

**AMRITA SCHOOL OF ARTIFICIAL ENGINEERING**

**AMRITA VISHWA VIDYAPEETHAM**

COIMBATORE - 641 112

**April - 2025**

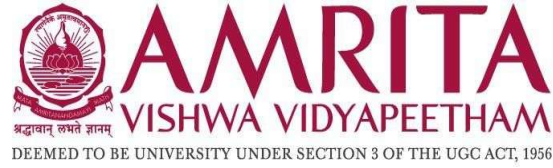
**B.TECH ARTIFICIAL INTELLIGENCE IN DATA SCIENCE  
AND MEDICAL ENGINEERING**

**AI-BASED PREDICTIVE MODEL FOR GENETIC DISEASE  
RISK VISUALIZATION**

**24AIM112 Molecular Biology & Basic Cellular Physiology**

**24AIM115 Ethics, Innovative Research Business & IPR**

**AMRITA VISHWA VIDYAPEETHAM**  
COIMBATORE - 641 112



**BONAFIDE CERTIFICATE**

This is to certify that the report entitled “**AI-Based Predictive Model for Genetic Disease Risk Visualization**” submitted by:

<b>Dharshini K.</b>	<b>CB.AI.UAIM24111</b>
<b>Esha R.</b>	<b>CB.AI.UAIM24112</b>
<b>Harsshitha S.</b>	<b>CB.AI.UAIM24115</b>
<b>Vaishnavi P.</b>	<b>CB.AI.UAIM24149</b>

for the final project of 2<sup>nd</sup> semester in **B.TECH ARTIFICIAL INTELLIGENCE IN DATA SCIENCE AND MEDICAL ENGINEERING** is a Bonafide record of the work carried out at Amrita School of Artificial intelligence, Coimbatore.

*Submitted for the final evaluation on 23-04-25*

**FACULTY:**

**Mrs. Reshma Sanal**

**FACULTY:**

**Dr. Neelesh Ashok**

## CONTENTS

1.Abstract .....	4
2.Introduction.....	4-6
3.Literature review.....	6-10
4.Methodology.....	10-12
5.Results .....	13-20
6.Discussions.....	21
7.Conclusion.....	22
8.References.....	22-23

# AI-Based Predictive Model for Genetic Disease Risk Visualization

## 1. Abstract:

Genetic disorders are the result of mutation in the deoxyribonucleic acid (DNA) sequence which can develop or inherited from parents. The predictions can be made by knowing all 3 billion DNA sequences which is humanly impossible, thereby it will lead to human error and wrong predictions. In order to overcome this many machine learning (ML) models have been developed to predict the disease more efficiently. The potential of such models can be utilized to predict genetic disorders at an early stage using the genome data for timely treatment. In our project, using the genome data we have predicted various genetic diseases: Bipolar Disorder, Congenital Heart Disease, Glaucoma, and Parkinson's Disease. We have used XGBoost algorithm for model training and have achieved an accuracy of 96% with this a user-interface dashboard was developed to display the predicted disease and its associated risk graphically. The user data is saved in the database. Thus, the developed and designed user-interface dashboard can effectively and efficiently handle the databases to detect patterns and future prediction diagnosis of the specified genetic diseases.

## 2. Introduction:

Genetic disorders are often rare and notoriously difficult to diagnose. On average, it takes between five and 10 years from the onset of symptoms to pinpoint the exact genetic cause of a rare condition. The long and arduous diagnostic journey often delays treatment, and it typically ends up being costly and isolating[1]. Genome-wide association studies (GWAS) have fundamentally changed how we understand the genetic basis of complex diseases. By scanning the entire genome for small variations, especially single nucleotide polymorphisms (SNPs), GWAS allows researchers to pinpoint genetic differences that are more common in people with a particular disease compared to those without it. These studies generate key metrics such as risk alleles, allele frequencies, p-values, and odds ratios that help quantify the strength and significance of these genetic associations [2].

### 2.1 What Are Genetic Disorders?

Genetic disorders happen when there are changes in our DNA, either passed down by parents or developed on their own. Some disorders follow clear inheritance patterns, while others are much harder to pin down, especially when many genes and lifestyle factors are

involved.

## 2.2 Challenges Of Traditional Diagnostic Methods

Looking at DNA manually or gene-by-gene is time-consuming and not always effective. Traditional methods of gene-by-gene analysis are often insufficient to capture these intricate patterns, particularly when dealing with polygenic disorders or multifactorial diseases where multiple genetic and environmental factors interplay.

## 2.3 Artificial Intelligence For Disease Prediction

Machine Learning models are great for identifying the hidden patterns in large datasets like in genetics. They can predict disease risks based on genetic markers alongside other patient information (like age, height, and weight) to predict the likelihood of a disease.

In our project, we used the XGBoost algorithm, which works well with structured data, to train a model that predicts genetic diseases with high accuracy.

## 2.4 Dataset Source

For our project, we downloaded the dataset from the Genome-Wide Association Studies (GWAS) website. GWAS datasets provide comprehensive genetic data that allows researchers to identify genetic variants associated with various diseases. The key genetic risk indicators in our model include:

- **Risk Allele:** A genetic variant associated with increased disease susceptibility. It signifies a higher likelihood of developing a condition when present in an individual's genotype.
- **Risk Frequency:** The proportion of individuals in a population who carry a specific risk allele, indicating how common that allele is in the studied group.
- **P-Value:** A statistical measure that quantifies the probability of obtaining the observed data (or more extreme results) under the null hypothesis (i.e., no true association).
- **Odds Ratio (OR):** A measure of the strength of association between a risk allele and disease, showing how much more likely individuals with the allele are to develop the disease compared to those without it.

## 2.5 Genetic Factors Influencing Major Disorders

- **Parkinson's Disease**

Parkinson's disease is linked to mutations in genes like *LRRK2* and *SNCA*, which are involved in the regulation of dopamine-producing neurons. These genetic alterations lead to neurodegeneration, causing motor dysfunction and other neurological symptoms.

