Visvesvaraya Technological University, Belagavi – 590018

SEMINAR REPORT (18CSS84)
ON

Genomic Intelligence: The Impact of AI in Uncovering Genetic Disorders

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE & ENGINEERING

Submitted by

Joyline Rencita Dsouza **4SO20CS073**

Under the Guidance of

Dr Saumya Y M
Associate Professor, Department of CSE

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
ST JOSEPH ENGINEERING COLLEGE
An Autonomous Institution
(Affiliated to VTU Belagavi, Recognized by AICTE, Accredited by NBA)
Vamanjoor, Mangaluru - 575028, Karnataka

2023-24

Abstract

Genetic disorders, arising from DNA mutations or changes in chromosomes, present significant health challenges. As the population expands, there's a notable surge in these disorders. Several types of commonly known diseases are related to hereditary gene mutations. Genetic testing aids patients in making important decisions in the prevention, treatment, or early detection of hereditary disorders. In the last ten years, there's been progress in customizing medical treatments based on people's unique genetics, called precision genomics-based medicine. The idea is to provide personalized and effective healthcare. However, accurately predicting illness risks, especially with methods like polygenic risk scores, is still tricky. To tackle this, researchers are looking into machine learning algorithms, which are good at predicting risks of complicated diseases. These algorithms are better because they can handle a variety of information.

The research methodology involves training the Machine Learning Model using a medical dataset. The model's predictive capabilities, experimental work, and subsequent results and discussions form the study's core. In this systematic review, we aimed to identify the methodological trends and the ML application areas in rare genetic diseases. The scope for future work lies in refining and expanding the model's capabilities to encompass a broader range of genetic disorders. This scoping review aims to address this gap and explores the use of machine learning in investigating rare diseases.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

A genetic disorder occurs when a gene is altered. This disruption can be hereditary or spontaneous, and some genetic disorders cannot be prevented [1]. Because they cannot be prevented, they can only be discovered when symptoms arise. This makes them extremely important to understand. Let's look at some common examples.

Article [5] categorizes genetic disorders into four main classes: Single-Gene Disorders (SGMDs), Chromosomal Disorders (CD), Complex Genetic Diseases, and Mitochondrial Disorders. Single-Gene Disorders (SGMDs) result from mutations in a specific gene, such as sickle cell anemia, impacting a singular gene's function. Chromosomal Disorders involve abnormalities in chromosome structure or number, leading to missing or duplicated chromosomes. Complex Genetic Diseases arise from malfunctions in multiple genes, influenced by both genetic inheritance and environmental factors. Mitochondrial Disorders, affecting mitochondrial DNA, disrupt energy production and contribute to diverse health issues [1].

Understanding these classes is crucial for comprehensive insights into genetic disorders and their management.In this we examine the current applications of machine learning in the context of rare diseases. By offering insights into the intersection of machine learning and rare diseases, our review aims to guide future research directions, identifying research gaps and potential areas for further investigation [3].

1.2 Traditional Approaches

Traditional approaches for detecting genetic disorders rely on well-established techniques such as karyotyping, fluorescence in situ hybridization (FISH), and polymerase chain reaction (PCR). These methods have been pivotal in diagnosing chromosomal abnormalities,

known genetic mutations, and structural rearrangements for decades. However, they may be labor-intensive, time-consuming, and limited in their ability to detect subtle variations or novel genetic factors. Despite these limitations, traditional approaches remain integral to clinical diagnostics, providing valuable insights into genetic diseases alongside newer genomic technologies.

1.2.1 Fluorescence In Situ Hybridization (FISH)

Fluorescence In Situ Hybridization (FISH) is a sophisticated molecular cytogenetic technique crucial for identifying specific chromosomal abnormalities and genetic disorders. It begins with the design of fluorescently labeled DNA probes tailored to target precise DNA sequences within the chromosomal region of interest. These probes are then introduced to cellular samples, such as blood or tissue samples, where they selectively bind to complementary DNA sequences within the chromosomes. Utilizing fluorescence microscopy, the labeled probes emit distinct fluorescent signals, which are then meticulously analyzed by cytogeneticists or molecular geneticists. Through this analysis, chromosomal abnormalities associated with genetic disorders can be accurately identified [4].
For example, in the case of Down syndrome, FISH can detect the presence of an extra copy of chromosome 21 (trisomy 21) by visualizing three fluorescent signals corresponding to chromosome 21 instead of the usual two signals. This method offers high specificity and sensitivity in diagnosing genetic disorders, aiding in accurate diagnosis and genetic counseling for affected individuals.

1.2.2 Polymerase Chain Reaction (PCR)

Amplifying specific DNA sequences. Initially, PCR separates double-stranded DNA into single strands. Then, primers bind to the target DNA, allowing DNA polymerase to generate complementary strands. Through numerous cycles of denaturation, annealing, and extension, PCR exponentially increases the amount of target DNA, aiding in detecting genetic anomalies. Analysis of the amplified DNA, often through gel electrophoresis, reveals the presence or absence of DNA fragments associated with genetic disorders. This reliable technique is crucial for diagnosing genetic conditions and guiding patient care.
One example of a genetic disorder that can be detected using Polymerase Chain Reaction (PCR) is Cystic Fibrosis (CF). PCR can be utilized to amplify specific regions of the CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) gene, which is associated with CF [4].

1.2.3 Pedigree Analysis

Pedigree analysis is a fundamental tool in genetics that aids in understanding the inheritance patterns of genetic traits and disorders within families. It involves the construction

of a family tree or pedigree chart, detailing the relationships and phenotypic characteristics of individuals across generations. By examining patterns of inheritance, such as autosomal dominant, autosomal recessive, X-linked, or mitochondrial inheritance, geneticists can infer the likelihood of an individual inheriting or passing on a particular genetic disorder. Pedigree analysis plays a crucial role in diagnosing genetic diseases, predicting recurrence risks, and guiding genetic counseling and family planning decisions.

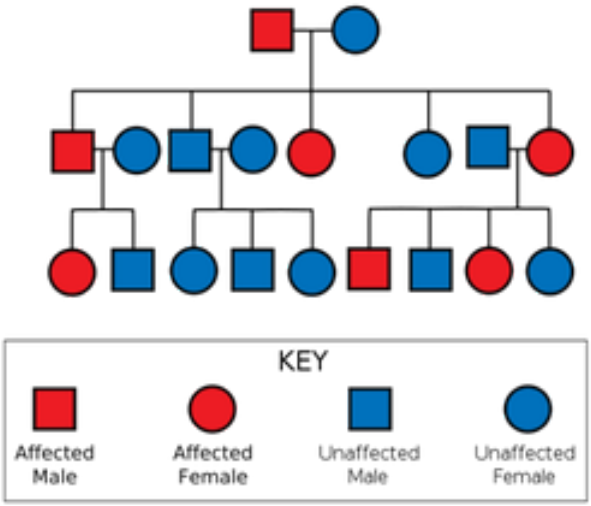

Figure 1.1: Pedigree Analysis.

In Figure 1.1, the inheritance pattern of Down syndrome, it can arise spontaneously from cell division errors or be passed down from parents with chromosomal changes. Pedigree charts visually represent how the condition is inherited within families, aiding in genetic counseling sessions. Understanding these inheritance patterns is vital for providing support to affected families and assessing the likelihood of recurrence in future generations.

1.2.4 Cytogenetic Analysis

Cytogenetic analysis is a pivotal technique in genetics, focusing on the study of chromosomes to detect structural and numerical abnormalities associated with genetic disorders. It involves the visualization and analysis of chromosomes under a microscope to identify anomalies such as deletions, duplications, inversions, or translocations. This analysis provides crucial insights into the genetic basis of diseases, guiding diagnosis, prognosis, and treatment decisions [4]. By examining a patient's karyotype, cytogeneticists can pinpoint chromosomal abnormalities, such as trisomies or chromosomal rearrangements.
One example of a disease diagnosed through cytogenetic analysis is Turner syndrome. In this condition, individuals typically have only one X chromosome (monosomy X) instead of the usual two. Cytogenetic analysis can reveal this chromosomal abnormality through examination of the patient's karyotype, confirming the diagnosis of Turner syndrome.

