SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I6P113

Volume 12 Issue 6, 158-183, June 2025

Original Article

A Hybrid Ensemble-Based Binary Classifier for Early and Interpretable Detection of Genetic Disorders

Ishdeep[1], Neetu Rani[2]

[1,2]*Computer Science & Engineering, Chandigarh University, Mohali, Punjab, India.*

[1]*Corresponding Author : ishdeep.singla@gmail.com*

Received: 07 April 2025 Revised: 08 May 2025 Accepted: 09 June 2025 Published: 27 June 2025

Abstract *- Genetic disorders often stem from environmental or inherited DNA mutations, and early detection significantly improves life expectancy and reduces long-term healthcare costs for the commonwealth. Machine learning has proven effective in predicting and diagnosing such disorders, enabling treatment before the disorder hits a critical point. This research focuses on enhancing diagnostic accuracy using ensemble and bagging algorithms across three major genetic disorder groups while also making it economically inexpensive and less time-consuming by coining a hybrid ensemble-based binary classifier, the "Binary Multi-Model Disorder Classifier (BMMDC)", a novel approach, which addresses limitations of the current standard multiclass classifiers, achieving an average of 95% accuracy over all disorders while also increasing the interpretability using explainable artificial intelligence.*

Keywords *- Machine learning, BMMDC, Genetic Disorder, Gradient Boosting, LightGBM, XGBoost.*

1. Introduction

As history shows, life on Earth has always been accompanied by genetic variations and mutations [1]. These variations guarantee that a particular species can adapt to the surrounding environment, acquire immunity and resistance to various infections, maintain a diverse gene pool, and relieve the dangers associated with inbreeding depression, thus helping life thrive [2].

Table 1. Biomarker organization

Category	Parameter	Measurement Unit / Description
Complete Blood Count (CBC)	Red Blood Cell (RBC) Count	Measured in million cells per microliter (mcL)
	White Blood Cell (WBC) Count	Measured in thousand cells per microliter (mcL)
Vital Signs	Respiratory Rate	Measured in breaths per minute
	Heart Rate	Measured in beats per minute
Perinatal Factors	Birth Asphyxia	Indicates oxygen deprivation at birth
	Folic Acid Supplementation	Details on peri-conceptional intake
Maternal History	Serious Maternal Illness	History of significant health issues during pregnancy
	Radiation Exposure	Exposure to X-rays during pregnancy
	Substance Abuse	Use of harmful substances during pregnancy
	Assisted Conception	Use of In Vitro Fertilization (IVF) or Assisted Reproductive Technology (ART)
Obstetric History	Number of Previous Abortions	Count of prior pregnancy losses
		History of anomalies in previous pregnancies
Current Observations	Birth Defects	Any congenital anomalies present
	Blood Test Results	Findings from specific blood tests

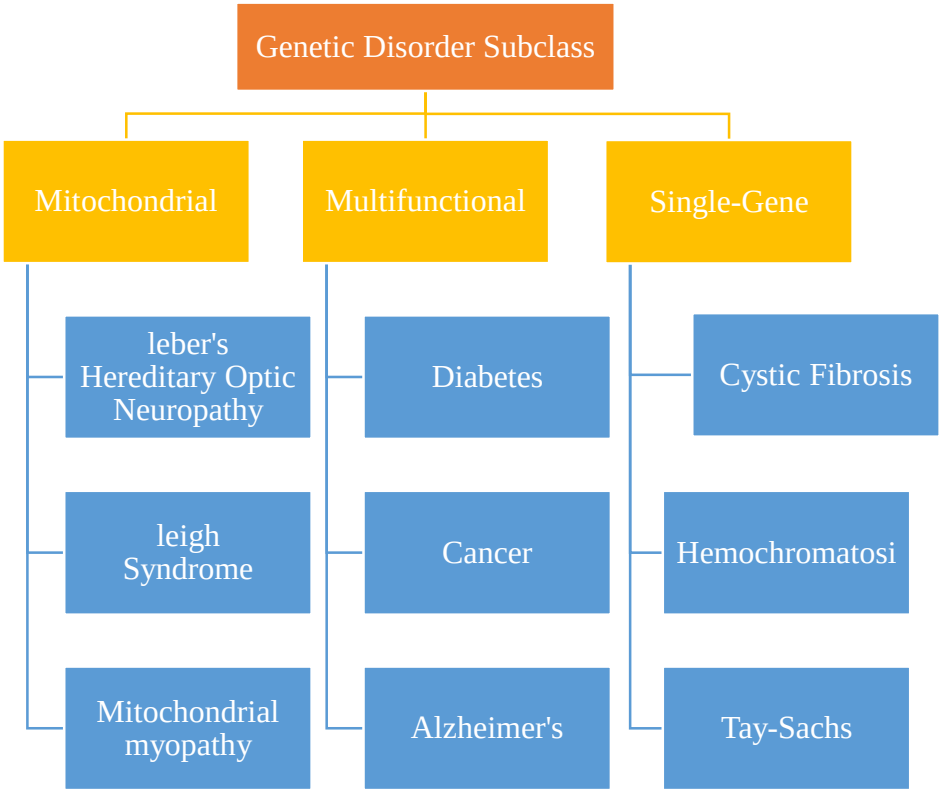

Fig. 1 Genetic disorders and subtypes

1.1. About Genetic Disorders

Although crucial for the development of healthy life and life forms, if these mutations "go wrong", they can lead to a plethora of fatal and sometimes incurable disorders. These genetic mutations may affect a single gene or multiple genes. Either way, they cause terrible defects in the patient's genetic constitution, eventually leading to a drastic impact on the individual's health, which is majorly associated with their genetic makeup and family history, causing a lower quality and expectancy of life [3]. Genetic disorders can be commonly diagnosed in the form of developmental delays, physical or mental defects or in the form of poor health conditions. In short, genetic disorders are health problems caused by one or more anomalies in the genome.

1.2. Disorders Focused in this Study

This study focuses on three major categories: 1) single gene [4], 2) multifactorial [5], and 3) mitochondrial disorders [6], followed by nine disorder subclasses in these three categories as described in Figure 1.

1.3. Useful Biomarkers in Disorder Identification

In diagnosing the nine aforementioned disorders, the diagnostic parameters would be categorized into six broad categories, each further divided into relevant subtypes. This structured approach enables a deeper understanding of the root causes of these conditions, ultimately aiding in selecting key biomarkers essential for training machine learning-based disorder classification models. Table 1 provides an overview of the biomarkers under consideration.

1.4. Agenda of this Study

The primary goal of this study is to develop a robust and efficient machine learning-based classification model that can accurately predict the presence or likelihood of genetic disorders using simplified diagnostic and clinical data. Rather than relying on complex and expensive genomic sequencing techniques, this research uses routinely available medical parameters, including blood test results, perinatal and maternal history, vital signs, and observable birth anomalies, to detect potential genetic abnormalities.

The specific objectives of the study are:

- To identify and organize clinically relevant biomarkers from standard medical reports that can be reliable predictors for genetic disorders.
- To design and implement an ensemble-based classification model, particularly the proposed Binary Multi-Model Disorder Classifier (BMMDC), that overcomes the limitations of traditional single-model classifiers.
- To evaluate and compare the performance of advanced machine learning algorithms in predicting three broad categories of genetic disorders—Single-Gene, Multifactorial, and Mitochondrial.
- To improve diagnostic accuracy and efficiency, aiming for a high prediction accuracy while reducing computational overhead and test area.
- To propose a cost-effective and accessible alternative to genetic sequencing by utilizing readily available clinical data for early and accurate detection of genetic disorders.

This study aims to bridge the gap between cutting-edge machine learning technologies and their practical application in clinical genetics, ultimately contributing to better patient outcomes through early intervention and personalized treatment strategies.

2. Literature Review

Genomics and healthcare have been completely transformed due to the incorporation of Machine Learning (ML) [7]. Many recent scientific publications have critically assessed the application of AI and ML techniques and determined their advantages, disadvantages, and aspects to be improved.

Researchers in [8] investigated the application of ML in genetics and how these approaches utilize extensive and intricate genomic data towards the goal of precision medicine. This study described 32 AI and ML techniques previously documented in 24 different genomic studies, claiming their usefulness in diagnosing inflammatory bowel disease, systemic lupus erythematosus, and some cancers. Important algorithms like RF, SVM, Gradient Boosting, and XGBoost also performed successfully in disease classification alongside their biomarkers.

Earlier research reported success in cancer stratification and biomarker discovery with RNA sequence and whole genome sequencing data types using these algorithms. Significant positive impacts have been made in genomics studies due to access to publicly available datasets, such as the TCGA and Gene Expression Omnibus. At the same time, the study notes some challenges that should raise concerns, such as erratic output from different datasets, algorithmic bias, and inadequate generalization capability. To address these challenges, the authors propose more harmonized data processing and reporting methods and greater incorporation of diverse datasets to improve the models' generalizability and ethical scrutiny.

Researchers in [9] have developed a groundbreaking method for predicting genetic disorders based on machine learning techniques. Disorders that relate to the human genome are likely to remain a challenge to global health due to the mutations occurring in DNA. The authors developed a multi-label, multiclass classifier using novel feature engineering and a classifier chain approach.

To increase the accuracy of their model, the authors used ensemble learning methods of Extra Trees (ET) and Random Forest (RF) to maximize the prediction potential of their model. Using Extreme Gradient Boosting (XGB), they obtained a high alpha-evaluation score of 92% and macro accuracy of 84%. One of the innovations of the study was the incorporation of the classifier chain approach, which utilizes the output of previous classifiers to influence future outputs. Exploratory Data Analysis (EDA) revealed additional information, such as possible relationships between certain genetic disorders and blood cell counts. A study in [10] also put forth a system for classification, yet they traded generalizability for accuracy, thus reducing the application areas for their classifier. Authors in [11] proposed AI and big data techniques in areas of cancer medicine, including NGS, radiomics and digital pathology. NGS is one of the prominent techniques in cancer precision medicine. Therefore, genomic abnormalities and mutations culminating in oncogene activation can be profiled.

Other developments, such as RNA sequencing, have made it possible to obtain gene signatures of expression profiles and other molecular changes that would be useful for companion diagnostics and drug development. As per the study, AI also assists in further pinpointing important disease areas by introducing interpretable visualization of diagnostic images, in case AI tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) render such images interpretable. Another approach, shallow whole-genome sequencing, cheaper than the traditional one, was found to enhance detection capacity whilst maintaining its sensitivity. However, despite the many improvements, the authors note worrying trends, such as data fragmentation, algorithm biases, and AI non-explainability. The combination of Ethical issues of data protection, privacy and fairness of algorithms creates even more problems for accepting artificial intelligence in the clinical environment.

Researchers in [12] critically review the metrics used in evaluating ML models in genomics, which fall into three major categories: clustering, classification, and regression. Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) are typical metrics for clustering tasks. In cases where datasets have balanced clusters, ARI performed well, whereas in imbalanced datasets, the performance of AMI is significantly better, particularly for rare diseases. Intrinsic validation metrics, such as the Silhouette Index (SI) and Davies-Bouldin Index (DBI), are also discussed, though their limitations in handling irregular genomic distributions are noted. Classification tasks for disease prediction are typically assessed concerning precision, F1 score, recall, and Matthew's Correlation Coefficient (MCC).

MCC seems more robust in the presence of imbalanced datasets. Although less frequent in genomics, regression metrics are quite important for trait prediction of continuous traits. MAE and RMSE are discussed in this context, where outliers tend to increase the variance of the method. The authors underline the importance of task-specific metrics selection for genomics in achieving reliability and better interpretability. They suggest a standardized protocol, which will consistently standardize the interpretation across studies.

Authors in [13] reported implementing a new hybrid model that integrates Feynman Concordance and Interpolated Nearest Centroid methods for improved accuracy in genomic disorder prediction. Their model worked on enhanced class discrimination as well as

decision boundary refinement in high-dimensional genomic datasets. Their model performed well with a concordance index of 0.89, showing strong results for some rare genetic disorders. Still, digressive genomic sequence dependencies and intricate calculations as parts of the framework might hinder pragmatic use in under-resourced clinical settings, which often rely on routine biomarker data rather than complex genomic sequencing. The reviewed literature documents considerable progress in the application of AI/ML in genomics and health, including the prediction of genetic diseases and cancer diagnosis. Many problems still need to be addressed, including that despite advancements in AI-driven healthcare solutions, several research gaps persist in early genetic disorder detection. The limited availability of annotated and diverse datasets significantly impedes the development of robust predictive models.

Additionally, the issue of class imbalance, where rare disorders are underrepresented, affects model generalization. A critical gap exists in the interpretability of hybrid models, which are often complex and challenging for clinicians to trust. The absence of well-defined strategies for selecting and combining neural network architectures also limits performance across heterogeneous genetic conditions. Moreover, while hybrid ensemble models have the potential to improve accuracy, their practical design and validation remain underexplored. Lastly, inconsistent validation protocols and inadequate data-sharing frameworks obstruct progress in model deployment and reproducibility.

Addressing these gaps would enhance the impeded autonomy and fairness of machine learning in genomics and healthcare. Issues such as data privacy and fairness of algorithms remain important to guarantee that AI-based solutions will be implemented equitably.

3. Proposed Framework

The proposed study first aims to develop a high-performance multiclass classifier that is readily available in the market, with some preprocessing enhancements followed by a new approach to disorder classification that enhances predictive accuracy. This framework would undergo a series of tests to ensure its fitness and applicability to precision medicine. This study will follow the workflow depicted in Figure 2 to ensure seamless execution.

3.1. Biomarker–Disorder Correlation

After establishing a proper workflow for our study, this study will move towards the first data collection phase, *"The biomarker correlation identification".* This section will identify the significance of the biomarkers discussed in section 1.3 for the aforementioned nine disorders. Table 2 comprises all the biomarker correlations with the specified disorders.

The biomarkers chosen for each disorder align with their respective pathophysiology, helping improve diagnostic accuracy, particularly when integrated into machine learning models. Henceforth, from the following section onward, this study will move towards dataset identification and development of the classifiers, followed by the evaluation, testing and validation pipeline. Considering the review done on multiple research studies, diagnosis of most of the disorders can be done quickly with the help of the proposed methodology. However, some of them, like hemochromatosis and Tay-Sachs, have the problem that most features are not correlated enough with the disorders. Using a standard ML algorithm would not be fruitful because the decision factors would be very diluted amongst the features, making any average classifier confused about the presence of these disorders, which calls for a more dedicated diagnosis methodology.

