GENETIC DISORDER PREDICTION BASED ON MACHINE LEARNING

Submitted By

Shahriar Islam Sazid

ID: 202-35-644

Department of Software Engineering

Daffodil International University

Supervised By

Dr. Md. Fazla Elahe

Assistant Professor & Associate Head

Department of Software Engineering

Daffodil International University

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Spring 2024

Approval

APPROVAL

This thesis titled on "**GENETIC DISORDER PREDICTION BASED ON MACHINE LEARNING**", submitted by **Shahriar Islam Sazid (ID: 202-35-644)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Fazla Elahe 10.7.24

Chairman

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been Md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Mostafiz Khan
Managing Director
Tecognize Solutions Limited

Declaration

Declaration

I announce that I am rendering this study document under Dr. Md. Fazla Elahe, Assistant Professor & Associate Head, Department of Software Engineering, Daffodil International University. I therefore, state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised By

Fazla Elahe 24.07.24

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Daffodil International University

Submitted By

Sazid

Shahriar Islam Sazid
ID: 202-35-644
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

I extend my heartfelt gratitude to the Almighty Allah for His abundant blessings, which have enabled me to successfully complete my final year thesis.

I am deeply indebted to **Dr. Md. Fazla Elahe,** Assistant Professor & Associate Head of the Department of Software Engineering at Daffodil International University. His profound knowledge and unwavering dedication to the field of Machine Learning have been instrumental in guiding me through this research journey. I am immensely grateful for his patience, intellectual guidance, continuous encouragement, meticulous supervision, invaluable constructive criticism, insightful counsel, and unwavering support throughout the process. His thorough review of numerous iterations and subsequent corrections have significantly enriched the quality of my work.

I would also like to express my sincere appreciation to **Dr. Imran Mahmud**, Head of the Software Engineering Department at DIU, as well as the other esteemed professors and staff members who have generously contributed to my academic development. Their guidance and assistance have been invaluable in facilitating the successful defense of my thesis and achieving academic excellence.

I am deeply grateful to my parents for their enduring patience, love, and unwavering support throughout my educational journey. Their belief in me has been a constant source of motivation and inspiration.

To all those who have played a part in shaping my academic and personal growth, I extend my heartfelt thanks. Your support and guidance have been instrumental, and I am honored to acknowledge your contributions in this thesis.

Abstract

Genetic disease caused by mutations in DNA or chromosome abnormalities are major global health concerns. These illnesses, which are frequently inherited, cover a broad spectrum of conditions that have significant effects on a person's physical, mental, and social well-being as well as those of their family. The frequency of genetic diseases is still rising, which is made worse by a general lack of knowledge about the significance of genetic testing, even though advancements in genetic testing enable early detection and well-informed decision-making. Most of these ailments are incurable, meaning that continuous observation is required. Although early discovery can have a major impact on disease management and patient outcomes, predicting genetic abnormalities in advance is a tough but important task. With the use of large patient medical datasets, machine learning algorithms present a promising avenue for the extremely accurate identification of genetic diseases. This work proposes a complex genetic maladies prediction system that contains machine learning methods, such as Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression(LR),Gradient Boosting(GB) and Support Vector Machine (SVM). Using datasets covering genetic disorders prediction and shown their subclasses, our ensemble method combines several machine learning models, including Support Vector Machine, Gradient Boosting, Random Forest, K-Nearest Neighbors (KNN). The dataset is graphically evaluated for patterns in genetic illnesses, symptoms, inherited genes, and birth abnormalities after undergoing extensive preprocessing to resolve missing information. Prior to using machine learning models, which are assessed using the following metrics accuracy, recall, precision and F1 score, features are first selected by standardization. To evaluate the performance of classification, confusion matrices are produced.The Ensemble Model provide the highest accuracy of 99.59% for genetic disorders prediction. Other traditional model accuracy like, K-Nearest Neighbors (KNN) 97.54%, Decision Tree 91.9%, Random Forest 96.72%, Gradient Boosting 93.55%, Support Vector Machine 86.3%,Logistic Regression 86.3% according to the results. Ensemble model technique more accurately finding the genetic disease prediction signals a paradigm shift in preventive disease management and individualized treatment by showcasing a pragmatic and competitive strategy. This discovery has the potential to revolutionize the healthcare sector by combining cutting-edge computational methods with extensive medical data, bringing in a new era of personalized care and proactive wellness.

Table of Contents

Table of Figures

Table of Tables

Chapter 1

1.1 Research Background: Genetic problems, caused by abnormalities in DNA or mutations in chromosomes, are becoming a more significant global health issue. These diseases, which are frequently handed down through the generations, have a significant influence on the health of individuals on all levels. Specifically, the growing global population and a lack of knowledge on genetic testing have led to a rise in the prevalence of these illnesses. Many genetic diseases have no effective treatment, which makes continuous observation necessary and emphasizes the significance of early detection. Predicting genetic disorders accurately is important yet difficult. Advances in machine learning in recent times provide fresh prospects for enhancing diagnostic precision. Machine learning algorithms have the potential to improve the prediction and classification of genetic diseases through the analysis of massive datasets obtained from patient medical records. Our objective is to create a sophisticated machine learning model that can forecast genetic diseases and the subtypes that make them up. Based on comprehensive patient medical histories, this model uses a variety of machine learning approaches, such as Random Forest (RF), Support Vector Machine (SVM), K-nearest Neighbors (KNN), and Decision Tree (DT),Logistic Regression (LR) , Gradient Boosting and Hybrid Model to determine the probability of a genetic condition.

1.2 Problem Statement: This is a massive problem around the world. Most accurately early disease prediction does not exist. To resolve this issue we have started our research.

1.3 Research Questions:

RQ1: How can precise medical histories be used to identify the existence of genetic deficiencies in patients using machine learning techniques?

RQ2: How does the impact of integrating multiple machine learning techniques on the accuracy based on patient data?

RQ3: How does the inclusion of advanced machine learning algorithms enhance the early diagnosis and treatment of genetic diseases in comparison to conventional diagnostic techniques?

1.4 Research Objective:

- Build a machine learning model which can help to predict the genetic disorder.
- Using multiple machine learning techniques to enhance the model accuracy.
- Demonstrate how, in comparison with typical approaches, modern machine learning algorithms enable earlier prediction and improved management of genetic diseases.

1.5 Introduction: The genetic disorders stem from mutations in the genome or changes in gene structure, as the genome plays a crucial role in carrying biological information that affects an organism's structure and functions. These diseases arise from abnormalities carried on by mutations in the DNA sequence that generates the genes. The genome's data, as the repository of vital information, is essential for examining genetic abnormalities that lead to a variety of illnesses. Within the specialty area of bioinformatics known as genomics, the structure and anomalies of genomes are studied.Genetic diseases divide into three categories: complex or multifactorial inheritance disorders, chromosomal or mitochondrial inheritance disorders, and single gene inheritance disorders. These divisions are predicated on the type and source of the genetic modifications. Each category illuminates the complex connection between the appearance of transmissible diseases and DNA structure.Genetics plays an increasingly important role in determining treatment outcomes and illness susceptibility in the complex system of human health. The genetic landscape of disease is broad and diverse, ranging from rare single gene inheritance syndromes to intricate multifactorial problems. The area has advanced because of recent developments in genetic technology, which have made it possible to collect enormous volumes of genetic data and have opened the door to previously unexplored insights into the inherited causes of disease.

Single gene inheritance disorders influence almost every area of physiological functioning and are characterized by mutations in a single gene. These disorders can present in a variety of clinical ways. Even though these conditions are uncommon on their own, they greatly increase the burden of disease worldwide when combined. Multifactorial genetic illnesses, on the other hand, pose a significant challenge to both researchers and physicians due to their intricate interaction between genetic and environmental components. Through the use of advanced analytical tools that can

deconstruct the complicated chain of genetic interactions, these disorders' intricate genetic signatures can be identified.

The distinctive maternal inheritance pattern of mitochondrial gene inheritance diseases presents significant diagnostic and interventional problems. Mutations in mitochondrial non nuclear DNA cause these illnesses, which frequently appear as multisystemic syndromes affecting high-energy organs. To improve patient outcomes and develop tailored therapeutic strategies, it is imperative to comprehend the multiple mechanisms operating mitochondrial disorders.

Polygenic diseases, also known as multifactorial genetic disorders, result from complex interactions between genetic abnormalities affecting many genes and different environmental and dietary factors. In contrast to single-gene illnesses, these conditions entail a network of genetic polymorphisms that collectively contribute to the disease phenotype, rather than having a clear inheritance pattern.Diabetes, Alzheimer's disease, and cancer are a few examples of multifactorial genetic illnesses. In many disorders, environmental triggers, dietary habits, exposure to chemicals, physical activity, and other lifestyle factors interact with genetic predispositions to impact the onset and course of the disease. Due to the intricate nature of these relationships, a thorough understanding of both genetic and non-genetic components is necessary for the prediction and treatment of multifactorial disease.

Amidst this intricacy, machine learning has become a potent instrument for extracting knowledge from big genomic datasets. More precise disease prediction and risk assessment are now possible thanks to the extraordinary effectiveness of machine learning algorithms, in particular, in identifying significant patterns from complicated biological data. Researchers hope to fully realize the promise of precision medicine and usher in a new era of individualized healthcare by utilizing the predictive powers of algorithms.Nevertheless, there are still a lot of obstacles to overcome in spite of machine learning's potential in genetic analysis. Innovative methods for data processing and interpretation are required due to the significant challenges posed by the volume and complexity of genetic data. Furthermore, the fact that polygenic risk scores and other existing approaches rely solely on oversimplified models of genetic inheritance limits their predictive power.

In recent years, deep learning and machine learning have been utilized effectively in a range of biological scenarios. Algorithms based on deep learning and machine learning can handle large data sets with significant noise, complexity, and imperfection with only

a few educated estimates about probability distributions and data production methods. While classic statistical methods take an inferential approach, the techniques used in machine learning and deep learning have as their main objective prediction. On the other hand, the capacity to forecast the danger of complex diseases has improved due to machine learning algorithms. The reason for this increase in prediction abilities is the capacity of machine learning algorithms to handle multi-dimensional guidance.

