JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 08 Issue 01
2024

Research Article
https://doi.org/10.56979/801/2024

Machine Learning-Based Multi-Factorial Genetic Disorder Prediction System

Maya Bint Yousaf[1], Sadia Abbas Shah[2] , Syed Younus Ali[3], Huma chaudhry[3], Ahmad Ibne Yousaf[4], Khurram Aziz[5], Khizra Hashmat[6], and Misbah Akram[3]

[1]Lecturer Department CS & IT, Minhaj University Lahore, Lahore, 5700, Pakistan.
[2]School of System and Technology, Department of Software Engineering, University of Management and Technology, Lahore, Pakistan.
[3]School of Software Engineering, Minhaj University Lahore, Lahore, 5700, Pakistan.
[4]School of Human, Nutrition & Dietetics / Assistant Registrar, Minhaj University Lahore, Lahore, 5700, Pakistan.
[5]Service Engineer, Automation Technology, University of Engineering and Technology, Lahore, Pakistan.
[6]School of Human, Nutrition & Dietetics, Minhaj University Lahore, Lahore, 5700, Pakistan.
*Corresponding Author: Maya Bint Yousaf. Email: mayayousaf.csit@mul.edu.pk

Received: August 11, 2024 **Accepted:** October 12, 2024

Abstract: A genetic disorder is a medical illness caused by an error in a person's DNA. Genes are hereditary units that carry instructions for the body's development, functioning, and upkeep. Mutations or alterations in these genes can cause genetic illnesses, affecting how the body's cells create proteins or carry out certain functions. This study proposed the multi-factorial genetic disorder prediction system using machine learning algorithms specifically Decision trees and Naïve Bayes. The study covered three diseases cancer, cystic fibrosis and diabetes. Diabetes, cystic fibrosis, and cancer are frequently the result of a complicated interaction of genetic, environmental, and behavioural variables. This proposed model can identify those who are more likely to develop certain diseases, allowing for early intervention and personalized preventive care. These proactive healthcare methods not only improve patient outcomes but also contribute to the general efficiency and efficacy of the healthcare system, supporting a change towards more targeted and efficient healthcare delivery. The system aims to predict the likelihood of an individual developing a system that predicts based on various factors such as family history, lifestyle and environmental factors. This study evaluated a disorder-based combined dataset from Kaggle containing 43 attributes and 9467 rows and achieved 96.95% accuracy in the training phase and 94.25% in the testing phase. In predicting the likelihood of developing a multi-factorial genetic disorder. This system can potentially assist healthcare professionals in providing personalized preventive care to individuals at high risk of developing a genetic disorder. This proposed model helped improve the overall healthcare system and the well-being of the individuals.

Keywords: Cancer; Cystic Fibrosis; Diabetes; ; Decision Tree; Machine Learning; Naïve Bayes.

1. Introduction

A multi-factorial genetic disorder also known as a polygenic disorder is caused by a composition of multi-genetic and environmental factors that encompasses the interaction of several genes. These genetic traits may be inherited from either one or both parents. This disorder inheritance pattern is not simple instead the risk is the manifestation of the cumulative effect of multiple genetic variants interplaying together. The proposed model is based on machine learning for the prediction of multi-factorial genetic disorders that are used for diseases (Diabetes, Cystic Fibrosis, and Cancer) prediction [1].

Timely prediction of multi-factorial genetic disorders is critical since an individual's life and progress depend upon it. Various solutions already exist but they still need a reliable and efficient prediction model to overcome the challenges faced by this disease. For this, machine learning approaches which is used in

this study for efficient and timely prediction of multi-factorial genetic disorder disease (Rahman et al. 2022).

The multi-factorial genetic disorder involves various stages and components to recognize the risk of having this disorder like diabetes, cystic fibrosis and cancer for individuals. Diabetes shows problems in working of the insulin that cause chronic hyperglycemia and severe vascular complications. Cystic Fibrosis is a multifactorial protein misfiling disease with a major effect on respiratory function. In cancer, the genetic predisposition can interact with external factors like exposure to carcinogens to further contribute to the danger of developing this disease [2].

Diverse alternative solutions are already present but still need to enhance the multi-factorial genetic disorder prediction system. The proposed model is based on a Machine Learning model that uses a Naive Bayes and Decision Tree (DT) using a featured-based dataset from Kaggle which contains 43 attributes and 9467 rows. The proposed model contains parameters (White blood cell, maternal illness, radiation exposure, substance abuse, birth asphyxia, family history and many more given in the dataset) and gives high accuracy compared to earlier approaches. This method involves the replacement of the missing or defective gene with a functional one, enabling the body to produce the correct enzyme or protein and thereby addressing the fundamental cause of the disease. Predicting disease genes stands as a crucial matter in biomedical research. Initially, strategies based on annotation were introduced to address this issue. Initial signs of genetic diseases can be found in ancient cultures. People noticed odd characteristics and noticed that some disorders tended to run in families. However, the processes behind these results remained a mystery [3].

Most individuals do not experience adverse effects from their faulty genes as they inherit two copies of nearly all genes – one from the mother and the other from the father. The only exception to this generalization applies to genes located on the male chromosomes. In males, where one X and one Y chromosome are present, the former is inherited from the mother, and the latter from the father. Consequently, each cell possesses only one copy of the genes located on these sex chromosomes [4]. In the majority of instances, having one normal gene is adequate to prevent the manifestation of disease symptoms. In the case of recessive genes, if they have a normal counterpart, it can perform all the necessary functions. Only when individuals inherit two copies of the same recessive gene from their parents will a disease manifest. Combination of genetic and environmental factors Environmental toxins can induce disease by impacting genes identifying the genes linked to a disease is valuable for both disease prevention and treatment. Additionally, it plays a crucial role in comprehending the biological functions of genes [5].

1.1. Machine Learning

Machine learning (ML) involves advanced computer algorithms that mimic human intelligence and decision-making by learning from their surroundings in recent years, this method has gained popularity in neuroscience and cognition. Numerous studies have emphasized the effectiveness of ML in these fields. ML is a discipline centered on two interrelated inquiries: The investigation of machine learning is significant both for tending to these essential logical and designing inquiries furthermore for the highly functional program it has provided and managed across a wide range of applications [6-8].

