International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229273

Prediction of Genetic Disorders using Machine Learning

Sadichchha Naik, Disha Nevare, Amisha Panchal, Dr. Chhaya Pawar
Department of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, India

Article Info
Volume 9, Issue 3
Page Number : 395-401

Publication Issue
May-June-2022

Article History
Accepted : 20 April 2022
Published : 04 May 2022

ABSTRACT

A genetic disorder is a health condition that is usually caused by mutations in DNA or changes in the number or overall structure of chromosomes. Several types of commonly-known diseases are related to hereditary gene mutations. Genetic testing aids patients in making important decisions in the prevention, treatment, or early detection of hereditary disorders. With increasing population, studies have shown that there has been an exponential increase in the number of genetic disorders. Genetic disorders impact not only the physical health, but also the psychological and social well-being of patients and their families. Genetic disorders have powerful effects on families. Like many chronic conditions, they may require continual attention and lack cures or treatments. Low awareness of the importance of genetic testing contributes to the increase in the incidence of hereditary disorders. Many children succumb to these disorders and it is extremely important that genetic testing be done during pregnancy. In that direction, the project aims to predict Genetic Disorder and Disorder Subclass using a Machine Learning Model trained from a medical dataset. The model being derived out of a predictor and two classifiers, shall predict the presence of genetic disorder and further specify the disorder and disorder subclass, if present.

Keywords: Genetic disorder, Machine Learning, Medical dataset

I. INTRODUCTION

India is the sixth largest country in the world in terms of its geographical area and the second largest country in population density. The people of the country are diverse in terms of their social, linguistic, cultural, and racial backgrounds. Evolutionarily, the Indian subcontinent has been a corridor for different migratory waves arising from Africa, through land as well as coastline routes.

Genetic disorders impact not only the physical health, but also the psychological and social well-being of patients and their families. Genetic disorders have powerful effects on families. Like many chronic conditions, they may require continual attention and lack cures or 2 treatments. They have implications for the health of relatives, so a genetic diagnosis for one family member may mean other blood relatives are also at risk, even if they currently show no symptoms. In addition to the medical implications, genetic

disorders present emotional challenges and special reproductive implications.

Precision genomics-based medicine has arisen in the last decade to provide patients with personalised and effective treatment based on their genetic characteristics. To realise the full potential of precision medicine, researchers are aiming to capitalise on advances in genomics to further develop individualised medical care through increasingly accurate illness risk prediction models. Despite recent advances, the outcomes of polygenic risk score are still limited due to the present methodologies.

Machine learning algorithms, on the other hand, have improved the ability to predict the risk of complicated diseases. The ability of machine learning algorithms to handle multi-dimensional data accounts for this boost in predicting skills.

According to studies, the number of genetic disorders has increased exponentially as the population has grown. Genetic disorders are becoming more common due to a lack of understanding about the need of genetic testing. The project's goal is to use a Machine Learning Model trained on a medical dataset to predict Genetic Disorder and Disorder Subclass.

II. PURPOSE

With increasing population, studies have shown that there has been an exponential increase in the number of genetic disorders. Low awareness of the importance of genetic testing contributes to the increase in the incidence of hereditary disorders. In that direction, the project aims to predict Genetic Disorder and Disorder Subclass using a Machine Learning Model trained from a medical dataset.

The project commences with a thorough analysis of the information available and deriving key insights that lead to identification and association of medical data with the disease. The prime objective is to build a predictor and classifier model that can determine the probability of genetic disease based on the provided patient data.

III. LITERATURE REVIEW

To thoroughly search recent literature on deep learning applications in disease prediction, we carefully reviewed previous works, searched popular sites: Google Scholar, PubMed, IEEE Xplore, and PMC, and examined related online blogs and tutorials, such as GitHub (http://github.com/), Kaggle (http://www.kaggle.com/), and Cross Validated (https://stats.stackexchange.com/). Plenty of methods have been proposed in disease prediction using genomic data. Due to the large number of predictors (i.e., gene transcripts), the main approach in disease detection/prediction is to first obtain a subset of gene transcripts (e.g., a few top gene transcripts in transcript-wise tests) ora subset of representations of gene transcripts (e.g., a few top principal components), and then to predict disease status based on the selected transcripts or representations using machine learning algorithms. We identified Five papers published between July 2019 to August 2020, which applied Machine learning methods in Genetic Disease Prediction using genomic data.