Chapter 2

Methedology

In the methodology section,we conduct a thorough comparison of various approaches used in the field. By analyzing factors such as accuracy, efficiency, and robustness, we identify the most promising techniques. Through systematic experimentation and validation, we strive to develop a robust approach that surpasses previous methodologies.

The initial step involves the collection of genomic data, including information from individuals with known genetic disorders and those without. Rigorous preprocessing is then applied, incorporating data cleaning, normalization, and addressing missing values to establish a reliable and standardized dataset. Subsequently, relevant features representing genomic variations associated with genetic disorders are carefully selected. Feature engineering techniques, such as dimensionality reduction methods, manage the complexity of the dataset while preserving essential information.

In this section, we investigate various methods utilized for predicting genetic disease outcomes [5]. Through our analysis, we have identified the most effective approach to date. Table 2.1 presents a comparative analysis of various proposed approaches discussed

Table 2.1: Comparative analysis of different proposed approaches.

Year	Technique	Training Time	Macro Accuracy	Hamming Loss	Evaluation Score
2020	SVM	7.10	73	0.22	88
2020	KNN	0.01	70	0.25	86
2020	KNN	0.01	70	0.25	86
2020	RF	2.48	82	0.14	90
2021	KNN	0.01	70	0.25	86
2022	ETRF + XGB	3.59	84	0.12	92

in [1]. The data showcases the training time, macro accuracy, Hamming loss, and evaluation score for each technique employed across different years. Notably, the ETRF + XGB method stands out with its commendable accuracy, achieving a macro accuracy of 84%, a Hamming loss of 0.12, and an evaluation score of 92%.

Chapter 3

Genetic Disease Detection

In this study, we propose a comprehensive approach for predicting genetic disorders using a multi-label multi-class classifier chain approach and a novel feature engineering technique called ETRF (Extreme Trees Random Forest). Our model aims to leverage the rich information present in genome and genetic datasets to accurately classify various genetic disorders and their subclasses. Through a detailed workflow, we demonstrate how our approach effectively handles the complexities of multi-label multi-class classification in genetic data analysis.

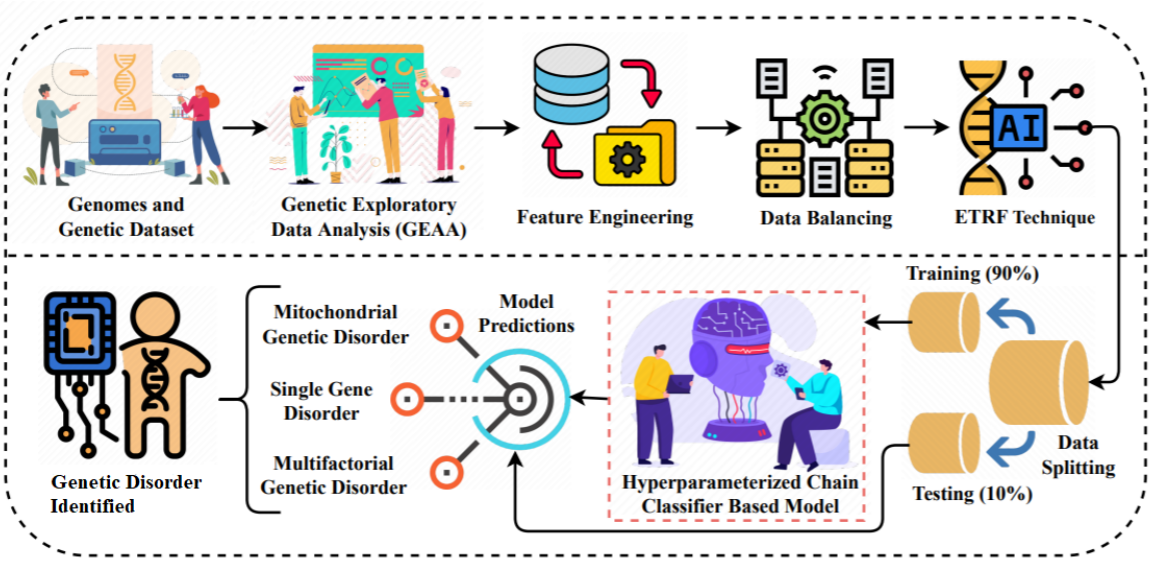

Figure 3.1: Illustration of the proposed approach

Figure 3.1 represents our proposed approach [5] for predicting genetic disorders and types of disorders involves a comprehensive methodological analysis aimed at leveraging machine learning techniques to analyze genomic data. Here's an elaboration of the proposed methodology:

3.1 Genomes and Genetic Dataset

The genomes dataset comprises medical information from individuals with genetic disorders, including children and adults. It's a multi-label multi-class dataset, this dataset contains a total of 44 attributes. Each attribute provides valuable information for analyzing and understanding genetic disorders and their subclasses.

Table 3.1: The genomes dataset features descriptive analysis

SlNo.	Feature	Count	SlNo.	Feature	Count
1	Patient Id	31,548	23	Follow-up	29,382
2	Patient Age	30,121	24	Gender	29,375
3	Mother's side genes	31,548	25	Birth asphyxia	29,409
4	Inherited from father	30,691	26	Autopsy shows birth defect	30,522
5	Maternal gene	25,015	27	Place of birth	29,424
6	Paternal gene	31,548	28	Folic acid details	29,431
7	Blood cell count	31,548	29	H/O serious maternal illness	29,396
8	Patient First Name	31,548	30	H/O radiation exposure	29,395
9	Family Name	12,540	31	H/O substance abuse	29,353
10	Father's name	31,548	32	Assisted conception IVF	29,426
11	Mother's age	25,512	33	History of anomalies	29,376
12	Father's age	25,562	34	No. of previous abortion	29,386
13	Institute Name	24,406	35	Birth defects	29,394
14	Location of Institute	31,548	36	White Blood cell count	29,400
15	Status	31,548	37	Blood test result	29,403
16	Respiratory Rate	26,513	38	Symptom 1	29,393
17	Heart Rate	26,535	39	Symptom 2	29,326
18	Test 1	29,421	40	Symptom 3	29,447
19	Test 2	29,396	41	Symptom 4	29,435
20	Test 3	29,401	42	Symptom 5	29,395
21	Test 4	29,408	43	Genetic Disorder	19,937
22	Test 5	29,378	44	Disorder Subclass	19,915

Table 3.1 serves as a comprehensive presentation of the descriptive analysis of the genomes dataset.The 'genetic disorder' attribute serves as the primary classification, while 'disorder subclass' further refines sub-classes, enabling precise categorization of individuals into specific genetic disorders

3.2 Genetic Exploratory Data Analysis (GEDA)

We begin by analyzing the genomes dataset, which contains medical information of patients with genetic disorders.
The dataset consists of 44 attributes including patient demographics, genetic information, and medical history. We conduct exploratory data analysis (EDA) to uncover hidden patterns and understand the distribution of genetic disorders and subclasses.

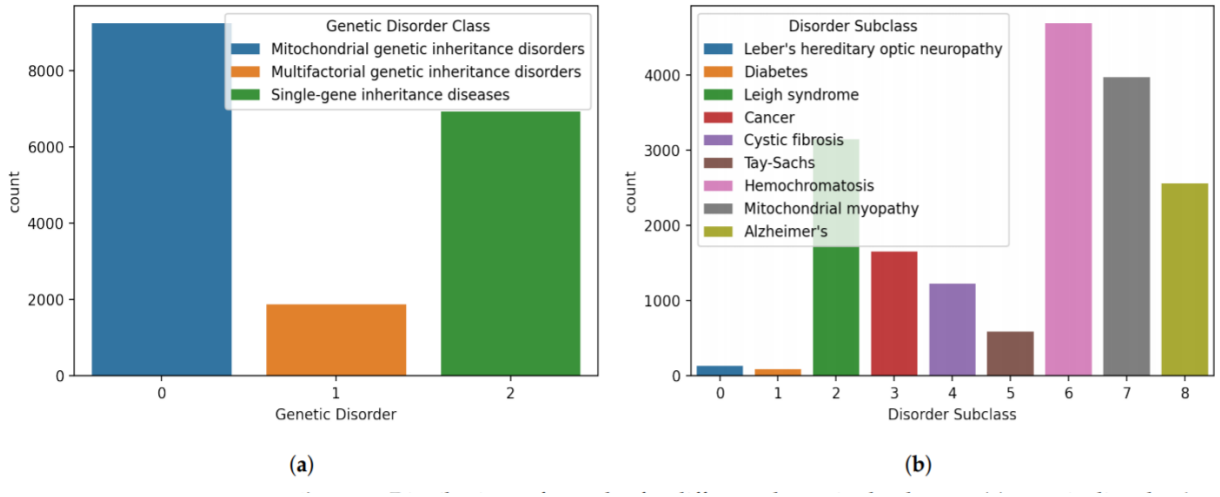

Figure 3.2: Distributions of samples for different classes in the dataset, (a) genetic disorders' main classes, and (b) genetic disorder subclasses.