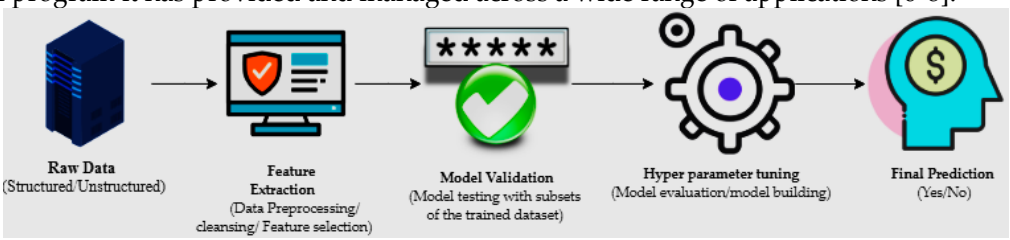

Figure 1. The primary mechanisms of machine learning [8]

ML is a discipline centered on two interrelated inquiries: Machine learning research is important not only tar addressing these important logical and design questions but also for the highly effective program it has produced and managed across a variety of applications. Machine learning represents a branch of artificial intelligence (AI) that concentrates on creating algorithms and statistical models capable of enhancing their performance on specific tasks through experience, all without explicit programming [9,10]. Essentially, Machine learning gives computers the ability to learn from data so they can anticipate or decide without direct human intervention. The process involves several key stages, starting with the collection of

relevant and representative data for the given task. During the training phase, the chosen algorithm is exposed to this data, allowing it to identify patterns, relationships, and trends. The algorithm adjusts its internal parameters to optimize performance. Subsequently, the model's capacity to generalize and produce precise predictions is assessed using fresh, untested data, facilitating an assessment of its performance. Once the model demonstrates satisfactory results, it can be deployed to make predictions or decisions on new, real-world data. Machine learning encompasses three primary types: In supervised learning, the system receives training from labelled data; unsupervised learning, which identifies patterns in unlabeled data; and reinforcement learning, where the model learns optimal strategies through interactions with an environment and feedback in the form of rewards or penalties. This versatile technology finds applications in various domains, including natural language processing, image and audio recognition, and recommendation systems, and autonomous cars, contributing to significant advancements in addressing complex problems [9-10]

1.1.1. Types of Machine Learning

Regardless of the learning algorithm employed, a key objective is to theoretically characterize the capabilities of distinct learning algorithms and the intrinsic challenges associated with specific learning problems. This encompasses grasping the algorithm's capacity to learn accurately from a particular type and volume of training data, its resilience to errors in modeling assumptions or training data, and the determination of whether a successful algorithm can be devised for a learning problem given a specified volume of training data, or if the nature of the problem inherently poses significant challenges [11].

Efforts to theoretically describe ML algorithms are used to the integration of statistical and computational theories, aiming to describe the pattern difficulty (how much data is needed for accurate learning) and computational complexity (how much computation is required) to understand their dependence on factors such as the algorithm's representation of what it learns. Development theory, with upper and lower bounds, has proven particularly useful in recent years [12].

Broadly speaking, machine learning algorithms fall into three types: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Like any methodology, there are different approaches to training AI algorithms, each with its merits and drawbacks. To understand these, let's examine the kind of information they consume. Machine learning has two types of data: labelled and unlabeled. The labelled dataset contains both input and output boundaries, but it needs significant individual effort to label. In contrast, an unlabeled dataset has only boundaries in a machine-understandable format, reducing the need for individual labor but requiring difficult structures.

1.1.2. Supervised learning

One kind of machine learning is supervised learning, in which the algorithm is trained on labelled data. Although precise labelling of data is essential for the method's effectiveness, supervised learning can prove highly effective under the appropriate conditions. In this approach, a machine learning algorithm is provided with a concise training dataset. This dataset, a subset of a larger dataset, gives the algorithm a basic grasp of the data points, the problem, and the answer. It will encounter. Throughout the training process, the algorithm develops a solid comprehension of how the data functions and the relationship between input and output. Subsequently, the final dataset is employed to further train the solution, enabling the algorithm to continuously improve and identify new patterns and relationships even after deployment [13].

1.1.3. Unsupervised Learning

Unsupervised learning alludes to the utilization of Artificial Intelligence (A-I) calculations to distinguish designs in informational indexes having informational focal points that are neither designated nor arranged. The computations are thus authorized to define, label, or any external guidance in performing that task. In essence, unsupervised learning enables the system to identify patterns within datasets independently An AI system will classify unsorted data using similarities and differences in unsupervised learning. Even in the absence of provided classifications. Unsupervised learning calculations are capable of carrying out more complex errands than controlled learning frameworks. Furthermore, one way to assess Al's understanding is to expose a framework to solo learning [14].

1.1.4. Reinforcement Learning

Reinforcement learning draws inspiration from how individuals acquire information in their daily activities. It involves a computational process that evolves and learns from new situations through

experimentation. Positive outcomes are incentivized or reinforced, while negative outcomes are penalized or discarded. In the framework of reinforcement learning, the algorithm is placed in an environment with an agent and a system

During each iteration of the algorithm, the output result is presented to the agent, which assesses whether the outcome is favorable. If the algorithm discovers the correct solution, the agent reinforces the solution by providing a reward. In cases where the outcome is suboptimal, the algorithm is compelled to iterate until it finds an improved result. Typically, the reward system is directly tied to the effectiveness of the outcome [15]. The answer is not a fixed number in common reinforcement learning applications, like determining the shortest path between two points on a map. Timely prediction of multi-factorial genetic disorders is critical since an individual's life and progress depend upon it. Various solutions already exist but still need a reliable and efficient prediction model to overcome the challenges faced by this disease. For this, machine learning approaches is used in this study for efficient and timely prediction of multi-factorial genetic disorder disease. Machine learning-based Multifactorial Genetic Disorder prediction systems hold the potential to significantly enhance the diagnosis, treatment, and prevention of Multifactorial Genetic Disorder. These systems offer the capability to identify individuals at risk of developing a Multifactorial Genetic Disorder, create personalized treatment plans, and pinpoint novel drug targets. However, several challenges persist, including data availability, data integration, model interpretability, and model validation. As machine learning algorithms advance and more data becomes accessible, the future holds promise for increasingly innovative and effective ML-based Multifactorial Genetic Disorder prediction systems.