In the context of this dissertation, we set out to close the gap between state of the art genetic studies and clinical practice by utilizing machine learning to improve personalized medicine and disease prediction. Our goal is to create strong predictive models that can correctly identify people who are at risk of inheriting diseases by thoroughly examining genomic data and applying machine learning techniques. We aim to equip doctors with the necessary tools to provide patients throughout the world with more accurate and efficient care by utilizing the complementary capabilities of genetics and machine learning. Furthermore, studies show that a lack of knowledge about the significance of genetic testing has contributed significantly to the rise in the prevalence of genetic diseases along with population expansion. By using a machine learning model trained on medical data to predict genetic diseases and their subclasses, this study seeks to close this knowledge gap and advance the understanding and treatment of these conditions.

Chapter 2

2. Literature Review:

We extensively searched popular platforms including Google Scholar, IEEE Xplore, and Springer, as well as carefully reviewed prior research, in order to compile the most recent literature on the use of deep learning in disease prediction. We also looked into important web resources such as Cross Validated (https://stats.stackexchange.com/), Kaggle (http://www.kaggle.com/), and GitHub (http://github.com/) blogs, tutorials, and repositories. Many methods have been proposed in the field of genetic data-based disease prediction. Considering the large number of predictors, which are gene transcripts, the most common approach is to first identify a subset of these transcripts. This selection procedure could involve extracting a subset of representations, like the leading principle components, or identifying the top gene transcripts based on transcript-wise testing. After that, these particular transcripts or their representations are subjected to machine learning algorithms that forecast the disease condition.99.9% of the genetic material that makes up humans is shared by all individuals; our DNA is made up of roughly 6 billion molecular pairs of bases (A, T, C, and G). A mere one percent of our genetic makeup differs from person to person, which accounts for the distinctions that make each individual unique. There are at least 4 million unique genetic differences among individuals, and these variations known as genetic mutations account for them .The prediction of pathologies is a topic of continuing discussion among researchers. Some specialists argue that there are not enough genetic differences to reliably predict disease risk, or that most disorders are not exclusively genetic. For example, one cannot completely trace conditions like strokes and cardiovascular disorders to a single or numerous genetic abnormalities.The amount of genetic data being collected has increased dramatically due to improvements in typing and testing techniques. Still unclear, nevertheless, are the exact mechanisms by which genetic differences contribute to the development of disease. In spite of continuous endeavors to map chromosomal alleles and cancer variants, the majority of these variants continue to have inadequate genomic characterizations. Association studies were the mainstay of early efforts to find disease-associated genes without the use of experimental techniques. These investigations evaluate the likelihood of detecting particular genes in a living thing in relation to the expected random distribution.Strong phenotypic and comorbidity characteristics are

exhibited by disorders with neighboring genetic areas, as previous investigations have shown . It has been proposed that the fact that distinct disruptions within a single disease module frequently produce manifestations that are comparable makes genetic data especially valuable. Furthermore, there is a strong correlation between transcription factor networks and protein interactions in phenotypic networks, which are networks of genes that are nodes connected by edges representing associated phenotypic states. Furthermore, disorders that are far apart in the interactome typically result in different phenotypes . By combining these many kinds of data, several approaches have been created to forecast genes linked to disease. Combining the data into a single graph and using it for predictive modeling is one method of doing this. Genes linked to disorders can be found inside these module clusters; similar approaches have been used to predict disorder modules, which provide a similar problem. For example, Liu et al. [9] use expression network partitions and genetic data analysis to identify disease components, whereas Ghiassian et al. [10] use immediate neighbor analysis in nutrient interaction networks to constantly incorporate genes into categories. Genetic risk prediction has been demonstrated to have long-term effects on populations as well as individuals [11].

However, genetic risk prediction has only been possible with recent developments in high-density genotyping technologies. For instance, genes linked to cardiovascular disease may potentially influence intermediate outcomes such as smoking, hypertension, or dyslipidemia [12]. Genetic differences included in intermediate variables may become less significant when these factors are included in a dependent variable, according to basic principles of scientific research. However, genetic variants that contribute to yet-to-be-identified pathways or processes with unquantifiable intermediary components can improve illness prediction above and beyond known risk factors. It is more likely for some diseases than others to have novel, unidentified pathways. An important consideration is that the identification of new genes may lead to the discovery of intermediate biomarkers and novel etiological networks, which may prove to be more accurate disease predictors than the genetic variants that initially produced them.The massive availability of functional genomic data, fueled by cutting-edge technologies that offer enormous volumes of behavioral information among biological constituents, has made it easier to construct network-based strategies and improved application methods, such as protein structure analysis. Many important biomedical problems have recently seen the successful application of machine learning methods. Genetic code interpretation, genetic analysis classification, gene regulatory network inference , therapeutic target

prediction, identification of epigenetic relationships in disease statistics , and pharmacological advances are a few of these. Genes linked to disease have also been predicted using machine learning. In order to anticipate novel genetic diseases, this issue is usually treated as a classification problem, where models are constructed utilizing biological data, medical history information, and recognized genetic abnormalities. As a result, more useful methods have been developed. Notably, it has been suggested to use only positive data to train unary classifiers .The combination of superior machine learning methodologies with extensive genomic data exhibits considerable potential in augmenting our aptitude to anticipate genes linked to illnesses. These developments increase our knowledge of hereditary illnesses and open the door to personalized medicine, in which a patient's unique genetic profile can guide customized preventative, diagnostic, and therapeutic methods.

We found seven new papers that use effective methods to predict genetic deficiencies; they were published between 2020 and 2023. These researchers analyze genomic data using machine learning approaches in an effort to increase the precision of risk assessment and disease prediction.

Table 2.1: Literature Review of the previous studies

Sr No.	Author	Year	Model	Accuracy	Limitation
1	Duc Le et al	2020	NB,SVM, ANN	63%,68%, 71%	Does not discuss the scalability and computational requirements.
2	Xiaomei *et al*	2021	DT,SVM, RF,KNN	78.1%,81.4 %,78.6%, 76.1%	High throughput sequencing data.
3	Ali, Furqan *et al*.	2022	ETRF+ XGB	92%	Less number of Model used in this paper.
4	Sadichchha *et al*.	2022	KNN, CATBOOST, RF, XGBOOST	79.71%, 75.36%, 89.4%, 86.1%	Limited discussion on the validation and evaluation of the model's performance.

5	Nasir *et al*.	2022	ANN,KNN, SVM	85.7%,84.9 %,84.3%	Lack of specific details about the medical parameters.
6	Taher *et al*.	2022	SVM , KNN	92.8%, 92.5%	Does not compare the proposed model with other existing models.
7	Venkat et al.	2023	RF	HF-90.9%, AF-95%, CVD-95.9%	Model's accuracy in capturing heart-specific molecular events.

Previous studies have shown that the majority of research on genetic diseases is based on genome sequencing. On the other hand, it has been determined that genome sequencing has certain important limits.

- Genome regions may be processed below the DNA sequence's minimal coverage level during the prediction procedure. It is possible that the base calls won't be able to pinpoint a person's true genetic condition if there is not enough depth of coverage.
- Limited discussion of the model validation and evaluation.
- Some papers do not show the scalability and computational problem.
- One major obstacle still to be overcome is ensuring the clinical validity of the predictions.

To improve the genetic disorder prediction accuracy, we overcome these shortcomings in our suggested model. To enhance the therapeutic relevance and predictability of the results, we plan to include medical histories of the patients. By merging clinical knowledge with genetic data, this integration enables a more thorough method that yields a more precise and individualized prognosis of genetic disorders.

Chapter 3

3. Methodology:

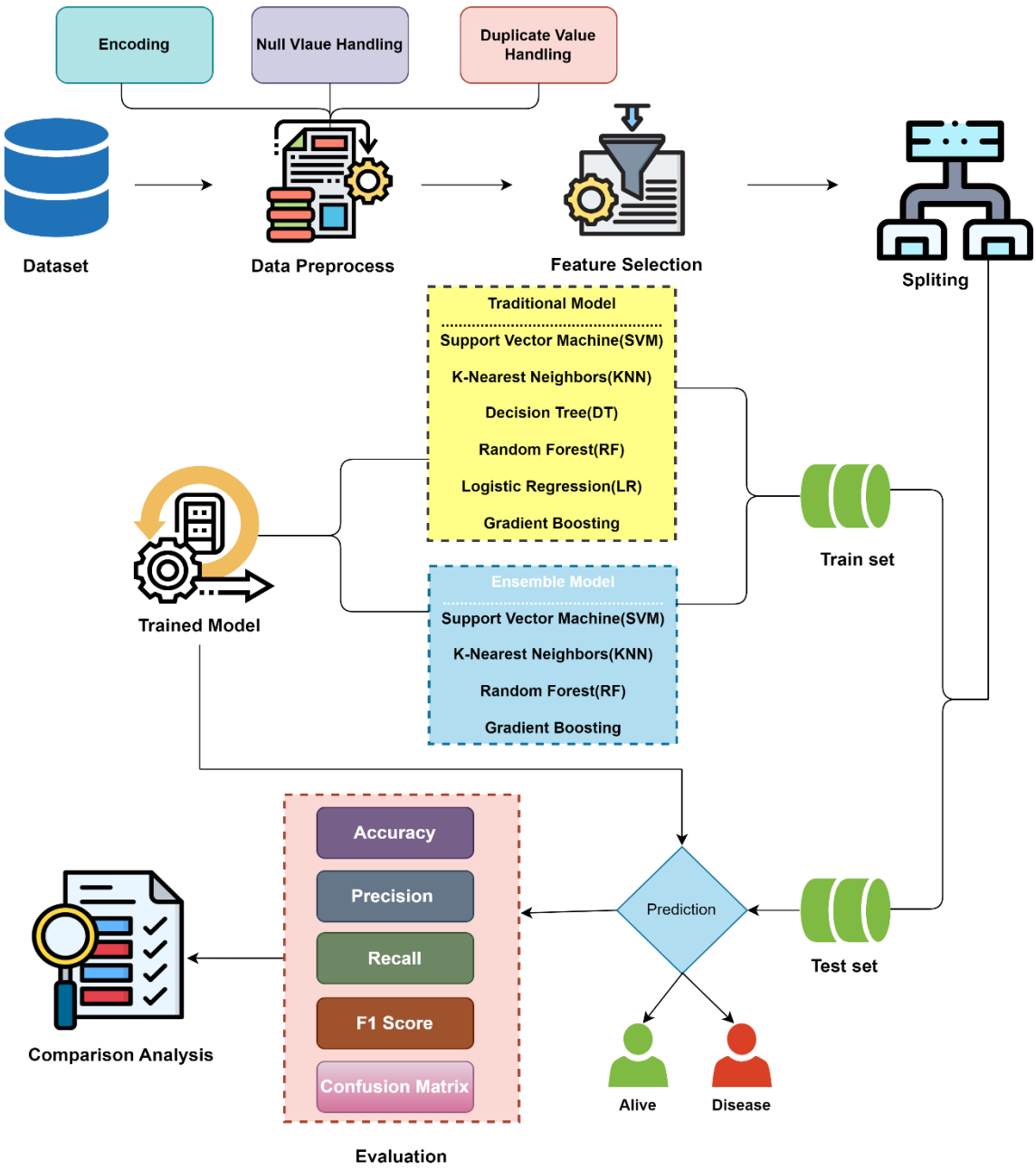

Figure 3.1: Proposed methodological analysis to predict the genetic disorder.