3.2. Identifying a Dataset with the Above-Mentioned Biomarkers

To develop a cost-effective and highly accurate approach for diagnosing genetic disorders, curating a dataset encompassing the above-listed parameters is essential. By integrating a diverse set of biomarkers ranging from haematological and vital sign indicators to maternal and perinatal factors, it can be ensured that the dataset remains both comprehensive and generalizable. This will, in turn, increase the effectiveness of machine learning models in determining and classifying genetic disorders. Hopefully, organizing a dataset incorporating these biomarkers will enable us to achieve more accurate early detection, customized treatment plans, and favourable patient health indices.

3.2.1. Hardware and Software Requirements

This study utilized Jupyter Notebook within Visual Studio Code (VS Code) for interactive coding and visualization [82, 83]. The software framework was based on Python 3.11.0 [84] with key libraries: Pandas 2.2.3 for data manipulation [85], NumPy 2.0.2 for numerical computations [86], Scikit-learn 1.5.2 for machine learning [87], Matplotlib 3.9.3 and Seaborn 0.13.2 for visualization [88, 89], and Light GBM (4.6.0), XGBoost 2.1.3 and Imbalanced-learn 0.0 for advanced modeling [90, 91]. These tools enabled efficient preprocessing, analysis, and interpretation of medical datasets.

3.2.2. Exploratory Data Analysis (EDA)

After an extensive search for a comprehensive dataset that captures all relevant factors, a dataset that effectively generalizes laboratory test results has been identified. This dataset provides a well-rounded representation of lab tests, making it suitable for training a machine learning classifier. This dataset enhances the model's ability to make accurate and meaningful predictions across the nine targeted disorders by ensuring a diverse range of features and observations. The dataset was collected from a study by NCBI. It consists of 22083 patients' data, measuring 45 parameters describing the medical lab report of the patients in the USA. The dataset is a complex blend of patient demographics and lab tests. It comprises categorical, continuous, and Boolean data, as displayed in Table 3.

Table 3. Data distributions and column specifications

Data Type	Columns
Category	Respiratory Rate (breaths/min), Heart Rate (rates/min), Genetic Disorder, H/O radiation exposure (x-ray), H/O substance abuse, assisted conception IVF/ART, H/O serious maternal illness, History of anomalies in previous pregnancies, Disorder Subclass, Follow-up, Autopsy shows birth defect (if applicable), Gender, Status, Birth asphyxia, Birth defects, Blood test result
Boolean	Genes on mother's side, Test 3, Maternal gene, Paternal gene, inherited from father, Test 2, Test 4, Symptom 4, Test 5, Symptom 1, Symptom 2, Test 1, Symptom 3, Symptom 5, Parental consent
String	Patient ID, Patient First Name, Place of birth, Family Name, Father's name, Institute Name, Location of Institute, Folic acid details (peri-conceptional)
Float	Patient Age, Blood cell count (mcL), Mother's age, Father's age, No. Of previous abortion, White Blood cell count (thousand per microliter)

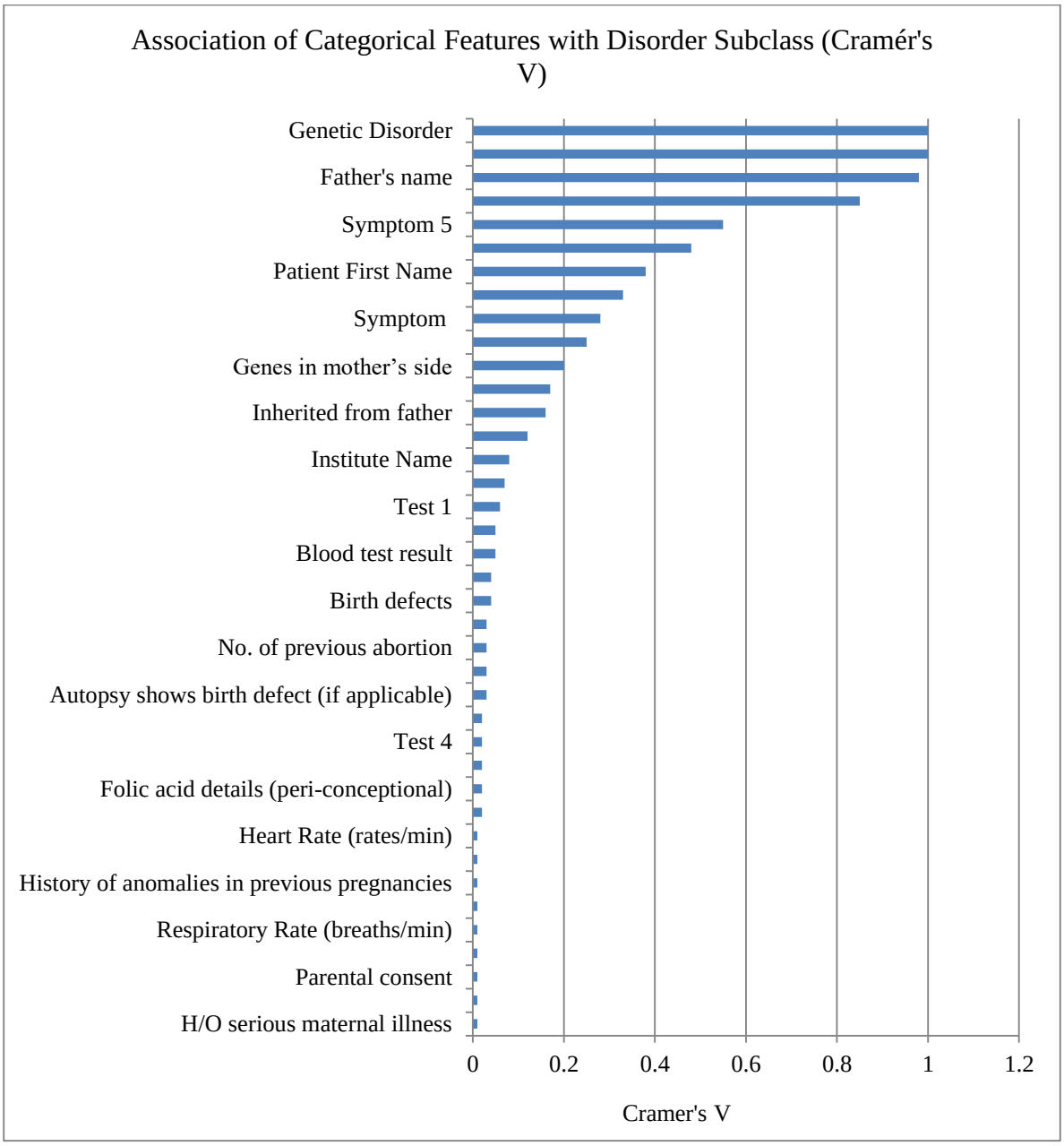

Fig. 3 Correlation of categorical features with disorder subclass

3.2.3. Data Distribution

The data frame has multiple numeric features with a diverse range of values. For instance, Mother's and Father's ages have the largest values, often between 30 and 60 years. In contrast, features like the WBC count and blood cell counts, though represented with smaller numerical values, are measured in thousands and millions, respectively.

Then comes the number of previous abortions and the cell (blood cell and WBC) count that, despite being represented in small magnitude decimals, are measured in thousands and millions. This causes a range difference, which could bias machine learning models toward prioritizing features with larger values.

3.2.4. Correlation Analysis

A structured correlation analysis is necessary to understand the features contributing to decision-making. This section will analyze the correlation of categorical, continuous and Boolean features with the disorder subclass (categorical feature).

Categorical Features

Categorical features in this dataset are those that have some predefined unique values, like "Yes"/ "No", "Normal"/ "Abnormal", and have been analyzed using the chi-square tests and Cramér's V [92].

The Chi-Square Test for Independence assesses whether there is a significant association between two categorical variables. In this case, it tests if the categorical feature (e.g., a demographic or genetic factor) is independent of the disorder subclass.

Equation 1: Chi-Square Test for Independence

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where:

- O_{ij}: Observed frequency
- E_{ij}: Expected frequency

$$E_{ij} = \frac{(RowTotal) \times (ColumnTotal)}{GrandTotal}$$

Once the Chi-Square Test for Independence is performed, Cramér's V is used to quantify the strength of the association between the categorical feature and the disorder subclass, as shown in Figure 3.

Equation 2: Cramér's V to Measure the Strength of the Categorical Association

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

Where:

- k: The smaller number of rows or columns in the contingency table.
- n: Total number of observations.

Continuous Features

For continuous features, Pearson's Correlation Coefficient has been used to assess the relationship with the disorder subclass, as shown in Figure 4.

Equation 3: Pearson's Correlation Coefficient

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual data points for the two variables being compared.
- $\underline{x} \wedge \underline{y}$ are the means of the variables x and y, respectively.

3.2.5. NULL Identification

Despite being diverse and robust, the dataset shows 95884 missing cells, averaging 2130 missing cells in each feature. On considering the class population of the dataset from Table 3, it can be seen that in the columns, some disorders like cancer and Alzheimer's have very few occurrences compared to others.

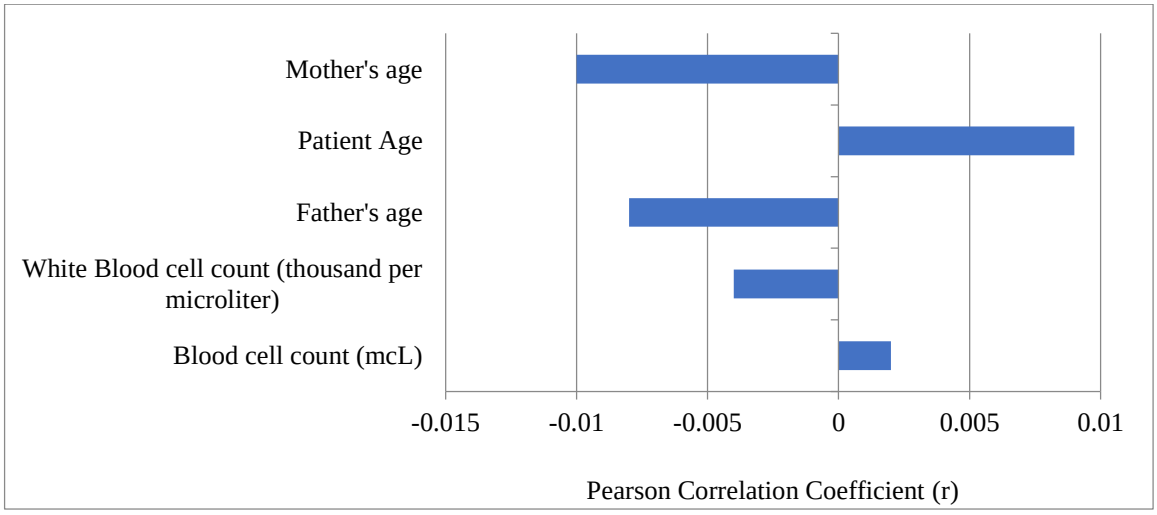

Fig. 4 Correlation of continuous features with disorder subclass

Table 3. Disorder subclass occurrences

Disorder Subclass	Occurrences
Leigh syndrome	5160
Mitochondrial myopathy	4405
Cystic fibrosis	3448
Tay-Sachs	2833
Diabetes	1817
Hemochromatosis	1355
Leber's hereditary optic neuropathy	648
Alzheimer's	152
Cancer	97

The nulls cannot be dropped in these classes thoughtlessly to make the data compatible with the machine learning model because that would reduce the data diversity for such lowly populated subclasses. Moreover, it would increase the already high-class imbalance that would tempt the model to make incorrect decisions due to developing a bias. Thus, tailored imputation strategies like selective mean imputation, hot deck imputation, and assumptive imputations have been developed for the features carrying nulls, as described in Table 4.

Table 4. NULL imputation strategies

Column Name	Strategy Used	Value Filled	Context
Patient Age	Mean by Disorder Subclass	Mean value per subclass	Handles age-specific disorder subclass
Blood cell count (mcl)	Mean by Disorder Subclass	Mean value per subclass	Numeric lab measurement
Mother's age	Mean by Disorder Subclass	Mean value per subclass	
Father's age	Mean by Disorder Subclass	Mean value per subclass	
No. of previous abortions	Mean by Disorder Subclass + Round to 0	Rounded mean (nearest integer)	Treated as a discrete integer value
White Blood cell count (thousand per microliter)	Mean by Disorder Subclass	Mean value per subclass (rounded to 2 decimals)	WBC lab value
History of anomalies in previous pregnancies	Mean by Disorder Subclass, then default	Mean or "No"	Inconsistency: used means first, later filled with "No"
Genetic Disorder	Mapping based on Disorder Subclass	Inferred value (e.g., "Mitochondrial...")	The custom dictionary used for mapping subclass to genetic disorder
Symptom 1 to Symptom 5	Fill missing values with the default value.	"0"	Assumes symptom absence if missing
Assisted conception IVF/ART	Fill missing values with the default value.	"No"	Categorical flag
Birth defects	Fill missing values with the default value.	"None"	Indicates absence of birth defects
Parental consent	Fill missing values with the default value.	"No"	Binary category
H/O substance abuse	Fill missing values with a default value.	"No"	Medical history
H/O serious maternal illness	Fill missing values with a default value.	"No"	Medical history
H/O radiation exposure (x-ray)	Fill missing values with a default value.	"No"	Medical history
Remaining Object Columns (if any)	Fill the missing with a placeholder value.	"void"	Catches unexpected string-based nulls
Rows with Disorder Subclass as NaN	Row removal	Dropped entirely	Ensures the label is present and accurate for supervised training

Table 5. Exclusion parameters

Parameter	Reason for Exclusion
Patient ID	Only a unique identifier is used to mark a patient for every patient in the format (PID0x6418).
Name of Patient, family and father	These names are only identifiers for a child below 14 years of age. The name of (or related to) any child does not determine any disorder inheritance pattern unless that particular family has some peculiar genetic constitution making them susceptible to specific disorders. This is not the case here and can easily be addressed on an individual case basis.
Name of the institute, location of institute and birthplace	Just like the patient names and family names, this is similarly unnecessary in disorder prediction unless it is given that the birthplace or the institute has some history of being a cause of genetic disorders.
Test 1 to 5	The tests only tell us about the tests conducted, not about their results. For example, Tests 1, 2, 3, and 5 have all values set to 0.0, and Test 4 has all values set to 1. Moreover, calculating the correlation values results in them being 0.0.