1.2. Decision Tree

A decision tree algorithm stands as a widely employed machine learning technique, serving purposes in both classification and regression tasks. Its functioning revolves around the recursive division of a dataset into subsets, guided by the most influential attribute. This process constructs a structure that resembles a tree, with each internal node representing a choice made in response to a certain trait. Individual branch emanating from these nodes signifies an outcome of that decision, and every leaf node represents the ultimate predicted outcome [16].

The interpretability of decision trees is one of its benefits. Decision trees, unlike some other machine learning algorithms, can be visualized and comprehended by people. This is critical for a wide range of applications, including medical diagnosis, fraud detection, and legal decision-making, where openness and explain ability are critical. Without considerable preparation, decision trees can handle both numerical and categorical data, as well as missing values and outliers.

However, decision trees have several drawbacks. They are susceptible to overfitting, particularly when the tree is large or the data is noisy. This might result in poor generalization performance on previously unknown data. Various pruning strategies, such as reduced error pruning, cost complexity pruning, and minimal description length pruning, can be employed to alleviate this. Another difficulty with decision trees is their sensitivity to the splitting criterion and feature order. Various criteria and ordering might result in various trees and forecasts. To solve this, ensemble approaches like as bagging, boosting, and random forests may be used to integrate numerous trees and enhance prediction resilience and accuracy [18] [21].

1.3. Naïve Bayes

Naive Bayes is a probabilistic algorithm frequently employed in machine learning for classification purposes. Its foundation lies in Bayes' theorem, a formula for determining the probability of a hypothesis given observed evidence. Despite its simplifying assumption of independence among features, Naive Bayes has demonstrated effectiveness in a variety of practical applications in the real world [17].

Naive Bayes classifiers are widely used in a variety of applications, including text classification, spam filtering, and recommendation systems. They are known for their simplicity, efficiency, and effectiveness, especially when dealing with large datasets. Despite their oversimplified assumptions, they can often produce results that are comparable to or even better than more complex machine learning models.

2. Literature Review

Gazal et al. (2022) [18] proposed Machine Learning with Supervision using K-nearest neighbor (KNN) and Support Vector Machine (SVM). In this study, the developed model uses a multi-factorial genetic

inheritance disorder dataset. The proposed model achieved good accuracy but the limitation is that the dataset was limited and only diseases were covered.

Behravan, Hartikainen, Tengstrom, Kosma and Mannermaa (2022) described to Predict breast cancer using machine learning approaches. The developed model uses the KBCP dataset. In this study, the proposed model achieved 78.00% accuracy. The limitation is that the small parameter, restricted and imbalanced dataset is used and the accuracy is also low [19].

Ghazal et al. (2019) presented the framework of facial image analysis for Deep learning is being used to identify face characteristics of genetic diseases, Deep Gestalt, which computer vision and deep learning methods are used. In this study, the developed model uses the Casia-Web-Face dataset. The presented model obtained good accuracy However, there is a limitation in this study is the lack of comparison to the other method or human experts in some experiments. The result is restricted to patients with very specific syndromes and is not transmittable [20].

Battineni, Sagaro, Nalini, Armenta and Tayebati (2019) described the comparative analysis of machine-learning classifiers like Naive Bayes, Logistic Regression and Random Forest for prediction of Type-2Diabetes.In this study, the Pima India Diabetes dataset (PIDD) is used. This research focuses on a follow-up investigation centred on cross-validation methods. The developed model achieved Naïve Bayes with 81%, Logistic Regression with 83% and Random Forest with 82% accuracy. The model achieved different accuracy by using supervised learning approaches. The use of better convolutional and deep learning approaches with large datasets will improve the performance [22].

Osker, Pahikkala and Aittokallio (2013) describe Exploring Machine Learning and Network Perspectives for Disease Risk Prediction: Understanding Genetic Variants and Their Interactions. Using advanced feature selection algorithms implemented, LASSO (least absolute shrinkage and selection operator. The generated model in this study makes use of a high-dimensional dataset. But the limitation is that binary disease outcomes will not provide the most authentic study phenotypes. These all may have important effects on forecast production [22].

Kavakiotis, Tsave, Salifoglou, Maglaveras, Vlahavas and Chouvarda (2016) presented the Application of machine learning and data mining techniques in diabetes research. In this study, clinical datasets were used. To contribute insight into the utilization of advanced computational approaches for comprehending and managing diabetes. This contribution furnishes valuable tools for the diagnosis and treatment and more research of diabetes within the field. The developed model achieved 89% accuracy. The limitation is that the research only covers one disease and a restricted dataset [23].

Nguyen and Ho (2011) presented identifying genes associated with disease using semi-supervised learning and protein-protein interaction networks. In this study, gene expression dataset. It highlights the effectiveness of semi-supervised learning in the detection of disease-related genes, particularly when integrating information from protein interactions. The developed model achieved 82% accuracy. However, the label data is restricted. The study is only covering a particular disease [24].

Battineni, Sagaro Chinatalapudi and Amenta (2020) presented the predictive model in chronic disease diagnosis using machine learning applications. The research is anticipated to provide insights into the potential of machine learning in advancing diagnostic procedures for chronic diseases, with a focus on personalized and predictive methodologies. The proposed model achieved 79% accuracy. The model achieved low accuracy and lacked of optimal dataset and needs to improve the model approaches in a better way [25].

Gulande and Awale (2022) presented the Systematic Study of Gen Profiles Analysis Methods in Disease Classification using a machine learning algorithm. In this study, the developed model uses a dataset which is downloaded from GO DATABASE. The accuracy of this study is different for individual classification of diseases. The limitation is that the rare event detection techniques and their applications are not endorsed by the authors [26].