3.1 Dataset: The dataset is used in multiple numbers of papers like Raza et al. (2022) , Ghazal et al. (2022) , Rahman et al. (2022) , Nasir et al. (2022) , Naik et al. (2022) . So, we collected the dataset from these papers and applied our studies.The study applied a dataset that was obtained from paper. It included 22,083 patient records and 38 features

that are pertinent to the prediction of genetic diseases. Different data normalization approaches were used throughout the data preprocessing stage to address and replace any missing values, guaranteeing the consistency and integrity of the dataset for further analysis.At an initial sight, many of the dataset's factors seem non-statistical and ignored. But by using machine learning techniques well, it is feasible to find the important factors and get rid of the ones which aren't required.

3.2 Data Analysis: A number of parameters are included in the dataset: patient ID, name, blood cell count, respiratory rate, heart rate, folic acid levels, WBC count, any symptoms,history of anomalies etc. The goal variables for model training in this structured dataset are the disorder or not. Figure 2 gives an idea for the disorder people who are not affected by these chromosomal diseases.

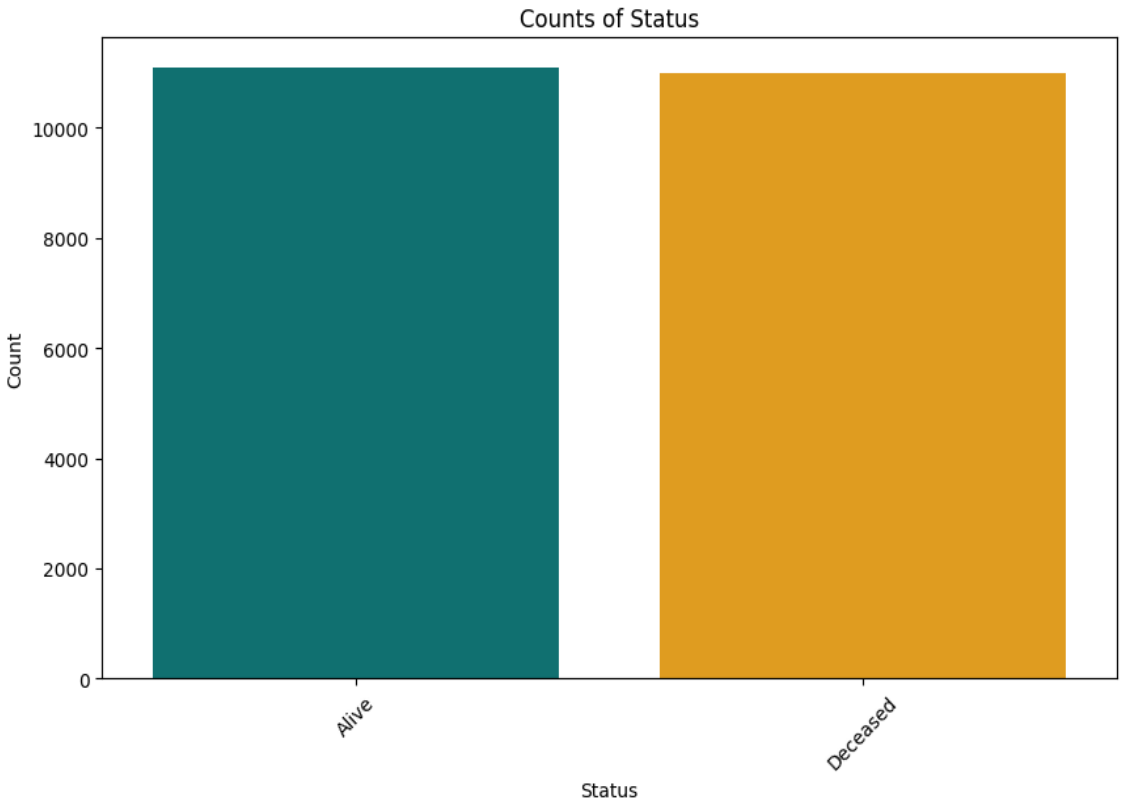

Figure 3.2: Status for Genetic Disorder

Genetic diseases are divided into three categories in this dataset: single gene inheritance disorders, multifactorial genetic inheritance disorders, and mitochondrial genetic inheritance disorders. When compared to the other two categories, mitochondrial genetic inheritance problems are more common among them. This prevalence points to a strong

focus on mitochondrial abnormalities in the dataset, which could have an impact on the study's methodology and results. For the purpose of creating focused machine learning models that can reliably forecast and distinguish between the various genetic inheritance patterns, it is essential to comprehend the distribution of these disease kinds. This realization also emphasizes the necessity of using specific methods to treat the distinctive traits and intricate issues connected to mitochondrial diseases. Figure 3 gives us a statistical view of these types.

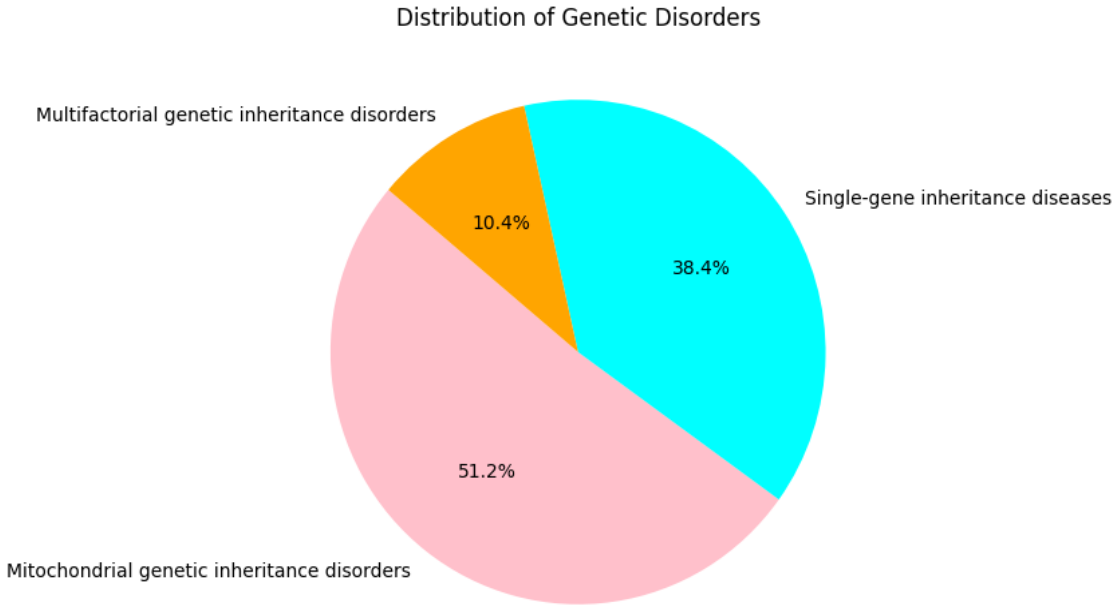

Figure 3.3: Genetic Disorder Class.

There are nine different classes within the subclass category that correspond to different genetic disorders: hemochromatosis, mitochondrial myopathy, cancer, cystic fibrosis, diabetes, Leigh syndrome, hemochromatosis, hemochromatosis, and Leber's hereditary optic neuropathy. Interestingly, the dataset's distribution values for Leber's hereditary optic neuropathy and diabetes are the lowest, suggesting a dearth of data samples for these specific conditions. In a similar vein, Tay-Sachs likewise presents a comparatively small sample size.Because there are insufficient samples for some illnesses, the model may find it difficult to learn and generalize patterns as a result of this imbalance in the distribution of data among subclasses. This could present problems for both model evaluation and training. To guarantee the model's resilience and precision in forecasting illnesses with lower representation, it becomes imperative to rectify this imbalance using methods like class weighting, stratified sampling, or data augmentation. Insights into

illness prevalence, diagnostic rates, and data collecting biases may also be found by analyzing the causes of sample scarcity for certain disorders.

This knowledge will help to guide future data collection initiatives and research objectives in genetic disorders.Figure 4 contains a graphical view which shows about the subclasses name and which are the highest and lowest points.