The Figure 3.2 (a) illustrates distributions of genetic disorders' main classes, while the Figure 3.2 (b) depicts distributions of genetic disorder subclasses in the dataset.
The Next step is Gene analysis for disorder main class, it evaluates the correlation between inherited genes and the likelihood of specific genetic disorders, including maternal, paternal, maternal side, and paternal inheritance which is shown in Figure 3.3 below .

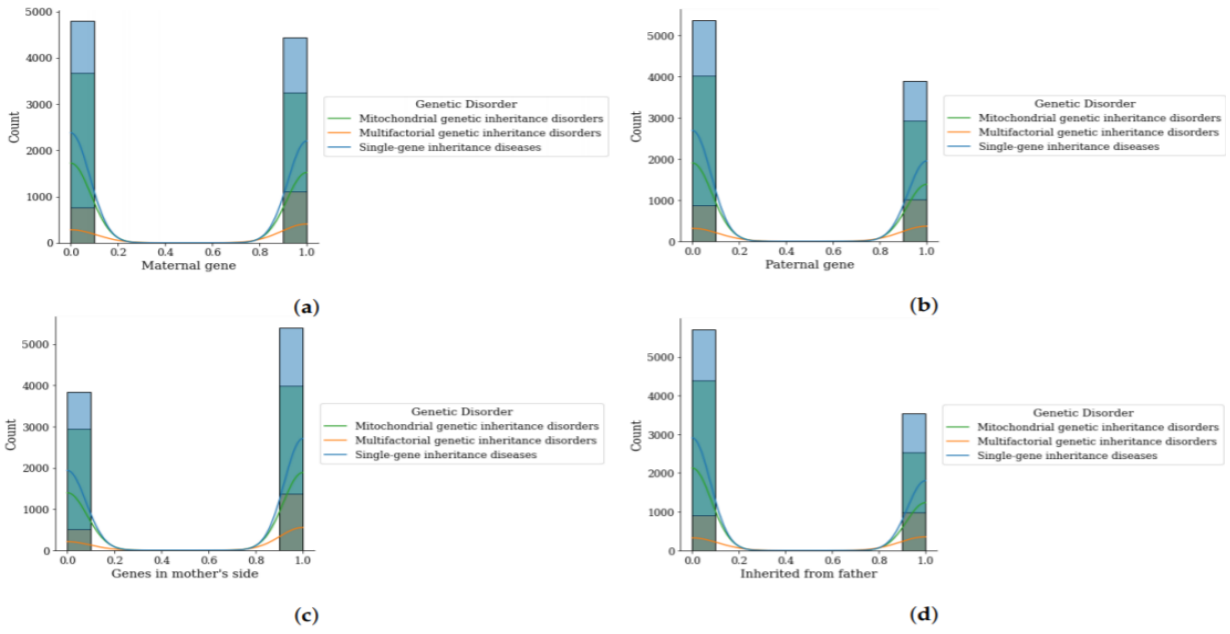

Figure 3.3: Data distribution by genetic disorder category, (a) maternal gene, (b) paternal gene, (c) genes from mother side, and (d) inherited from father.

This analysis highlights genes as significant factors influencing the occurrence of genetic disorders. Specifically, it indicates that when the values of maternal and paternal genes

are 0 or 1, there is a higher probability of mitochondrial disorders, whereas single-gene disorders have a lower likelihood. Furthermore, Figure 3.3(c) and 3.3 (d) demonstrate that mitochondrial disorders are more likely when genes inherited from the mother's side and father are at values of 0 or 1.
Next step we proceed for the Gene analysis for disorder sub-class

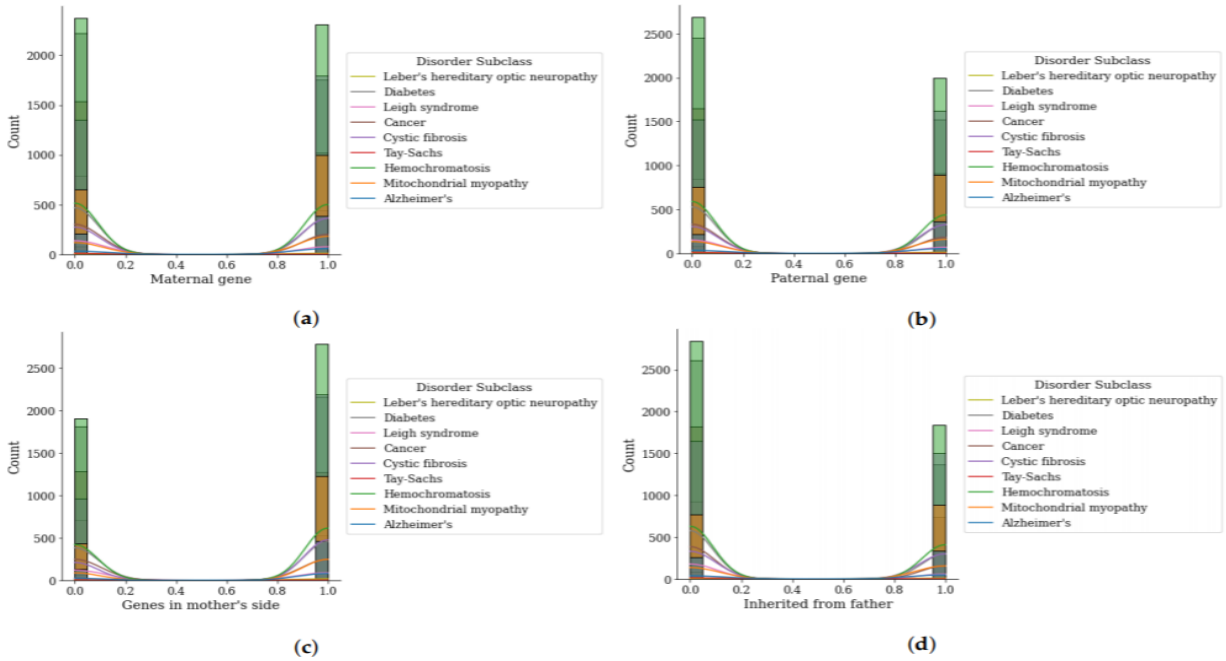

Figure 3.4: Data distribution by genetic disorder sub-category, (a) maternal gene, (b) paternal gene, (c) genes from mother side, and (d) inherited from father.

The analysis presented in Figure 3.4 illustrates that there is a correlation between the genetic makeup, particularly the maternal and paternal genes, and the occurrence of specific disorders within the subclass category.For instance, the analysis reveals that diabetes disorder occurs more frequently when both maternal and paternal genes have values of 0 or 1, whereas the likelihood of Leigh syndrome is lower across all genes within the dataset.
The age analysis is conducted to understand the correlation between age and the occurrence of genetic disorders, considering the age of both parents and patients.

Figure 3.5 illustrates the relationship between age and the occurrence of genetic disorders. For example, the analysis indicates that there is a higher likelihood of genetic disorders when the mother's age falls between 20 and 60 years, while a lower probability is observed when the mother is younger than 20 years. Similarly, the father's age between 20 and 70 years is associated with an increased chance of genetic disorders. An instance of this correlation is that individuals within a certain age range, such as those aged between 15 to 30 years, exhibit a higher incidence of genetic disorders.