Furthermore, the lowest order of correlation of the parameters, as shown in Figures 3 and 4, has been dropped, particularly based on hit and trial, due to the overall low correlations of all the available parameters, leaving us with non-redundant features only.

3.3. Multiclass Classifier

From this section onwards, this study will move towards developing the multiclass classifier that will act as the benchmark for improving our study.

3.3.1. Data Encoding

Following the data cleaning phase, the next step involves transforming categorical text data into a numerical format suitable for machine learning algorithms. This conversion ensures that models can process the data efficiently and extract meaningful patterns. One commonly used technique for this transformation is label encoding, which converts categorical values into numerical representations based on lexicographical order. Each unique text category is assigned a distinct integer value, allowing the model to interpret the data in a structured manner. While label encoding is particularly useful for ordinal categorical variables, where the order of categories holds significance, it may introduce unintended ordinal relationships in nominal categorical variables. In such cases, alternative encoding methods, such as One-Hot Encoding, may be more appropriate to prevent misinterpretations by the model. The dataset is effectively prepared for machine learning algorithms by applying appropriate encoding techniques and enhancing model performance and interpretability. The category-specific encoding procedure has been recorded in Table 6.

3.3.2. Data Balancing

The dataset exhibits significant class imbalance, as observed in Table 3, where disorders like Leigh syndrome and mitochondrial myopathy have thousands of occurrences, whereas Alzheimer's and cancer have fewer than 200 cases. This imbalance can develop a bias in machine learning models, leading them to favour majority classes while underperforming minority classes. Data balancing techniques were applied to mitigate this and ensure a fair representation of all classes. Synthetic Minority Over-sampling Technique, an oversampling technique known as SMOTE, was used to generate synthetic data points for underrepresented classes. SMOTE creates artificial examples by interpolating between existing instances, preventing overfitting while improving generalization.

Table 6. Data encoding strategies

Column Name	Encoding Method
Gender, Birth asphyxia, Autopsy shows birth defect, H/O serious maternal illness, H/O radiation exposure (x-ray), H/O substance abuse, Assisted conception IVF/ART, History of anomalies in previous pregnancies, Status, Genetic Disorder, Disorder Subclass.	One Hot Encoding
Follow-up, Blood test result, Respiratory Rate (breaths/min), Heart Rate (rates/min), Birth defects	Label Encoding

3.3.3. Data Standardization

The dataset contained various numerical features, such as parental age rate and blood cell counts, which were measured in different units and scales. The Z-score normalization was applied to eliminate discrepancies in magnitude, as shown in Equation (4).

Equation 4: Z-Score normalization

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

Where:

- X is the original value,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

3.4. Model Selection and Evaluation Metrics

From all the analyses performed above, it is confirmed that the features exhibit low correlations with one another. This can be attributed to the dataset being a

combination of both categorical and numerical data of multiple natures (demographic/ medical/ Boolean), which means that a combination of all multiple factors causes any particular disorder, not just one. This characteristic increases the complexity of training machine learning models, as the absence of strong correlations limits straightforward feature extraction and relationship modelling. Additionally, significant variability within classes complicates the problem even more. For example, different feature values are observed in patients with the same disorder subclass.

This variation makes it difficult for the simpler machine learning models. Given these features, the focus should instead be on non-linear models because linear classifiers are inappropriate for the task due to their weakness in interpreting complex data. This dataset's characteristics are best handled by models that implement ensemble and boosting strategies. These methods do particularly well in complex datasets through advanced feature selection and feature weighting strategies to control for complexity and variability. Combining weak learners or sequentially refining multiple classifiers enables the ensemble and boosting models to capture the subtle structures paramount for the classification and prediction tasks in the present dataset. The following section will move towards understanding the classifiers that can be used.

3.4.1. Working of XGB Classifier

XGB (Extreme Gradient Boosting) is an advanced gradient boosting algorithm that constructs multiple weak decision trees, with each tree improving on the errors of its predecessors. This iterative boosting process enhances both efficiency and accuracy. Each tree represents a function that is gradually optimized in the following iterations, as the next tree works to correct its predecessor's mistakes, thus leading to a drastic increase in the accuracy and robustness of the model over time, as shown in Equation (5).

Equation 5: Core function for the XGB classifier

$$Obj(\Theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where:

- $\sum_{i=1}^{n} L(y_i, \hat{y}_i)$ represents the loss function
- f_k marks a decision tree in the model.

$\Omega(f_k)$ is the regularization term that is used to prevent overfitting.

3.4.2. Working of Light Gradient Boosting Machine (LGBM) Classifier

The key mechanism of the LGBM Classifier involves constructing an ensemble of decision trees sequentially. Each tree corrects errors from the previous one by minimizing a loss function using gradient descent. The formula for the predicted output at iteration t is given in Equation (6).

Equation 6: Gradient Descent

$$\hat{y} = \hat{y}_{t-1} + \eta \cdot f_t(x)$$

Where:

- $\hat{y}_{t-1}$ is the prediction from the previous iteration.
- η is the learning rate that controls the contribution of each tree.
- $f_t(x)$ is the decision tree built at iteration *t.*

3.4.3. Working of Random Forest (RF)

RF is an ensemble method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Each tree is trained on a random subset of data and features, introducing diversity. For classification, predictions are based on the majority vote and the average of the outputs for regression. The process starts with a dataset, which is divided into subsets. A decision tree is built independently for each subset, resulting in multiple trees (n-trees). Each tree generates its classification output (Result-1, Result-2, etc.). The classifier then uses majority voting to aggregate these results, where the most common prediction among the trees becomes the final output.

Important Equations for RF:

- Gini Index (Impurity Measure): This is given by Equation (7), which measures node impurity, where lower values indicate purer nodes.

$$G = 1 - \sum_{i=1}^{C} p_i^2$$ *Equation 7: Finding the Gini index.*

- Bootstrap Sample: Equation (8) represents the creation of a bootstrap sample consisting of randomly selected data points (features and labels) from the original dataset to train each decision tree.

Equation 8: Bootstrap creation

$$D_b = \{(x_1, y_1), (x_2, y_2), .. (x_n, y_n)\}$$

Where:

- $x_i \wedge y_i$ are the features and labels of the data points, respectively.

- Majority Voting (Final Classification): Equation (9) shows how the final class prediction is determined by taking the most common class label among the predictions made by all decision trees in the forest.

Equation 9: Equation for a mode of a series

$$\hat{y} = mode(T_1(x), T_2(x), .., T_B(x))$$

Where:

- $T_1(x), T_2(x), T_3(x), .., T_B(x)$ are the class labels predicted by each of the B trees.

Mode selects the most frequent class among the tree outputs.

3.5. Model Evaluation

This classification task uses a dataset of high variability and class imbalance, so a robust evaluation

framework is necessary to ensure accurate model assessment. The primary concern while making the evaluation was keeping in mind the uneven class distribution, for which four metrics, namely accuracy score, precision and recall score, were considered. The confusion matrix in Table 7 clarifies all the parameters used in the evaluation metrics.

Table 7. Confusion matrix

	True Class	
Predicted class	TP: True Positive	FP: False Negative
	FP: False Positive	TP: True Negative

Accuracy measures the overall correctness of the model by calculating the proportion of correctly classified instances out of the total number of samples. Despite its simplicity, the accuracy score can be unreliable in cases of class imbalance. If one class dominates the dataset, a model may achieve high accuracy by simply predicting the majority class while failing to recognize minority class instances. To check the proportion of correct optimistic predictions, precision comes into play. A high precision score indicates that the model has a low false positive rate, which is crucial in scenarios where misclassification of the positive class is costly (e.g., medical diagnosis). Afterwards, the recall would aid us in measuring the ability of the model to correctly identify all actual positive instances in the dataset, which can be done via recall score. A high recall score means the model identifies the most positive instances, minimizing the number of false negatives. Using the F1 score in this context is extremely helpful because it ensures that false negatives and positives are accounted for. For this specific dataset, a high F1 score means that the model accurately identifies positive instances and avoids false positive errors, which is a high recall and precision. The evaluation also included cross-validated scores across multiple folds to guarantee that the given train-test data split did not bias the model's performance. Cross-validation directly evaluates how well the model will generalize to an independent dataset by training it on different portions and testing it on portions not used during training. This approach mitigates overfitting issues and ensures the model is robust across various data distributions.

3.6. Model Validation

After the evaluation, validation of the classifiers is highly crucial to ensure that the model follows the proper medical diagnostic logic rather than making superficial and unethical guesses, which is highly unsuitable for a system aimed at precision medicine. To aid with model comprehension, SHAP (Shapley Additive exPlanations) would be implemented on the entirety of the framework at every step of validation. This would help depict the relevance of specific variables in determining the presence of any disorder, thus allowing clinicians to comprehend the reasons underlying such predictions and making the system more useful. Two types of plots would be considered: the global bar plot and the local waterfall plot.

3.7. Global Bar and Summary Plots

The global summary bar chart SHAP plots would combine all feature importance for all data points into one summary. It will emphasize features such as the age of the patient, lab results, or medical history, which would have the most tremendous impact on determining whether or not a specific disorder exists. This overview helps to appreciate the general model's decision logic, which helps verify that the classifier functions on the medical and logical premises.

3.7.1. Local Waterfall Plot

These plots help summarize individual patient data, including all details and allow physicians detailed access to how certain features affect the prediction for every patient. Patients might have unique data, requiring special explanations to construct reliable diagnoses.

For example, a model determining whether a patient is likely to have a particular condition gives a reason for each prediction, and physicians can check the local SHAP values to see which symptoms or test results were most responsible for the prediction.

4. Development and Testing

4.1. Multiclass Classification System

A multiclass classifier can be developed using the methodology described in the earlier sections, which would then undergo rigorous testing and validation to prove its authenticity in the following sections.

4.2. Model Testing

A multiclass classifier was developed using an ensemble of models (LightGBM, XGBoost, and Random Forest) and tested over 5 folds of cross-validation over multiple evaluation parameters, as shown in Table 8.

Table 8. Evaluation of the ensemble models in developing the multiclass classifier using cross-validation.

Metric	LightGBM	XGBoost	Random Forest
Cross-Validation Mean F1-Score (Weighted)	0.92	0.92	0.91
Cross-Validation Standard Deviation	0	0	0
Test Set F1-Score	0.84	0.84	0.79
Test Set Accuracy	0.84	0.84	0.8
Macro Average F1-Score (Test)	0.81	0.8	0.75

Table 9. Results of evaluation metrics for the ensemble classifiers.

	LGBM			XGBoost			Random Forest		
Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Alzheimer's	0.71	0.52	0.6	0.59	0.45	0.51	0.46	0.21	0.29
Cancer	0.81	0.89	0.85	0.81	0.89	0.85	0.85	0.89	0.87
Cystic fibrosis	0.95	0.96	0.96	0.96	0.95	0.96	0.93	0.95	0.94
Diabetes	0.96	0.97	0.97	0.95	0.97	0.96	0.94	0.97	0.96
Hemochromatosis	0.78	0.63	0.7	0.78	0.62	0.69	0.77	0.5	0.6
Leber's hereditary optic neuropathy	0.86	0.82	0.84	0.89	0.81	0.85	0.85	0.75	0.8
Leigh syndrome	0.79	0.8	0.8	0.79	0.81	0.8	0.73	0.81	0.77
Mitochondrial myopathy	0.78	0.77	0.78	0.79	0.78	0.78	0.77	0.69	0.73
Tay-Sachs	0.81	0.87	0.84	0.8	0.88	0.84	0.74	0.85	0.8

A more detailed class-specific performance metric has been provided in Table 9, showing that the ensemble methods are the right choice for this framework, although they have a lot of grey areas in their class-specific performances that need to be corrected in future sections.

4.3. Validating the Multiclass Classifier

The summary plot of the multiclass classifier in Figure 5 shows the global feature importance of different genetic and demographic factors. Most genetic disorders are inherited from the parents hereditarily in either an autosomal dominant or autosomal recessive manner. Furthermore, advanced parental age can be associated with an increased risk of new genetic mutations in offspring (de novo). For example, older paternal age has been linked to a higher incidence of new mutations due to the continuous division of sperm cells over a man's lifetime. These are not inherited mutations but new changes that can lead to genetic disorders. Similarly, as individuals age, the cumulative effect of genetic predispositions and environmental factors can increase the risk of developing genetic disorders. For instance, the risk of Alzheimer's disease and other age-related conditions can be influenced by both genetic factors and ageing processes.

The model prioritizes genetic inheritance (maternal and paternal genes), parental age, and status, reinforcing its alignment with known hereditary disease patterns. With the overall testing now done, individual testing would be prioritized.

This classifier has been tested on hundreds of patient data sets with satisfactory results; two of them have been discussed below: a patient with Alzheimer's and a patient with Leigh Syndrome. Figure 6 shows that the "paternal gene" is given the highest priority, followed by the "maternal gene". Subsequently, in Figure 7, all the maternally linked features have been given the highest priority for detecting a patient with Leigh syndrome (i.e., a maternally inherited disorder). Thus, it can be confirmed that the multiclass classifier makes logical decisions instead of unethical guesses.

