VISVESVARAYA TECHNOLOGICAL UNIVERSITY,
JNANASANGAMA, BELAGAVI – 590018

TECHNICAL SEMINAR REPORT(18CSS84)

on

Genomic Intelligence: The Impact of AI in Uncovering Genetic Disordersl

Submitted in partial fulfillment of the award of degree

Bachelor of Engineering

in

Computer Science & Engineering

Submitted by

Joyline Rencita Dsouza

4SO20CS073

Under the Guidance of

Dr Saumya Y M

Department of Computer Science and Engineering

St Joseph Engineering College

Mangaluru - 575028

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 [1]. 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 [1].

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 [2]. 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 [3].

Table of Contents

Abstract i

Table of Contents iii

List of Figures iv

List of Tables v

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Traditional Approaches . . . 1
  - 1.2.1 Fluorescence In Situ Hybridization (FISH) . . . 2
  - 1.2.2 Polymerase Chain Reaction (PCR) . . . 2
  - 1.2.3 Pedigree Analysis . . . 2
  - 1.2.4 Cytogenetic Analysis . . . 3

2 Methedology 4

3 Genetic Disease Detection 5
- 3.1 Genomes and Genetic Dataset . . . 6
- 3.2 Genetic Exploratory Data Analysis (GEDA) . . . 6
- 3.3 Data Normalization and Feature Engineering . . . 9
- 3.4 Data Balancing . . . 11
- 3.5 Data Splitting . . . 11
- 3.6 Applied Learning Techniques . . . 11
- 3.7 Multi-Label Multi-Class Chain Classifier Approach . . . 12
- 3.8 Novel ETRF Feature Engineering Approach . . . 13
- 3.9 Results . . . 14

4 Applications 16

5 Advantages and Disadvantages 18
- 5.1 Advantages . . . 18
- 5.2 Disadvantages . . . 19

6 Conclusion and Future Work **20**

References **21**

List of Figures

1.1 Pedigree Analysis 3

3.1 Illustration of the proposed approach 5
3.2 Distributions of samples for different classes in the dataset, (a) genetic disorders' main classes, and (b) genetic disorder subclasses. 7
3.3 Data distribution by genetic disorder category, (a) maternal gene, (b) paternal gene, (c) genes from mother side, and (d) inherited from father. . . 7
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. 8
3.5 Age analysis of patients for the disorder category. 9
3.6 Feature correlation analysis graphs of genomes data. 10
3.7 The architectural analysis of the multi-label multi-class classifier chain approach. 12
3.8 The architecture analysis of ETRF technique for hybrid feature set formation mechanism. 13
3.9 The comparative analysis of different data split ratios without proposed technique using balanced data. 14
3.10 The comparative analysis of different data split ratios with proposed technique using balanced data. 14

List of Tables

2.1 Comparative analysis of different proposed approaches. 4

3.1 The genomes dataset features descriptive analysis 6
3.2 Performance analysis of machine learning models using an balanced dataset with a data split of 80:20. 15

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.

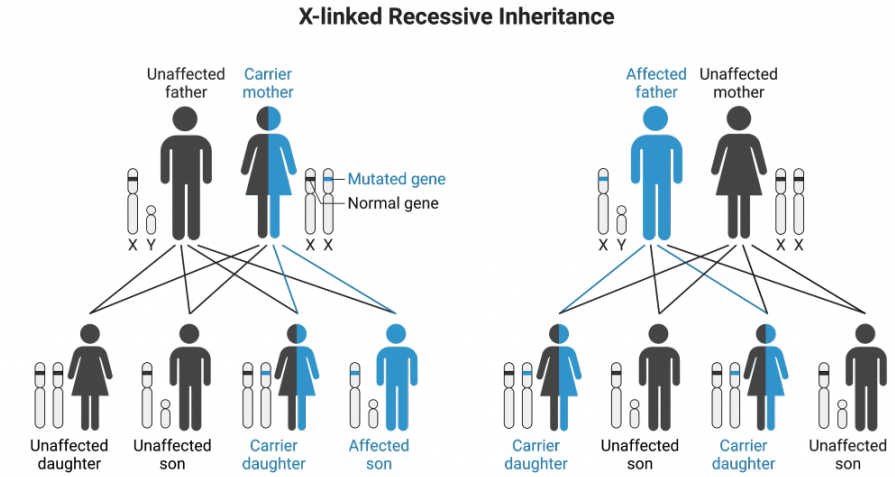

Figure 1.1: Pedigree Analysis

In Figure 1.1, the X-linked recessive inheritance of color blindness is illustrated. If a father is color blind (X-linked), daughters can be carriers. If a carrier daughter has sons, they have a 50 percent chance of being color blind. Pedigree charts track inheritance patterns, aiding diagnosis and genetic counseling for families with color vision deficiency.

