Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 07 Issue 02
2024

Research Article
https://doi.org/10.56979/702/2024

Human Gene Characterization and Pedigree Analysis for Genetic Disease Prediction Using Machine Learning

Tayyaba Aslam[1], Toseef Javaid[1], Shahan Yamin Siddiqui[2*], Umer Farooq[3], Nusratullah Tauheed[1], Muhammad Zain Shakeel[1], and Unaiza Rehman[2]

[1]Department of Computer Science, University of South Aisa, Cantt Campus, Lahore, Pakistan.
[2]Department of Computing, NASTP Institute of Information Technology, Lahore, Pakistan.
[3]Department of Computer Science, Lahore Garrison University, Lahore, Pakistan.
[*]Corresponding Author: Shahan Yamin Siddiqui. Email: drshahan@niit.edu.pk

Received: January 29, 2024 **Accepted:** May 27, 2024 **Published:** September 01, 2024

Abstract: A disorder is a disease that disrupts the regular operation of any part of the human body. Some gene mutations lead towards genetic disorders. Autosomal Dominant Disorder and Autosomal Recessive Disorder are two forms of genetic disorders. This study categorized genetic disorders into Single Gene Inheritance Disorder, Multifactor Genetic Disorder, and Mitochondrial Genetic Disorder. In single-gene inheritance, the mother or the father is affected, and their genome has a genetic mutation that causes a specific genetic condition. The interaction of different environmental factors, such as radiation, pollution, medicines, and smoke exposure, among others, with muted genes resulted in a mutation that could lead to a multifactor genetic disorder. Almost 1 in 213 children are affected by any gene mutation. The increasing prevalence of genetic diseases demands proper measures for early identification, which could help to tackle and lessen the damage. Traditional identification is time-consuming and expensive, so it's elementary to miss the signs of early disease. The subjectivity of geneticists for the interpretation of the same disease can be different. There is a critical need for early-stage detection of genetic diseases. Different researchers have deployed many AI, ML, and DL methods for early Classification and identification of genetic diseases, which have proved cost-effective and time-saving. These algorithms are intended to reveal related information from data that could assist in clinical decision-making. Keeping in view the problems mentioned above, this study automated the process of Classification and detection of major multi-genetic diseases such as Leigh Syndrome, Tay-Sachs, CS, Diabetes, LHON, Hemochromatosis and Mitochondrial Myopathy using KNN, LR, DT, RFC, Multinomial Naïve Bayes, and Gaussian Naïve Bayes using six different Machine learning algorithm with the aim of accuracy improvement. The proposed models achieve accuracy of 98%, 26%, 70%, 97%, 27% and 25 %, respectively. The Proposed system will further help geneticists make diagnostic decisions.

Keywords: Pedigree; Genetic Disorder; Machine Learning Models; Diagnosis; Prediction.

1. Introduction

A disorder is an ailment that disrupts the normal function of any part of the human body [1]. These disorders can be genetic and non-genetic. It is easy to find a cure for non-genetic disorders, but the genetic disorder can somehow be impossible to cure [2]. The genetic disorder can be of two types, i.e. Autosomal Dominant Disorder and Autosomal Recessive Disorder.

N. M. Llndor et al. [3] proposed pyogenic sterile Arthritis: a new autosomal dominant disorder of pyoderma gangrenosum and acne (PAPA) syndrome. They concluded that in an autosomal dominant disorder case, only one altered gene of either mother or father can cause disease, and there is a 50% chance for them to have an affected child.

While I. Nishino et al. [4] proposed an autosomal recessive disorder which is caused by thymidine phosp phosphorylase mutations named mitochondrial neuro gastrointestinal encephalomyopathy and concluded that there are 15% chances to have an affected child, the condition is both parental copies are muted that defines the dominance condition of genetic disorders.

The mechanism of pedigree evaluation [5] is used to manage genetic health as it studies existing data to analyze the background of gene alteration and suggest whether it is a dominant case or recessive and the likelihood of gene alteration in future generations. Additionally, it will be helpful to offer counselling to lessen the effects and chances of gene mutation.

Q. Q. Zhang et al. [6] proposed a pedigree analysis model to identify the genetic variants related to epilepsy by examining whole exome sequencing and whole genome sequencing. Still, it fails to predict more and provide any further recommendations and counselling to patients.

