Syrian Arab Republic
Ministry of Higher Education
and Scientific Research
Syrian Virtual University

الجمهورية العربية السورية
وزارة التعليم العالي والبحث العلمي
الجامعة الافتراضية السورية

جامعة الافتراضية السورية
SYRIAN VIRTUAL UNIVERSITY

Prediction of the therapeutic response in psoriasis patients using artificial intelligence tools

A thesis submitted as a fulfillment of requirements for a Master's degree in Bioinformatics

By:

Nadine Alkhoury

nadine_270685

Supervised by:

Dr. Louay Saleh

2025

Abstract

Background: Psoriasis is a chronic inflammatory disease involving both immune dysregulation and environmental factors, with a global prevalence of 2-3%. The introduction of TNF-alpha inhibitors previously used for other immune-mediated conditions like rheumatoid arthritis marked a transformative shift in psoriasis treatment. However, despite their efficacy, 30-40% of psoriasis patients fail to respond to anti-TNF-alpha therapy. This underscores the critical need for reliable predictive tools to assess individual treatment responses, enabling personalized therapeutic decisions.

Aim: This study aims to develop a machine learning model based on DNA methylation profiles to predict Anti-TNF-α response in psoriasis patients, distinguishing responders from non-responders.

Materials and Methods: Using Google Colab, five machine learning models Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and Multi-Layer Perceptron Classifier (MLPClassifier) were trained on DNA methylation data from 70 psoriasis patients. The cohort was stratified into:

- 49Anti-TNF-α responders (PASI improvement ≥90%)
- 21Anti-TNF-α non-responders (PASI improvement <70%)

The methylation dataset was sourced from the NCBI's GEO database (Accession:[GSE151278]).

Results: Among the evaluated models, Random Forest (RF) exhibited the highest predictive performance, with a (CV accuracy of 0.750 and test-accuracy: 0.785, precision: 0.835, recall: 0.785, F1: 0.735).Notably, the three most influential variables in our model mapped to genomic loci where differential methylation patterns could potentially regulate the expression of genes encoding proteins directly implicated in psoriasis pathogenesis .

Conclusions: Our machine learning analysis of DNA methylation data identified Random Forest as the optimal predictor of anti-TNF-α response in psoriasis patients (79% accuracy). The top predictive loci were biologically relevant to psoriatic pathways, suggesting clinical potential for treatment stratification. Further validation in larger cohorts could enhance predictive utility.

Keywords: Psoriasis- DNA Methylation- Machine Learning

ملخص

خلفية البحث: الصدفية هي مرض مناعي ذاتي مزمن ينتج عن تداخل العديد من العوامل البيئية والمناعية والجينية. تتراوح نسبة الإصابة به عالميا من (2-3%). شكل استخدام مثبطات عامل نخر الورم ألفا (anti-TNF-α) المستخدمة سابقا لعلاج أمراض مناعية أخرى مثل التهاب المفاصل الروماتيدي نقلة نوعية في علاج هذا المرض، ولكن على الرغم من أهمية هذه الزمرة العلاجية في علاج المرضى إلا أن (20-30%) من المرضى لا يستجيبون للعلاج، مما أبرز الحاجة الملحة إلى وجود أدوات تنبؤية تساهم في التنبؤ بمدى الاستجابة العلاجية بحيث تسهم في دعم العلاج الفردي واتخاذ قرارات علاجية صحيحة.

هدف البحث: تهدف هذه الدراسة إلى تطوير نموذج تعلم آلي يعتمد على تباين مثيلة الحمض النووي DNA للتنبؤ بالاستجابة لـ(anti-TNF-α) لدى مرضى الصدفية.

أدوات وطرائق البحث: باستخدام بيئة عمل (Google Colab)، تم تدريب خمسة نماذج تعلم آلي هي Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Support) (Vector Machine (SVM), and Multi-Layer Perceptron Classifier (MLPClassifier) على بيانات مثيلة الحمض النووي DNA لـ 70 مريض صدفية ينتمون إلى مجموعتين: الأولى مؤلفة من 49 مريض استجابوا للعلاج بشكل جيد، المجموعة الثانية مؤلفة من 21 مريض غير مستجيب للعلاج. تم استيراد هذه البيانات من قاعدة بيانات GEO التابعة ل NCBI معرفة بالرقم [GSE151278].

النتائج: أظهر نموذج الغابات العشوائية (Random Forest) أعلى أداء تنبؤي بنتائج (CV accuracy : 0.750 test- accuracy: 0.785, precision: 0.835, recall: 0.785, F1: 0.735). كما قمنا في البحث باستخلاص أهم ثلاثة مواقع مثيلة ساهمت في التنبؤ في هذا النموذج وربطها بمواقعها الجينية فتبين أنها تنتمي إلى مواقع ذات صلة مباشرة بمرض الصدفية مما يؤكد أن التنبؤ ارتكز على أساس بيولوجي ولم يكن نتيجة صدفة إحصائية.

الخلاصة: حددت هذه الدراسة نموذج الغابات العشوائية (Random Forest) كأفضل نموذج تعلم آلي للتنبؤ باستجابة مرضى الصدفية ل (anti-TNF-α) بالاعتماد على بيانات مثيلة الحمض النووي DNA بدقة 79%. كما تم تحديد أهم مواقع المثيلة المؤثرة في عملية التنبؤ وربطها بموقعها الجيني ومناقشة تأثيرها البيولوجي في المسارات الإمراضية لمرض الصدفية مما يقترحها هدفا للدراسات الجزيئية المستقبلية . يمكن للدراسات المستقبلية التي تشمل عدد أكبر من المرضى أن تعزز فائدة ودقة هذا النموذج وتجعله قابلا للتطبيق السريري.

الكلمات المفتاحية: مرض الصدفية- مثيلة DNA- التعلم الآلي

TABLE OF CONTENTS

Table of figures

Table of tables

Table of Abbreviations

IL-17	Interleukin-17
IL-23	Interleukin-23
TNF	Tumor necrosis factor-alpha
NF-κB	Nuclear Factor Kappa B
JAK-STAT	Janus Kinase - Signal Transducer and Activator of Transcription
IRF	Interferon Regulatory Factor
Ig G	Immunoglobulin G
TH1	T Helper 1 Cells
ML	Machine Learning
5-MC	5-MethylCytosine
N6-mA	N6-methyladenine
7-mG	7-methylguanine
WGBS	Whole Genome Bisulfite Sequencing
RRBS	Reduced Representation Bisulfite
PNR	Partial Non-Response
CRP	C-Reactive Protein
PASI	Psoriasis Area and Severity Index
PsA	Psoriatic Arthritis
CNVs	Copy Number Variations
ER	Excellent Response
PR	Partial Response
IDAT files	Illumina BeadArray Data Files
ChAMP	Chip Analysis Methylation Pipeline
MyCAN	Myeloid Cancer Pane
CpG	Cytosine-phosphate-Guanine

CBS	Circular Methylated Sites
DMSs	Different Methylated Sites
HRs	Hazards Ratios
AUC	Area Under the Curve
BSA	Body Surface Area
PsO	Psoriasis
PSoHO	Psoriasis and Holistic Health Outcomes
GEO	Gene Expression Omnibus
NCBI	National Center for Biotechnology Information
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
LR	Logistics Regression
MLP	Multi-Layer Perceptron
CV	Cross Validation
PCA	Principal Component Analysis
Chr	Chromosome
PKC	Protein Kinase C
MAPK	Mitogen Activated Protein Kinase

CHAPTER 1

PREFACE

1.1. Introduction

Psoriasis is a chronic immune related skin disorder .This disorder is observed in roughly (2-3)% of individuals globally. Studies indicate that 30% of psoriasis patients suffer from psoriatic arthritis, and nail lesions occur in 50% of cases (1).

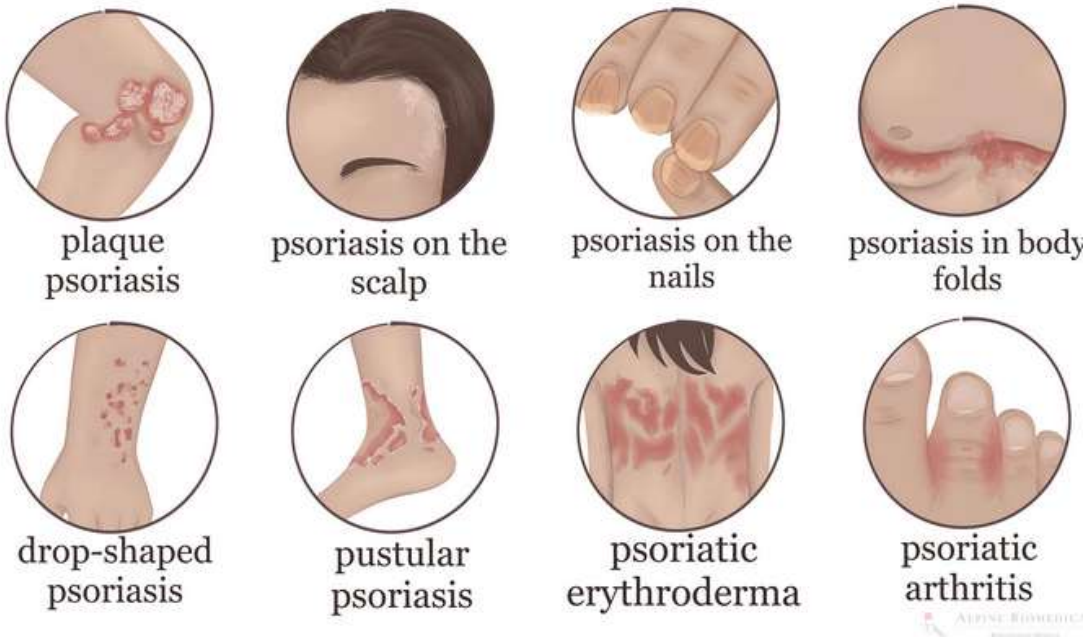


Figure 1.1. Types of Psoriasis

Psoriasis results from a combination of immunological genetic and environmental triggers: (1)

Genetic and environmental triggers: current genomic research has uncovered more than 63 genetic loci that show significant association with psoriasis pathogenesis. Key environmental triggers of psoriasis encompass metabolic factors (obesity, diet), lifestyle habits (smoking, alcohol), physical trauma, medication reactions, and infectious agents

Immunological triggers: Psoriasis manifests through cellular and molecular mechanisms: epidermal keratinocytes exhibit dysregulated proliferation and

aberrant differentiation, while immune cells (particularly T-cells and dendritic cells) infiltrate the dermis. At the molecular level, these cells generate excessive pro-inflammatory cytokines (e.g., IL-17, IL-23, TNF-α), creating a self-sustaining inflammatory microenvironment within psoriatic lesions. While epidermal keratinocytes normally function as the primary physical and immunological barrier, psoriatic keratinocytes exhibit profound dysregulation. Their accelerated proliferation stems resulting in immature cells with deficient lipid and keratohyalin production. Crucially, these dysfunctional keratinocytes engage in pathological crosstalk with innate and adaptive immune cells - particularly dendritic cells, monocyte-derived macrophages, and tissue-resident memory T cells - establishing a self-perpetuating inflammatory circuit. Psoriasis pathogenesis involves complex immunomodulatory networks converging on key signaling cascades. Critical pathways including NF-κB, JAK-STAT, and interferon regulatory factor (IRF) systems become activated, driving inflammatory gene transcription in keratinocytes and immune cells. (1)