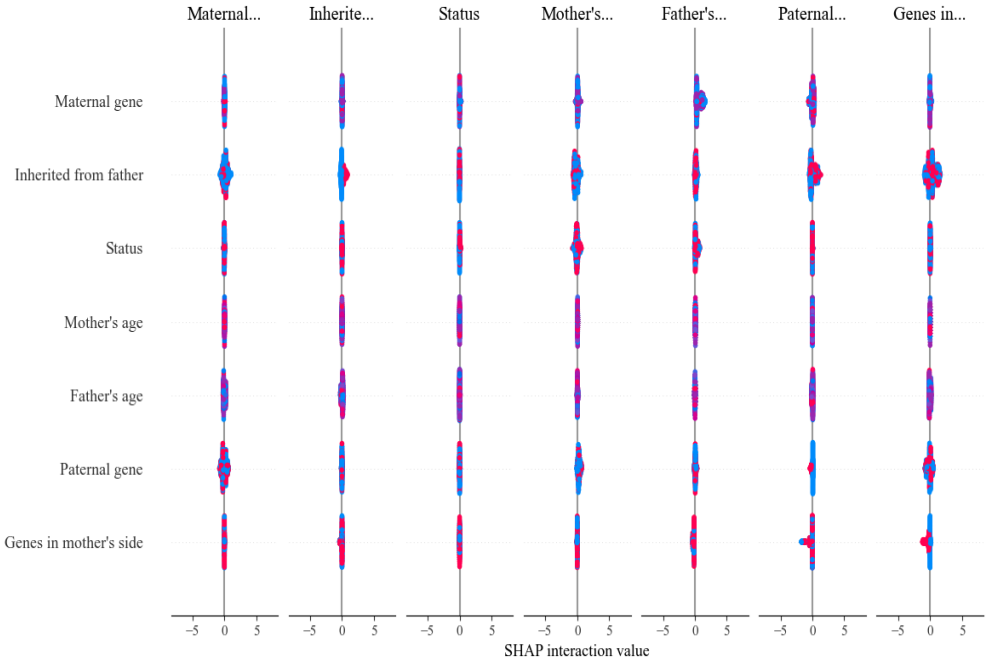

Fig. 5 Summary plot for multiclass classifier

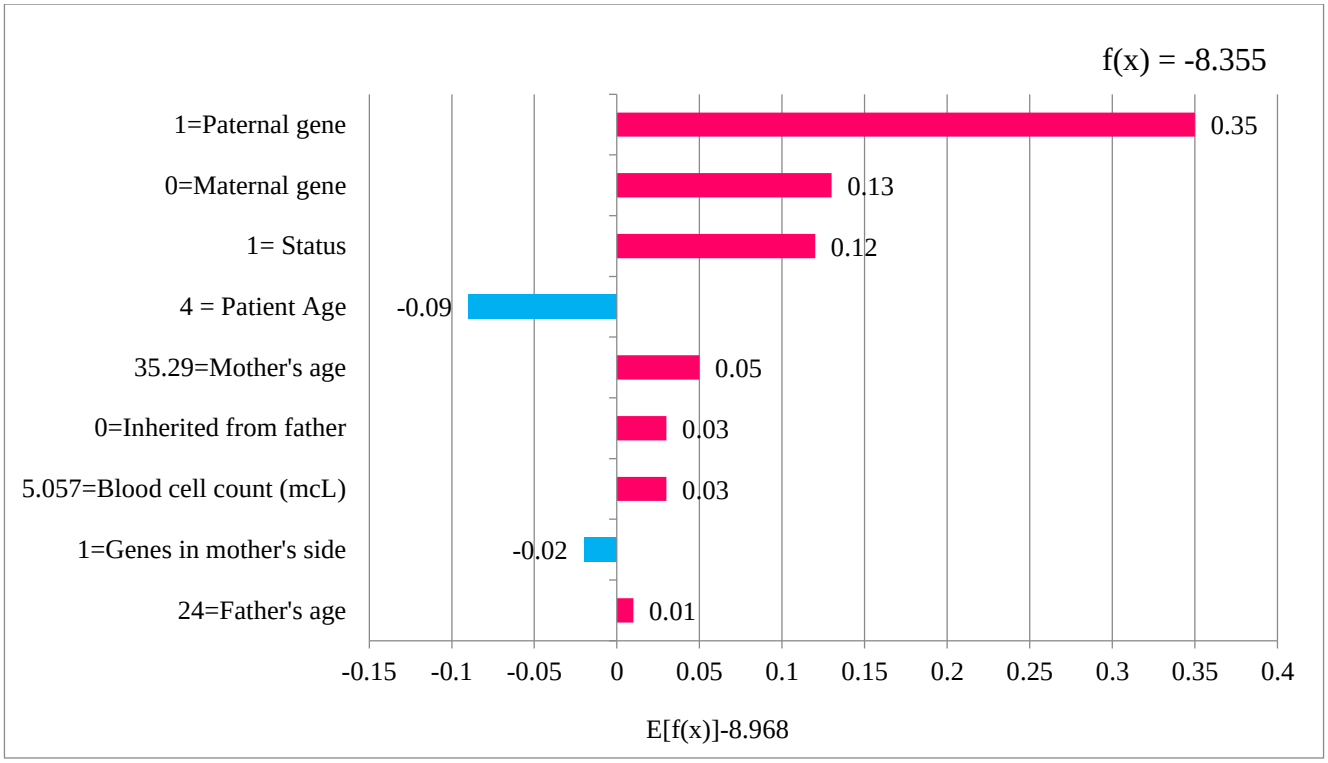

Fig. 6 Waterfall plot for a patient with alzheimer's disease

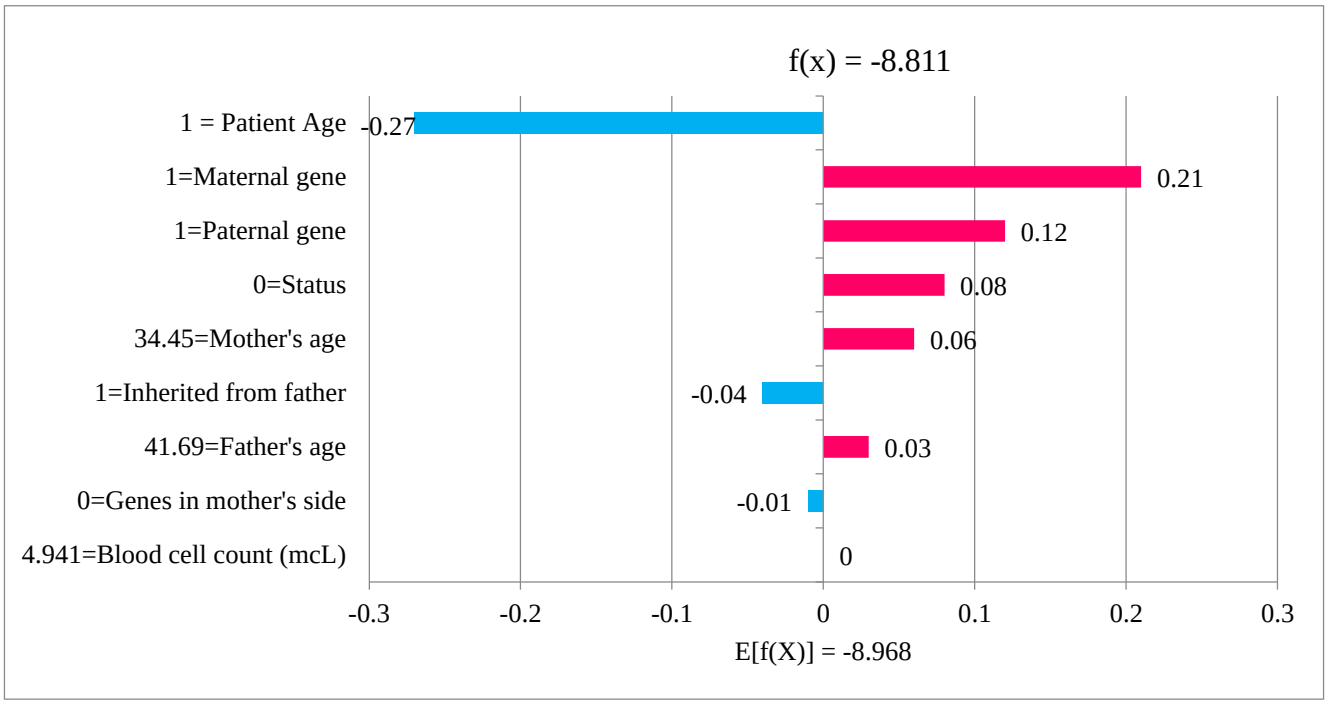

Fig. 7 Waterfall plot for a patient diagnosed with leigh syndrome

The waterfall plot in Figure 6 for a patient diagnosed with Alzheimer's demonstrates that parental genes contribute the most to the prediction, followed by genetic factors across the family, as depicted in Table 2. Using this logic, the actual cause of the disorder among the patients can be verified, and new biomarkers can even be found that are yet to be uncovered.

Following a similar analogy as Figure 6, figure 7 also reveals that the genetic and age-related factors play the most significant role in the prediction. This confirms the model's reliability in identifying genetic disorders by correctly emphasizing key features associated with Leigh Syndrome's mitochondrial inheritance pattern.

4.4. Understanding the Problems

The previous sections were focused on developing the multiclass classifier, but it had a critical flaw that hampered its classification ability due to the severe data imbalance of this dataset, as shown in Table 10.

Table 10. Predictive capacity of multiclass classifier

Class	F1-Score
Alzheimer's	0
Cancer	0.83
Cystic Fibrosis	0.94
Diabetes	0.93
Hemochromatosis	0.33
LHON	0.76
Leigh Syndrome	0.61
Mitochondrial myopathy	0.58
Tay Sachs	0.75

For specific disorders, the classifier had high precision and recall abilities, while other disorders were severely misclassified due to class imbalance. This was a primary constraint because a product designed for medical diagnosis needs to work accurately for all medical conditions. The first revision used the One Versus Rest Classifier (OVRC) algorithm, separating multiclass classification problems into several binary classification problems. This method involves fitting a classifier for each disorder's prediction, predicting one disorder against all others. This allows for more specific and intricate predictions since this approach is advantageous when handling class imbalances or rare disorder categories. Nonetheless, even with ensemble techniques like LightGBM as the base estimator for the OVR algorithm, it enhanced the performance relative to the standalone LightGBM classifier, providing only slight performance improvements.

4.5. Introducing the BMMDC

Considering how grave a misdiagnosis could be and understanding that simplistic machine learning algorithms would not make it suitable for diagnostics, it was critical to pursue a different classification approach to improve predictive efficacy. The development of the Binary Multi-Model Disorder Classifier (BMMDC) is a predictive mechanism that emerged from the need to improve the accuracy and generalizability of genetic disorder predictors, as illustrated in Figure 8.

4.5.1. Data Encoding/ Binarization

BMMDC uses an ensemble of 9 binary classifiers independently trained on separate disorder prediction tasks instead of a single multiclass classifier. This approach synergizes the merits of ensemble and binary classification, producing a more precise and focused model. The design rationale for BMMDC was to allow equitable attention to each disorder throughout training, thereby identifying the adverse impact of class imbalance suffered by previous models. The first step in operationalizing the BMMDC framework was redesigning the training dataset to fit a binary classification approach. In the prepared dataset, the target column was modified so that each disorder-specific classifier had separate "positive" instances for that particular disorder and all other instances as "other". For example, while training a classifier for Leigh Syndrome, all patients diagnosed with Leigh Syndrome were marked as positive while those with any other disorder were placed in "other." This process was repeated for all nine disorders to create nine data frames.

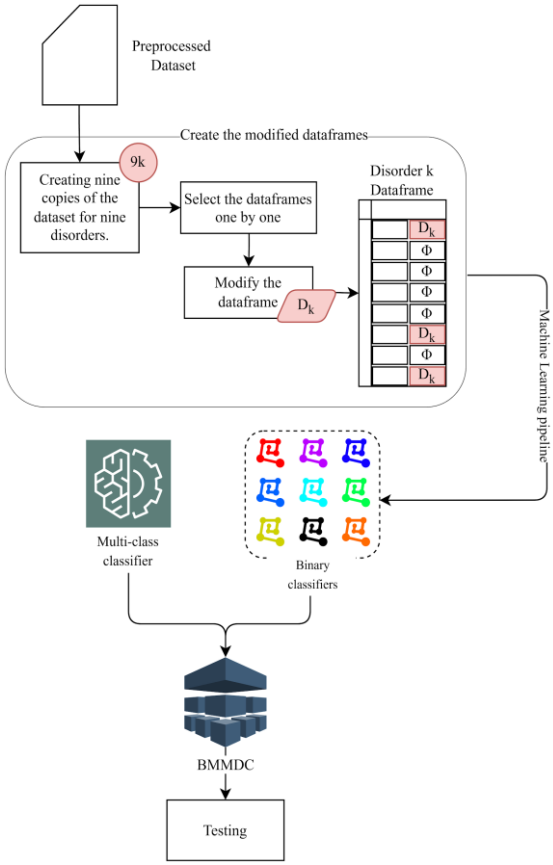

Fig. 8 Data binarization and BMMDC development

4.5.2. Data Balancing

The above encoding created an additional challenge of class imbalance, as some disorders like Cancer and Alzheimer's had significantly fewer occurrences. To address this challenge, a mobile balancing strategy was applied whereby a sample equal to the positive cases of the target disorder from the "other" category was selected randomly to match the target number. This ensured that each classifier was trained on an optimally balanced dataset, which would otherwise lead to unwanted bias towards the overrepresented class and improve generalization. Moreover, the chances of overfitting were low because even the smallest class still contained 194 instances across 33 features, providing sufficient data for robust model training.

4.5.3. Model Training

After structuring the datasets appropriately, nine binary classifiers were created to correspond to a specific disorder. As potential base models, XGBoost and LGBM classifiers were evaluated, and their performance was analyzed in terms of precision, recall, and F1-score, as discussed in previous sections. It can be seen from the

results presented in Tables 8 and 9 that both models performed relatively well; however, LGBM slightly surpassed XGBoost in precision and recall, justifying its use in BMMDC. Given that precision is paramount in medical diagnostics (as false positives could lead to unnecessary treatments and patient distress), LGBM was finalized as the base classifier for all nine binary models. With the binary classifiers in place, the decision-making logic of BMMDC was formulated to ensure optimal predictive performance. The fundamental principle governing the classifier's predictions is that binary classifier outputs take precedence over the multiclass classifier. This means that if a Leigh Syndrome binary classifier returns a "True" prediction, it is given priority over any conflicting multiclass classification output. The rationale behind this prioritization is that a disorder-specific binary classifier, trained exclusively to detect a particular condition, is inherently more specialized and accurate than a generalized multiclass model.

4.5.4. Evaluating the Models

The binary classifiers are evaluated here as well, using the same metrics as before, and a drastic improvement in accuracy can be observed here compared to the previous. Both models achieve exceptionally high accuracy, consistently exceeding 0.94 for all disorders. This indicates that the BMMDC framework effectively distinguishes between the presence and absence of different disorders with high reliability. Furthermore, from Tables 11 and 12, it can be concluded that this approach has successfully countered class-specific performance problems.