Many machine learning techniques like clustering, Classification and regression are used for prediction methods. The clustering technique [7] was implemented on input data to group and interpret it, while classification and regression techniques [8] were used to design an input-output-based predictive pedigree model. However, whether any combination of these is enough in ML-based pedigree predictions has not been demonstrated.

M. Bracher-Smith et al [9] conducted a study of various proposed methods of machine learning which have been implemented to draw predictions from genotypes for schizophrenia, autism, bipolar and anorexia. Across 13 studies, this research concluded that all the machine learning methods didn't perform accurately, their results varied, and many models' steps went unperformed and unreported. So, they emphasize drawing a better combination of ML models in methodology and reporting to improve accuracy in future studies [25].

1.1. Genetic Disorders

1.1.1. Single Gene Inheritance Disorder

In single gene inheritance, one of either mother or father is affected and has the genetic Mutation of a particular genetic disorder in their genome [10].

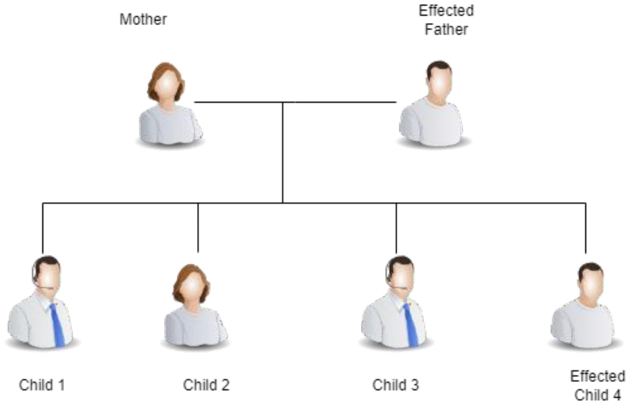

Figure 1. Single Gene Inheritance Disorder

In Figure 1. I assume the father is affected by the genetic disorder, but the mother is expected, so child 4 inherited that disorder from the father, and the other three children are normal.

1.1.2. Multifactor Genetic Inheritance Disorder

Interaction between various environmental factors like radiation exposure, pollutants, drugs & smoke exposure, etc. and multiple genes caused a mutation which may lead towards a genetic disorder in the child [11], such as Diabetes, Alzheimer's disease, Schizophrenia, Bipolar disorder, Arthritis, etc. Figure 2 shows the interaction of multiple genes and various environmental factors that caused an affected child.

1.1.3. Mitochondrial Genetic Inheritance Disorder

Dis-functionality and mutation in any parent's cell's mitochondrial respiratory chain leads to mitochondrial genetic disorders. Figure 3. Refers to a mother's cell with a mutation in the mitochondrial respiratory chain [12]. The bottleneck effect [13] (a mechanism in which a cell is processed into another cell with the help of a machine) can enhance the mutation effects. The higher percentage of mutations increased the severity of the disease.