- **Congenital Heart Disease (CHD)**

Congenital Heart Disease is often caused by mutations in genes such as *NKX2-5* and *GATA4*, which are essential for proper heart development. These mutations can disrupt the normal formation of heart structures during embryonic development, leading to various congenital heart defects.

- **Bipolar Disorder**

Bipolar disorder has a strong genetic component, with variations in genes like *ANK3* and *CACNA1C* playing a significant role in regulating mood and neuronal function. These genetic variants influence brain signaling pathways, contributing to the onset of mood swings. Bipolar disorder is characterized by extreme mood swings, with two distinct phases: mania and depression.

- **Glaucoma**

Glaucoma, a condition that damages the optic nerve, is associated with mutations in the *MYOC* gene. These mutations affect intraocular pressure regulation and the function of the trabecular meshwork in the eye, increasing the risk of optic nerve damage and vision loss.

### **3.Literature Review:**

CardioRiskNet represents a significant advancement in cardiovascular risk prediction by leveraging AI's adaptability and interpretability to overcome the limitations of conventional methods, potentially enabling earlier and more personalized interventions in cardiovascular care. Key contributions of the paper include the introduction of this hybrid model that combines XAI, active learning, and attention mechanisms, outperforming

traditional risk assessment tools and providing a powerful, interpretable tool for healthcare professionals to manage CVD risk more effectively. The model was evaluated on a real-world Heart Failure Clinical Records Dataset and demonstrated superior performance metrics: 98.7% accuracy, 98.7% sensitivity, 99% specificity, and an F1-Score of 98.7%. These results indicate that CardioRiskNet not only improves prediction accuracy but also provides explainable insights, enhancing clinical utility [3].

This research makes a significant contribution to the field of genetic disorder prediction by effectively leveraging ensemble feature extraction and classifier chains to handle complex multi-label problems. The authors propose a unique approach of combining class probabilities from Extra Trees (ET) and Random Forest (RF) classifiers to create an enriched feature set, which enhances model training and predictive power. The XGB model achieved the best results, with a 92%  $\alpha$ -evaluation score and 84% macro accuracy, outperforming several state-of-the-art approaches in both predictive accuracy and computational efficiency. The approach balances predictive performance with computational efficiency, making it a promising tool for early diagnosis and treatment planning in genomic medicine [4].

The research presents a valuable contribution to the field of cervical cancer diagnosis by introducing CHAMP, a web-based tool that leverages machine learning to improve risk assessment. The study employs several machine learning algorithms, including XGBoost, SVM, Naive Bayes, AdaBoost, Decision Tree, and K-Nearest Neighbors, to analyze cervical cancer databases and identify patterns for future prediction. The tool is built using Python 3.9.0 with Flask, focusing on user interaction and accessibility. CHAMP is designed to handle cervical cancer databases efficiently, detect patterns for future diagnosis, evaluate and optimize processes, and provide personalized and intuitive data analysis for informed decision-making. Ada-Boost 97.3%, K-Nearest Neighbour 98.3%, Naïve Bayes 24.6 %, Support Vector Machine 97.2%, XGBoost 97.5%, Decision Tree 97.8% of accuracies have been achieved [5].

The paper introduces a promising machine-learning model that effectively integrates genetic and non-genetic factors to improve disease risk prediction, demonstrating superior performance compared to existing methods for CAD and T2D. PRSIMD applies posterior regularization, using Kullback-Leibler (KL) divergence as a loss function to align the posterior distribution with prior knowledge. Mendelian Randomization is used to identify

causal non-genetic risk factors. Evaluated on coronary artery disease (CAD) and type 2 diabetes (T2D) using the UK Biobank, PRSIMD outperformed existing PRS algorithms, showing an average 24% increase in AUROC. It also outperformed composite risk scores (CRS) that combined PRS and non-genetic factors, increasing AUROC by at least 0.013 for CAD and 0.035 for T2D [6].

The prediction of genetic disorders is given special emphasis in this study on the various machine learning models used, with commendable degrees of accuracy. A Deep Learning-Powered Advanced Genome Disorder Prediction Model has given 91% accuracy, whereas Random Forest and XG Boost gave accuracy of 88%. Preprocessing operations done in the study included the handling of missing values and outlier detection. The study emphasizes the inaccessibility of healthcare due to challenges posed by a growing global population and further increases in genetic disorders. Future research attempts to enlarge the classification by the inclusion of more genetic variations, along with advanced genetic tests and clinical data to formulate better diagnostic tools [7].

This study talks about the comparison of gene expression data between schizophrenia and bipolar disorder using ROC curves, emphasizing the performance of different algorithms such as SVM, NSC, and others. It gives graphical plots of true and false positive rates for both conditions, as well as p-value histograms that indicate distribution biases in test outcomes. Major findings show expression profile differences, with certain genes having high p-values, indicating their possible significance in explaining these mental illnesses [8].

The purpose of this study was to use machine learning algorithms, namely SVM and KNN, to predict diseases like cancer, dementia, and diabetes or for that matter any disease that may have a genetic link or history of the patient. The proposed model achieved a testing accuracy of classification of 92.5% with SVM, endorsing the utility of putting together multifactorial genetic data. The study puts forward the possible application for early prediction of diseases to better patient outcomes and underlines the urgent need for continual improvement and refinement of such models with more genetic data input and improved algorithms, perhaps even transfer learning [9].



The study presents the ethical and regulatory issues surrounding polygenic risk scores (PRSs) and AI-driven PRS in medicine. It calls for standardizing regulatory frameworks, contrary to the prevailing disarray in the regulations. Among ethical challenges are informed consent, explainability, and the role of ethics committees in the governance of AI technologies. The study also emphasizes incorporating bioethics education into the training of healthcare professionals so they can consider social implications and foster transparency in AI applications. The authors emphasize the urgent need for cooperation among all stakeholders in confronting the AI challenge in medicine [10].

