EXPLORING MACHINE LEARNING APPROACHES IN PREDICTION OF AUTOIMMUNE DISORDERS USING EPIGENETIC MODIFICATIONS

A Thesis Submitted
In Partial Fulfilment of the Requirements for the Degree of

MASTER OF SCIENCE
in
BIOTECHNOLOGY

by
SHREYA KOHLI
(23/MSCBIO/44)

Under the Supervision of

PROF. YASHA HASIJA
Professor and Head of Department
Department of Biotechnology
Delhi Technological University

Department of Biotechnology

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India
June, 2025

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, Prof. Yasha Hasija, for her constant support, encouragement, and invaluable guidance throughout this research. Her insightful direction and unwavering belief in my work significantly shaped the course of this study.

I am also sincerely thankful to the Department of Biotechnology, Delhi Technological University (DTU) for providing the necessary facilities and resources that were crucial for the successful completion of this research.

My heartfelt thanks go to the PhD scholars Ms. Akanksha Bisht and Ms. Khushi Yadav for their immense support and guidance in my day-to-day work. Their expertise, patience, and mentorship were truly invaluable.

Lastly, I would like to thank my family, especially my brother Aayush and sister Anika, for their unwavering support, motivation, and constant encouragement. Their love and camaraderie played a vital role in helping me complete this thesis.

Shreya Kohli
23/MSCBIO/44

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, Shreya Kohli, bearing Roll No. 23/MSCBIO/44 hereby certify that the work which is being presented in the thesis entitled "**EXPLORING MACHINE LEARNING APPROACHES IN PREDICTION OF AUTOIMMUNE DISORDERS USING EPIGENETIC MODIFICATIONS.**" in partial fulfilment of the requirements for the award of the Degree of Master of Science , submitted in the Department of Biotechnology , Delhi Technological University is an authentic record of my own work carried out during the period from January 2025 to May 2025 under the supervision of Prof. Yasha Hasija.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that **SHREYA KOHLI (**23/MSCBIO/44) has carried out their search work presented in this thesis entitled **"EXPLORING MACHINE LEARNING APPROACHES IN PREDICTION OF AUTOIMMUNE DISORDERS USING EPIGENETIC MODIFICATIONS"** for the award of **Master of Science** from Department of Biotechnology, Delhi Technological University, Delhi, under my supervision. The thesis embodies the results of original work, and studies are carried out by the student herself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:

Prof. Yasha Hasija
Professor and Head of Department
Department of Biotechnology
Delhi Technological University
Delhi - 110042

EXPLORING MACHINE LEARNING APPROACHES IN PREDICTION OF AUTOIMMUNE DISORDERS USING EPIGENETIC MODIFICATIONS

Shreya Kohli

ABSTRACT

Epigenetic Modifications occur due to an interplay between the genetic and environmental factors and are an important cause for the occurrence of autoimmune disorders where there is disruption of immune tolerance and the body starts to attack its own cells. DNA methylation, histone modifications, and ncRNAs are the major modifications observed in autoimmune disorders such as Systemic Lupus Erythematosus, Rheumatoid Arthritis, Multiple Sclerosis etc. Due to the recent advancements in Artificial intelligence and Machine learning, training of epigenetic data can be used to obtain better predictive and analytical responses. In this thesis we assess the application of various Machine learning techniques such as Supervised, Unsupervised and Deep learning in models after preprocessing and evaluation, which are implied to find the predictive capabilities of the epigenetic data in the diagnosis of autoimmune conditions. We used a DNA methylation profile dataset for Systemic Lupus Erythematosus and applied machine learning algorithms such as Random Forest, Support vector machines, Logistic Regression, XGBoost, Naive Bayes and Artificial neural networks and their evaluation metrics were obtained. It was concluded that Support vector machines worked the best for model development and gave an AUROC of 0.97. Gaps in the current knowledge along with future implications are also highlighted.

LIST OF PUBLICATIONS

A Conference paper entitled ‘Comprehensive Review of AI and Machine Learning Approaches for Prediction of Epigenetic Modifications in Autoimmune Disorders’ has been accepted for publication by Springer in “Information and Communication Technologies” series. It will be presented in the 2nd International Conference on Artificial Intelligence and Sustainable Computing 2025 (AISC 2025) on 24-26 July 2025.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS & ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
ANA	Antinuclear Antibody
ANN	Artificial Neural Network
β	Beta
CRP	C-Reactive Protein
CT	Computed Tomography
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
ESR	Erythrocyte Sedimentation Rate
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
ncRNA	Non-Coding Ribonucleic Acid
PCA	Principal Component Analysis
RA	Rheumatoid Arthritis
RF	Random Forest
RL	Reinforcement Learning
SLE	Systemic Lupus Erythematosus
SVM	Support Vector Machine

CHAPTER – 1

INTRODUCTION

Autoimmune disorders pose a significantly big health challenge across the world with approximately more than 10% of the global population, i.e. 1 out of every 10 people suffering from different autoimmune conditions. These are a set of conditions where the immune system starts to attack the body's own cells by mistake. This so-called ‘self-attack' can lead to damage and inflammation in various body parts or even systemically [1]. The etiology of autoimmune conditions is highly complex, and epigenetic modifications play a very significant role for the same.

Epigenetic modifications do not include any alterations to the underlying DNA sequence but are heritable changes which affect gene expression. Simply put, these epigenetic changes can cause dysregulation of the immune system leading to attack on body's own cells and tissues. Due to the reversible and dynamic properties of epigenetic alterations along with their sensitivity to environmental factors, it makes them indicators of disease status and its ability to interact with the environment [2]. This leads to epigenetic markers being used as potentially responsive biomarkers for disease prediction, progression, and therapy in complex diseases like autoimmune disorders.

In this era of technology, artificial intelligence (AI) and machine learning (ML) have revolutionized healthcare, thereby improving patient care and quality of life. Various ML algorithms help to uncover hidden patterns that may indicate early signs of autoimmune disorders. Since these disorders exhibit overlapping symptoms, ML models can be trained to identify specific epigenetic patterns and thereby aid in early and accurate diagnosis and prognosis. These models have predictive power which means that they can not only detect disorders but also help in predicting disease progression and future risks. Since epigenetic modifications are influenced by environmental and lifestyle factors, these ML techniques can help in providing personalized approaches to tackle autoimmune disorders.

CHAPTER – 2

LITERATURE REVIEW

2.1 Autoimmune Disorders

Autoimmune disorders represent a diverse group of conditions characterized by the body’s immune system targeting its own tissues and organs, leading to chronic inflammation and tissue damage [3]. In these conditions, the immune system loses its ability to distinguish between "self" and "non-self. They can range from relatively mild to life-threatening and can affect virtually any part of the body. There are more than 80 known autoimmune disorders which affect various parts of the body. Based on the organ getting affected we can classify it into organ-specific, which target a single organ or gland like Hashimoto's thyroiditis (thyroid), Graves' disease (thyroid), Type 1 diabetes (pancreas), or systemic, which affects multiple organ and tissues, like rheumatoid arthritis, Systemic lupus erythematosus (SLE) [4]. The etiology of autoimmune disorders is multifaceted, involving an interplay of genetic predisposition, environmental factors including infections and stress, & epigenetic modifications have emerged as crucial players [5].

2.2 Epigenetic Modifications

The heritable changes to the expression of a gene, which occur without any changes to the underlying DNA sequence refer to as epigenetic modifications. They influence the reading of a DNA sequence, which either turns the genes "on" or "off" or modulates their expression levels accordingly. Epigenetic changes are reversible, which makes them attractive targets for therapeutic and diagnostic intervention. Epigenetic modifications can generally be classified into DNA methylation, histone modification and non-coding RNAs which have been found to play a key role in the regulation of gene expression [6].

- DNA methylation

  In DNA methylation, a methyl group (CH3) is added to the 5th carbon atom of a cytosine base. This is typically observed in cytosine residues that are followed by a guanine residue, creating CpG dinucleotides.[7] They are not evenly distributed & tend to cluster in regions called CpG islands, which are generally found in the region where the promoters are present. Methylation of CpG sites or islands can lead to gene silencing & is carried out using enzymes called DNA methyltransferases (DNMTs).

- Histone modifications

  They include post-translational modifications of histone proteins, around which the DNA wraps, including acetylation, methylation, phosphorylation, ubiquitination, and sumoylation. Acetylation of histones and methylation of histones promote an open chromatin state (Euchromatin) that facilitates active transcription. Conversely, when chromatin gets condensed (heterochromatin), the repression of gene activity is seen [8]. The process of histone acetylation is mediated by histone acetyltransferases and reversibly by histone deacetylases, which alters gene expression.

- Non-coding RNAs (ncRNAs)

  They are RNA molecules which have regulatory functions but are not translated into proteins . They can be classified into long non-coding RNA's which are longer than 200 nucleotides with the ability to regulate gene expression and cellular differentiation, and small non-coding RNA's which are shorter than 200 nucleotides called MicroRNA's which can bind to mRNA and inhibit their translation. SiRNA's are also a part of the small category that play an important role in RNA interference which cause gene silencing [9].

2.3 Specific Autoimmune disorders & their Epigenetics

- **Rheumatoid Arthritis (RA)**

  It is a chronic inflammatory and autoimmune disease which primarily affects the synovial joints causing swelling and pain. The onset, severity and epigenetics of the condition is affected by environmental factors along with genetic predisposition. Extensive hypomethylation of DNA has been found particularly in synovial fibroblasts (RASF) and peripheral blood mononuclear cells (PBMCs) [10]. Genes expression involved in immune regulation such as TNF and IL-6. Histone modifications like acetylation and deacetylation impact pro-inflammatory cytokine regulation. miRNA's like miR-155 which is upregulated, also impacts both methylation and histone modification resulting in dysregulation of the immune cells [11].

- **Systemic Lupus Erythematosus (SLE)**

  It is a chronic autoimmune disorder which affects multiple organ systems including the skin, joints, kidneys, heart, and central nervous system. Increased production of autoantibodies which lead to inflammation and damage. Aberrant DNA methylation is a hallmark of SLE i.e. hypomethylation which causes over expression of certain genes coding for CD70 on B-cells, perforin, KIR etc. in patients with active lupus [12]. Abnormal methylation pattern causes overactivation of the type I interferon (IFN-α) system which is suspected to contribute to heightened interferon response [13]. Both histone acetylation by histone acetyltransferases (HATs) and methylation, in T cells and B cells causes overexpression of pro-inflammatory cytokines and autoantibodies. Altered miRNA expression like miR-146a has been implicated in regulating the NF-κB and regulate certain key immune genes. **miR-21** is overexpressed in SLE and several tumor suppressor genes and regulators of apoptosis are its targets [14].

- **Type 1-Diabetes**

  It is an autoimmune disorder wherein the person's own immune system starts to attack the beta cells in pancreas which produce insulin. Insulin helps in the regulation of blood sugar (glucose) and leads to hyperglycemia in the blood. Disruption of the DNA methylation patterns in the genes coding for B cells causes abnormality in their proper functioning and may lead to increased apoptosis of these cells [15]. Certain immune cells such as T cells and dendritic cells get affected by changes in DNA methylation patterns which leads to failure of immune tolerance. Genes like FOXP3 which regulates T Regulatory cells get impaired development due to the above. Histone acetylation and methylation of promoter regions of T cells affect the activation of loss of self-tolerance and activation of pro inflammatory genes. Regulation of inflammation generally is seen to involve certain miRNAs like miR-21 and miR-146a. Dysregulation of miR-375 might lead to beta-cell dysfunction or their apoptosis since this miRNA plays an important role in regulating beta-cell function [16].

- **Multiple Sclerosis (MS)**
  It is a chronic autoimmune disease which affects the central nervous system (CNS) including both brain & spinal cord. It occurs when the immune system starts attacking the myelin sheath by mistake which results in demyelination, inflammation and loss of nerve function eventually. Modifications in DNA methylation patterns in B and T cells are the main cause for the disease pathogenesis. Overexpression of interleukin-2 (IL-2) causes the destruction of myelin, genes coding for Th17 cells contribute to inflammation and tissue damage, genes which are involved in immune tolerance like FOXP3 are also aberrantly methylated [17]. Abnormal histone acetylation affects T-cells and microglia which leads to increased expression of pro inflammatory cytokines. Activation of pro-inflammatory pathways due to increased levels of miR-155 and reduced expression of miR-146a (helps in inflammation decrease) is observed in MS patients [18].

2.4 Current detection methods for autoimmune disorders

A combination of methods is used in the detection of autoimmune disorders since no one test can diagnose all the conditions. Initially a detailed medical history is enquired by the doctor which includes appearance of signs and symptoms, disease progression, family history etc. A physical examination is done only if any physical clinical manifestations like signs of inflammation are present.

Laboratory-Based Methods

A variety of laboratory tests are done to generate evidence of any inflammation, dysfunction or production of autoantibody in the body.

1. **General Inflammatory Markers:** Indication of systemic inflammation by non-specific markers is done but they do not confirm an underlying autoimmune condition. For monitoring the disease activity and deciding the treatment response is aided by these markers.

    a. **Erythrocyte Sedimentation Rate or ESR:** It measures the rate at which red blood cells settle in a test tube. Inflammation is indicated by Elevated ESR, but it can also be affected by numerous factors.

    b. **C-Reactive Protein or CRP:** It is a reactant protein (acute-phase) which is produced by the liver in response to inflammation. CRP levels are more potent markers as they rise and fall more rapidly than ESR, making it a highly sensitive marker to detect systemic or acute inflammation [19].

2. **Autoantibody Detection:** To detect the presence of autoantibodies is a common practice for the diagnosis of autoimmune disorders. Autoantibodies are antibodies which are against the body's own proteins, cells, or tissues and are a hallmark of the disease.

    a. **Antinuclear Antibody or ANA Test:** This is a screening test for detection of antinuclear antibodies in the blood, which are autoantibodies that target substances found in the nucleus of one's own cells. This could be a sign of an autoimmune disorder. A positive ANA is usually obtained for Systemic Lupus Erythematosus (SLE), Sjögren's syndrome, Rheumatoid arthritis, Scleroderma etc. Many healthy people can test falsely positive for this test due to low titers, increased antinuclear antibodies with increasing age or chronic inflammatory diseases [20].

b. **Specific Autoantibody Tests:** More targeted tests are performed after a positive ANA or strong clinical symptoms which imply a specific disease. These may include:

   i. **Anti-double-stranded DNA or anti-dsDNA antibodies Detection:** Highly specific for SLE [21]

   ii. **Anti-Smith (anti-Sm) antibodies detection:** Highly specific but less sensitive for SLE.

c. **Detection of Rheumatoid Factor:** An autoantibody against the Fc region of IgG. It is present in a high percentage of Rheumatoid Arthritis (RA) patients and is an autoantibody against the Fc region of IgG. It is not a specific marker for detection [22].

d. **Detection of Anti-cyclic Citrullinated Peptide antibodies:** Anti CCP antibodies are highly specific for RA and appear before the occurence of clinical symptoms. They can be very helpful in disease diagnosis and prognosis [23].

e. **Detection of Thyroid Antibodies:** Anti-thyroglobulin and anti-thyroid peroxidase (anti-TPO) antibodies are key markers for autoimmune thyroid diseases like Hashimoto's thyroiditis and Graves' disease [24].

Imaging Techniques

Visualization of any inflammation or structural damage can be used in the detection of autoimmune disorders. Bone and joint erosions, common in conditions like RA can be detected using X-rays. MRI or Magnetic Resonance Imaging can provide soft tissue views, which can be used to detect and identify lesions in the brain or spinal cord in cases of multiple sclerosis. Computed Tomography or CT scans provide cross-sectional information about organ damage, which may occur after a long prognosis. Ultrasounds also act as a useful tool and are used in assessing joint synovitis, or any glandular changes seen in cases of Sjögren's Syndrome. These methods act as an additive diagnostic method which can be coupled with other lab-based techniques to provide collectively visual evidence in the monitoring of disease progression [25].

Histopathology (Biopsy)

Checking histopathology, via tissue biopsy, where direct microscopic examination of tissues affected is performed gives detailed diagnostic information. This method is crucial for identifying characteristic inflammatory patterns, immune cell infiltrates, and the deposition of immunoglobulins or complement components that confirm an autoimmune process [26].

2.5 Limitations of using conventional techniques

Though conventional methods for detection of autoimmune disorders are used increasingly, but they possess several limitations:

1. Lack of Specificity: Since common markers, such as ESR, CRP, or even ANA, are non-specific, they indicate only general inflammation or immune activation but do not definitively point to a specific autoimmune disease. This leads to ambiguity in diagnosis and potential false positives [27].

2. Late Detection: After the manifestation of clinical symptoms and a significant dysregulation of the body's immune system, only then will these tests show positive results for a particular autoimmune disorder. For early and effective intervention along with appropriate preclinical prediction, these tests then prove to be unreliable. This also delays diagnosis and complicates the disease prognosis [28].

3. Variable Sensitivity: The use of different autoantibodies in diagnosis of various autoimmune conditions can vary significantly. Certain autoantibodies are not specific but have high sensitivity like in the ANA test, while some others like anti-Sm used in SLE are highly specific but are only present in a few patients and majorly absent in others.

4. Snapshot View: All the conventional, lab-based tests only offer a static overview of the immune system status of the patient at a specific moment in

time. As a result, it is unable to detect or capture the highly dynamic nature of autoimmune diseases. The failure to detect small shifts in the body which precedes the illness is also a big challenge posed by these methods [29].

5. Overlapping Symptoms: Due to the overlapping of symptoms between various autoimmune disorders, the process of clinical differentiation becomes very difficult and challenging. Overlapping with non-autoimmune conditions is also frequently observed, which makes the process of diagnosis and detection a very tedious task [30].

6. Limited Mechanistic Insight: These tests often present no or limited insight into the molecular and cellular mechanisms behind the autoimmune processes. They majorly just inform about any immune system dysregulation and make the process of development of targeted therapy difficult to achieve.

2.6 Significance of epigenetic studies in prediction of autoimmune disorders

Epigenetics play a significant role in predicting autoimmune disorders even before the occurrence of clinical symptoms. It is a highly evolving and advancing area of research, even though it is not very prevalent as standard clinical practice. The following reasons can be enumerated by highlighting its significance:

1. Preclinical Detection: Conventional techniques identify markers like autoantibodies, while the epigenetic changes occur years before the onset of clinical symptoms of a particular autoimmune disorder. These changes are the earliest molecular footprints of the dysregulation of the body’s immune system and are better for early prediction. They can also be referred to as "pre-symptomatic" signals, and medical intervention can be taken before any irreversible damage occurs [31].

2. Bridging Genetics and Environment: Epigenetic changes are affected by environmental factors such as stress, diet, toxins, smoking, pollution and can bridge the gap between genetic predisposition and environmental triggers. Genetics, however, cannot justify why some people develop these conditions and while others don't, even among genetically identical twins. These epigenetic changes can lead to aberrant regulation of immune-related genes.

3. Higher Specificity: Conventional markers are non-specific and lead to ambiguity and false positives. In epigenetic studies we can identify highly specific signatures like DNA methylation patterns which are unique to a particular autoimmune disorder [32].

4. Better Therapeutic Targets: Epigenetic modifications help us to get deep insights into the gene regulatory pathways which are affected in case of autoimmune conditions unlike the conventional methods. This information plays an essential role in the development of novel targeted therapies which are used to correct or reverse the condition rather than just suppressing the symptoms [33].