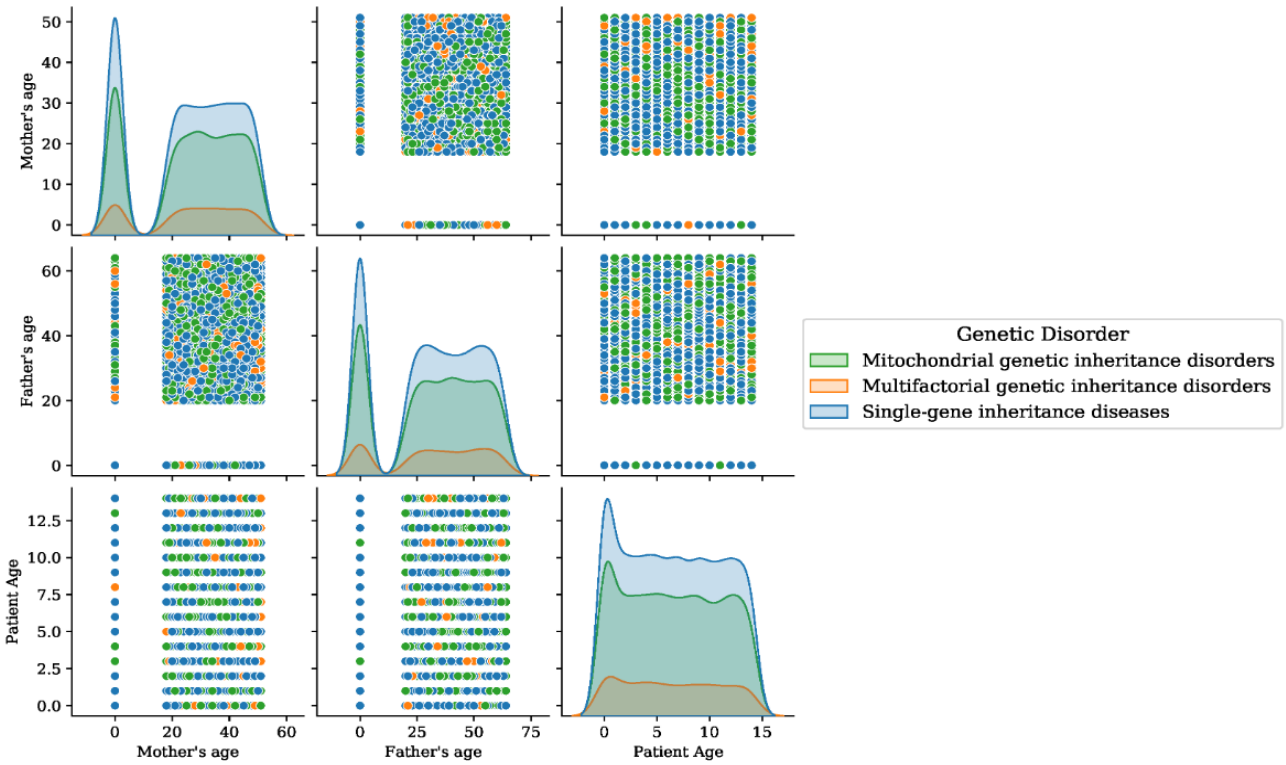

Figure 3.5: Age analysis of patients for the disorder category.

3.3 Data Normalization and Feature Engineering

Data normalization is the important step in preparing the dataset for machine learning models. It involves scaling numerical features to a standard range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. Normalization ensures that features with different scales contribute equally to the model training process and prevents any single feature from dominating the learning algorithm due to its larger magnitude [5]. Feature engineering plays a pivotal role in preparing the genomes dataset for machine learning analysis. Through careful selection and elimination of features based on their relevance, the dataset is refined to include only the most informative attributes for predicting genetic disorders. Decision tree models aid in this process by identifying feature importance, guiding the removal of less significant variables. Null values are addressed by filling them with zeros, ensuring data completeness and integrity. Categorical features undergo encoding to transform them into numerical representations, facilitating the application of machine learning algorithms. This meticulous preprocessing ensures that the dataset is well-structured and optimized for subsequent modeling tasks. By enhancing the quality and relevance of the dataset, feature engineering contributes to the overall effectiveness and accuracy of predictive models in genetic disorder analysis.

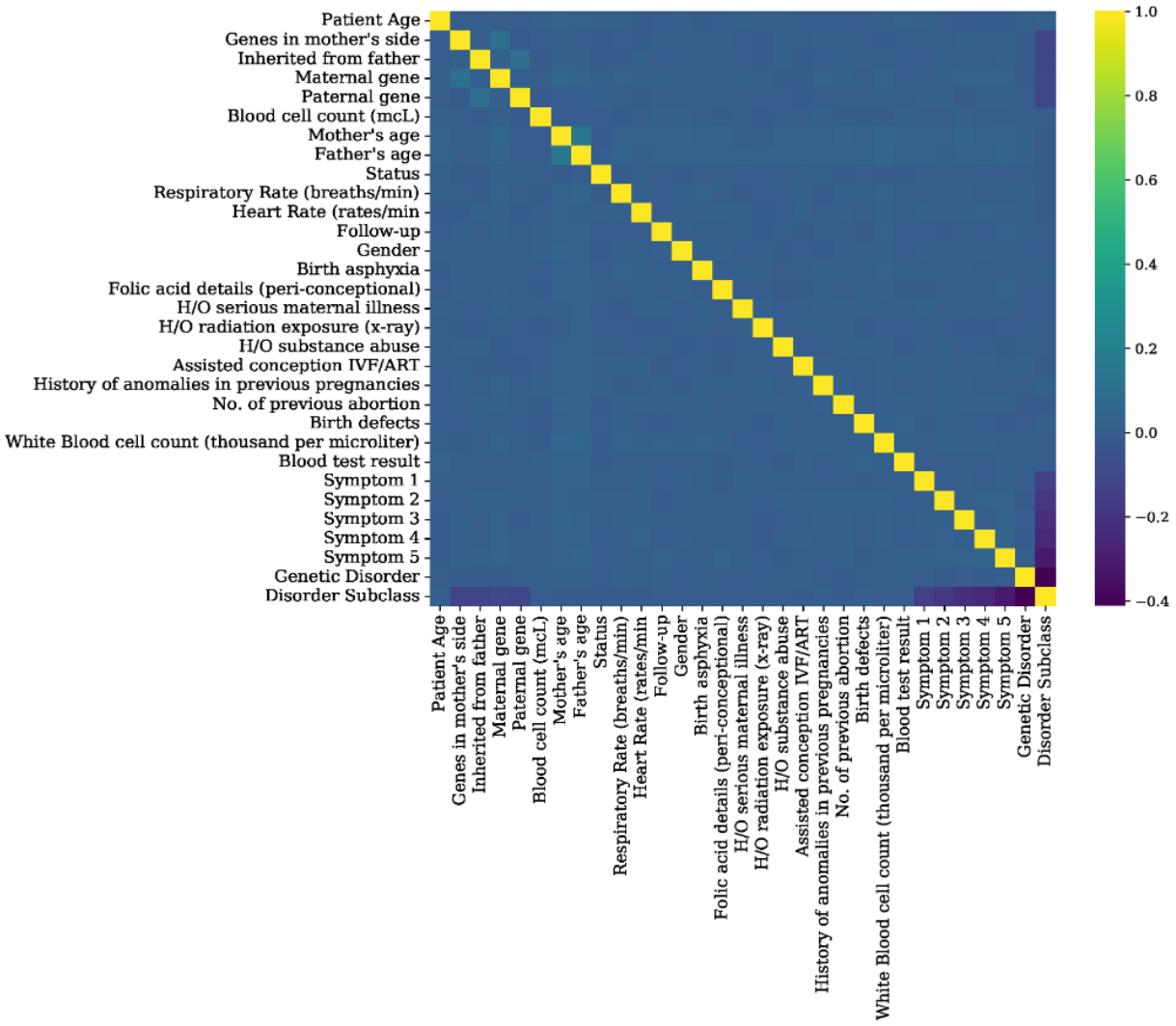

Figure 3.6: Feature correlation analysis graphs of genomes data.