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. Further details on this successful method will be provided in the following chapter.

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

Table 2.1: Comparative analysis of different proposed approaches.

Table 2.1 presents a comparative analysis of various proposed approaches discussed 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.

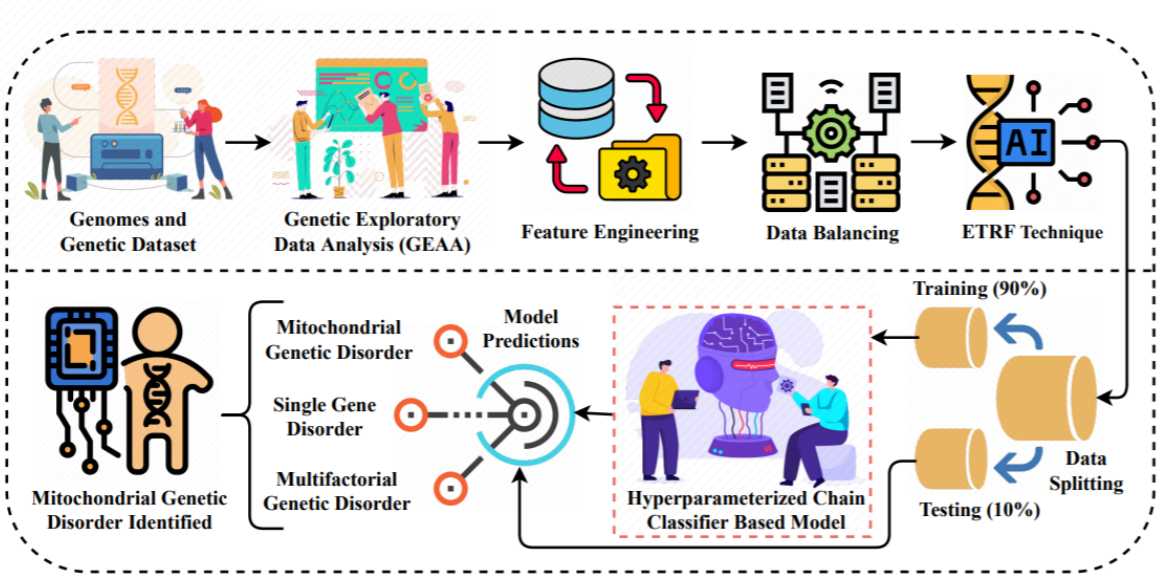

Figure 3.1: Illustration of the proposed approach

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 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.

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: The genomes dataset features descriptive analysis

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 subclasses, 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.

The Figure 3.2 (a) illustrates distributions of genetic disorders' main classes, while the

