Predicting genetic disorder and disorder subclass based on familial history and its effects

B Pravena
Department of Computer Science
PES University
Bangalore, India
bpravena11@gmail.com

Lavanya Yavagal
Department of Computer Science
PES University
Bangalore, India
lavanyamiliy@gmail.com

Swarnamalya A S
Department of Computer Science
PES University
Bangalore, India
swarna.pes@gmail.com

Varna Satyanarayana
Department of Computer Science
PES University
Bangalore, India
varna.satya@gmail.com

***Abstract*— Detecting the genetic disorder subclass in children can help with early medical intervention, therefore helping patients with disorders to live a better quality of life in the future. Our project predicts the genetic disorder subclass a child may have based on hereditary factors such as the presence of a certain defective gene in the mother or father or their families, parents' age, pregnancy factors such as periconceptional folic acid details, history of substance abuse, serious illness, anomalies in previous pregnancies or exposure to radiation. Also given the outcomes of 5 masked tests and the presence of 5 masked symptoms, blood cell count, gender, presence of birth asphyxia, and whether autopsy shows any birth abnormalities in the patient, we predict what genetic disorder the child may have. We made use of the Gradient Boosting model to predict the genetic disorder subclass with an accuracy of 75.9% which we use to classify the genetic disorder that the patient may have. Given the genetic disorder, we can predict the disorder subclass with an accuracy of 88.87% using the same model.**

I. Introduction

A genetic disorder is a category of diseases that includes certain types of birth defects, chronic diseases, developmental problems, and sensory deficits that are inherited from one or both parents. This happens when there is a defect in the code of the gene(s), this can be due to mutation in a gene, missing parts in a chromosome, extra or missing chromosomes, or too many or too few sex chromosomes. Genetic disorders are vastly due to fetal exposure to certain drugs, chemicals, and radiation. Pregnancy factors such as the mother's age during conception, alcoholism, and diabetes, and infertility of either parent are some of the causes.

Due to a lack of understanding about the need for genetic testing, hereditary disorders are becoming more common. Often kids are affected as a result of these illnesses, thus genetic testing during pregnancy is critical. Genetic counseling helps understand the risk of you or your child developing a genetic disorder and know the ground realities of causes and effects disorders have. This project was done with the aim of providing an easy tool to perform genetic counseling with minimal provided data such as information about patients and their hereditary factors such as maternal and paternal genes and genes present on the mother's and father's side, a brief description of the corresponding symptoms (masked) and health information such as blood cell count, respiratory rate, autopsy showing birth defect, periconceptional folic acid details and contains the outcomes of masked tests of the patient. With this information, we predict the disorder subclass a patient might have. Each disorder subclass belongs to one of three genetic disorders: Mitochondrial genetic inheritance dis- orders, Multifactorial genetic inheritance disorder, and Single- gene inheritance diseases. Once we predict the genetic disorder subclass, we infer the genetic disorder the patient may have. We made use of the gradient boosting classifier model on the dataset Genomes and Genetics[29] to predict the disorder subclass.

Detecting disorders early on may help with treatments that may help improve the quality of life of the patient. It is very important for expecting parents to understand the risks that may come with their genetic history and lifestyle. Using this tool could help doctors identify the genetic disorder the patient may have.

II. Previous Work

Many approaches to classify patients into subgroups based on hereditary information and symptoms were looked into. One of the approaches included using heterogeneous decision trees seen in [11]. They make use of self-adapting hetero- geneous trees to find the best-hidden tree structure for an accurate prediction. This works with a 95% accuracy but due to the large computational complexity, it forces heavy feature selection and other simplifications or restrictions to algorithm implementation. One more approach using Nearest Consensus Clustering Classification was seen in [12]. This worked on classifying patients into their belonging cohorts. Their proposed model integrates decision trees, consensus clustering, and single linkage metrics and handles such as high variance due to sampling variance and method bias. This has