2.7 Artificial Intelligence and Machine Learning in Data Analysis and Prediction

Artificial Intelligence or AI is the computer's ability to emulate or copy the human thought process and use it for real world scenarios along with surpassing the aspects of human intelligence. It simulates and augments the human intelligence capabilities to achieve more. Machine learning (ML) is a subset of AI, which uses algorithms to learn from data and identify intricate patterns to make necessary predictions. All the above capabilities of AI and ML are vital in the field of biology to analyze and learn from highly complex datasets which are not in the capacity of traditional computational methods. Some of the ML techniques include (in Fig 1):

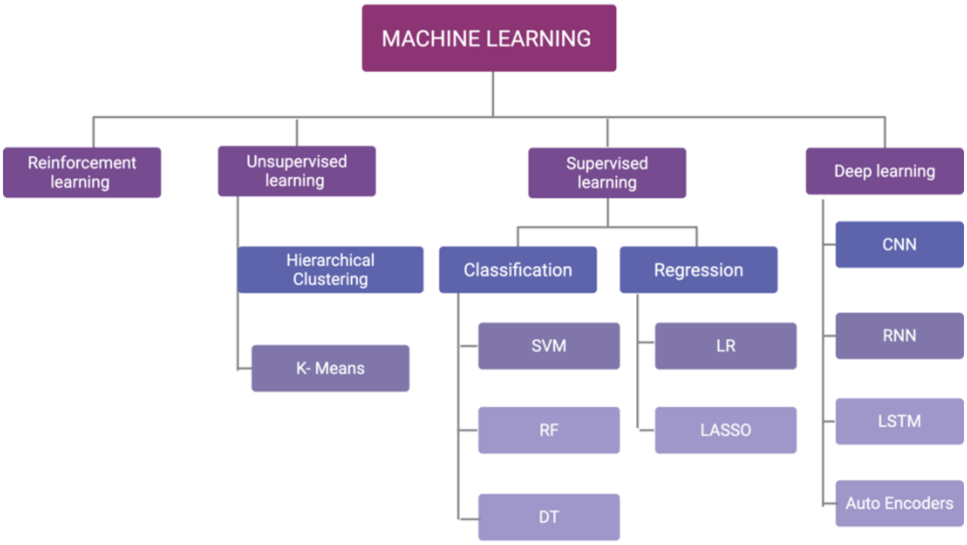


Fig.1. Types of Machine Learning techniques

- Supervised learning is a type of ML where the input is labelled. It uses a training data set and is generally used to predict outcomes. It is classified into two types when data mining ie classification and regression. When classifying the data, these techniques help in predicting a category or a class label which includes SVM (Support Vector Machines), RF (Random forests), decision trees etc. [34]. While on the other hand, Regression algorithms focus on predicting a continuous numerical value like LR (linear or logistic regression), LASSO (least absolute shrinkage and selection operator regression) polynomial regression [35].

- Unsupervised learning does not need the input data to be labelled and is generally used for the analysis function. It can help in tasks of clustering and association along with dimensionality reduction, helping to find correlation between individual variables/classes within a data set [34]. Clustering Algorithms (e.g., K-means, Hierarchical Clustering) are employed for grouping unlabeled data based on their similarities or differences. These methods help to locate regions of similar methylation patterns or histone modifications. Association is another type of data mining technique which uses a certain set of rules to find correlation between classes in a data set. Dimensionality reduction is a learning technique that reduces the amount of complex data points and helps in reducing it to a manageable size. It is done for large scale epigenetic data to identify significant modifications. PCA (Principal Component Analysis) & Partial least squares discriminant analysis can be used to achieve the same [35].

- Deep learning or Deep neural network is a subset of machine learning that uses artificial neural networks which are inspired by the function of the human brain with multiple layers to learn from data. It is a constantly evolving field and consists of various techniques such as CNN (convolutional neural network), RNN (recurrent neural network), LSTM (long short-term memory), Autoencoders, etc. [34]. CNNs are used to analyze spatial relationships ie to identify patterns in DNA sequences specifically regulatory elements. While RNN's are used for sequential data to predict modification over a period of time. LSTMs selectively remember or forget information over long sequences and help in identifying dynamics of epigenetic changes [36]. Autoencoders help in the compressed representation of data to predict patterns in large epigenetic data.
- MLP (Multi-layer Perceptrons) is a simpler form of deep learning model which can perform complex learning tasks to predict non-linear relationships. It is a feedforward type of neural network where information moves in a single direction and doesn't loop back. It could be applied to predict the presence and relationship between different epigenetic modifications.
- GBM (Gradient Boosting Machines) can utilize highly complex datasets to predict outcomes such as disease onset and prognosis. Techniques like XGBoost and LightGBM are widely used where combination of predictions of many weak models like decision trees is done to make predictions [36].
- RL (Reinforcement Learning) is also a learning technique in which models learn from environmental feedback and make decisions to optimize an output.

The model learns through trial and error, i.e. by receiving feedback. It is an emerging field and can be used to analyze the genome for predicting epigenetic modifications by receiving feedback [35].

- TL (Transfer Learning) is also being used highly since it can transfer knowledge gained while solving one problem to a different but related problem. This is used when limited labelled data is available for a certain task. It can train large epigenetic datasets and turn them into disease specific datasets where limited data for epigenetic modifications maybe present.
- Graph Neural Networks (GNNs) are a type of neural network that works with data structured in graphs, making them quite effective in establishing relationships between different biological entities such as genomic regions. These networks can be used to predict how changes in epigenetic marks at a certain region might affect distant genomic regions or genes [37].

2.8 AI and ML for Autoimmune disease prediction

Artificial intelligence (Al) and machine learning (ML) have emerged as significant contributors to advancing research, in the domain of prediction of autoimmune disorders. These technologies not only play a crucial role in identifying and interpreting complex epigenetic modifications but also help to indicate disease onset or progression. The application of AI and ML in epigenetic research is immensely important in deciphering complex biological datasets, so as to extract significant patterns pertinent to autoimmune disorders.

AI can help in decoding the epigenetic variability in autoimmune disorders by adopting the following ways:

- **Pattern Recognition:** Epigenetic modifications such as DNA methylation or histone modifications are highly complex and vary in intensity. Computer algorithms can recognize intricate patterns within this data and help in defining the presence of the autoimmune disorder in a family.
- **Prediction of Epigenetic Markers:** Machine learning algorithmic models can be trained in order to predict the presence or absence of specific epigenetic markers and

generalize them for populations. They can be a certain DNA sequence, transcription factor binding sites, etc.

- **Integration of Multi-Omic Data:** After different genomic, transcriptomic, and proteomic data is obtained after experimentation, AI can be used to integrate them all to obtain a better comprehensive study of epigenetics. This can further help in making disease prediction and identifying potential drug targets.

2.9 Applications of ML in prediction of Autoimmune disorders using epigenetic modifications

2.9.1 Applications of ML in Rheumatoid Arthritis

Diagnosis of RA diagnosis by the identification of epigenetic changes can be challenging. Certain DNA methylation patterns, at CpG sites, can serve as biomarkers where blood samples from patients suffering from RA, Osteoarthritis (OA) and Healthy Controls (HC) undergo targeted DNA methylation sequencing and some steps like in Fig. 2. The study identified 16 CpG sites after application of 6 ML learning techniques including Logistic Regression (logit model), RF, SVM, Naive Bayes, AdaBoost and Learning Vector Quantization (LVQ). Here a 10-fold cross-validation was applied on the data. Identification of SMAD3 which distinguished patients with RA from the HC, with an AUC value of 0.95 [38].

Exosome-derived microRNAs (exomiRNAs) can also act as good candidates for early diagnosis of autoimmune diseases. Deregulated exomiRNAs were used to distinguish between patients with early RA and HC sTWEAK (Synovial fluid tumor necrosis factor-like weak inducer of apoptosis) and exomiR-451a are signature biomarkers which were statistical analyzed. Prediction of exomiRNA gene-targets interactions was done using miRNet software. For analysis, Statistical software SPSS Statistics 24.0 package and R software were used where a p-value < 0.05 was considered significant. Partial Least Square Discriminant Analysis (PLS- DA), variable importance in projection (VIP) analysis and logistic regression analysis were applied on the data. An AUC of 0.983, 100 % specificity, and 85.7 % sensitivity was obtained [39].

2.9.2 Applications of ML in Systemic Lupus Erythematosus

Increased hypomethylation is in SLE patients and can be used as a prominent epigenetic tool in the early diagnosis of the disease. A study was done to systematically compare epigenome-wide DNA methylation changes along with detection of certain disease-specific alterations where patients with SLE and healthy controls were taken. Linear regression models with Bonferroni correction were applied to find differentially methylated CpG sites (DMCs). Lower methylation levels at many sites, in the genes of Type I IFN pathway which could prove to be an essential biomarker. Genes where SLE-specific differential DNA methylation was identified were FADD and CASP-8 in the NFκB activation pathway used the Functional gene ontology analysis. After the application of Random Forest classifier, the reported AUC value was 0.96 and showed a significant potential [40]. Abnormal IFI44L promoter methylation showed 88.5% sensitivity and 97.1% specificity for SLE detection can also be used in ML model training in the prediction of the disease using epigenetic markers [41].

Since ncRNAs also are key epigenetic markers, they can act as non-invasive biomarkers which can reflect disease condition in lupus nephritis. Differentially expressed lncRNAs in both inactive HC and active patients were analysed. 6 lncRNAs associated with LN flares with False Discovery Rate < 0.05 were found out of which NRIR and KLHDC7B-DT appeared to be key regulators involved in IFN-related pathways using WGCNA- Weighted Gene Co- expression Network Analysis. This uses Pearson correlation along with clustering algorithms such as hierarchical clustering. Since the above ncRNAs are linked with key pathways, we can use them in epigenetic prediction after combining them further with ML models to train data sets and develop algorithms which can take testing data sets and provide necessary outcomes [42].

2.9.3 Applications of ML in Type 1 Diabetes

DNA methylation patterns are the most common epigenetic changes identified which can be used in detecting T1D risk and progression. A predictive model which discriminates HC with the diseased, based on methyl- haplotypes within the Insulin Gene Promoter (IGP) region using ML techniques. The methylation status was identified using targeted deep bisulfite sequencing on 10 CpG sites. Principal Component Analysis (PCA) was done to Dimensionally reduce the data. The training data set included 3 principal components which were applied with 5-fold cross-validation to five different supervised machine learning classifiers which include RF, DT, SVM, Logistic Regression and Naive Bayes. The best results were achieved in Naive Bayes which gave an accuracy of 0.90 and an AUC of 0.96. The Random Forest model also performed well with an accuracy of 0.87 and an AUC of 0.93 [43].

Since altered lncRNA expression patterns are often seen in T1D, prediction of diagnosis on the basis of a lncRNA- based signature with the help of ML models can prove highly effective. After data collection, lncRNA's expression levels are mapped and preprocessed. Differential expressions were found using Bonferroni correction between the T1D patients and HC where $p < 0.01$. Support Vector Machine model with a sigmoid kernel was employed along with a Random Forest algorithm, which ranked the importance of different lncRNAs. The result after a 3-fold cross validation were 26 specific lncRNAs (26LncSigT1DM) which were all down regulated in T1D patients showed high AUC values of 0.9973, 0.9641, 0.9556 which implied excellent performance. Also, the diagnostic ability of this lncRNAs signature was validated with an AUC value of 0.825 in an independent patient SVM model [44].

2.9.4 Applications of ML in Multiple Sclerosis

Prediction in DNA methylation patterns can also help in investigating disease severity and also serve as biomarkers. After the epigenome wide association study, differentially Methylated Positions (DMPs) were identified 1708 correlated CpG sites selected with a False Discovery Rate (FDR) of 0.05, for modeling between patients with mild MS and severe MS. Three different models were compared, where the first two models used Elastic Net (EN) regression with the only difference in input. The third model was developed using weighted methylation risk score (wMRS). The second model with methylation data as input got an AUC of 0.91 and implied the correlation between the DNA methylation and MS severity by outperforming models based on clinical data alone (AUC=0.74) or a methylation risk score (AUC=0.77) [45].

MS prediction using serum Exosome MicroRNAs as epigenetic markers which are non-invasive, can be done by applying ML algorithms. These miRNAs are identified between active disease versus patients with quiescent disease after 6 months treatment with fingolimod and follow the steps as in Fig. 3. Empirical Bayes method was employed along with a p-value < 0.05 which identified 15 distinct miRNAs. In Univariate predictive modeling, Logistic Regression was used where each miRNA was taken individually. A modest overall power of individual miRNAs was found with an average accuracy of 77%. Multivariate Predictive Modeling used Random Forest with combinations of miRNAs which resulted in better predictive performance giving an accuracy of 92% [46].

CHAPTER – 3

METHODOLOGY

In this study we aim to develop a ML model for the prediction of Systemic Lupus Erythematosus (SLE) using epigenetic data. The proposed methodology of the experiment includes a number of steps like data collection, preprocessing, data splitting, feature selection, model selection and finally model evaluation and its validation as depicted in fig.2. All the steps in the process of analyses were conducted using Python along with its associated scientific computing libraries.

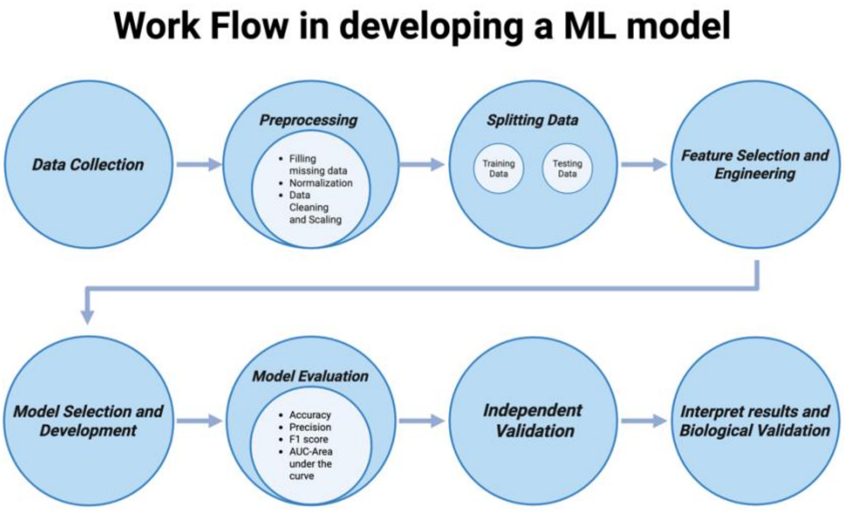
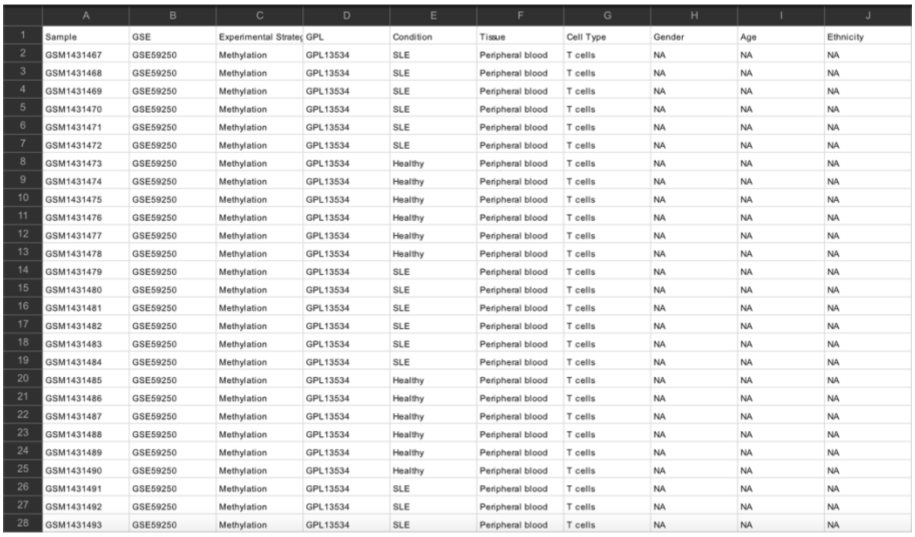


Fig 2. Workflow in ML Model Development

3.1 Step 1 – Data Collection

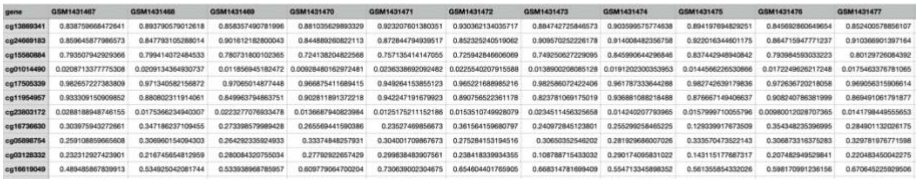
The process of developing a machine learning model starts with the first and most crucial step of acquiring an appropriate dataset. In this case, since epigenetic modifications are the basis on which the prediction is made, gene methylation data was obtained from ADEx– Autoimmune Disease Explorer which is a "comprehensive database for integrated analysis of omics data in autoimmune diseases." The data obtained corresponds to the Gene Expression Omnibus (GEO) dataset GSE59250 which includes DNA methylation data. This methylation profile was obtained after using the platform- Illumina HumanMethylation450 Bead Chip which gave β values for individual CpG sites. These values show the proportion of methylation at a particular CpG site in the given sample and ranges from 0 being unmethylated to 1 being fully methylated.

The dataset contained 149 samples which were obtained from isolating T cells from peripheral blood of both healthy and patients suffering from SLE states as in fig.3 and showed β values of 421947 CpG sites for each individual as in fig.4.