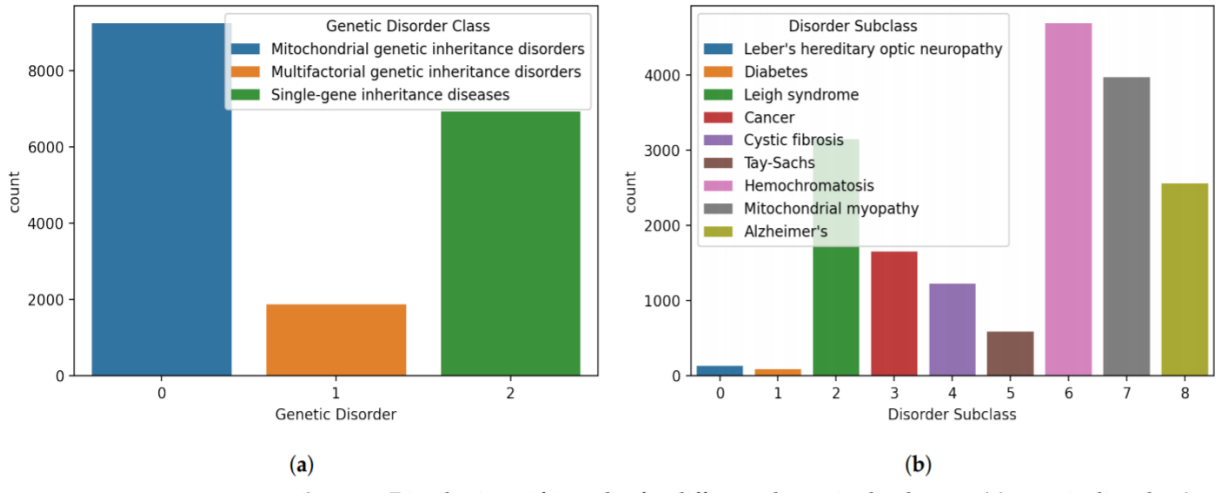

Figure 3.2: Distributions of samples for different classes in the dataset, (a) genetic disorders' main classes, and (b) genetic disorder subclasses.

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 .

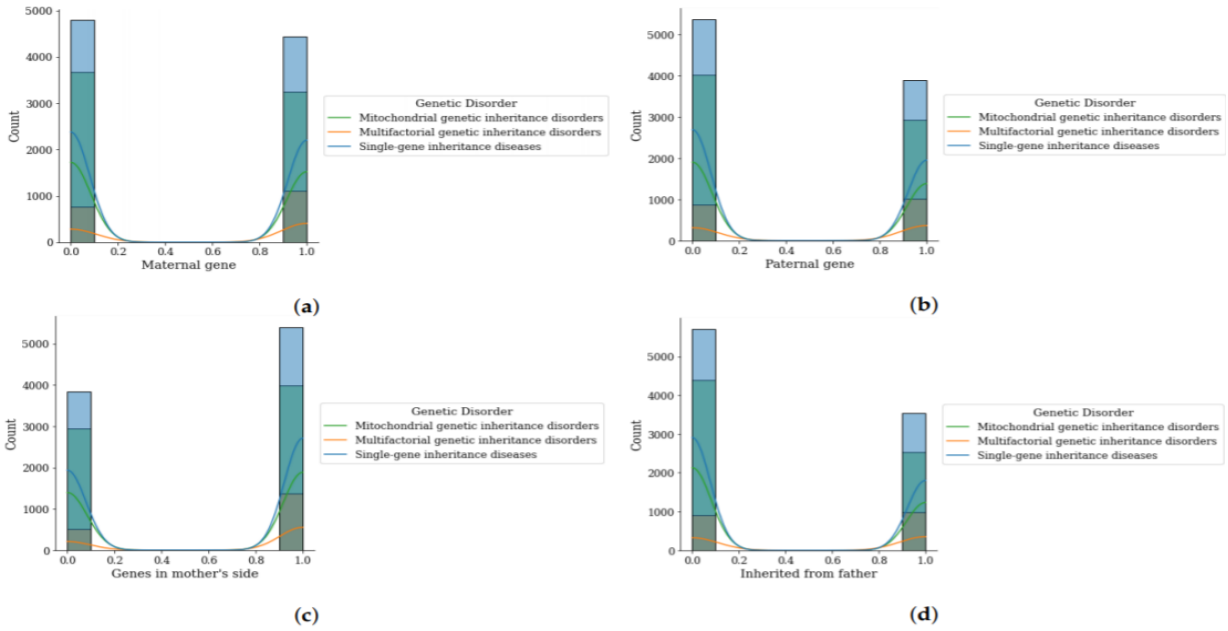

Figure 3.3: Data distribution by genetic disorder category, (a) maternal gene, (b) paternal gene, (c) genes from mother side, and (d) inherited from father.