an accuracy of 75. In [14], they developed a complementary machine-learning approach based on a human brain-specific gene network to present a genome-wide prediction of autism risk genes. They used Evidence-weighted Support Vector Machine disease-gene classifier and KNN methods. Though the model gave an accuracy of 89%. Another approach [15] proposed a method that applies graph embedding representation and GCN on a heterogeneous gene-disease graph to predict the gene-disease associations. To ensure the generalization ability, they add cluster loss and dropout of adjacent matrices into training. Another method [16] that aims to predict gene-disease associa- tions, which are considered semi-supervised learning problems because of the lack of negative samples. Results reveal that the performance of LFMHSR is 7% more efficient than that of other existing approaches. In [14], they developed a complementary machine-learning approach based on a human brain-specific gene network to present a genome-wide prediction of autism risk genes. They used Evidence-weighted Support Vector Machine disease-gene classifier and KNN methods. Though the model gave an accuracy of 89%. Another approach [15] proposed a method that applies graph embedding representation and GCN on a heterogeneous gene-disease graph to predict the gene-disease associations. To ensure the generalization ability, they add cluster loss and dropout of adjacent matrices into training. Another method [16] that aims to predict gene-disease associa- tions, which are considered semi-supervised learning problems because of the lack of negative samples. Results reveal that the performance of LFMHSR is 7% more efficient than that of other existing approaches.

Based on our research of one of the previous papers [23], which used SVM, Logical Regression, Best-estimate, the results indicate that the atypical subtype should be incorporated in future treatment, genetic and other etiologic studies of major depression. One of the drawbacks was that the subsample of specific subtypes of depression reduced the power to detect the specificity of mood disorder subtypes. The performed models actually gave them an accuracy of 95%. Another paper [24] used the models - Genome-Wide Association Study (GWAS) and Multivariate Analysis of Variance (MANOVA). One limitation of the study was the power reduction caused by the exclusion of relatives, rather than (LMM) with a term for kinship as measured by the (GRM). In [25], the results showed that the drivers that influence genetic testing among adults with IMDs and parents/caregivers of affected minors include the desire for knowledge, a benefit to science, recommendations by doctors and cascade testing. This study used Fisher's Exact Test, t-test but one of the limitations was that the percentage of correct responses was way too less. The limitations of this study include our inability to verify whether or not genetic testing took place.

In the paper [26], gene expression profiling which has been widely used to characterize the status of a cell to reflect the health of the body, to diagnose genetic diseases, etc predicted using the XGBoost model. It outperforms the D-GEX algorithm and is better than the traditional machine learning algorithms. It's also less likely to overfit, faster than traditional tree models, and has a higher practical value. This model gave an accuracy of 71.8%. In the paper [27], three phases are proposed: first, all sources are merged into a single network, then it partitions the integrated network according to edge density introducing a notion of edge type to distinguish the parts and finally, employ a novel node kernel suitable for graphs with typed edges. They show how the node kernel can generate a large number of discriminative features that can be efficiently processed by linear regularized machine learning classifiers. This gave them an accuracy of about 80%. In the paper [28], they examined the relationship between the two approaches for genetic risk prediction and showed that risk prediction based on a GRS (Genetic Risk Score) is mathematically equivalent to NBC (Naive Bayesian Classifier), when the same SNPs (single nucleotide polymorphisms) with the same mode of inheritance are used in the models. The equivalence is based on the fact that both models essentially base the prediction on a weighted average of ORs of the individual SNPs.

III. Proposed Solution

Machine Learning is categorized into 3 primary categories based on how an algorithm learns to become more accurate in its predictions - supervised, unsupervised and semi-supervised learning. Depending on what type of data is being predicted, the type of algorithm is chosen accordingly. In supervised learning, the algorithms are trained based on a labeled dataset depending upon their characteristics. The classes can either be continuous, or discrete. One of the most common applications of machine learning algorithms is classification, which is used in everyday life in order to solve large-scale real-world problems. This study focuses on the classification of Genetic Disorders as 'Mitochondrial genetic inheritance disorders', 'Multifactorial genetic inheritance disorder' and 'Single-gene inheritance diseases' based on the Disorder Subclass which are 'Cystic fibrosis', 'Diabetes', 'Leigh syndrome', 'Cancer', 'Tay-Sachs', 'Hemochromatosis' and 'Mitochondrial myopa- thy'.

A. Preprocessing

We do not consider the observations where 'Disorder Sub- class' is not mentioned, hence they are unknown to us. We first consider only the necessary features from the dataset and hence drop the columns 'Patient Id', 'Family Name', 'Patient First Name',' Father's name', 'Institute Name', 'Location of Institute', 'Place of birth', 'Parental consent' and 'Follow-up' as all these columns are not relevant to predicting the disorder a patient might have. As each disorder subclass belongs to one of the three genetic disorders: Mitochondrial genetic in- heritance disorders, Multifactorial genetic inheritance disorder, and Single-gene inheritance diseases, any rows in the dataset with missing genetic disorder is filled by inferring it from the disorder subclass mentioned. Any rows that have missing disorder subclasses are dropped. The NaN values in all other columns in the data frame are filled by grouping the data by the Genetic Disorder and Disorder subclass and filling missing values in each group with mode if the column is of categorical type and mean if the column is of continuous type. The columns which have continuous values are normalized using a minmax scaler.

