# BCSE497J Project-I

# GeneWise: Predictive Modelling for Genetic Disorders Using SNP-Based Feature Engineering

| | |
|---|---|
| **22BDS0049** | **TANAY SAXENA** |
| **22BDS0053** | **SHWETA SANJAY THAKUR** |

Under the Supervision of

**VIJAYASHERLY V**

Associate Professor Grade 1

School of Computer Science and Engineering (SCOPE)

**B.Tech.**

*in*

**Computer Science and Engineering**

**(with specialization in Data Science)**

## School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

September 2025

# ABSTRACT

GeneWise is a machine learning project that leverages DNA data, specifically patterns in Single Nucleotide Polymorphisms (SNPs), to predict an individual's risk of developing genetic disorders. It utilizes real-world genomic datasets to clean and process high-dimensional genetic data, extract meaningful features, and train predictive models such as Logistic Regression, Decision Trees, and Gradient Boosting. To manage the complexity of genomic data, the project employs dimensionality reduction techniques like Principal Component Analysis (PCA) and selects the most relevant genetic markers using statistical methods and domain knowledge from genomics. For interpretability, model explanation tools such as SHAP are used to highlight which genes most influence the predictions. The goal of GeneWise is to build an accurate, interpretable, and data-driven system that bridges machine learning and genomics, thereby supporting early diagnosis and enhancing understanding of inherited disease risks. This work contributes to advancements in personalized and precision healthcare.

# TABLE OF CONTENTS

# INTRODUCTION

## 1.1 Background

Genetic disorders are often caused by variations in DNA sequences, many of which can be traced to Single Nucleotide Polymorphisms (SNPs). SNPs are the most common type of genetic variation and serve as important biomarkers for disease risk prediction. With the rapid growth of genomic datasets and advances in machine learning, there is a unique opportunity to identify patterns in SNP data that can improve the early diagnosis of genetic disorders.

Traditional approaches to genetic risk assessment often rely on limited markers or family history, while computational genomics allows the integration of high-dimensional SNP datasets to uncover disease-associated variations at scale.

## 1.2 Motivations

The motivation for this project arises from the growing demand for precision healthcare solutions. Early and accurate detection of genetic risks enables timely interventions, lifestyle modifications, and personalized treatment strategies. However, analysing genomic datasets is challenging due to their high dimensionality, noise, and the complex interactions between SNPs.

Existing predictive models often struggle with interpretability, making it difficult for clinicians to trust the results. By leveraging advanced machine learning models, dimensionality reduction techniques, and interpretability frameworks such as SHAP, this project aims to provide both predictive accuracy and transparency, addressing an urgent need in medical genomics.

## 1.3 Scope of the Project

The scope of GeneWise includes developing a robust pipeline for processing genomic datasets, extracting SNP-based features, and building predictive models for genetic disorder risk assessment. The project will:

- Apply feature engineering and dimensionality reduction techniques to manage high-dimensional SNP data.
- Train machine learning models such as Logistic Regression, Decision Trees, and Gradient Boosting to predict genetic disorder risks.
- Use explainable AI methods to highlight significant SNPs and provide biological insights.
- Evaluate the models on performance metrics such as accuracy, precision, recall, and F1-score.
- The project is intended as a proof-of-concept framework that can be extended for real-world clinical decision support systems in personalized medicine.

# PROJECT DESCRIPTION AND GOALS

## 2.1 Literature Review

Machine Learning in Genomics

Advancements in high-throughput sequencing technologies have generated vast amounts of genomic data, creating both opportunities and challenges in biomedical research. Machine learning (ML) approaches have been increasingly applied to analyze such data for tasks like disease classification, biomarker discovery, and patient stratification. Early studies demonstrated that linear models, such as Logistic Regression, can identify disease-associated genetic variants by modelling binary outcomes (presence or absence of disease) from genomic features. More recent work employs ensemble techniques such as Random Forests and Gradient Boosting to capture non-linear interactions between genetic markers, improving predictive performance on complex traits [1,2]. These methods, coupled with robust preprocessing and feature engineering, have proven critical in transforming raw genomic data into clinically actionable insights.