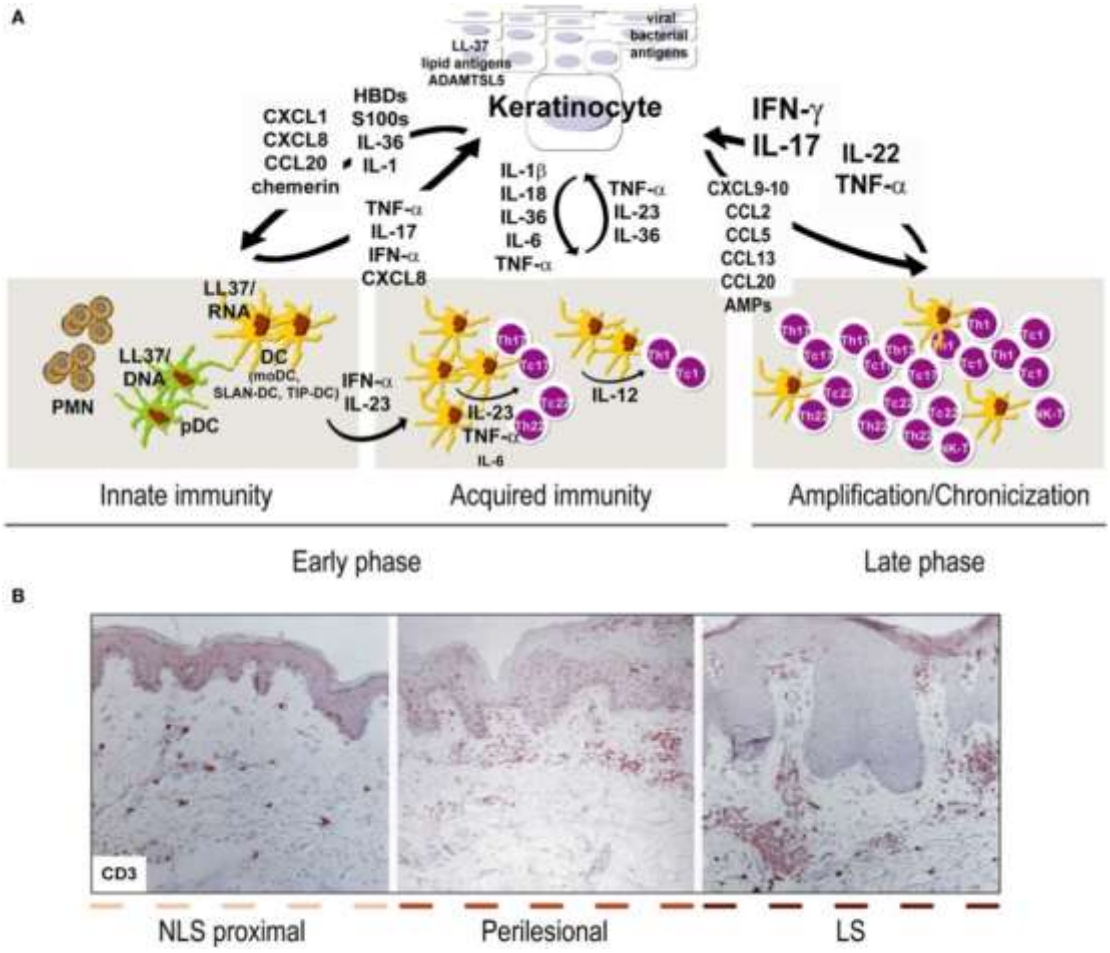


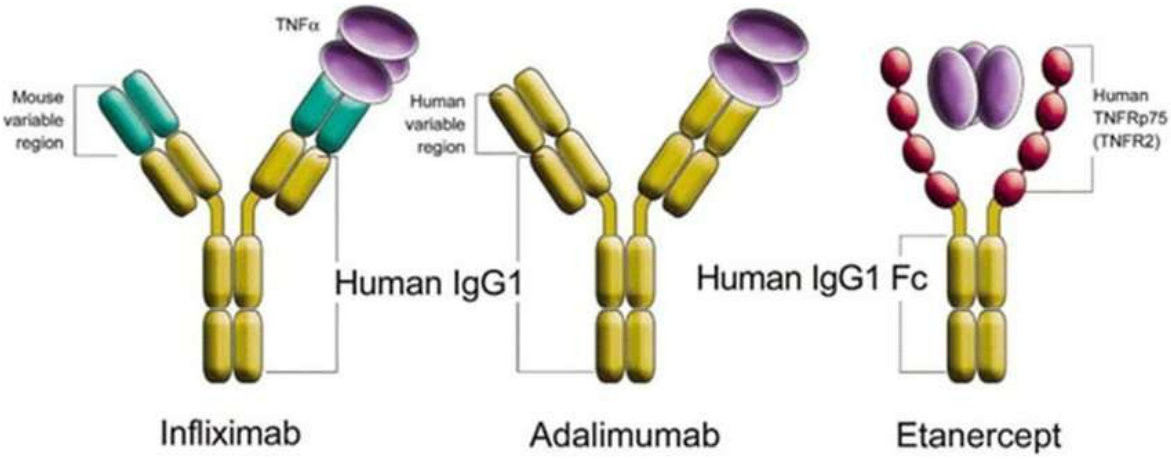
Figure 1.2 Aberrant interplay of keratinocytes and immune cells in psoriasishttps://en.wikipedia.org/wiki/en:Creative_Commons (2)

Advances in the understanding of psoriasis pathogenesis have led to the discovery and development of an expanding array of targeted molecules, which demonstrate improved clinical outcomes and better quality of life for patients.

A major breakthrough in psoriasis treatment was first achieved with the use of tumor necrosis factor (TNF) inhibitors, which had already been approved for other inflammatory conditions, such as rheumatic diseases. (3)

Etanercept	Etanercept is a bioengineered fusion protein that combines two key components: Two soluble TNF receptor domains These bind to both free-floating (soluble) and cell-surface (membrane-bound) TNF-α, blocking its activity. An IgG1 Fc fragment This stabilizes the molecule and extends its lifespan in the bloodstream. As a dimer, Etanercept can neutralize two TNF-α molecules simultaneously, effectively competing with natural TNF receptors and reducing inflammation. TNF-α itself is produced by immune cells such as dendritic cells, Th1/Th17/Th22 lymphocytes, macrophages, and even skin cells (keratinocytes), playing a major role in psoriasis pathogenesis (3)
Infliximab	Infliximab is a chimeric monoclonal antibody composed of: Murine-derived variable regions (for precise TNF-α binding) Human-derived IgG1 constant regions (to reduce immune rejection) It works by binding and neutralizing both soluble and membrane-bound TNF-α, blocking its inflammatory effects. (3)
Adalimumab	Adalimumab is a fully human IgG1 monoclonal antibody that specifically targets and neutralizes both soluble and membrane-bound TNF-α, similar to infliximab. Fully human structure (reduces immunogenicity compared to chimeric antibodies like infliximab). (3)

Table1.1. TNF-α inhibitors



Anti-TNF-α agents (4)

Although Anti-TNF-α agents (e.g., adalimumab, infliximab, etanercept) are effective in treating psoriasis and psoriatic arthritis. However, 30-40% of patients exhibit either primary non-response(PNR) or secondary non-response (SNR), limiting treatment efficacy.

Personalized medicine represents a transformative approach in psoriasis management, enabling tailored therapeutic strategies that optimize treatment efficacy while minimizing adverse effects and reducing healthcare costs. (5)

Despite considerable progress in elucidating the pathogenesis of psoriasis, the implementation of individualized genetic profiling remains constrained by prohibitive costs, time limitations, and the scarcity of highly specific biomarkers capable of predicting treatment response. Consequently, a standardized biomarker panel would likely prove more reliable than single-target gene analysis. Furthermore, the development of a comprehensive algorithm integrating both genotypic and phenotypic patient data could significantly enhance diagnostic and therapeutic (5).

The global proliferation of data derived from medical devices and electronic health records has facilitated the advancement of machine learning (ML) technologies. These innovations are poised to play a pivotal role in developing personalized psoriasis bio panels. Such ML-driven approaches aim to integrate multidimensional patient data, including genetic profiles, phenotypic characteristics, comorbid conditions (which may contraindicate certain therapies), and histories of treatment failure (which may help delineate distinct psoriasis endotypes). This integrated framework promises to enhance clinical decision-making by enabling physicians to select optimal, patient-specific therapeutic strategies (5) .

To date, despite numerous studies investigating biomarkers predictive of response to biologic therapies, no consensus has been established regarding a standardized

panel suitable for routine clinical implementation. The integration of artificial intelligence (AI) to develop algorithms that synthesize individual genotypic and phenotypic data represents a transformative approach to holistic patient management, enabling truly personalized therapeutic strategies. However, further research is required to validate and optimize this paradigm (5).

Historically, the majority of research on inter-individual variability in drug response has centered on genetic polymorphisms that alter transcription factor binding sites. However, emerging evidence highlights the role of heritable, epigenetic modifications such as DNA methylation, his tone modifications, and non-coding RNA regulation in modulating gene expression and pharmacodynamic outcomes independently of DNA sequence variation. These mechanisms contribute significantly to phenotypic diversity in drug metabolism, efficacy, and toxicity, underscoring the need for integrative genomic and epigenomic approaches in precision medicine. (6)

DNA methylation stands as one of the most extensively studied epigenetic modifications governing gene expression regulation. Notably, this heritable molecular marker frequently occurs in genomic regions encoding pharmacologically relevant proteins, including:

(1) Drug-metabolizing enzymes (e.g., cytochrome P450 superfamily)

(2) Membrane transport proteins (e.g., ABC transporters)

(3) Molecular drug targets (e.g., receptor proteins)

DNA methylation, primarily occurring as 5-methylcytosine (5-mC), serves as a key epigenetic regulator of gene silencing, which can be reversed through active or passive demethylation processes. While 5-mC dominates eukaryotic DNA methylation, minor modifications such as N6-methyladenine (N6-mA) and 7-methylguanine (7-mG) also contribute to epigenetic regulation, though their roles remain less understood. Methylation-induced transcriptional suppression can be dynamically modulated, influencing critical biological processes, including drug metabolism, cellular differentiation, and disease pathogenesis. (6)

The Illumina DNA methylation microarray platforms, such as the HumanMethylation450 BeadChip (450K array) and the Infinium Methylation EPICBeadChip (850K array), along with whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS), represent the most widely utilized high-throughput technologies for genome-wide DNA methylation profiling at single-nucleotide resolution. (7)

1.2. Problem Statement:

- **High Non-Response Rates to Anti-TNF-α Therapy**: Although Anti-TNF-α agents (e.g., adalimumab, infliximab, etanercept) are effective in treating psoriasis and psoriatic arthritis, 30–40% of patients exhibit either partial non-response (PNR) or total non-response. This limits treatment efficacy, leading to prolonged disease activity, reduced quality of life, and increased healthcare burdens.
- **Limitations of Current Predictive Approaches:** Existing studies primarily rely on clinical or serum biomarkers (e.g., CRP, PASI scores) and statistical models to predict treatment response. However, these methods lack sufficient accuracy for individualized predictions, as they fail to fully capture the complex genetic and epigenetic mechanisms underlying non-response.
- **Underexplored Role of DNA Methylation in Treatment Response:** Recent evidence suggests that *DNA* methylation patterns may influence Anti-TNF-α responsiveness, but most studies have analyzed these genetic factors using traditional statistical methods (e.g., regression models). This approach overlooks the potential of machine learning (ML) to detect non-linear interactions and improve predictive performance.

1.3. Objectives:

The primary goal of this study is to construct a predictive machine learning model utilizing DNA methylation signatures to stratify Anti-TNF-α responders and non-responders among psoriasis patients. More precisely:

- **Identify Differential Methylation Patterns:** Investigate and compare DNA methylation profiles in psoriasis patients who respond to Anti-TNF-α therapy versus non-responders (PNR/SNR) to pinpoint epigenetically significant loci.
- **Develop Machine Learning Predictive Model:** Design and train an interpretable ML model using methylation data to classify patients into responders and non-responders with high accuracy.
- **Translate Findings into Potential Biomarkers:** Extract and prioritize top predictive methylation markers to propose a minimal epigenetic signature for future clinical use in personalized treatment selection.

CHAPTER 2

Theoretical Background

2.1. Introduction

Psoriasis is a persistent inflammatory dermatological condition, has a worldwide prevalence of 2–3%. This disease is often linked to several comorbid conditions, such as psoriatic arthritis (PsA), cardiovascular disorders, and depressive illness.