Algamal and Lee (2018) proposed sparse logistic regression in two stages for efficient gene selection in high-dimensional microarray data categorization using sparse logistic regression. The presented model uses four public-domain high-dimensional gene expression datasets in this study. The proposed model obtained 95.13% accuracy but the limitation is that the dataset is restricted and a small group of genes were covered for this research [27].

Khan et al. (2021) presented the model for brain disease diagnosis using Machine learning and deep learning approaches the authors are anticipated to delve into the foundational principles of these approaches, offering insights into recent advancements in the field. The exploration may encompass how these technologies contribute to enhancing the precision and efficiency of diagnosing various brain disorders. In this study, the developed model uses open-source datasets with twenty-two different sets of data formats. The proposed model achieved good accuracy but has restricted datasets to improve the classification accuracy of this model, it needs to extend their datasets [28].

Le, Hoai and Kwon (2015) presented a Novel Disease Gene Prediction using classification-based machine learning methods in which DT, KNN, NB, ANN, SVM, and RF algorithms were used. The developed model uses an omics dataset. In this study, the proposed model achieved good accuracy and RF showed better performance. Moreover, this research likely involves an assessment and comparison of the effectiveness in the identification of genes associated with diseases. The primary objective may involve evaluating the accuracy, efficiency, and robustness of various methods in predicting novel disease genes with restricted datasets [29].

3. Proposed Methodology

In this phase, the machine learning framework helps to predict the multi-factorial genetic disorder and single-gene inheritance disorder. A disorder-based dataset is taken from Kaggle for the identification of selected diseases like diabetes, cancer and cystic fibrosis. The model contains 2 phases, the training and validation phase where training gets 70% and validation get 30% dataset. During the training phase, the first layer is the acquisition layer where data is acquired by the IOMT device. Input variables are white blood cells, maternal illness, radiation, substance abuse, birth asphyxia, family history and others in the dataset.

The dataset is taken from Kaggle which is a combination dataset of genetic disorder and disorder subclass containing folders like train, csv, test, and CSV and sample submission.csv. The data is raw which is why it needs to be clean so it is directed to the next layer which is a pre-processing layer. In this layer, moving average and bin means methods are used for the conversion of raw data into processed form. As we know that the bin mean and moving average methods is used to clean the raw data because the bin method and the moving average approach are data cleaning strategies that are used to improve the quality and interpretability of datasets. The moving average approach, which is particularly useful for smoothing time-series data, involves calculating the average of a predetermined number of consecutive data points, resulting in a moving average that minimizes short-term fluctuations or noise. Use of this strategy to more effectively find and comprehend long-term patterns in data, decreasing the effects of outliers or random fluctuation. This results in a more precise portrayal of the underlying patterns, allowing for more accurate assessments and informed decision-making.

After that, processed data are directed to the application layer where it is sub-divided into the prediction layer and performance layer. Two algorithms which are decision tree (DT) and Naïve Bayes (NB) are used in the prediction layer and the performance layer, miss rate, sensitivity, specificity and accuracy are measured to check the performance for prediction as shown in figure 1. If the model is trained successfully then the data are saved in the database otherwise the model can be retrained. After that the validation phase starts, where the model is tested whether it is predicted accurately for further recommendation or deletion.

Decision Tree and Naïve Bayes are classification algorithms of machine learning models but decision tree is also used for regression tasks. The decision tree predicts the target value based on the input parameters. Naïve Bayes are good and efficient for high-dimensional data but assume features are independent.

3.1. Training Phase

The first layer of the training phase is the data acquisition layer where the data can be acquired from the IOMT devices.

3.1.1. IOMT Device of Dataset

IOMT, or the Internet of Medical Things, pertains to the integration of medical devices and applications that can communicate with healthcare information technology systems. These IOMT devices

are intended to capture, transmit, and receive health data, allowing for real-time monitoring and analysis to improve healthcare management.

A dataset, on the other hand, is nothing more than a collection of data. A dataset in the context of IOMT could include assembled medical or health-related information gathered from various devices or sources.

When referring to an 'IOMT Device of Dataset,' refers to a medical device that generates or is linked to a dataset within the Internet of Medical Things. This device essentially captures and transmits health data, adding vital information to a larger dataset.