The article elaborates on the ethics of the emergence of artificial intelligence (AI) in genomic medicine, particularly during pregnancy and pediatrics. The paper has identified the benefits as well as the risks that may accrue from the use of AI here. Although AI can manage genomic data, areas of concern may lie around trust, decision-making and patient autonomy in delicate environments such as recognizing fetal abnormalities and those of critically ill children. Ethical principles such as non-maleficence and transparency are recommended. AI tools should complement human judgment and thus, accountability can only rest on healthcare practitioners and institutions [11].

Former genome-wide polygenic score studies have not effectively stratified risk for common diseases. However, with larger datasets and sophisticated algorithms, current improvements are proving successful in identifying at-risk subpopulations for diseases like coronary artery disease and type 2 diabetes at levels comparable to rare monogenic mutations. The study validated GPSs in the UK Biobank, demonstrating significant risk stratification potential and proposing integration of polygenic risk assessments into clinical care in order to improve management and treatment strategies for patients through genetic predisposition [12].

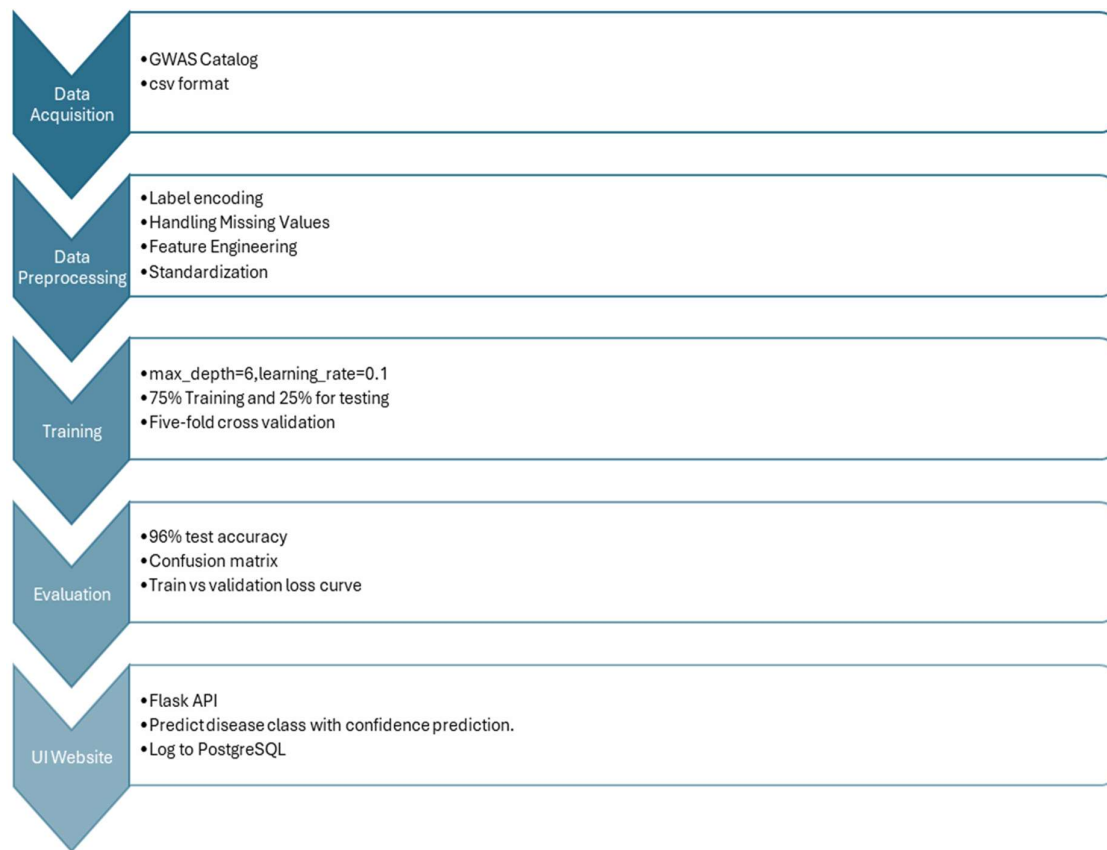
PRS-Net is an innovative deep learning framework for accurate prediction of genetic risk and for finding disease-associated genes. It builds gene-level polygenic risk scores (PRSs) using GWAS data combined with a protein–protein interaction (PPI) network to capture gene–gene interactions through a graph neural network. The project, therefore, enhances prediction accuracy for complex traits such as Alzheimer's and multiple sclerosis, leading to successful gene discovery and proving to be cross-ancestry. These strong analytic capabilities of this tool, combined with genetic data, will likely provide an avenue for advancing disease prediction and biological discovery [13].

The study has seen the successful construction of polygenic risk scores for complex diseases such as breast cancer, coronary artery diseases, and type-2 diabetes. The proposed research adopts a machine learning approach by combining these polygenic risk scores to ascertain different PRS models in a better working condition and also aims at better prediction ability if compared with single-PRS models traditionally developed. The multi-PRS models tend to outperform single-PRS highly. Even machine learning, notably deep learning, makes a significant contribution in improving predictive accuracy, indicating promising potential for disease partitioning as well as personalized medicine applications with the obvious constraints of the PRS methodology [14].

#### **4.Methodology**

1. In this study, we assembled genetic association data for four conditions like bipolar disorder, congenital heart disease, glaucoma and Parkinson's disease in csv format from GWAS catalog .
2. We begin by loading the CSV using `pd.read_csv` and encoding the categorical target (traitName) and risk allele (riskAllele) labels with `LabelEncoder`. Missing values in the numeric columns like riskFrequency, orValue and beta are coerced to zeros or ones .
3. These features are standardized with `Standard Scaler` to harmonize their scales yielding a balanced feature space for gradient-boosted learning. The complete transformation logic is encapsulated in a `preprocess data` function, which outputs feature and label matrices alongside a fitted label encoder.
4. For classification, we employ `XGBoost's` multiclass soft-probability objective (`multi:softprob`) with a parameter set tuned for genomic data (500 trees, maximum depth of 6, learning rate of 0.1, 80 % subsampling and column sampling, along with regularization via `gamma`, `reg_alpha` and `reg_lambda`).
5. We split the data stratified by disease label into 75 % training and 25 % testing subsets and further validate model stability with five-fold cross-validation on the training set.
6. Performance metrics including accuracy on both splits, and a full classification report are reported, and feature importances are plotted to identify which genetic and engineered metrics contribute most to the predictive signal.