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

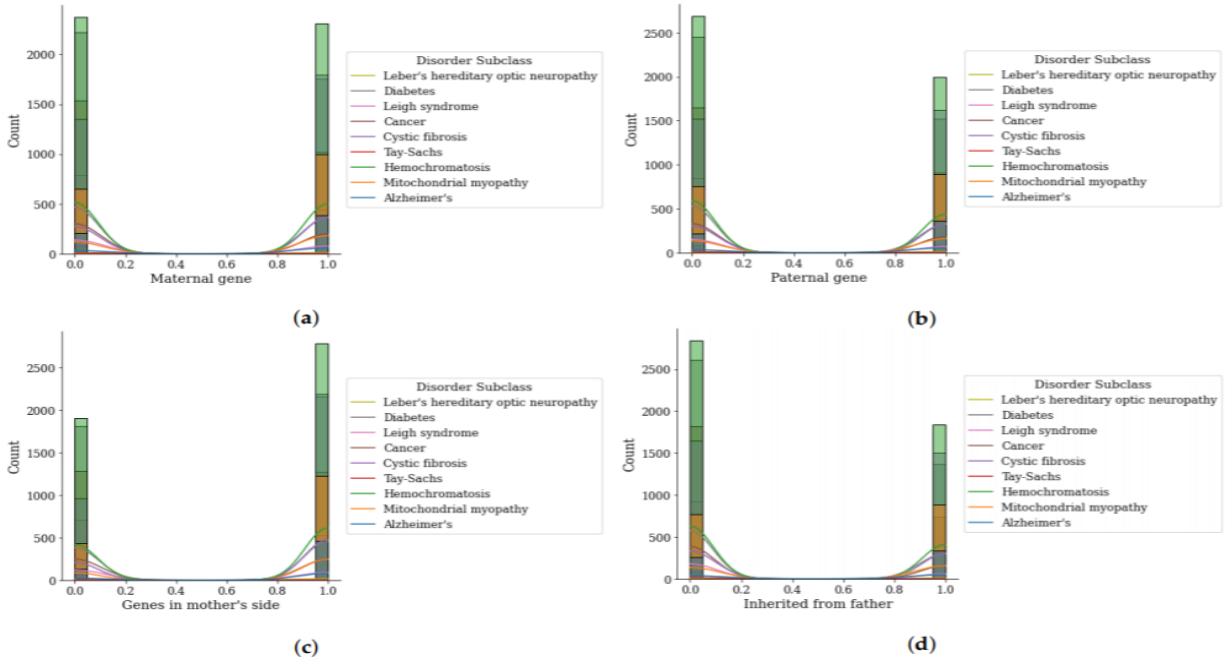

Figure 3.4: Data distribution by genetic disorder sub-category, (a) maternal gene, (b) paternal gene, (c) genes from mother side, and (d) inherited from father.

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.