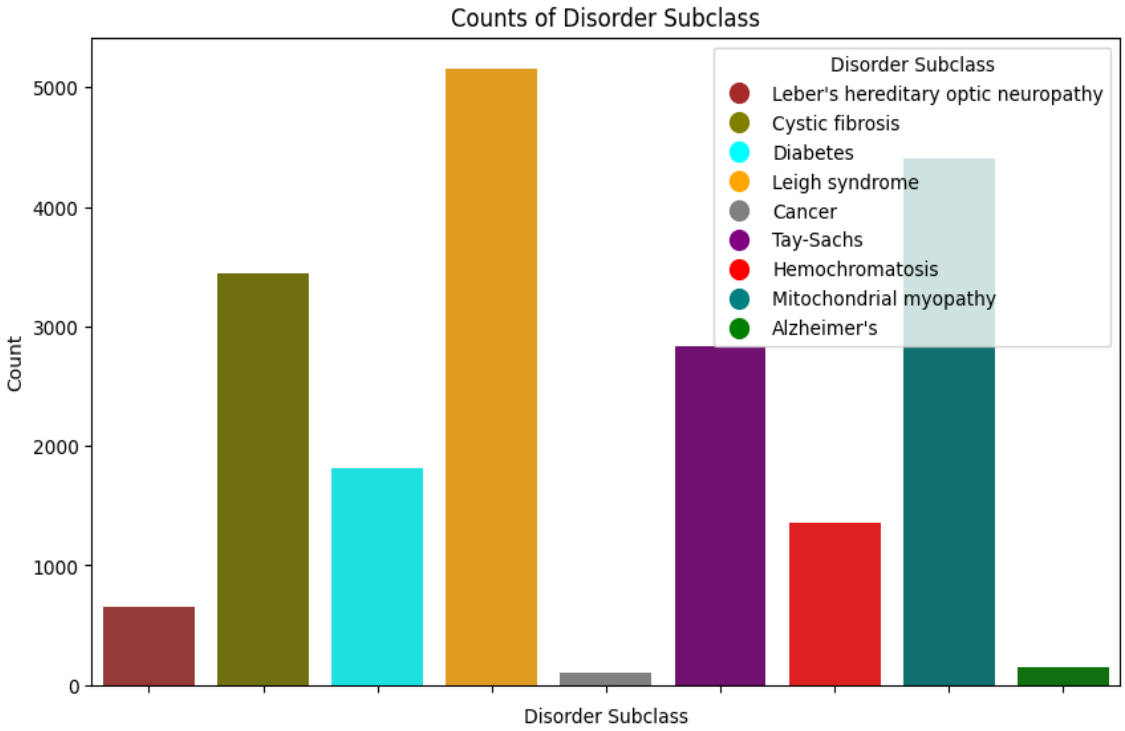

Figure 3.4: Different types of Diseases.

3.3 Feature Engineering:

3.3.1 Feature Selection

Optimizing a model's performance requires knowing which feature selection strategy is best for it. It's critical to first define the sorts of input and output variables involved in order to choose the optimal approach. We mostly used Filter feature selection methods in our work:

Filter Method: This method, which is independent of the learning algorithm, uses statistical metrics to choose features during the preprocessing step. The filter approach streamlines the model by eliminating duplicate and unnecessary features through the evaluation of many metrics. The effectiveness of filter approaches in terms of processing time and their ability to withstand overfitting are two important advantages. In particular, we use the missing value ratio to compare the feature set to a threshold value. The number

of missing values in each column divided by the total number of observations yields the missing value ratio. Features that are deemed unreliable and are removed from the model are those that are over the threshold ratio. This procedure makes sure that the dataset is stable and devoid of a lot of noise, which could harm the model's ability to function.We make sure that our model is effective and efficient by concentrating on these feature selection techniques, using just the most pertinent data for prediction employment.

3.3.2 Outlier Handling

The Interquartile Range (IQR) approach is a reliable technique that's often used to remove outliers from datasets. Data points known as outliers are those that differ noticeably from the bulk of observations. These data points have the ability to distort analytical results and produce false conclusions. By concentrating on the middle section of the data distribution, the IQR approach aids in the identification of these anomalies.

IQR Method Utilized for Outlier Removal Process:

- Determining Numerical Column Identification:The first step in the process is to determine which numerical columns in the dataset need to be checked for outliers. Usually, these are columns with continuous data points that may show values that are extreme.
- Quartile Calculation:Now compute the first quartile (Q1), or the 25th percentile, and the third quartile (Q3), or the 75th percentile, for each numerical column. The dispersion and central tendency of the data can be better understood with the aid of these quartiles.
- The process of Determining the Interquartile Range (IQR) to calculate the IQR, deduct Q1 from Q3 (IQR = Q3 - Q1). This range captures the core region where the majority of data points are located, measuring the middle 50% of the data.
- Determining Boundaries for Outliers: The IQR is used to characterize outliers. To be more precise, an outlier is any data point that is above the upper bound (Q3 + 1.5 * IQR) or below the lower bound (Q1 - 1.5 * IQR). The purpose of these thresholds is to filter away extreme values while retaining the majority of the data points.
- Selecting Outliers: Information that deviates from the computed boundaries is considered an outlier, and it is then eliminated from the dataset. This process assists in removing anomalies from the dataset that can skew the study.

- In our work datasets are free of extreme values that could distort the results by employing the IQR approach, which will produce more accurate and dependable results in data analysis and machine learning applications.

3.3.3 Null Value Handling

In our large number of dataset lot's of missing values which are affecting our data to ready for the training model. The numerical features are missing value handling using mean all of this specific feature and replace the null value . For categorical values null value replacing the mode of this feature and removing all null values.

3.3.4 Feature Encoding

In the field of machine learning, datasets often have many labels dispersed over several columns. These labels can be represented by numbers or words, with words being the most readable format for people. However, a conversion process is necessary in order for machines to effectively understand and use this data. Label encoding is the process of converting word-based labels into numerical values. Machine learning algorithms can handle data more effectively and enable better analysis and decision-making by transforming labels into a numeric representation. In order for computers to identify patterns, relationships, and insights that would be less obvious in a textual format, this mathematical transformation is essential. As a result, label encoding is a crucial stage in getting data ready for efficient machine learning model training and usage.

3.3.5 Standardization

Standardizing datasets is an essential process for numerous machine learning techniques. These models often assume that the characteristics follow a conventional normal distribution, which is characterized by a Gaussian distribution with a mean of zero and a variance of one. Without this uniformity, the models may perform weakly or generate incorrect results.The StandardScaler utility class provides a simple and effective approach for normalizing features in a dataset that resembles an array. This utility operates by calculating the difference between the mean and each characteristic, then thereafter adjusting the features to get a standard deviation of one. This procedure puts the data into a form where each feature contributes equally to the model's performance.Standardization is crucial as it guarantees that every characteristic is measured on a comparable scale, preventing features with wider ranges from

overpowering those with narrower ranges. Equalization is crucial for algorithms that depend on distance measurements, such as k-nearest neighbors (KNN) and support vector machines (SVM).Implementing the StandardScaler has consistently provided favorable results across many datasets and machine learning workloads. By standardizing the features, the performance of the estimators is boosted, resulting in more accurate and dependable predictions. This preprocessing step is consequently a basic technique in the preparation of data for machine learning applications.

3.4 Trained Model

This method uses one classifier to predict the specific genetic disorder.An 80:20 train-test split has been used to train these models. To assure optimal performance, classifiers with a proven track record of managing data with a high number of features and class imbalance have been selected based on the input parameters. The main models that are emphasized for this intent are:

3.4.1 Support Vector Machine: The preceding concepts are represented by Support Vector Machines (SVMs). We must examine two areas in order to comprehend this: the hypothesis spaces that SVMs use and the loss functions that they apply. SVMs are traditionally thought to solve the learning problem by locating the "optimal" hyperplane. The linear SVM model, in which the hyperplane is defined in the space of the input data x, is the most basic type of SVM model. The hypothesis space in this case is made up of hyperplanes that look like this:

$$f(x) = w.x + b.$$

In their most basic form, Support Vector Machines (SVM) find a hyperplane in a transformed space called the feature space that is specified by a kernel K rather than in the original input space x. In this feature space, the kernel K establishes a dot product. "Hyperplanes" within this induced feature space make up the hypothesis space for the Support Vector Machine (SVM) when using the kernel K. Another way to think of these hyperplanes is as a collection of functions inside the Reproducing Kernel Hilbert Space (RKHS), which is given by K. The pertinent literature is recommended reading for readers seeking a comprehensive discussion on RKHS.

$$\{ f : || f || \ K_k^2 < \}$$

where $||f||_k^2$ is the function's RKHS norm and K is the portion of the kernel that determines the RKHS. The RKHS norm of the functions of the sort f(x) = w.x + b, for instance, is just the norm of w, that is,$||f||_k^2$ = ||w||^2. This is the case for the linear case discussed above, where K is the kernel of functions K(x1, x2) = x1.x2.
SVM really takes into account subsets of this domain, namely collections of the type:

$$\{ f : || f || K_k^2 \quad A \quad ^2 \}$$

The hypothesis spaces in the Statistical Learning Theory (SLT) framework are structured by a constant called A, the value of which increases with the complexity of the hypothesis space. Finding the solution with the "optimal" RKHS norm, that is, determining the ideal value for A is the aim of the Support Vector Machine (SVM).

In summary, Support vector machines (SVM) are learning models that strive to minimize empirical error while also taking the "complexity" of the hypothesis space into account, in accordance with the concepts of Statistical Learning Theory (SLT) as explained. The Reproducing Kernel Hilbert Space (RKHS) norm of the solution, represented as $||f||_k^2$, is minimized in order to control this complexity. SVMs accomplish this in practice by striking a balance between the complexity of the hypothesis space and empirical error. Formally, this balance can be attained by resolving the following minimization issues:

SVM classification $\quad min\ ||f||_k^2 \ + C_{i=1}^{l} \ | \ 1 \ yf(x)|_+$

SVM Regression $\quad min\ ||f||_k^2 \ + C_{i=1}^{l} \ | \ y_i \ - \ f(x_i \ )|$

3.4.2 K-Nearest Neighbors: The widely used classification method K-nearest neighbors (KNN) is renowned for its ease of use and computational effectiveness. It is preferred over more sophisticated algorithms due to its simplicity of interpretation and very short computation time. The selection of the parameter k, which establishes the quantity of nearest neighbors utilized in the classification procedure, is a crucial component of KNN.Accurate classification results in this approach depend on figuring

out what the ideal value of k should be. The usual method for doing this is to evaluate the validation and training error rates for various k values. Through the assessment of these error rates, practitioners can determine the optimal k value to balance variance and bias, resulting in a model that performs well when applied to previously unseen data.