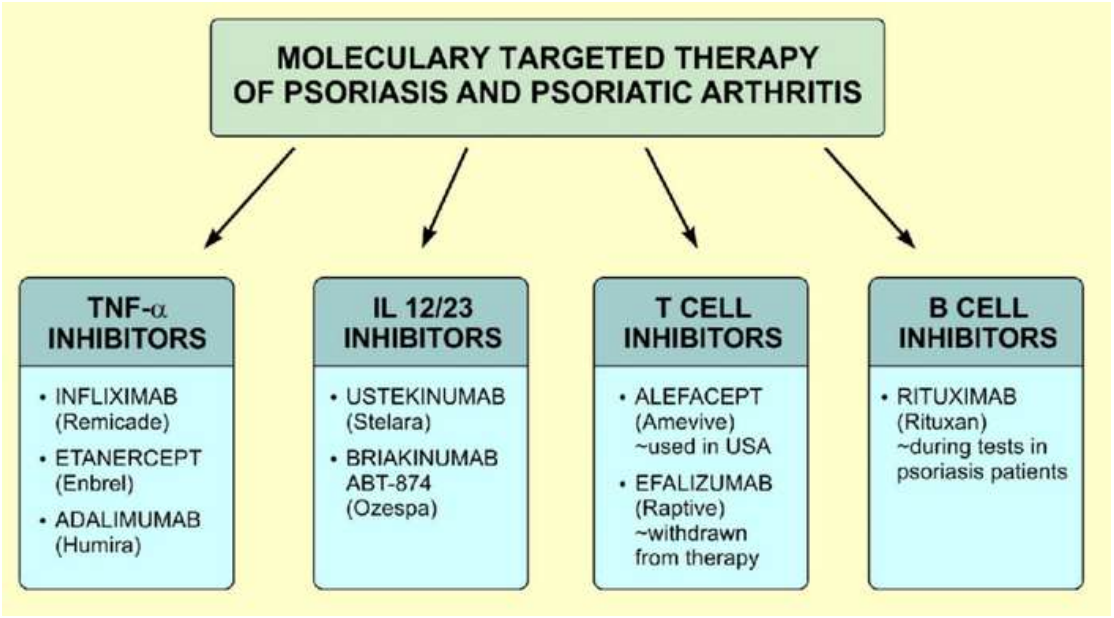


Figure 2.1. molecularly targeted therapy of psoriasis

In mild-to-moderate psoriasis, symptoms are frequently controlled with topical therapies and/or phototherapy. However, patients with severe disease typically require systemic treatments, including biologic agents.

The advent of biologic therapies has revolutionized the treatment paradigm for psoriasis. Before their introduction, achieving disease remission often entailed prolonged trials of topical and systemic agents, accompanied by considerable risks of drug-related toxicity. In contrast, contemporary biologic agents including tumor necrosis factor (TNF) inhibitors, interleukin (IL)-17A inhibitors, and IL-23/IL-12/23 inhibitors demonstrate markedly superior efficacy, with up to 80% of patients attaining PASI 90 responses and up to 90% achieving PASI 75 responses.

Optimal treatment selection for psoriasis patients often involves trial and error, with some requiring several drug switches before achieving long-term efficacy. However, each unsuccessful attempt raises the risk of discontinuation due to inefficacy.

2.2. Review of literature

1) Ancor SG, Reolid A, Fisas LH, Munoz-Aceituno E, Llamas-Velasco M, Sahuquillo-Torralba A, Botella-Estrada R, Garcia-Martinez J, Navarro R, Dauden E, Francisco AS. **DNA copy number variation associated with anti-tumour necrosis factor drug response and paradoxical psoriasiform reactions in patients with moderate-to-severe psoriasis.** Actadermato-venereologica. 2021 May 4;101(5):689.

Although biologic agents targeting tumor necrosis factor (TNF) demonstrate efficacy in psoriasis treatment, 30–50% of patients exhibit either non-response or paradoxical psoriasiform reactions. This study investigates potential DNA copy number variations (CNVs) as predictive biomarkers for anti-TNF therapeutic response or the development of TNF inhibitor-induced psoriasiform eruptions. CNVs are structural genomic variants characterized by reduced (deletion) or elevated (duplication/insertion) copies of specific DNA sequences, which may alter gene dosage and regulatory landscapes.

Blood samples were collected from 70 patients with moderate-to-severe psoriasis who were treated with anti-TNF agents (adalimumab, infliximab, or etanercept). Treatment response was clinically evaluated, and patients were stratified into two groups based on therapeutic outcomes:

- Excellent responders (ER, n=49): Patients demonstrating optimal clinical improvement.
- Partial responders (PR, n=21): Patients exhibiting suboptimal or limited therapeutic response.

DNA was extracted from blood samples, followed by genome-wide DNA methylation profiling using the high-density Infinium HumanMethylation450 BeadChip array. Methylation data were recorded as IDAT files for downstream analysis.

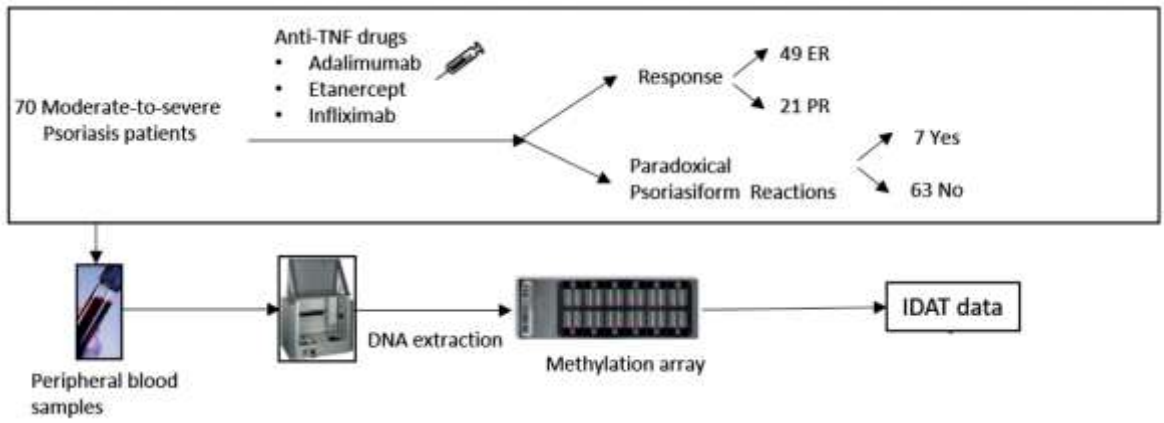


Figure 2.2. Workflow performed in this study.

Raw IDAT files were processed using two specialized R packages: conumee and the Chip Analysis Methylation Pipeline (ChAMP). These tools were employed to detect copy number variations (CNVs) based on methylation array output. Specifically, CNVs were derived from ChAMP using the 'myCAN' function. The ChAMP pipeline integrates methylated and unmethylated probe intensity values for each cytosine-phosphate-guanine (CpG) site, followed by intensity normalization using a series of controls obtained from the minfi package (healthy reference genomes). (minfiData: Example data for the Illumina Methylation 450k array. R package version 0.36.0)

Following initial data processing, the conumee package performs two distinct DNA partitioning operations: bins and segments. Bins represent contiguous 15-CpG genomic regions, with a fixed count of 15,820 bins per patient. Segments are larger homogeneous regions of consistent copy number variation, identified through the Circular Binary Segmentation (CBS) algorithm. These segments range from 100,000 to 6,000,000 base pairs in size, with the number varying across patients.

Following CNV identification, significant bins and segments have been mapped to genomic coordinates using the R bedr package to identify overlapping genes. These genes have been subsequently analyzed for pathway enrichment using EnrichR to determine relevant signaling pathways.

For comparative analysis:

1. A custom scripts to cross-validate CNVs have been developed and called by different packages and identify representative CNV regions.
2. Methylation intensity values per bin have been compared between patient groups (ER vs PR) using Student's t-tests.
3. Segment comparisons have been restricted to:

- Identical genomic coordinates (same start/end positions)
- Minimum recurrence (≥2 patients per group)
- Length-matched regions

To address multiple testing:

- Bonferroni correction has been applied
- This conservative approach has minimized false discovery while maintaining detection power.

This study demonstrates that:

1. **Therapeutic Response**: Clinical response to adalimumab correlates significantly with specific CNV patterns ($p<0.05$).
2. **Adverse Effects**: Development of cutaneous complications shows a strong association with distinct CNV profiles.
3. **Predictive Biomarkers**: Statistically significant CNVs ($p<0.05$) were identified as potential biomarkers for:
    - Predicting adalimumab treatment efficacy
    - Anticipating adverse drug reactions

2) Ovejero-Benito MC, Cabaleiro T, Sanz-García A, Llamas-Velasco M et al. **Epigenetic biomarkers associated with antitumour necrosis factor drug response in moderate-to-severe psoriasis.**2018 Mar;178(3):798-800. PMID: 28369750

Recent studies have revealed that epigenetic changes, particularly DNA methylation, play a role in the development of psoriasis. DNA methylation is a heritable and dynamic covalent modification that occurs at cytosine-phosphate-guanine (CpG) sites and can influence gene expression. While anti-tumor necrosis factor-alpha (anti-TNF-α) therapies such as(adalimumab, etanercept, and infliximab) are effective treatments for moderate-to-severe psoriasis, approximately 30–50% of patients show an insufficient response. This study is the first to investigate potential epigenetic biomarkers that may predict patient response to anti-TNF therapy.

This study included 70 White patients with moderate-to-severe plaque psoriasis who were treated with anti-TNF therapy. Patients were selected and divided into two groups:

- excellent responders (ER) : whose achieved >=90% improvement
- partial responders (PR) : whose achieved <70% improvement

DNA methylation profiling was performed using the Illumina Infinium HumanMethylation450 BeadChip array. The ChAMP pipeline was employed for methylation data analysis. All analyses were performed in R. Differential methylation was assessed using a moderated t-test, adjusted for batch effects. The test statistic was computed as the ratio of the methylation β-value (or M-value) to its standard error.

For categorical variables, the study applied the moderated t-test, while linear regression models were used to evaluate associations between methylation levels (M-values) and continuous variables, such as PASI scores at 3- and 6-month follow-ups.

Results:

No differentially methylated sites (DMSs) were identified between patients exhibiting an excellent response and those with a partial response to anti-TNF therapy. Similarly, no significant DMSs were observed when comparing excellent and partial responders to either infliximab or etanercept. However, three CpG sites were found to be hypermethylated in partial responders (n = 4) compared to excellent responders (n = 21) to adalimumab treatment.

Linear regression analysis revealed no significant association between baseline PASI or PASI at 3 months and the methylation levels (m-values) of any analyzed CpG sites. However, a positive correlation was observed between PASI at 6 months and the m-values of **cg09141835**, suggesting that hypermethylation at this site may be associated with a poorer response to anti-TNF therapy.

Variable	Number	Comparison	Total	Hyper	Hypo
(A) EWAS association analysis performed and the results obtained					
Anti-TNF drug global response	70	ER (N = 49) vs PR (N = 21)	–	–	–
Adalimumab response	25	ER (N = 21) vs PR (N = 4)	3	3	–
Etanercept response	27	ER (N = 16) vs PR (N = 11)	–	–	–
Infliximab response	18	ER (N = 12) vs PR (N = 6)	–	–	–
Baseline PASI	70	LC	–	–	–
PASI at 3 months	70	LC	–	–	–
PASI at 6 months	69	LC	3	1	2

(B) Significant correlation between Differentially Methylated Sites in ER (n=21) and PR (n=4) patients to adalimumab

CpG site	CHR	Gene Name	adj. P-value
cg23132469	1	TAS1R2	0·014
cg18837178	5	NA	0·003
cg05221720	6	COL9A1	0·049

(C) Significant correlation between PASI at 6 months and DNA methylation values

CpG site	CHR	Gene Name	adj. P-value
cg09141835	5	CBFA2T3	0·001
cg23446055	16	PRELID2	0·002
cg03242666	17	PMP22	0·026

Figure 2.3. Results from Ovejero et al 2018

3) Amy X. Du, Zarqa Ali , Kawa K. Ajgeiy, Maiken G. Dalager, Tomas N. Dam, Alexander Egebjerg, Christoffer V. S. Nissen, Lone Skov, Simon Francis Thomsen, Sepideh Emam, Robert Gniadecki1. **Machine Learning Model for Predicting Outcomes of Biologic Therapy in Psoriasis.**Journal of the American Academy ofDermatology doi: 10.1016/j.jaad.2022.12.046

Objective: To evaluate and compare the predictive accuracy of a conventional risk factor-based frequentist statistical model versus machine learning algorithms in estimating the 5-year probability of biologic therapy discontinuation.