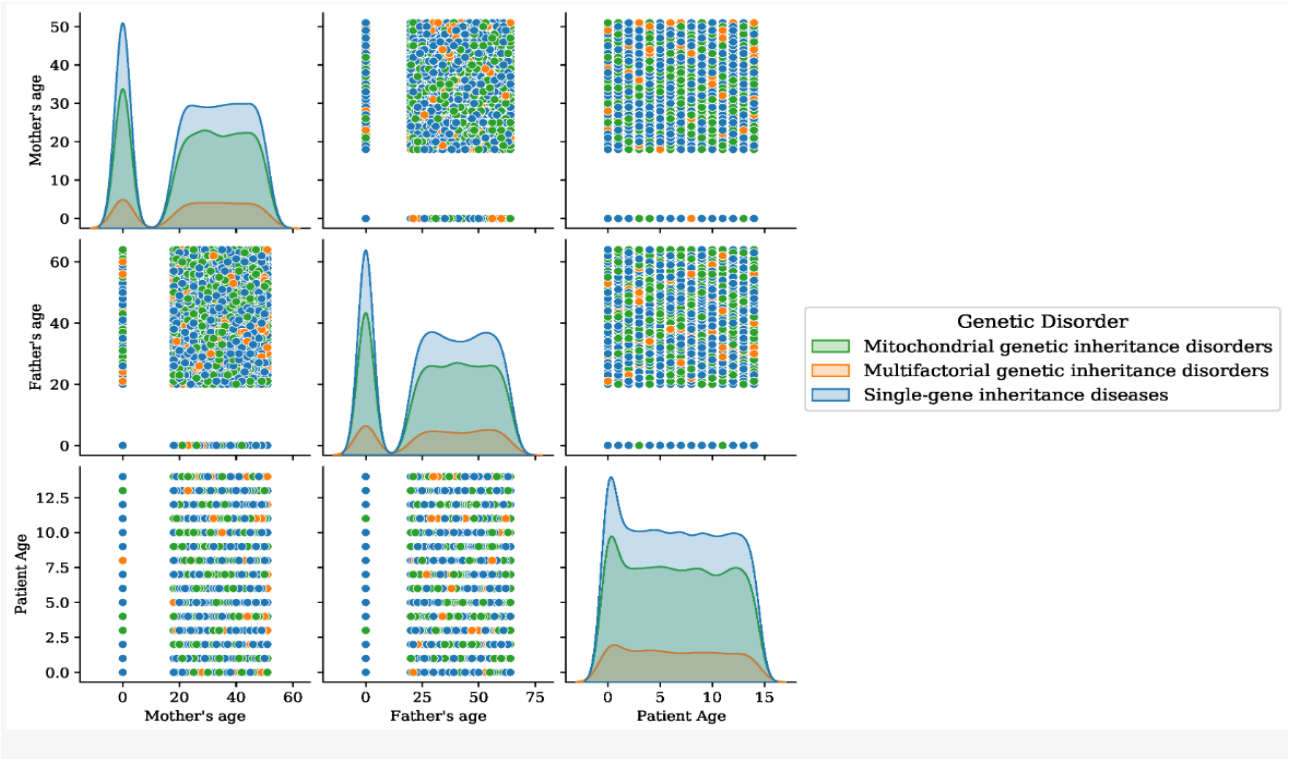

Figure 3.5: Age analysis of patients for the disorder category.

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.