SNP-Based Predictive Models

Single Nucleotide Polymorphisms (SNPs) represent the most abundant form of human genetic variation and play a pivotal role in disease susceptibility. Several computational studies have focused on predicting disease risk using SNP panels derived from genome-wide association studies (GWAS). Approaches range from polygenic risk scores (PRS), which aggregate weighted SNP effects, to ML-driven feature selection frameworks that identify disease-relevant variants. For example, Chen et al. [3] applied Decision Tree classifiers on SNP data to predict Type II Diabetes susceptibility, achieving high interpretability but limited scalability. Similarly, Wei et al. [4] employed Gradient Boosting Machines (GBMs) on schizophrenia SNP datasets, demonstrating improved accuracy compared to classical GWAS-based scoring. These studies highlight the potential of ML in leveraging SNP-level data but also expose challenges such as high dimensionality, redundancy among markers, and model overfitting.

Dimensionality Reduction in High-Dimensional Genetics

Genomic datasets are characterized by a “large p, small n” problem, where the number of features (SNPs) vastly exceeds the number of samples. This imbalance leads to computational inefficiency and risk of spurious correlations. To address this, dimensionality reduction techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are widely used to project SNP data into lower-dimensional subspaces that preserve genetic variation [5]. More advanced approaches, such as autoencoders and manifold learning, have also been explored to capture non-linear patterns in genomic features [6]. However, while these methods improve computational feasibility and reduce noise, they often compromise interpretability, making it difficult for clinicians to trace reduced features back to specific SNPs.

Explainable Artificial Intelligence in Healthcare

The adoption of ML in genomics faces a critical barrier: interpretability. Clinical decision-making requires transparency in how predictive models arrive at their conclusions. Explainable Artificial Intelligence (XAI) frameworks, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), have recently been applied to genomic prediction tasks to provide insights into which SNPs or genetic regions contribute most strongly to disease risk predictions [7]. For instance, Lundberg et al. [8] showed how SHAP can highlight the importance of individual genomic features in predicting patient survival outcomes, enabling biologically meaningful interpretation. These methods not only improve trust in ML models but also facilitate hypothesis generation for further biological research.

Limitations in Prior Works

Despite promising progress, several limitations persist in the literature. Many SNP-based predictive models rely heavily on feature selection heuristics or dimensionality reduction methods that may obscure biologically relevant markers. Scalability also remains a challenge, as models trained on one cohort may fail to generalize to diverse populations. Furthermore, while XAI tools have introduced interpretability, integrating them seamlessly with genomic pipelines is still an evolving practice. These challenges highlight the need for frameworks that combine robust feature engineering, predictive accuracy, and interpretability, which forms the basis for the GeneWise project.

## 2.2 Research Gaps

From the existing body of work in SNP-based predictive modelling and genomic machine learning, the following key research gaps have been identified:

- **High-dimensionality of genomic data**:
  Most existing studies struggle with the “large p, small n” problem, where the number of SNP features vastly exceeds the available sample size. This often leads to overfitting or reliance on aggressive dimensionality reduction that may discard biologically relevant markers.
- **Limited generalizability across populations**:
  Many predictive models are trained on specific cohorts (e.g., European ancestry) and fail to generalize effectively to genetically diverse populations, reducing clinical applicability.

- **Inadequate integration of feature selection and interpretability:**
  While some works employ statistical methods for SNP selection, they often lack interpretability frameworks to explain why certain markers influence predictions, making the models less useful in medical decision-making.
- **Trade-off between accuracy and interpretability:**
  Models with high predictive accuracy (e.g., ensemble methods) are frequently black-box in nature, while interpretable models (e.g., decision trees, logistic regression) often compromise on accuracy. Achieving both simultaneously remains a challenge.
- **Scalability of computational approaches:**
  SNP datasets are extremely large, requiring scalable preprocessing and training pipelines. Many prior works have not sufficiently addressed the computational efficiency required for real-world clinical deployment.
- **Lack of comprehensive workflows:**
  Existing studies often focus on isolated tasks (feature selection, prediction, or interpretation) without integrating them into a robust, end-to-end pipeline that can support clinical genomics applications.