	A	B	C	D	E	F	G	H	I	J
1	Sample	GSE	Experimental Strateg	GPL	Condition	Tissue	Cell Type	Gender	Age	Ethnicity
2	GSM1431467	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
3	GSM1431468	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
4	GSM1431469	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
5	GSM1431470	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
6	GSM1431471	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
7	GSM1431472	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
8	GSM1431473	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
9	GSM1431474	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
10	GSM1431475	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
11	GSM1431476	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
12	GSM1431477	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
13	GSM1431478	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
14	GSM1431479	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
15	GSM1431480	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
16	GSM1431481	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
17	GSM1431482	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
18	GSM1431483	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
19	GSM1431484	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
20	GSM1431485	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
21	GSM1431486	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
22	GSM1431487	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
23	GSM1431488	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
24	GSM1431489	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
25	GSM1431490	GSE59250	Methylation	GPL13534	Healthy	Peripheral blood	T cells	NA	NA	NA
26	GSM1431491	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
27	GSM1431492	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA
28	GSM1431493	GSE59250	Methylation	GPL13534	SLE	Peripheral blood	T cells	NA	NA	NA

Fig.3 Metadata with SLE vs Healthy state



gene	GSM1431467	GSM1431468	GSM1431469	GSM1431470	GSM1431471	GSM1431472	GSM1431473	GSM1431474	GSM1431475	GSM1431476	GSM1431477
cg13869341	0.838759668472641	0.893790579012618	0.858357490781996	0.881035629893329	0.923207601380351	0.930362134035717	0.884742725846573	0.903599575774638	0.894197694829251	0.845692860649654	0.852400578856107
cg24669183	0.859645877986573	0.847793105288014	0.901612182800043	0.844889260822113	0.872844794939517	0.852325240519062	0.909570252226178	0.914008482356758	0.922016344601175	0.864715947771237	0.910366901397164
cg15560884	0.793507942929366	0.799414072484533	0.780731800102365	0.724138204822568	0.757135414147055	0.725942846606069	0.749250627229095	0.845990644296846	0.837442948940842	0.793984593033223	0.80129726084392
cg01014490	0.0208713377775308	0.0209134364930737	0.011856945182472	0.00928480162972481	0.0236338692092482	0.0225540207915588	0.0138900208085128	0.0191202300353953	0.0144566226530866	0.0172249626217248	0.0175463376781065
cg17505339	0.982657227383809	0.971340582156872	0.97065014877448	0.966875411689415	0.949264153855123	0.965221688985216	0.982586072422406	0.961787333644288	0.982742639179836	0.972636720218058	0.969056315906614
cg11954957	0.933309150909852	0.880802311914061	0.849963794863751	0.902811891372218	0.942247191679923	0.890756522361178	0.823781069175019	0.936881088218488	0.876667149406637	0.908240786381999	0.869491061791877
cg23803172	0.0288188948746155	0.0175366234940307	0.0223277076933478	0.0136687940823984	0.0125175211152186	0.0153510749928079	0.0234511456325658	0.0142402077793965	0.0157999710055796	0.00980012028707365	0.0141798449555653
cg16736630	0.303975943272661	0.347186237109455	0.273398579989428	0.265569441590386	0.23527469856673	0.361564159680797	0.240972845123801	0.255299258465225	0.129339917673509	0.354348235396995	0.284901132026175
cg05898754	0.259108859665608	0.306960154094303	0.264292335924933	0.33374848257931	0.304001709867673	0.275284153194516	0.30650352546202	0.281929686007026	0.333570473522143	0.306873316375283	0.329781976771598
cg03128332	0.232312927423901	0.216745654812959	0.280084320755034	0.27792922657429	0.299838483907561	0.238418339934355	0.108788715433032	0.290174095831022	0.143115177687317	0.207482949529841	0.220483450042275
cg16619049	0.489485867839913	0.534925042081744	0.533938968785957	0.609779064700204	0.730639002304675	0.654604401765905	0.668314781699409	0.554713345898352	0.561355854332026	0.598170991236156	0.670645225929506
cg05475702	0.80217048920165	0.740185645459092	0.764821534894956	0.802395539488532	0.796043278803888	0.810940625274214	0.732722833926127	0.807755519641047	0.782321489761917	0.77002326466623	0.779664228934948
cg18147296	0.895761419161372	0.907884533297327	0.933795164871839	0.918465188878975	0.904964251808066	0.885449702589235	0.87446223749512	0.91943398875097	0.903178841140491	0.880162694341109	0.883254478892216
cg05597748	0.880404070858335	0.882535845400374	0.884144558142797	0.901019959210741	0.894133864633528	0.891263513797397	0.889450135956366	0.902550123252552	0.906393183339009	0.864799817076173	0.86499401603378
cg00111697	0.820192681301445	0.82470986608116	0.852373490790504	0.818965087792674	0.82028178004298	0.821252026849077	0.817377668858998	0.842700763691248	0.817148941844752	0.79528527444515	0.816869175528057
cg10988531	0.83119580665861	0.847416995690179	0.901571799339495	0.895528112963446	0.886107566198428	0.809106693863545	0.798346307033871	0.901710652802469	0.8540090003046	0.822910733797515	0.866770313783136
cg13938959	0.750033504428873	0.72539914929176	0.66040112306183	0.73053889180365	0.695461715209297	0.771741215953186	0.670343990594099	0.69441086651672	0.535387431954709	0.669265086918009	0.753605036716377
cg12445832	0.515898974409922	0.582773338768381	0.436985480189641	0.558864095007027	0.467514282441358	0.613759503957889	0.46914899804184	0.485028367365351	0.38116854866749	0.503448252669274	0.634061240116671
cg23999112	0.690901838832305	0.798477404486874	0.721559451122566	0.743659803419369	0.730413840316156	0.819761628676908	0.717329630563467	0.682595519090885	0.620604618690251	0.758552732283426	0.79629302125226
cg11527153	0.943936503508654	0.92621564406981	0.93705894850855	0.952223433093231	0.909591169291608	0.910730194288359	0.953148811441482	0.919629705556277	0.910646871063883	0.923375530969697	0.940214909225522
cg27573606	0.97165206300099	0.97543656629661	NaN	0.967256197428503	0.972544508629008	0.969700177381606	0.976543587497296	0.976950149150902	NaN	0.979458331098767	0.969104796881662
cg04195702	0.92677777605941	0.893044626035154	0.921129791020995	0.911277389096405	0.931760417017634	0.873987016621241	0.913848267407509	0.902848298373999	0.863538362682342	0.916108214871582	0.88078529998044
cg08128007	0.903341406831801	0.878184990595867	0.927895854385252	0.900805829983252	0.917520009956654	0.965258963448012	0.932876415977277	0.883820421859889	0.951762503422071	0.941494697611942	0.916735245596202
cg23733394	0.408655022540767	0.604996863889471	0.260318888331158	0.648423105396377	0.41193740010286	0.763992441388846	0.397478506438148	0.286407498585052	0.826960152591647	0.646374026236937	0.653893783174091
cg13371836	0.0266044691789052	0.0188059941869083	0.0218192790750085	0.0226566305085057	0.0261688810882477	0.0469072854419745	0.0117145839551749	0.0251627746611899	0.0277310360203031	0.019974287919362	0.0195965513096548
cg04407431	0.00375760569215433	0.00544728686701768	0.00603686295044704	0.00314923399036255	0.00544166480418874	0.0046051542100991	0.00395049853590839	0.0106065113744594	0.00700020196510009	0.00421673525141516	0.00493015174751739
cg13565718	0.953847040230835	0.960228924280835	0.939465328272536	0.966668888601173	0.969625284444677	0.975547411220081	0.941809153973126	0.965207139906473	0.953983045925462	0.937448086837448	0.960673435401767
cg01097950	0.814545325598355	0.778274378691689	0.840828073905128	0.812560065310449	0.832376203039802	0.7701517508998	0.798191333376061	0.789770657290571	0.788567687682938	0.781240712911403	0.747949595832009
cg08900511	0.905304528886591	0.869948833479318	0.896989578936853	0.924694789936628	0.93143638032564	0.935887000191386	0.88883770845454	0.878854190687036	0.909521401143531	0.903543050442456	0.884592086019572
cg04410003	0.957935725230393	0.957937131724662	0.940028066690754	0.960591677085697	0.946370180196783	0.956203609410147	0.961137478681877	0.956470329266039	0.943040713408734	0.938708582219115	0.927994834130909
cg18762746	0.972155981891043	0.961825413885969	0.955967242559891	0.971550592224032	0.979242207059076	0.981809967769861	0.974407886998108	0.967363357303105	0.954615206116049	0.958321402387819	0.968845790452907
cg12726839	0.543029470111481	0.4275561371661	0.283822763324988	0.481776638536462	0.359744835736808	0.445213926641116	0.441436374403639	0.467132856902195	0.58239905281289	0.411832995730898	0.452124556078922
cg04364728	0.31541066366436	0.306051542485413	0.35908848781586	0.542994025625694	0.258130453695348	0.384394258143164	0.286693573146183	0.409045671483327	0.292204654624181	0.243775787682255	0.407430471580246
cg05151709	0.57655186368238	0.585912700319486	0.645340561151418	0.686350711949545	0.583663216253835	0.668511284240936	0.500594895154294	0.588333433943132	0.504531731812753	0.605372802839537	0.614443354858248
cg08546707	0.729120984680011	0.709306699075624	0.762076878724914	0.797418299971835	0.723747798965672	0.834287037940869	0.745927311823935	0.731254401373808	0.694288661857222	0.736694324856791	0.777642868771769
cg06154531	0.917452179554156	0.875822409163669	0.882169078943511	0.907469304370081	0.912350430748082	0.89727697676256	0.906207513110899	0.914155296121834	0.884809447734022	0.927127472451912	0.883393010345064
cg01782097	0.447271714313731	0.435135700054149	0.49967883853237	0.451081935473143	0.488682561221844	0.436594307548129	0.493094208617303	0.509921161503643	0.418688913077939	0.465529951315169	0.403368699653145
cg20825023	0.441191253602469	0.405253907409159	0.459190478463299	0.360532057852407	0.443998707220621	0.386813940539652	0.379344610058416	0.415133173461953	0.367076592739515	0.398119860095398	0.350461650396659

Fig.4 β values of CpG sites

3.2 Step 2 – Data Preprocessing

This is a critical step in the development of a machine learning pipeline and prepares the raw data into a usable format by transforming, organizing and cleaning the data into suitable formats. It is done to handle missing values, normalize the data, or remove any duplicates if present, which ultimately is important to improve the quality of the dataset obtained. Data preprocessing is generally done with complex and high-dimensional biological data like DNA methylation in this case where it transforms raw methylation array signals into more normalized and clean forms so that used for further down streaming steps. It includes different steps as in Table 1.

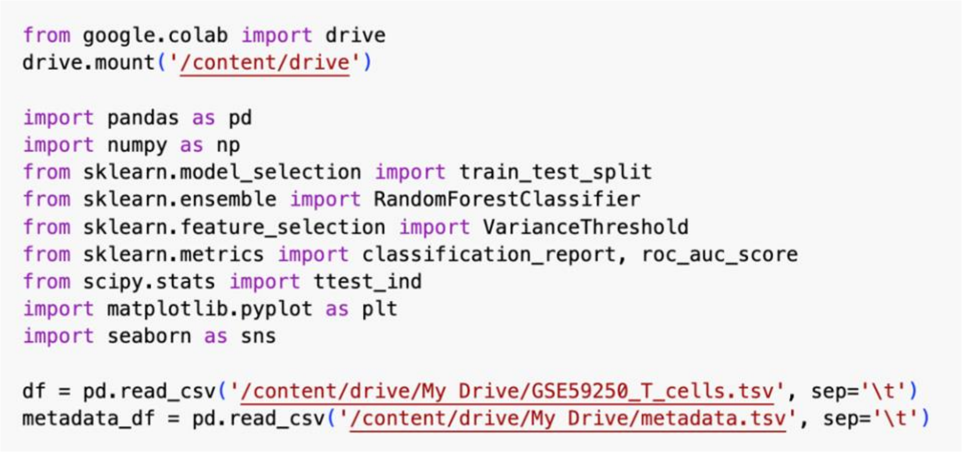
Table 1. Different steps in Data Preprocessing

Category	**Techniques**
Data Cleaning	Handling Missing Values, Handling Noisy Data and Outliers, Removing Duplicates
Data Transformation	Normalization, Scaling, Feature Engineering, Categorical Encoding, Log Transformation
Data Reduction	Dimensionality Reduction, Feature Selection, Feature Extraction, Data Compression, Aggregation
Data Integration	Combination of data from different sources

In this code, many preprocessing steps were performed in the process of data analysis.

1. **Loading Data**:

   Here, both the data files i.e. the main data fig. 4 (GSE59250_T_cells.tsv) and metadata fig.3 (metadata.tsv) were loaded into pandas Data Frames.

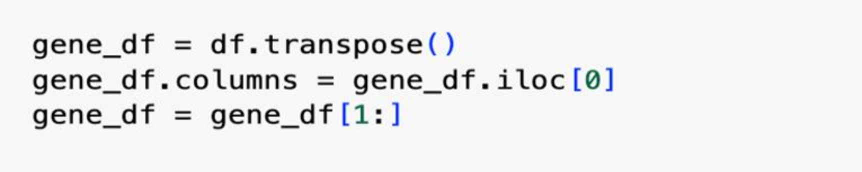


```
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import VarianceThreshold
from sklearn.metrics import classification_report, roc_auc_score
from scipy.stats import ttest_ind
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('/content/drive/My Drive/GSE59250_T_cells.tsv', sep='\t')
metadata_df = pd.read_csv('/content/drive/My Drive/metadata.tsv', sep='\t')
```

Fig.5 Loading data and data frames

2. **Transposing Data**:

   This step was taken to reformat the dataset, by altering the orientation of the table. The main data DataFrame (df) was transposed to make the samples into rows and features like CpG β values into columns and obtained the new (gene_df). This was followed by removing the redundant header after the first row was set as column header. The result obtained showed the - Shape of transposed beta data with samples as rows: (149, 421947).

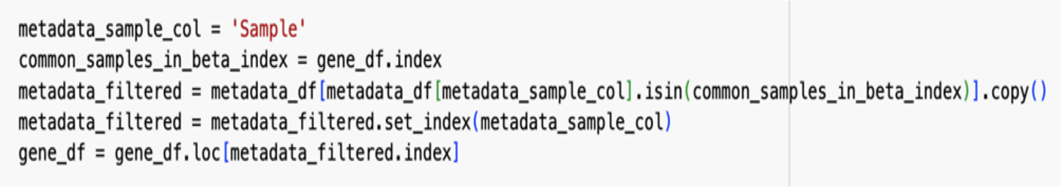


```
gene_df = df.transpose()
gene_df.columns = gene_df.iloc[0]
gene_df = gene_df[1:]
```

Fig.6 Data Transposition

3. **Aligning the sample order**:

   We identified the column which contained the sample IDs in the filtered metadata which was obtained from the transposed Data Frame (gene_df). The step of realignment was performed on the transposed Data Frame (gene_df) to match the sample order of the filtered metadata Data Frame (metadata_filtered). This is a highly crucial step as it ensures the samples are in the exact order with the gene expression data.

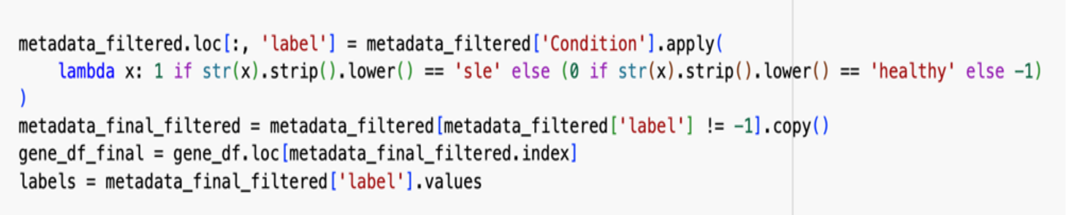


```python
metadata_sample_col = 'Sample'
common_samples_in_beta_index = gene_df.index
metadata_filtered = metadata_df[metadata_df[metadata_sample_col].isin(common_samples_in_beta_index)].copy()
metadata_filtered = metadata_filtered.set_index(metadata_sample_col)
gene_df = gene_df.loc[metadata_filtered.index]
```

Fig.7 Data Alignment

4. **Labeling and Filtering Samples**:

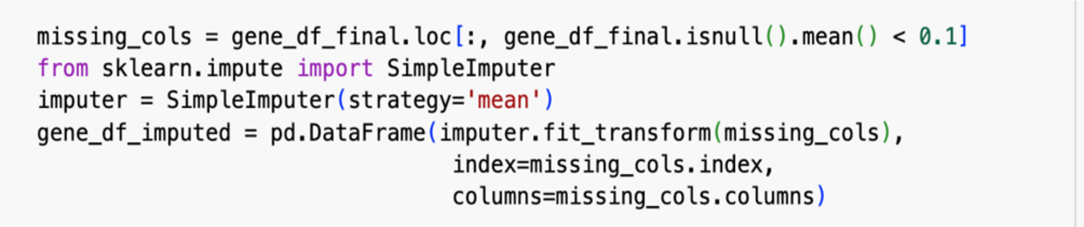
   A new label column was created in the filtered metadata to map the patient conditions between SLE and healthy to numeric forms - 1 for 'SLE' and 0 for 'Healthy'. Thereafter, the samples were filtered with a (-1) label for any other condition. This step was done to define the binary classification problem. The realignment step was performed once again thereby extracting the final labels in metadata_final_filtered (gene_df_final). Shape of β data after filtering for SLE/Healthy: (149, 421947)



```python
metadata_filtered.loc[:, 'label'] = metadata_filtered['Condition'].apply(
    lambda x: 1 if str(x).strip().lower() == 'sle' else (0 if str(x).strip().lower() == 'healthy' else -1)
)
metadata_final_filtered = metadata_filtered[metadata_filtered['label'] != -1].copy()
gene_df_final = gene_df.loc[metadata_final_filtered.index]
labels = metadata_final_filtered['label'].values
```

Fig.8 Data Labelling

5. **Handling Missing Values**: Features were dropped from the final filtered data which had more than 10% of missing values (user-defined threshold). This was done to simplify the dataset as features with a high percentage of missing data can cause problems since they are uninformative. The remaining missing values after the 10% dropping were filled using the mean of each feature. This is done since the ML algorithms cannot handle NaN values. Shape of β data after dropping features with >10% missing: (149, 414402).



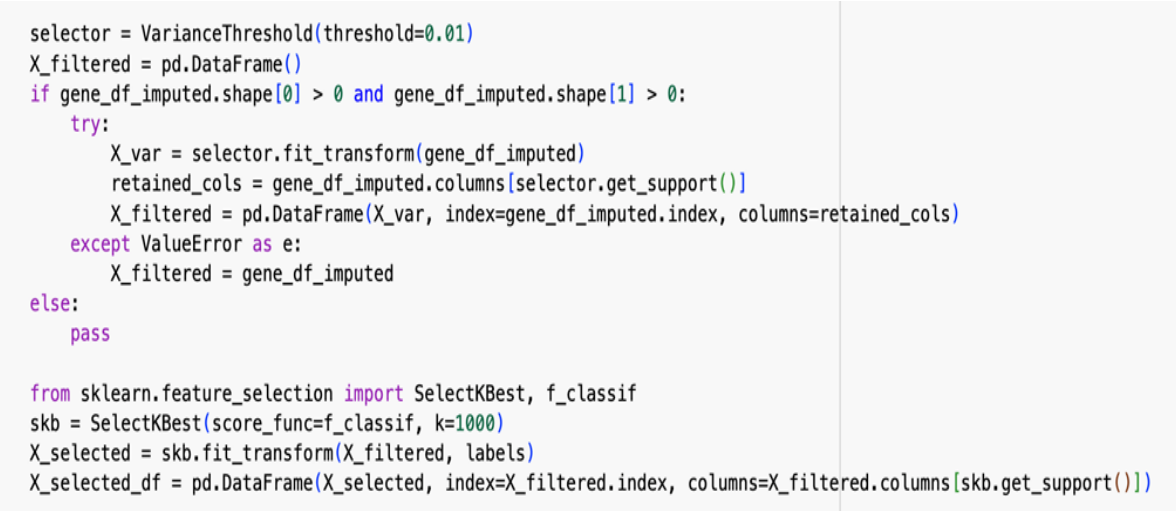
```python
missing_cols = gene_df_final.loc[:, gene_df_final.isnull().mean() < 0.1]
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean')
gene_df_imputed = pd.DataFrame(imputer.fit_transform(missing_cols),
                               index=missing_cols.index,
                               columns=missing_cols.columns)
```

Fig.9 Dropping missing values

3.3 Step 3 - Feature Selection

Variance Thresholding: The features which had a variance below the 0.01 threshold were removed. This step was performed to remove the dimensionality of the data since the features which have very little variation across the samples are likely uninformative. Reduction in computational cost and better model performance are its uses. Shape of data after variance thresholding: (149, 11136)

Statistical Test: Application of SelectKBest with the f_classif score function was done to select the top 1000 features. This is based on the F-value of a one-way ANOVA test between the features, which can help in classification. After SelectKBest: (149, 1000)

