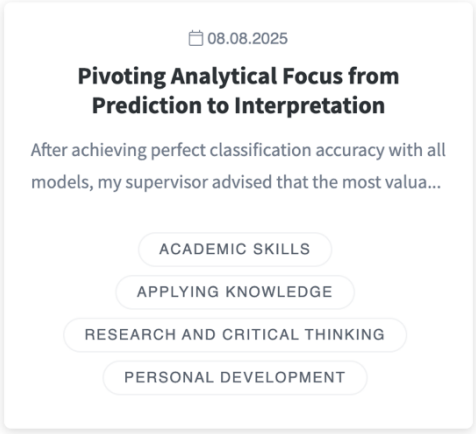
Result: This pivot transformed the dissertation from a simple report of high-accuracy into a more sophisticated investigation of model interpretability and the *multiplicity problem*. This improved the project's originality and impact, leading directly to the identification of the five-gene consensus signature, which became the central finding of the study.

Reflection: This feedback taught me a crucial lesson in scientific storytelling: the most obvious result is not always the most interesting one. The true insight came from interrogating why the models worked so well. The key skill I developed was analytical synthesis. I learned to look beyond surface-level metrics to find deeper, interpretable findings and then connect them back to the existing scientific literature for validation. This was a key step in my development from a data analyst into a data scientist.

Associated skills:

Applying knowledge, Research and critical thinking, Personal development, Academic skills

mySkills screengrab:



Appendix F: Code and Computational Environment

Files submitted (code and data):

- `pre-processing.ipynb` – loads GSE54456, performs inspection, $\log_2(\text{RPKM} + 1)$ transform, variance-based gene selection, scaling and PCA visualisation.
- `model_training_evaluation.ipynb` – trains LR, RFC and SVM; runs cross-validation and test evaluation; produces confusion matrices, ROC curves, coefficients / feature importances and SHAP analyses; builds the consolidated consensus ranking and performs pairwise / multiway feature overlap analyses.
- `Psoriasis_disease_GSE54456_RPKM.csv` – expression matrix used in this study (GSE54456).
- `requirements.txt` – full, pinned Python package list for reproducibility.

Key environment (versions):

- Python 3.13.5
- Core libraries (full list in `requirements.txt`):
    - `pandas` **2.3.0**
    - `scikit-learn` **1.7.0**
    - `matplotlib` **3.10.3**
    - `seaborn` **0.13.2**
    - `shap` **0.48.0**

Data availability:

The dataset originates from the NCBI Gene Expression Omnibus (GEO) repository (NCBI, n.d.) and corresponds to accession GSE54456 (Li et al., 2014b).

Notes:

- Figures and tables shown in the dissertation are generated by the notebooks.
- Environment replicability depends on these exact versions; use the provided `requirements.txt` when running.