Figure 3.6 illustrates the feature analysis correlation graph of the genome data, providing insights into the relationships between different attributes.

The selected features are encoded with appropriate categorical data values to prepare the dataset for analysis. Binary features like 'genes in mother's side', 'inherited from father', 'maternal gene', and others, are mapped to 1 for 'Yes' and 0 for 'No'. Features such as 'H/O radiation exposure (X-ray)' and 'H/O substance abuse' are mapped to 1, 0, and -1, corresponding to 'Yes', 'No', and 'Not applicable', respectively. 'Status' is mapped to 0 for 'deceased' and 1 for 'alive'. Categorical features like 'respiratory rate (breaths/min)' and 'heart rate (rates/min)' are replaced with 0 for 'normal' and 1 for 'Tachypnea'. 'Follow-up' is mapped to 0 for 'Low' and 1 for 'High'. 'Gender' is encoded as 0 for 'male', 1 for 'female', and 2 for 'ambiguous'. 'Birth asphyxia' values are replaced with 0, 0, 0, and 1 for 'No record', 'Not available', 'No', and 'Yes', respectively. Similarly, 'birth defects' is mapped to 0 for 'singular' and 1 for 'multiple', while 'blood test result' is replaced with 0 for 'normal' and 1 for 'abnormal'.

3.4 Data Balancing

To enhance the accuracy of applied learning techniques, dataset balancing is implemented. This approach ensures that learning models are trained on an equal number of data samples, promoting efficient results [5]. Prior to balancing, the dataset comprises 10,202, 2071, and 7664 data samples for mitochondrial genetic inheritance disorders, multi-factorial genetic inheritance disorders, and single-gene inheritance classes, respectively. Balancing involves randomly dropping data samples from other classes to match the count of the lowest class.

3.5 Data Splitting

Data splitting is employed to divide the dataset into training and test sets, mitigating learning model over fitting and promoting generalization. In our experiments, various split ratios (0.7:0.3, 0.8:0.2, 0.85:0.15, and 0.9:0.1) are applied for the genomes dataset [5]. These ratios facilitate cross-validation, allowing the assessment of machine learning techniques' performance and identifying the optimal split for achieving the most effective learning model.

3.6 Applied Learning Techniques

Several machine learning models are applied to analyze the performance of the proposed feature engineering approach. Eight well-known machine learning models, which are reported to show good performance for tasks similar to genetic disorder prediction, are utilized [5]. The main focused algorithms are listed below:

- **Decision Tree Classifier (DTC)**: DTC is a supervised learning algorithm used for classification tasks. It constructs a tree-like structure where inner nodes represent data attributes and leaf nodes contain outcome labels. The algorithm aims to minimize generalization error by selecting optimal decision trees based on measures like information gain and Gini index.

- **Random Forest Classifier (RFC):** RFC is an ensemble learning technique based on multiple decision trees. It aggregates predictions from these trees using majority voting, reducing overfitting and improving classification performance compared to individual classifiers.

- **Extra Trees Classifier (ETC):** ETC is another ensemble-based, bagged decision tree technique similar to RFC. It reduces model variance by employing a random split selection of values and a meta estimator that fits randomized decision trees on sample datasets, leading to improved accuracy and reduced overfitting.

- **Logistic Regression (LR):** LR is a statistical learning method for classifications, often used for multilabel classification tasks. It predicts dependent categorical variables using independent variables and maps predicted outputs to probabilities using the Sigmoid function.

- **Multi-layer Perceptron (MLP):** MLP is a classification algorithm based on feedforward neural networks with multiple fully connected layers. It uses stochastic gradient descent to optimize the loss function and has shown superior performance for various tasks despite its simplicity compared to more complex models.

- **K-Nearest Neighbors (KNN):** KNN is a non-parametric supervised learning technique that predicts the class of data based on the similarity of nearest neighbors. It groups data points based on their proximity, with classification performed using distance metrics such as Euclidean distance.

- **XGBoost (XGB):** XGB is a flexible and efficient classification algorithm based on boosting techniques. It employs parallel gradient boosting tree technique to solve classification problems and uses regularization to reduce overfitting, leading to improved predictive performance.

- **Support Vector Classifier (SVC):** SVC is a supervised learning algorithm primarily used for classification tasks. It finds the optimal hyperplane that separates input data points into different classes by maximizing the margin between them, with predictions made using a hypothesis function based on support vectors.

3.7 Multi-Label Multi-Class Chain Classifier Approach

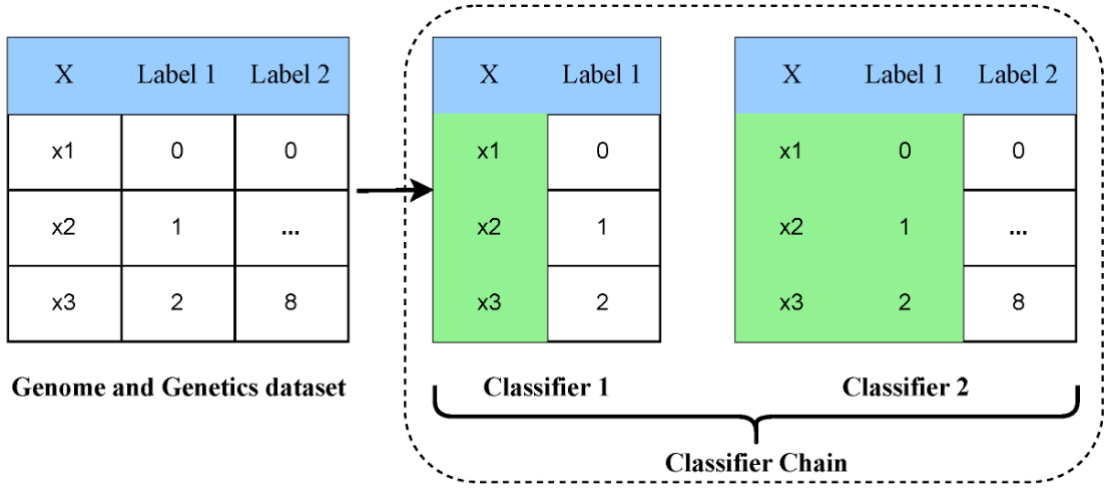

Figure 3.7: The architectural analysis of the multi-label multi-class classifier chain approach.

Figure 3.7 illustrates the architectural analysis of the multi-label multi-class classifier chain approach. The Multi-Label Multi-Class Chain Classifier Approach utilizes a connected chain of multiple classifiers to address the multi-label multi-class classification task of genetic disorders and their subclasses. In this approach, each classifier in the chain predicts the presence or absence of a specific label (i.e., a genetic disorder or subclass) based on the input data and the predictions made by preceding classifiers in the chain [5]. The classifier chain technique preserves label correlations within the dataset by considering the order specified by the chain during both training and testing phases. During training, each classifier learns to predict its assigned label based on the input features and the predictions of preceding classifiers. During testing, the predictions made by earlier classifiers in the chain are incorporated as input features for subsequent classifiers.

3.8 Novel ETRF Feature Engineering Approach

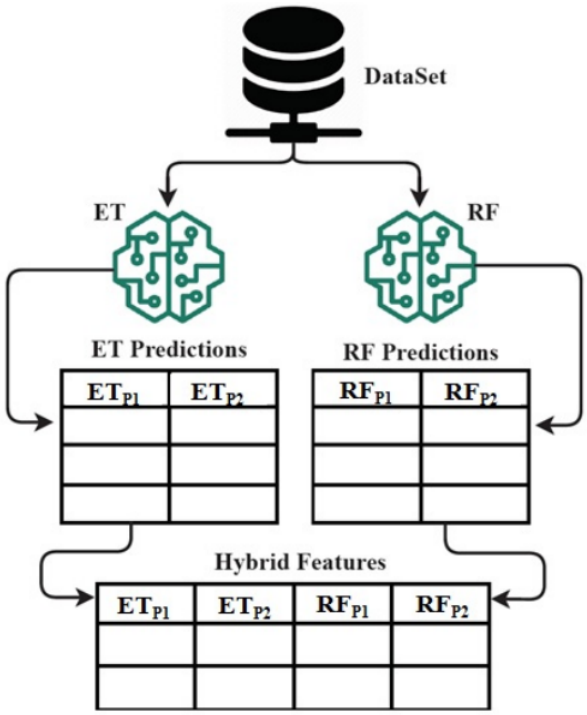

Figure 3.8: The architecture analysis of ETRF technique for hybrid feature set formation mechanism.