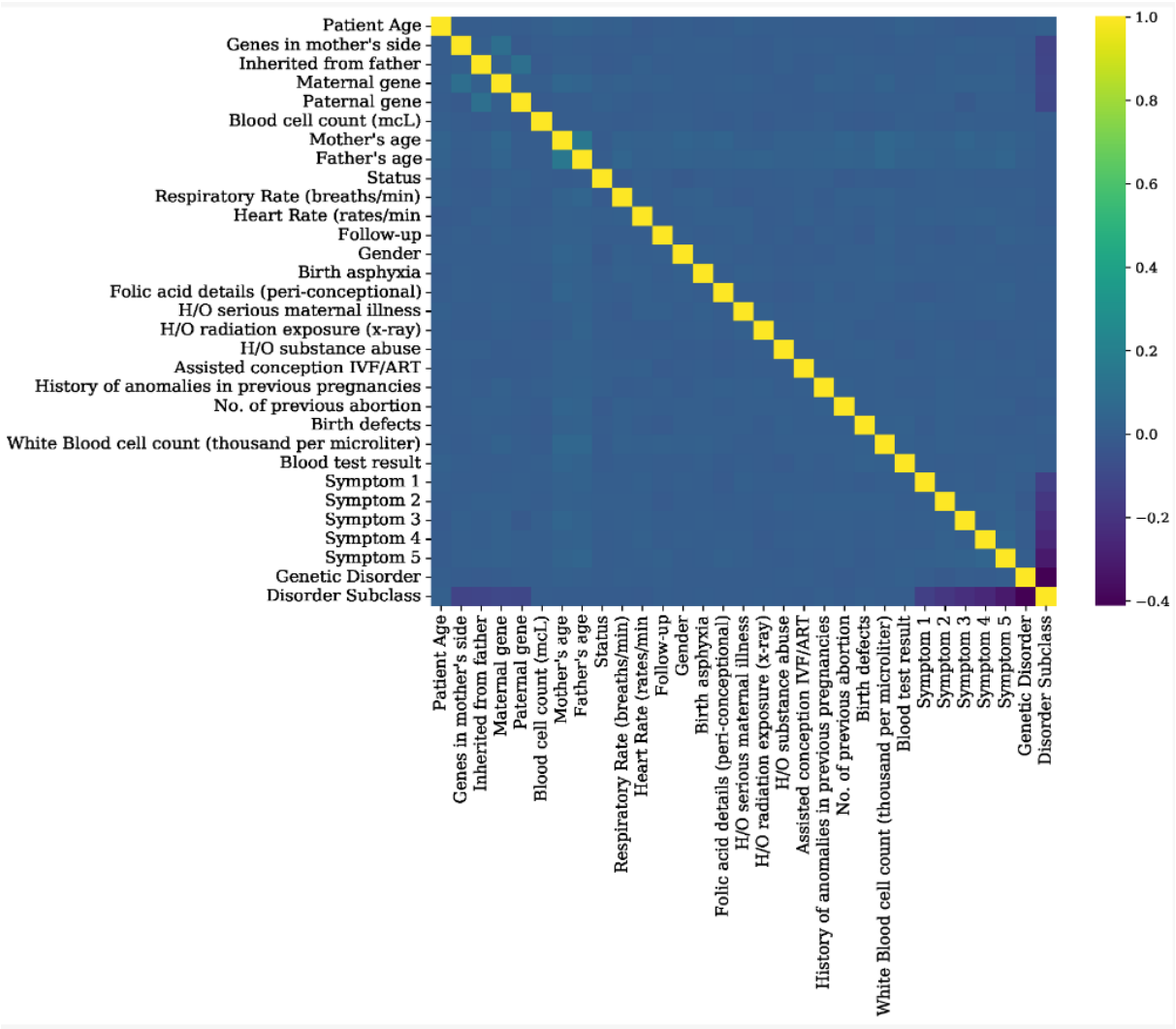

Figure 3.6: Feature correlation analysis graphs of genomes data.

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

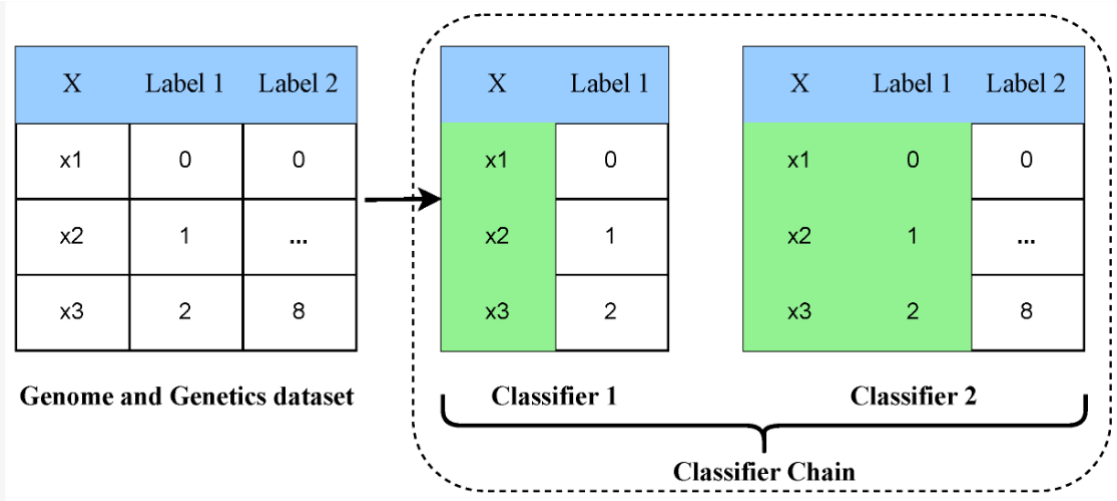

Figure 3.7: The architectural analysis of the multi-label multi-class classifier chain approach.

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

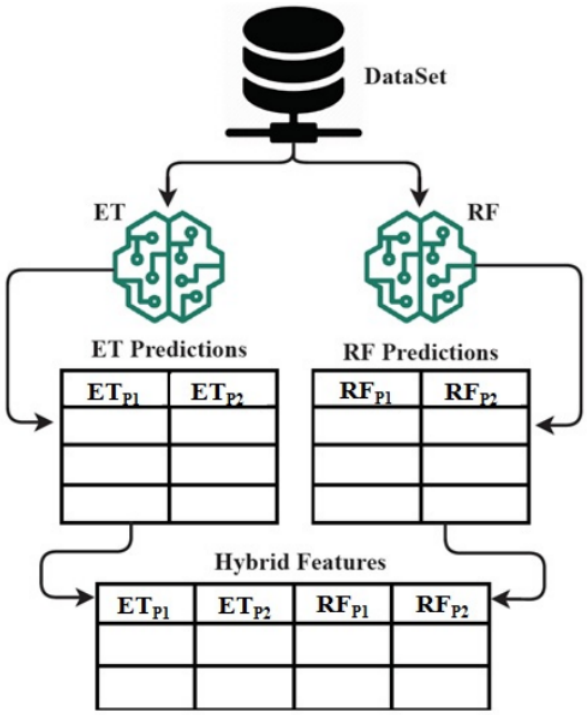

Figure 3.8: The architecture analysis of ETRF technique for hybrid feature set formation mechanism.

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.