B. Visualization

We performed the chi-square test to find the correlation between all the variables. We found that the 'inherited from father' column has a positive correlation of 0.57 with Genetic Disorder, with a correlation value of. Blood test results in- herited from the father have a strong correlation of 0.87. The 'Genes on mother side' was also found to be strongly corre- lated to respiratory rate, heart rate, gender, folic acid details, H/O serious maternal illness, H/O radiation exposure, assisted conception, history of anomalies in previous pregnancies, number of previous abortion and birth defects. Paternal gene and Genetic disorder have a high correlation of 0.95 value. Paternal gene is also strongly correlated to birth defects, a number of previous abortions, history of anomalies in previous pregnancies, assisted conception, respiratory rate, heart rate, gender, folic acid details, H/O serious maternal illness, and H/O radiation exposure. This verifies that genetic disorders are formed due to parental history. The outliers seen in the blood cell count were not removed as we perceive it as important information to help classify the patient into the right genetic disorder. The correlation between the continuous columns was also found but no strong correlation exists between the variables. The categorical values were then encoded using one- hot encoding.

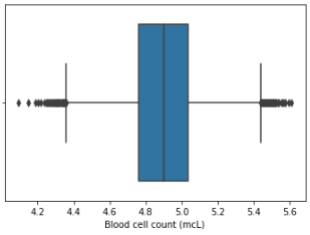

Fig. 1. Outliers of Blood Cell Count (mcL).

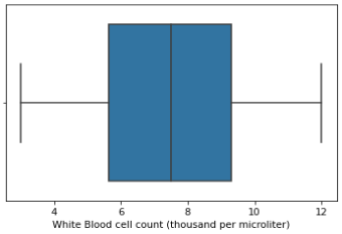

Fig. 2. Outliers of White Blood Cell Count (mcL)

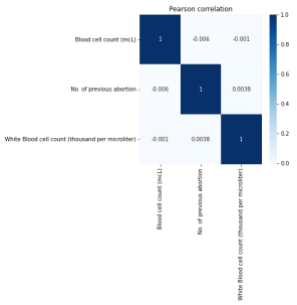

Fig. 1. Pearson Correlation

C. Classification Models

For our analysis, we have used 9 models depending upon each of its unique characteristics to tackle the problem from different angles.

1) XGBoost: XGBoost handles feature selection with a high number of variables. It is usually used to deal with miss- ing data imputation, regression, and classification problems. It can also handle noisy data and outliers very well and is designed for better speed and performance. This algorithm is usually considered powerful as it is scalable, which drives fast learning through distributed and parallel learning and offers efficient memory usage.

2) Random Forest: This classifier contains a number of decision trees on the subsets of the dataset and then takes average to improve the predictive accuracy. The greater the number of trees in the forest, the higher the accuracy. Random Forest allows randomized feature selection from the dataset to construct Decision Trees. This randomization then decreases the effects of correlation and reduces overfitting.

3) Decision Tree Classification: This is a supervised learn- ing technique that can be used for both regression and clas- sification problems. It is a tree-structured classifier, where the attributes of the dataset are represented by the internal nodes, the decision rules are represented by branches, and the outcome by the leaf nodes. We use the CART (Classification and Regression Tree) algorithm to build a decision tree. It is a graphical representation that, based on given conditions, gets all possible solutions.

4) Gaussian Naive Bayesian Classifier: This classification is as good as Decision Tree and Neural Network algorithms. This classifier is a variant of Naive Bayes that supports continuous data and follows Gaussian normal distribution. For this classifier, it is assumed that the continuous values associated with each class are distributed according to a normal or Gaussian distribution. In this approach, the data is described by a Gaussian with no co-variance between dimensions. This model can be fit by calculating the mean and standard deviation of the points with each label.

5) Multilayer Perceptron: Multilayer Perceptron is a feed-forward artificial neural network. It is composed of one or more layers of neurons/input nodes that are connected as directed graphs between the input and output layers. In this classification, the data is fed to the input layer, there may be one or more hidden layers and predictions are made on the output layer. This classifier uses backpropagation for training the network. It is a deep learning method. Every link has some weights, which are randomly initialized and fine-tuned during the training phase.