Methodology: Data were extracted from the Danish national psoriasis registry (DermBio), which included 6,172 treatment courses involving anti-TNF agents (etanercept, infliximab, adalimumab), ustekinumab, guselkumab, and anti-IL-17 therapies (secukinumab, ixekizumab) across 3,388 unique patients. Cox proportional hazards regression was employed to calculate hazard ratios (HRs) for all available predictive factors. For machine learning (ML) approaches, multiple models were trained to predict 5-year drug discontinuation risk using 10 routinely collected clinical features. Model training incorporated a 5-fold cross-validation framework. Predictive performance was evaluated using the area under the receiver curve (AUC).

Results: Ustekinumab and ixekizumab demonstrated the lowest 5-year discontinuation rates among the evaluated biologics. Additional predictors of treatment persistence included male sex and biologic-naïve status. The conventional risk factor-based predictive model achieved modest discrimination (AUC = 0.61).

In contrast, the optimal machine learning approach (gradient boosted trees) showed superior predictive performance (AUC = 0.85).

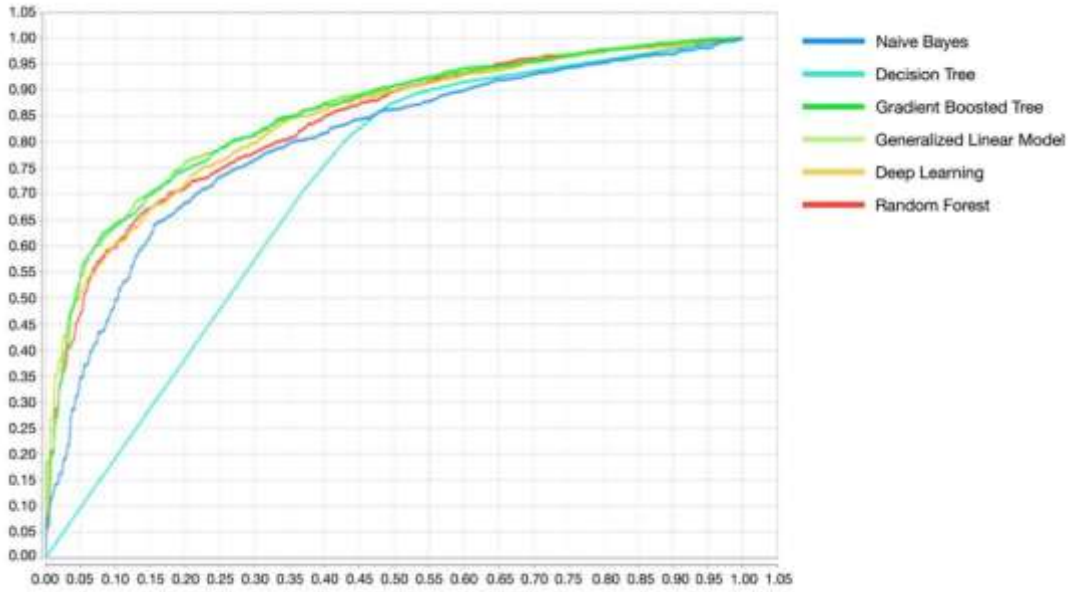


Figure 2.4. AUC Curve

Conclusion:

In this study, two distinct approaches for predicting biologic therapy discontinuation are developed and validated: a conventional risk factor-based statistical model and a machine learning (ML)-based predictive tool. Our findings demonstrate the superior performance of the ML algorithm, which shows promising potential as a clinical decision-support tool for personalized treatment selection in psoriasis. This advanced predictive model may enhance therapeutic decision-making by enabling dermatologists to optimize biologic selection and improve patient counseling through individualized risk assessment.

4) (8), **Multivariable Predictive Models to Identify the Optimal Biologic Therapy for Treatment of Patients with Psoriasis at the Individual Level**. JAMA Dermatol, August 17, 2022;158;(10):1149-1156. doi:10.1001/jamadermatol.2022.3171

Objective: To determine the most effective biologic therapy for psoriasis patients using predictive statistical modeling and machine learning approaches.

Methodology: This nationwide cohort study utilized data from Danish national registries, with DERMBIO serving as the primary data source. The study population comprised adult patients receiving biologic therapy for moderate-to-severe

psoriasis. Data processing and analyses were conducted from spring 2021 through spring 2022.

This study employed unsupervised learning to identify clinically meaningful patient clusters using routinely collected registry data. Statistical methods and supervised machine learning algorithms were subsequently applied to:

1. predict treatment discontinuation (binary outcome) within 1-3 years,
2. classify patients according to their optimal biologic therapy (multiclass outcome) based on treatment persistence.

Results: Using a success criterion of 3 years of sustained treatment, this study analyzed 2034 patients with a total of 3452 treatment series. The majority of treatment series involved male patients (2147, 62.2%), with most originating from Denmark (3190, 92.4%). Additionally, 2414 (69.9%) of the patients had completed education beyond primary school. The average age at psoriasis diagnosis was 24.9 years, while the average age at the start of biologic therapy was 45.5 years.

In predicting the most effective cytokine target (e.g., interleukin-17 inhibition), gradient-boosted decision trees achieved an accuracy of 63.6%, while logistic regression reached 59.2%. The top 2accuracy improved to 95.9% and 93.9%, respectively.

For predicting specific successful drugs, gradient boosting showed an accuracy of 48.5%, compared to 44.4% for logistic regression. The top 2 accuracy was 77.6% (gradient boost) and 75.9% (logistic regression), while the top 3 accuracy reached 89.9% and 89.0%, respectively.

5) April W. ARMSTRONG, Elisabeth RIEDL, Patrick M. BRUNNER, Stefano PIASERICO, Willie I. VISSER, Natalie HAUSTRUP, Bruce W. KONICEK, Zbigniew KADZIOLA, Mercedes NUNEZ, Alan BRNABIC and Christopher SCHUSTER. (9)**.** (9)**.** ActaDermato-Venereologica, 2024.DOI: 10.2340/actadv.v104.40556

Objective: Despite the availability of extensive clinical data, the selection of biologic therapies for patients with moderate-to-severe psoriasis (PsO) remains largely based on a trial-and-error approach. While modern biologics achieve high rates of skin clearance (PASI90/100) in many patients, suboptimal initial therapy can delay effective disease control and impact long-term outcomes. This study evaluated predictors of complete skin clearance (PASI100) at the following time points:

1. Week 12 (short-term response)
2. Month 12 (long-term response)
3. Week 12 with durability through Months 6 and 12 (sustained response)

A secondary objective was to analyze predictor variables (e.g., demographic, clinical, or molecular biomarkers) to:

- Quantify their association with PASI100 likelihood
- Compare differences in efficacy across biologic classes (e.g., anti-TNF, IL-17/23 inhibitors)

Methodology and Results: Using machine learning and advanced statistical methods, this study analyzed a sub-population of 1,917 patients with moderate-to-severe psoriasis (PsO) from the PSoHO dataset who were treated with biologics. Researchers identified 14 novel predictor variables, which were combined with 12 additional variables previously linked to treatment response in the literature, resulting in a total of 26 potential predictors.

A subsequent logistic regression analysis revealed three significant predictors associated with achieving at least one of the three PASI100 outcomes (complete skin clearance at Week 12, Month 12, or sustained response):

1. Nail Psoriasis: The absence of nail involvement emerged as a strong predictor for two different PASI100 outcomes, underscoring its clinical relevance.
2. Hypertension
3. Body Surface Area (BSA) involvement

This study underscores the persistent difficulty in defining reliable clinical predictors of treatment response in moderate-to-severe psoriasis. Despite this challenge, the absence of nail involvement emerged as the most robust and clinically actionable marker from real-world evidence to predict optimal biologic therapy outcomes.

Critical Takeaways for Practice:

1. Nail PsO Assessment as a Decision Tool

    - Routine screening for psoriatic nail disease should be prioritized during clinical evaluations
    - Its absence serves as a practical indicator for higher likelihood of PASI100-level responses

2. Therapeutic Strategy Implications

- Findings advocate for tailored biologic selection based on nail involvement status
- Reinforces need for comprehensive baseline assessments beyond skin severity alone

This evidence transforms a simple bedside observation (nail examination) into a stratification tool for precision medicine in psoriasis management.

2.3. Research gap

Despite advances in ML applications for precision medicine, no study has yet developed an ML-based predictive framework using methylation profiles to stratify psoriasis patients by their likelihood of responding to Anti-TNF-α therapy. Bridging this gap could enable earlier identification of non-responders, reduce trial-and-error prescribing, and optimize therapeutic outcomes.

While numerous studies have focused on predicting factors influencing treatment response, most rely on either statistical approach. This study aims to address a critical gap by leveraging machine learning models topredict treatment response based on genetic methylation data. By doing so, we propose a more robust and data-driven approach to personalize psoriasis therapy, overcoming the limitations of traditional statistical methods and enhancing predictive accuracy for clinical decision-making.

CHAPTER 3

Materials and Methods

3.1. The Dataset

Link	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151278	
Title	Genome-wide DNA methylation analysis of peripheral blood samples of moderate-to-severe psoriasis patients treated with anti-TNF drugs	
Publish Date	May 28, 2020	
Summary	Genome wide DNA methylation profiling of peripheral blood samples of moderate-to-severe psoriasis patients treated with anti-TNF drugs. Patients were distributed on Excellent Responders (ER) if they achieved PASI90 (a 90% reduction with respect to baseline PASI) at 3 and 6 months of treatment with anti-TNF drugs and Partial responders if they did not achieve a PASI75 (a 75% reduction with respect to baseline PASI) at 3 and 6 months of treatment. The Illumina Infinium 450k Human DNA methylation Beadchip v1.2 was used to obtain DNA methylation profiles across approximately 485,000 CpGs in 49 ER and 21 PR which were obtained from peripheral blood samples of anti-TNF drug treated patients. We have searched for pharmaco epigenetic biomarkers of anti-TNF response in moderate-to-severe psoriasis patients.	
Samples	70	
Organization name	Instituto de Investigación Sanitaria la Princesa (IIS-IP) MADRID, Spain	
Features (inputs)	Age	
	Age at initiation	
	Gender	Male/female
	Treatment Type	Adalimumab- Etanercept- infliximab
Labels (Targets)	Response to Anti-TNF Therapy	Responder (1) Non-Responder (0)

Table 3.1.The data set

Blood samples were collected from adult patients diagnosed with moderate-to-severe plaque psoriasis, as defined by the Spanish Academy of Dermatology and Venereology Psoriasis Working Group guidelines. These patients were undergoing treatment with anti-TNF agents (adalimumab, infliximab, or etanercept) and provided written informed consent. The study protocol and consent forms adhered to Spanish regulations on biomedical and clinical research and were approved by the Ethics Committee for Clinical Research of Hospital Universitario de la Princesa.

To enhance the contrast in treatment outcomes, patients with extreme phenotypic responses to anti-TNF therapy were selected. They were categorized into two groups:

1. Excellent responders (ER): Patients who achieved a PASI90 response (≥90% improvement from baseline Psoriasis Area and Severity Index score) at both 3 and 6 months.
2. Partial responders (PR): Patients who failed to reach a PASI75 response (≥75% improvement from baseline PASI score).

DNA Extraction and Methylation Analysis: Genomic DNA was isolated from peripheral blood samples using the MagNA Pure® System (Roche Applied Science, Penzberg, Germany). DNA integrity was assessed using the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Subsequently, 1,000 ng of genomic DNA underwent bisulfite conversion using the EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA).