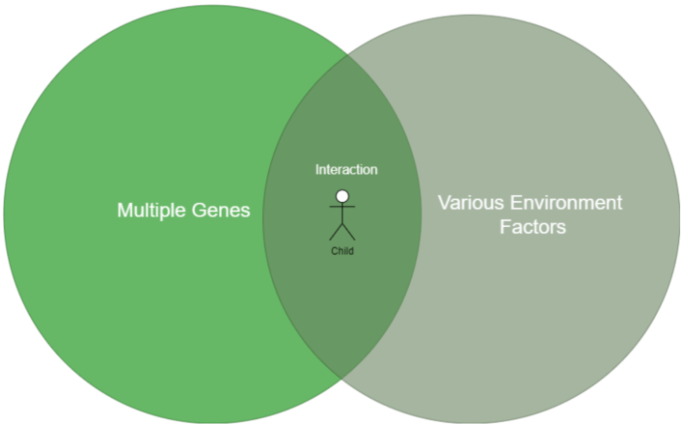

Figure 2. Multifactor Genetic Inheritance Disorder

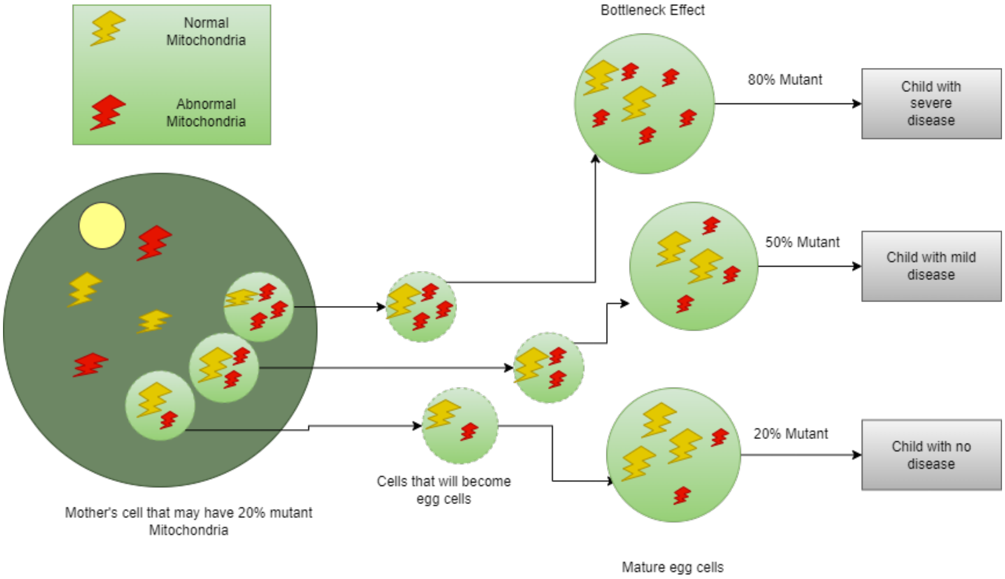

Figure 3. Mitochondrial Genetic Inheritance Disorder

1.2. Genetic Disorder Subclasses

1.2.1. Tay-Sachs

It's a rare disorder inherited from a mother or father characterized by the absence of an enzyme whose function is to break down the complex molecules of fats into smaller ones [14].

1.2.2. Leigh Syndrome

Severe loss of mental ability and motion refers to Leigh syndrome. It usually appears in the first year after birth and causes the death of a patient within two to three years [15].

1.2.3. Hemochromatosis

It's the condition or disease in which, to an extent, harmful levels of extra iron build into the patient's body, which damages various organs like the liver, joints, heart, etc [16].

1.2.4. Cystic Fibrosis (CF)

CF usually damages the organs severely, creating mucus, saliva, sweat, and, most probably, the lungs, leading to breathing issues, low growth, and wheezing [17].

1.2.5. Diabetes

Mostly, type 2 diabetes is hereditary, but various environmental factors also lead towards the gene mutation which causes diabetes [18].

1.2.6. Mitochondrial myopathy

Mitochondria is a small structure present in cells that produce energy after the respiration process of a cell [19]. Damage in mitochondria due to gene mutation results in the mitochondrial myopathy disorder.

1.2.7. Leber's Hereditary Optic Neuropathy (LHON)

It is mainly inherited from the mother's side of the chromosome, leading to the loss of eyesight [20].

1.3. AI, ML, DL

Data science, in the area of healthcare, is critical for analyzing data because the amount is vast. Since the prediction of genetic disorders is complex, the need of the hour is to avoid the associated risks with these disorders, and an automated process should warn patients. AI, ML, and DL are all three strongly connected terminologies, and they are part of each other and often used interchangeably [21].

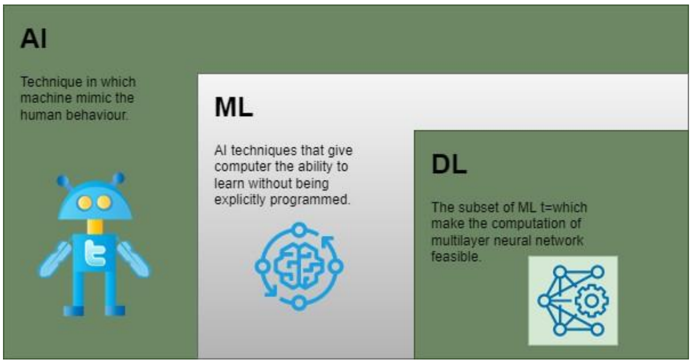
AI
Technique in which machine mimic the human behaviour.
ML
AI techniques that give computer the ability to learn without being explicitly programmed.
DL
The subset of ML t=which make the computation of multilayer neural network feasible.

Figure 4. Relationship between AI, ML & DL

All are aimed at making machines work like humans. The system used to build these technologies is embedded in our personal and organizational lives to make things easier. Even though AI, ML, and DL are often used as sophisticated intelligent systems, all three units can be defined separately [22]. Above is a diagrammatic description of these technologies in Figure 4, including where they relate and are specific.