Table 11. The performance of BMMDC when using XGBoost

Disorder	Precision	Recall	F1-Score	Accuracy
LHON	0.94	0.94	0.94	0.94
Cystic Fibrosis	0.97	0.97	0.97	0.97
Diabetes	0.99	0.99	0.99	0.99
Leigh Syndrome	0.9	0.89	0.89	0.89
Cancer	1	0.99	0.99	0.99
Tay-Sachs	0.96	0.96	0.96	0.96
Hemochromatosis	0.95	0.95	0.95	0.95
Mitochondrial Myopathy	0.9	0.9	0.9	0.9
Alzheimer's	0.96	0.96	0.96	0.96

Table 12. The performance of the BMMDC when using LGBM

Disorder	Precision	Recall	F1-Score	Accuracy
LHON	0.93	0.93	0.93	0.93
Cystic Fibrosis	0.97	0.97	0.97	0.97
Diabetes	0.99	0.99	0.99	0.99
Leigh Syndrome	0.91	0.9	0.9	0.9
Cancer	1	0.99	0.99	0.99
Tay-Sachs	0.96	0.96	0.96	0.96
Hemochromatosis	0.95	0.94	0.94	0.94
Mitochondrial Myopathy	0.91	0.91	0.91	0.91
Alzheimer's	0.96	0.96	0.96	0.96

Here, it can be seen that LGBM is more suited for disorder prediction than XGB, even just by a small margin.

4.5.5. Validating the Models

However, as done in the multiclass classifier, this section also calls for its need for validations to be suitable for medicine. Like the multiclass classifier, the SHAP framework has been used to explain binary classifier results. This step significantly increased the trustworthiness of the binary classifiers, considering its relatively new approach. Like the previous section, the following section collectively displays the SHAP plots for the binary classifiers. Due to the large number of test cases for all nine disorders, only two of the 9 disorders here would be considered for validation.

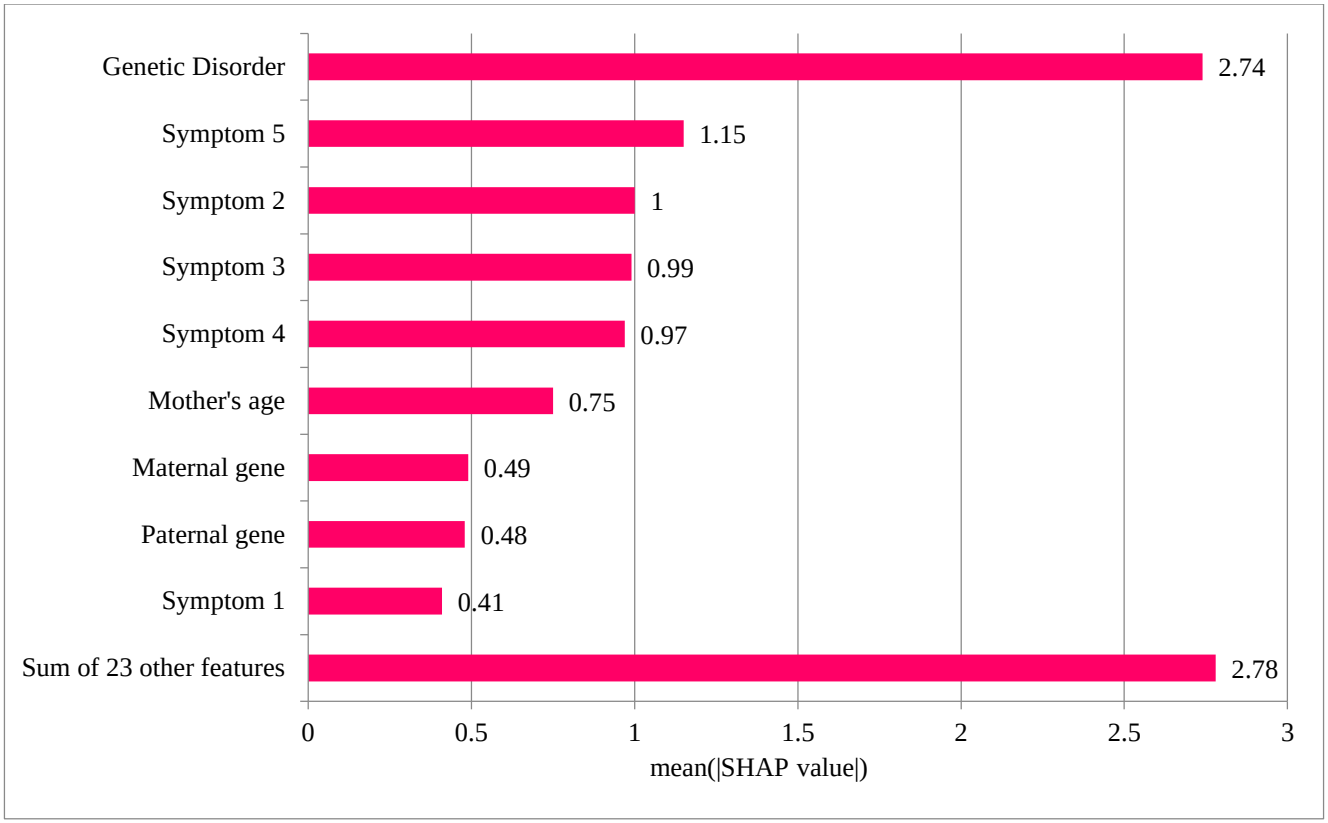

Fig. 9 Global bar plot for alzheimer's classifier

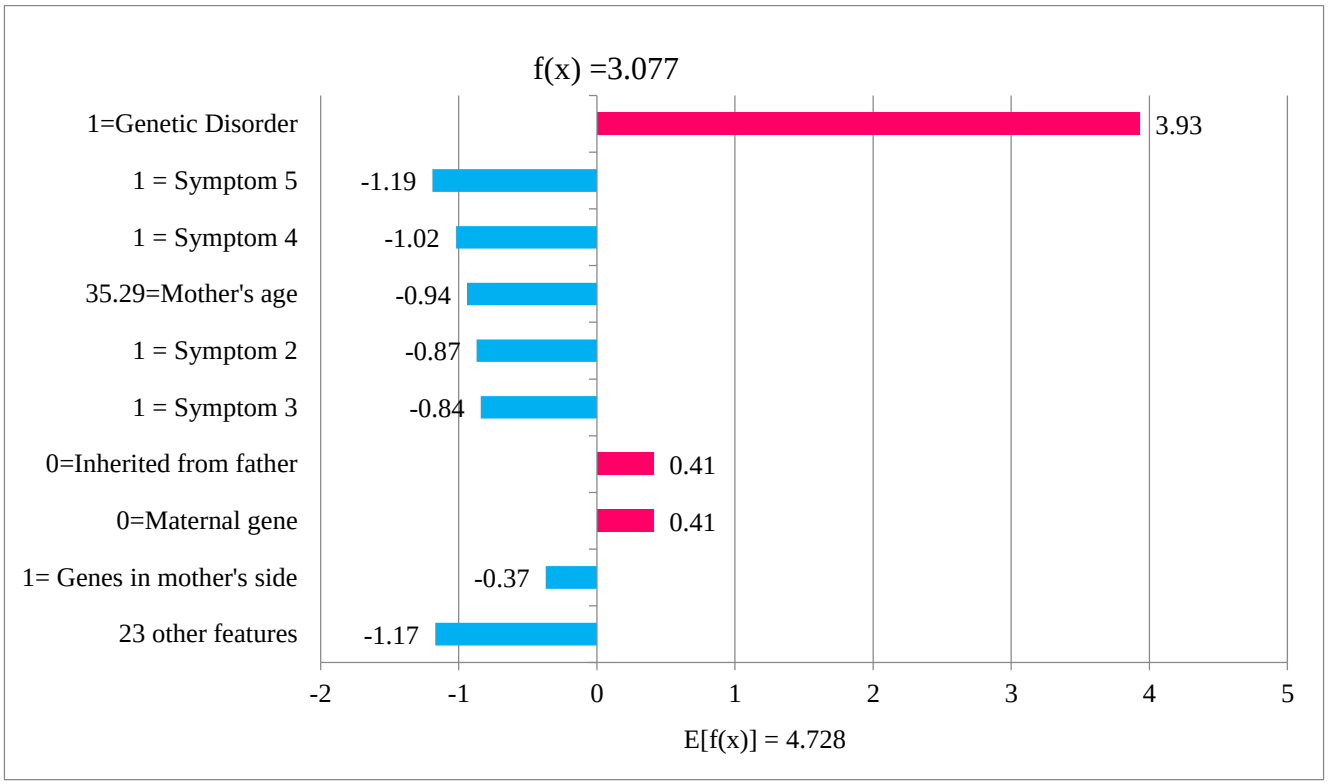

Fig. 10 Waterfall plot for a patient with alzheimer's

Figure 9 shows that symptoms, maternal age and genetic factors have been given the highest priority in making the diagnosis of Alzheimer's, which is correct according to the diagnostic logic.

Figure 10 illustrates the model's reasoning when predicting Alzheimer's for a specific patient with Alzheimer's, following the same prioritization pattern as observed in the global bar plot in Figure 9, which confirms it is a logical diagnosis.

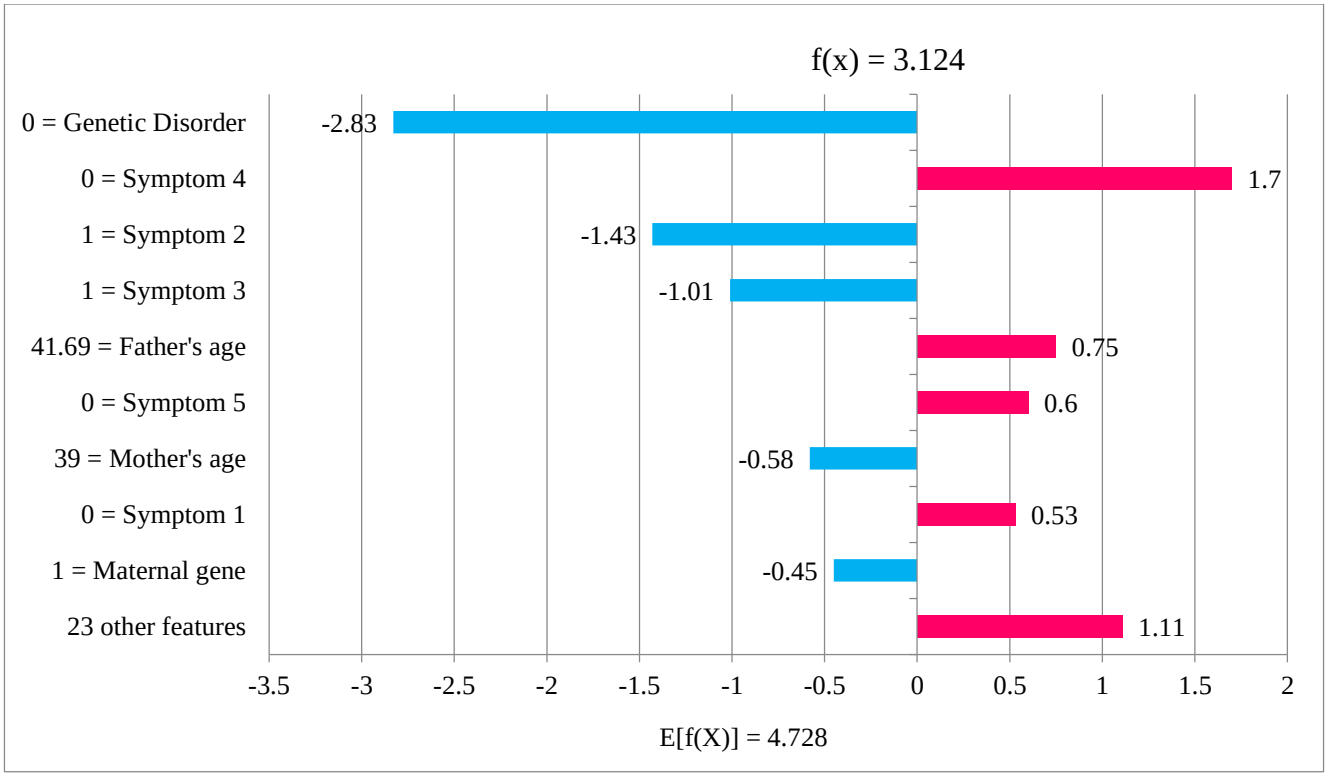

Fig. 11 Waterfall plot for a patient without alzheimer's

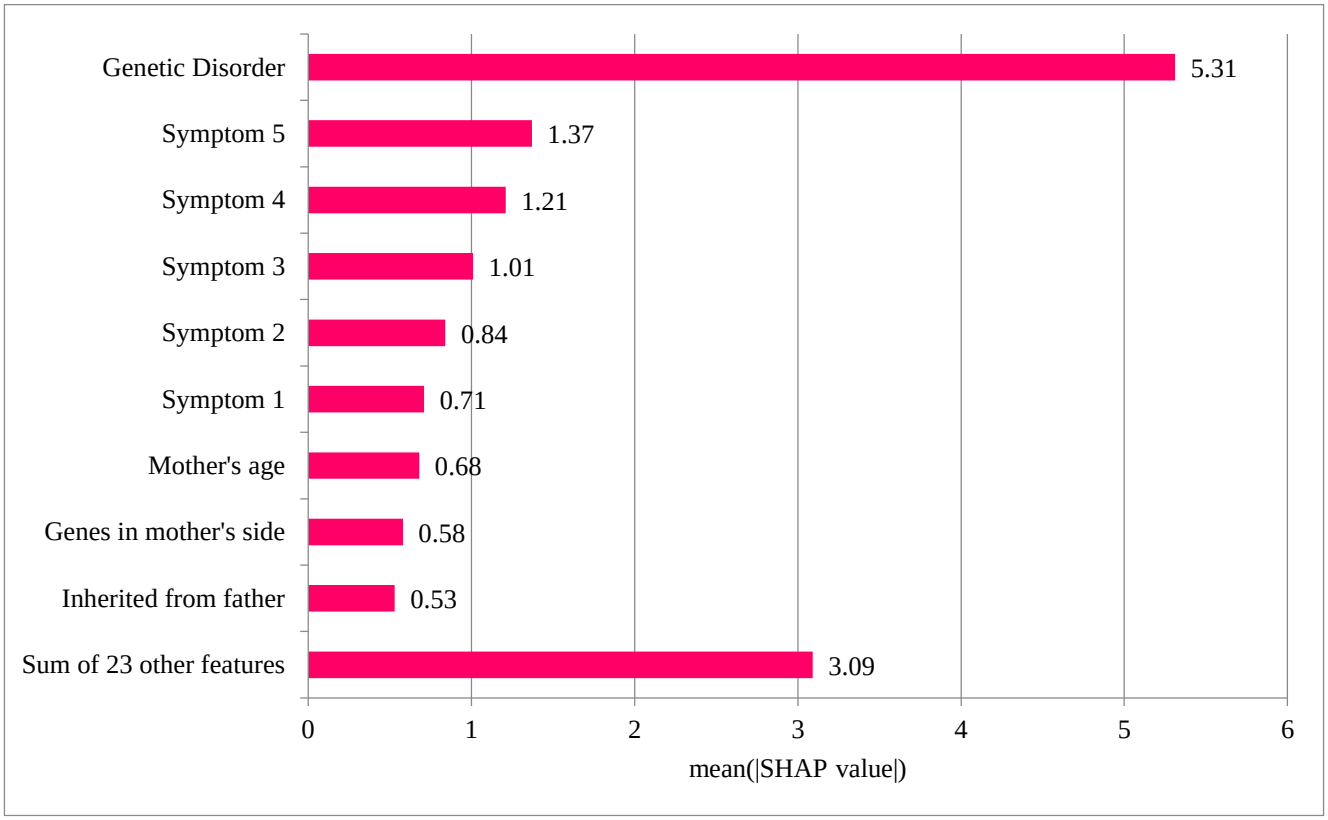

Fig. 12 Global bar plot for cancer classifier

Figure 11 shows that when the classifier is applied to a patient without Alzheimer's, the feature importance rankings are completely inverted compared to Figure 10, ensuring that the model correctly identifies the absence of Alzheimer's and suggests an alternative diagnosis.