Genome-wide DNA methylation profiling was conducted using the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, CA, USA) following the manufacturer's protocol. The methylation data generated in this study have been deposited in NCBI's Gene Expression Omnibus (GEO) and are publicly available under the accession number **GSE151278:**(https://www.ncbi.nlm.nih.gov/geo/). (10)

3.2. Tools Used

a) Computational Environment:

All machine learning workflows were implemented in Python 3.8+ using Google Colaboratory (Colab), a cloud-based Jupyter notebook platform.

b) Libraries & Dependencies:

pandas	A powerful data manipulation and analysis library, offering data structures like Data Frames and Series for handling structured data.
numpy	A fundamental library for numerical computing in Python, providing support for large, multi-dimensional arrays and matrices, along with mathematical functions
Matplotlib.pyplot	A popular plotting library for creating static, interactive, and animated visualizations in Python.
gzip	A Python library for compressing and decompressing files
scikit-learn (sklearn)	A popular machine learning library for Python, providing simple and efficient tools for data mining and data analysis
IO	A built-in Python module providing core tools for input/output (I/O) operations, supporting file handling, streams, and in-memory buffers.
stringIO	A built-in Python module (part of IO) that allows treating strings as file-like objects in memory, supporting read/write operations like a file.
sklearn.decomposition	A submodule in scikit-learn for dimensionality reduction and matrix factorization techniques
PCA (Principal Component Analysis)	A linear dimensionality reduction technique in sklearn .decomposition that transforms data into orthogonal components (ordered by variance)
sklearn.preprocessing	A scikit-learn submodule for data preprocessing and feature scaling, essential for preparing data before machine learning modeling
StandardScaler	A preprocessing tool that standardizes features
LabelEncoder	A preprocessing tool that encodes categorical labels (strings or integers) into numerical values

sklearn.model_selection	A scikit-learn submodule for model evaluation, selection, and hyperparameter tuning
train_test_split	A function to split datasets into random training and testing subsets, commonly used for model validation.
sklearn.ensemble	A scikit-learn submodule for ensemble learning, combining multiple base models to improve predictive performance and robustness
Random Forest Classifier	An ensemble learning method that constructs multiple decision trees during training and outputs the majority vote (classification) or average prediction (regression).
sklearn.metrics	A scikit-learn submodule for model evaluation, providing functions to measure performance for classification, regression, clustering, and more
classification_report	A function that generates a text summary of key classification metrics, including precision, recall, F1-score, and support for each class.
accuracy_score	A function that computes the accuracy classification score, i.e., the fraction of correct predictions out of all predictions made.
Requests	A popular, user-friendly HTTP library for Python, designed for making web requests

Table3.2. Libraries & Dependencies

3.3. Workflow and Data Preprocessing

Data has been loaded directly to colab by defining dataset URL and local file path then using Python's requests library to fetch the data. Pandas.read_csv has been used directly with URL for gzipped files.

Metadata parsing: Geo files are characterized by a distinctive format that begins with metadata lines starting with an exclamation mark (!).A function has been defined to handle this format, which systematically segregates metadata (denoted by '!' prefixes) from primary data entries. The implemented pipeline successfully imported the metadata (marked by '!' prefixes) and expression data matrices, establishing a complete dataset for processing.

Expression Data Head (first 5 rows, first 5 columns):

ID_REF	GSM4570206	GSM4570207	GSM4570208	GSM4570209	GSM4570210
cg00000029	0.449950	0.483412	0.383950	0.499973	0.368064
cg00000108	0.919370	0.889380	0.879016	0.885480	0.884407
cg00000109	0.720833	0.675689	0.682002	0.686516	0.728309
cg00000165	0.285152	0.212334	0.200122	0.184203	0.192691
cg00000236	0.683391	0.650948	0.686666	0.585181	0.701103

Table3.3. Expression Data Head (first 5 rows, first 5 columns)

Initial data visualization: To ensure data integrity, all expression values were verified to be numeric prior to analysis. Subsequently, two key visualizations were generated:

Plot 1: Distribution of All Expression Values. A density plot displaying the global distribution of beta value across the entire dataset.

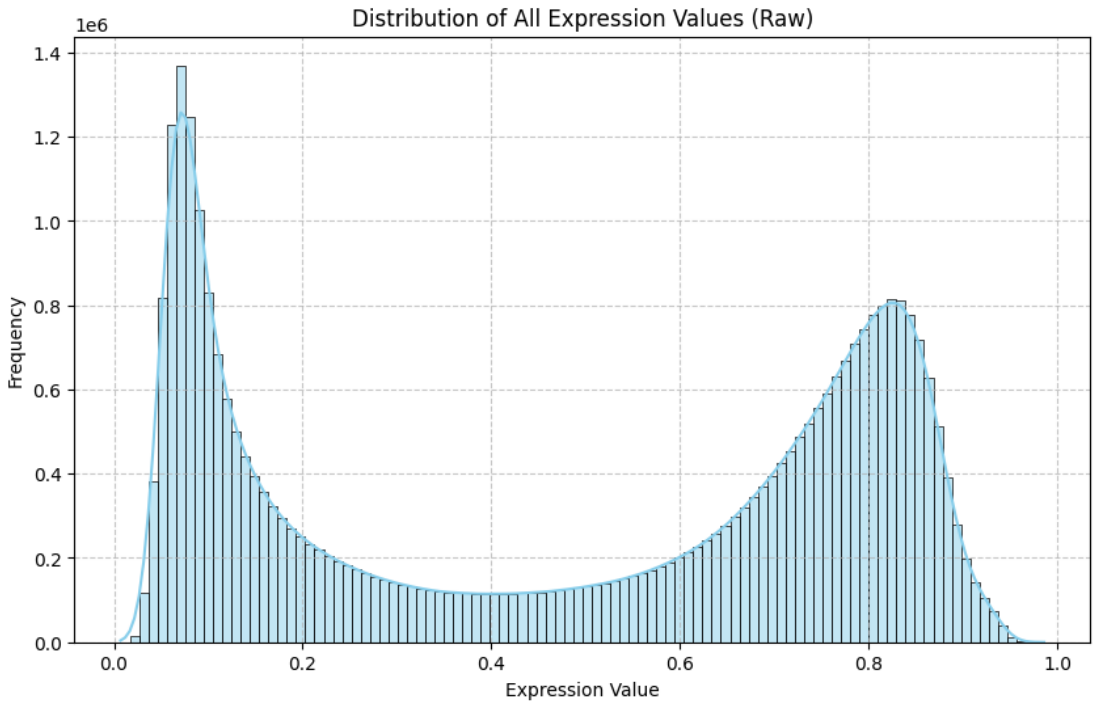


Figure 3.1. Data Distribution

Plot 2: Distribution of Expression Values for 20 Samples – A (violin plot) illustrating the variation in expression profiles across a subset of 20 representative samples.

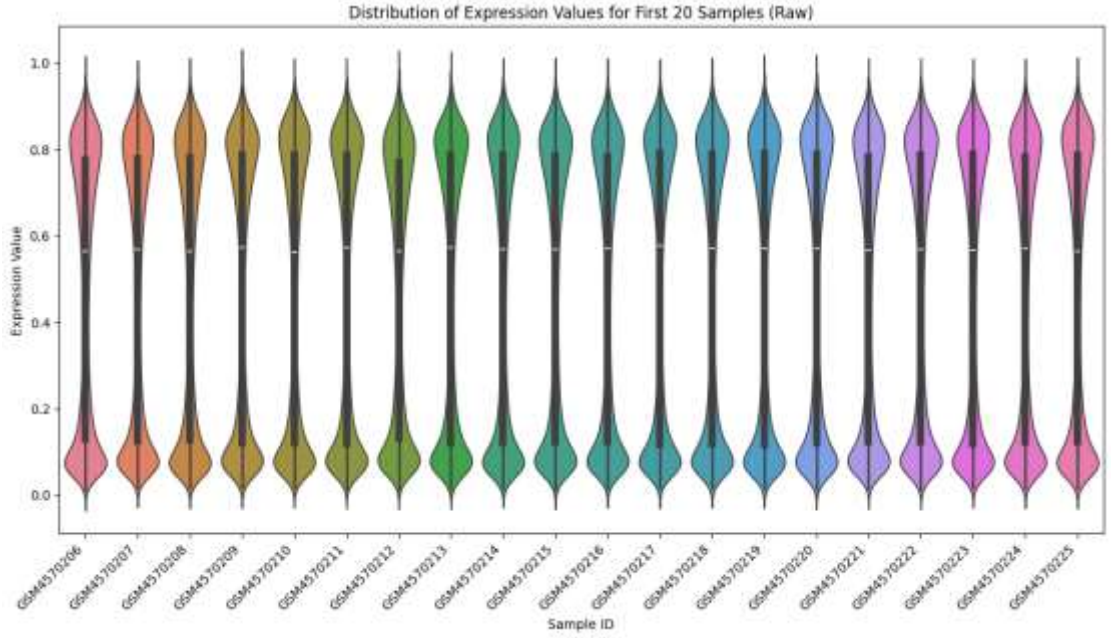


Figure 3.2. violin plot

Data Preprocessing:

The machine learning model expects samples as rows and features as columns; hence, the data matrix has been transposed**.**

Subsequently, the independent variables (age, gender, age at initiation,drug) and the dependent variable (response) have been extracted from metadata.

Sample IDs have been extracted from the index (X) to ensure consistent tracking of observations throughout the analysis.

Label encoder has been applied to transform categorical variables into numerical representations

ID_REF	Gender	Age	Age_at_Initiation	cg00000029	cg00000108
GSM4570206	1	57	42	0.449950	0.919370
GSM4570207	0	38	23	0.483412	0.889380
GSM4570208	0	62	18	0.383950	0.879016
GSM4570209	1	48	29	0.499973	0.885480
GSM4570210	1	66	37	0.368064	0.884407

Table 3.4. Data matrix

All features have been standardized using scikit-learn'sStandardScaler

--- Standardizing Features (X) ---

ID_REF	Gender	Age	Age_at_Initiation	cg00000029	cg00000108
GSM4570206	0.792406	0.685028	1.404207	0.070306	2.500579
GSM4570207	-1.261980	-0.625889	-0.245612	0.761601	0.790761
GSM4570208	-1.261980	1.030007	-0.679775	-1.293154	0.199858
GSM4570209	0.792406	0.064067	0.275383	1.103738	0.568417
GSM4570210	0.792406	1.305989	0.970044	-1.621342	0.507228

Table3.5. Standardizing Features (X)

- **Perform the mean and variance**

To ensure robust feature selection, we calculated

the **mean** and **variance** of DNA methylation levels (β-values) across all samples for each CpG site.