## 2.3 Objectives

The primary objective of the *GeneWise* project is to design and implement a predictive modelling framework that leverages SNP-based feature engineering and machine learning to assess genetic disorder risks. To achieve this overarching goal, the following specific objectives are defined:

- **Data Acquisition and Preprocessing**
  - Collect and curate publicly available genomic datasets containing SNP profiles and disease phenotypes.
  - Apply preprocessing techniques such as data cleaning, normalization, and encoding of SNP variants for ML compatibility.
- **Feature Engineering and Dimensionality Reduction**
  - Implement statistical and machine learning–based feature selection methods to identify disease-relevant SNPs.
  - Apply dimensionality reduction techniques (e.g., PCA) to manage high-dimensional data while retaining maximum variance.
- **Predictive Modelling**
  - Train and evaluate multiple machine learning algorithms (Logistic Regression, Decision Trees, Gradient Boosting) on SNP data.
  - Optimize hyperparameters to achieve high predictive accuracy, precision, recall, and F1-score.
- **Interpretability and Explainability**
  - Integrate explainable AI frameworks such as SHAP to provide transparent model outputs.

    - Highlight the most influential SNPs contributing to genetic disorder predictions for biological validation.
- **System Integration and Workflow Development**
    - Develop an end-to-end pipeline encompassing data preprocessing, feature engineering, model training, evaluation, and interpretation.
    - Ensure the pipeline is modular, reproducible, and extensible for future integration with larger genomic datasets.
- **Evaluation and Benchmarking**
    - Compare the proposed system's performance with baseline models (e.g., traditional GWAS-based risk scoring).
    - Assess the clinical utility of the framework by analyzing its interpretability and scalability for real-world applications.

## 2.4 Problem Statement

The rapid expansion of genomic data, particularly SNP datasets, has created new opportunities for understanding genetic contributions to disease. However, the inherent high dimensionality of SNP data, coupled with noise and redundancy, poses significant computational and statistical challenges. Traditional genome-wide association studies (GWAS) have identified disease-linked SNPs, but their predictive power is often limited when applied to individual-level risk assessment.

Existing machine learning models have shown promise in leveraging SNP data for disease prediction, yet they face critical limitations. Highly accurate models, such as ensemble methods, often function as "black boxes," offering little insight into the biological basis of predictions. Conversely, interpretable models sacrifice accuracy, limiting their usefulness in clinical contexts. Furthermore, many models are trained on homogeneous cohorts, reducing their generalizability to diverse populations and restricting their clinical applicability.

This creates a research gap in developing predictive frameworks that balance accuracy, interpretability, and scalability. A robust system that integrates feature engineering, dimensionality reduction, advanced ML models, and explainability tools is essential to transform SNP data into clinically meaningful predictions. Addressing this problem is vital to enable early detection of genetic disorders and to support the broader goal of precision and personalized healthcare.

## 2.5 Project Plan

The development of *GeneWise* follows a structured project plan designed to ensure systematic progress from data collection to predictive modelling and interpretability. The project is divided into distinct phases, each with well-defined tasks and deliverables:

1. **Phase I – Literature Review and Problem Understanding**

   - Conduct an extensive survey of existing research in SNP-based prediction and machine learning in genomics.
   - Identify research gaps and define project objectives.
   - Finalize tools, datasets, and methodologies to be used.

2. **Phase II – Data Acquisition and Preprocessing**

   - Acquire publicly available genomic datasets (SNP data with phenotype/disease labels).
   - Perform preprocessing including data cleaning, missing value handling, encoding SNPs, and normalization.
   - Partition data into training, validation, and testing sets.

3. **Phase III – Feature Engineering and Dimensionality Reduction**

   - Apply statistical tests and filter-based methods to identify significant SNPs.
   - Use PCA or similar methods to reduce dimensionality while retaining informative variation.
   - Document feature importance and prepare datasets for modelling.