6) Bagging Regressor: Bagging Regressor is an ensem- ble meta-estimator that fits base regressors, each on various subsets of the dataset. It then aggregates their individual predictions, either by voting or averaging, to form a final prediction. This can be used as a way to reduce the variance by introducing randomization into its construction procedure and making an ensemble from it.

7) Support Vector Machine: Support Vector Machine seeks to discover a hyperplane among classes and tries to maximize the space between the hyperplane and the opposing classes. It chooses the extreme vectors/points to create the hyperplane, maximizing the margin distance providing some reinforcement so that the predictions can be made with more confidence. It is highly preferred as it produces significant accuracy with very little computation power.

8) *Gradient Boost:* Gradient Boost is one of the most powerful techniques to build predictive models. It involves three elements: a loss function, a weak learner, and an additive model to add weak learners to minimize the loss function. This is one of the best-boosting algorithms which is used to minimize bias error of the model. It is used to not only predict continuous target variables but also categorical target variables. The best estimator for this boosting is fixed and that is Decision Stump.

9) *Ensemble Learning Methods:* The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account and average those models to produce one final model. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

D. Cross-Validation and Hyperparameter Tuning

Apart from setting up the feature space and fitting the model, parameter tuning is a crucial task in finding the model with the highest predictive power. Many strategies for tuning parameters exist. Usually while doing studies, parameters such as the number of trees, tree depth, and the learning rate are considered to be crucial.

The parameters being tuned include 'n_estimators', 'learning rate', and 'max_depth'. 'n_estimators' captures the number of trees that we have added to the model. A higher number can be computationally expensive but is better to learn the data. But this also slowed down the training process considerably, hence we did a parameter search to find the sweet spot. As we change the 'learning rate', ' n estimator' also should be readjusted. (A 10-fold decrease in 'learning rate' should go in line with an approximate 10-fold increase in 'n_estimator'.) 'max_depth' bounds the maximum depth of the tree. We used the obtained results to tune the max_depth parameter. The deeper the tree, the more splits it has, capturing more information. We also found out that 'max_leaf_nodes' = k gives comparable results to 'max_depth' = k-1 but is significantly faster to train at the expense of a slightly higher training error. Performing hyperparameter tuning was a major part of our analysis. It helped increase accuracy for the right values. For reference, we have included the accuracy of the other models as well. The model's performance did improve significantly with the tuning. The rates of 'true positive' and 'true negative' increased.

IV. Experimental Results

We used GradientBoostClassifier from sklearn in Python to execute this model. Apart from space and fitting the model, Boosting usually deals with bias-variance trade-offs, unlike bagging which only deals with high variance models and hence is considered to be more effective for our dataset. A prediction with an accuracy of about approximately 75% was acquired using the Gradient Boost model.

Hence, it was convenient to start with these three parameters and then move to the subsamples. We propose to test using 10-fold cross-validation on the entire set of observations and evaluate the models based on accuracy.

The Gradient Boost being used is an ensemble model, where the overall parameters are divided into 3 categories:

a) Tree-specific Parameters: These affect each individual tree in the model.

b) Boosting Parameters: These affect the boosting operation in the model.

c) Miscellaneous Parameters: Other parameters for overall functioning.

Hence, applying the ensemble modeling, the different methods were integrated using a voting protocol, helping us propose a better way to model different gene expression datasets and thereby giving us a significant increase in the accuracy when compared to applying each base model individually.

In this study, experiments showed that the Gradient Boosting Classifier model achieved a significantly higher overall accuracy than the existing XGB- Classifier, RandomForestClassifier, GradientBoostingClassifier, SVM, DecisionTreeClassifier, MLPClassifier, and Ensemble model. The Gradient boosting model gives us the best accuracy of 75.9% while predicting Disorder Subclass.

We found that the disorder subclass can be more accurately predicted while using the Genetic Disorder the patient already has to train the model, giving us an accuracy of 88.7%. This tool can be great to have as a way to predict what genetic disorder and disorder a child may have so that it can be diagnosed and treated early allowing the patient to lead a better lifestyle in the future.

We also found that filling in nan values in the columns by grouping that data frame by Genetic Disorder and Disorder subclass and taking the mean (if continuous values) or mode (if categorical values) of the column increases the accuracy of the model significantly.

V. Conclusion

The whole purpose of this study was to discover a reliable way to predict the "Genetic Disorder" to reduce the risk of error prediction when not enough samples were used for modeling. The base models that we used are LogisticRegression, XGBClassifier, RandomForestClassifier, GradientBoostingClassifier, SVM, DecisionTreeClassifier, MLP- Classifier with the best parameters after tuning each model. The distribution of our modeling performance indicated that different datasets required different methods in order to give the best results.