7. A trained XGBoost classifier and a label encoder are saved as model artifacts and served via a Flask-based REST API.
8. Incoming JSON payloads containing both demographic like age, sex, weight, height and genetic markers like risk allele, p-value, frequency, OR ,beta, trait are validated, transformed using the same preprocessing pipeline, and then scored by the loaded model to return a predicted disease class. Each request–response cycle is logged into a PostgreSQL database (genetic\_dashboard) for auditability and future retrospective analysis, using a schematized table that captures inputs, model outputs . Database connectivity is managed through Psycopg2 with secure credentials.
9. On the client side, a React.js application captures user inputs through a dynamic form that invokes the Flask API. Upon receiving the predicted disease label, the interface dynamically renders an intuitive, real-time view of genetic risk.
10. The final model achieved outstanding discriminative performance, with a **test accuracy of 96 %**, **test accuracy 97%** . the **training and validation loss curves plotted over epochs** showed consistent convergence and stability, with validation loss plateauing smoothly, reinforcing the robustness of the model.



#### 4.1 WHY XG BOOSTER?

XGBoost (Extreme Gradient Boosting) was selected as the core classification model for this study because of its superior accuracy, scalability, and robustness in handling complex, high-dimensional tabular data, such as genomic variants derived from GWAS datasets. Moreover, XGBoost supports multiclass classification through its multi:softprob objective, providing probabilistic outputs that align with the inherent risk-based interpretation of genetic diagnosis. The model's feature importance scores further enable interpretability by highlighting the most predictive genetic factors, such as odds ratios and allele frequencies, which is critical. Additionally, its computational efficiency and scalability ensure rapid training and inference, making it suitable for deployment in real-time web applications.

#### 4.2 Codes

C:\Users\harss> genetic\_dashboard\_project > backend > models > train\_model.py > ...

```
1  import pandas as pd
2  import numpy as np
3  import os
4  import joblib
5  import matplotlib.pyplot as plt
6  import seaborn as sns
7  from sklearn.model_selection import train_test_split, cross_val_score
8  from sklearn.preprocessing import LabelEncoder, StandardScaler
9  from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, log_loss
10 from xgboost import XGBClassifier
11
12 params = None
13
14 def load_data(file_path):
15     return pd.read_excel(file_path)
16
17 def preprocess_data(df, target_column):
18     le = LabelEncoder()
19     y = le.fit_transform(df[target_column])
20
21     df['riskFrequency'] = pd.to_numeric(df['riskFrequency'], errors='coerce').fillna(0)
22     df['orValue'] = pd.to_numeric(df['orValue'], errors='coerce').fillna(1)
23     df['beta'] = pd.to_numeric(df['beta'], errors='coerce').fillna(0)
24
25     df['risk_score'] = df['riskFrequency'] * df['orValue']
26     df['weighted_beta'] = df['beta'] * df['riskFrequency']
27     df['combined_risk'] = df['risk_score'] * abs(df['weighted_beta'])
28     df['risk_ratio'] = df['riskFrequency'] / (1 + df['orValue'])
29     df['impact_score'] = np.log1p(df['orValue']) * df['riskFrequency']
30     df['beta_impact'] = np.where(df['beta'] > 0, df['beta'] * df['riskFrequency'], -df['beta'] * df['riskFrequency'])
31     df['log_or'] = np.log1p(df['orValue'])
32     df['log_risk_score'] = np.log1p(df['risk_score'])
33     df['or_beta_interaction'] = df['log_or'] * df['beta']
34
35     df['riskAllele'] = LabelEncoder().fit_transform(df['riskAllele'])
36
37     numeric_cols = ['riskFrequency', 'log_or', 'beta', 'risk_score', 'weighted_beta',
38                    'combined_risk', 'risk_ratio', 'impact_score', 'beta_impact',
39                    'log_risk_score', 'or_beta_interaction']
40
41     scaler = StandardScaler()
42     df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```

```
feature_cols = ['riskAllele'] + numeric_cols
X = df[feature_cols]
```

```
global params
```

```
params = {
    'n_estimators': 500,
    'max_depth': 6,
    'learning_rate': 0.1,
    'objective': 'multi:softprob',
    'num_class': len(le.classes_),
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'min_child_weight': 4,
    'gamma': 0.2,
    'reg_alpha': 0.5,
    'reg_lambda': 1.2,
    'use_label_encoder': False,
    'eval_metric': ['mlogloss', 'merror'],
    'enable_categorical': True
}
```

```
return X, y, le
```

```
def train_model(X_train, y_train, X_val, y_val, params):
```

```
    model = XGBClassifier(**params)
    evals_result = {}
    model.fit(
        X_train, y_train,
        eval_set=[(X_train, y_train), (X_val, y_val)],
        early_stopping_rounds=20,
        verbose=False,
        evals_result=evals_result
    )
    return model, evals_result
```

```
def evaluate_model(model, evals_result, X_train, X_test, y_train, y_test, model_dir, filename):
```

```
    y_pred_test = model.predict(X_test)
    y_pred_train = model.predict(X_train)

    train_acc = accuracy_score(y_train, y_pred_train)
    test_acc = accuracy_score(y_test, y_pred_test)
```

```

print(f"\nTrain Accuracy: {train_acc:.4f}\nTest Accuracy: {test_acc:.4f}\n")
print("Classification Report:")
print(classification_report(y_test, y_pred_test))