Figure 3.8 illustrates the architectural analysis of ETRF technique for hybrid feature set formation mechanism. The ETRF technique, combining ET and RF algorithms, is explored for feature extraction in predicting genetic disorders. Genomes data samples are separately processed by ET and RF algorithms, and class predicted probabilities are extracted. These probabilities are then combined to form a hybrid feature set, serving as input for learning models to predict genetic disorders and their types. This innovative approach enhances predictive accuracy by leveraging the strengths of both algorithms.

3.9 Results

Results and evaluations of the proposed research approach are examined in this section. The machine learning models are utilized to predict the genetic disorders and types of genetic disorders.

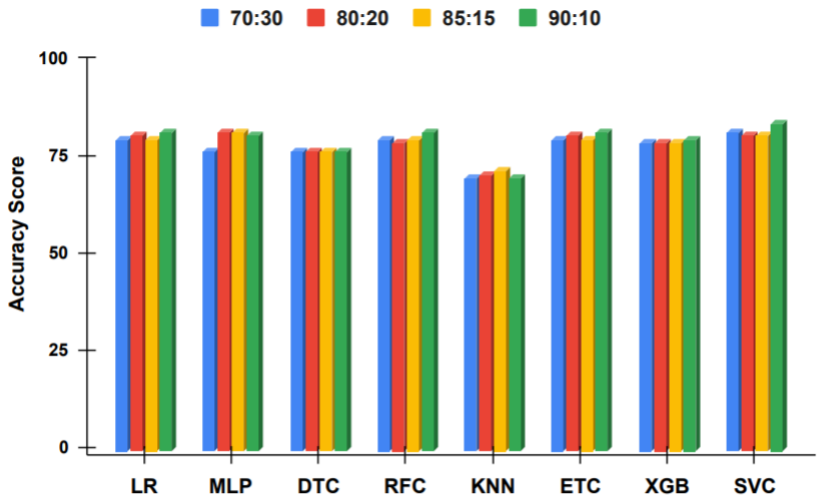

Figure 3.9: The comparative analysis of different data split ratios without proposed technique using balanced data.

Figure 3.9 illustrates the comparative analysis of different data split ratios without the proposed technique, specifically focusing on balanced data. The analysis indicates fluctuations in accuracy, precision, recall, and F1 scores, emphasizing the sensitivity of model outcomes to the distribution of training and testing data.

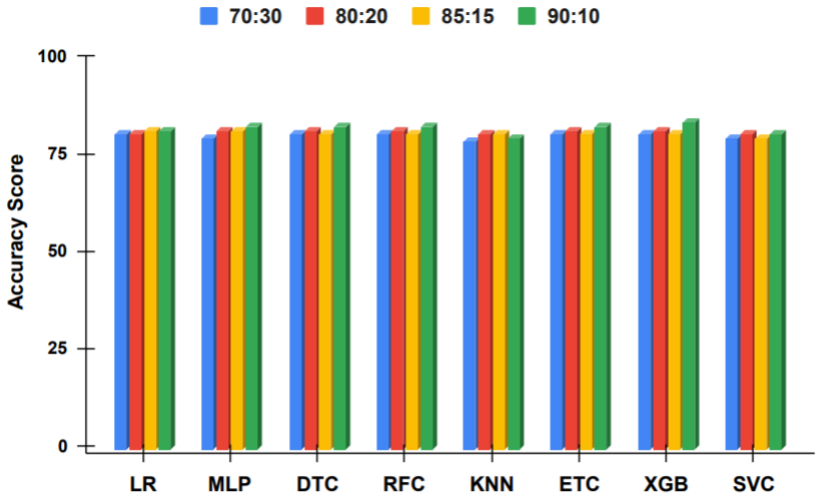

Figure 3.10: The comparative analysis of different data split ratios with proposed technique using balanced data.

Figure 3.10 illustrates the comparative analysis of different data split ratios with the proposed technique, specifically focusing on balanced data. This suggests that the introduced approach effectively mitigates the sensitivity of machine learning models to

changes in training and testing set distributions. The improved and sustained metrics such as accuracy, precision, recall, and F1 scores underscore the reliability and robustness achieved by employing the proposed technique in handling balanced data across different split configurations.

Table 3.2: Performance analysis of machine learning models using an balanced dataset with a data split of 90:10.

Models	Label 1				Label 2			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Results without Proposed Technique								
LR	53	57	42	37	34	24	17	14
MLP	57	56	49	48	36	31	24	24
DTC	50	44	45	44	30	23	24	24
RFC	58	56	46	47	38	39	23	25
KNN	46	34	33	32	22	13	12	12
ETC	59	56	48	49	38	40	26	28
XGB	57	52	49	50	36	33	26	28
SVC	51	40	41	35	31	15	19	15
Results with Proposed Technique								
LR	66	77	71	68	43	40	38	37
MLP	67	76	73	71	45	45	40	40
DTC	67	76	73	72	45	50	41	41
RFC	67	76	73	73	45	51	41	41
KNN	62	71	71	70	41	50	42	42
ETC	67	76	73	72	45	50	41	41
XGB	67	76	73	72	45	50	42	42
SVC	65	76	70	66	43	41	38	36

Table 3.2 presents a comprehensive performance analysis of machine learning models applied to a balanced dataset, employing an 90:10 data split. The table offers a detailed comparison of key metrics, including accuracy, precision, recall, and F1 score, providing a nuanced understanding of the models' effectiveness. The observed variations in these metrics underscore the impact of the proposed approach on enhancing the models' overall performance under the specified balanced dataset conditions.

Chapter 4

Technical Discussion

The Multi-Label Multi-Class Chain Classifier (ML-MC-Chain) approach is a strategy used in machine learning for scenarios where instances may be associated with multiple labels, each of which belongs to multiple classes. It involves constructing a chain of binary classifiers, where each classifier predicts the presence or absence of a specific label while considering the labels predicted by previous classifiers. This sequential prediction process enables the model to capture complex dependencies between labels and classes effectively.

When integrated into the ETRF framework, the ML-MC-Chain approach improves its capacity for multi-label multi-class classification. Each tree acts as a binary classifier for a specific label, with the algorithm constructing multiple trees during training. This hybrid approach enables ETRF to effectively exploit label and class dependencies, enhancing accuracy and robustness in classification tasks.

Let's predict one genetic disease along with its subclass using the ClassifierChain approach with Extra Trees (ET) and Random Forest (RF) classifiers. By analyzing patient data such as age and physiological measurements, we aim to predict the occurrence of specific genetic diseases and their subclasses.

Instance Details:

- Patient Age: 2
- Blood cell count: 4000
- Genes in mother's side: 1
- Inherited from father: 0
- Gender: 0(Male)
- Heart Rate: 0(Normal)

Prediction Steps:

1. ET(P1) Prediction:

- ET(P1) receives the input features for a specific instance.Then uses the provided input features to predict the presence of genetic disorders. It employs the Extra Trees classifier algorithm for this task.
- Extra Trees classifier is trained on a dataset containing various input features and corresponding labels indicating the presence or absence of genetic disorders.
- ET(P1) uses a decision tree ensemble approach to make predictions. It constructs multiple decision trees based on random subsets of features and data samples. These decision trees collectively make predictions, and the final prediction is determined by a majority vote or averaging mechanism.
- ET(P1) utilizes the learned patterns and relationships from the training data to predict the likelihood of the input instance belonging to various genetic disorder categories.
- After predicting the genetic disorder categories, ET(P1) assigns probabilities to each predicted category. Here it assigns 0.8 for Mitochondrial genetic inheritance.
- ET(P1) employs a predefined probability threshold to determine the final prediction. Otherwise, if the predicted probability is below the threshold, ET(P1) may classify it as not having the disorder.
- The final output predicts the presence of the genetic disorder class "Mitochondrial genetic inheritance" based on the provided input features.