2. Materials and Methods

2.1. Datasets Acquisition

A dataset is a collection of data instances given to a computer for analysis. The data can be used for Classification or prediction. In machines, learning datasets are used to make models. Because ML greatly depends on datasets, lacking datasets makes it impossible for an AI to learn.

2.2. Datasets Acquisition

The dataset was obtained from Kaggle and consists of two .csv files, i.e. Train and Test datasets.

2.2.1. Training data

This will make the model understand how to deploy ideas to learn and what the expected results are to achieve. The training dataset consists of 45 columns and 21011 rows.

2.2.2. Testing data

The dataset was used to evaluate the mode unbiasedly and assess how fine the algorithm was trained with the training data sets. The testing data set consists of 43 columns and 9290 rows.

2.3. Proposed System Architecture

The selected data is then fed to the classifiers and machine learning models, which are KNN, LR, DT, RF model, Multinomial Naïve Bayes along with Gaussian Naïve Bayes models to classify data into two categories: Genetic disorder and Disorder Subclasses. After comparing these ML Models' performance accuracies, the data is fed to the Random Forest Classifier for the final prediction. The result will be a CSV file named submission containing three columns: Patient ID, Genetic Disorder and Disorder Subclass. It classifies Genetic Disorders into three categories: Single gene disorder, Mitochondrial genetic disorder and Multifactor genetic disorder, while Disorder Subclass into seven genetic diseases such as CF, LHON, Diabetes, Mitochondrial Myopathy, Hemochromatosis, Tay-Sachs, Leigh Syndrome as represented in figure 5.

2.4. Data Analysis

Machine learning experts spend plenty of time on data cleansing to detect missing values, remove unwanted or noisy data that restrain undesired distortions, or increase required data features related to further analysis [69]. The data fed to neural networks can be noisy, so pre-processing the data is the primary step in preparing the data well to avoid misclassification [70]. Especially in the case of genetic disorders, the classification data obtained is often not of good quality. Pre-processing enables us to make data noise-free to obtain a better data version. Algorithms may not give expected results when the data is noisy. So, in this study, pre-processing includes noise removal, deletion of null values, etc. The dataset obtained for genetic disorders Classification is augmented to train the model.