Table 1. Accuracies of predicting Disorder Subclass and Genetic Disorder

Training model	XGBoost	Bagging Regressor	Ensemble	Gradient Boosting
Without Genetic Disorder (in %)	71.05	73.03	74.94	75.9

Training model	Multinomial Logistic regression	XGBoost	Random Forest	Decision Tree Classifier	Gaussian Naive Bayes Classifier	Bagging Regressor	Ensamble	Gradient Boosting
With Genetic Disorder (in %)	79	85.01	85.89	80.21	77.52	81.66	87.24	88.87

Table 2. Accuracies of predicting Disorder Subclass with Genetic Disorder

VI. REFERENCES

[1] https://medium.com/analytics-vidhya/constructing-heat-map-for-chi-square-test-of-independence-6d78aa2b140f
[2] https://www.datasciencelearner.com/gradient-boosting-hyperparameters-tuning/
[3] https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/
[4] https://www.datacamp.com/community/tutorials/xgboost-in-python
[5] https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powerful-models-cc89e60b7a85
[6] https://www.datacamp.com/community/tutorials/random-forests-classifier-python
[7] https://towardsdatascience.com/decision-tree-algorithm-for-multiclass-problems-using-python-6b0ec1183bf5
[8] https://kidshealth.org/en/parents/about-genetics.html
[9] https://www.childrenshospital.org/conditions-and treatments/conditions/g/genetic-disorders
[10] https://towardsdatascience.com/decision-tree-algorithm-for-multiclass-problems-using-python-6b0ec1183bf5
[11] Czajkowski, Marcin Jurczuk, Krzysztof Kretowski, Marek. (2021). Ac-celerated evolutionary induction of heterogeneous decision trees for gene expression-based classification. 946-954. 10.1145/3449639.3459376.
[12] Alyousef, A.A., Nihtyanova, S., Denton, C. et al. Nearest Consensus Clustering Classification to Identify Subclasses and Predict Disease. J Healthc Inform Res 2, 402–422 (2018). https://doi.org/10.1007/s41666-018-0029-6
[13] Ping Luo, Yuanyuan Li, Li-Ping Tian, Fang-Xiang Wu, Enhancing the prediction of disease–gene associations with multimodal deep learning,Bioinformatics, Volume 35, Issue 19, 1 October 2019,
[14] Krishnan A, Zhang R, Yao V, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci. 2016;19(11):1454-1462. doi:10.1038/nn.4353
[15] L. Zhu, Z. Hong and H. Zheng, ”Predicting gene-disease associations via graph embedding and graph convolutional networks,” 2019 IEEE
[16] Xiangxiang Zeng, Ningxiang Ding and Quan Zou, ”Latent fac- tor model with heterogeneous similarity regularization for predicting gene-disease associations,” 2016 IEEE International Conference
[17]https://machinelearningmastery.com/evaluate-gradient-boosting-models-xgboost-python/
[18]https://www.analyticsvidhya.com/blog/2016/02/completeguide-parameter-tuning-gradient-boosting-gbm-python/
[19] https://panjeh.medium.com/scikit-learn-hyperparameter-optimization-for-mlpclassifier-4d670413042b
[20] https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/
[21] https://www.datasciencelearner.com/gradient-boosting hyperparameters-tuning/
[22] https://www.datacareer.de/blog/parameter-tuning-in-gradient-boosting-gbm/
[23]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6333444/pdf/elife-39856.pdf
[24]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6333444/pdf/elife-39856.pdf
[25]https://www.sciencedirect.com/science/article/pii/S2214426920300793
[26]https://www.frontiersin.org/articles/10.3389/fgene.2019.01077/full
[27]https://academic.oup.com/bioinformatics/article/36/9/2649/5716660?login=true203318324
[28]https://www.frontiersin.org/articles/10.3389/fgene.2012.00026/fullh1
[29] https://www.kaggle.com/aryarishabh/of-genomes-and-genetics-hackerearth-ml-challenge

VII. CONTRIBUTION

B Pravena: SVM, XGBoost, Gradient Boosting, Ensemble, preprocessing

Lavanya Yavagal: Decision Tree Classifier, Gaussian Naive Bayes Classifier, Gradient Boosting, Ensemble, preprocessing

Swarnamalya A S: Gradient Boosting, hyperparameter tuning, Multilayer Perceptron, Ensemble, preprocessing

Varna Satyanarayana: Random Forest, Bagging Regressor, Gradient Boosting, Ensemble, preprocessing