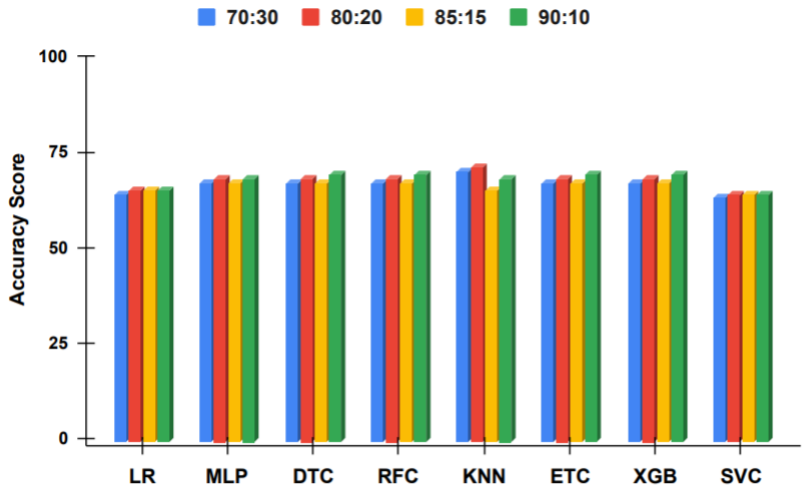

Figure 3.9: The comparative analysis of different data split ratios without proposed technique using balanced data.

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.

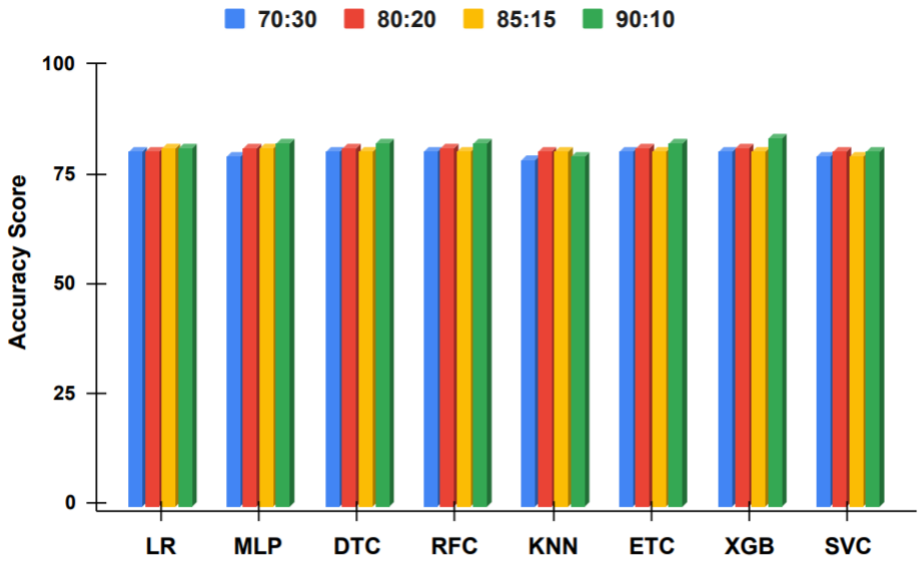

Figure 3.10: The comparative analysis of different data split ratios with proposed technique using balanced data.

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.

Technique	Label 1				Label 2			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Results without Proposed Technique								
LR	53	52	42	37	33	19	17	14
MLP	60	54	50	51	37	27	24	25
DTC	49	43	43	43	29	21	21	21
RFC	58	54	45	45	36	27	21	22
KNN	47	36	34	34	22	12	12	12
ETC	59	55	48	49	37	34	23	25
XGB	57	52	48	49	36	32	26	28
SVC	50	45	35	32	19	17	16	12
Results with Proposed Technique								
LR	65	75	71	68	43	39	39	37
MLP	67	75	73	72	45	48	40	39
DTC	67	75	73	72	45	55	41	41
RFC	67	75	73	72	45	54	41	40
KNN	55	69	69	66	37	43	42	40
ETC	67	75	73	72	45	55	41	41
XGB	67	75	73	72	45	55	41	41
SVC	65	75	70	67	43	40	38	36

Table 3.2: Performance analysis of machine learning models using an balanced dataset with a data split of 80:20.

Table 3.2 presents a comprehensive performance analysis of machine learning models applied to a balanced dataset, employing an 80:20 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

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.