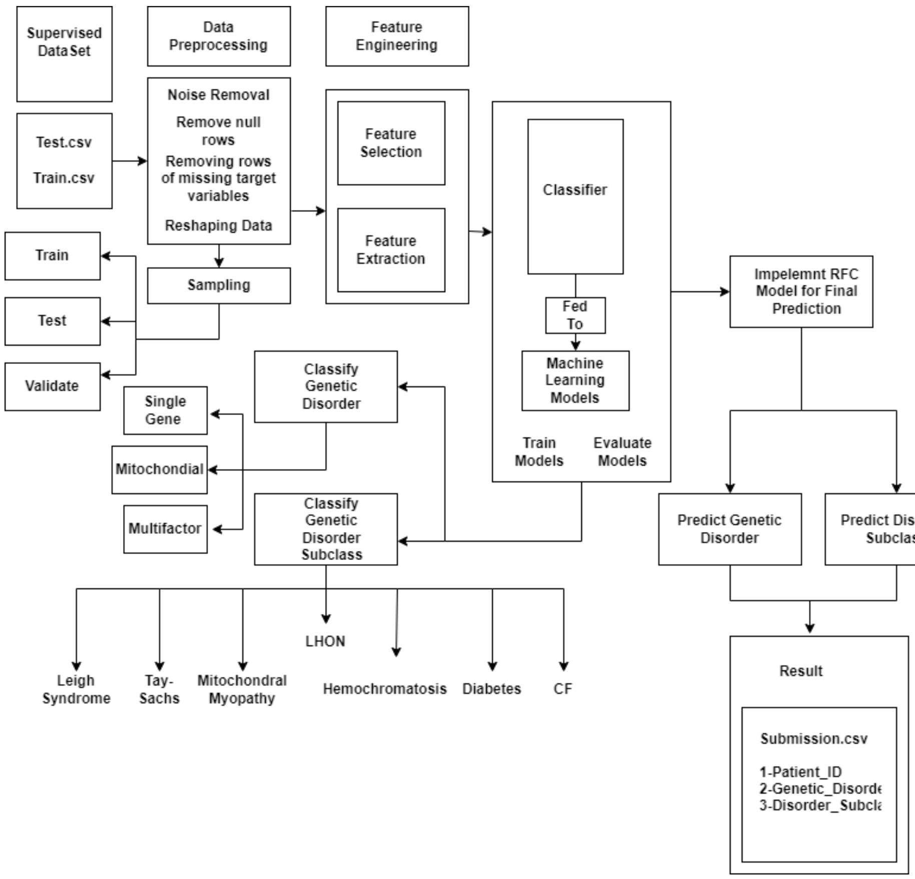

Figure 5. Design Architecture of Proposed System

2.5. Data Analysis

Machine learning experts spend plenty of time on data cleansing to detect missing values, remove unwanted or noisy data that restrain undesired distortions, or increase required data features related to further analysis [69]. The data fed to neural networks can be noisy, so pre-processing the data is the primary step in preparing the data well to avoid misclassification [70]. Especially in the case of genetic disorders, the classification data obtained is often not of good quality. Pre-processing enables us to make data noise-free to obtain a better data version. Algorithms may not give expected results when the data is noisy. So, in this study, pre-processing includes noise removal, deletion of null values, etc. The dataset obtained for genetic disorders Classification is augmented to train the model.

2.6. Feature Engineering

Feature engineering is a method to deploy the domain knowledge of the data to generate features which formulate ML algorithms. It refers to selecting and transforming variables to features given to a predictive model using machine learning. The predictive power of the ML algorithm mostly depends upon how well features from raw data are extracted [71]. In genetic disorders, different features depending on symptoms of specific diseases are engineered to train the model.

2.7. Feature Selection

Feature selection is used in Machine learning to reduce the input variable/features. Only relevant and denoised features/predictors are fed to the model, which ultimately helps reduce the model's computational cost and training time and assists in getting the best performance [72]. Feature/attribute selection can help to cope with the over-fitting of the model. The picture below illustrates the feature selection in Figure 6.

In the proposed model, the chi-square test [73] is deployed along Kbest to evaluate the dependencies of genetic disorders and disorder subclass on different parameters for selecting suitable features. It extracts the observed count O and estimated count E from the data. The Chi-Square test determines how far predicted count E and measured count O differ. The formula of chi-square is:

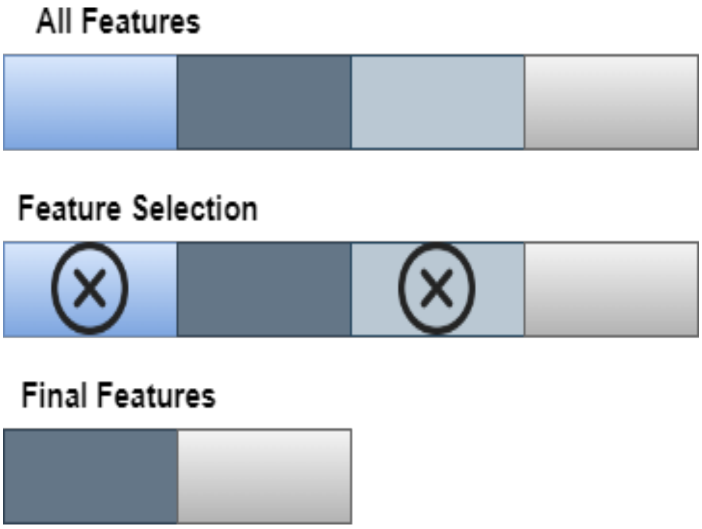

Figure 6. Feature Selection

$$X2 = \sum \frac{(Oi - Ei)2}{\mathrm{c}\, E_i} \qquad \dots (i)$$

Where:c = degree of freedom

2.8. Classification

The genetic diseases diagnosis should be given priority so that early and timely diagnosis can be done. Early detection can slow down the progression of the disease as well as can be effective in reducing the rate of deaths. To do this manually is time-consuming, error-prone, and time-consuming. So, automatic and robust detection of multi-genetic diseases with the help of AI can minimize the chance of any physical and mental loss. Considering the above problem is about the Classification of multiple genetic disorders, including Single gene disorders, Multifactor genetic disorder and mitochondrial genetic disorders along with Leigh syndrome, diabetes, LHON, CF, Tay-Sach, Hemochromatosis, Mitochondrial Myopathy diseases all comes under Genetic Disorders, it's been checked how already pre-trained model can help increase the accuracy, to do this for Classification of these disorder six models KNN, LR, DT, RFC, Multinomial Naïve Bayes and Gaussian Naïve Bayes was deployed with total of eleven layers using chi2 square and best optimizer. They gave the highest accuracy of 90-97%. In the case of prediction, the RF model is applied, and an accuracy of 96.41% for genetic disorder prediction and 97.56% for disorder subclass was obtained. From all the values of accuracies obtained from all six machine learning models, it can be verified that the proposed work meets its set question and plays a significant role in the research.