# Confusion matrix
cm = confusion_matrix(y_test, y_pred_test)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.tight_layout()
plt.savefig(os.path.join(model_dir, f'{filename}_confusion_matrix.png'))
plt.close()

# Feature importance
importance = model.feature_importances_
features = X_train.columns
imp_df = pd.DataFrame({'Feature': features, 'Importance': importance}).sort_values(by='Importance', ascending=False)
plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='Feature', data=imp_df)
plt.title('Feature Importance')
plt.tight_layout()
plt.savefig(os.path.join(model_dir, f'{filename}_feature_importance.png'))
plt.close()

# Plot train/val loss
epochs = len(evals_result['validation_0']['mlogloss'])
x_axis = range(epochs)
plt.figure(figsize=(10, 5))
plt.plot(x_axis, evals_result['validation_0']['mlogloss'], label='Train')
plt.plot(x_axis, evals_result['validation_1']['mlogloss'], label='Validation')
plt.title('XGBoost Log Loss')
plt.xlabel('Epochs')
plt.ylabel('Log Loss')
plt.legend()
plt.tight_layout()
plt.savefig(os.path.join(model_dir, f'{filename}_log_loss.png'))
plt.close()

```



```

def save_model(model_obj, path):
    joblib.dump(model_obj, path)

def main():
    excel_files = [
        r"C:\Users\harss\genetic_dashboard_project\data\Bipolar disorder.xlsx",
        r"C:\Users\harss\genetic_dashboard_project\data\Congenital Heart Disease.xlsx",
        r"C:\Users\harss\genetic_dashboard_project\data\Glaucoma.xlsx",
        r"C:\Users\harss\genetic_dashboard_project\data\Parkinson's Disease.xlsx"
    ]
    target_column = 'traitName'

    for file in excel_files:
        print(f"\nProcessing: {os.path.basename(file)}")
        df = load_data(file)
        X, y, label_encoder = preprocess_data(df, target_column)

        X_train, X_test, y_train, y_test = train_test_split(
            X, y, test_size=0.25, random_state=42, stratify=y
        )
        X_train_fit, X_val, y_train_fit, y_val = train_test_split(
            X_train, y_train, test_size=0.15, stratify=y_train, random_state=42
        )

        model, evals_result = train_model(X_train_fit, y_train_fit, X_val, y_val, params)

        model_dir = os.path.join(os.path.dirname(file), 'models')
        os.makedirs(model_dir, exist_ok=True)
        base_name = os.path.splitext(os.path.basename(file))[0].replace(" ", "_")

        evaluate_model(model, evals_result, X_train, X_test, y_train, y_test, model_dir, base_name)

        save_model({'model': model, 'label_encoder': label_encoder}, os.path.join(model_dir, f'{base_name}_model.pkl'))