2. ET(P2) Prediction:

- ET(P2) utilizes the Extra Trees classifier algorithm to make predictions, similar to ET(P1). Before making predictions, ET(P2) is trained on a dataset containing various input features and corresponding labels indicating the subclasses of genetic disorders.
- ET(P2) relies on the prediction made by ET(P1) regarding the genetic disorder. The genetic disorder predicted by ET(P1) serves as crucial input for ET(P2) in predicting the subclass.
- Considering the genetic disorder predicted by ET(P1) and other input features, ET(P2) predicts the likelihood of the input instance belonging to different subclasses of genetic disorders.
- ET(P2) assigns probabilities to each predicted subclass of genetic disorders. For example, it may predict "Leber's hereditary optic neuropathy" with a probability of 0.9 and assign lower probabilities to other subclasses.

- ET(P2) may utilize a predefined probability threshold to determine the final prediction of the subclass.
- The final output of ET(P2) is the prediction of the subclass of the genetic disorder based on the provided input features, such as "Leber's hereditary optic neuropathy".

3. RF(P1) Predictions:

- RF(P1) utilizes the Random Forest classifier algorithm to predict the presence of genetic disorders.
- Before making predictions, the Random Forest classifier is trained on a dataset containing various input features and labels indicating the presence or absence of genetic disorders.
- RF(P1) employs decision trees to make predictions, combining the results for the final prediction. It constructs an ensemble of decision trees, each trained on a random subset of features and data samples.
- Based on the learned patterns from the training data, RF(P1) predicts the likelihood of the input instance belonging to different genetic disorder categories. (P1) assigns probabilities to each predicted genetic disorder category. For example, it may assign a probability of 0.75 for "Mitochondrial genetic inheritance" and lower probabilities to other genetic disorders.
- The threshold in RF(P1) is set by evaluating the model's performance, usually by analyzing the ROC curve or using cross-validation. It helps balance between correctly identifying cases of "Mitochondrial genetic inheritance" (sensitivity) and avoiding false positives. Instances with predicted probabilities above this threshold are classified as having the disorder.
- The final output predicts the presence of the genetic disorder class "Mitochondrial genetic inheritance" based on the provided input features.

4. RF(P2) Predictions:

- RF(P2) utilizes the Random Forest classifier algorithm to predict subclasses of genetic disorders based on features similar to ET(P2).
- Like RF(P1), RF(P2) employs decision trees to make predictions by creating an ensemble of trees trained on random subsets of features and data samples.
- Predictions made by RF(P2) entail determining the likelihood of the input instance belonging to different subclasses of genetic disorders based on learned patterns from the training data.
- RF(P2) assigns probabilities to each predicted subclass, with higher probabilities indicating a greater likelihood of the instance belonging to that subclass.

- The threshold for RF(P2) is determined through model evaluation, often involving the analysis of performance metrics like the ROC curve or using cross-validation, with instances classified into subclasses if their predicted probabilities exceed this threshold.
- The final output of RF(P2) is the prediction of the subclass of the genetic disorder based on the provided input features, such as "Leber's hereditary optic neuropathy."

5. Combining Predictions:

- Obtain predictions from both Extra Trees (ET) and Random Forest (RF) classifiers for the target variable (genetic disorder and its subclass).
- ET(P1) achieved an accuracy of 80%, while RF(P1) achieved 75% accuracy. We normalize these accuracies to obtain weights for each classifier's predictions.
- ET(P1)'s weight is approximately 0.516, and RF(C1)'s weight is approximately 0.484, ensuring they sum up to 1. ET(P1) = 0.8 / (0.8 + 0.75) = 0.516 and Weight for RF(P1) = 0.75 / (0.8 + 0.75) = 0.484.
- These weights reflect the relative reliability or performance of each classifier in predicting the presence of "Mitochondrial genetic inheritance" in patients.
- If we set a threshold of 0.75 for combined probabilities, any instance with a total probability exceeding this value (for example, if ET(P1) predicts 0.8 and RF(P1) predicts 0.75, resulting in a total probability of 1.55) would be considered as having the disorder.
- After combining predictions and applying a threshold, we assess the model's performance on a validation dataset.

6. Final Prediction:

- Based on these predictions, the instance is predicted to have "Mitochondrial Genetic Inheritance Disorder" with the subclass of "Leber's hereditary optic neuropathy".

Chapter 5

Applications

AI plays a crucial role in enhancing gene editing techniques, predicting genetic diseases, and discovering novel treatments. By leveraging AI algorithms, researchers can optimize gene therapy approaches, identify individuals at risk of genetic disorders, and gain deeper insights into the underlying mechanisms of genetic diseases, ultimately advancing precision medicine initiatives. Here are the listed applications mentioned below.

- **Pharmacogenomics:** AI-driven pharmacogenomics analyzes genomic data to predict individual drug responses, enabling personalized treatment plans. By identifying genetic variations affecting drug metabolism and efficacy, it enhances precision medicine. This tailored approach minimizes adverse reactions and maximizes treatment effectiveness, revolutionizing patient care and drug development.
- **Biomedical Research:** AI algorithms analyze genetic data to pinpoint disease-associated mutations, aiding in the discovery of novel therapeutic targets, accelerating drug development, and paving the way for personalized medicine approaches.
- **Agriculture:** AI algorithms analyze genetic data in crops and livestock to improve breeding strategies, enhance yield, and develop disease-resistant varieties.
- **Gene Editing Optimization:** AI algorithms utilize CRISPR-Cas9 to precisely target disease-causing mutations, improving the efficiency and safety of gene therapy. By optimizing guide RNA design and minimizing off-target effects, AI enhances the effectiveness of genome editing, promising more accurate treatment of genetic diseases.
- **Genomic Data Sharing Platforms:** AI-powered platforms facilitate secure and interoperable sharing of genomic data among researchers and healthcare providers. By leveraging machine learning for data analysis and privacy protection, these platforms accelerate collaboration and the translation of genetic insights into clinical practice, advancing personalized medicine.

- **Protein Structure Prediction:** AI algorithms predict the three-dimensional structures of proteins encoded by genetic variants, aiding in understanding mutation impact on disease development. By accurately modeling protein structures, AI contributes to drug discovery efforts and the development of targeted therapies for genetic disorders.

- **Forensic Genetics:** AI assists in analyzing DNA evidence for identifying suspects, victims, and familial relationships in forensic investigations. By automating complex DNA analysis tasks, AI improves the accuracy and efficiency of criminal justice applications, ensuring reliable outcomes in legal proceedings.

- **Non-Invasive Prenatal Testing:** AI-based analysis of cell-free fetal DNA in maternal blood samples enables the detection of chromosomal abnormalities and genetic disorders in the fetus. This non-invasive approach offers a safer alternative to traditional prenatal diagnostic methods, empowering expectant parents with early, accurate information about their baby's health.

- **Identification of Disease-Causing Mutations:** AI algorithms identify genetic mutations associated with disease by correlating genomic data with clinical phenotypes. Through sophisticated analysis techniques, these algorithms pinpoint candidate mutations implicated in disease pathogenesis, guiding further research and therapeutic development.

- **Drug Repurposing:** AI-driven approaches predict drug-gene interactions and identify existing drugs with therapeutic potential for genetic disorders. By repurposing drugs for new indications based on genomic data analysis, AI accelerates the development of targeted therapies, potentially bringing effective treatments to patients more rapidly and cost-effectively.

- **Population Screening Programs:** AI-driven screening programs analyze large-scale genomic datasets to identify individuals at higher risk of common genetic disorders. By integrating machine learning algorithms, these programs enable targeted preventive measures and population-level interventions, ultimately improving public health outcomes. AI prioritizes individuals for further genetic testing or intervention based on their genetic risk profile, allowing for more efficient allocation of healthcare resources and reducing the burden of genetic diseases on healthcare systems.

Chapter 6

Advantages and Disadvantages

6.1 Advantages

The integration of artificial intelligence into genomic analysis allows for swift and thorough processing of extensive genetic data, leading to earlier and more accurate diagnoses of genetic disorders. AI's efficiency in handling large datasets results in expedited and precise diagnoses, paving the way for tailored treatment approaches.