TABLE I

LITERATURE REVIEW

Sr. No	Title	Year of Publication	Accuracy
1	Analysis for Disease Gene Association Using Machine Learning	Aug 2020	93.8%
2	Predicting Diabetes in Healthy Population	Aug 2019	84.1%

	through Machine Learning		
3	Down Syndrome Prediction Using a Cascaded Machine Learning Framework Designed for Imbalanced and Feature-Correlated Data	July 2019	95%

The selected study analyses some novel computational methods for the identification of genes associated with diseases. Four diseases are studied with respect to their gene association - Thalassemia, Diabetes, Malaria, and Asthma. Extreme learning machine is a neural network system that only enables data to move one direction across several layers. The weights of the network are constant during the validation process, where the model is tested. The DELM model integrates the input layer, several hidden layers, and one output layer.

A machine learning scheme that can identify healthy subjects that are at an increased risk of developing type-2 diabetes. The outcomes of the study show that high values of glucose observed at the 2 h mark during the OGTT may strongly indicate the potential risk of future development of type-2 diabetes.

Cascaded Machine Learning Framework has been proposed specially for Imbalanced and Feature-correlated data.

Framework is based on three complementary stages –

1. Prejudgement with isolation forest technique

2. Model assemble by voting strategy

3. Final judgement using logistic regression approach

IV. METHODOLOGY

Prediction of Genetic Disorders can be a complex process due to a number of parameters that result in one. There are many factors that are yet to be addressed and do not seem statistical at first. But, by proper use of machine learning techniques, one can relate the parameters involved and discard unnecessary ones.

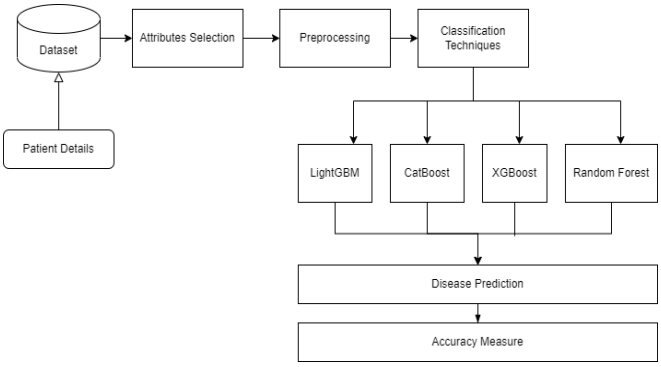

Figure 1. Proposed Methodology

A. Exploratory Data Analysis (EDA)

The dataset contains parameters like patient id name, blood cell count, respiratory rate, heart rate, folic acid details, serious maternal illness, WBC count of any symptoms, history of anomaly, etc. The dataset used is a structured dataset with disorder class and disorder subclass being target variables for training the model. Table II depicts relation between two target variables viz. 'Genetic Disorder' and 'Disorder Subclass'.

TABLE III
TARGET VARIABLES

Genetic Disorder	Disorder Subclass
Mitochondrial genetic inheritance disorders	Leber's hereditary optic neuropathy
	Leigh syndrome
	Mitochondrial myopathy
Multifactorial	Alzheimer's

genetic inheritance disorders	Cancer
	Diabetes
Single-gene inheritance diseases	Cystic fibrosis
	Hemochromatosis
	Tay-Sachs

Figure 2. gives an idea about the Genetic Disorder instances present across the three major classes. The figure shows an imbalance among classes where Multifactorial genetic inheritance disorders capped as low as 10.4% of the entire dataset.

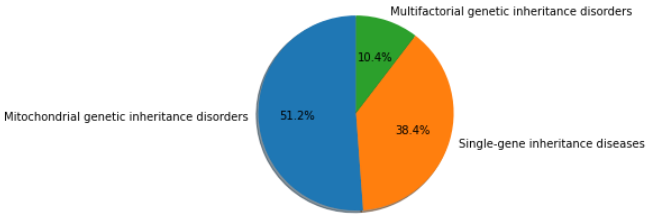

Figure 2. Genetic Disorder Class Instances

Furthermore, figure 3 expands the view by depicting a drastic imbalance among disorder subclasses.

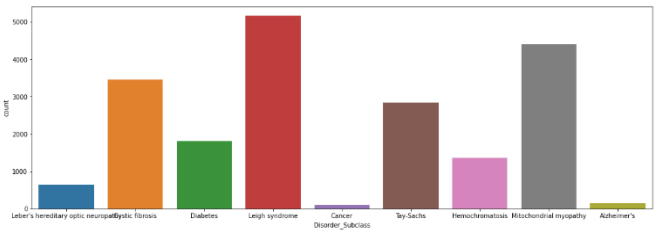

Figure 3. Disorder Subclass Instances

B. Feature Engineering

The raw dataset includes 22,084 instances of patient data built across 32 features. Table III gives a brief description of the existing features in the dataset.