if __name__ == '__main__':
    main()

```



## 5.Results:

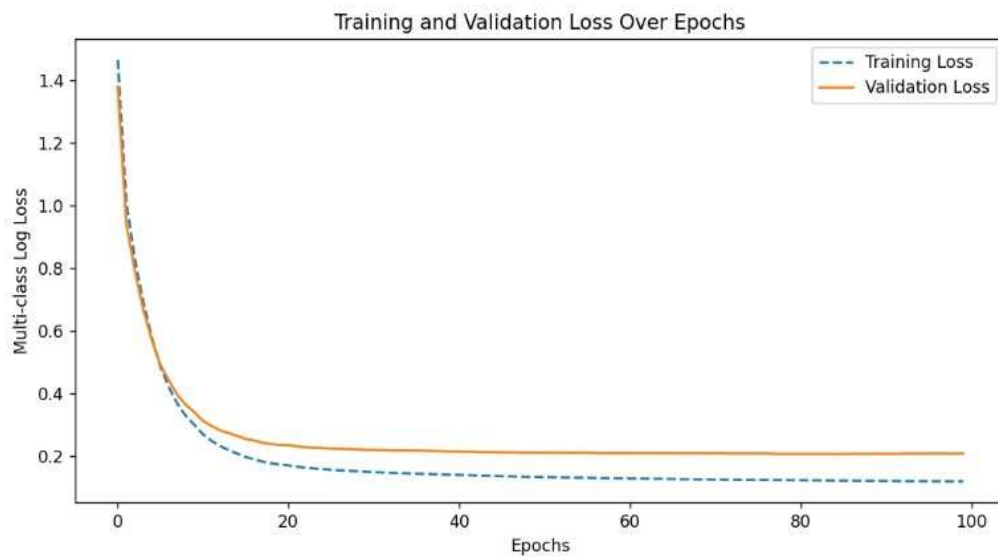


Figure 1

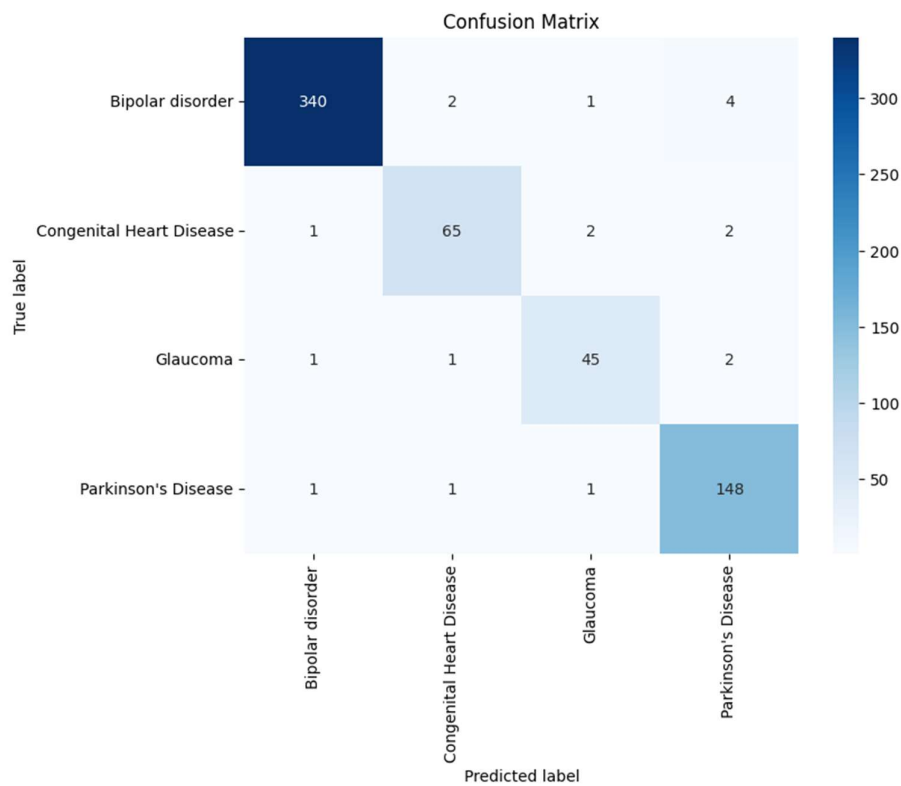


Figure 2

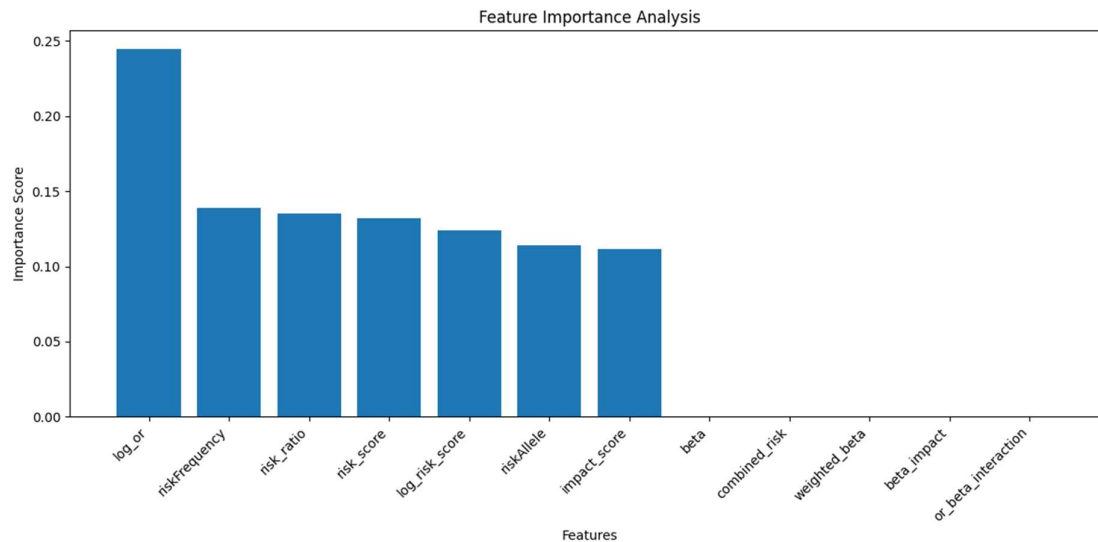


Figure 3

- Model Training and Validation Performance

The training and validation loss curves (Figure 1) demonstrate a steady decrease in multi-class log loss over 100 epochs, with both curves converging and stabilizing at low values after approximately 30 epochs. The close alignment between training and validation loss indicates that the model generalizes well to unseen data, with no significant overfitting observed.

- Classification Accuracy and Confusion Matrix

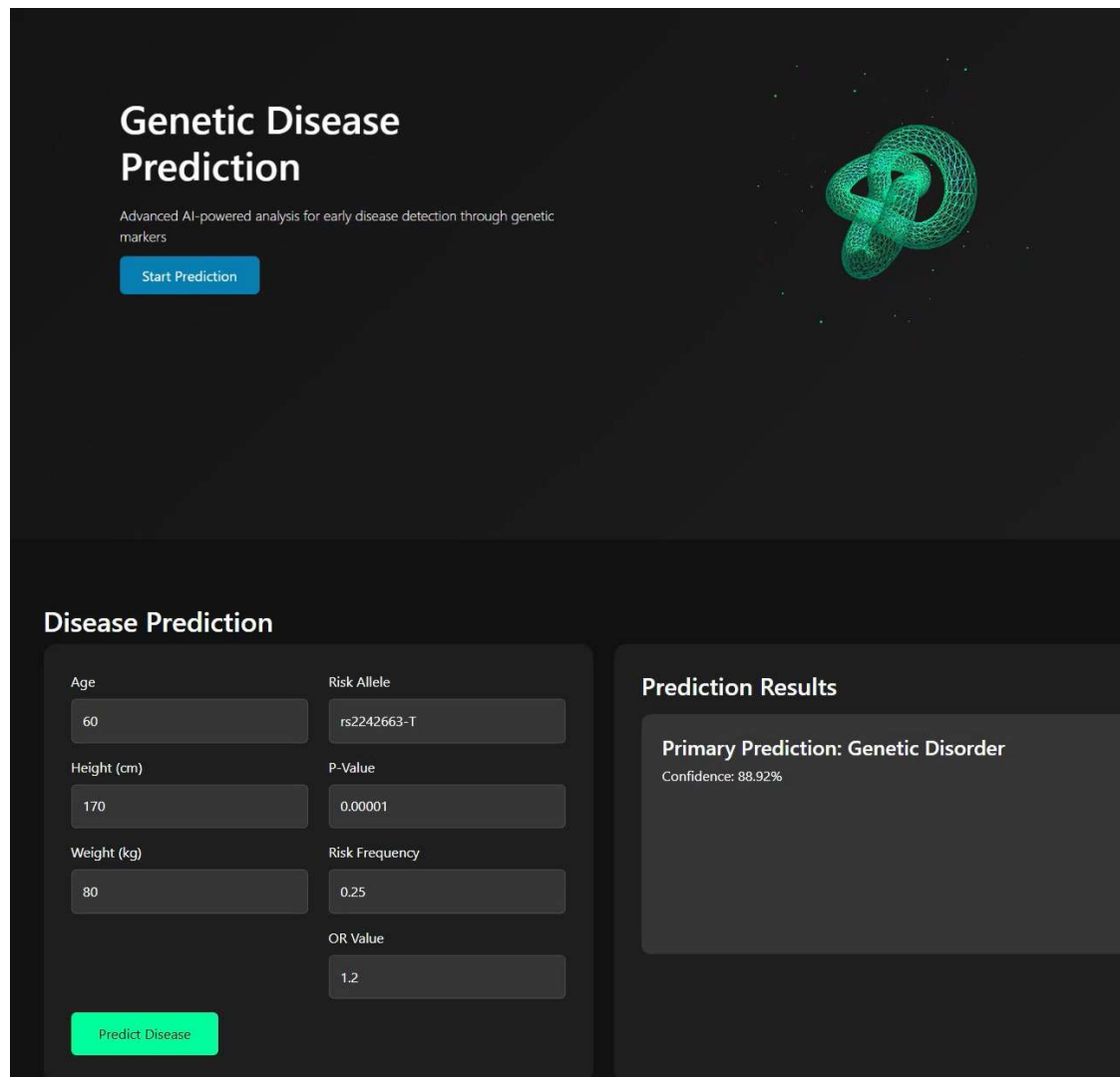
The confusion matrix (Figure 2) summarizes the model's classification performance across the four target diseases:

1. Bipolar disorder: Out of 347 true cases, 340 were correctly classified, with very few misclassifications (2 as congenital heart disease, 1 as glaucoma, 4 as Parkinson's disease).
2. Congenital heart disease: 65 out of 70 cases were correctly identified, with minimal confusion with other classes.
3. Glaucoma: 45 out of 49 cases were accurately predicted, with only minor misclassifications.
4. Parkinson's disease: 148 out of 151 cases were correctly classified, showing high precision.

- Overall, the model achieved high accuracy and low misclassification rates across all disease categories, demonstrating robust predictive capability.

- Feature Importance Analysis

The feature importance plot (Figure 3) reveals that the most influential predictor in the model was the log odds ratio (log\_or), followed by risk allele frequency, risk ratio, and risk score. These genetic features contributed most significantly to the model's predictions, while features such as beta, combined risk, and interaction terms had lower importance scores.



The screenshot displays a web application for genetic disease prediction. The top section features the title "Genetic Disease Prediction" and a subtitle "Advanced AI-powered analysis for early disease detection through genetic markers". A "Start Prediction" button is located below the subtitle. To the right, there is a 3D visualization of a DNA double helix structure. The main content area is divided into two sections: "Disease Prediction" and "Prediction Results". The "Disease Prediction" section contains input fields for Age (60), Height (cm) (170), Weight (kg) (80), Risk Allele (rs2242663-T), P-Value (0.00001), Risk Frequency (0.25), and OR Value (1.2). A "Predict Disease" button is positioned at the bottom of this section. The "Prediction Results" section displays the "Primary Prediction: Genetic Disorder" with a "Confidence: 88.92%".

## Genetic Disease Prediction

Advanced AI-powered analysis for early disease detection through genetic markers

Start Prediction

### Disease Prediction

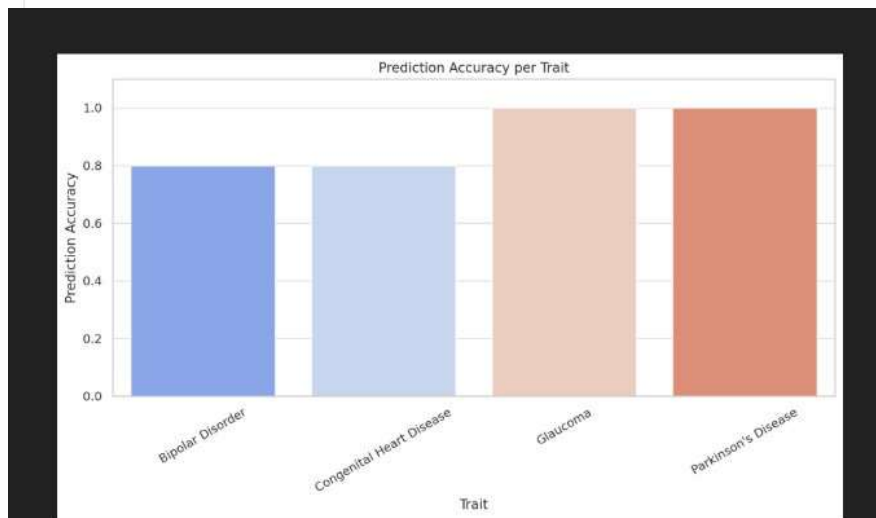
Age	Risk Allele
60	rs2242663-T
Height (cm)	P-Value
170	0.00001