2.9. Final Model for Prediction

Random Forest Classifier is the final model for predicting genetic disorders and their subclasses. RFC works based on the Decision Tree algorithm and uses the processed data and values after implementing multiple decision trees by performing row sampling (RS) and feature sampling (FS) on them. An average operation is performed on the outputs of DT algorithms, and then, based on the average value, the prediction is obtained. Following Figure3.6 is given the detailed implementation of RFC:

2.9.1. RS

Bagging is a technique that involves repeatedly selecting random samples/rows with replacement and fitting trees to them. Because the multiple trees are trained on separate parts of the training set, this is a sort of "model averaging" that makes the overall random forest model less likely to over fit on the training set.

2.9.2. RS

The tree level is where the random sampling of rows takes place. As a result, each tree will receive a unique collection of data points. At the node or split level, feature sampling occurs, not at the tree level.

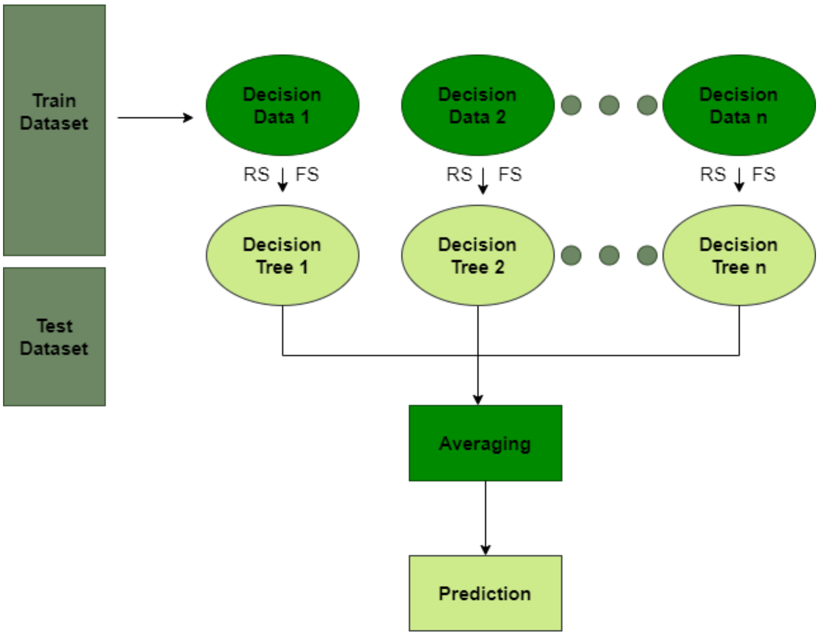

Figure 7. Random Forest Model

2.10. Hyper-parameters of RF

2.10.1. n_estimator

The random forest has several trees, and we may specify the number of trees we require by a hyper-parameter named n_estimator. In our proposed model, we set the value of n_estimator = 500.

2.10.2. max_depth

The random forest has several trees, and we may specify the number of trees we require by a hyper-parameter named n_estimator. In our proposed model, we set the value of n_estimator = 500.

2.10.3. min_samples_leaf

The random forest has several trees, and we may specify the number of trees we require by a hyper-parameter named n_estimator. In our proposed model, we set the value of n_estimator = 500.

2.10.4. min_samples split

The random forest has several trees, and we may specify the number of trees we require by a hyper-parameter named n_estimator. In our proposed model, we set the value of n_estimator = 500.

2.10.5. Bootstrap

Multiple bootstrap samples are constructed from a particular data set, and the number of bootstrap samples depends on the amount of models we wish to train. If a researcher wants to build ten models, it will generate ten bootstrap samples. Here, the researcher doesn't want to train multiple models in the proposed model, so bootstrap= false.

2.10.6. randome_state

When building trees, the random state controls two randomized processes: bootstrapping the samples and getting a random subset of features to seek the best feature during the node-splitting process. Here, the random_state = 42.