TABLE IIIII
ATTRIBUTES OF THE DATASET

Column name	Column description
Patient Id	Represents the unique identification number of a patient
Patient Age	Represents the age of a patient
Genes in mother's side	Represents a gene defect in a patient's mother
Inherited from father	Represents a gene defect in a patient's father
Maternal gene	Represents a gene defect in the patient's maternal side of the family
Paternal gene	Represents a gene defect in a patient's paternal side of the family
Blood cell count (mcL)	Represents the blood cell count of a patient
Patient First Name	Represents a patient's first name
Family Name	Represents a patient's family name or surname
Father's name	Represents a patient's father's name
Mother's age	Represents a patient's mother's name
Father's age	Represents a patient's father's age
Institute Name	Represents the medical institute where a patient was born
Location of Institute	Represents the location of the medical institute
Status	Represents whether a patient is deceased
Respiratory Rate	Represents a patient's

(breaths/min)	respiratory breathing rate
Heart Rate (rates/min)	Represents a patient's heart rate
Parental consent	Represents whether a patient's parents approved the treatment plan
Follow-up	Represents a patient's level of risk (how intense their condition is)
Gender	Represents a patient's gender
Birth asphyxia	Represents whether a patient suffered from birth asphyxia
Autopsy shows birth defect (if applicable)	Represents whether a patient's autopsy showed any birth defects
Place of birth	Represents whether a patient was born in a medical institute or home
Folic acid details (periconceptional)	Represents the periconceptional folic acid supplementation details of a patient
H/O serious maternal illness	Represents an unexpected outcome of labor and delivery that resulted in significant short or long-term consequences to a patient's mother
H/O radiation exposure (x-ray)	Represents whether a patient has any radiation exposure history
H/O substance abuse	Represents whether a parent has a history of drug addiction
Assisted conception IVF/ART	Represents the type of treatment used for infertility
History of anomalies in previous pregnancies	Represents whether the mother had any anomalies in her previous pregnancies
No. of previous abortion	Represents the number of abortions that a mother had
Birth defects	Represents whether a patient has birth defects
White Blood cell count (thousand per microliter)	Represents a patient's white blood cell count
Blood test result	Represents a patient's blood test results
Genetic Disorder	Represents the genetic disorder that a patient has
Disorder Subclass	Represents the subclass of the disorder

1) Standardization: Standardization of datasets is a common requirement of many machine learning estimators implemented; they might behave badly if the individual features do not more or less look like standard normally distributed data: Gaussian with zero mean and unit variance.

   StandardScaler utility class is a quick and easy way to perform the operations on an array-like dataset. Standardize features by removing the mean and scaling to unit variance. The method has been actively implemented and has proven to give appropriate results.

2) Feature Selection: It is important to understand which feature selection method will work properly for the model. To know this, we need to first identify the type of input and output variables. In the project we mainly used two types of feature selection Techniques –
   i) *Filter Method*: Statistical metrics are used to choose features in the Filter Method. This method chooses the features as a pre-

processing step and does not rely on the learning algorithm. By rating distinct metrics, the filter approach filters away unnecessary features and superfluous columns from the model. Filter methods have the advantage of requiring less processing effort and not overfitting the data. The value of the missing value ratio can be used for evaluating the feature set against the threshold value. The formula for obtaining the missing value ratio is the number of missing values in each column divided by the total number of observations. The variable having more than the threshold value can be dropped.

ii) *Random Forest*: Different tree-based feature selection approaches assist us in determining the value of features and provide a means for picking features. In this case, feature importance indicates which feature is more important in model construction or has a significant impact on the target variable. Random Forest is a tree-based method that aggregates a different number of decision trees and is a sort of bagging algorithm. It ranks the nodes across all trees based on their performance or decrease in impurity (Gini impurity). Nodes are ordered according to impurity levels, allowing for tree trimming below a certain node. The nodes that remain form a subset of the most significant features.

Correlation describes the relationship between the features and the goal variable. Correlation can be positive (an increase in one feature's value improves the value of the target variable) or negative (a reduction in one feature's value decreases the value of the target variable) (increase in one value of feature decreases the value of the target variable). Figure 4 is a heatmap of correlated characteristics using the seaborn library to find which features are most connected to the target variable.