4. **Phase IV – Predictive Modelling**

   - Train baseline ML models (Logistic Regression, Decision Trees).
   - Extend to advanced models (Gradient Boosting, Random Forest).
   - Tune hyperparameters and compare performance using accuracy, precision, recall, and F1-score.

5. **Phase V – Model Interpretability and Explainability**

   - Apply SHAP (Shapley Additive Explanations) to analyse SNP importance.
   - Generate visualizations to explain predictions at both global and individual levels.
   - Validate interpretability findings with biological relevance.

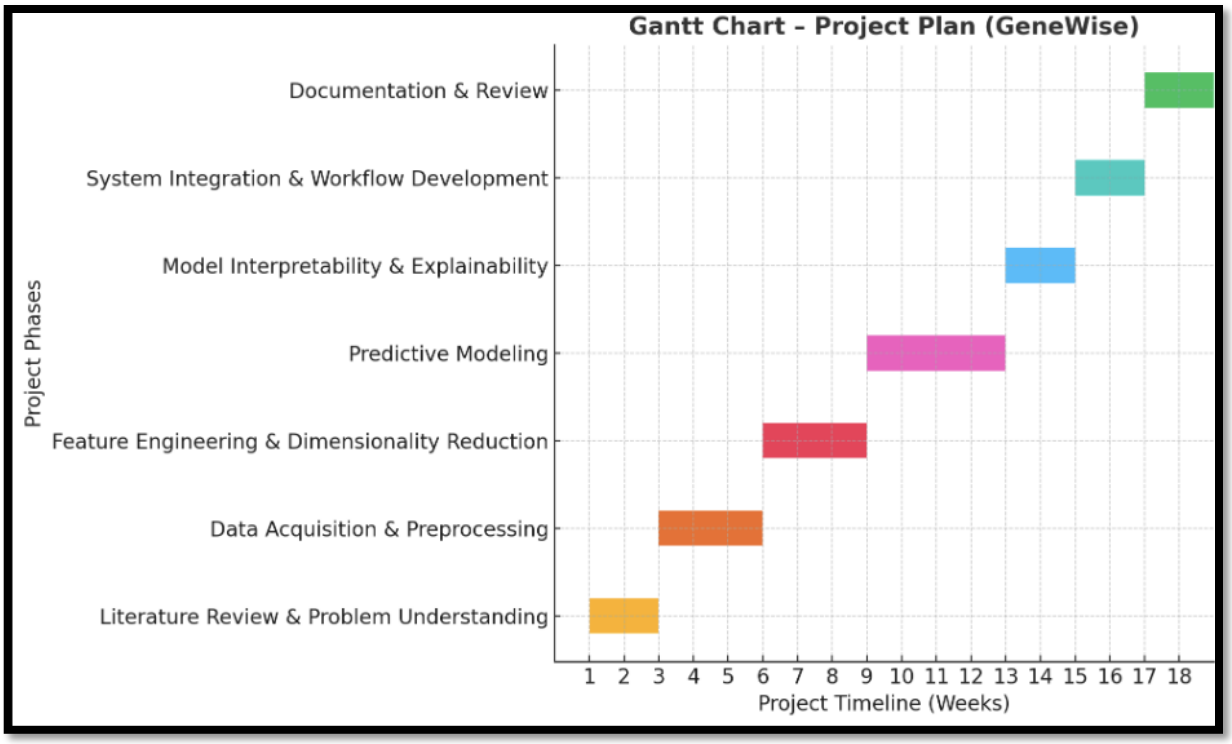
6. **Phase VI – System Integration and Workflow Development**

- Develop an end-to-end pipeline that integrates preprocessing, feature engineering, model training, evaluation, and interpretation.
- Ensure modularity and reproducibility for extension to larger datasets.

### 7. Phase VII – Documentation and Review

- Prepare Review-2 report with results, workflows, and implementation progress.
- Submit hard copy of report as per HOD's instructions.
- Incorporate feedback for further refinement.