Distance Calculation:

$$\mathrm{d}(\mathrm{x}, \mathrm{x}_i) = \| \mathrm{x} - \mathrm{x}_i \|$$

Selecting **k** Nearest Neighbors**:**

$$\{(\mathrm{x}_{(1)}, y_{(1)}), \ldots, (\mathrm{x}_{(k)}, y_{(k)})\}$$

Weight Calculation (if weighted KNN)**:**

$$\mathrm{w} = \frac{1}{\mathrm{d}(\mathrm{x},\mathrm{x}_i)+\epsilon}$$

where w_i is the weight of the i-th nearest neighbor and ϵ is a small constant to avoid division by zero.

Classification Decision:

$$\mathrm{y} = arg \max_{\mathrm{c}\in\mathrm{C}} \sum_{i,y_i} w_i$$

where c is the set of all possible classes.

Parameter Selection**:**

$$k_{\text{optimal}} = \arg min_k(\text{Validation Error(k)} + \text{Training Error(k)})$$

Text categorization using the KNN method has a wide range of applications. The KNN methodology is a simple and efficient method for text classification tasks like sentiment analysis, document categorization, and spam detection. KNN can handle text data efficiently and deliver dependable classification results when feature representations and suitable preparation methods are used.In conclusion, K-nearest neighbors is a flexible approach for classification that is appreciated for its effectiveness and simplicity. The model's performance is greatly influenced by the choice of the k parameter, and the ideal value is determined by carefully analyzing the training and validation error rates. Because

KNN is flexible and efficient with text, it is widely used in the text categorization field.For the standard K-nearest neighbors (KNN) algorithm, a number of modifications and improvements have been suggested with the goal of increasing its efficacy and efficiency, especially when it comes to text classification.A flexible KNN algorithm that combines a weighting scheme and the K-variable method is one modification. It was discovered that this combination greatly increased text classification efficiency. This method may better adapt to the properties of the data by dynamically varying the value of k and adding weighted distances, which improves classification performance [18].A method that is more accurate and efficient for classification tasks is produced by combining eager learning with KNN classification in another revision. This method improves the accuracy of predictions and expedites the classification process by building a model and preprocessing the data [16].

Furthermore, a unique KNN classification algorithm that combines aspects of evidence-based and model-based theories has been presented. The time-consuming aspect of traditional lazy learning in KNN is addressed by this hybrid approach. This approach delivers enhanced efficiency while preserving the robustness and accuracy of the classification findings by utilizing both model-based and evidence-based reasoning [17].In a number of changes have been made to the KNN algorithm to improve its effectiveness and precision in text categorization applications. These modifications make use of strategies including pre-processing, hybrid reasoning approaches, and dynamic parameter change to solve particular problems and enhance overall performance.

3.4.3 Decision Tree: A decision tree is a fundamental tool in decision-making processes because it provides an organized depiction of different options and their possible outcomes. A decision tree, like a comprehensive decision assistance tool, captures the various aspects surrounding a scenario through its branching structure and leaf nodes. It maps out decisions and their potential outcomes inside its branches, taking into account utility values, event probability, resource costs, and other relevant variables. A decision tree is essentially an algorithm's visual representation, which makes it easier to comprehend convoluted decision-making processes.

Different kinds of decision trees can be used, depending on the particular circumstances and intended goals:

- Classification Tree: his type of decision tree predicts categorical outcomes based on input data. It uses estimated probabilities to guide decision-makers towards the most predictable outcomes.

$$\text{Output} = \arg\max_{c \in C} P(c \mid x)$$

where P(c|x) is the conditional probability of class ccc given input x.

- Regression Tree: Useful for predicting continuous outcomes from various data sources, often used in fields like real estate for future value predictions.

$$output = \sum_{i=1}^{1} \frac{1}{N}$$

where y_i are the observed values and N is the number of observations in a terminal node.

- Decision Tree Forests: Decision tree forests offer a strong framework for precisely predicting particular outcomes by combining several decision trees. This method reduces individual tree biases and increases forecast accuracy by utilizing the combined wisdom of several decision trees.

$$output = \frac{1}{T}\sum_{t=1}^{T} Tree(x)$$

where T is the number of trees and Tree(x) is the prediction from the t-th tree.

- Classification and Regression Tree (CART): By taking dependent factors into account, CART algorithms enable logical assumptions about outcomes. Through the integration of classification and regression methodologies, CART models provide a holistic approach to forecasting, including a wide range of industries from banking to healthcare. Integrates both classification and regression methodologies.

$$Output = \arg\max_{c \in C} P(c \mid x) \qquad for\ classification$$

$$Output = \sum_{i=1}^{1} \frac{1}{N} \qquad for\ regression$$

Essentially, decision trees provide decision-makers with a flexible toolkit that allows them to negotiate complex decision landscapes and reach well-informed decisions in a variety of fields. Whether used for probabilistic prediction, trend estimation, or

dependency analysis, decision trees are an indispensable tool in the toolbox of decision support techniques.

3.4.4 Random Forest: A number of tree-based feature selection strategies are essential for determining the importance of features and helping to choose which ones to use when building models. These methods demonstrate which features have the greatest influence on the target variable and provide insightful information about the value of each feature.Of them, Random Forest is a well-known tree-based approach that effectively uses the potential of ensemble learning. Through the combination of many decision trees, Random Forest creates a kind of bagging algorithm by utilizing the combined knowledge of various trees.By rating each node throughout all trees according to its performance or the impurity reduction typically gauged using measures like Gini impurity Random Forest determines how important a feature is. The impurity I of a node can be measured as:

$$I = 1 - \sum_{i=1}^{c} {p_i}^2$$

Where p_i is the proportion of samples belonging to class i at the node. Tree cutting below specific nodes is made easier by the systematic ordering of nodes based on their impurity levels. A subset of the most important features is produced as a result of this procedure, and these features are nonetheless crucial to the predictive ability of the model.

Using a set of independent and different decision trees created using randomization principles, Random Forest is a powerful ensemble classifier. It is essentially the process of building a large number of randomized decision trees, each with a different set of random parameters. This formulation of Random Forest can be represented as:

$$\{h(x, \theta_k)\}_{k=1}^{L}$$

where x is the input data and θ_k is a set of mutually independent random vectors.

The concept of randomization in feature selection and data sampling is central to the Random Forest approach. A random subset of the sample dataset is picked for training, and a random subset of features is chosen from the pool of features accessible for each decision tree. The resilience and capacity for generalization of the entire ensemble are enhanced by this approach, which guarantees diversity among the individual trees.

Several crucial steps are included in the Random Forest creation algorithm:

1. Bootstrap Sampling: k different training sets are formed by randomly selecting samples from the training dataset with replacement. Each set D_k is used as the foundation for building a decision tree. The data that are out-of-bag (OOB) are samples that are not chosen during the process.

$$D_k \sim D$$

2. Feature Randomization: A random subset of m features is selected at each decision tree node. To choose the best characteristic for node splitting, the algorithm uses these features, which adds unpredictability to the decision-making process.

$$F_m \subseteq F$$

3. Unrestricted Growth: To ensure maximum feature space exploration and minimize bias, each decision tree in the Random Forest ensemble is allowed to grow to its full potential without pruning.

4. Ensemble Formation: The Random Forest model is created by combining the separate decision trees. The collective wisdom of the constituent trees is then employed to classify unknown data using this ensemble model.

$$\hat{y} = arg\max_{c} \sum_{k=1}^{L} h_k(x) = c$$

To put it simply, Random Forest creates a diversified ensemble of decision trees by utilizing randomness in both feature selection and data sampling. Combining these trees' predictions allows Random Forest to achieve strong classification performance that can handle complicated datasets and reduce overfitting inclinations. Random Forest offers a useful way to discover and rank the most important features in addition to providing a strong foundation for feature selection. Data scientists and practitioners may make more informed judgments when developing and analyzing models thanks to Random Forest's capacity to harness the combined insights of numerous decision trees, which improves the reliability and interpretability of feature importance assessments.

3.4.5 Logistic Regression: One powerful technique that is mainly used for classification tasks is logistic regression. Logistic regression works with data structured in groups or clusters, each of which represents a unique category, as opposed to linear regression, which works with data points ordered linearly. Think of these clusters as

dispersed piles, where each pile represents a specific class and all of the data points in a pile have the same category label. The key to logistic regression is defining the categorization boundary, which is represented by the regression formula. By use of an optimization procedure, the training classifier attempts to identify the best regression coefficients in this equation, so precisely defining the classification border. [3] [4].

$$Logit(p) = \ln \frac{p}{1-p} = \beta_0 + \beta_0 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

When using logistic regression for classification, the algorithm analyzes a wide range of inputs and uses a predetermined function to produce matching outputs. This tool streamlines the intricate decision-making process by effectively classifying the supplied data. In situations involving binary classification, the function often returns 0 or 1, which stand for the two different classes.

$$P(y = 1 \mid x) = \sigma(\beta_0 + \beta_0 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

where σ(z) is the sigmoid function:

$$\sigma(z) = \frac{1}{1 - e^{-z}}$$

However, the function has to follow several restrictions to guarantee mathematical tractability and usefulness. More specifically, the dependent variable is limited to either 0 or 1, and its argument range extends from positive infinite to negative infinity. These requirements are met by a multitude of functions, the simplest but intrinsically most restrictive of which is the 0-1 step function:

$$f(x) = 1 \quad \text{if } x \geq 0$$
$$f(x) = 1 \quad \text{if } x < 0$$

The step function is simple, but it is not smooth at the transition point, which makes it difficult to manipulate and analyze mathematically. Therefore, even if logistic regression offers a flexible framework for classification problems, selecting the function that controls its operation needs careful thought, striking a balance between mathematical flexibility and simplicity to guarantee good performance in a variety of scenarios.

3.4.6 Gradient Boosting: Among the tree-based ensemble approaches, boosting is one of the most efficient strategies since it can store the labels and weights of the leaf nodes, making prediction interpretation easier. Gradient Boosting,

introduced by Chen et al. [25], has become a popular and useful method among different boosting algorithms .Gradient Boosting is based on the principle of combining weak learners iteratively to build a robust, strong learner. Every weak learner focuses on areas where earlier iterations have failed to identify the underlying patterns in the data, progressively adding to the model's overall prediction power. Decisions on classification are based on the residuals or errors from previous iterations during the Gradient Boosting process. Let $F_{m(x)}$) denote the model prediction at iteration m. The residuals r_i^m at iteration m are computed as:

$$r_i^m = y_{i- F_{m-1}(x_i)}$$

where y_i is the true label and x_i is the input data.