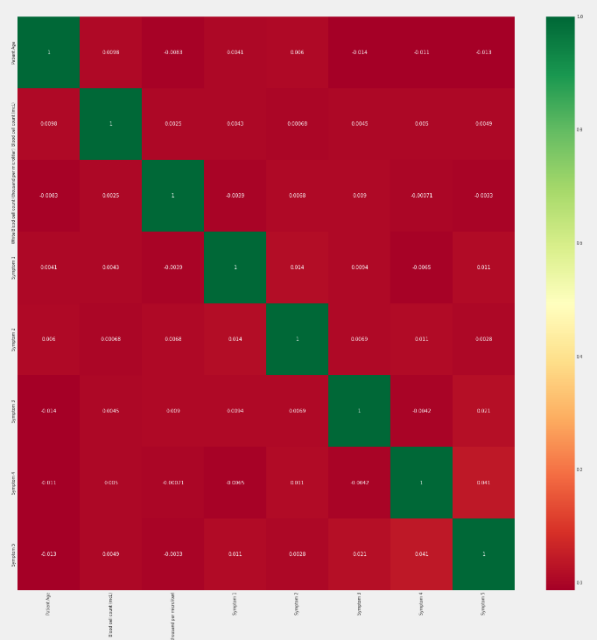

Figure 4. Feature Correlation Heatmap

From the methods stated, we selected the best attributes listed in Table IV.

TABLE IVV
SELECTED ATTRIBUTES OF THE DATASET

Sr. No.	Column name
1	Patient Age
2	Genes in mother's side
3	Inherited from father
4	Maternal gene
5	Paternal gene
6	Blood cell count (mcL)
7	Respiratory Rate (breaths/min)
8	Heart Rate (rates/min)
9	Gender

10	Birth asphyxia
11	Autopsy shows birth defect (if applicable)
12	Folic acid details (periconceptional)
13	H/O serious maternal illness
14	H/O radiation exposure (x-ray)
15	H/O substance abuse
16	Assisted conception IVF/ART
17	History of anomalies in previous pregnancies
18	No. of previous abortion
19	Birth defects
20	Blood test result

3) Feature Encoding: In machine learning, we usually deal with datasets that contain multiple labels in one or more than one column. These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labelled in words. Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated.

C. Training the model

Two classifiers for the prediction of genetic disorder and the disorder subclass wherein first classifier is for genetic disorder prediction and second is for Disorder Subclass are targeted. For training, a train-test split of 80:20 has been used. According to the input parameters, the related classifiers that have a remarkable history in classifying data with high feature count and class imbalance have been selected for optimum results.

The mainly focused algorithms are as follows:

1) K-Nearest Neighbors: It tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closest to the test data.
2) CatBoost: It is a high-performance open-source library for gradient boosting on decision trees.
3) XGBoost: It is efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. XGBoost dominates structured or tabular datasets on classification and regression predictive modelling problems.
4) LightGBM: It uses a histogram-based algorithm which bucket continuous feature (attribute) values into discrete bins. These speeds up training and reduces memory usage.
5) Random Forest: It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.

V. RESULTS

The said algorithms were trained and tested for both classifiers and the results are noted in Table V.

TABLE V
ACCURACY COMPARISON

Model	Classifier - 1		Classifier - 2	
	Train	Test	Train	Test
K-Nearest Neighbor	73.47	60.59	79.71	68.02
CatBoost	62.2	55.98	75.36	69.86
Random	81.9	57.9	89.4	69.6

Forest				
LightGBM	66.9	55.3	90.1	70.8
XGBoost	95.28	56.6	86.1	70.1
XGBoost + RF	78.1	72.6	83.2	88.7

We find that accuracy of the combination of XGBoost and Random Forest is much more efficient as compared to other algorithms. It can be concluded that this combination is best among others with 88% accuracy and the comparison is shown in Figure 5.

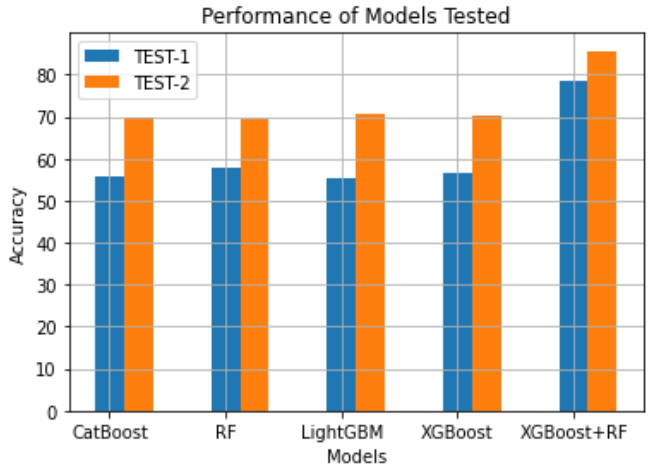

Figure 5. Performance Evaluation

VI. CONCLUSION

It is believed that predicting genetic disorders at an early phase of its advent becomes important for a healthy population, to maximise comfort of the patient and retard its growth. Early detection and medical interventions can prevent many severe complications. This report runs from the fundamentals of the Human Genome and depicts the reason why it is necessary to diagnose the genetic disorders. The available literature relating to this subject has been analysed and reviewed which has helped in comprehending various approaches to build the said model. Furthermore, a proposed methodology for building the model has been presented with a brief description of the same. Finally, system requirements for execution and evaluation of various classifier and prediction algorithms have been specified.