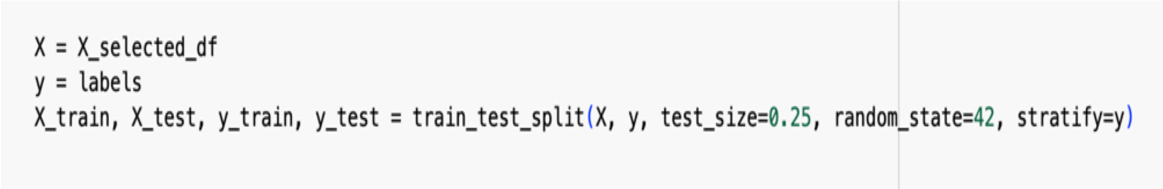


```
selector = VarianceThreshold(threshold=0.01)
X_filtered = pd.DataFrame()
if gene_df_imputed.shape[0] > 0 and gene_df_imputed.shape[1] > 0:
    try:
        X_var = selector.fit_transform(gene_df_imputed)
        retained_cols = gene_df_imputed.columns[selector.get_support()]
        X_filtered = pd.DataFrame(X_var, index=gene_df_imputed.index, columns=retained_cols)
    except ValueError as e:
        X_filtered = gene_df_imputed
else:
    pass

from sklearn.feature_selection import SelectKBest, f_classif
skb = SelectKBest(score_func=f_classif, k=1000)
X_selected = skb.fit_transform(X_filtered, labels)
X_selected_df = pd.DataFrame(X_selected, index=X_filtered.index, columns=X_filtered.columns[skb.get_support()])
```

Fig.10 Application of statistical test

3.4 Step 4 - Data Splitting

Splitting of the preprocessed data (X_selected_df) was done into training and testing data sets using the train_test_split function. The test size considered here was 0.25. The function of the training set is to train the model while the testing set evaluates the performance of the data. The random_state function ensured that the same split was made every time the code was run. We also applied stratify=y to ensure that a proper balance is maintained in both the sets.



```
X = X_selected_df
y = labels
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42, stratify=y)
```

Fig 11. Train/Test split

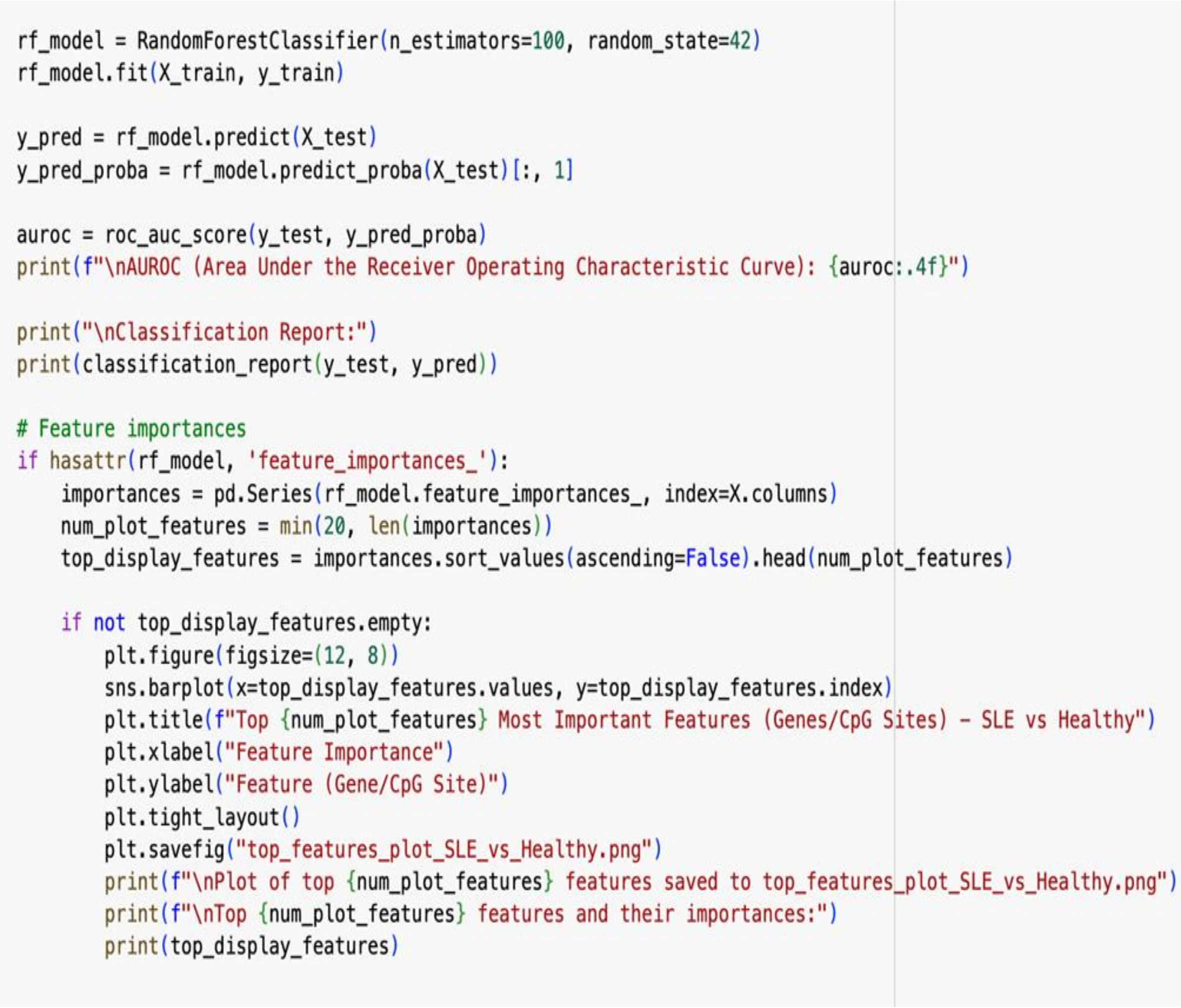
3.5 Step 5 – Algorithm Selection and Model Development

A number of algorithms were applied to the data obtained after training and test split such as Random Forest, XGBoost, Support Vector Machines, Logistic Regression and ANN- Artificial neural networks. The code which was run is added below each algorithm used. Along with finding feature importance i.e. the CpG sites which have played the most important role in the classification, wherever possible.

- **Random Forest**

  In order to handle large and complex datasets, Random Forest is a widely used choice as a supervised ML algorithm. It has high accuracy and can handle high dimensionality with reduction in any overfitting. The principal basis of Random Forest is to train many individual decision trees, followed by the combination of their outputs to give a more stable prediction. Being an ensemble learning method, it combines 2 different techniques of Bagging-Bootstrap Aggregating along with Feature Randomness.

  Bagging or Bootstrap Aggregating is a technique where variance is introduced among the trees by the training of individual trees on different bootstrap samples. These samples are taken from the original data by random sampling along with replacement. All the outputs are then aggregated, and a final prediction is then made. Here for classification, a majority vote is always taken while for regression, an average is taken. Feature Randomness is the technique used to find out the best split wherein only random subset of features at each split in every individual tree is considered. [47] Similarity between trees due to one dominant feature is prevented by this step by decorrelation of trees which provides more robustness to the whole algorithm.

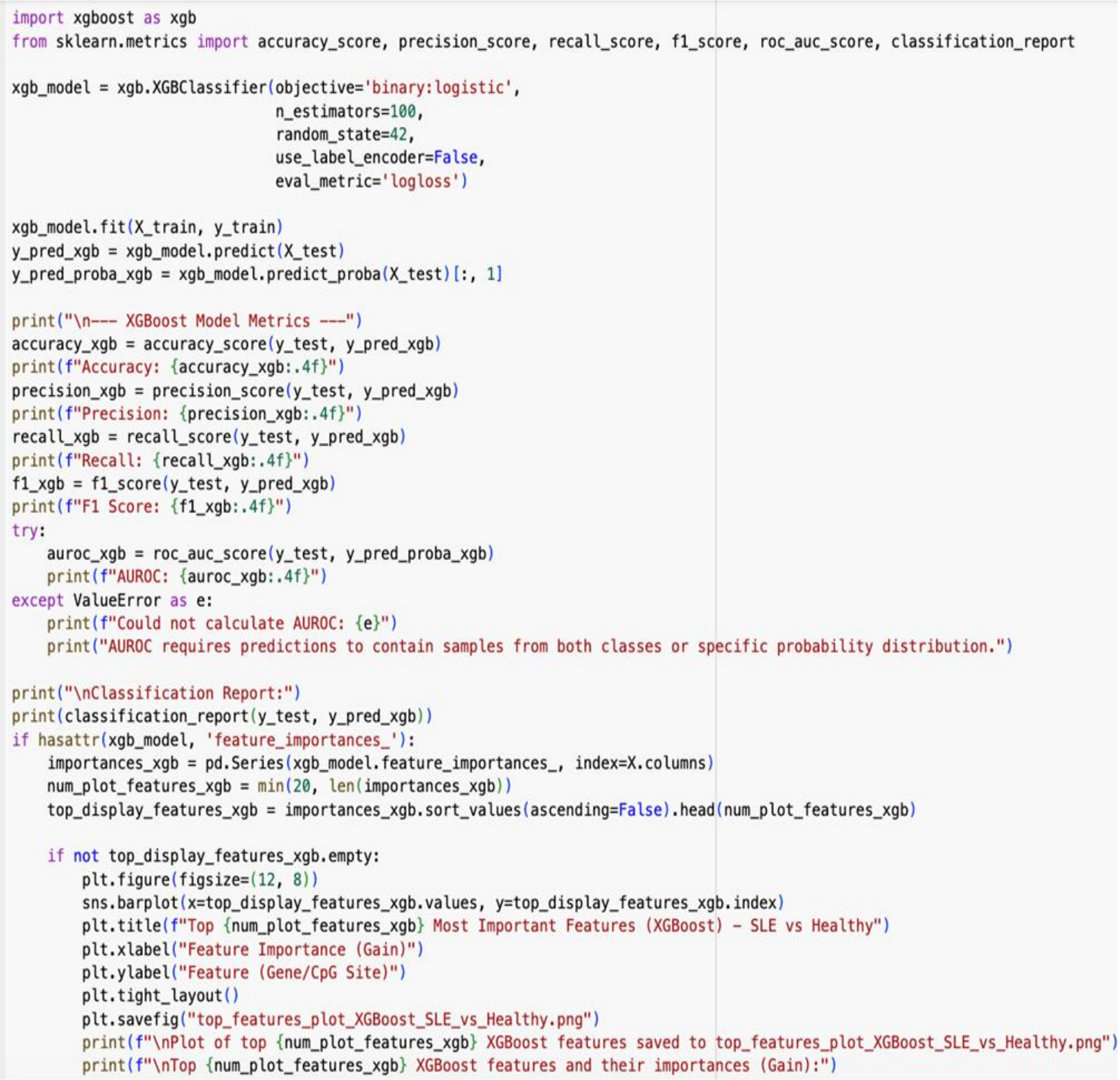


```
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

y_pred = rf_model.predict(X_test)
y_pred_proba = rf_model.predict_proba(X_test)[:, 1]

auroc = roc_auc_score(y_test, y_pred_proba)
print(f"\nAUROC (Area Under the Receiver Operating Characteristic Curve): {auroc:.4f}")

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

# Feature importances
if hasattr(rf_model, 'feature_importances_'):
    importances = pd.Series(rf_model.feature_importances_, index=X.columns)
    num_plot_features = min(20, len(importances))
    top_display_features = importances.sort_values(ascending=False).head(num_plot_features)

    if not top_display_features.empty:
        plt.figure(figsize=(12, 8))
        sns.barplot(x=top_display_features.values, y=top_display_features.index)
        plt.title(f"Top {num_plot_features} Most Important Features (Genes/CpG Sites) - SLE vs Healthy")
        plt.xlabel("Feature Importance")
        plt.ylabel("Feature (Gene/CpG Site)")
        plt.tight_layout()
        plt.savefig("top_features_plot_SLE_vs_Healthy.png")
        print(f"\nPlot of top {num_plot_features} features saved to top_features_plot_SLE_vs_Healthy.png")
        print(f"\nTop {num_plot_features} features and their importances:")
        print(top_display_features)
```

Fig. 12 Code for Random Forest

- **XGBoost**

  eXtreme Gradient Boosting is a highly sophisticated and optimized implementation of gradient boosting, which consistently performs well on structured data. Speed and Scalability are its biggest advantages and are designed to handle large datasets for computational efficiency. XGBoost is a popular choice used for different **tasks such** as regression and classification and is famous for being highly flexible. This works by boosting the built trees in a sequential order. Each new tree works in order to correct the errors made by the previous ones. The main idea is to focus on the residuals or differences which occurred in the previous trees and further improve the model. The steps of regularization and advanced pruning make it ‘extreme’ and help improve performance. It has the capacity to handle sparse data which has many missing values by skipping over the missing entries, which is a built-in feature. This step helps to speed up computation and reduce memory access time [48].

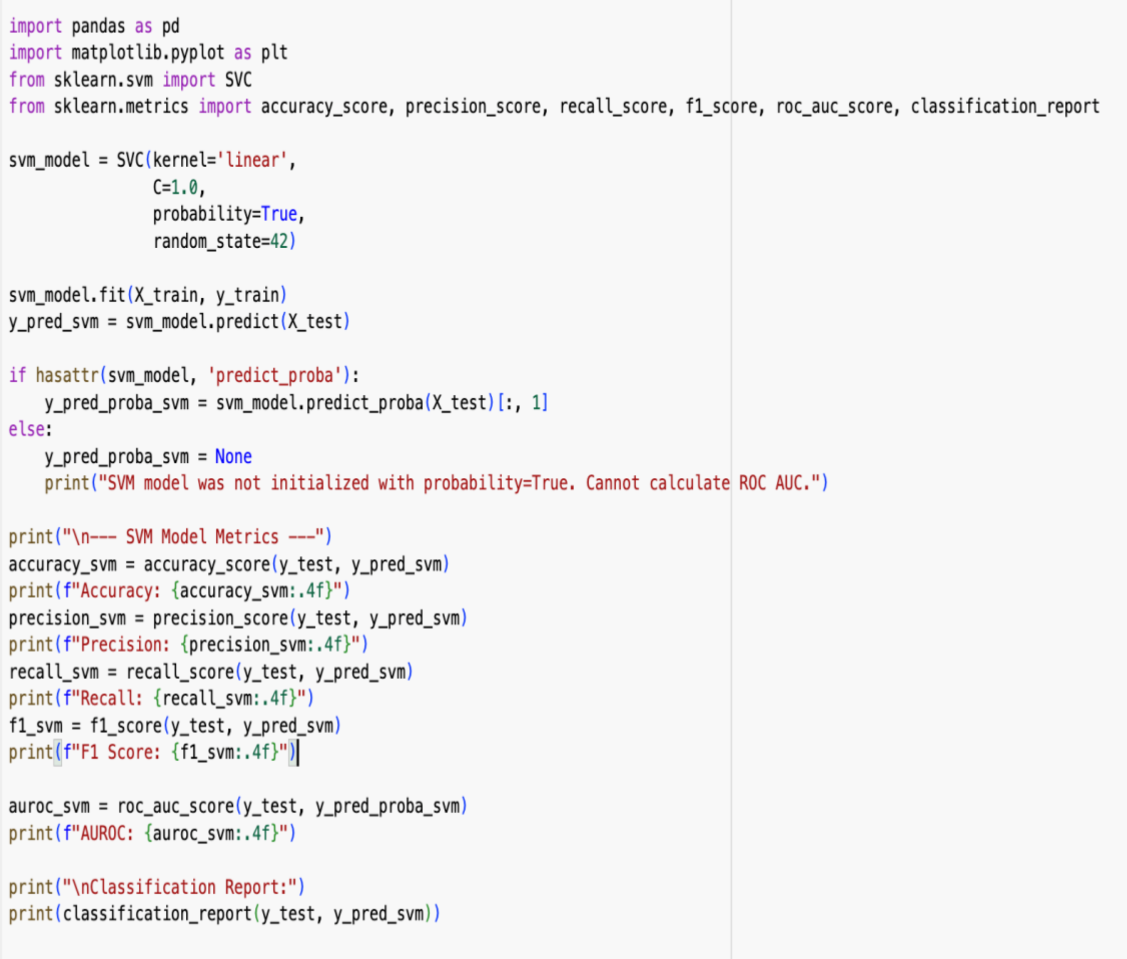


```
import xgboost as xgb
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, classification_report

xgb_model = xgb.XGBClassifier(objective='binary:logistic',
                              n_estimators=100,
                              random_state=42,
                              use_label_encoder=False,
                              eval_metric='logloss')

xgb_model.fit(X_train, y_train)
y_pred_xgb = xgb_model.predict(X_test)
y_pred_proba_xgb = xgb_model.predict_proba(X_test)[:, 1]

print("\n--- XGBoost Model Metrics ---")
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)
print(f"Accuracy: {accuracy_xgb:.4f}")
precision_xgb = precision_score(y_test, y_pred_xgb)
print(f"Precision: {precision_xgb:.4f}")
recall_xgb = recall_score(y_test, y_pred_xgb)
print(f"Recall: {recall_xgb:.4f}")
f1_xgb = f1_score(y_test, y_pred_xgb)
print(f"F1 Score: {f1_xgb:.4f}")
try:
    auroc_xgb = roc_auc_score(y_test, y_pred_proba_xgb)
    print(f"AUROC: {auroc_xgb:.4f}")
except ValueError as e:
    print(f"Could not calculate AUROC: {e}")
    print("AUROC requires predictions to contain samples from both classes or specific probability distribution.")

print("\nClassification Report:")
print(classification_report(y_test, y_pred_xgb))
if hasattr(xgb_model, 'feature_importances_'):
    importances_xgb = pd.Series(xgb_model.feature_importances_, index=X.columns)
    num_plot_features_xgb = min(20, len(importances_xgb))
    top_display_features_xgb = importances_xgb.sort_values(ascending=False).head(num_plot_features_xgb)

    if not top_display_features_xgb.empty:
        plt.figure(figsize=(12, 8))
        sns.barplot(x=top_display_features_xgb.values, y=top_display_features_xgb.index)
        plt.title(f"Top {num_plot_features_xgb} Most Important Features (XGBoost) - SLE vs Healthy")
        plt.xlabel("Feature Importance (Gain)")
        plt.ylabel("Feature (Gene/CpG Site)")
        plt.tight_layout()
        plt.savefig("top_features_plot_XGBoost_SLE_vs_Healthy.png")
        print(f"\nPlot of top {num_plot_features_xgb} XGBoost features saved to top_features_plot_XGBoost_SLE_vs_Healthy.png")
        print(f"\nTop {num_plot_features_xgb} XGBoost features and their importances (Gain):")
```

Fig.13 Code for XGBoost

- **Support Vector Machines**

  SVM’s are a type of supervised ML algorithms which are considered essential in conducting both **regression** and **classification** tasks. The basic principle is aimed at finding an optimal **hyperplane** which can separate different classes in a particular dataset. These are highly versatile and robust and have the ability to work with high dimensional spaces (high number of features present). This algorithm works by drawing an optimal hyperplane which will separate the data points by the maximum possible margin. Margin is the distance between the closest data points and the hyperplane. The differentiation between "healthy" and "SLE" in this case is a binary classification problem where the SVM will draw the decision boundary or hyperplane [49].