Figure 12, just as 11, shows that symptoms and inheritance patterns have been given the highest priority in determining the diagnosis, which is correct according to the diagnostic logic.

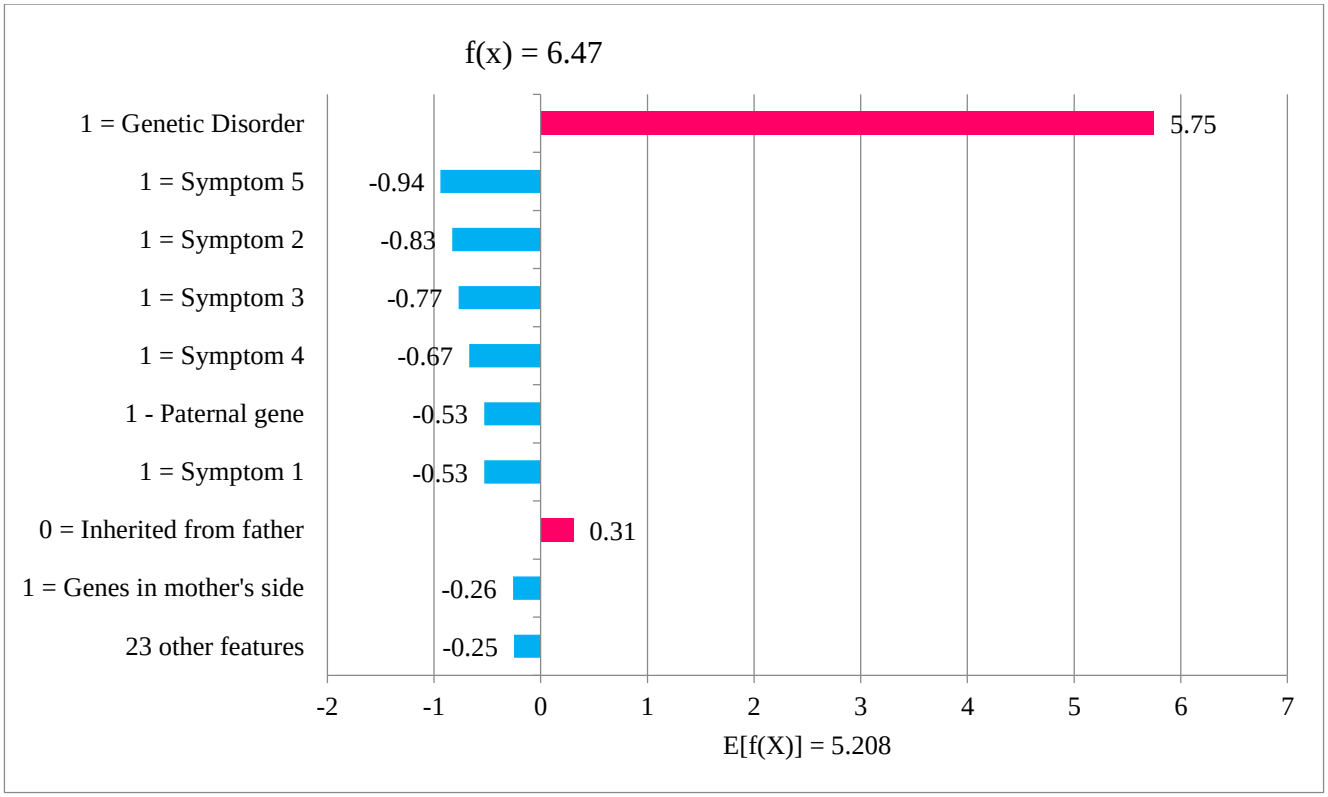

Fig. 13 Waterfall plot for a patient with cancer

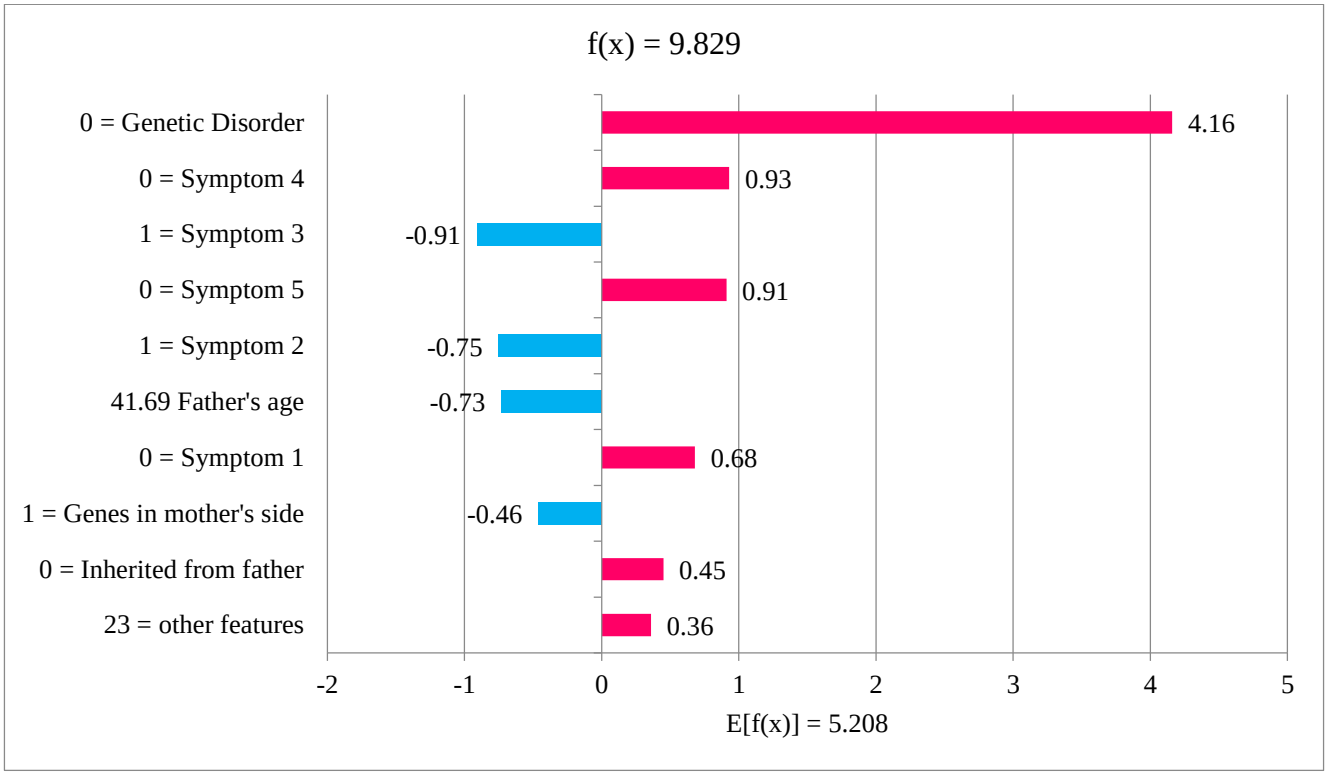

Fig. 14 Waterfall plot for a patient without cancer

Figure 13 illustrates the most significant factors influencing the model's decision. The classifier prioritizes symptoms as the primary diagnostic criteria, followed by genetic information consistent with standard cancer diagnostic strategies. The decision is driven by all key symptoms, confirming the model's reasoning that symptomatology is the primary factor leading to cancer prediction.

Again, just like Figure 11, Figure 14 also shows a drastic change in the feature contributions, thus leading to this particular diagnosis.

5. Results and Discussion

The outcomes from the multiclass classifier and the Binary Multi-Model Disorder Classifier (BMMDC) offer a critical understanding of machine learning's capability in diagnosing genetic disorders. The SHAP analysis provided a breakdown of the important features in analyzing the model's decision-making. The classifiers utilize inheritance factors, parental age, and symptoms, reinforcing medical trust. Among other features, the assessment of feature importance Figures 5-7 showed that genetic components, especially maternal and paternal genes, were unrelenting to be the most impactful in decision-making for the multiclass classifier. The classifier accurately detected the inheritance patterns of Alzheimer's and Leigh Syndrome, proving its ability to recognize the hereditary traits.

The binary classifiers in BMMDC, on the other hand, provided greater refinement in disorder identification. Training each classifier independently on one disorder protected the system from misclassification risks due to class imbalance. As Figures 9-14 illustrated, the binary classifiers displayed rational and logical choices focusing on clinical symptomatology, genetic indicators, and other socio-demographic variables. Nevertheless, feature selections, such as radiation exposure history in predicting cystic fibrosis, suggest that this classifier needs refinement. The classifiers make reasonable and medically relevant predictions while providing unconditional support to clinical judgment. The outcomes of this study have been critically assessed using the works of other researchers.

Equation 1: Comparative Analysis Based on State-of-the-Art Criteria

Criterion	Nasir et al. (2022) [10]	Singh et al. (2024)	Raza et al. (2023) [9]	Proposed Study
Dataset & Domain	Genetic Disorder	Genomic Disorder Prediction	Genetic Disorder	Genetic Disorder
Model / Methodology	ANN, SVM, KNN; Linear regression	Feynman Concordance & Interpolated Nearest Centroid	XGBoost / Chain classifiers + Extra Trees	LightGBM / Binarization + Ensemble Learning
Key Metrics	85.7% training accuracy, 84.9% testing	Concordance Index: 0.89	92% α-score, 84% macro accuracy	Precision: 0.953; Recall: 0.953; F1-Score: 0.95; Accuracy: 0.95
Unique Contributions	Linear regression for feature selection	Novel use of Feynman Concordance; advanced centroid interpolation	Class probabilities from ET/RF as hybrid features	Hybrid binary classification ensemble approach
Limitations	Limited interpretability, smaller feature set	Potential complexity of methodology; dependent on genomic data availability	Computationally intensive for large datasets	There is slight computational overhead due to multiple classifiers
Clinical Use Case	Early screening using medical history	Enhanced genomic prediction accuracy for disorders	Multi-disorder subclass prediction	Multi-disorder subclass prediction
Scalability	Limited to pre-selected features	Moderate (requires robust genomic data)	Moderate (balanced datasets needed)	Highly scalable, effective even with minimal datasets
Interpretability	Low (black-box ANN models)	Moderate (specific genomic insights provided)	Moderate (feature importance maps)	High (SHAP-based explainability)

The comparative analysis shows that the proposed framework outperforms the other studies in multiple fields, thus making it useful in precision medicine.

6. Conclusion

This research highlights the great promise of machine learning in solving the problems encountered in diagnosing genetic disorders, particularly regarding accuracy and data imbalance issues. This study formulated a new framework called the BMMDC or the Binary Multi-Model Disorder Classifier by applying robust ensemble techniques such as Random Forest, LightGBM, and XGBoost. Its primary intention was to design the model so that even with sub-optimal classification decisions, the best-case scenario would narrow the output to two possible disorders, at worst, three in very few cases (~1/80). This would assist effectively in the diagnosis of the disorders. Thus, a genetic disorder classifier with an average accuracy of 95%, which, in comparison with other methods currently available, is remarkably high, has been created. Thus, this system can become acceptable for clinical deployment for the commonwealth.

6.1. Future scope

Despite the noteworthy accomplishments of the BMMDC, there is still some need for improvements to overcome its specific limitations and foreseen enhancements, such as false positives. The enhancement of model performance aimed at rare genetic disorders will be the primary focus of the upcoming research, as it will employ data augmentation approaches and seek new methods to address the problem of extreme class imbalance. Furthermore, infusing real-time patient data and health records may enhance the model's applicability and robustness. With the advancement of deep learning, we can develop even more accurate and reliable prediction systems, and more interpretability techniques can be developed to shed more light on the importance of the features and their interrelations to understand the cause and effect in an even more transparent way. In conclusion, transitioning BMMDC from a local tool to a cloud-enabled system or an integrated electronic medical records system design would support clinical practitioners' rapid and easy adaptation for wide-scale utilization in genetic screening and individualized medicine implementation.