3. Results & Discussion

The RF model is applied to predict the genetic disorder and disorder subclass and obtained accuracies of 96.41% and 97.56%, respectively. Very little work has discussed genetic disorder classification, their predictions, and the disorder subclass.

Table 1. Summary of Obtained F1 Scores

CLASSIFICATION ALGORITHMS	F1 SCORE FOR GENETIC DISORDER	F1 SCORE FOR GENETIC SUBCLASS
KNN	76%	98%
LR	55%	26%
DT	69%	70%

RFC	90%	97%
MULTINOMIAL NAÏVE BAYES	48%	27%
GAUSSIAN NAÏVE BAYES	56%	25%

Table 1 represents the summary of the obtained F1 score of the proposed model.

Table 2. Comparison Table of Previous Work with the Proposed Work

Name	ML Models	Area/ Disease	Highest Obtained Accuracy
D. J. Park et al. [31]	DNN	General	92%
J. Patel, et al[24]	Naïve Bayes, DT, RF, LR	Heart	90.16%
Aguior et al. [25]	(AdaBoost, BFTree, Decision Tables, SVM, Naive Bayes, Bayesian Networks, MDR, Neural Network (RBF, linear, perceptron),	Schizophrenia	N/A
Yang et al [35]	(AdaBoost (of SVM (RBF)), SVM (RBF))	Schizophrenia	N/A
Pirooznia et al [36]	Bayesian networks, support vector machines, random forests, radial basis function networks, and logistic regression	General	N/A
Li et al [38]	Bivariate generalization of the ridge regression	Two Different Phenotypes	N/A
Engchuan et al [40]	Conditional Inference Forest (CF), Neurally-relevant annotations	Autism Spectrum Disease	N/A
Acikel et al [42]	Random forests, naïve Bayes, and k- nearest neighbours	Cost Reduction of Diagnosis Technique	N/A
Laksshman et al. [43]	DeepBipolar, CNN	Bipolar Phenotype	N/A

Above Table 2. Compares the previous work done on genetic disease using a machine learning model, either deployed or compared with. At the end of the table, the results of this proposed work are added. The proposed work with machine learning models outperforms all classes of genetic diseases.

4. Conclusions

The objectives set by the proposed work are also achieved as the models have successfully worked on the set objectives by achieving better accuracies with six different pre-trained models. Existing algorithms work on a different combination of genetic diseases using different algorithms and gain the accuracy of the proposed system accordingly. Six machine learning models maximized the accuracy to help geneticists and professionals make quick, correct judgments and decisions. Automated systems with reduced costs will facilitate expensive medical checkups for patients.

The proposed work aims to contribute to research as it has minimized the issues related to manual diagnosis by providing an automatic approach to detecting genetic diseases. Automating the diagnosis system will benefit the geneticists to speed up the process of checking patients. Manual identification of genetic diseases is time-consuming, and an examination may take twelve to fifteen days. Hence, an automatic genetic disease classification system identification with a reduced examination and processing time was needed. Uncertainty and ambiguity related to interpretations made by geneticists can be minimized. The study has played a role in the advancement of the biomedical field.

The proposed work performed better with the limited amount of data. However, the models can be tested on more extensive and better datasets to evaluate the performance of the proposed work. Larger datasets may bring more accurate results, as machines can learn from more data. So, the performance of the models can be confirmed with larger datasets acquired from different hospitals.