## Gantt Chart – Project Plan (GeneWise)



## Work Breakdown Structure – GeneWise

| Phase | Tasks / Subtasks |
|---|---|
| **Phase I: Literature Review & Problem Understanding** | • Survey existing SNP-based ML models • Identify research gaps • Finalize tools, datasets, and methodology |
| **Phase II: Data Acquisition & Preprocessing** | • Collect genomic SNP datasets • Clean data and handle missing values • Encode SNPs (numerical format) • Normalize dataset • Split into training/validation/test sets |
| **Phase III: Feature Engineering &** | • Statistical SNP selection (Chi-square, ANOVA) • Dimensionality reduction (PCA) • Prepare reduced feature set for modeling |

| | |
|---|---|
| **Dimensionality Reduction** | |
| **Phase IV: Predictive Modeling** | • Train baseline models (Logistic Regression, Decision Trees) • Extend to advanced models (Gradient Boosting, Random Forest) • Hyperparameter tuning and cross-validation • Evaluate metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC) |
| **Phase V: Model Interpretability** | • Apply SHAP for global & local feature importance • Visualize influential SNPs • Validate biological significance |
| **Phase VI: System Integration & Workflow Development** | • Build end-to-end pipeline integrating all modules • Ensure modularity and reproducibility • Automate reporting and visualization |
| **Phase VII: Documentation & Review** | • Prepare Review-2 report (as per HOD template) • Submit hard copy to review panel • Incorporate feedback for refinement |

# REQUIREMENT ANALYSIS (SRS)

## 3.1 Functional Requirements

The *GeneWise* system should provide the following core functionalities:

1. **Data Acquisition and Preprocessing**

   - Import SNP-based genomic datasets from publicly available repositories.
   - Handle missing values, perform normalization, and encode categorical SNP variants.
   - Partition datasets into training, validation, and testing sets.

2. **Feature Engineering and Dimensionality Reduction**

   - Select statistically significant SNPs using domain-informed methods.
   - Apply dimensionality reduction techniques (e.g., PCA) to manage data sparsity.
   - Retain biologically relevant features for downstream modelling.

3. **Predictive Modelling**

   - Train and evaluate machine learning models such as Logistic Regression, Decision Trees, and Gradient Boosting.
   - Perform hyperparameter tuning to optimize model performance.
   - Generate performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

4. **Model Interpretability**

    - Apply SHAP to quantify feature (SNP) importance.
    - Generate visual explanations to support clinician understanding.
    - Provide both global (overall feature influence) and local (per-patient prediction) interpretations.

5. **System Integration and Reporting**

    - Integrate preprocessing, modelling, and explainability into an end-to-end pipeline.
    - Allow automated reporting of results and feature importance.
    - Provide outputs in a structured format suitable for further validation or clinical research.