ID_REF	Mean	Variance
ID_REF		
cg00000029	0.446546	0.002377
cg00000108	0.875510	0.000312
cg00000109	0.717971	0.000911
cg00000165	0.220072	0.001042
cg00000236	0.659683	0.001223

Table3.6.Mean and Variance

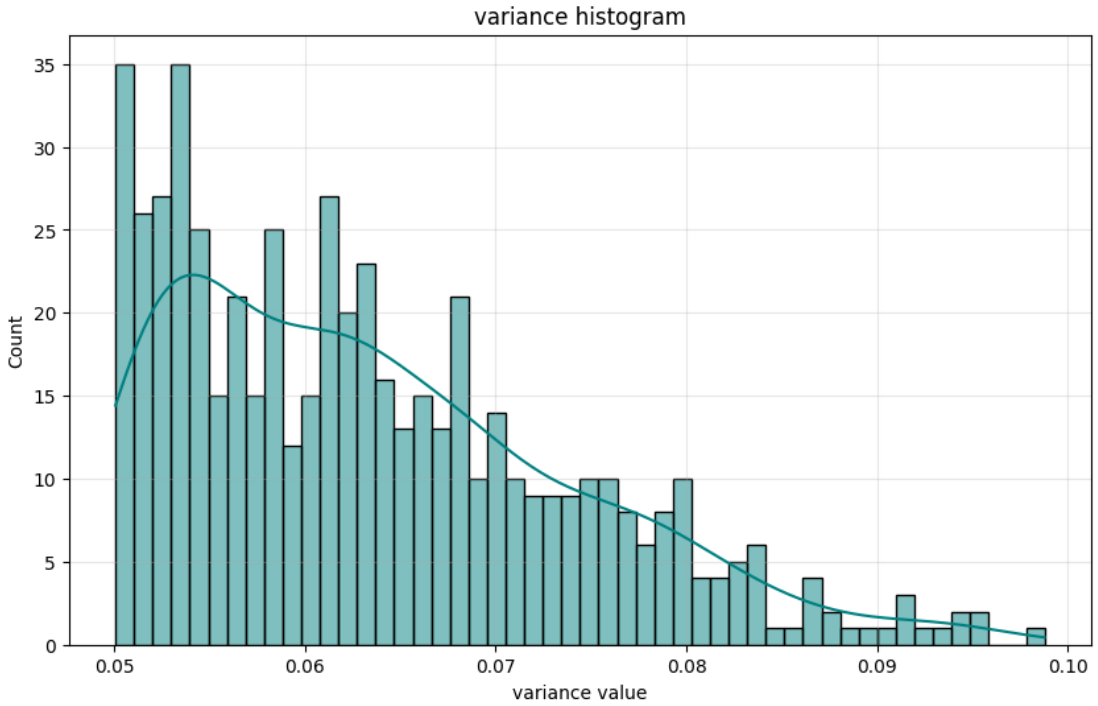


Figure 3.3. Samples with variance values>0.05

3.4. Chapter Closure

In a preliminary analysis of the data, we observe that:

DNA methylation levels are not normally distributed; instead, they tend to cluster into either high or low values. Generally, low methylation levels (hypomethylation)

in promoter regions are associated with active gene expression, whereas high methylation levels (hypermethylation) are linked to gene silencing. However, the relationship between methylation and gene expression can vary depending on genomic context, such as enhancers or gene bodies, where methylation may have different regulatory role.

DNA methylation profiling identified >430,000 CpG sites per patient, annotated as cg x (where "x" is an 8-digit probe ID).

Chapter 4

Prediction Using Machine Learning

4.1. Methodology

4.1.1. Predictive Models

Several predictive models have been trained and evaluated, with priority given to established algorithms demonstrating high potential for performance accuracy.

- logistic regression
- Decision Tree Classifier
- Random Forrest Classifier
- Support Vector Machine – SVM
- Multi-Layer Perceptron Classifier (MLPClassifier)

Logistic Regression

Logistic Regression is a statistical model used for classification that predicts the probability of an instance belonging to a class. While commonly used for binary classification (e.g., Yes/No), it can be extended to multiclass problems (3+ classes) using:

Binary Logistic Regression uses the sigmoid function to output a probability between 0 and 1. Equation:

$$P(y = 1 \mid \mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^T \mathbf{X})}}$$

Multinomial Logistic Regression (Softmax): Generalizes to multiple classes by assigning probabilities using the softmax function. One-vs-Rest (OvR): Trains multiple binary classifiers (one per class).

Properties of logistic regression:

- Interpretable (coefficients show feature importance).
- Works for both binary and multiclass tasks.

Limitations:

- Poor performance on non-linear decision boundaries. (11)

Decision Tree Classifier

A Decision Tree splits data into branches based on feature values to classify instances. It uses rules like Gini impurity or entropy to select optimal splits. Trees are intuitive (mimic human decision-making) but prone to overfitting without constraints (e.g., max depth).

For node m, the Gini impurity is:

$$G_m = 1 - \sum_{k=1}^{K} p_{mk}^2,$$

where Pmk is the proportion of class k in node m.

Key Properties:

- Non-parametric: No assumptions about data distribution.
- Transparency: Rules are human-readable (unlike "black-box" models).

Limitations:

- High variance; small data changes alter tree structure. (12)

Random Forest Classifier

Random Forest is an ensemble method that builds multiple decision trees on random subsets of data and features, then aggregates their predictions (majority vote for classification). It reduces overfitting and improves accuracy compared to single trees.

A Random Forest is an **ensemble** of *B* decision trees trained on bootstrap samples of the data (**bagging**). Each split uses a random subset of features (size $m \approx \sqrt{p}$ for p features).

Properties:

- Variance reduction: Averaging over trees decreases overfitting.
- Robustness: Handles noisy data and outliers.

Limitations**:**

- Computationally expensive for large *B*. (13)

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression. It works by finding the optimal hyperplane that maximizes the margin between classes in a high-dimensional feature space. SVM can handle both linear and non-linear decision boundaries using kernel functions.

Linear SVM (Hard Margin): Minimizes$1/2\|w\|^2$ subject to

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

where **w** is the weight vector and *b* is the bias.

Soft Margin (C-SVM): Introduces slack variables ξi to handle misclassifications:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

where *C* controls the trade-off between margin width and classification error.

Kernel Trick: Maps data to a higher-dimensional space using kernel functions (e.g., RBF, polynomial):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad \text{(RBF kernel)}$$

Advantages:

- Effective in high-dimensional spaces.
- Robust to overfitting, especially with small datasets.
- Versatile (works with linear and non-linear data via kernels).

Limitations:

- Computationally intensive for large datasets.

- Requires careful tuning of hyperparameters (e.g., C, γ).

Multi-Layer Perceptron Classifier (MLPClassifier)

The MLPClassifier (Multi-Layer Perceptron Classifier) is a feedforward artificial neural network (ANN) used for supervised learning tasks, particularly classification. It consists of multiple layers of interconnected neurons (nodes) that learn non-linear decision boundaries through backpropagation and gradient descent optimization.

Components:

- **Architecture:**
  - Input Layer: Receives feature vectors.
  - Hidden Layers: One or more layers with activation functions .
  - Output Layer: Uses Softmax (for multi-class) or Sigmoid (for binary) activation.
- **Mathematical Formulation (Forward Pass):**

  For layer l:

$$\mathbf{z}^l = \mathbf{W}^l \mathbf{a}^{l-1} + \mathbf{b}^l, \quad \mathbf{a}^l = \sigma(\mathbf{z}^l)$$

- where $\mathbf{W}l$ =weights, $\mathbf{b}l$ =biases, σ =activation function.
- **Loss Function:** Cross-entropy (classification).
- **Optimization:** Stochastic Gradient Descent (SGD) or Adam.

Advantages:

- Handles non-linear relationships via hidden layers.
- Flexible architecture (adaptable to various problems).

Limitations:

- Prone to overfitting (requires regularization like dropout or L2 penalty).
- Computationally expensive for large networks.

(14).

4.1.2. Stratified K-Fold Cross-Validation

Stratified K-Fold is a cross-validation technique that preserves the class distribution (stratification) in each fold. It partitions the dataset into K subsets (*folds*) of approximately equal size, ensuring that each fold maintains the same percentage of samples for each class as in the original dataset.

Features:

- Preserves Class Balance: Critical for imbalanced datasets.
- Reduces Bias: Provides more reliable performance estimates than standard K-Fold.
- Model Evaluation: Each fold serves as a validation set once while the remaining $K-1$ folds are used for training.

Advantages:

- More accurate performance estimation for classification tasks.
- Mitigates overfitting in imbalanced scenarios.

Limitations:

- Computationally intensive for large K.
- Not suitable for regression (use standard K-Fold instead)

(15)

4.2. Implementation and Results :

After performing the initial preprocessing, the data has been ready for the training phase. As mentioned in the previous chapter, we have used the Google Colab environment. We have split the data into 80% for training and 20% for testing, increasing the training set size due to the limited original dataset size. Additionally, we have employed **Stratified K-Fold cross-validation (K=5)** to handle the class imbalance, as the number of treatment responders is significantly higer than non-responders.

To reduce dimensionality and remove low-variance features, Variance Threshold filtering has been applied. Through repeated model training and statistical evaluation, the variance threshold range of 0.005-0.8 has been identified as the most impactful range.

A subset of previously extracted metadata features (e.g., age, gender, age at initiation, drug) was identified as non-predictive. These features were excluded from final models.

Model Performance Evaluation

Following hyperparameter optimization and cross-validation, the trained models demonstrated the following performance metrics on the held-out test set:

Model	CV Mean Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1
RForest	0.750	0.785	0.835	0.785	0.735
Logest_Reg	0.713	0.785	0.835	0785	0.735
NN	0.680	0.785	0.774	0.785	0.775
SVM	0.696	0.714	0.510	0.714	0.595
DTree	0.521	0.642	0.669	0.642	0.653