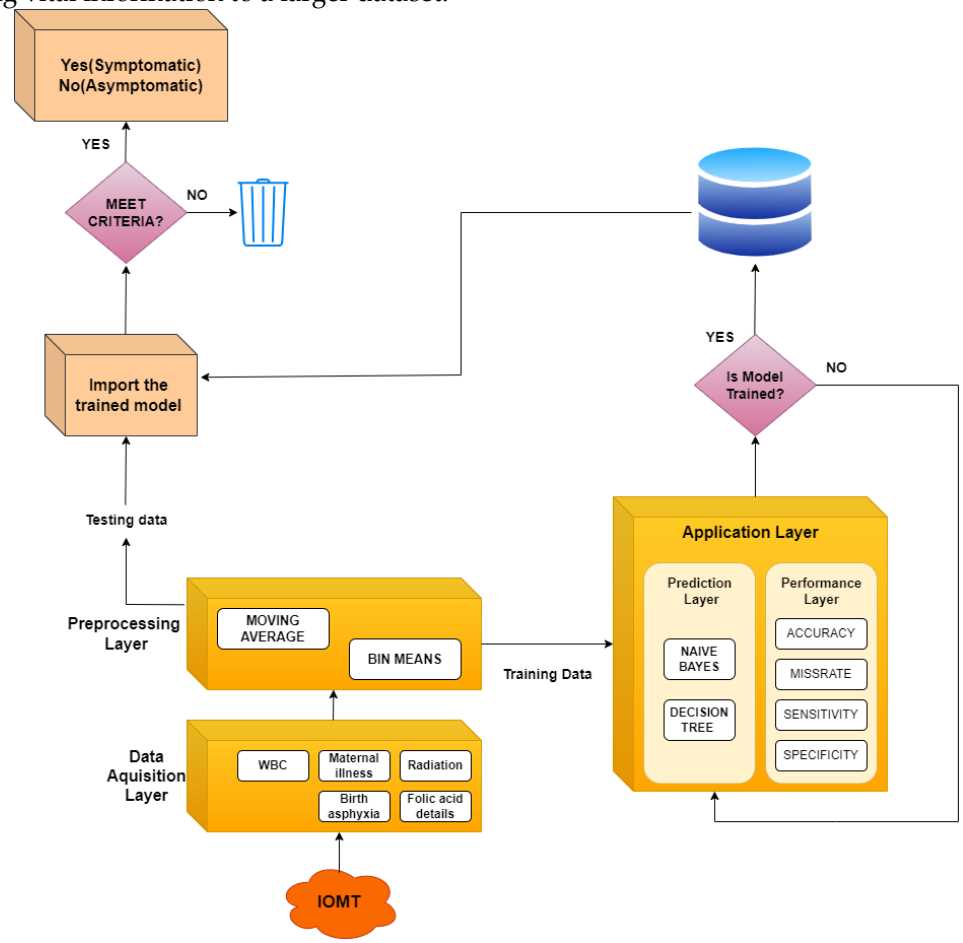

Figure 2. Proposed Multi-Factorial Genetic Disorder Using ML

3.1.2. Data Acquisition Layer

The data acquisition layer is a stage in the training phase for machine learning models in the context of IoMT (Internet of Medical Things) systems. This layer collects data from numerous medical devices, sensors, and other sources linked to the IoMT network. Data can be acquired via wired or wireless connections and can comprise physiological signals, patient health information, environmental data, and other pertinent metrics. Because the obtained data serves as the foundation for the future training of machine learning models, the data collecting procedure should assure its correctness, completeness, and dependability. This layer may also include data cleaning, normalization, and feature extraction algorithms to prepare the data for the training phase. The accuracy of the data collected

Moreover, The Data Acquisition Layer is a critical component of information technology systems that is responsible for methodically collecting and aggregating data from various sources. Its major job is to collect raw data from sensors, instruments, or devices and process, analyses, and store it. Essentially, the Data-collecting Layer serves as a vital interface between the physical world and digital systems, ensuring successful raw data collecting and initial processing within larger system architecture.

Table 1. Parameters of Dataset

Effective Parameter	Specification	Normal Ranges

White Blood Cells	These are the cells raised during infection or inflammation.	4500 to 11,000 WBCs per microliter of blood.
Maternal Illness	Any mental or physical illness that is directly related to pregnancy and childbirth.	Any type of illness occurring between weeks 1 to 42 weeks of gestation.
Radiation	Energy is used to treat cancer. And to kill cancerous cells.	Radiation
Birth Asphyxia	A condition in which a baby doesn't receive enough O2 and nutrients before, during or right after birth.	Birth Asphyxia
Folic Acid	It is Vitamin B. It helps the body to make new cells. Important nutrient for pregnant ladies.	Folic Acid

The proposed model's effective factors are WBC, maternal sickness, radiation, birth hypoxia, and folic acid.

3.1.3. Preprocessing Layer

Raw data is prepared for deeper analysis and application in the preprocessing layer. It cleans, filters, and transforms data to improve its quality and make it compatible with the following steps in processing. This layer in the proposed model deals with moving average and bin means procedures to ensure that the data is in the optimal shape for meaningful analysis and comprehension across the system.

3.1.3.1. Moving Average Method

In the preprocessing layer, the moving average approach is frequently used to smooth time-series data. Find the average of a bunch of neighboring data points. This average becomes the new value for a given data point. This procedure aids in decreasing bumps and making the data more understandable at each stage of the process.

3.1.3.2. Bin Mean Method

A preprocessing layer's bin mean approach organizes data by classifying it into bins and calculating the average value for each category. This makes the data easier to deal with and allows you to detect patterns and trends in different value ranges.

3.2. Application Layer

The application layer is further subdivided into two layers: prediction and performance. In the prediction layer, the proposed model is trained using various machine learning algorithms, and in the performance layer, the proposed model is evaluated using various methods.

3.2.1. Prediction Layer

The two methods Nave Bayes and Decision Tree are utilized to train the dataset in the prediction Layer.

3.2.1.1. Naive Bayes

Naive Bayes is a popular machine learning method for categorizing objects. It predicts which category a data point belongs to base on its features. It calculates the likelihood of a category given the features of the data using Bayes' theorem. Naive Bayes assumes that the features are independent, which simplifies the calculation. It is used in tasks such as text categorization, spam filtering, and mood analysis. There are three common types: Gaussian, multinomial, and Bernoulli. While Naive Bayes is quick and effective for moderate-sized datasets, its performance is dependent on the assumption of feature independence and the quality of features chosen. As a result, when using Naive Bayes in machine learning models, these factors must be carefully considered.

The Bayes Theorem is used to represent the Naive Bayes.

Let X is the class and Y is the vector of the feature.

$$P(X|Y) = \frac{P(Y|X).P(X)}{P(Y)} \quad (1)$$

3.2.1.2. Decision Tree

A decision tree is a popular machine learning model that can handle both classification and regression tasks. It works by dividing the dataset into smaller groups iteratively based on the most influential feature at each stage of the process. A decision tree uses a recursive process to divide a dataset into subsets.