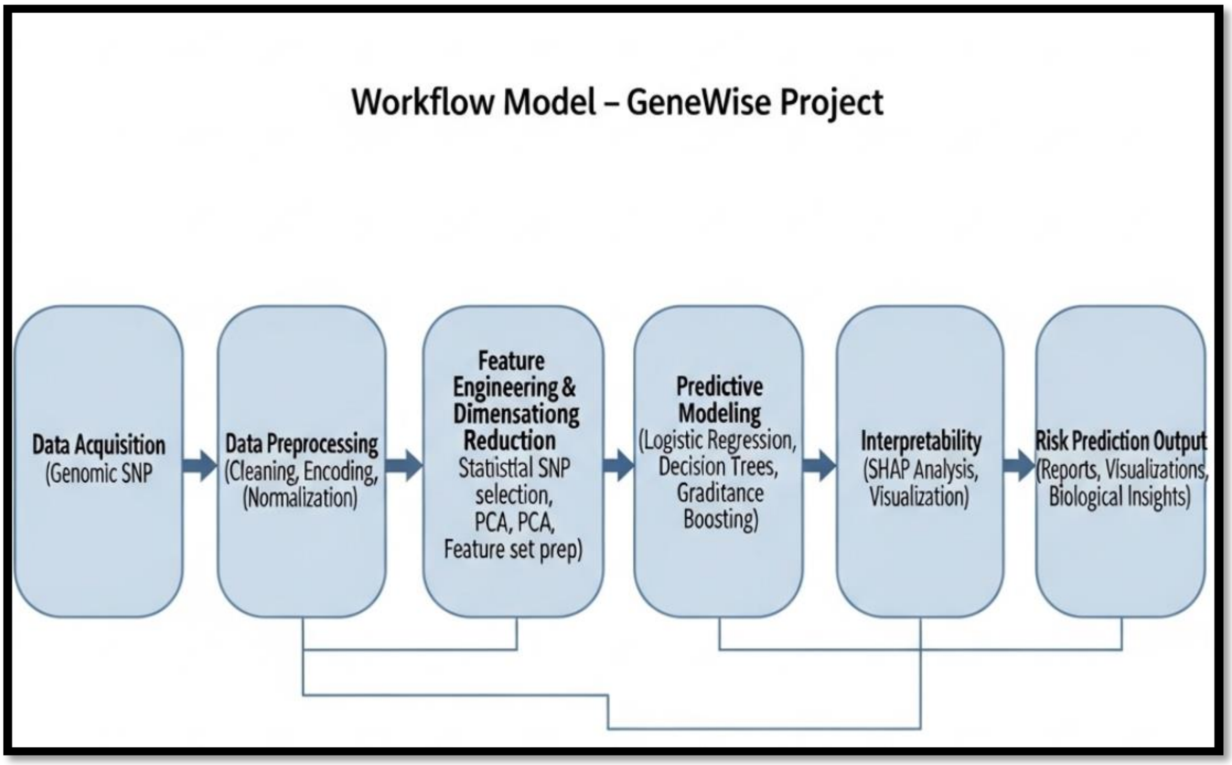
## 3.2 Non-Functional Requirements

The system should also satisfy key non-functional requirements to ensure robustness and usability:

- **Performance:** The models should train and infer within reasonable computational limits, considering high-dimensional SNP datasets.
- **Scalability:** The pipeline must be extensible to larger genomic datasets without major re-engineering.
- **Accuracy:** Prediction accuracy should be benchmarked against baseline models (GWAS-based risk scores).
- **Interpretability:** Model outputs should be explainable, enabling clinicians to understand genetic risk predictions.
- **Reproducibility:** Results should be reproducible across runs, ensured via controlled randomness and proper documentation.
- **Usability:** The system workflow should be modular, allowing researchers to modify individual components (e.g., model choice, feature selection) easily.
- **Reliability:** The system should provide consistent results under repeated experiments.
- **Data Privacy:** While working on publicly available datasets, ensure that any sensitive genomic information is handled in compliance with ethical guidelines.

| Category | Specification |
|---|---|
| **Hardware** | |
| Processor | Intel Core i5/i7 (10th Gen or above) / Equivalent AMD Ryzen |
| RAM | Minimum 8 GB (16 GB recommended for large genomic datasets) |
| Storage | 256 GB SSD (minimum), 512 GB preferred for handling genomic data files |
| GPU (Optional) | NVIDIA GPU with CUDA support (e.g., GTX 1660 or higher) for accelerated ML tasks |
| Operating System | Windows 10/11, Linux (Ubuntu 20.04+), or macOS Monterey+ |
| **Software** | |
| Programming Language | Python 3.9+ |
| Libraries/Frameworks | NumPy, Pandas, Scikit-learn, Matplotlib/Seaborn, SHAP, XGBoost/LightGBM |
| Development Environment | Jupyter Notebook / Google Colab / VS Code |
| Version Control | Git & GitHub for collaboration and version management |
| Documentation Tools | Microsoft Word/LaTeX for reports, draw.io / Lucid chart for workflow diagrams |

## Workflow model


Workflow Model – GeneWise Project
Data Acquisition (Genomic SNP
Data Preprocessing (Cleaning, Encoding, (Normalization)
Feature Engineering & Dimensationg Reduction Statistial SNP selection, PCA, PCA, Feature set prep)
Predictive Modeling (Logistic Regression, Decision Trees, Graditance Boosting)
Interpretability (SHAP Analysis, Visualization)
Risk Prediction Output (Reports, Visualizations, Biological Insights)

# MODULE DESIGN AND IMPLEMENTATION

The *GeneWise* system is structured into modular components, each addressing a specific stage of the predictive modelling pipeline. The modular design ensures clarity, reproducibility, and extensibility for future enhancements. The current implementation has achieved significant progress, with approximately 40–50% of the pipeline completed.