Table4.1 Model Performance

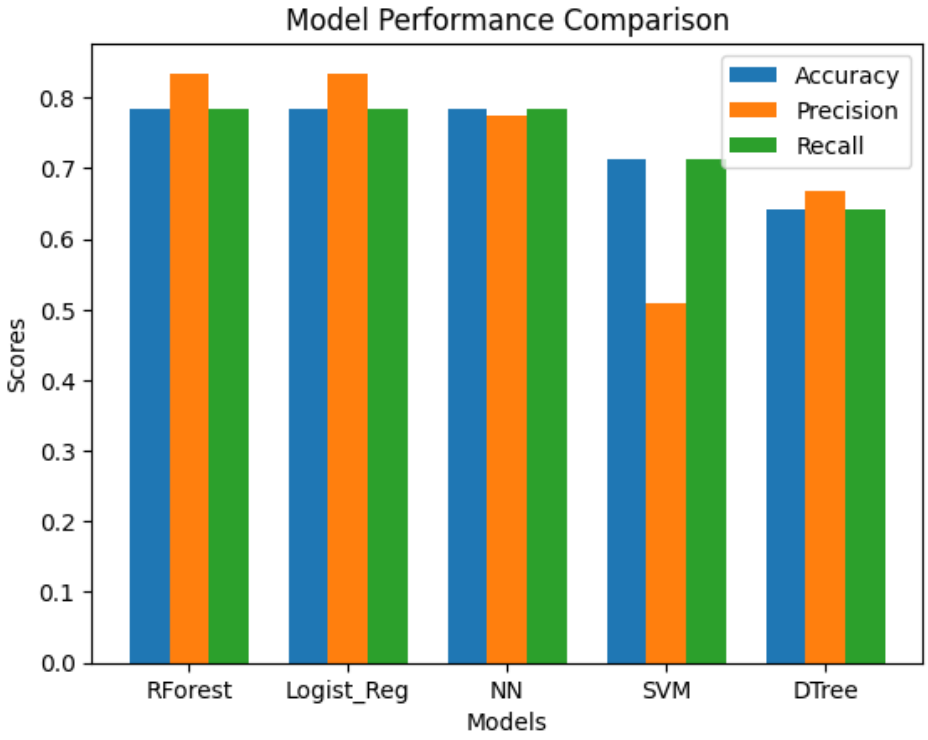


Figure4.1 Model performance comparison

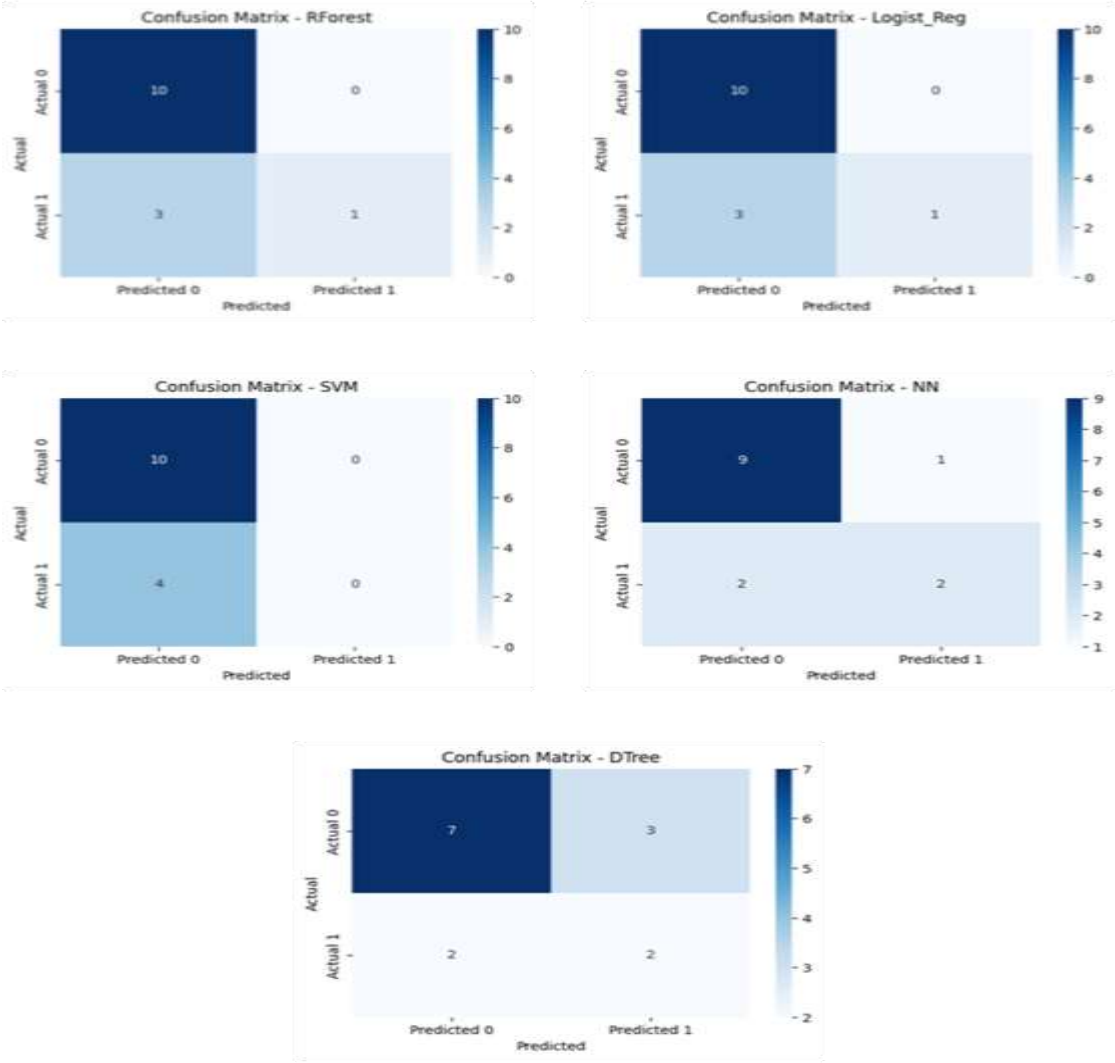


Figure4.2. Model confusion matrix

Random Forest and **Logistic Regression** achieved the highest cross-validation (CV) accuracy and identical test accuracy (**0.7857**). However, Random Forest demonstrated greater stability (lower standard deviation).

The **Neural Network** attained the highest **F1-score (0.7755)** despite its lower CV accuracy, suggesting a better balance between precision and recall compared to other models.

SVM struggled with low **precision (0.5102)**, significantly reducing its F1-score, even though its recall was acceptable.

Decision Tree performed the weakest across most metrics.

Based on rigorous evaluation (CV accuracy, precision, recall, and F1-score), Random Forests emerged as the most effective model.

Variable Importance Analysis for the Best Model (Random Forest)

In the Random Forest model, assessing variable importance helps identify which features have the most significant impact on predictions**.** The following 25 CpG sites were identified as the most influential variables in the predictive model, ranked by their importance scores:

ID_REF	variable	variance	importance
cg19430537	cg19430537	0.005051	0.015181
cg16045423	cg16045423	0.005765	0.011622
cg09969882	cg09969882	0.012825	0.007985
cg26547816	cg26547816	0.005669	0.007942
cg14638919	cg14638919	0.030298	0.007700
cg15935227	cg15935227	0.019368	0.007478
cg10075506	cg10075506	0.024824	0.006818
cg16586594	cg16586594	0.009493	0.006793
cg00169354	cg00169354	0.023266	0.006658
cg17471939	cg17471939	0.008031	0.006549
cg00168694	cg00168694	0.005731	0.006379
cg05645557	cg05645557	0.006956	0.005951
cg21790587	cg21790587	0.007325	0.005869
cg17833169	cg17833169	0.010260	0.005844
cg15519096	cg15519096	0.005003	0.005663
cg13410614	cg13410614	0.006876	0.005520
cg12738248	cg12738248	0.030968	0.005513
cg10701640	cg10701640	0.006511	0.005511
cg12728606	cg12728606	0.005570	0.005422
cg03292213	cg03292213	0.010938	0.005385
cg04269043	cg04269043	0.005830	0.005311
cg18584424	cg18584424	0.006924	0.005311
cg03666441	cg03666441	0.005157	0.005300
cg18757828	cg18757828	0.022755	0.005176
cg17906168	cg17906168	0.006565	0.005158

Table 4.2.Top 25 Most Important CpG Sites

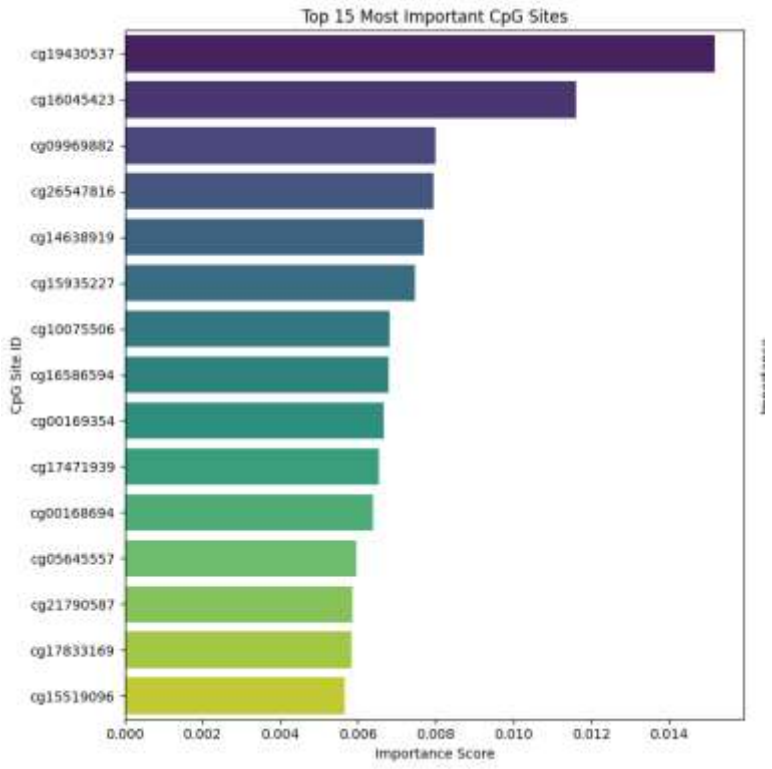


Figure4.3.Top 15 Most Important CpG Sites

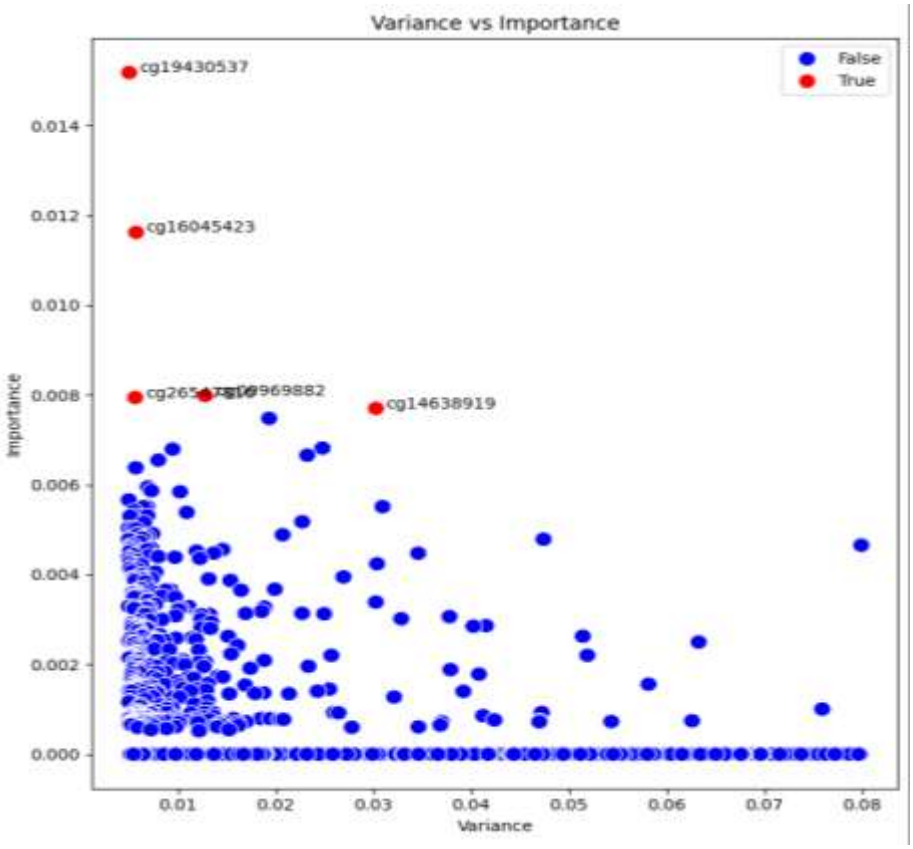


Figure4.4.Variance vs Importance

4.3. Improvement

Some trials are made to enhance the results:

- PCA, a linear dimensionality reduction technique, was applied but failed to enhance performance. This implies that either (1) key predictive features were correlated with low-variance components, or (2) non-linear feature interactions dominate the data's discriminative structure.
- Dimensionality reduction was performed by selecting features with variance thresholds between 0.005 and 0.08. This optimized feature subset led to significant performance improvements, suggesting that:
    1. Features with very low variance (<0.005) were likely noise-dominated and their removal enhanced model robustness.
    2. Retaining features within this variance range preserved discriminative patterns while eliminating redundancy.
    3. The upper threshold (0.08) effectively prevented high-variance features from dominating the feature space.

4.4. Discussion

Our findings demonstrate that the Random Forest model outperformed other approaches in predicting treatment response (CV Mean Accuracy 0.750, test accuracy 0.785, test prcecision0.835, test recall0785, F1 0.735) highlighting its efficacy in capturing complex interactions among variables. Key methylation sites were identified as top predictors. The three most influential DNA methylation sites were examined to validate that the model's predictions were biologically meaningful and not due to random noise. Using the UCSC database, the genomic positions of this methylation sites and the associated genes were identified.

1) First site: cg19430537

The CpG site **cg19430537** (**chr17**:74,128,860-74,128,860) .This site is part of a robust epigenetic signature specific to $CD8^+$T cells, capable of accurately inferring cell-type-specific methylation from bulk blood data (16). Consequently, its association with treatment response in psoriasis likely reflects a direct role in modulating $CD8^+$T cell biology, a key player in psoriatic inflammation, thereby providing a compelling biological rationale for its predictive value.

2) Second site: cg16045423

The CpG site **cg16045423** (chr22:39,378,346-39,378,346) is located within the *APOBEC3B* gene body. This gene is a member of the cytidine deaminase gene family. Given the crucial role of the PKC/classical NF-κB pathway in the transcriptional regulation of, *APOBEC3B* in cancer cells (17) and considering that this pathway is hyperactivated in psoriatic plaques, we propose that *APOBEC3B* expression may serve as a biomarker for predicting response to anti-TNF therapy.

.