- **Gene Editing Optimization:** AI algorithms can optimize gene editing techniques such as CRISPR-Cas9 for precise targeting of disease-causing mutations, enhancing the efficiency and safety of gene therapy approaches.
- **Genomic Data Sharing Platforms:** AI-powered platforms for secure and interoperable sharing of genomic data can facilitate collaboration among researchers and healthcare providers, accelerating the discovery and translation of genetic insights into clinical practice.
- **Protein Structure Prediction:** AI algorithms can predict the three-dimensional structures of proteins encoded by genetic variants, aiding in understanding the functional impact of mutations and their role in disease development.
- **Forensic Genetics:** AI assists in analyzing DNA evidence to identify suspects, victims, and familial relationships in forensic investigations, improving accuracy and efficiency in criminal justice applications.
- **Evolutionary Biology:** AI-driven analysis of genomic data from diverse species helps uncover evolutionary patterns, genetic adaptations, and biodiversity conservation strategies.
- **Genetic Engineering in Agriculture:** AI algorithms can optimize genetic modifications in crops to enhance yield, nutritional value, and resistance to pests and diseases, contributing to food security and sustainability.

- **Non-Invasive Prenatal Testing:** AI-based analysis of cell-free fetal DNA in maternal blood samples can detect chromosomal abnormalities and genetic disorders in the fetus, offering a non-invasive alternative to traditional invasive prenatal diagnostic methods.

- **Drug Repurposing:** AI-driven approaches can predict drug-gene interactions and identify existing drugs with potential therapeutic benefits for genetic disorders, facilitating drug repurposing efforts and accelerating the development of targeted therapies.

- **Genomic Data Sharing Platforms:** AI-powered platforms for secure and interoperable sharing of genomic data can facilitate collaboration among researchers and healthcare providers, accelerating the discovery and translation of genetic insights into clinical practice.

- **Population Screening Programs:** ML-driven screening programs analyze large scale genomic datasets to identify individuals at higher risk of common genetic disorders, enabling targeted preventive measures and population-level interventions.

Chapter 5

Advantages and Disadvantages

5.1 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 ability of AI algorithms to swiftly and efficiently analyze vast amounts of genomic data can potentially reveal patterns and associations that might evade detection through manual analysis alone.

- **Improved Diagnosis Accuracy:** By discerning subtle patterns in genomic data, AI systems can enhance the accuracy and timeliness of diagnoses for genetic disorders, complementing human expertise.

- **Cost-Effectiveness:** Automated genomic analysis powered by AI has the potential to reduce the time and resources required for genetic testing, making it more accessible and affordable for patients.

- **Personalized Medicine:** Leveraging AI-driven genomic analysis, healthcare providers can develop tailored treatment plans aligned with an individual's genetic profile, optimizing therapy effectiveness and targeting specific genetic conditions.

- **Faster Insights:** AI algorithms expedite the identification of disease-causing genetic variants, facilitating quicker diagnoses and potentially enabling earlier interventions for patients.

- **Research Advancements:** Through the analysis of large-scale genomic datasets, AI contributes to accelerating genomic research by uncovering novel genetic associations and generating hypotheses for further investigation.

- **Genetic Counseling Support:** AI tools offer valuable assistance to genetic counselors in interpreting complex genomic data, thereby facilitating informed discussions with patients and families, enhancing genetic counseling processes.

5.2 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 utilization of AI in genomic analysis heightens worries about data privacy and security, as safeguarding sensitive genetic information from unauthorized access and misuse becomes paramount.

- **Ethical Considerations :** The integration of AI in genomic analysis raises ethical dilemmas concerning consent, transparency, and the ethical utilization of genetic information, necessitating stringent regulation and oversight.

- **Technical Complexities and Errors :** Technical complexities and algorithmic errors could arise in AI-driven genomic analysis, introducing uncertainties and vulnerabilities in the diagnostic process.

- **Interpretation Challenges:** AI-generated insights from genomic data may be difficult to interpret or validate, leading to uncertainty and potential errors in diagnosis and treatment decision-making.

- **Data Integrity and Quality:** Ensuring the integrity and quality of genomic data used in AI-driven analysis is crucial, as inaccuracies or inconsistencies in data collection, processing, or storage could undermine the reliability of diagnostic results.

- **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, leading to variability in patient care quality.

Chapter 6

Conclusion and Future Work

Advances in machine learning can significantly improve diagnosis, treatment and prognosis of rare disease patients [3]. 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 [1].

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 [2].

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