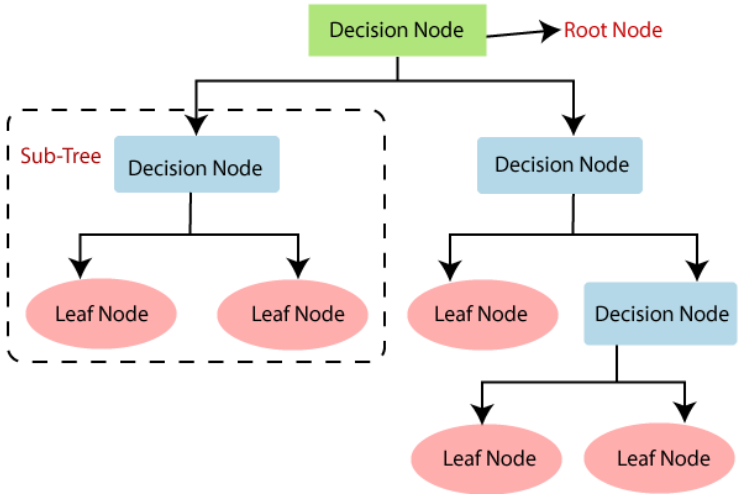

Figure 3. Decision Tree (Jijo & Abdulazeez,2021)

It begins with a root node that represents the full dataset, then chooses the most influential feature based on factors such as information gain and splits the data accordingly. This is repeated for each subset, building decision nodes and branches until stopping requirements are satisfied, such as maximum depth or a minimum amount of data points. The terminal points, also known as leaf nodes, represent the final decision. Follow the decision path from the root to a leaf based on feature values to anticipate a new data point. The interpretability of decision trees is appreciated because it provides insights into decision-making.

3.3. Performance Layer

In this layer, the performance of the proposed model is measured using different methods.

3.3.1. Accuracy

Accuracy is a way of checking how often the model gets its predictions right. Specifically, in classification tasks, accuracy is determined by the number of correct predictions divided by the total number of predictions. The Mathematical formula of Accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

In simpler terms, accuracy tells how well the model performs by looking at how many times it predicts things correctly compared to the total predictions it makes.

3.3.2. Miss Rate

In machine learning, the miss rate is a measure of how many real positive instances the model predicts as negative. It is calculated as the number of false negatives divided by the total number of actual positive instances.

Mathematical form of Miss Rate:

$$Miss\ Rate = \frac{FN}{FN+TP} \quad (3)$$

In simpler terms, the miss rate shows the percentage of positive cases that the model fails to identify.

3.3.3. Sensitivity

In machine learning, sensitivity, also known as recall or true positive rate, is a model that identifies all the actual positive cases. Calculate sensitivity by dividing the number of true positives (instances correctly identified as positive) by the sum of true positives and false negatives (instances wrongly labelled as negative).

The Mathematical formula of sensitivity is:

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

It's an important metric, especially in situations where catching every positive case is crucial, such as when it comes to medical diagnosis, missing a positive instance can have catastrophic repercussions.

3.3.4. Specificity

In machine learning, specificity, the actual negative rate, often known as the ability of a model to accurately recognize negative instances. The specificity measure is determined by the ratio of true negatives (correctly identified negative instances) to the sum of true negatives and false positives (instances wrongly classified as positive).

The Mathematical formula of specificity is:

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

3.4. Validation Phase

After that the validation phase starts, where the model is tested whether it is predicted accurately for further recommendation or deletion. The trained dataset is imported into the database. The proposed model is checked in the validation process to help the identification of patients with Cancer, Diabetes and Cystic fibrosis. If the model is trained successfully, it identifies the relevant patients of relevant diseases in Yes or No form. If the proposed model does not identify the patients, then the parameters are terminated.

A model-based classification method in healthcare provides benefits such as faster and more accurate diagnosis, improved patient outcomes, and more effective resource utilization. The suggested methodology quickly identifies individuals who require additional testing and therapy, allowing for early intervention for better results. However, the model's efficacy is dependent on the accuracy and precision of the validation data. These elements must be carefully considered during validation to ensure reliability. To ensure accurate and dependable predictions in real-world settings, ongoing monitoring and optimization are required.

4. Results and Discussions

4.1. Tools and Techniques

The purpose model of predicting the multi-factorial genetic disorders, imitation based on machine learning is carried out MATLAB R2023a. The dataset has three inputs and two outputs feature.

4.2. Result of the Proposed Model

The total dataset consists of 31,548 and the training dataset comprises of 22083 and the testing dataset comprises of 9465.

Table 2. Comparison of the training and testing accuracy

	Training	**Testing**
Accuracy	96.95%	94.25%
Miss Rate	3.05%	5.75%

Accuracy is the percentage of true predictions made by the model. The Miss Rate represents the proportion of incorrect predictions. The model achieved a training accuracy of 96.95% and a testing accuracy of 94.25%.This result indicates that the model's ability to predict is more accurate when it is tested with new data.

In this study, the model's accuracy is evaluated in both the training and testing phases. Accuracy refers to the proportion of correct predictions made by the model. It is an essential metric in evaluating the performance of a classification model. The comparison of training and testing accuracy is crucial as it helps to determine the model's ability to generalize well to new data. Ideally, the testing accuracy should be similar to the training accuracy, demonstrating that the model has discovered the underlying patterns and can make accurate predictions on unseen data.

The training accuracy in this study varies between 1.2 and 0.2, while the testing accuracy also ranges between 1.2 and 0.2. Let’s explore the concepts of training and testing accuracy. Training accuracy refers to the model's performance on the training data, which the model has been trained on. High training accuracy indicates that the model has learned the patterns present in the training data and can make accurate predictions on it. Testing accuracy, on the other hand, refers to the model's performance on the testing data, which the model has not seen before. It provides an unbiased estimate of the model's ability to generalize to new data.