- **Enhanced Data Analysis:** The rapid and thorough analysis enabled by AI algorithms allows researchers to uncover intricate patterns and associations within vast genomic datasets that might otherwise go unnoticed with manual analysis methods alone. These insights hold the potential to unlock new understandings of genetic diseases and inform targeted therapeutic interventions.

- **Improved Diagnosis Accuracy:** AI systems excel at detecting subtle patterns in genomic data, augmenting the diagnostic process for genetic disorders with increased accuracy and efficiency. By complementing human expertise, AI contributes to more timely and precise diagnoses, empowering healthcare professionals to provide better patient care.

- **Cost-Effectiveness:** Automated genomic analysis driven by AI streamlines the testing process, potentially reducing both the time and financial resources required for genetic testing. This increased efficiency makes genetic testing more accessible and affordable for patients, ultimately improving overall healthcare accessibility and equity.

- **Personalized Medicine:** AI-powered genomic analysis enables healthcare providers to tailor treatment plans according to an individual's unique genetic profile. By leveraging this personalized approach, healthcare professionals can optimize treatment effectiveness and address specific genetic conditions, leading to better patient outcomes and quality of life.

- **Faster Insights:** AI algorithms expedite the identification of disease-causing genetic variants, accelerating the diagnostic process and enabling earlier interventions for patients. This swift turnaround time for insights enhances patient care by allowing for timely treatment adjustments and proactive management strategies.
- **Research Advancements:** AI-driven analysis of large-scale genomic datasets fuels advancements in genomic research by uncovering novel genetic associations and generating hypotheses for further investigation. By expediting the research process, AI contributes to the rapid expansion of our understanding of genetic disorders and potential therapeutic targets.
- **Genetic Counseling Support:** AI tools provide valuable support to genetic counselors by assisting in the interpretation of complex genomic data. This aid enhances the genetic counseling process, enabling counselors to have more informed discussions with patients and families regarding genetic risks, testing options, and treatment considerations.
- **Pattern Recognition:** AI algorithms excel in recognizing complex patterns within genomic, gene expression, and epigenetic data. Through extensive training on large datasets, these algorithms can identify subtle variations and correlations, crucial for understanding genetic mutations and their implications in disease.
- **Variant Classification:** AI-driven systems classify genetic variations based on their characteristics and predicted impact. By categorizing variants into different classes, these systems help prioritize variants for further investigation, distinguishing between benign polymorphisms and pathogenic mutations.
- **Integrative Analysis:** AI tools comprehensively analyze diverse genomic data types, including DNA sequences, gene expression profiles, and epigenetic modifications. By integrating multiple layers of information, these tools offer a comprehensive understanding of genetic landscapes, unraveling complex disease mechanisms.

6.2 Disadvantages

The integration of AI in genomic analysis presents several disadvantages, including concerns regarding data privacy, potential biases in algorithms, and challenges in interpretation. Furthermore, limited accessibility, ethical considerations, and the potential for over diagnosis or misdiagnosis are important factors to consider. Here are some of the listed disadvantages:

- **Data Privacy Concerns:** The integration of AI in genomic analysis raises significant concerns regarding data privacy and security. Safeguarding sensitive genetic

information from unauthorized access and misuse becomes paramount, necessitating robust data protection measures and stringent regulatory oversight to ensure patient confidentiality and trust in healthcare systems.

- **Ethical Considerations:** The utilization of AI in genomic analysis introduces complex ethical dilemmas surrounding consent, transparency, and the responsible use of genetic information. Issues such as informed consent for genetic testing, equitable access to genomic technologies, and the potential for discrimination based on genetic predispositions require careful consideration and robust ethical frameworks to mitigate risks and uphold patient autonomy and justice.

- **Technical Complexities and Errors:** AI-driven genomic analysis may encounter technical complexities and algorithmic errors, leading to uncertainties and vulnerabilities in the diagnostic process. Challenges such as algorithm biases, model overfitting, and insufficient training data can compromise the accuracy and reliability of AI-generated insights, necessitating ongoing validation and refinement of machine learning models to enhance diagnostic accuracy and minimize errors.

- **Interpretation Challenges:** AI-generated insights from genomic data may be challenging to interpret or validate, leading to uncertainty and potential errors in diagnosis and treatment decision-making. Complex genetic interactions, rare variants, and incomplete understanding of the genetic basis of certain diseases can pose challenges in interpreting AI-driven diagnostic results, requiring collaboration between AI systems and healthcare professionals to ensure accurate and clinically relevant interpretations.

- **Data Integrity and Quality:** Ensuring the integrity and quality of genomic data used in AI-driven analysis is crucial for reliable diagnostic results. Inaccuracies or inconsistencies in data collection, processing, or storage can undermine the reliability of AI-generated insights, potentially leading to misdiagnosis or inappropriate treatment decisions. Rigorous data quality control measures, standardized data collection protocols, and robust data governance frameworks are essential to maintain data integrity and enhance the reliability of AI-driven diagnostic processes.

- **Lack of Standardization:** The absence of standardized protocols and guidelines for AI-driven genomic analysis may result in inconsistencies in data interpretation and treatment recommendations. Without clear standards for data processing, analysis methodologies, and reporting practices, there is a risk of variability in patient care quality and diagnostic accuracy across different healthcare settings. Establishing standardized practices and regulatory frameworks is essential to ensure consistency, reproducibility, and reliability in AI-driven genomic analysis.

- **Over-reliance on Technology:** Over-reliance on AI without clinical validation may lead to misdiagnoses. Balancing AI insights with clinical judgment is crucial for safe and effective genetic disease detection, emphasizing the importance of human oversight in healthcare decision-making.
- **Limited Generalization:** AI models trained on specific datasets may struggle to generalize findings to diverse populations or novel genetic variants. This limitation can hinder the effectiveness of AI-driven genetic analysis in real-world clinical settings, requiring continuous adaptation and validation to ensure applicability across diverse patient populations.

Chapter 7

Conclusion and Future Work

Advances in machine learning can significantly improve diagnosis, treatment and prognosis of rare disease patients. It is believed that predicting genetic disorders at an early phase of its advent becomes important for a healthy population, to maximise comfort of the patient and retard its growth. Early detection and medical interventions can prevent many severe complications.

On the other hand, the goals of AI/ML algorithms in RDs using sequencing data are broad, ranging from patient stratification to the identification of possible pathogenic combinations of variants. However, we found common patterns in these goals when configuring the datasets with which these models are trained, identifying key features for each of the objectives. Finally, we identified possible future challenges, such as the use of CNV to train the AI/ML models, or the application of AI/ML for the stratification of patients with non-neoplastic RDs. Thus, this systematic review can be used as a reference for further studies, supporting the development of future ML models in the diagnosis of rare genetic diseases.

References

[1] Sadichchha Naik, Disha Nevare, Amisha Panchal, Dr. Chhaya Pawar, "Prediction of Genetic Disorders using Machine Learning", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X, Volume 9, Issue 3, pp.01-09, May-June-2022. Available at https://doi.org/10.32628/IJSRST229273

[2] P. Roman-Naranjo, A.M. Parra-Perez, J.A. Lopez-Escamez https://doi.org/10.1016/j.jbi.2023.104429 "A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases.", vol.143, July 2023, 104429.

[3] Julia Schaefer, Moritz Lehne, Josef Schepers,Fabian Prasse and Sylvia Thun."The use of machine learning in rare diseases: a scoping review". Orphanet J Rare Dis 15, 145 (2020) https://doi.org/10.1186/s13023-020-01424-6. f

[4] Gang Wang, Yuyan Xu, Qintao Wang, Yi Chai, Xiangwei Sun, Fan Yang, Jian Zhang, Mengchen Wu, Xufeng Liao, Xiaomin Yu, Xin Sheng, Zhihong Liu, Jin Zhang, "Rare and undiagnosed diseases: From disease-causing gene identification to mechanism elucidation", Fundamental Research, Volume 2, Issue 6, 2022, pp 918-928, ISS 2667-3258, https://doi.org/10.3390/genes14010071

[5] 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