Through a series of evaluations of each feature's effect on these residuals, the algorithm progressively improves its predictions until it reaches the desired degree of accuracy.

$$F_m(x) = F_{m-1}(x) + \gamma h(x; \theta m)$$

where γ is the learning rate, controlling the contribution of each weak learner to the overall model.

The process of iterative refinement guarantees that the model gains proficiency in identifying intricate associations present in the data, culminating in the production of exceptionally precise forecasts.

3.5 Ensemble Model: Ensemble methods have become a potent tool in machine learning, utilizing the aggregate intelligence of numerous models to improve accuracy and adaptability in the quest for greater prediction performance. This work uses an ensemble model that combines approaches from Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting. Each of these techniques adds special qualities to the group and helps create a wide range of prediction skills. KNN relies on the closeness of data points to make decisions, making it an easy-to-use method for classification tasks due to its simplicity and effectiveness. On the other hand, SVM is skilled at managing both linear and nonlinear interactions within the data since it makes use of the idea of locating an ideal hyperplane in a high-dimensional environment. By creating many decision trees and mixing their predictions, Random Forest, a tree-based ensemble technique, performs exceptionally well in feature selection and classification. Gradient Boosting, on the other hand,

iteratively strengthens model predictions by concentrating on areas where earlier iterations failed, progressively improving the model's comprehension of intricate data patterns. Through the use of a voting classifier mechanism, the ensemble model integrates several disparate methodologies, leveraging the combined insights of each model to produce a predictive framework that is more resilient and precise. By integrating the complementing characteristics of several methodologies, this holistic approach not only improves overall performance but also lessens the limitations of any one model. This study attempts to obtain improved prediction performance by integrating SVM, Random Forest, Gradient Boosting, and KNN in an ensemble framework, enhancing machine learning's capacity to address real-world problems.

3.6 Evaluation Parameter: Different types of evaluation metrics are used for this assessment of genetic disorder prediction models. Multiple evaluation parameters, such as the True Positive (TP) rate, False Positive (FP) rate, precision, recall, F-measure, and the area under the Receiver Operating Characteristic (ROC) curve, are used to evaluate the performance of different computational models.

- True Positive Rate (TPR): This metric assesses the extent to which an algorithm detects people who are legitimately suffering with the disease.
- False Positive Rate (FPR): This is the capacity of any algorithm to accurately forecast individuals who are not afflicted with the disease.
- Precision: Precision is defined as the proportion of appropriately recognized pertinent elements to all elements that the algorithm successfully retrieves.
- Recall: The ratio of relevant elements to the total of both relevant and non-relevant features is indicated by the fraction.
- ROC Area: A visual aid to evaluate classifier performance, the ROC curve depicts the connection between the true positive rate and false positive rate across various variable thresholds.

In this case, TP stands for the number of genes properly recognized as disease genes, FN for the number of genes incorrectly identified as non-disease genes, TN for the number of genes correctly identified as non-disease genes, and FP for the number of genes incorrectly identified as disease genes.

Table 3.1: Evaluation Criteria Explanation

Parameter Type	Equation
True Positive Rate	$$TRP = \frac{TP}{TP + FN}$$
False Positive Rate	$$FRP = \frac{FP}{FP + TN}$$
Precision	$$precision = \frac{TP}{TP + FP}$$
Recall	$$Recall = \frac{TP}{TP + FN}$$
F Score	$$F\ measure = 2 \times \frac{REC \times PREC}{PREC + REC}$$ $$F_{0.5} = 1.25 \times PREC \times \frac{REC}{0.25 \times PREC + REC}$$ $$F_1 = 2 \times PREC \times \frac{REC}{PREC + REC}$$ $$F_2 = 5 \times PREC \times \frac{REC}{4 \times PREC + REC}$$

In the above table short form of TP,FN, TN,FP,PREC and REC denoted respectively in the below:

- TP = True Positive
- FN = False Negative
- TN = True Negative
- FP = False Positive
- PREC = Precision
- REC = Recall
- TRP = True Positive Rate
- FRP = False Positive Rate

Chapter 4

4 Results and Discussion:

Multiple machine learning models are applied in this work.The suggested model was trained and tested using a variety of machine learning techniques. A number of criteria, including F1 score, precision, recall, accuracy of classification were employed to assess these algorithms. Preprocessing the data, which includes replacing missing values and dividing the data into training and testing stages, is the first step in the proposed model.The outcomes of the proposed model are explained in more detail below, with a focus on different prediction parameters. The confusion matrices for training and testing across several machine learning methods are first presented in the first phase. The second phase then offers a comparison analysis regarding their parameters.

Table 4.1 : Different model evaluation

Model	Accuracy	Precision	Recall	F1 Score
KNN	97.54%	0.9764	0.9764	0.9764
RF	96.72%	0.9692	0.9672	0.9671
GB	93.55%	0.9428	0.9355	0.9352
LR	86.03%	0.8906	0.8603	0.8573
SVM	86.03%	1.0000	0.7178	0.8357
DT	91.09%	0.9243	0.9109	0.9101

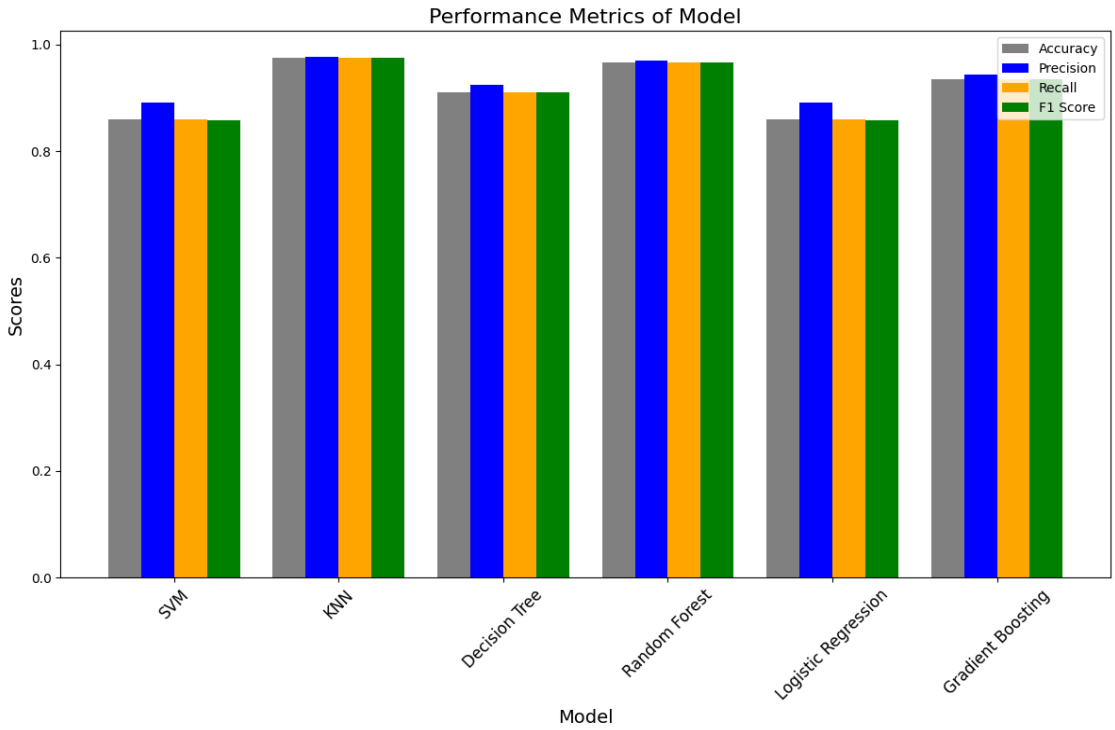

Figure 4.1: Evaluation of the all traditional model.

4.1 KNN, or K-Nearest Neighbors

With an astounding accuracy of 97.54%, the K-Nearest Neighbors (KNN) algorithm demonstrated the highest overall performance. It obtained recall, precision, and an almost flawless F1 score, all at roughly 0.9764. These consistent measures demonstrate how well KNN predicts accurate results while upholding a fair trade-off between recall and precision. KNN's confusion matrix demonstrated 2214 true positives, 2066 true negatives, 2 false positives, and 106 false negatives, highlighting the system's resilience in correctly classifying and eliminating situations.

4.2 Random Forest (RF)

With an accuracy of 96.72%, the Random Forest (RF) algorithm also showed good performance. Its scores were 0.9692, 0.9671, and 0.9672 for precision, recall, and F1 score, respectively. Based on these parameters, it can be concluded that RF has a slightly lower level of reliability than KNN. 2216 true positives, 2028 true negatives, 0 false positives, and 144 false negatives were reported by the RF model's confusion matrix. Because RF uses an ensemble technique, which efficiently integrates several decision trees to boost performance, it can be said to have excellent accuracy and generalizability.

4.3 Boosting Gradient (GB)

Gradient Boosting performed admirably, achieving 93.55% accuracy. Its precision was 0.9428, recall was 0.9355, and F1 score was 0.9352, indicating that it could correctly categorize data. Still, it was not as good as RF and KNN. The confusion matrix for the GB model showed 2216 true positives, 1889 true negatives, 0 false positives, and 283 false negatives, indicating that although GB is a good model, its dependability might not be as good as that of the best models in this comparison.

4.4 (LR) Logistic Regression

With a recall of 0.8603, precision of 0.8906, and F1 score of 0.8573, Logistic Regression (LR) yielded an accuracy of 86.03%. These measures are less than those of the other models, suggesting that although LR is faster and easier to use, it may not adequately represent the intricacy of the data. LR's limitations in precision and recall were evident in the confusion matrix, which had 2216 true positives, 1559 true negatives, 0 false positives, and 613 false negatives.