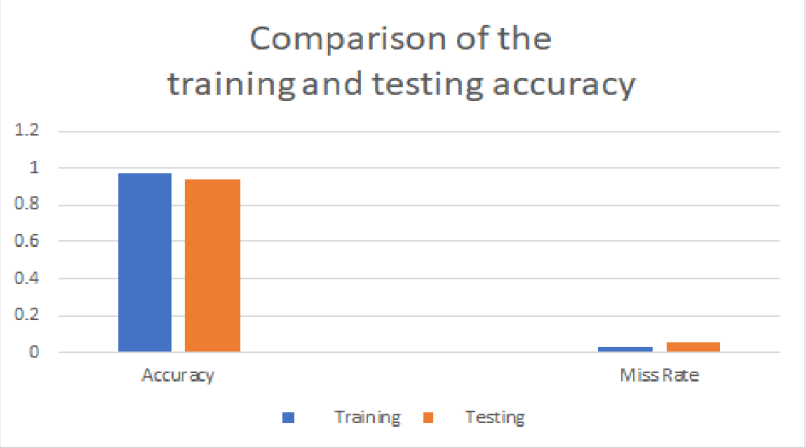

Figure 4. Graph of comparison of training and testing accuracy

4.3. Confusion Matrix

The confusion matrix is a visualization tool that allows us to analyze the performance of a classification model. In this case, we have a model that predicts whether a patient has disease or not, based on symptoms. The matrix consists of four quadrants, representing the different combinations of predicted and actual results:

- True Positive (TP): The model correctly predicted that the patient has disease.
- True Negative (TN): The model correctly predicted that the patient does not have disease.
- False Positive (FP): The model incorrectly predicted that the patient has disease.
- False Negative (FN): The model incorrectly predicted that the patient does have disease.

Table 3. Confusion matrix for Training

N=22083	Osymptomatic	Oasymptomatic
Isymptomatic=14353	13916	437
Iasymptomatic=7730	236	7494

The given text appears to be a confusion matrix for a binary classification problem, where the goal is to predict whether an individual is symptomatic or asymptomatic based on some data.

The true positives (13916) represent the number of samples that were correctly classified as symptomatic. The false positives (236) represent the number of samples that were incorrectly classified as symptomatic, while they were actually asymptomatic. The false negatives (437) represent the number of samples that were incorrectly classified as asymptomatic, while they were actually symptomatic. The true negatives (7494) represent the number of samples that were correctly classified as asymptomatic

In the confusion matrix for the training set, the predicted values are represented by the columns, while the actual values are represented by the rows. The number of cases in each quadrant represents the frequency of that particular combination.

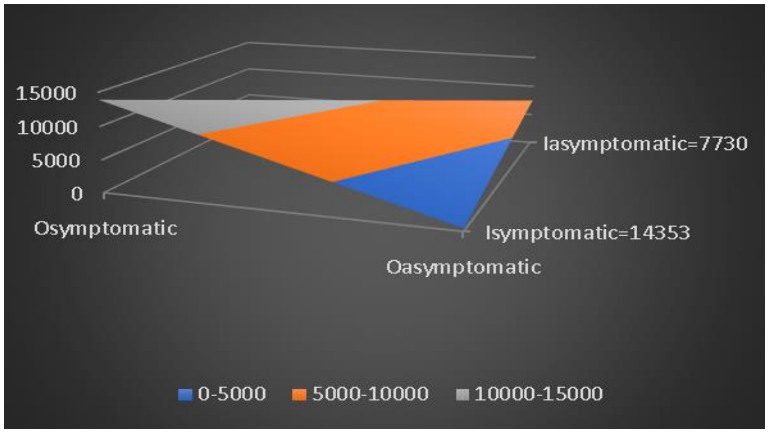

Figure 5. Confusion Matrix for raining

In this specific confusion matrix, we can observe that the model performed well in correctly identifying the patients who had the disease. However, it incorrectly predicted that some patients did not have the disease. Overall, the confusion matrix provides a clear visualization of the model's performance

Table 4. Confusion Matrix for Testing

N=9465	Osymptomatic	Oasymptomatic
Isymptomatic=6152	5798	354
Iasymptomatic=3313	190	3123

The given text appears to be a confusion matrix for a binary classification problem, where the goal is to predict whether an individual is symptomatic or asymptomatic based on some data.

From the table, we can see that the classifier correctly identified 3313 individuals as asymptomatic (true positives), but it also incorrectly identified 190 individuals as asymptomatic who were actually symptomatic (false positives). Similarly, the classifier correctly identified 5798 individuals as symptomatic (true negatives), but it also incorrectly identified 354 individuals as symptomatic who were actually asymptomatic (false negatives).

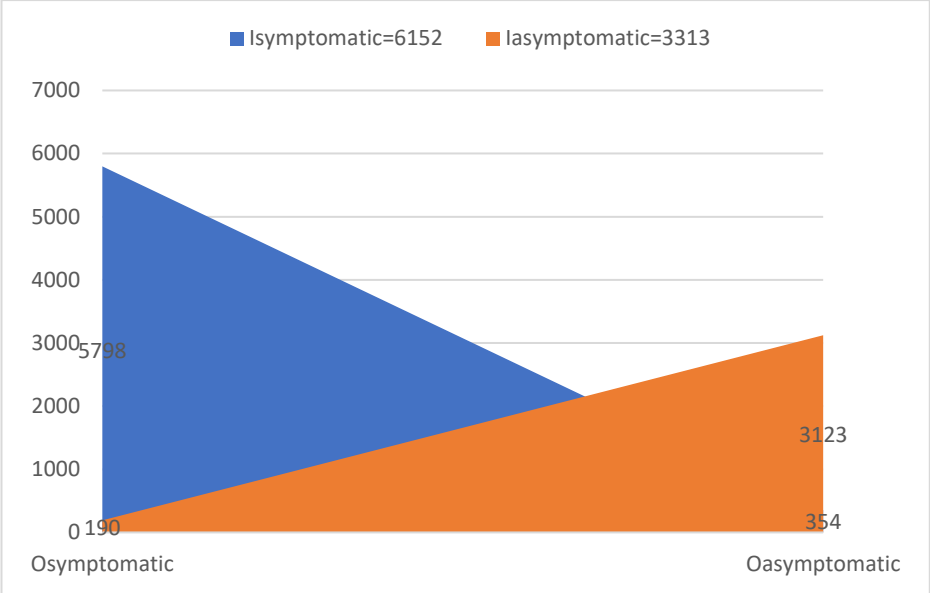

Figure 6. Confusion Matrix for Testing

A confusion matrix provides insights into the effectiveness of the test. The matrix categorizes patients into true positives (TN), true negatives (FP), false positives (FN), and false negatives (TN).