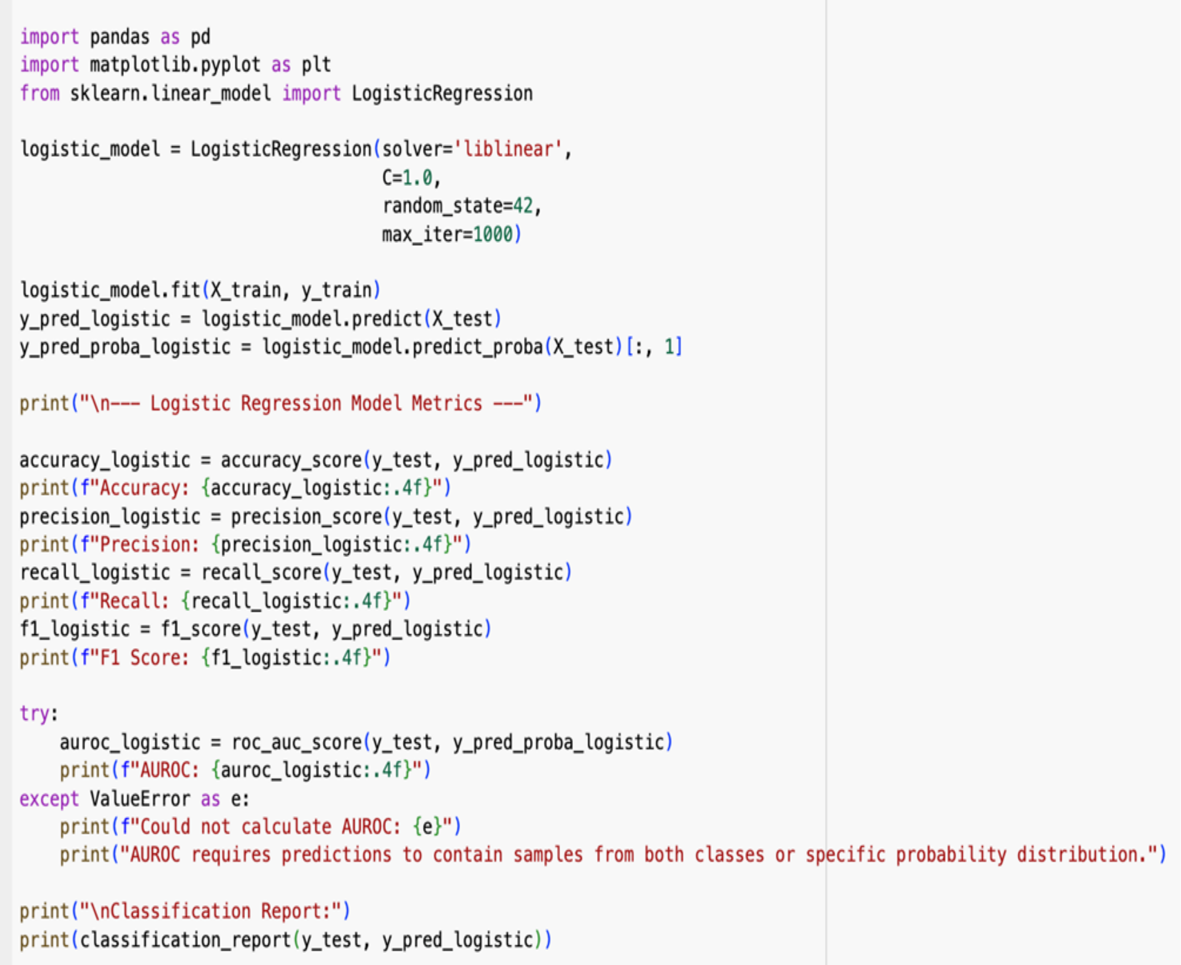


```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, classification_report

svm_model = SVC(kernel='linear',
                C=1.0,
                probability=True,
                random_state=42)

svm_model.fit(X_train, y_train)
y_pred_svm = svm_model.predict(X_test)

if hasattr(svm_model, 'predict_proba'):
    y_pred_proba_svm = svm_model.predict_proba(X_test)[:, 1]
else:
    y_pred_proba_svm = None
    print("SVM model was not initialized with probability=True. Cannot calculate ROC AUC.")

print("\n--- SVM Model Metrics ---")
accuracy_svm = accuracy_score(y_test, y_pred_svm)
print(f"Accuracy: {accuracy_svm:.4f}")
precision_svm = precision_score(y_test, y_pred_svm)
print(f"Precision: {precision_svm:.4f}")
recall_svm = recall_score(y_test, y_pred_svm)
print(f"Recall: {recall_svm:.4f}")
f1_svm = f1_score(y_test, y_pred_svm)
print(f"F1 Score: {f1_svm:.4f}")

auroc_svm = roc_auc_score(y_test, y_pred_proba_svm)
print(f"AUROC: {auroc_svm:.4f}")

print("\nClassification Report:")
print(classification_report(y_test, y_pred_svm))
```

Fig.14 Code for SVM

- **Logistic Regression**

  This ML algorithm is generally used for tasks related to **classification** and is easy to understand and interpret due to its simplicity. It is a linear model which is less computationally intensive and less sensitive to the outliers present. LR is used in prediction of binary outcomes like 0/1 in our case and can also be used with multi class classification datasets. The main basis of this technique is the sigmoid function or logistic function which forms an S shaped curve and converts the linear output into probability. The linear relationship between the logit or log-odds, which are natural logarithm of the odds of an event and the independent variables or features, makes it a more generalized model [50].



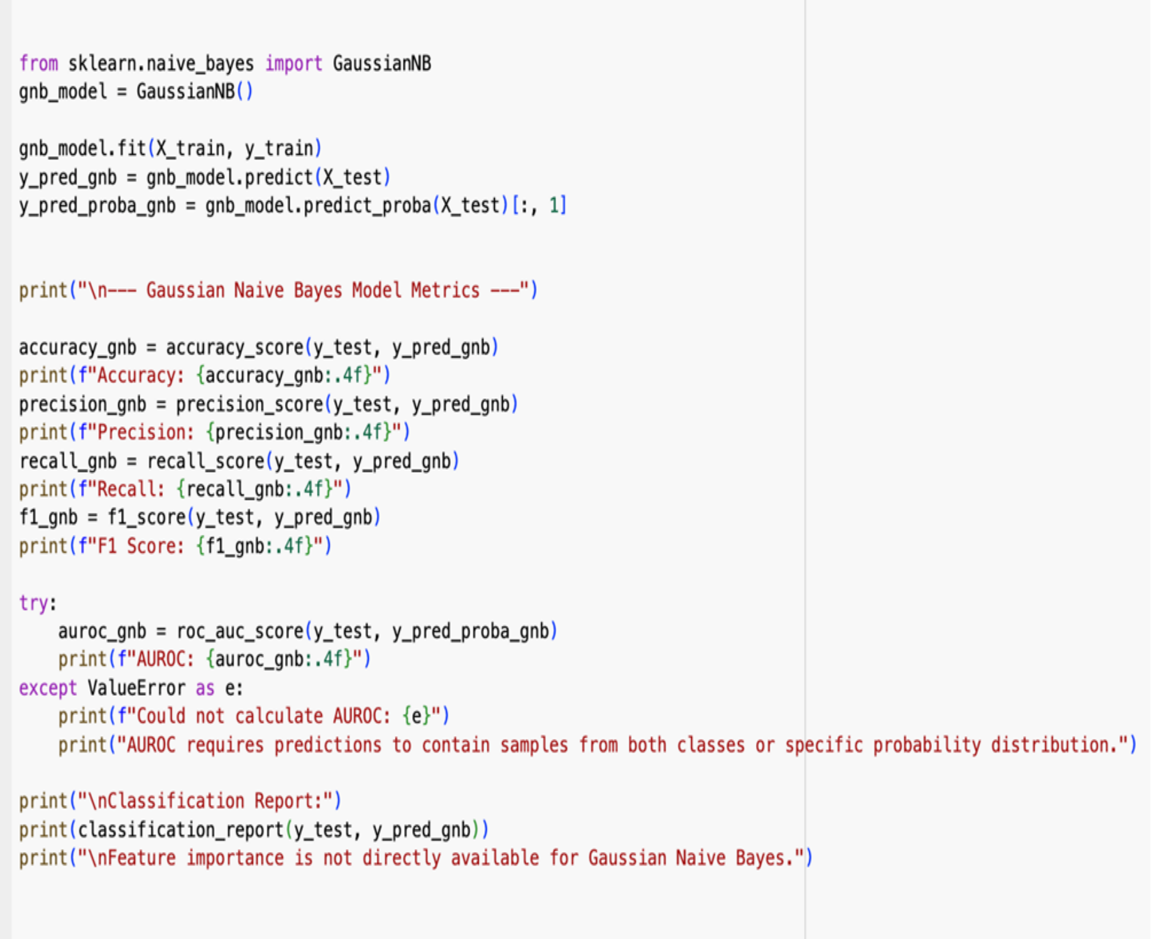
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression

logistic_model = LogisticRegression(solver='liblinear',
                                    C=1.0,
                                    random_state=42,
                                    max_iter=1000)

logistic_model.fit(X_train, y_train)
y_pred_logistic = logistic_model.predict(X_test)
y_pred_proba_logistic = logistic_model.predict_proba(X_test)[:, 1]

print("\n--- Logistic Regression Model Metrics ---")

accuracy_logistic = accuracy_score(y_test, y_pred_logistic)
print(f"Accuracy: {accuracy_logistic:.4f}")
precision_logistic = precision_score(y_test, y_pred_logistic)
print(f"Precision: {precision_logistic:.4f}")
recall_logistic = recall_score(y_test, y_pred_logistic)
print(f"Recall: {recall_logistic:.4f}")
f1_logistic = f1_score(y_test, y_pred_logistic)
print(f"F1 Score: {f1_logistic:.4f}")

try:
    auroc_logistic = roc_auc_score(y_test, y_pred_proba_logistic)
    print(f"AUROC: {auroc_logistic:.4f}")
except ValueError as e:
    print(f"Could not calculate AUROC: {e}")
    print("AUROC requires predictions to contain samples from both classes or specific probability distribution.")

print("\nClassification Report:")
print(classification_report(y_test, y_pred_logistic))
```

Fig. 15 Code for Logistic Regression

- **Naive Bayes**

  This is a probability-based classifier which works on the basis of Bayes Theorem. This theorem helps in the calculation of the probability of a class or feature where it takes into consideration any prior knowledge related to any class. It is called ‘naive’ due to the assumption of conditional independence between features. The first step is the calculation of probability of each class present in the training data followed by the calculation of posterior probability for unseen data points. There are three different types of Naive Bayes

Classifiers such as Gaussian, Multinomial and Bernoulli. It shows great potential with high-dimensional data and is very fast to train with easy implementation [51]. Feature importance is not available for this technique.

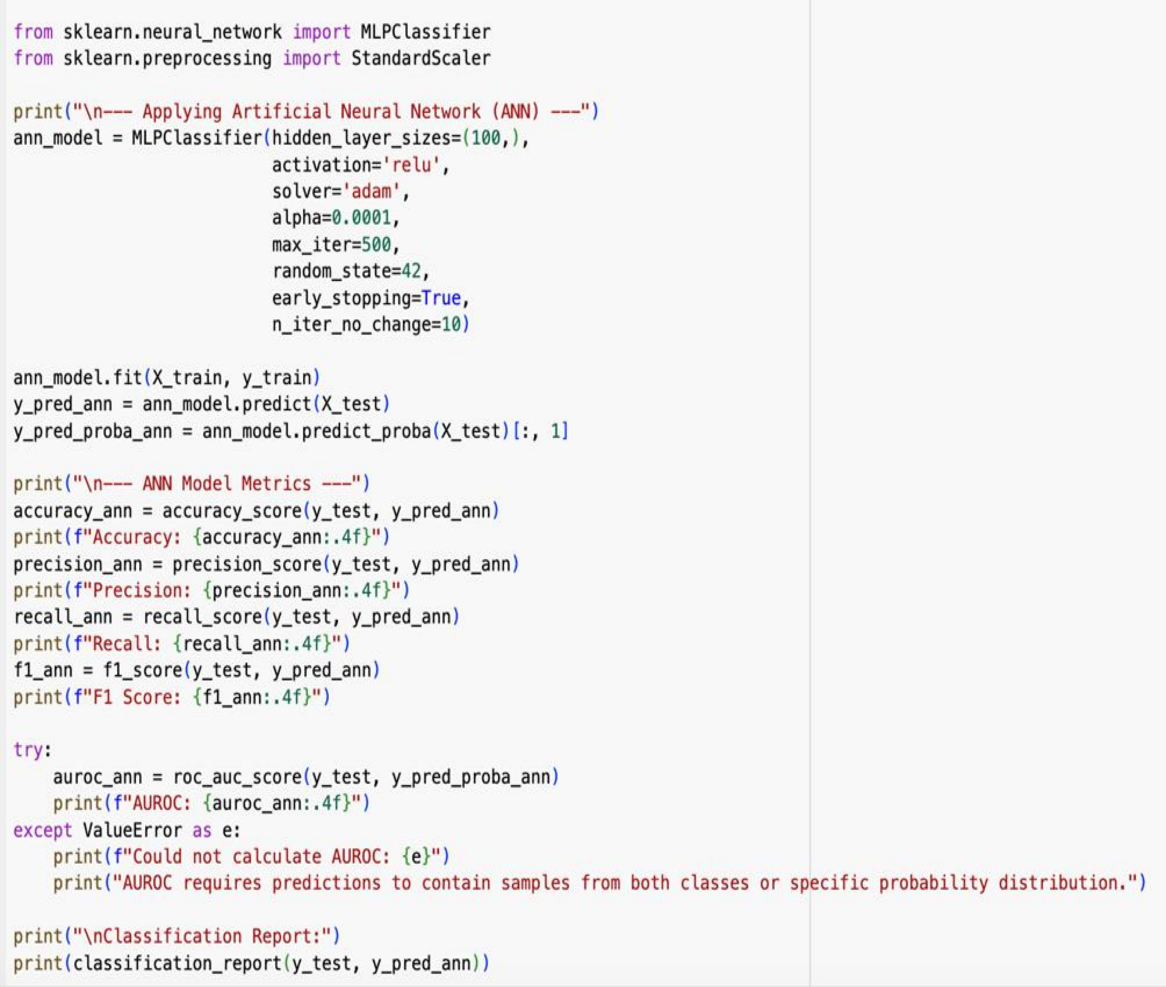


```
from sklearn.naive_bayes import GaussianNB
gnb_model = GaussianNB()

gnb_model.fit(X_train, y_train)
y_pred_gnb = gnb_model.predict(X_test)
y_pred_proba_gnb = gnb_model.predict_proba(X_test)[:, 1]


print("\n--- Gaussian Naive Bayes Model Metrics ---")

accuracy_gnb = accuracy_score(y_test, y_pred_gnb)
print(f"Accuracy: {accuracy_gnb:.4f}")
precision_gnb = precision_score(y_test, y_pred_gnb)
print(f"Precision: {precision_gnb:.4f}")
recall_gnb = recall_score(y_test, y_pred_gnb)
print(f"Recall: {recall_gnb:.4f}")
f1_gnb = f1_score(y_test, y_pred_gnb)
print(f"F1 Score: {f1_gnb:.4f}")

try:
    auroc_gnb = roc_auc_score(y_test, y_pred_proba_gnb)
    print(f"AUROC: {auroc_gnb:.4f}")
except ValueError as e:
    print(f"Could not calculate AUROC: {e}")
    print("AUROC requires predictions to contain samples from both classes or specific probability distribution.")

print("\nClassification Report:")
print(classification_report(y_test, y_pred_gnb))
print("\nFeature importance is not directly available for Gaussian Naive Bayes.")
```

Fig. 16 Code for Naive Bayes

- **Artificial Neural Networks**

  ANNs are ML algorithms which are specifically designed to classify data, make predictions and recognize patterns by understanding the complex relationships between inputs and outputs. It mimics the human brain by being composed of a number of layers of artificial interconnected neurons. Just as the biological neural networks work, these neurons make the information flow from the input layer to the output layer via the feed forward connection. The basis of ANN includes several processes such as forward propagation, loss calculation, back propagation etc. This helps in refining the technique to make accurate predictions. Parallel processing, being robust and easy removal of noisy data are some of the main advantages of using this technique in model development [52]. This technique also does not show any feature importance.



```
from sklearn.neural_network import MLPClassifier
from sklearn.preprocessing import StandardScaler

print("\n--- Applying Artificial Neural Network (ANN) ---")
ann_model = MLPClassifier(hidden_layer_sizes=(100,),
                          activation='relu',
                          solver='adam',
                          alpha=0.0001,
                          max_iter=500,
                          random_state=42,
                          early_stopping=True,
                          n_iter_no_change=10)

ann_model.fit(X_train, y_train)
y_pred_ann = ann_model.predict(X_test)
y_pred_proba_ann = ann_model.predict_proba(X_test)[:, 1]

print("\n--- ANN Model Metrics ---")
accuracy_ann = accuracy_score(y_test, y_pred_ann)
print(f"Accuracy: {accuracy_ann:.4f}")
precision_ann = precision_score(y_test, y_pred_ann)
print(f"Precision: {precision_ann:.4f}")
recall_ann = recall_score(y_test, y_pred_ann)
print(f"Recall: {recall_ann:.4f}")
f1_ann = f1_score(y_test, y_pred_ann)
print(f"F1 Score: {f1_ann:.4f}")

try:
    auroc_ann = roc_auc_score(y_test, y_pred_proba_ann)
    print(f"AUROC: {auroc_ann:.4f}")
except ValueError as e:
    print(f"Could not calculate AUROC: {e}")
    print("AUROC requires predictions to contain samples from both classes or specific probability distribution.")

print("\nClassification Report:")
print(classification_report(y_test, y_pred_ann))
```

Fig. 17 Code for ANN

3.6 Step 6 – Model Evaluation

Model evaluation is done with the help of different quantitative measures which are used as metrics to assess the performance of a ML algorithm applied to a dataset. They give us information about how well a model works and then generalizes unseen data and can also be used to compare different models on the basis of its strengths and weaknesses. Different metrics give information about different aspects and need to be selected on the basis of the data and the output desired.

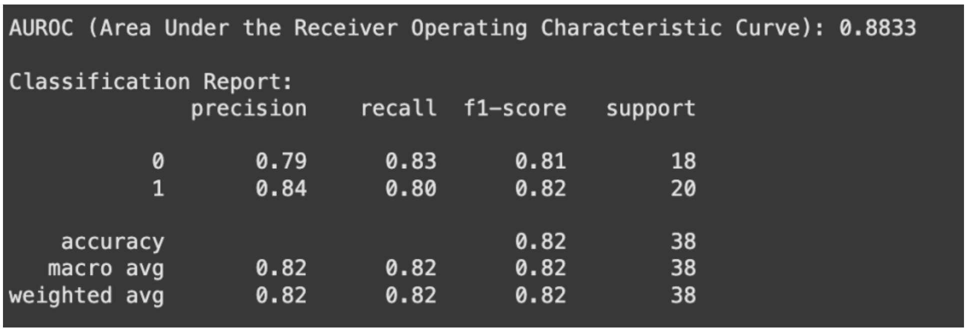
1. Accuracy - It is a measure of effectiveness of the classification done by the ML algorithm on the data.
2. Precision - It is the percentage of events that were accurately predicted to be positive.
3. F1 Score - It is a parameter which is the average of accuracy and recall and hence detects the effectiveness of any classification system, when the evaluation of a model's performance is done in case of binary classification tasks.
4. Recall - Recall counts determining which of the affirmative examples had accurate labels applied
5. AUCROC – It is referred to as the Area Under the Receiver Operating Characteristic Curve. It is the ability of the ML classifier ability to perform class differentiation. It is independent of any chosen classification threshold. A better discriminatory power is inferred from higher AUC.

CHAPTER – 4

RESULT AND DISCUSSION

After running the code using the above-mentioned ML algorithms, the result was evaluated on the basis of the values of different metrics obtained along with feature importance where the top 20 CpG sites which played the most significant role in classification were obtained and plotted.

4.1 Random Forest