References

[1] David W. Pfennig, *Phenotypic Plasticity & Evolution: Causes, Consequences, Controversies*, Taylor & Francis, 2021. [Google Scholar] [Publisher Link]

[2] Tiffany A. Kosch et al., "Genetic Approaches for Increasing Fitness in Endangered Species," *Trends in Ecology & Evolution*, vol. 37, no. 4, pp. 332-345, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mor Hanany, Carlo Rivolta, and Dror Sharon, "Worldwide Carrier Frequency and Genetic Prevalence of Autosomal Recessive Inherited Retinal Diseases," *Proceedings of the National Academy of Sciences*, vol. 117, no. 5, pp. 2710-2716, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Nicholas Lench et al., "The Clinical Implementation of Non-Invasive Prenatal Diagnosis for Single-Gene Disorders: Challenges and Progress Made," *Prenatal Diagnosis*, vol. 33, no. 6, pp. 555-562, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[5] Dahlak Daniel Solomon et al., "Extensive Review on the Role of Machine Learning for Multifactorial Genetic Disorders Prediction," *Archives of Computational Methods in Engineering*, vol. 31, pp. 623-640, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Gráinne S. Gorman et al., "Mitochondrial Diseases," *Nature Reviews Disease Primers*, vol. 2, pp. 1-22, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sameer Quazi, "Artificial Intelligence and Machine Learning in Precision and Genomic Medicine," *Medical Oncology*, vol. 39, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sreya Vadapalli et al., "Artificial Intelligence and Machine Learning Approaches using Gene Expression and Variant Data for Personalized Medicine," *Briefings in Bioinformatics*, vol. 23, no. 5, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ali Raza et al., "Predicting Genetic Disorder and Types of Disorder Using Chain Classifier Approach," *Genes*, vol. 14, no. 1, pp. 1-31, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Muhammad Umar Nasir et al., "Single and Mitochondrial Gene Inheritance Disorder Prediction using Machine Learning," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 953-963, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Zodwa Dlamini et al., "Artificial Intelligence (AI) and Big Data in Cancer and Precision Oncology," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2300-2311, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Catriona Miller et al., "A Review of Model Evaluation Metrics for Machine Learning in Genetics and Genomics," *Frontiers in Bioinformatics*, vol. 4, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sofia Singh et al., "Enhancing Genomic Disorder Prediction through Feynman Concordance and Interpolated Nearest Centroid Techniques," *Scientific Reports*, vol. 14, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] C.R. Haldeman-Englert et al., *GeneReviews®*, Europepmc, 1993. [Google Scholar] [Publisher Link]

[15] Shuo Yang et al., "Long-Term Outcomes of Gene Therapy for the Treatment of Leber's Hereditary Optic Neuropathy," *EBioMedicine*, vol. 10, pp. 258-268, 2016. [Google Scholar] [Publisher Link]

[16] Shibani Kanung et al., "Mitochondrial Disorders," *Annals of Translational Medicine*, vol. 6, no. 24, pp. 1-27, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] Svetlana Yovinska et al., "e-Posters EP01 Reproductive Genetics," *European Journal of Human Genetics*, vol. 31, pp. 91-344, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Bryce A. Pasqualotto et al., "Galactose-Replacement Unmasks the Biochemical Consequences of the G11778A Mitochondrial DNA Mutation of LHON in Patient-Derived Fibroblasts," *Experimental Cell Research*, vol. 439, no. 1, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mohammed Hasan Barrak et al., "Pathophysiology, the Biochemical and Clinical Significance of Lactate Dehydrogenase," *International Journal of Health and Medical Research*, vol. 3, no. 7, pp. 440-443, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jia-Der Ju Wang et al., "Sleep and Breathing Disturbances in Children with Leigh Syndrome: A Comparative Study," *Pediatric Neurology*, vol. 136, pp. 56-63, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Inn-Chi Lee, and Kuo-Liang Chiang, "Clinical Diagnosis and Treatment of Leigh Syndrome Based on SURF1: Genotype and Phenotype," *Antioxidants*, vol. 10, no. 12, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Albert Z. Lim et al., "Natural History of Leigh Syndrome: A Study of Disease Burden and Progression," *Annals of Neurology*, vol. 91, no. 1, pp. 117-130, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[23] Klervie Loiselet et al., "Cerebral Blood Flow and Acute Episodes of Leigh Syndrome in Neurometabolic Disorders," *Developmental Medicine & Child Neurology*, vol. 63, no. 6, pp. 705-711, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[24] Grayson Beecher, Mark D. Fleming, and Teerin Liewluck, "Hereditary Myopathies Associated with Hematological Abnormalities," *Muscle & Nerve*, vol. 65, no. 4, pp. 374-390, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[25] Gina Perez Giraldo et al., "Neuromyelitis Optica Spectrum Disorder With Comorbid Complex 1 Mitochondrial Disease, Leukopenia And Neutropenia, A Challenging Case With Difficult Management (P7-1.005)," *Neurology*, vol. 98, no. 18, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[26] Tina D. Jeppesen et al., "Exercise Testing, Physical Training and Fatigue in Patients with Mitochondrial Myopathy Related to mtDNA Mutations," *Journal of Clinical Medicine*, vol. 10, no. 8, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[27] Kavya Bharathidasan et al., "Mitochondrial Myopathy in a 21-Year-Old Man Presenting with Bilateral Lower Extremity Weakness and Swelling," *Journal of Primary Care & Community Health*, vol. 14, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[28] Puneet K. Samaiya, Sairam Krishnamurthy, and Ashok Kumar, "Mitochondrial Dysfunction in Perinatal Asphyxia: Role in Pathogenesis and Potential Therapeutic Interventions," *Molecular and Cellular Biochemistry*, vol. 476, no. 12, pp. 4421-4434, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[29] Vincenzo Tragni et al., "Personalized Medicine in Mitochondrial Health and Disease: Molecular Basis of Therapeutic Approaches based on Nutritional Supplements and Their Analogs," *Molecules*, vol. 27, no. 11, pp. 1-49, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[30] Yi Shiau Ng et al., "Mitochondrial Disease in Adults: Recent Advances and Future Promise," *The Lancet Neurology*, vol. 20, no. 7, pp. 573-584, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[31] Po-Yu Lin et al., "Exacerbation of Myopathy Triggered by Antiobesity Drugs in a Patient with Multiple Acyl-CoA Dehydrogenase Deficiency," *BMC Neurology*, vol. 21, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[32] Raziye Melike Yildirim, and Emre Seli "Mitochondria as Therapeutic Targets in Assisted Reproduction," *Human Reproduction*, vol. 39, no. 10, pp. 2147-2159, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[33] Alessia Adelizzi et al., "Fetal and Obstetrics Manifestations of Mitochondrial Diseases," *Journal of Translational Medicine*, vol. 22, pp. 1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[34] Rajeev Bhatia, Bruce H. Cohen, and Neil L. McNinch, "A Novel Exercise Testing Algorithm to Diagnose Mitochondrial Myopathy," *Muscle & Nerve*, vol. 63, no. 5, pp. 715-723, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[35] Yaqi Wang et al., "The Relationship between Erythrocytes and Diabetes Mellitus," *Journal of Diabetes Research*, vol. 2021, no. 1, pp. 1-9, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[36] Atsuhiko Kawabe et al., "WBC Count Predicts Heart Failure in Diabetes and Coronary Artery Disease Patients: A Retrospective Cohort Study," *ESC Heart Failure*, vol. 8, no. 5, pp. 3748-3759, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[37] Giovanni Corona et al., "Diabetes is Most Important Cause for Mortality in COVID-19 Hospitalized Patients: Systematic Review and Meta-Analysis," *Reviews in Endocrine and Metabolic Disorders*, vol. 22, pp. 275-296, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[38] Antonino Tuttolomondo et al., "Assessment of Heart Rate Variability (HRV) in Subjects with Type 2 Diabetes Mellitus with and Without Diabetic Foot: Correlations with Endothelial Dysfunction Indices and Markers of Adipo-Inflammatory Dysfunction," *Cardiovascular Diabetology*, vol. 20, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[39] Chun-Hua Liu et al., "Effect of Birth Asphyxia on Neonatal Blood Glucose during the Early Postnatal Life: A Multi-Center Study in Hubei Province, China," *Pediatrics & Neonatology*, vol. 64, no. 5, pp. 562-569, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[40] Omid Asbaghi et al., "Folic Acid Supplementation Improves Glycemic Control for Diabetes Prevention and Management: A Systematic Review and Dose-Response Meta-Analysis of Randomized Controlled Trials," *Nutrients*, vol. 13, no. 7, pp. 1-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[41] Asher Ornoy et al., "Diabetes During Pregnancy: A Maternal Disease Complicating the Course of Pregnancy with Long-Term Deleterious Effects on the Offspring. A Clinical Review," *International Journal of Molecular Sciences*, vol. 22, no. 6, pp. 1-36, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[42] Jagdish Gopal Paithankar, Subash Chandra Gupta, and Anurag Sharma, "Therapeutic Potential of Low Dose Ionizing Radiation against Cancer, Dementia, and Diabetes: Evidences from Epidemiological, Clinical, and Preclinical Studies," *Molecular Biology Reports*, vol. 50, pp. 2823-2834, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[43] Siyu Zhu et al., "Adverse Childhood Experiences and Risk of Diabetes: A Systematic Review and Meta-Analysis," *Journal of Global Health*, vol. 12, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[44] Yehia Moustafa Ghanem et al., "Potential Risk of Gestational Diabetes Mellitus in Females Undergoing in Vitro Fertilization: A Pilot Study," *Clinical Diabetes and Endocrinology*, vol. 10, no. 1, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Linn Håkonsen Arendt et al., "Glycemic Control in Pregnancies Complicated by Pre-Existing Diabetes Mellitus and Congenital Malformations: A Danish Population-Based Study," *Clinical Epidemiology*, pp. 615-626, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[46] Matias Vaajala et al., "Previous Induced Abortion or Miscarriage is Associated with Increased Odds for Gestational Diabetes: A Nationwide Register-Based Cohort Study in Finland," *Acta Diabetologica*, vol. 60, pp. 845-849, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[47] Stephanie L. Marchincin et al., "Risk of Birth Defects by Pregestational Type 1 or Type 2 Diabetes: National Birth Defects Prevention Study, 1997–2011," *Birth Defects Research*, vol. 115, no. 1, pp. 56-66, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[48] Fausto Petrelli et al., "Red Blood Cell Transfusions and the Survival in Patients with Cancer Undergoing Curative Surgery: A Systematic Review and Meta-Analysis," *Surgery Today*, vol. 51, pp. 1535-1557, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[49] Pradeep S. Virdee et al., "The Association between Blood Test Trends and Undiagnosed Cancer: A Systematic Review and Critical Appraisal," *Cancers*, vol. 16, no. 9, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[50] Sakiko Fukui et al., "Association between Respiratory and Heart Rate Fluctuations and Death Occurrence in Dying Cancer Patients: Continuous Measurement with a Non-Wearable Monitor," *Supportive Care in Cancer*, vol. 30, pp. 77-86, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[51] Yu-Jie Su et al., "Prevalence and Risk Factors Associated with Birth Asphyxia Among Neonates Delivered in China: A Systematic Review and Meta-Analysis," *BMC Pediatrics*, vol. 24, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[52] Haidara Kherbek et al., "The Relationship between Folic Acid and Colorectal Cancer; A Literature Review," *Annals of Medicine and Surgery*, vol. 80, pp. 1-4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[53] Audrey Bonaventure et al., "Maternal Illnesses during Pregnancy and the Risk of Childhood Cancer: A Medical-Record based Analysis (UKCCS)," *International Journal of Cancer*, vol. 156, no. 5, pp. 920-929, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[54] Sheikh Ahmad Umar, and Sheikh Abdullah Tasduq, "Ozone Layer Depletion and Emerging Public Health Concerns-An Update on Epidemiological Perspective of the Ambivalent Effects of Ultraviolet Radiation Exposure," *Frontiers in Oncology*, vol. 12, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[55] Xu Ji et al., "Substance Use, Substance Use Disorders, and Treatment in Adolescent and Young Adult Cancer Survivors—Results from a National Survey," *Cancer*, vol. 127, no. 17, pp. 3223-3231, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[56] M. Condorelli et al., "Impact of ARTs on Oncological Outcomes in Young Breast Cancer Survivors," *Human Reproduction*, vol. 36, no. 2, pp. 381-389, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[57] Shinje Moon, Ka Hee Yi, and Young Joo Park, "Risk of Adverse Pregnancy Outcomes in Young Women with Thyroid Cancer: A Systematic Review and Meta-Analysis," *Cancers*, vol. 14, no. 10, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[58] Anders Husby, Jan Wohlfahrt, and Mads Melbye, "Pregnancy Duration and Ovarian Cancer Risk: A 50-Year Nationwide Cohort Study," *International Journal of Cancer*, vol. 151, no. 10, pp. 1717-1725, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[59] Jessica S.W Borgers et al., "Immunotherapy for Cancer Treatment during Pregnancy," *The Lancet Oncology*, vol. 22, no. 12, pp. 550-561, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[60] Le-Tian Huang et al., "Association of Peripheral Blood Cell Profile with Alzheimer's Disease: A Meta-Analysis," *Frontiers in Aging Neuroscience*, vol. 14, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[61] Usma Munawara et al., "Hyperactivation of Monocytes and Macrophages in MCI Patients Contributes to the Progression of Alzheimer's Disease," *Immunity & Ageing*, vol. 18, pp. 1-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[62] Showkat Ul Nabi et al., "Mechanisms of Mitochondrial Malfunction in Alzheimer's Disease: New Therapeutic Hope," *Oxidative Medicine and Cellular Longevity*, vol. 2022, no. 1, pp. 1-28, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[63] Yume Imahori et al., "Association of Resting Heart Rate with Cognitive Decline and Dementia in Older Adults: A Population-Based Cohort Study," *Alzheimer's & Dementia*, vol. 18, no. 10, pp. 1779-1787, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[64] Agata Tarkowska et al., "Preservation of Biomarkers Associated with Alzheimer's Disease (Amyloid Peptides 1-38, 1-40, 1-42, Tau Protein, Beclin 1) in the Blood of Neonates after Perinatal Asphyxia," *International Journal of Molecular Sciences*, vol. 24, no. 17, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[65] Cian Carey et al., "Hypertensive Disorders of Pregnancy and the Risk of Maternal Dementia: A Systematic Review and Meta-Analysis," *American Journal of Obstetrics and Gynecology*, vol. 231, no. 2, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[66] Kathleen B. Miller et al., "Ionizing Radiation, Cerebrovascular Disease, and Consequent Dementia: A Review and Proposed Framework Relevant to Space Radiation Exposure," *Frontiers in Physiology*, vol. 13, pp. 1-26, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[67] María P. Aranda et al., "The Relationship of History of Psychiatric and Substance Use Disorders on Risk of Dementia among Racial and Ethnic Groups in the United States," *Frontiers in Psychiatry*, vol. 14, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[68] Magdalena Pszczołowska et al., "Association between Female Reproductive Factors and Risk of Dementia," *Journal of Clinical Medicine*, vol. 13, no. 10, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[69] Laia Montoliu-Gaya et al., "Blood Biomarkers for Alzheimer's Disease in Down Syndrome," *Journal of Clinical Medicine*, vol. 10, no. 16, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[70] Theodoros Karampitsakos et al., "Increased Monocyte Count and Red Cell Distribution Width as Prognostic Biomarkers in Patients with Idiopathic Pulmonary Fibrosis," *Respiratory Research*, vol. 22, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[71] Monica Averna, Paola Melotti, and Claudio Sorio, "Revisiting the Role of Leukocytes in Cystic Fibrosis," *Cells*, vol. 10, no. 12, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[72] Gabriela Oates et al., "The Association of Area Deprivation and State Child Health with Respiratory Outcomes of Pediatric Patients with Cystic Fibrosis in the United States," *Pediatric Pulmonology*, vol. 56, no. 5, pp. 883-890, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[73] T. Spencer Poore, Jennifer L. Taylor-Cousar, and Edith T. Zemanick, "Cardiovascular Complications in Cystic Fibrosis: A Review of the Literature," *Journal of Cystic Fibrosis*, vol. 21, no. 1, pp. 18-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[74] Sanaz Vaziri et al., "Time to be Blunt: Substance Use in Cystic Fibrosis," *Pediatric Pulmonology*, vol. 59, no. 4, pp. 1015-1027, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[75] Mabel Aoun, Michel Jadoul, and Hans-Joachim Anders, "Erythrocytosis and CKD: A Review," *American Journal of Kidney Diseases*, vol. 84, no. 4, pp. 495-506, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[76] Roshini Kurian, Preethu Anand, and George Ghaly, "A Late and Complex Presentation of Hereditary Haemochromatosis," *Cureus*, vol. 14, no. 11, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[77] Jinling Wang et al., "Case Report: A Rare Case of Hereditary Hemochromatosis Caused by a Mutation in the HAMP Gene in Fuyang, China," *Frontiers in Medicine*, vol. 11, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[78] Michał Świątczak et al., "The Potential Impact of Hereditary Hemochromatosis on the Heart Considering the Disease Stage and Patient Age—The Role of Echocardiography," *Frontiers in Cardiovascular Medicine*, vol. 10, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[79] Danielle S. Kroll et al., "Elevated Transferrin Saturation in Individuals with Alcohol Use Disorder: Association with HFE Polymorphism and Alcohol Withdrawal Severity," *Addiction Biology*, vol. 27, no. 2, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[80] Ambrin Gull Shamas, "Primary Hereditary Haemochromatosis and Pregnancy," *GastroHep*, vol. 2023, no. 1, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[81] Corneliu Toader et al., "From Recognition to Remedy: The Significance of Biomarkers in Neurodegenerative Disease Pathology," *International Journal of Molecular Sciences*, vol. 24, no. 22, pp. 1-32, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[82] Thomas Kluyver et al., *Jupyter Notebooks—A Publishing Format for Reproducible Computational Workflows*, *Positioning and Power in Academic Publishing: Players, Agents, and Agendas*, IOS Press, pp. 87-90, 2016. [CrossRef] [Google Scholar] [Publisher Link]
[83] *Visual Studio Code Documentation*, Microsoft, 2023. [Online]. Available: https://code.visualstudio.com/docs
[84] *Python 3.11.0 Release Notes*, Python Software Foundation, 2024. [Online]. Available: https://www.python.org/downloads/release/python-3110/
[85] Wes McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, Austin, TX, pp. 56-61, 2010. [CrossRef] [Google Scholar] [Publisher Link]
[86] Charles R. Harris et al., "Array Programming with NumPy," *Nature*, vol. 585, pp. 357-362, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[87] Fabian Pedregosa et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.* [Google Scholar] [Publisher Link]
[88] John D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, pp. 90-95, 2007. [CrossRef] [Google Scholar] [Publisher Link]
[89] Michael L. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, vol. 6, no. 60, pp. 1-4, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[90] Tianqi Chen, and Carlos Ernesto Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]
[91] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas, "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1-5, 2017. [Google Scholar] [Publisher Link]
[92] Cencheng Shen, Sambit Panda, and Joshua T. Vogelstein, "The Chi-Square Test of Distance Correlation," *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, pp. 254-262, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[93] P. Yu-Wai-Man et al., "Inherited Mitochondrial Optic Neuropathies," *Journal of Medical Genetics*, vol. 46, pp. 148-158, 2009. [CrossRef] [Google Scholar] [Publisher Link]