The testing model was able to accurately diagnose 6152 cases of symptomatic infection, making it 99.44% effective in this category. On the other hand, the model was only 94.28% effective in correctly diagnosing the presence of a disease in asymptomatic cases.

5. Conclusion

Genetic disorders have always been so less predicted yet problematic for the patient suffering from them. The importance of detecting these multifactorial genetic disorders by using machine learning is a very useful and necessary thing in this century when we have advanced technologies and tools around the world of artificial intelligence. This study used disorders-based combined datasets from Kaggle to identify selected diseases i.e. diabetes, cystic fibrosis and cancer with effective parameters i.e. white blood cells, maternal illness, radiation, birth asphyxia and folic acid deficiency. In this study, 43 attributes and 9467 rows have been used—the proposed model of multifactorial genetic disorders trained with Naïve Biased and Decision Tree Algorithm. According to the results, it is clarified that the proposed model is far more accurate than previously used models for detecting diseases. In this model, we have emphasized single gene-based different genetic diseases along with multifactorial genetic disorders.

References

1. Okser, S., Lehtimäki, T., Elo, L.L., Mononen, N., Peltonen, N., Kähönen, M., Juonala, M., Fan, Y.M., Hernesniemi, J.A., Laitinen, T. and Lyytikäinen, L.P., 2010. Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young Finns study. PLoS genetics, 6(9), p.e1001146.
2. Parsa, N., 2012. Environmental factors inducing human cancers. Iranian journal of public health, 41(11), p.1.
3. Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A. and Ferrari, R., 2018. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in bioinformatics, 19(2), pp.286-302.
4. Snell, D.M. and Turner, J.M., 2018. Sex chromosome effects on male–female differences in mammals. Current Biology, 28(22), pp.R1313-R1324.
5. Ohno, S., 2013. Sex chromosomes and sex-linked genes (Vol. 1). Springer Science & Business Media.
6. Tyagi, A.K. and Chahal, P., 2022. Artificial intelligence and machine learning algorithms. In Research anthology on machine learning techniques, methods, and applications (pp. 421-446). IGI Global.
7. Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), p.160.
8. Asthana, P. and Hazela, B., 2020. Applications of machine learning in improving learning environment. Multimedia big data computing for IoT applications: concepts, paradigms and solutions, pp.417-433.
9. Sarker, I.H., 2022. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science, 3(2), p.158.
10. Clune, J., 2019. AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. arXiv preprint arXiv:1905.10985.
11. Zhou, L., Pan, S., Wang, J. and Vasilakos, A.V., 2017. Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, pp.350-361.
12. Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.
13. Muhammad, I. and Yan, Z., 2015. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. ICTACT Journal on Soft Computing, 5(3).
14. Haldorai, A., Ramu, A., Murugan, S., Haldorai, A., Ramu, A. and Murugan, S., 2019. Artificial intelligence and machine learning for future urban development. Computing and Communication Systems in Urban Development: A Detailed Perspective, pp.91-113.
15. Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), p.160.
16. Barros, R.C., Basgalupp, M.P., De Carvalho, A.C. and Freitas, A.A., 2011. A survey of evolutionary algorithms for decision-tree induction. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(3), pp.291-312.
17. Ren, J., Lee, S.D., Chen, X., Kao, B., Cheng, R. and Cheung, D., 2009, December. Naive bayes classification of uncertain data. In 2009 Ninth IEEE international conference on data mining (pp. 944-949). IEEE.
18. Nasir, M.U., Khan, M.A., Zubair, M., Ghazal, T.M., Said, R.A. and Al Hamadi, H., 2022. Single and mitochondrial gene inheritance disorder prediction using machine learning. Comput. Mater. Contin, 73, pp.953-963.
19. Gudhe, N.R., Behravan, H. and Mannermaa, A., 2023. A GRAPH-BASED REPRESENTATION LEARNING APPROACH FOR BREAST CANCER RISK PREDICTION USING GENOTYPE DATA.
20. Ghazal, T.M., Munir, S., Abbas, S., Athar, A., Alrababah, H. and Khan, M.A., 2023. Early detection of autism in children using transfer learning. Intelligent Automation & Soft Computing, 36(1), pp.11-22.

21. Sothe, C., De Almeida, C.M., Schimalski, M.B., La Rosa, L.E.C., Castro, J.D.B., Feitosa, R.Q., Dalponte, M., Lima, C.L., Liesenberg, V., Miyoshi, G.T. and Tommaselli, A.M.G., 2020. Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. GIScience & Remote Sensing, 57(3), pp.369-394.
22. Ahadi, A., Lister, R., Haapala, H. and Vihavainen, A., 2015, August. Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the eleventh annual international conference on international computing education research (pp. 121-130).
23. Sajjad, R., Khan, M. F., Nawaz, A., Ali, M. T., & Adil, M. (2022). Systematic analysis of ovarian cancer empowered with machine and deep learning: a taxonomy and future challenges. Journal of Computing & Biomedical Informatics, 3(02), 64-87.
24. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, pp.104-116.
25. Pope, W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L., Al-Atrache, Z., Alcoser, T.A., Alexander, L.M., Alfano, M.B., Alford, S.T., Amy, N.E. and Anderson, M.D., 2011. Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. PloS one, 6(1), p.e16329.
26. Alanazi, R., 2022. Identification and prediction of chronic diseases using machine learning approach. Journal of Healthcare Engineering, 2022.
27. Gulande, M.P. and Awale, R.N., 2021. Systematic Study of Gen Profiles Analysis Methods in Disease Classification. Annals of the Romanian Society for Cell Biology, pp.15361-15371.
28. Algamal, Z.Y. and Lee, M.H., 2019. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Advances in data analysis and classification, 13(3), pp.753-771.
29. Khan, P., Kader, M.F., Islam, S.R., Rahman, A.B., Kamal, M.S., Toha, M.U. and Kwak, K.S., 2021. Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. Ieee Access, 9, pp.37622-37655.
30. Le, D.H., Xuan Hoai, N. and Kwon, Y.K., 2015. A comparative study of classification-based machine learning methods for novel disease gene prediction. In Knowledge and Systems Engineering: Proceedings of the Sixth