```
AUROC (Area Under the Receiver Operating Characteristic Curve): 0.8833

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.83      0.81        18
           1       0.84      0.80      0.82        20

    accuracy                           0.82        38
   macro avg       0.82      0.82      0.82        38
weighted avg       0.82      0.82      0.82        38
```

Fig. 18 Performance metrics for Random Forest

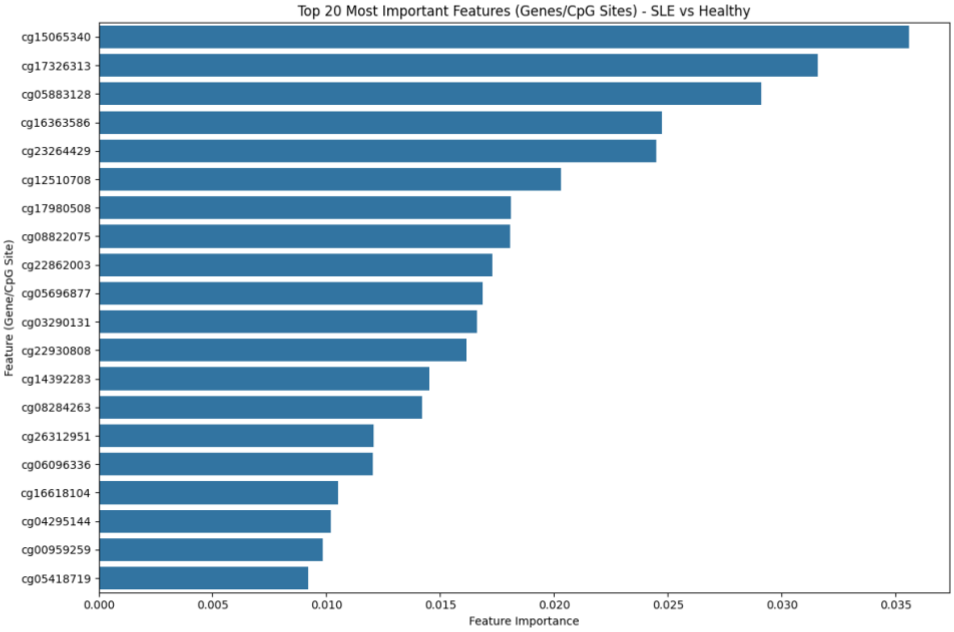
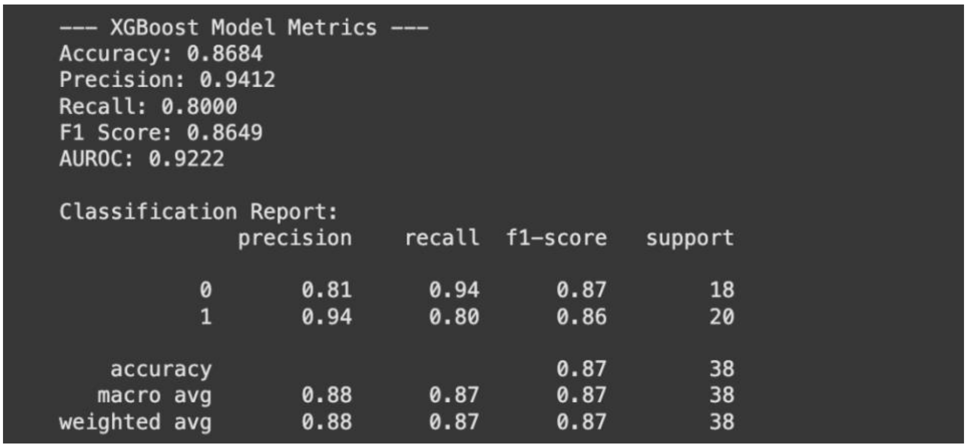


Fig. 19 Bar Graph showing Top 20 CpG sites using Random Forest

4.2 XGBoost



```
--- XGBoost Model Metrics ---
Accuracy: 0.8684
Precision: 0.9412
Recall: 0.8000
F1 Score: 0.8649
AUROC: 0.9222

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.94      0.87        18
           1       0.94      0.80      0.86        20

    accuracy                           0.87        38
   macro avg       0.88      0.87      0.87        38
weighted avg       0.88      0.87      0.87        38
```

Fig. 20 Performance metrics for XGBoost

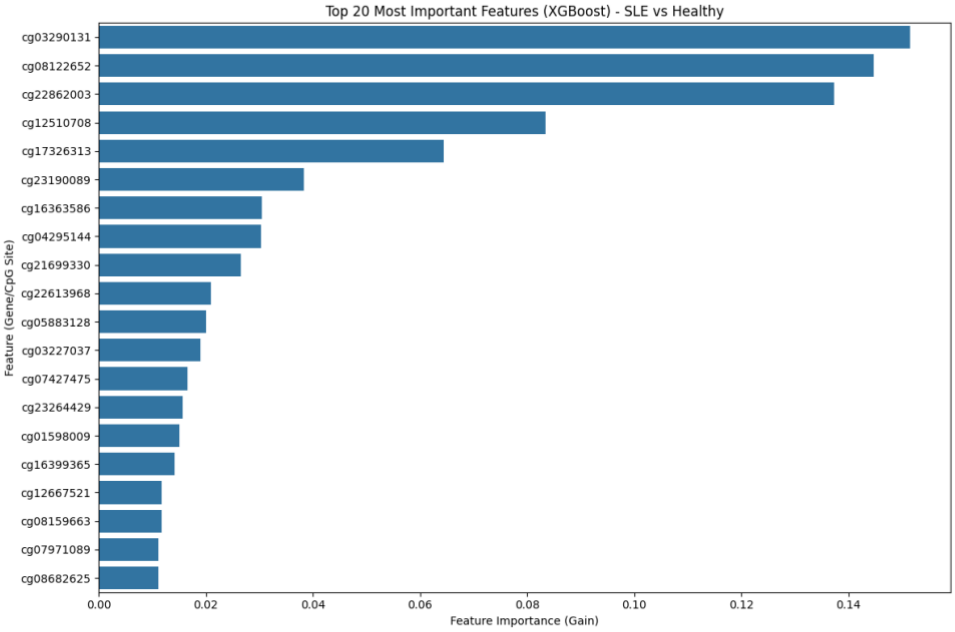
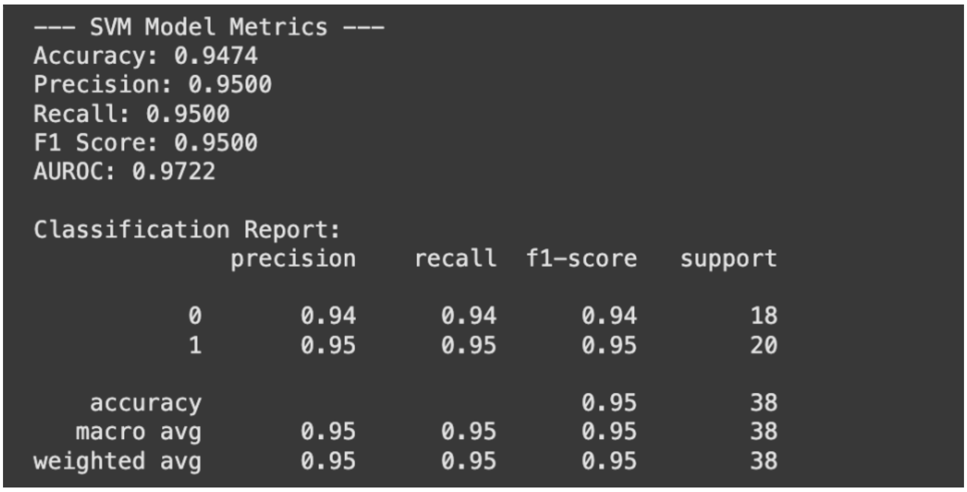


Fig. 21 Bar Graph showing Top 20 CpG sites

4.3 Support Vector Machines



```
--- SVM Model Metrics ---
Accuracy: 0.9474
Precision: 0.9500
Recall: 0.9500
F1 Score: 0.9500
AUROC: 0.9722

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94        18
           1       0.95      0.95      0.95        20

    accuracy                           0.95        38
   macro avg       0.95      0.95      0.95        38
weighted avg       0.95      0.95      0.95        38
```

Fig. 22 Performance metrics for SVM

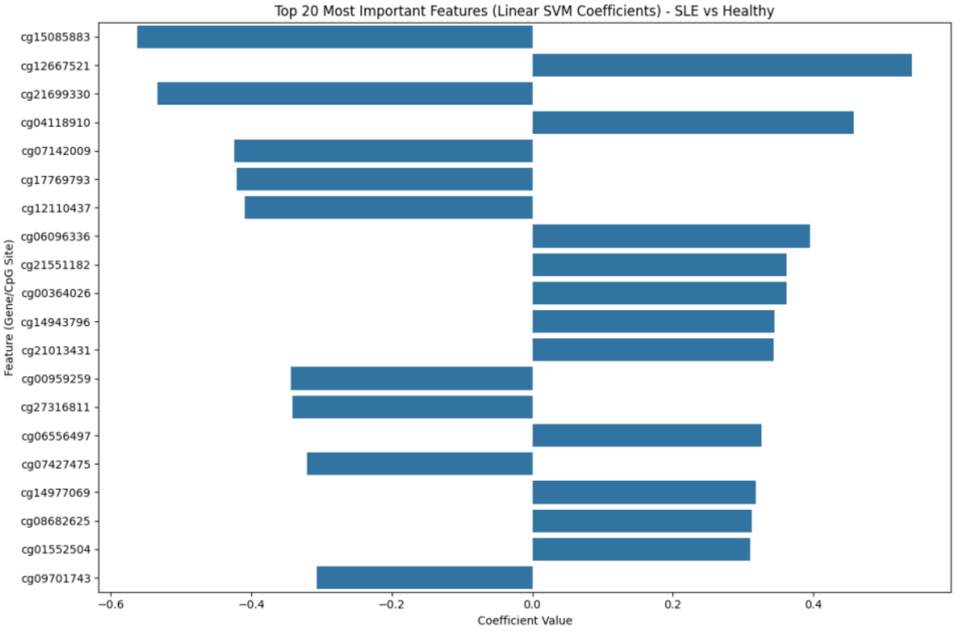
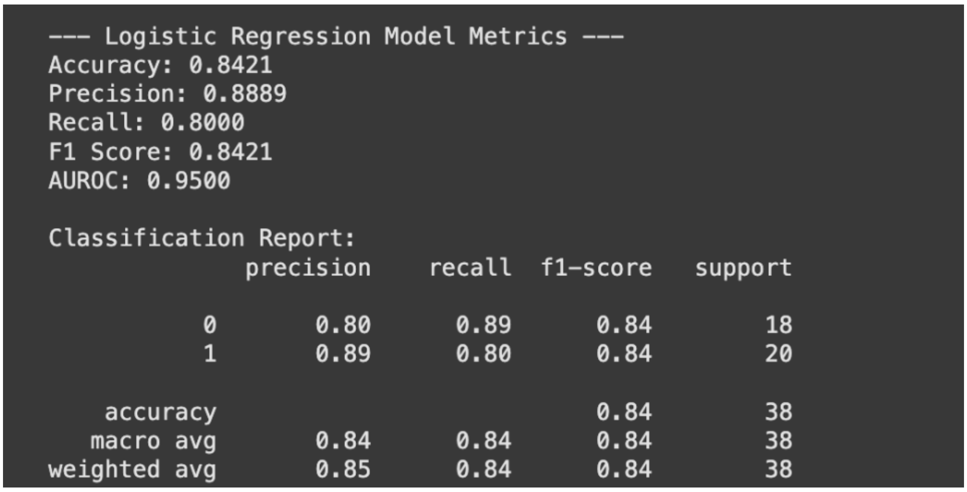


Fig. 23 Bar Graph showing Top 20 CpG sites using SVM

4.4 Logistic Regression



```
--- Logistic Regression Model Metrics ---
Accuracy: 0.8421
Precision: 0.8889
Recall: 0.8000
F1 Score: 0.8421
AUROC: 0.9500

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.89      0.84        18
           1       0.89      0.80      0.84        20

    accuracy                           0.84        38
   macro avg       0.84      0.84      0.84        38
weighted avg       0.85      0.84      0.84        38
```

Fig. 24 Performance metrics for Logistic Regression

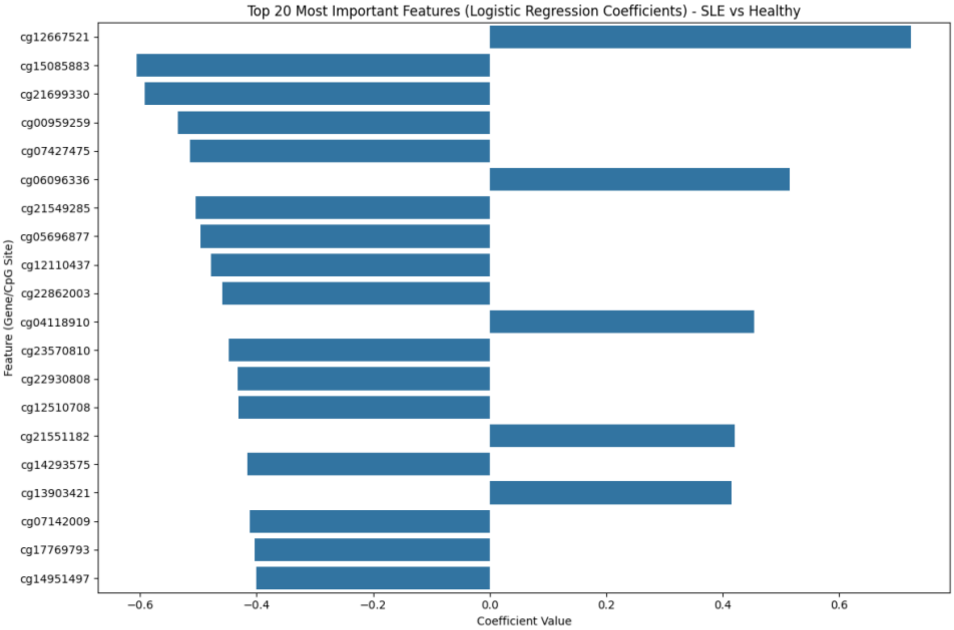
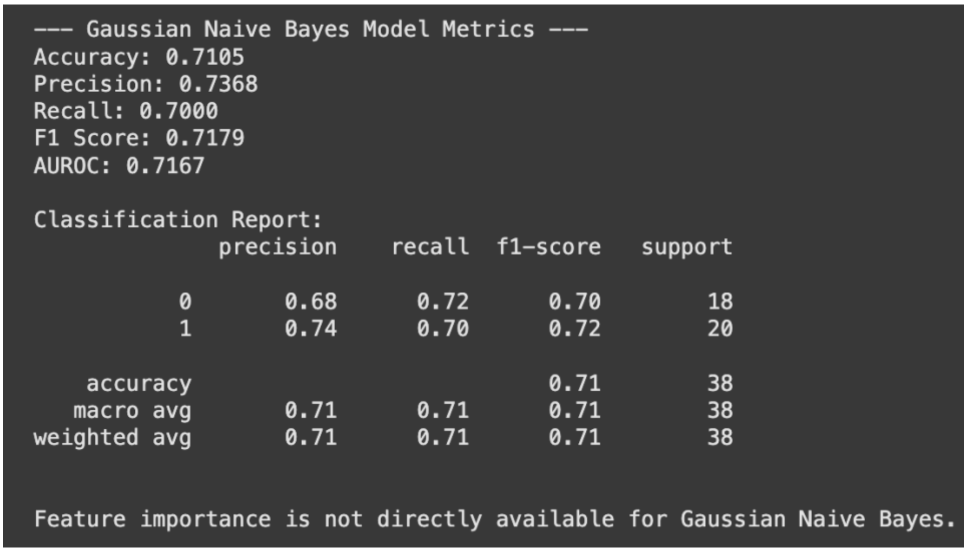


Fig. 25 Bar Graph showing Top 20 CpG sites using Logistic Regression

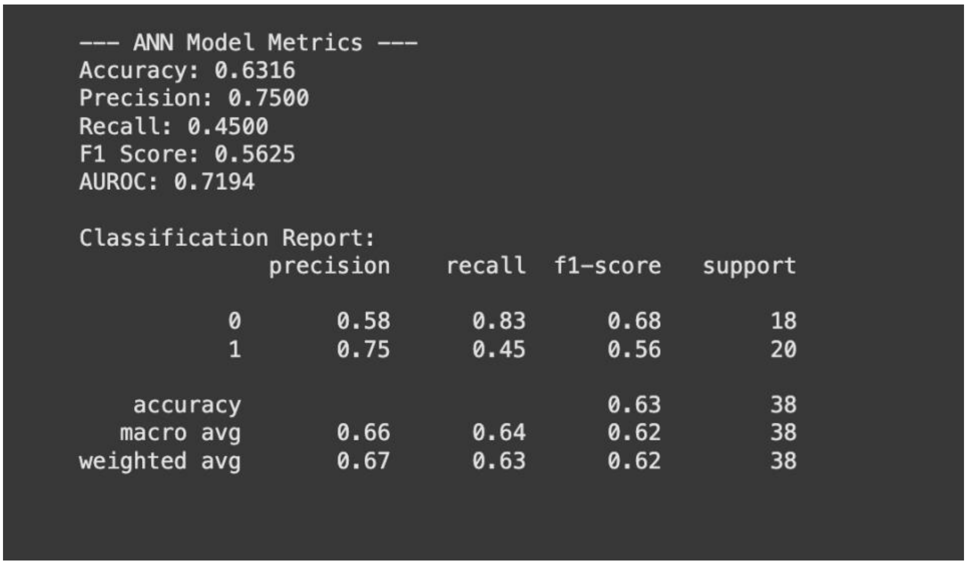
4.5 Naive Bayes



```
--- Gaussian Naive Bayes Model Metrics ---
Accuracy: 0.7105
Precision: 0.7368
Recall: 0.7000
F1 Score: 0.7179
AUROC: 0.7167

Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.72      0.70        18
           1       0.74      0.70      0.72        20

    accuracy                           0.71        38
   macro avg       0.71      0.71      0.71        38
weighted avg       0.71      0.71      0.71        38


Feature importance is not directly available for Gaussian Naive Bayes.
```

Fig. 26 Performance metrics for Naïve Bayes

4.6 Artificial Neural Network



```
--- ANN Model Metrics ---
Accuracy: 0.6316
Precision: 0.7500
Recall: 0.4500
F1 Score: 0.5625
AUROC: 0.7194