Evaluation of the performance of machine learning methods and algorithms in predicting the genetic disorders and its subclasses can help in increasing the accuracy of the model. Analysis of the dataset has further allowed us to extract optimum feature sets for fitting the model. By enduring a few literature reviews, we were able to devise a concrete methodology which is believed to obtain the best accuracy possible. The strength of machine learning data modelling in disease prediction lies in its handling of interactive high-dimensional data.

VII. FUTURE SCOPE

The model can be tried and tested for real medical diagnosis by health departments, clinics and hospitals. The results of this study may also further assist in lab experiments on genetic disorders. The project can be made more accessible and easier to use by adding a Graphical Interface like a website or a mobile application. The application of such disease predictive models to diverse clinical populations will clarify the performance and limitations of proposed predictive models and improve medical practice. While prediction will continue to be challenging, future investigations promise to provide a wealth of information, some of which will be clinically useful if considered in the appropriate context.

VIII. REFERENCES

[1]. Lvovs, D.; Favorova, O.O.; Favorov, A.V. (2012). "A Polygenic Approach to the Study of Polygenic Diseases". Acta Naturae. 4(3): 59–71. doi:10.32607/20758251-2012-4-3- 59-71. ISSN 2075-8251. PMC 3491892. PMID 23150804.

[2]. Bick, David; Bick, Sarah L.; Dimmock, David P.; Fowler, Tom A.; Caulfield, Mark J.; Scott, Richard H. (March 2021). "An online

compendium of treatable genetic disorders". American Journal of Medical Genetics. Part C, Seminars in Medical Genetics. 187 (1): 48–54. doi:10.1002/ajmg.c.31874. ISSN 1552-4876. PMC 7986124. PMID 33350578.

[3]. Kumar, Pankaj; Radhakrishnan, Jolly; Chowdhary, Maksud A.; Giampietro, Philip F. (2001-08-01). "Prevalence and Patterns of Presentation of Genetic Disorders in a Pediatric Emergency Department". Mayo Clinic Proceedings. 76 (8): 777–783. doi:10.4065/76.8.777. ISSN 0025-6196. PMID 11499815.

[4]. Jackson, Maria; Marks, Leah; May, Gerhard H.W.; Wilson, Joanna B. (2018-12-03). "The genetic basis of disease". Essays in Biochemistry. 62 (5): 643–723. doi:10.1042/EBC20170053. ISSN 0071-1365. PMC 6279436. PMID 30509934

[5]. Sleator RD (August 2010). "An overview of the current status of eukaryote gene prediction strategies''. Gene. 461 (1–2): 1–4. doi:10.1016/j.gene.2010.04.008. PMID 20430068.

[6]. Ejigu, Girum Fitihamlak; Jung, Jaehee (2020-09-18). "Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing". Biology. 9 (9): 295. doi:10.3390/biology9090295. ISSN 2079-7737. PMC 7565776. PMID 32962098.

[7]. M. Sikandar et al., "Analysis for Disease Gene Association Using Machine Learning," in IEEE Access, vol. 8, pp. 160616-160626, 2020, doi: 10.1109/ACCESS.2020.3020592.

[8]. T. Akter et al., "Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders," in IEEE Access, vol. 7, pp. 166509-166527, 2019, doi: 10.1109/ACCESS.2019.2952609.

[9]. H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani and K. Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 2019, pp. 567-570, doi: 10.1109/CBMS.2019.00117.

[10]. L. Li, W. Liu, H. Zhang, Y. Jiang, X. Hu and R. Liu, "Down Syndrome Prediction Using a Cascaded Machine Learning Framework Designed for Imbalanced and Feature-correlated Data," in IEEE Access, vol. 7, pp. 97582-97593, 2019, doi: 10.1109/ACCESS.2019.2929681.

[11]. Alharbi, Noorh H.; Bameer, Rana O.; Geddan, Shahad S.; and Alharbi, Hajar M. (2020) "Recent Advances and Machine Learning Techniques on Sickle Cell Disease," Future Computing and Informatics Journal: Vol. 5 : Iss. 1 , Article 4.

[12]. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003). KNN Model-Based Approach in Classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003

Cite this article as :

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