4.5 SVM or support vector machine

Additionally, the SVM model's accuracy was 86.03%. Its F1 score was 0.8357 and recall was 0.7178, demonstrating a considerable mismatch despite its superb precision of 1.0000. This suggests that although SVM is quite accurate at predicting positive cases, it is not able to identify every positive instance, which may provide issues in situations where a high recall rate is essential. 2216 true positives, 1559 true negatives, 0 false positives, and 613 false negatives were displayed in the SVM confusion matrix.

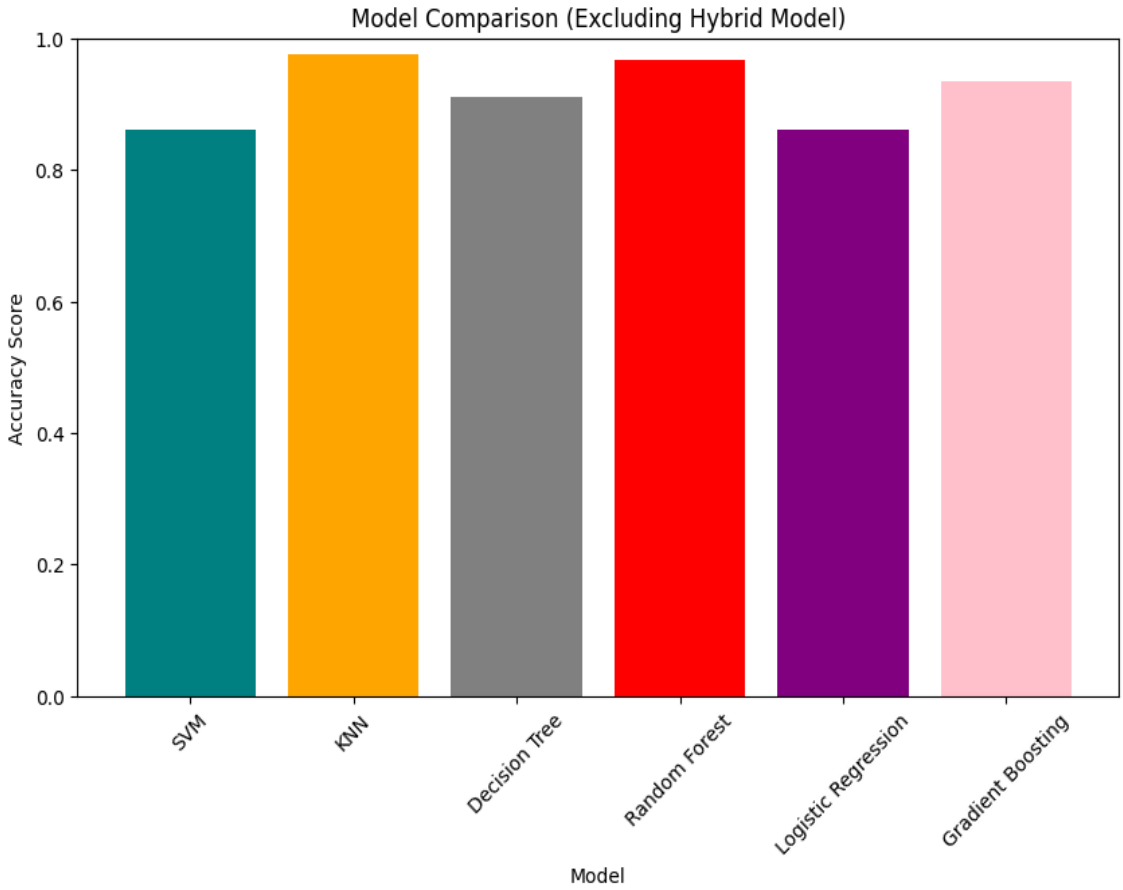

Figure 4.2: Model Comparison with accuracy measurement.

4.6 Decision Tree (DT)

With a precision of 0.9243, recall of 0.9109, and an F1 score of 0.9101, the Decision Tree (DT) algorithm yielded an accuracy of 91.09%. Despite its strong performance, DT was not able to outperform KNN and RF, most likely because of its propensity to overfit the training set. There were 2216 true positives, 1781 true negatives, 0 false positives, and 391 false negatives according to the DT confusion matrix.

4.7 Performance of Ensemble Models

With a VotingClassifier combining KNN, SVM, Random Forest, and Gradient Boosting classifiers, the proposed ensemble model was able to attain a phenomenal accuracy of 99.59%. By utilizing soft voting, which averages a predicted probability of each classifier to get a final choice, this hybrid strategy capitalizes on the capabilities of each individual classifier. This produces a model that is more reliable and accurate than any of the individual classifiers that were evaluated previously.

Table 4.2 : Model performance comparison.

Prediction Model	Accuracy
NB,SVM,ANN (Duc - Hau et al. 2020)	63%,68%,71%
DT,SVM,RF,KNN (Xiaomei et al. 2021)	78.1%,81.4%,78.6%,76.1%
ETRF+XGB (Ali et al. 2022)	92%
KNN,CATBOOST,RF,XGBOOST (Sadichchhha et al. 2022)	79.71%,75.36%, 89.4%, 86.1%
ANN,KNN,SVM (Nasir et al. 2022)	85.7%,84.9%, 84.3%
SVM, KNN (Taher et al. 2022)	92.8%,92.5%
Ensemble Model (Our work)	**99.59%**

It appears from comparing the performance of different prediction models across studies that substantial progress has been made. Results for Naive Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were presented by Duc - Hau et al. (2020), who attained accuracies of 63%, 68%, and 71%, respectively. With accuracies of 78.1%, 81.4%, 78.6%, and 76.1%, respectively, Xiaomei et al. (2021) showed improvements with Decision Tree (DT), SVM, Random Forest (RF), and K-Nearest Neighbors (KNN) models. With an accuracy of 92%, Ali et al. (2022) presented the ETRF+XGB model, demonstrating the potency of ensemble methods. The accuracy of the KNN, CatBoost, RF, and XGBoost models were reported by Sadichchhha et al. (2022) to be 79.71%, 75.36%, 89.4%, and 86.1%, respectively. Furthermore, Taher et al. (2022) obtained accuracies of 92.8% and 92.5% for SVM and KNN models, while Nasir et al. (2022) reported results for ANN, KNN, and SVM with accuracies of 85.7%, 84.9%, and 84.3%, respectively. With an astounding accuracy of 99.59%, our ensemble model far beats current cutting-edge methods, demonstrating its effectiveness and practical applicability.

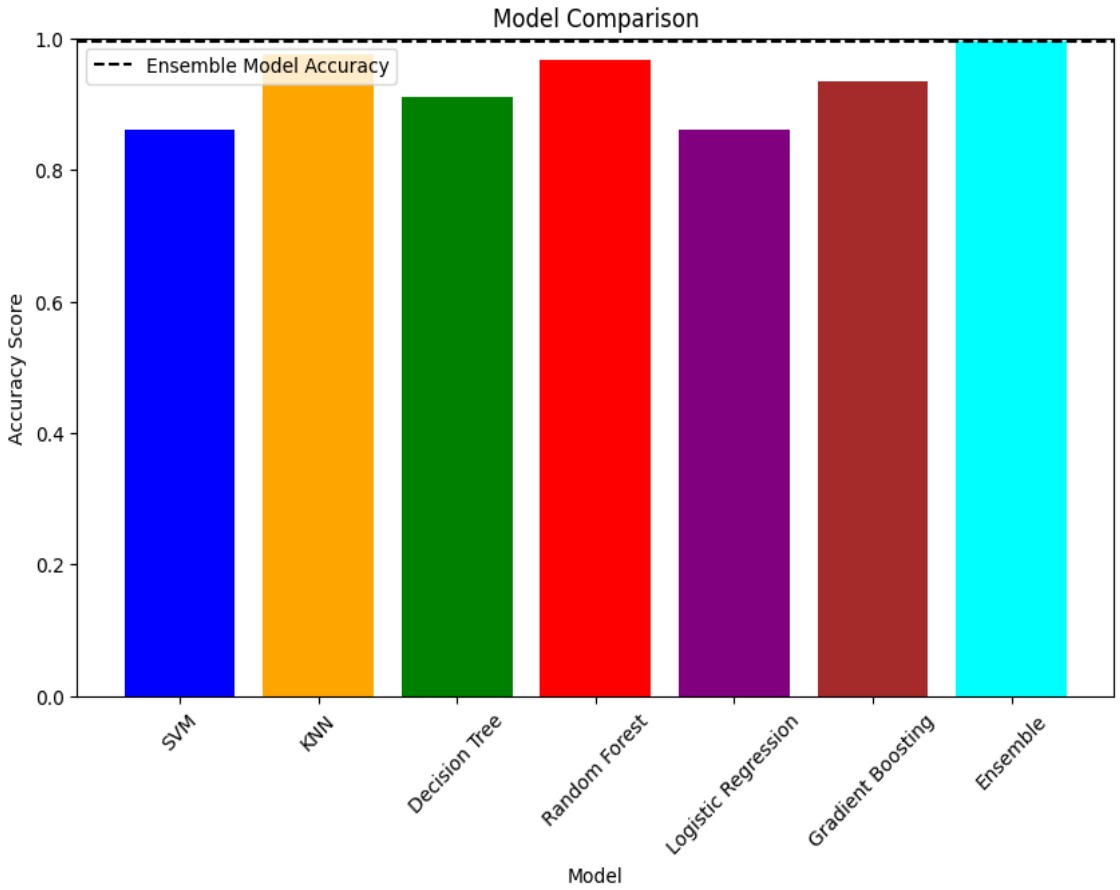

Figure 4.3 : Performance comparison of all models.

The remarkable accuracy of 99.59% achieved by the hybrid model highlights the efficacy of merging different classifiers using soft voting. The combination of all the classifiers' distinct capabilities creates a model that is more reliable and accurate than any one classifier working alone. By combining the complementing advantages of Gradient Boosting, Random Forest, SVM, and KNN, this method overcomes the shortcomings of each technique separately and produces better results. The success of the hybrid model shows how ensemble methods may be used to increase classification accuracy and reliability, which makes them a useful strategy for challenging data processing work.