Classification Report:
              precision    recall  f1-score   support

           0       0.58      0.83      0.68        18
           1       0.75      0.45      0.56        20

    accuracy                           0.63        38
   macro avg       0.66      0.64      0.62        38
weighted avg       0.67      0.63      0.62        38
```

Fig. 27 Performance metrics for ANN

Table.2 Evaluation metrices of ML techniques

ML Techniques	Accuracy	AUROC	Precision	Recall	F1-score
Random Forest	0.82	0.8833	0.84	0.80	0.82
XGBoost	0.8684	0.9222	0.9412	0.8	0.8649
SVM	0.9474	0.9722	0.95	0.95	0.95
Logistic Regression	0.8421	0.95	0.8889	0.80	0.8421
Naive Bayes	0.7104	0.7167	0.7368	0.70	0.7179
ANN	0.6316	0.7194	0.75	0.45	0.5625

- Based on the above data it can be interpreted that SVM (Support Vector Machine) is the best-performing model. It has the highest AUROC=0.9722, which indicates excellent discriminatory power. It also exhibits the highest recall, F1-score and precision for the identification of positive SLE cases correctly. There is also a minimization of false positives and negatives. It also has the highest overall accuracy.
- A strong performance is also seen by Logistic Regression and XGBoost, by evaluating the high AUROC scores of 0.95 and 0.92 respectively. XGBoost showed a high precision value for class 1 i.e. for SLE to be 0.94 but has a recall value lower than SVM. Logistic Regression can also be considered as good candidate for model development as it showed a good performer with a high AUROC.
- Random Forest showed a reasonably well performance but not as strong as SVM, Logistic Regression, or XGBoost in terms of the key evaluation metrics.
- Naive Bayes and ANN have the lowest performance metrics which includes AUROC, which indicates that they are not the most suitable for this task of classification with the current data of SLE.

CHAPTER – 5

CONCLUSION, GAPS & FUTURE SCOPE

5.1 Conclusion

The data obtained from the Gene Expression Omnibus (GEO) dataset GSE59250 which included DNA methylation data giving β values for individual CpG sites of 149 samples of both healthy and patients suffering from SLE was used and various ML algorithm such as Random Forest, XGBoost, SVM, Logistic Regression, ANN, Naive Bayes were successfully applied. Various evaluation metrics such as AUROC, Accuracy, Recall, Precision, and F1 Score were obtained to identify the most suitable model development. Based on the results and the discussion, it can be concluded that the SVM algorithm is the best fit technique for the process of model development in the prediction of SLE in patients based on the epigenetic data obtained from the database. It outperforms all the other models across the evaluation metrics obtained after the model was run for this task.

In conclusion, by using such epigenetic datasets to extract relevant patterns, AI and ML techniques have shown tremendous promise in prediction of autoimmune disorders. This method has several benefits including the combination of numerous data sources and offering a thorough evaluation of the disease risk and prognosis.

5.2 Gaps in current knowledge

Even though AI & ML have been increasing used in biology as they hold great potential in diagnostics, treatment & personalized medicine, they still have a number of limitations. The presence of large scale, high quality for the training of AI models is scarce. Also epigenetic data contains various types of modifications across the genome which includes datasets that are complex and high-dimensional which are difficult to comprehend. It is difficult to find consistent patterns due to the heterogeneity of autoimmune diseases as the patients have different phenotypes, disease progression, treatment etc. Many models including the Deep learning models lack interpretability and have a ‘black box’ nature ie they do not give any information on underlying mechanisms or why or how the particular autoimmune disease occurs.

Even though the prediction of epigenetic modifications can be done using AI models they still lack biological understanding and significance, i.e., how the disease progresses in an individual. Further complication to the use of AI & ML occurs due to ethical concerns. Patient consent for data procurement & algorithmic bias are some of the concerns which along with privacy issues where patient health information is linked with epigenetic data maybe disregarded by some. Integration of multi-omic data which includes genomics, transcriptomics, proteomics, and epigenomics requires certain computationally demanding techniques which ultimately decreases the power of prediction of AI models.

5.3 Future Perspectives

The emerging use of AI & ML in the prediction of autoimmune disorders using epigenetic modifications holds a promising future. By identification of Epigenetic modification patterns, we can improve disease prediction and can help in early diagnosis of a particular autoimmune disorder.AI models can identify and predict how a particular disease will progress which may aid in clinical interference at the right time for a personalized treatment. By integrating an individual's multi-omic data we can identify personalized treatment for the same. This includes testing AI models which can let us know which drugs are better suited for an individual treatment. After analysis of large datasets of epigenetic data, we can enhance novel biomarker discovery which are specific to the autoimmune conditions. Complex network of gene regulation can be used by ML algorithms to learn and decipher the role of DNA methylation, histone modifications etc. Many AI driven tools can be used in the real time monitoring of modifications and help us to learn about how a particular epigenetic change influences a gene regulatory cascade. A collaboration between AI developers and biologists can help in a better understanding of the results obtained from the models and make them more interpretable, high-dimensional and relevant. Even though we might face certain challenges at present but in the long run the use of AI will help humanity in better research and finding solutions to problems in a short time with better accuracy.

REFERENCES

[1] G. S. Cooper, M. L. K. Bynum, and E. C. Somers, “Recent insights in the epidemiology of autoimmune diseases: Improved prevalence estimates and understanding of clustering of diseases,” *Journal of Autoimmunity*, vol. 33, no. 3–4, pp. 197–207, Oct. 2009, doi: 10.1016/j.jaut.2009.09.008.

[2] S. Mu *et al.*, “Autoimmune disease: a view of epigenetics and therapeutic targeting,” *Frontiers in Immunology*, vol. 15, Nov. 2024, doi: 10.3389/fimmu.2024.1482728.

[3] N. G. Nia, E. Kaplanoglu, and A. Nasab, “Evaluation of artificial intelligence techniques in disease diagnosis and prediction,” *Discover Artificial Intelligence*, vol. 3, no. 1, Jan. 2023.

[4] R. Mazzone *et al.*, “The emerging role of epigenetics in human autoimmune disorders,” *Clinical Epigenetics*, vol. 11, no. 1, Feb. 2019.

[5] D. Fairweather, S. Frisancho-Kiss, and N. R. Rose, “Sex Differences in Autoimmune Disease from a Pathological Perspective,” *American Journal of Pathology*, vol. 173, no. 3, pp. 600–609, Aug. 2008.

[6] R. Mazzone *et al.*, “The emerging role of epigenetics in human autoimmune disorders,” *Clinical Epigenetics*, vol. 11, no. 1, Feb. 2019.

[7] B. Sun, L. Hu, Z.-Y. Luo, X.-P. Chen, H.-H. Zhou, and W. Zhang, “DNA methylation perspectives in the pathogenesis of autoimmune diseases,” *Clinical Immunology*, vol. 164, pp. 21–27, Jan. 2016.

[8] D. E. Handy, R. Castro, and J. Loscalzo, “Epigenetic modifications,” *Circulation*, vol. 123, no. 19, pp. 2145–2156, May 2011.

[9] A. E. A. Surace and C. M. Hedrich, “The role of Epigenetics in Autoimmune/Inflammatory Disease,” *Frontiers in Immunology*, vol. 10, Jul. 2019.

[10] M. Neidhart, “DNA methylation in synovial fibroblasts,” in *Elsevier eBooks*, 2015, pp. 381–393.

[11] E. Man and S. Evran, “Deacetylation of histones and non-histone proteins in inflammatory diseases and cancer therapeutic potential of Histone deacetylase inhibitors,” *Current Genomics*, vol. 24, no. 3, pp. 136–145, May 2023.

[12] D. M. Absher *et al.*, “Genome-Wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations,” *PLoS Genetics*, vol. 9, no. 8, p. e1003678, Aug. 2013.

[13] D. Fernandez, E. Bonilla, P. Phillips, and A. Perl, “Signaling abnormalities in systemic lupus erythematosus as potential drug targets,” *Endocrine Metabolic & Immune Disorders - Drug Targets*, vol. 6, no. 4, pp. 305–311, Dec. 2006.

[14] R. P. Singh *et al.*, “The role of miRNA in inflammation and autoimmunity,” *Autoimmunity Reviews*, vol. 12, no. 12, pp. 1160–1165, Jul. 2013.

[15] N. Parveen and S. Dhawan, “DNA methylation patterning and the regulation of beta cell homeostasis,” *Frontiers in Endocrinology*, vol. 12, May 2021.

[16] L. Eliasson, “The small RNA miR-375 – a pancreatic islet abundant miRNA with multiple roles in endocrine beta cell function,” *Molecular and Cellular Endocrinology*, vol. 456, pp. 95–101, Feb. 2017.

[17] N. Celarain and J. Tomas-Roig, “Aberrant DNA methylation profile exacerbates inflammation and neurodegeneration in multiple sclerosis patients,” *Journal of Neuroinflammation*, vol. 17, no. 1, Jan. 2020.

[18] Z. A. Aljawadi, M. A. N. Kashmoola, A. M. Almahdawi, A. R. Al-Derzi, and B. A. Abdul-Majeed, “Micro ribonucleic acids (20A, 146A, and 155) and forkhead box P3 genes expressions in multiple sclerosis patients,” *Multiple Sclerosis and Related Disorders*, vol. 26, p. 252, Nov. 2018.

[19] G. S. Cooper, M. L. K. Bynum, and E. C. Somers, “Recent insights in the epidemiology of autoimmune diseases: Improved prevalence estimates and understanding of clustering of diseases,” *Journal of Autoimmunity*, vol. 33, no. 3–4, pp. 197–207, Oct. 2009, doi: 10.1016/j.jaut.2009.09.008.

[20] D. S. Pisetsky, “Antinuclear antibody testing — misunderstood or misbegotten?,” *Nature Reviews Rheumatology*, vol. 13, no. 8, pp. 495–502, May 2017, doi: 10.1038/nrrheum.2017.74.

[21] M. E. Orme, A. Voreck, R. Aksouh, R. Ramsey-Goldman, and M. W. J. Schreurs, “Systematic review of anti-dsDNA testing for systemic lupus erythematosus: A meta-analysis of the diagnostic test specificity of an anti-dsDNA fluorescence enzyme immunoassay,” *Autoimmunity Reviews*, vol. 20, no. 11, p. 102943, Sep. 2021, doi: 10.1016/j.autrev.2021.102943.

[22] E. Benito-Garcia, P. H. Schur, and R. Lahita, “Guidelines for immunologic laboratory testing in the rheumatic diseases: Anti-Sm and anti-RNP antibody tests,” *Arthritis Care & Research*, vol. 51, no. 6, pp. 1030–1044, Dec. 2004, doi: 10.1002/art.20836.

[23] K. Nishimura *et al.*, “Meta-analysis: Diagnostic accuracy of Anti–Cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis,” *Annals*

of Internal Medicine, vol. 146, no. 11, p. 797, Jun. 2007, doi: 10.7326/0003-4819-146-11-200706050-00008.

[24] M. Abbassi-Ghanavati, B. M. Casey, C. Y. Spong, D. D. McIntire, L. M. Halvorson, and F. G. Cunningham, "Pregnancy outcomes in women with thyroid peroxidase antibodies," *Obstetrics and Gynecology*, vol. 116, no. 2, pp. 381–386, Jul. 2010, doi: 10.1097/aog.0b013e3181e904e5.

[25] M. Absinta *et al.*, "A lymphocyte–microglia–astrocyte axis in chronic active multiple sclerosis," *Nature*, vol. 597, no. 7878, pp. 709–714, Sep. 2021, doi: 10.1038/s41586-021-03892-7.

[26] R. Rao, V. Shetty, and K. Subramaniam, "Utility of immunofluorescence in dermatology," *Indian Dermatology Online Journal*, vol. 8, no. 1, p. 1, Jan. 2017, doi: 10.4103/2229-5178.198774.