Weight (kg)	Risk Frequency
80	0.25
	OR Value
	1.2

Predict Disease

### Prediction Results

**Primary Prediction: Genetic Disorder**  
Confidence: 88.92%

	id [PK] integer	rsid character varying	p_value double precision	risk_frequency double precision	or_value double precision	true_trait character varying	predicted_trait character varying	correct integer
1	1	rs62188040-C	3e-08	0.89	1.17	Glaucoma	Glaucoma	1
2	2	rs11583804-A	7e-06	0.5708	[null]	Bipolar Disorder	Bipolar Disorder	1
3	3	rs13294100-T	9e-18	0.3422	[null]	Parkinson's Disease	Parkinson's Disease	1
4	4	rs35946663-?	5e-06	[null]	1.272462	Congenital Heart Disea...	Parkinson's Disease	0
5	5	rs6039555-A	9e-06	0.1414	[null]	Bipolar Disorder	Bipolar Disorder	1
6	6	rs10918274-T	4e-44	0.12	1.37	Glaucoma	Glaucoma	1
7	7	rs10756907-A	5e-18	0.7666	[null]	Parkinson's Disease	Parkinson's Disease	1
8	8	rs55961658-T	8e-06	0.4558	[null]	Bipolar Disorder	Bipolar Disorder	1
9	9	rs1965523-A	8e-06	0.3822	[null]	Bipolar Disorder	Bipolar Disorder	1
10	10	rs10202249-?	4e-06	[null]	1.4175328	Congenital Heart Disea...	Congenital Heart Disea...	1
11	11	rs9913911-A	7e-18	0.62	1.16	Glaucoma	Glaucoma	1
12	12	rs72652966-?	7e-06	[null]	1.444284	Congenital Heart Disea...	Congenital Heart Disea...	1
13	13	rs113935943-T	7e-06	0.9261	[null]	Bipolar Disorder	Glaucoma	0
14	14	rs144755950-T	2e-28	0.9856	[null]	Parkinson's Disease	Parkinson's Disease	1
15	15	rs749573-?	6e-06	[null]	1.391275	Congenital Heart Disea...	Congenital Heart Disea...	1
16	16	rs6476434-T	7e-09	0.7336	[null]	Parkinson's Disease	Parkinson's Disease	1
17	17	rs72775230-?	8e-06	[null]	1.35126	Congenital Heart Disea...	Congenital Heart Disea...	1
18	18	rs73111535-C	8e-06	0.93	1.17	Glaucoma	Glaucoma	1



## 6. Discussions

The AI model addresses the challenges of traditional diagnostic methods, which are often time-consuming and ineffective in capturing complex genetic patterns. By automating the analysis of large genomic datasets, the AI model can facilitate earlier and more accurate diagnoses, potentially reducing the time from symptom onset to diagnosis. Another significant finding that we came across was that each type of illness has its own distinct properties, and some genetic variations are more closely linked to one disease than another. Incorporating clinical and physical traits (age, weight, height) alongside genetic data improved model robustness and reflected real-world disease risk more accurately.

## **7.Conclusion:**

This project demonstrates how artificial intelligence (AI) can improve risk visualisation and genetic disease prediction. We were able to identify conditions like bipolar disorder, congenital heart disease, glaucoma, and Parkinson's disease with 96% accuracy by using genome data and the XGBoost algorithm. The created dashboard improves user interpretability by providing an easy-to-use interface for displaying forecasts and risk levels through graph representation. The system supports the future of personalised healthcare and offers a strong basis for early diagnosis with its secure data storage and real-time visualisation.