## 4.1 Module Design

1. **Data Acquisition Module**
    - Handles loading of genomic SNP datasets from public repositories.
    - Ensures proper formatting of SNP matrices and phenotype labels.
2. **Data Preprocessing Module**
    - Performs missing value handling (e.g., mode imputation for SNP values).
    - Encodes SNP variations (AA, AT, TT → numerical encoding).
    - Applies normalization and data partitioning (train/validation/test).
3. **Feature Engineering & Dimensionality Reduction Module**
    - Implements statistical feature selection (chi-square test, ANOVA).
    - Integrates dimensionality reduction (Principal Component Analysis).
    - Retains biologically meaningful features for downstream tasks.
4. **Interpretability Module (Planned)**
    - SHAP framework for local and global interpretability.
    - Visualization of influential SNPs for biological insight.
5. **System Integration & Reporting Module (Planned)**
    - Integrates all modules into an end-to-end workflow.
    - Generates structured reports and visualizations.

## 4.2 Implementation Progress

- **Completed (~40%)**
    - Dataset preprocessing pipeline has been implemented, including SNP encoding and normalization.

  - Feature engineering using chi-square–based SNP filtering is functional.
  - Baseline predictive models (Logistic Regression, Decision Tree) have been trained and evaluated on test splits.
  - Initial results demonstrate promising accuracy, establishing feasibility.

- **In Progress (~30%)**

  - Gradient Boosting and Random Forest models are being integrated and tuned.
  - PCA-based dimensionality reduction is under testing for scalability.

- **Planned (~30%)**

  - SHAP-based interpretability integration.
  - End-to-end pipeline assembly.
  - Automated reporting and visualization generation.

## REFERENCE

1. **Ma K et al. (2025)** — *Integrating explainable machine learning and transcriptome data to identify genes distinguishing AMD (Age-related Macular Degeneration) controls and cases, with AUC-ROC of 0.80.*
   Nature
2. **Wang J et al. (2025)** — *Comprehensive review of ML model development, hyperparameter optimization, and interpretability across genomics (2011–2024).*
   ScienceDirect
3. **Wang C. et al. (2025)** — *Deep Learning and Explainable AI: new pathways to gain genetic insights; categorizes interpretability methods (input/model-based), critiques their limitations, and provides research directions.*
   arXiv
4. **Alme C. et al. (2025)** — *Applies deep learning (FFNs, CNNs, autoencoders) to SNP-based disease prediction using standardized preprocessing; demonstrates benefits of DL architectures across multiple GEO datasets.*
   SpringerLink
5. **Ye L. et al. (2025)** — *Introduces Ge-SAND: an explainable deep-learning framework using self-attention to model large-scale SNP interactions, achieving up to 20 % boost in AUC-ROC on Crohn's, schizophrenia, and Alzheimer's datasets.*
   BioMed Central
6. **Aljarallah N.A. et al. (2024)** — *Systematic review of genetics- and molecular-pathway-based ML models for neurological and speech disorder diagnosis, highlighting feature engineering, classification techniques, and limitations.*
   MDPI
7. **Ma J. et al. (2024)** — *Application of interpretable machine learning in computational biology; covers key advances, pitfalls, and recommendations.*
   Wikipedia

8. **Nunkesser R. (2025)** — *Presents highly interpretable prediction models for SNP data (logicFS, GPAS, logicDT), focusing on model interpretability and automatic model selection in genetic association tasks.*
SciTePress
9. **Wagle M.M. et al. (2024)** — *Review of interpretable deep learning applications in single-cell omics; discusses model limitations and future interpretability challenges.*
arXiv
10. **Gunter N. et al. (2023)** — *Proposes methods to encode pairwise SNP epistatic interactions into features to improve interpretability and preserve constituent information.*