[27] C. Castro and M. Gourley, "Diagnostic testing and interpretation of tests for autoimmunity," *Journal of Allergy and Clinical Immunology*, vol. 125, no. 2, pp. S238–S247, Jan. 2010, doi: 10.1016/j.jaci.2009.09.041.

[28] G. Frazzei, R. F. Van Vollenhoven, B. A. De Jong, S. E. Siegelaar, and D. Van Schaardenburg, "Preclinical Autoimmune Disease: a Comparison of Rheumatoid Arthritis, Systemic Lupus Erythematosus, Multiple Sclerosis and Type 1 Diabetes," *Frontiers in Immunology*, vol. 13, Jun. 2022, doi: 10.3389/fimmu.2022.899372.

[29] A. Kernder *et al.*, "Delayed diagnosis adversely affects outcome in systemic lupus erythematosus: Cross sectional analysis of the LuLa cohort," *Lupus*, vol. 30, no. 3, pp. 431–438, Jan. 2021, doi: 10.1177/0961203320983445.

[30] K. Conrad, *Autoantibodies in systemic autoimmune diseases: A Diagnostic Reference*. 2007.

[31] X. Mei, B. Zhang, M. Zhao, and Q. Lu, "An update on epigenetic regulation in autoimmune diseases," *Journal of Translational Autoimmunity*, vol. 5, p. 100176, Jan. 2022, doi: 10.1016/j.jtauto.2022.100176.

[32] J. Bohacek and I. M. Mansuy, "Epigenetic inheritance of disease and disease risk," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 220–236, Jul. 2012, doi: 10.1038/npp.2012.110.

[33] M. Shome, Y. Chung, R. Chavan, J. G. Park, J. Qiu, and J. LaBaer, "Serum autoantibodyome reveals that healthy individuals share common autoantibodies," *Cell Reports*, vol. 39, no. 9, p. 110873, May 2022, doi: 10.1016/j.celrep.2022.110873.

[34] R. J. Woodman and A. A. Mangoni, "A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future," *Aging Clinical and Experimental Research*, vol. 35, no. 11, pp. 2363–2397, Sep. 2023.

[35] A. Anghel, N. Papandreou, T. Parnell, A. De Palma, and H. Pozidis, "Benchmarking and optimization of gradient boosting decision tree algorithms," *arXiv (Cornell University)*, Jan. 2018.

[36] S. Graw *et al.*, "Multi-omics data integration considerations and study design for biological systems and disease," *Molecular Omics*, vol. 17, no. 2, pp. 170–185, Dec. 2020.

[37] J. Zhao *et al.*, "Identification of clinical characteristics biomarkers for rheumatoid arthritis through targeted DNA methylation sequencing," *International Immunopharmacology*, vol. 131, p. 111860, Mar. 2024.

[38] J. Zhao *et al.*, "Identification of clinical characteristics biomarkers for rheumatoid arthritis through targeted DNA methylation sequencing," *International Immunopharmacology*, vol. 131, p. 111860, Mar. 2024.

[39] S. Rodríguez-Muguruza, A. Altuna-Coy, S. Castro-Oreiro, M. J. Poveda-Elices, R. Fontova-Garrofé, and M. R. Chacón, "A serum biomarker panel of ExomiR-451A, ExomiR-25-3P and soluble TWEAK for early diagnosis of rheumatoid arthritis," *Frontiers in Immunology*, vol. 12, Nov. 2021.

[40] J. Imgenberg-Kreuz *et al.*, "Shared and unique patterns of DNA methylation in systemic lupus erythematosus and primary Sjögren's syndrome," *Frontiers in Immunology*, vol. 10, Jul. 2019.

[41] B. Zhang *et al.*, "A simple and highly efficient method of IFI44L methylation detection for the diagnosis of systemic lupus erythematosus," *Clinical Immunology*, vol. 221, p. 108612, Oct. 2020.

[42] G. Sentis *et al.*, "A network-based approach reveals long non-coding RNAs associated with disease activity in lupus nephritis: key pathways for flare and potential biomarkers to be used as liquid biopsies," *Frontiers in Immunology*, vol. 14, Jul. 2023.

[43] E. P. Kotanidou *et al.*, "Methylation haplotypes of the insulin gene promoter in children and adolescents with type 1 diabetes: Can a dimensionality reduction approach predict the disease?," *Experimental and Therapeutic Medicine*, vol. 26, no. 4, Aug. 2023.

[44] M. V. Joglekar *et al.*, "Applicability of a microRNA-based dynamic risk score for type 1 diabetes," *Research Square (Research Square)*, Oct. 2024.

[45] M. P. Campagna *et al.*, “Whole-blood methylation signatures are associated with and accurately classify multiple sclerosis disease severity,” *Clinical Epigenetics*, vol. 14, no. 1, Dec. 2022.

[46] S. Ebrahimkhani *et al.*, “Serum Exosome MicroRNAs Predict Multiple Sclerosis Disease Activity after Fingolimod Treatment,” *Molecular Neurobiology*, vol. 57, no. 2, pp. 1245–1258, Nov. 2019.

[47] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Jan. 2001, doi: 10.1023/a:1010933404324.

[48] T. Chen and C. Guestrin, “XGBoost,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

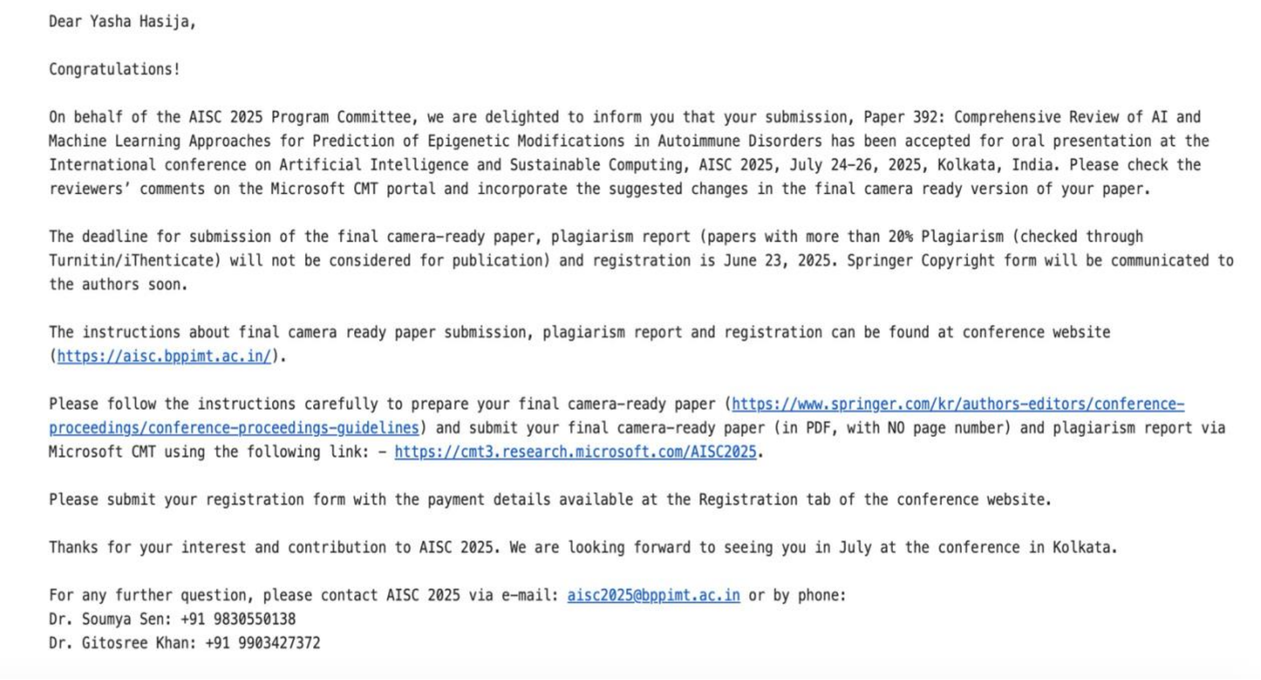
[49] R. G. Brereton and G. R. Lloyd, “Support Vector Machines for classification and regression,” *The Analyst*, vol. 135, no. 2, pp. 230–267, Dec. 2009, doi: 10.1039/b918972f.

[50] Q. Ma, “Recent applications and perspectives of logistic regression modelling in healthcare,” *Theoretical and Natural Science*, vol. 36, no. 1, pp. 185–190, Jul. 2024, doi: 10.54254/2753-8818/36/20240614.

[51] H. Duan, G. Hearne, R. Polikar, and G. L. Rosen, “The Naïve Bayes Classifier ++ for Metagenomic Taxonomic Classification—Query Evaluation,” *Bioinformatics*, vol. 41, no. 1, Dec. 2024, doi: 10.1093/bioinformatics/btae743.

[52] R. Azim, S. Wang, and S. A. Dipu, “CDSImpute: An ensemble similarity imputation method for single-cell RNA sequence dropouts,” *Computers in Biology and Medicine*, vol. 146, p. 105658, May 2022, doi: 10.1016/j.compbiomed.2022.105658.

PROOF OF PUBLICATION



Dear Yasha Hasija,

Congratulations!

On behalf of the AISC 2025 Program Committee, we are delighted to inform you that your submission, Paper 392: Comprehensive Review of AI and Machine Learning Approaches for Prediction of Epigenetic Modifications in Autoimmune Disorders has been accepted for oral presentation at the International conference on Artificial Intelligence and Sustainable Computing, AISC 2025, July 24-26, 2025, Kolkata, India. Please check the reviewers' comments on the Microsoft CMT portal and incorporate the suggested changes in the final camera ready version of your paper.

The deadline for submission of the final camera-ready paper, plagiarism report (papers with more than 20% Plagiarism (checked through Turnitin/iThenticate) will not be considered for publication) and registration is June 23, 2025. Springer Copyright form will be communicated to the authors soon.

The instructions about final camera ready paper submission, plagiarism report and registration can be found at conference website (https://aisc.bppimt.ac.in/).

Please follow the instructions carefully to prepare your final camera-ready paper (https://www.springer.com/kr/authors-editors/conference-proceedings/conference-proceedings-guidelines) and submit your final camera-ready paper (in PDF, with NO page number) and plagiarism report via Microsoft CMT using the following link: - https://cmt3.research.microsoft.com/AISC2025.

Please submit your registration form with the payment details available at the Registration tab of the conference website.

Thanks for your interest and contribution to AISC 2025. We are looking forward to seeing you in July at the conference in Kolkata.

For any further question, please contact AISC 2025 via e-mail: aisc2025@bppimt.ac.in or by phone:
Dr. Soumya Sen: +91 9830550138
Dr. Gitosree Khan: +91 9903427372

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis 'Exploring Machine Learning Approaches in Prediction of Autoimmune Disorders using Epigenetic Modifications'

Total Pages 45 Name of the Scholar: Shreya Kohli

Supervisor

Prof. Yasha Hasija
Professor and Head of Department
Department of Biotechnology
Delhi Technological University

Department of Biotechnology

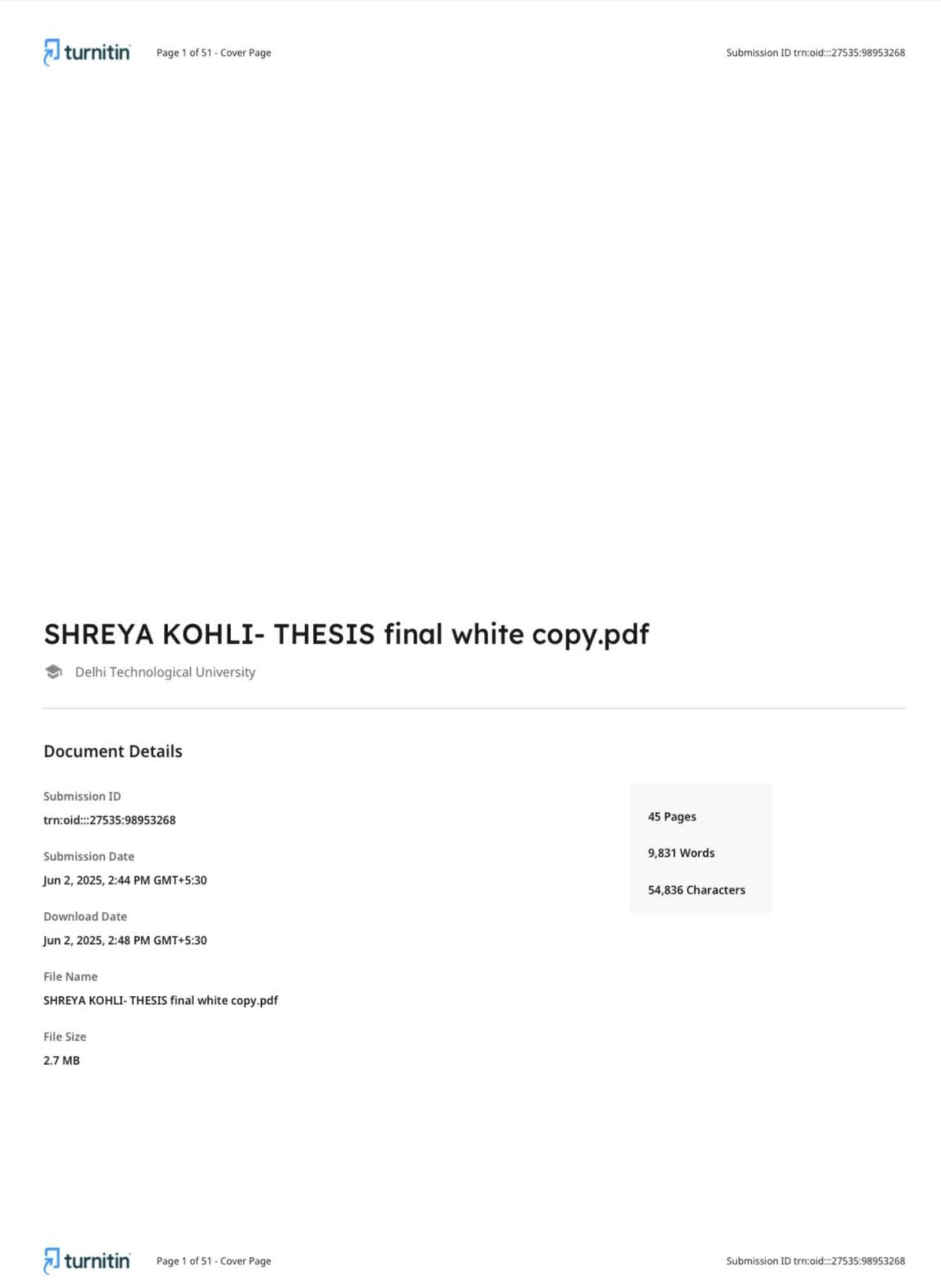
This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 8% Total Word Count: 9831

Date: 02 June 2025

Candidate's Signature **Signature of Supervisor**

PLAGIARISM AND AI REPORT



turnitin Page 1 of 51 - Cover Page Submission ID trn:oid:::27535:98953268

SHREYA KOHLI- THESIS final white copy.pdf

Delhi Technological University

Document Details

Submission ID
trn:oid:::27535:98953268

Submission Date
Jun 2, 2025, 2:44 PM GMT+5:30

Download Date
Jun 2, 2025, 2:48 PM GMT+5:30

File Name
SHREYA KOHLI- THESIS final white copy.pdf

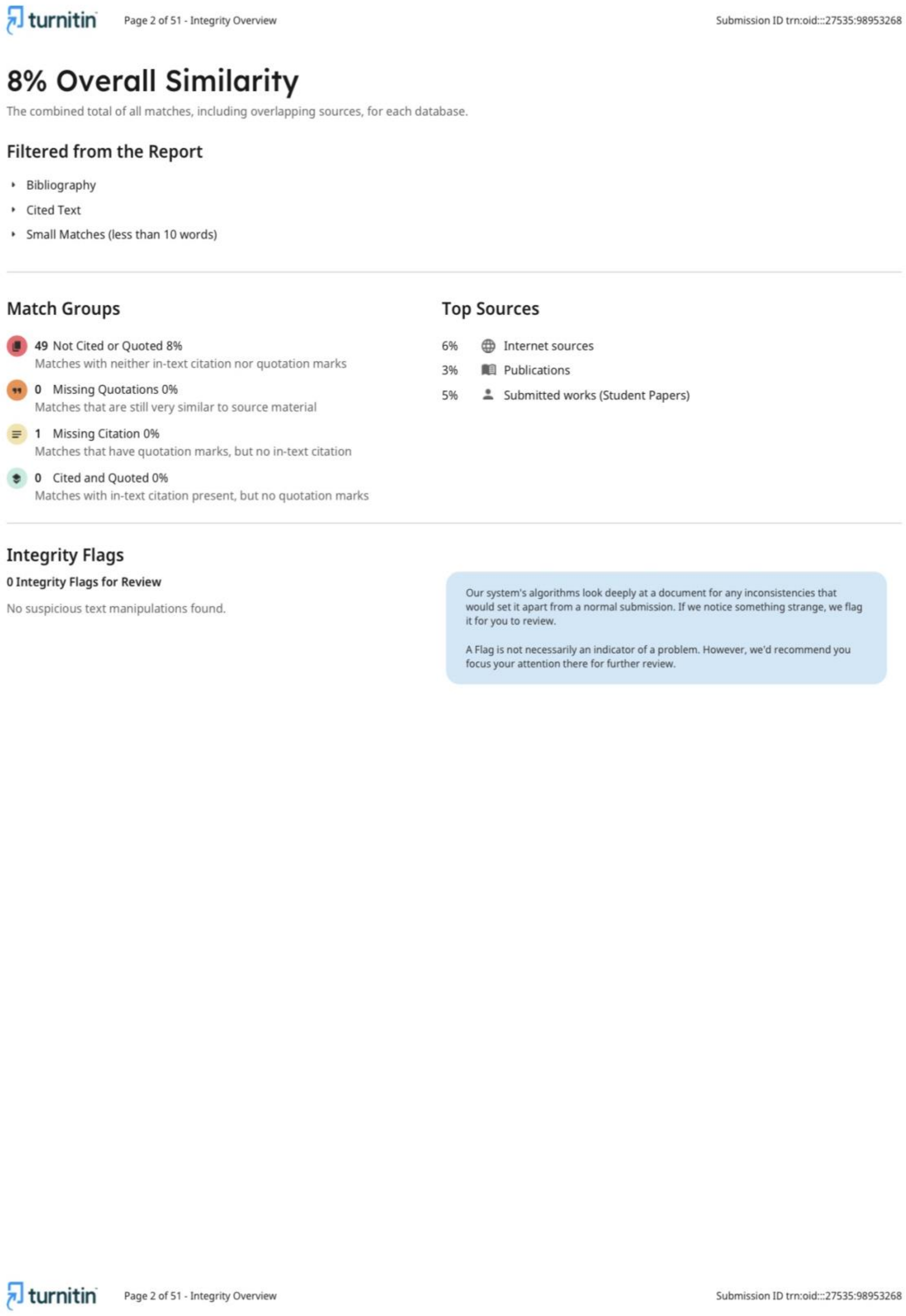
File Size
2.7 MB

45 Pages

9,831 Words

54,836 Characters

turnitin Page 1 of 51 - Cover Page Submission ID trn:oid:::27535:98953268



turnitin Page 2 of 51 - Integrity Overview Submission ID trn:oid:::27535:98953268

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 10 words)

Match Groups

- **49** Not Cited or Quoted 8%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **1** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

6%	Internet sources
3%	Publications
5%	Submitted works (Student Papers)

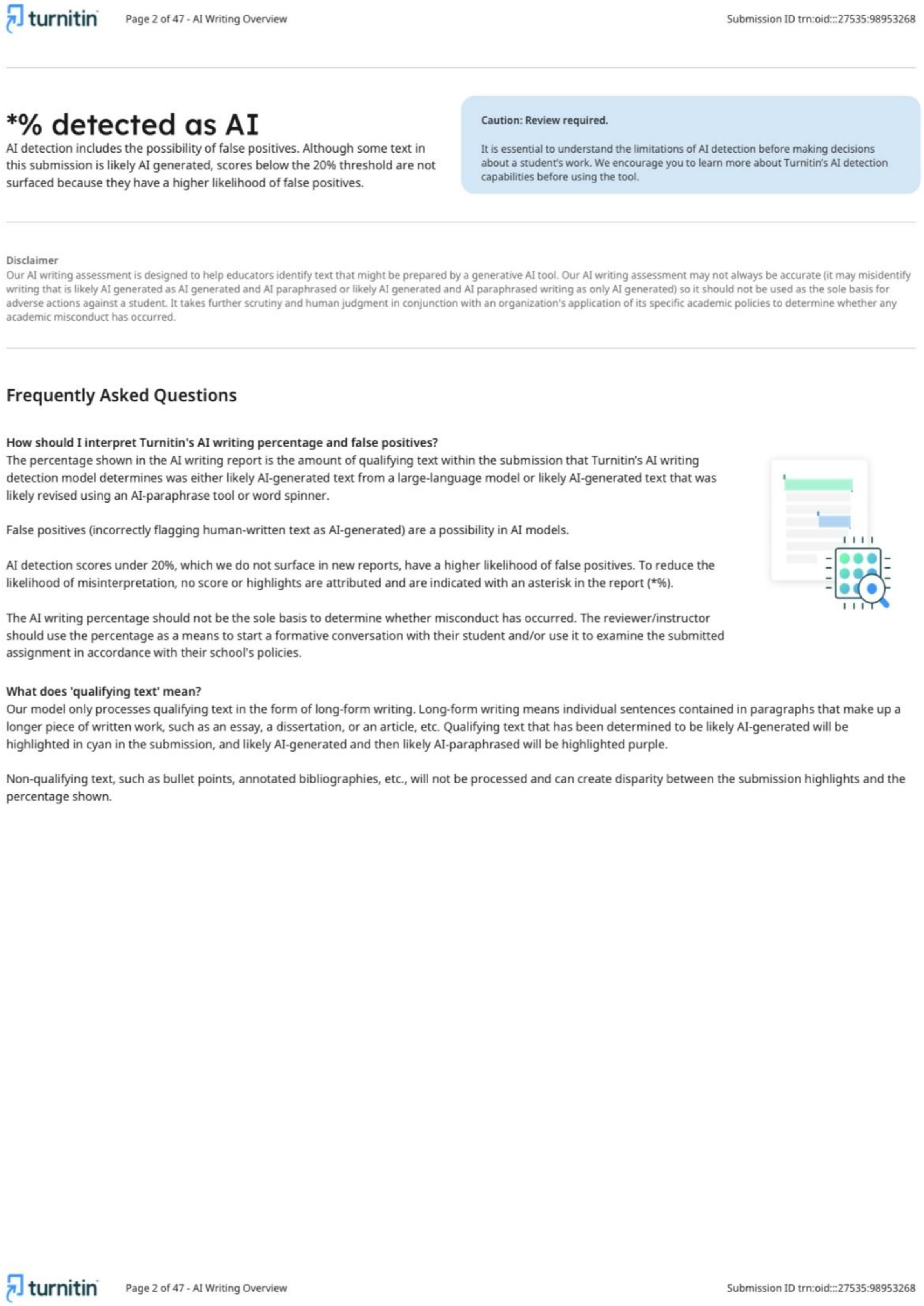
Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.



*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

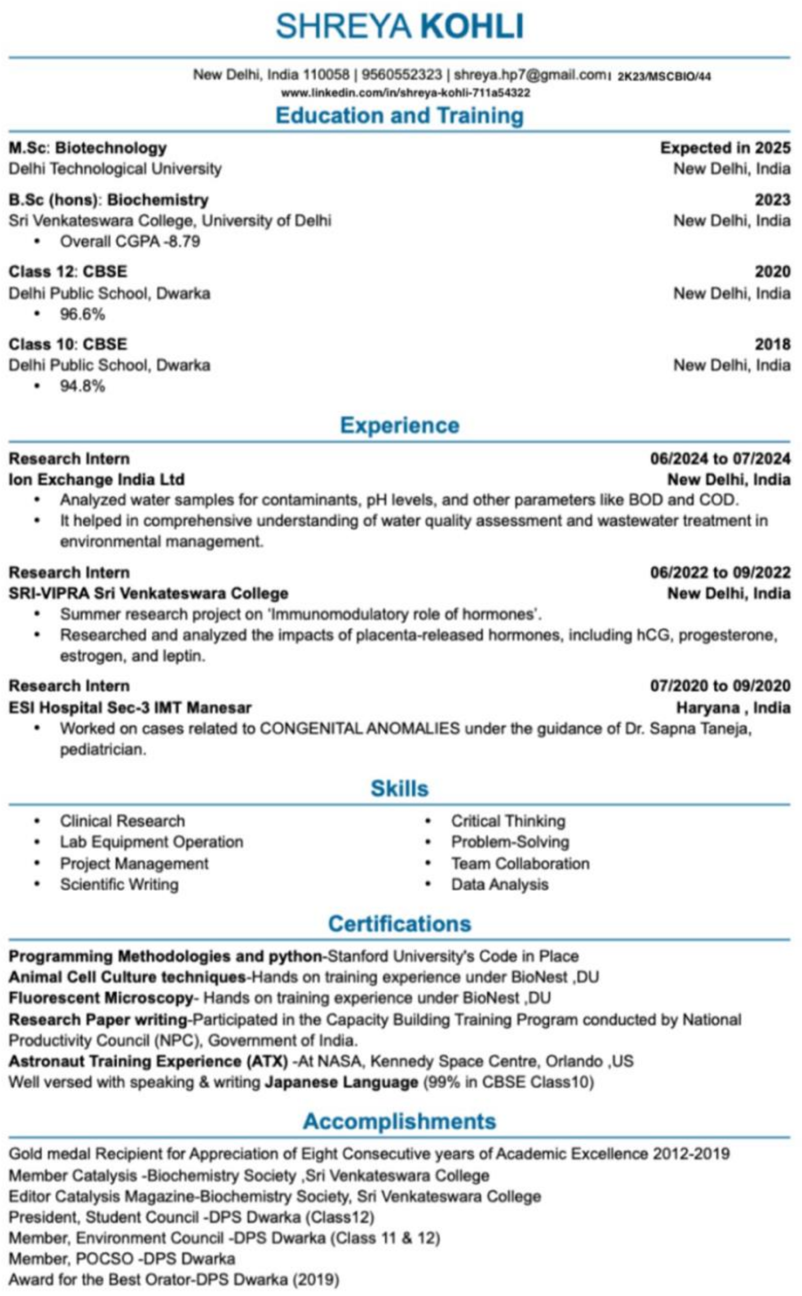
AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

CURRICULUM VITAE



SHREYA **KOHLI**

New Delhi, India 110058 | 9560552323 | shreya.hp7@gmail.com| 2K23/MSCBIO/44
www.linkedin.com/in/shreya-kohli-711a54322

Education and Training

M.Sc: Biotechnology **Expected in 2025**
Delhi Technological University New Delhi, India

B.Sc (hons): Biochemistry **2023**
Sri Venkateswara College, University of Delhi New Delhi, India

- Overall CGPA -8.79

Class 12: CBSE **2020**
Delhi Public School, Dwarka New Delhi, India

- 96.6%

Class 10: CBSE **2018**
Delhi Public School, Dwarka New Delhi, India

- 94.8%

Experience

Research Intern **06/2024 to 07/2024**
Ion Exchange India Ltd **New Delhi, India**

- Analyzed water samples for contaminants, pH levels, and other parameters like BOD and COD.
- It helped in comprehensive understanding of water quality assessment and wastewater treatment in environmental management.

Research Intern **06/2022 to 09/2022**
SRI-VIPRA Sri Venkateswara College **New Delhi, India**

- Summer research project on ‘Immunomodulatory role of hormones’.
- Researched and analyzed the impacts of placenta-released hormones, including hCG, progesterone, estrogen, and leptin.

Research Intern **07/2020 to 09/2020**
ESI Hospital Sec-3 IMT Manesar **Haryana , India**

- Worked on cases related to CONGENITAL ANOMALIES under the guidance of Dr. Sapna Taneja, pediatrician.

Skills

- Clinical Research
- Lab Equipment Operation
- Project Management
- Scientific Writing
- Critical Thinking
- Problem-Solving
- Team Collaboration
- Data Analysis

Certifications

Programming Methodologies and python-Stanford University's Code in Place
Animal Cell Culture techniques-Hands on training experience under BioNest ,DU
Fluorescent Microscopy- Hands on training experience under BioNest ,DU
Research Paper writing-Participated in the Capacity Building Training Program conducted by National Productivity Council (NPC), Government of India.
Astronaut Training Experience (ATX) -At NASA, Kennedy Space Centre, Orlando ,US
Well versed with speaking & writing **Japanese Language** (99% in CBSE Class10)

Accomplishments

Gold medal Recipient for Appreciation of Eight Consecutive years of Academic Excellence 2012-2019
Member Catalysis -Biochemistry Society ,Sri Venkateswara College
Editor Catalysis Magazine-Biochemistry Society, Sri Venkateswara College
President, Student Council -DPS Dwarka (Class12)
Member, Environment Council -DPS Dwarka (Class 11 & 12)
Member, POCSO -DPS Dwarka
Award for the Best Orator-DPS Dwarka (2019)