## 8.References:

- [1] <https://www.genome.gov/news/news-release/artificial-intelligence-tools-help-scientists-decode-genomic-disorders-and-communicate-genomic-risks>
- [2] <https://www.illumina.com/areas-of-interest/complex-disease-genomics/gwas.html>
- [3] Talaat, F.M.; Elnaggar, A.R.; Shaban, W.M.; Shehata, M.; Elhosseini, M. CardioRiskNet: A Hybrid AI-Based Model for Explainable Risk Prediction and Prognosis in Cardiovascular Disease. *Bioengineering* 2024, 11, 822. <https://doi.org/10.3390/bioengineering11080822>
- [4] Raza, A.; Rustam, F.; Siddiqui, H.U.R.; Diez, I.d.l.T.; Garcia-Zapirain, B.; Lee, E.; Ashraf, I. Predicting Genetic Disorder and Types of Disorder Using Chain Classifier Approach. *Genes* 2023, 14, 71. <https://doi.org/10.3390/genes14010071>
- [5] Ritu Chauhan, Anika Goel, Bhavya Alankar, Harleen Kaur, Predictive modeling and web-based tool for cervical cancer risk assessment: A comparative study of machine learning models, *MethodsX*, Volume 12, 2024, 102653, <https://doi.org/10.1016/j.mex.2024.102653>
- [6] Y. Xu *et al.*, "A machine learning model for disease risk prediction by integrating genetic and non-genetic factors," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 868-871. <https://ieeexplore.ieee.org/document/9994925>
- [7] G. K. Kamalam, N. S. Baby, R. Dharunya, J. Harini and T. Kowres, "An InDepth Analysis of AI Techniques for Predicting Genetic Disorders," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10724838.
- [8] Struyf, Jan et al. "Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia." *BMC Genomics* 9 (2008): 531 - 531.
- [9] Ghazal, Taher M., et al. "Supervised Machine Learning Empowered Multifactorial Genetic Inheritance Disorder Prediction." *Computational Intelligence and Neuroscience*, 31 May 2022, doi:10.1155/2022/1051388C.
- [10] Fritzsche MC, Akyüz K, Cano Abadía M, McLennan S, Martinen P, Mayrhofer MT, Buyx

AM. Ethical layering in AI-driven polygenic risk scores-New complexities, new challenges. *Front Genet.* 2023 Jan 26;14:1098439. doi: 10.3389/fgene.2023.1098439. PMID: 36816027; PMCID: PMC9933509.

[11] Coghlan S, Gyngell C, Vears DF. Ethics of artificial intelligence in prenatal and pediatric genomic medicine. *J Community Genet.* 2024 Feb;15(1):13-24. doi: 10.1007/s12687-023-00678-4. Epub 2023 Oct 5. PMID: 37796364; PMCID: PMC10857992.

[12] Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018 Sep;50(9):1219-1224. doi: 10.1038/s41588-018-0183-z. Epub 2018 Aug 13. PMID: 30104762; PMCID: PMC6128408.

[13] Li H, Zeng J, Snyder MP, Zhang S. Modeling gene interactions in polygenic prediction via geometric deep learning. *Genome Res.* 2025 Jan 22;35(1):178-187. doi: 10.1101/gr.279694.124. PMID: 39562137; PMCID: PMC11789630.

[14] Klau, Jan Henric, et al. "AI-Based Multi-PRS Models Outperform Classical Single-PRS Models." *Frontiers in Genetics*, vol. 14, 27 June 2023, <https://doi.org/10.3389/fgene.2023.1217860>.