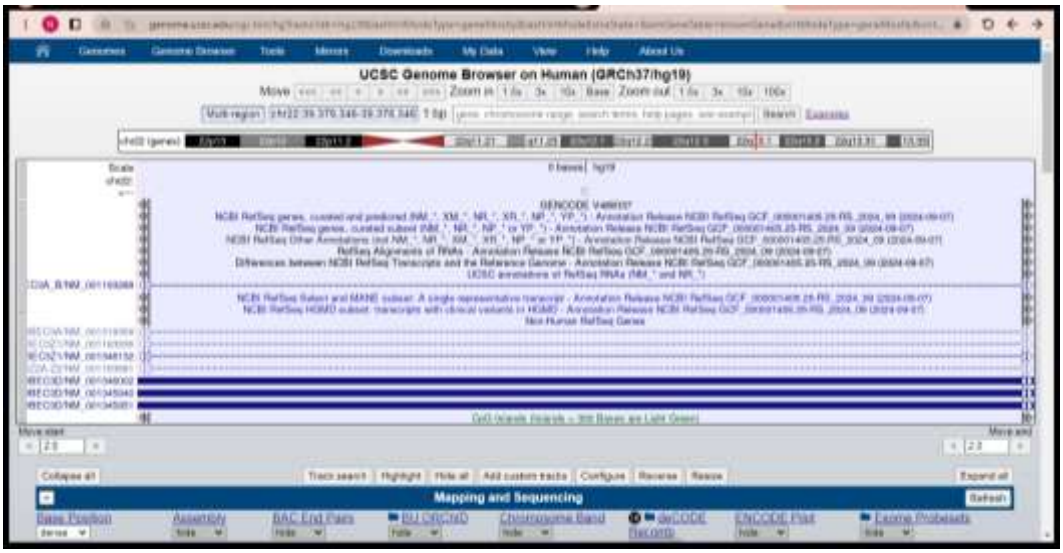


Figure 4.5 cg16045423 genomic position

3) Third site:cg09969882

The CpG site **cg09969882** (chr2:239346400-239346400) is located within the ***ASB1*** gene**.** Unlike typical SOCS box proteins that promote substrate degradation, ASB1 unexpectedly stabilizes its substrate, TAB2, by inhibiting its K48-linked ubiquitination. This enhancement of TAB2 stability leads to amplified activation of downstream NF-κB inflammatory pathway (18). This stabilization of TAB2 by ASB1 provides a potential molecular mechanism for the hyperactive inflammation observed in chronic diseases such as psoriasis. It may also explain the differential patient responses to therapies targeting upstream cytokines (e.g., anti-TNF)**.**

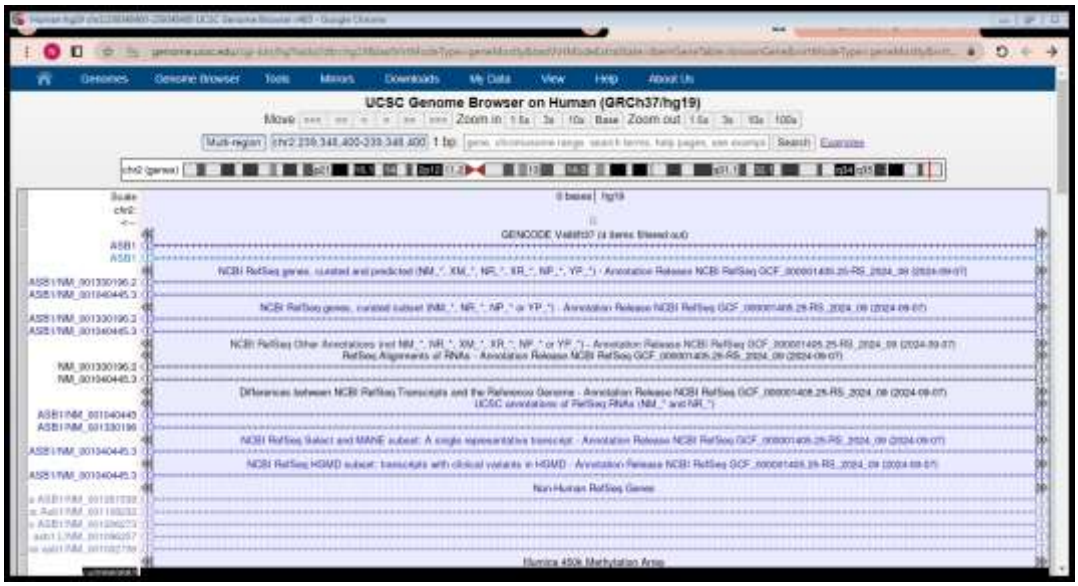


Figure.4.6 cg09969882 position

We observe that the top three variables in the model correspond to genomic loci where variation in methylation may affect the transcription of genes encoding proteins directly involved in the immunological and molecular mechanisms of psoriasis. This indicates that the model reflects biologically plausible associations, not merely artificial or noise-driven correlations**.**

Future studies should focus on these variables to further elucidate their precise role in (antiTNF-α response in psoriasis patients)**.**

CHAPTER 5

Conclusion and Future Prospects

5.1. Conclusion

In this study, we developed a machine learning framework leveraging DNA methylation data to predict anti-TNF-α response in psoriasis patients. Among the five models tested, Random Forest demonstrated the best performance (79% accuracy), highlighting its potential for clinical stratification. Notably, the top three predictive variables were biologically relevant, mapping to genomic loci implicated in psoriasis-related immune pathways. While these results are promising, further validation in larger cohorts and integration of multi-omics data could enhance predictive power. Our approach underscores the utility of AI in personalized dermatology, paving the way for more targeted therapeutic decisions.

5.2. Future Prospects

1. **Validation in Larger, Prospective, and Diverse Cohorts**: The performance of our model must be rigorously validated in larger, multi-center, prospective cohorts that encompass greater ethnic, genetic, and clinical diversity. This is essential to confirm generalizability, assess potential confounding factors, and ultimately ensure the model's robustness before clinical deployment.
2. **Model Optimization and Advanced Architectures:** Although Random Forest demonstrated superior performance, exploring more complex and sophisticated algorithms could yield further improvements.
3. **Multi-Omics Data Integration**: To move beyond a predictive model towards a mechanistic understanding, future work should integrate DNA methylation data with other molecular layers. A multi-omics approach incorporating matched transcriptomic (RNA-seq), proteomic, and genomic data would provide a more comprehensive systems biology view. This could unravel the functional consequences of epigenetic changes and identify master regulators of treatment response.

4. **Functional Characterization of Top Predictive Loci:** The biological relevance of our top predictors (e.g., cg19430537) is a major strength. We propose dedicated functional studies to elucidate their causal role. This could involve in vitro experiments.

REFERENCES

1. **Jen-Chih Tseng 1, Yung-Chi Chang 2, Chun-Ming Huang 3, Li-Chung Hsu 2,4,* and Tsung-Hsien Chuang 1,*.** Therapeutic Development Based on the Immunopathogenic. *Pharmaceutics 2021, 13, 1064. https://doi.org/10.3390/pharmaceutics13071064.* july 11, 2021.

2. **Albanesi C, Madonna S, Gisondi P and Girolomoni G.** The Interplay Between Keratinocytes and Immune Cells in the Pathogenesis of Psoriasis. 2018.

3. **Iversen, Kirsten Rønholt * and Lars.** Old and New Biological Therapies for Psoriasis. *international journal of molecular science.* november 1, 2017.

4. **Eng, Grith Petersen.** Optimizing biological treatment in rheumatoid arthritis with the aid of therapeutic drug monitoring. *Danish Medical Journal.* November 2016.

5. **Elisa Camela, Luca Potestio, Gabriella Fabbrocini, Angelo Ruggiero & Matteo.** New frontiers in personalized medicine in psoriasis. *Expert Opinion on Biological Therapy.* Augest 16, 2022.

6. **J. Tang*†a MD, Y. Xiong*†a MD, H.-H. Zhou*† MD PhD and X.-P. Chen*† MD PhD.** DNA methylation and personalized medicine. *Journal of Clinical Pharmacy and Therapeutics,.* August 17, 2014.

7. **Ting Wei1, †, Jinfu Nie2,†, Nicholas B. Larson 1, Zhenqing Ye1,.** CpGtools: a python package for DNA methylation. *Bioinformatics, 37(11), 2021, 1598–1599.* december 4, 2019.

8. **Mia-Louise Nielsen, MSc1, et al., et al.** Multivariable Predictive Models to Identify the Optimal Biologic Therapy for Treatment of Patients With Psoriasis at the Individual Level. *JAMA Dermatol.* august 17, 2022.

9. Identifying Predictors of PASI100 Responses up to Month 12 in Patients with Moderate-to-severe Psoriasis Receiving Biologics in the Psoriasis Study of Health Outcomes (PSoHO).

10. **Ancor SANZ-GARCÍA1, Alejandra REOLID2, Laura H. FISAS3, Ester MUÑOZ-ACEITUNO2, Mar LLAMAS-VELASCO2, Antonio.** DNA Copy Number Variation Associated with Anti-tumour Necrosis Factor Drug Response and Paradoxical Psoriasiform Reactions in Patients with Moderate-to-severe Psoriasis. *Acta Derm Venereol.* april 13, 2021.

11. **Hastie, T. (Trevor),Tibshirani, R. (Robert),Friedman, J. (Jerome).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* s.l. : Springer, 2009.

12. **Breiman, L. (Leo), Friedman, J. (Jerome), Stone, C. J. (Charles J.), & Olshen, R. A. (Richard A.** *Classification and Regression Trees.* s.l. : CRC Press, 1984.

13. **Breiman, L. (Leo.** Random Forests. *Machine Learning.* 2001.

14. **Goodfellow, I. (Ian), Bengio, Y. (Yoshua), & Courville, A. (Aaron).** *Deep Learning.* 2016.

15. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É.** Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR).* 2011.

16. **Daniel W. Kennedy, Nicole M. White,Miles C. Benton,Andrew Fox,Rodney J. Scott,Lyn R. Griffiths,Kerrie Mengersen,Rodney A. Le.** Critical evaluation of linear regression models for cell-subtype specific methylation signal from mixed blood cell DNA. *PLOS one.* december 20, 2018.

17. **Wataru Maruyama, Kotaro Shirakawa, Hiroyuki Matsui, Tadahiko Matsumoto,Hiroyuki Yamazaki, Anamaria D. Sarca, Yasuhiro Kazuma, Masayuki Kobayashi Keisuke Shindo, Akifumi Takaori-Kondo.** Classical NF-κB pathway is responsible for APOBEC3B expression in cancer cells. *Biochemical and Biophysical Research Communications.* 2016.

18. **Panpan Hou a, 1, Penghui Jia a,1, Kongxiang Yang b,1 , Zibo Li a , Tian Tian c, Yuxin Lina , Weijie Zeng a, Fan Xinga,Yu Chen b, Chunmei Li a, Yingfang Liu a, and Deyin Guo a.** An unconventional role of an ASB family protein inNF-κB activation and inflammatory response duringmicrobial infection and colitis. *immunology and inflammation.*

19. **: Ovejero-Benito MC, Cabaleiro T, Sanz-García A, Llamas-Velasco M et al.** Epigenetic biomarkers associated with antitumour necrosis factor drug response in moderate-to-severe psoriasis. *Br J Dermatol.* march 2018.

20. **Amy X. Du1, Zarqa Ali2 , Kawa K. Ajgeiy3, Maiken G. Dalager4, Tomas N. Dam5, Alexander Egebjerg2, Christoffer V. S. Nissen2, Lone Skov6, Simon Francis Thomsen2, Sepideh Emam7*, Robert Gniadecki1*†.** Machine Learning Model for Predicting Outcomes of Biologic Therapy in Psoriasis. *Journal of the American Academy of Dermatology .* december 7, 2021.

21. Support-Vector Networks. *Machine Learning .* septemper 1995.

22. **Matsui, H., Yokoyama, T., Sekiguchi, T., Iijima, D., Sunaga, S., Maniwa, Y., ... & Matsuo.** Interferon-γ induces APOBEC3B expression through JAK-STAT signaling in human keratinocytes. *Journal of Investigative Dermatology.* 2010.

23. **Lande, R., Botti, E., Jandus, C., Dojcinovic, D., Fanelli, G., Conrad, C., ... & Gilliet.** Plasmacytoid dendritic cells sense self-DNA coupled with antimicrobial peptide LL-37 in psoriasi. *Science Translational Medicine.* 2014.

24. **Hastie, T. (Trevor),Tibshirani, R. (Robert),Friedman, J. (Jerome).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* s.l. : Springer, 2009.