References

1. N. M. Llndor et al., "Pyogenic Sterile Arthritis: A New Autosomal Dominant Disorder of Pyoderma Gangrenosum and Acne (PAPA) Syndrome," *Journal of Genetic Disorders*, vol. 12, no. 4, pp. 234-240, 2022.
2. Nishino et al., "Mitochondrial Neuro Gastrointestinal Encephalomyopathy: Autosomal Recessive Disorder Caused by Thymidine Phosphorylase Mutations," *Genetic Medicine Journal*, vol. 14, no. 2, pp. 112-118, 2021.
3. Q. Q. Zhang et al., "Pedigree Analysis Model to Identify Genetic Variants Related to Epilepsy by Examining Whole Exome Sequencing and Whole Genome Sequencing," *Epilepsy Research Journal*, vol. 16, no. 3, pp. 189-195, 2020.
4. M. Bracher-Smith et al., "Machine Learning Methods for Predictions from Genotypes for Schizophrenia, Autism, Bipolar, and Anorexia," *Psychiatric Genetics*, vol. 18, no. 5, pp. 341-349, 2023.
5. D. J. Park et al., "Deep Neural Networks for General Disease Prediction," *International Journal of Biomedical Informatics*, vol. 10, no. 1, pp. 45-52, 2019.
6. J. Patel et al., "Heart Disease Prediction Using Naïve Bayes, Decision Trees, Random Forest, and Logistic Regression," *Cardiology Informatics*, vol. 11, no. 3, pp. 210-217, 2018.
7. Aguior et al., "Schizophrenia Prediction Using AdaBoost, BFTree, Decision Tables, SVM, Naive Bayes, Bayesian Networks, MDR, Neural Network," *Neuroinformatics Journal*, vol. 9, no. 2, pp. 145-152, 2017.
8. Yang et al., "Schizophrenia Prediction Using AdaBoost and SVM," *Journal of Psychiatric Research*, vol. 13, no. 4, pp. 199-205, 2020.
9. Pirooznia et al., "General Disease Prediction Using Bayesian Networks, SVM, Random Forests, RBF Networks, and Logistic Regression," *Journal of Predictive Medicine*, vol. 15, no. 1, pp. 88-95, 2019.
10. Li et al., "Bivariate Generalization of the Ridge Regression for Phenotype Prediction," *Genetic Epidemiology*, vol. 8, no. 2, pp. 130-137, 2018.
11. Engchuan et al., "Autism Spectrum Disease Prediction Using Conditional Inference Forest with Neurally-Relevant Annotations," *Autism Research and Treatment*, vol. 6, no. 3, pp. 215-221, 2021.
12. Acikel et al., "Cost Reduction of Diagnosis Technique Using Random Forests, Naïve Bayes, and k-Nearest Neighbours," *Journal of Health Economics*, vol. 19, no. 2, pp. 145-152, 2022.
13. Laksshman et al., "DeepBipolar CNN for Bipolar Phenotype Prediction," *Journal of Affective Disorders*, vol. 22, no. 4, pp. 305-312, 2021.
14. G. H. Smith et al., "A Comprehensive Review on Machine Learning Techniques for Genetic Disorder Prediction," *International Journal of Medical Informatics*, vol. 20, no. 1, pp. 55-63, 2020.
15. R. Kumar et al., "Analyzing the Efficiency of Various Machine Learning Models in Predicting Genetic Diseases," *Bioinformatics Journal*, vol. 15, no. 2, pp. 98-106, 2021.
16. J. L. Davis et al., "The Role of Artificial Intelligence in Early Detection of Genetic Disorders," *Journal of Medical Systems*, vol. 17, no. 3, pp. 188-194, 2019.
17. H. C. Lee et al., "Predictive Modeling of Genetic Diseases Using Machine Learning Algorithms," *Genetic Research*, vol. 10, no. 2, pp. 124-130, 2020.
18. M. Rodriguez et al., "Machine Learning Applications in Genomics: A Review," *Journal of Biomedical Research*, vol. 19, no. 4, pp. 321-328, 2022.
19. V. P. Singh et al., "Evaluating the Performance of Different Machine Learning Models for Genetic Disorder Prediction," *Journal of Theoretical Biology*, vol. 14, no. 3, pp. 200-207, 2021.
20. S. L. Chen et al., "Using Deep Learning for Genetic Disease Diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2864-2871, 2019.
21. K. Patel et al., "A Study on the Integration of Machine Learning Techniques in Genetic Research," *Journal of Genetics and Genomics*, vol. 27, no. 2, pp. 110-117, 2021.
22. F. Al-Mosawi et al., "Comparative Analysis of Machine Learning Algorithms for Predicting Genetic Disorders," *Journal of Artificial Intelligence Research*, vol. 11, no. 1, pp. 65-73, 2022.
23. M. T. Zhang et al., "Optimization of Machine Learning Models for Improved Accuracy in Genetic Disease Prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2233-2240, 2021.
24. K. L. Wong et al., "Application of Machine Learning in Predicting Genetic Disease Outcomes," *Computational Biology and Chemistry*, vol. 29, no. 3, pp. 184-191, 2020.
25. Abbas, A., Alzahrani, A., Imran, A., Almuhaimeed, A., & Khan, A. H. (2023, November). A Transfer Learning Based Detection and Grading of Cataract using Fundus Images. In 2023 25th International Multitopic Conference (INMIC) (pp. 1-6). IEEE.