Appendix

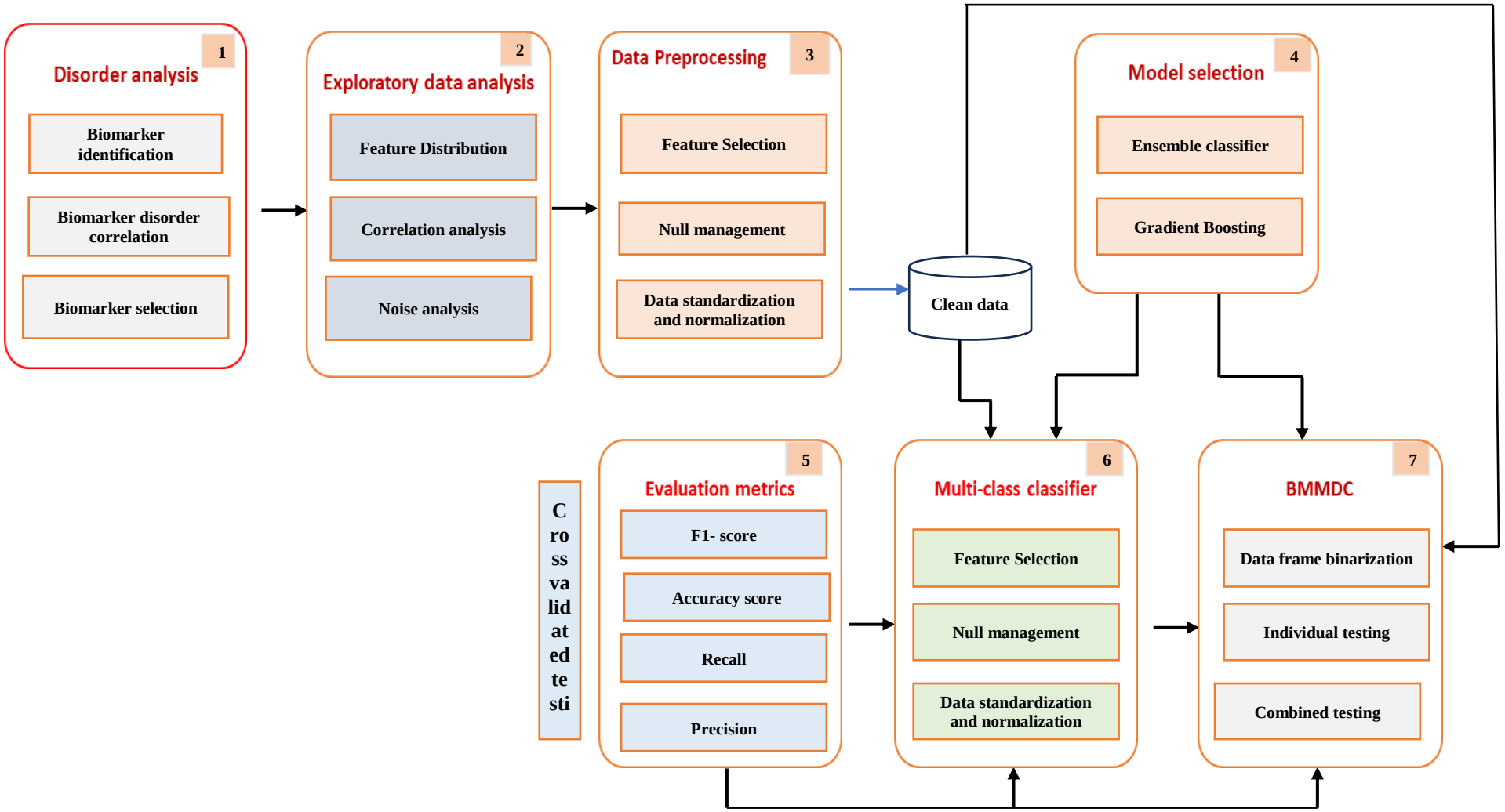

Fig. 2 The workflow of this study

Table 2. Biomarker disorder correlation

Parameter / Disorder	LHON	Leigh Syndrome	Mitochondrial myopathy	Diabetes	Cancer	Alzheimer's	Cystic fibrosis	Hemochromatosis	Tay-Sachs
Red Blood Cell (RBC) Count	Normal; possible link to MCV [93]	Normal; sometimes elevated	Reduced count, anaemia and oxidative stress [24]	Altered morphology & aggregation [35]	Reduced (anaemia, marrow infiltration) [48]	Altered RBC characteristics [60]	Normal to increased (chronic hypoxia) [70]	Normal; sometimes erythrocytosis or microcytosis [75][76]	Typically, normal
White Blood Cell (WBC) Count	Normal [14]	Normal; sometimes elevated	Leukopenia in Barth syndrome [25]	Elevated (inflammation, insulin resistance) [36]	Elevated (leukaemia, lymphoma, paraneoplastic response) [49]	Altered profiles (monocyte, lymphocyte shifts) [61]	Elevated (neutrophilia, chronic inflammation) [71]	Normal [75]	Normal
Respiratory Rate	Mitochondrial dysfunction affects ATP synthesis and ROS [15]	Abnormal breathing patterns (Cheyne-Stokes, hyperventilation) [20]	Exertional dyspnea, respiratory failure [26]	Altered in diabetic ketoacidosis (DKA) [37]	Elevated (infection, tumour impact on lungs) [50]	Linked to OSA, pneumonia, and muscle weakness. [62]	Increased (severity marker, exacerbations) [72]	Normal initially, later abnormal due to complications.[77]	Normal initially, abnormal in progression
Heart Rate	Some arrhythmias reported [16]	Arrhythmias, conduction defects, cardiomyopathy [20]	Bradycardia, tachyarrhythmias, and conduction issues [27]	Increased resting HR, reduced HRV (autonomic dysfunction) [38]	Increased (cancer-associated autonomic dysfunction, cardiotoxicity) [50]	Elevated resting HR, HRV changes (ANS dysfunction) [63]	Exercise intolerance increased HR on exertion [73]	Normal initially, later irregular (iron-induced arrhythmias) [78]	Normal initially, irregular later
Birth Asphyxia	No link	No direct link	Sometimes observed [28]	Increased risk in infants of diabetic mothers [39].	Increased leukaemia/solid tumour risk [51]	Perinatal oxygen deprivation linked to AD [64]	Possible false-positive CF screening (elevated IRT) [73]	No association	No association

Folic Acid Supplementation	Moderate to high correlation (mitochondrial metabolism) [17]	Limited data but slight correlation [21]	Strong correlation (mitochondrial function) [29]	Improves glycemic control.[40]	Incorrect intake may increase cancer risk [52]	No direct link	No association	No direct evidence	No direct evidence
Serious Maternal Illness	No link	No link	Significant correlation (metabolic/autoimmune disorders) [30]	Hypertensive disorders, severe maternal morbidity. [41]	Correlated with increased childhood cancer risk. [53]	Maternal metabolic cardiovascular illness linked to AD risk [65]	No confirmed relationship	No direct evidence	No direct evidence
Radiation Exposure	High correlation (mtDNA damage, prenatal risk) [18]	No link	No link	Pancreatic radiation exposure linked to diabetes risk.[42]	Prenatal exposure increases childhood cancer risk [54]	Increased AD risk (DNA damage, oxidative stress) [66]	No confirmed relationship	No direct evidence	No direct evidence
Substance Abuse	High correlation (alcohol, tobacco, mitochondrial stress)	There is no direct link, but it may have degenerative effects on the patient.	Mitochondrial toxicity (aminoglycosides, valproic acid, statins) [31]	Impairs pancreatic development increases diabetes risk. [43]	Alters epigenetics increases fetal cancer risk [55]	Linked to cognitive decline, AD risk [67]	Worsens respiratory issues [74]	Exacerbates liver damage [79]	No causal connection
Assisted Conception (IVF/ART)	No link	No link	May reduce disorder transmission via mitochondrial replacement therapies [32]	Possible diabetes risk due to epigenetic imprinting. [44]	Mixed evidence regarding childhood cancer risk [56]	No direct link	No confirmed relationship	No direct evidence	No direct evidence
History of Anomalies in Previous Pregnancies	No link	May suggest mitochondrial inheritance	Increased recurrence risk (mitochondrial inheritance) [33]	Indicative of maternal diabetes/GDM [45]	May reflect genetic/environmental cancer risk [57]	May influence maternal AD risk later	No confirmed relationship	Elevated recurrence risk (maternal iron dysregulation) [80]	No direct evidence

Number of Previous Abortions	No link	No link	Increased risk with more abortions [33]	Increased risk of spontaneous abortion [46]	Associated with increased cancer risk [58]	This may reflect immune dysregulation/genetic susceptibility. [68]	No confirmed relationship	No direct evidence	No direct evidence
Birth Defects	No link	Facial dysmorphisms, congenital lactic acidosis, hypotonia [22]	Pre-eclampsia, IUGR, polyhydramnios/ oligohydramnios, and preterm labour [33]	More frequent in infants of diabetic mothers. [47]	Severe defects in some cases [59]	Down syndrome linked to early-onset AD [69]	No direct connection	No direct evidence	No direct evidence
Blood Test Results	Regular metabolic panels may reveal mitochondrial dysfunction [19]	Lactic acidosis, elevated pyruvate & alanine [23]	Elevated lactate, variable CK, amino acid, and organic acid abnormalities [34]	Elevated fasting glucose, HbA1c, and insulin resistance markers. [35][36]	Anemia, WBC abnormalities, elevated LDH, tumor markers [49]	Peripheral metabolic changes in neurodegeneration [60] [61]	Elevated IRT levels (CF screening) [70] [71]	High ferritin, abnormal liver enzymes, coagulopathy [75][76]	Minimal peripheral biomarker changes, CNS degeneration evident [81]