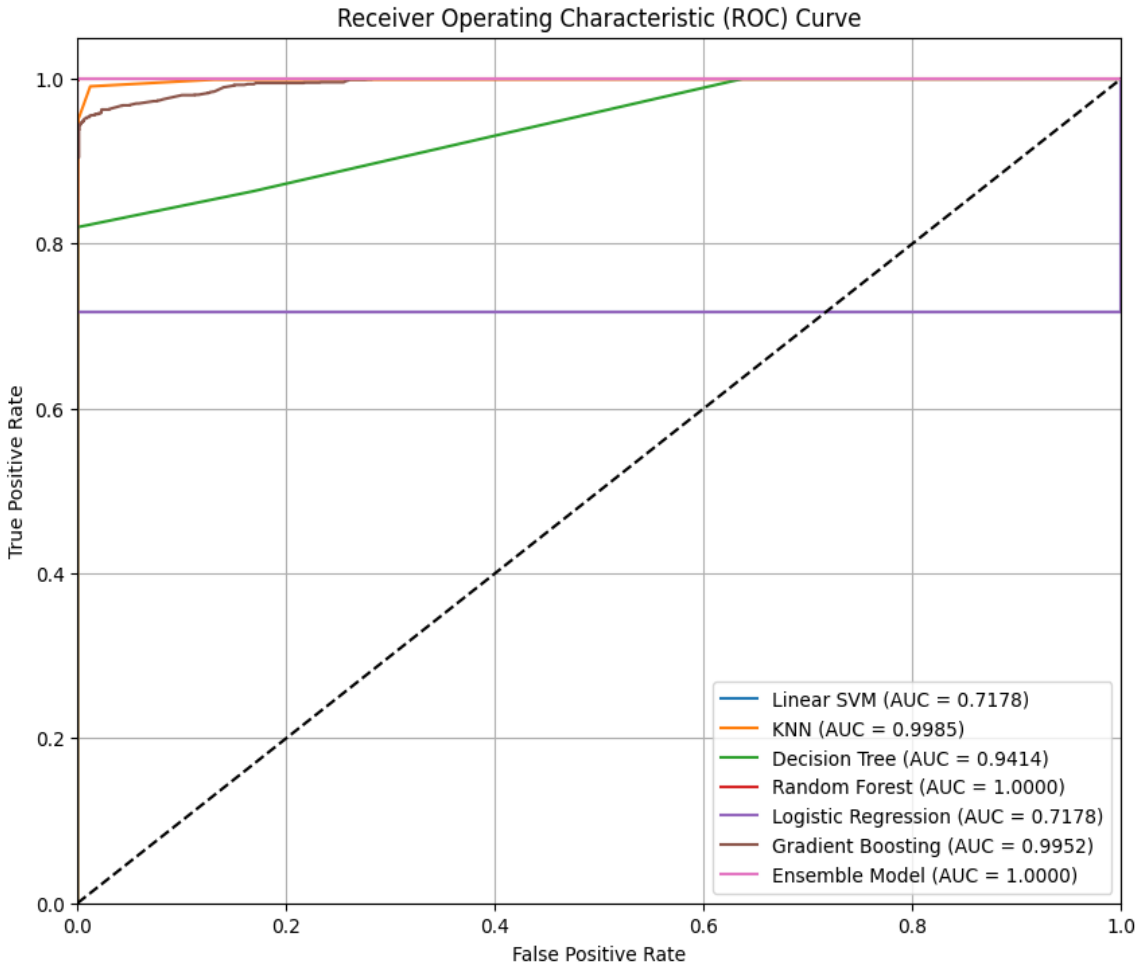

Figure 4.4: ROC curve analysis based on the performance model.

The AUC scores obtained from the ROC analysis unequivocally demonstrate the dominance of the hybrid ensemble model and Random Forest, with both earning flawless AUC scores of 1.00. These models are the most efficient classifiers for this dataset. KNN and Gradient Boosting had outstanding performance, achieving AUC scores of 0.99, which indicates their robust classification capabilities. The Decision Tree model demonstrates robust performance, with an AUC (Area Under the Curve) value of 0.94, but slightly lower than the leading models. Linear SVM and Logistic Regression, despite having AUC ratings of 0.71, exhibit only moderate performance. Therefore, they may not be the optimal options for tasks that demand strong discrimination. The analysis highlights the importance of ensemble approaches in attaining high levels of accuracy and dependability in classification problems.

Chapter 5

5 Conclusion:

Early detection of genetic diseases is essential for maximizing the course of these conditions and increasing individual well-being. We are aware of the significance of timely diagnosis of genetic diseases due to our comprehension of the complexities of the Human Genome. After doing a thorough literature review, we have learned about many approaches to creating predictive models in this field.We address this by putting out a methodology specifically designed for building this kind of model and emphasizing its importance and its effects. We also describe the system requirements needed to successfully implement and assess various classifier and prediction algorithms.. We have developed a strong technique to achieve the highest accuracy in disease prediction, drawing on extensive studies of the literature. The effectiveness of machine learning data modeling in this particular situation is demonstrated by its capacity to manage difficult, multifaceted data. The genetic disorder prediction limitation on the genetic sequences data for the prediction process.In future the model can be used by healthcare professionals to diagnose patients and support laboratory research on genetic disorders and enhancing accessibility through a graphical user interface, such as a website or mobile application, will facilitate use and it’s efficacy will become clear through testing on a variety of populations, advancing medical practice.Transfer learning approaches can be used to increase the model's performance and adaptability to various genetic diseases in multi-label, multi-class classification tasks.

References

1. Umar Nasir, M. *et al.* (2022) "Single and mitochondrial gene inheritance disorder prediction using machine learning," *Computers, materials & continua*, 73(1), pp. 953–963. doi: 10.32604/cmc.2022.028958.
2. Ghazal, T. M. *et al.* (2022) "Supervised machine learning empowered multifactorial genetic inheritance disorder prediction," *Computational intelligence and neuroscience*, 2022, pp. 1–10. doi: 10.1155/2022/1051388.
3. Naik, S. *et al.* (2022) "Prediction of Genetic Disorders using Machine Learning," *International Journal of Scientific Research in Science and Technology*, pp. 01–09. doi: 10.32628/ijsrst229273.
4. Rahman, A.-U. *et al.* (2022) "IoMT-based mitochondrial and multifactorial genetic inheritance disorder prediction using machine learning," *Computational intelligence and neuroscience*, 2022, pp. 1–8. doi: 10.1155/2022/2650742.
5. Le, D.-H. (2020) "Machine learning-based approaches for disease gene prediction," *Briefings in functional genomics*, 19(5–6), pp. 350–363. doi: 10.1093/bfgp/elaa013.
6. Sikandar, M. *et al.* (2020) "Analysis for disease gene association using machine learning," *IEEE access: practical innovations, open solutions*, 8, pp. 160616–160626. doi: 10.1109/access.2020.3020592.
7. Atta-Ur-Rahman *et al.* (2022) "Advance genome disorder prediction model empowered with deep learning," *IEEE access: practical innovations, open solutions*, 10, pp. 70317–70328. doi: 10.1109/access.2022.3186998.
8. Liu, Y. *et al.* (2014) "DiME: A scalable disease module identification algorithm with application to glioma progression," *PloS one*, 9(2), p. e86693. doi: 10.1371/journal.pone.0086693.
9. Ghiassian, S. D., Menche, J. and Barabási, A.-L. (2015) "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome," *PLoS computational biology*, 11(4), p. e1004120. doi: 10.1371/journal.pcbi.1004120.
10. Scheuner, M. T. and Rotter, J. I. (2006) "Quantifying the health benefits of genetic tests: The importance of a population perspective," *Nature*, 8, pp. 141–142.

11. Katherisan, S. *et al.* (2008) “Polymorphisms associated with cholesterol and risk of cardiovascular events,” *New England Journal of Medicine*, 358(12), pp. 1240–1249.
12. Venkat, V. *et al.* (2023) “Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine,” *Genomics*, 115(2), p. 110584. doi: 10.1016/j.ygeno.2023.110584.
13. Evgeniou, T. and Pontil, M. (2001) “Support Vector Machines: Theory and Applications,” in *Machine Learning and Its Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 249–257.
14. Moldagulova, A. and Sulaiman, R. B. (2017) “Using KNN algorithm for classification of textual documents,” in *2017 8th International Conference on Information Technology (ICIT)*. IEEE.
15. Dong, T., Cheng, W. and Shang, W. (2012) “The research of kNN text categorization algorithm based on eager learning,” in *2012 International Conference on Industrial Control and Electronics Engineering*. IEEE.
16. Guo, G., Ping, X. and Chen, G. (2006) “A fast document classification algorithm based on improved KNN,” in *First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC’06)*. IEEE.
17. Yunliang, Z. *et al.* (2009) “Flexible KNN algorithm for text categorization by authorship based on features of lingual conceptual expression,” in *2009 WRI World Congress on Computer Science and Information Engineering*. IEEE.
18. Moldagulova, A. and Sulaiman, R. B. (2016) “Classification of textual documents in r using knn algorithm,” *PONTE International Scientific Researchs Journal*, 72(12). doi: 10.21506/j.ponte.2016.12.35.
19. A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," *2011 IEEE Control and System Graduate Research Colloquium*, Shah Alam, Malaysia, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826
20. Provost, F., Hibert, C. and Malet, J. P. (2016) “Automatic classification of endogenous seismic sources within a landslide body using random forest algorithm //EGU,” *General Assembly Conference Abstracts*, 18.

21. Zou, X. *et al.* (2019) “Logistic regression model optimization and case analysis,” in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. IEEE.
22. Upadhyay, D. *et al.* (2021) “Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids,” *IEEE transactions on network and service management*, 18(1), pp. 1104–1116. doi: 10.1109/tnsm.2020.3032618.
23. Chen, T. and Guestrin, C. (2016) “XGBoost: A Scalable Tree Boosting System,” *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1603.02754.
24. Josi, R. M. and Minu, R. I. (2020) “Review of computational approaches to Parkinson’s disease gene prediction,” in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE.
25. Raza, A. *et al.* (2022) “Predicting genetic disorder and types of disorder using chain classifier approach,” *Genes*, 14(1), p. 71. doi: 10.3390/genes14010071.

