# MACHINE LEARNING TECHNIQUES TO IDENTIFY HEREDITARY ILLNESS

***Mrs.G.Sujatha[1]., P.Pallavi[2].,K.Swetha[3].,M.Sai Supraja[4].,S.Yasaswini[5]***
*1 Assistant Professor, Department Of CSE., Malla Reddy College Of Engineering For Women., Maisammaguda., Medchal., Ts, India (✉sujathamantra@gmail.Com)*
*2, 3, 4, 5 B.TechCSE, (19RG1A05N1, 19RG1A05L4, 19RG1A05M0, 19RG1A05N9),*
*Malla Reddy College Of Engineering For Women., Maisammaguda., Medchal., Ts, India*

## Abstract

*New technologies have made it much simpler to treat hereditary illnesses. One of the most difficult challenges of the post-genomic era is sifting through an enormous quantity of genetic data to identify disease-causing genes. It might be challenging to recognize biological markers in cases with complex disorders because of the sometimes extremely varied genotypes that are present. To establish such markers, machine learning algorithms are often sought after, but their efficacy is highly dependent on the quantity and quality of the data made accessible to them. The field of machine learning promises to enable computers to help humans by analyzing large and complicated data sets; its primary goal is to develop algorithms that become better with use. Gene Ontology (GO)-trained machine learning classifiers may improve and identify the genes that are involved in complicated illnesses; we built a supervised technique of machine learning to predict such genes; and we tested with the proposed algorithm. With a combination of principal component analysis (PCA),*

## Introduction

The last 10 years have seen remarkable development in the discovery of genes associated with types of neurological disorders associated with Medellin. Through human gene discovery is important for the affected patients, for the researchers focused on the disease and for the broader neuroscience community [10]. These gene discoveries may however be even more significant when collectively considered. First and foremost, the development of rare disease genetics has shown that central and peripheral nervous system dysfunction is a common result of genetic disorders with approximately 50% of all rare diseases (3000-3500 conditions) having some type of neurological abnormality [3].

The fact that neurological disorders display extraordinary genetic variation has also become apparent. It has become clear, eventually, that variable expressiveness and typical expressions are not the exception, but the rule when contemplating neurogenetic disorders [4]. Such findings have led us to the idea that genes associated with neurogenetic conditions needed something "extra" to exist. The machine learning field cares with developing and applying computer algorithms which improve with experience [6]. Machine learning algorithms use input file to perform as well as learning the way to recognize gene patterns in DNA.

There are numerous diseases which will find through genetics. A genetic disorder could also be a disease caused by one or more genome anomalies, especially a condition present from birth. Most of the genetic diseases are very rare and are affecting only a couple of thousand or millions. The genetic diseases may be heritable, i.e. inherited by the genes of the oldsters. The prediction of the mutations to occur the DNA helps to early swapping of the DNA cells and stop the disease the DNA mutation occurs [7]. The machine learning field cares with developing and applying computer algorithms which improve the experience. In this paper, the section 2 briefs the related work in the field of genetic diseases, section 3 explains the proposed genetic disease analyzer, the experimented results are discussed in the section 4, and the conclusion is discussed in section 5.

## Related Work

Many risks are associate while analyzing with the machine learning tools for genetic diseases analysis [10]. The study address whether the HSQ-23 successfully identify the people who will enjoy PHC consultations. Data from the baseline survey (n = 2056) has been used to analyze the measurement and the properties of the SQ-33 in order to determine the scaling properties for the complete range of subjects. Risk assessment is a crucial component of genetic therapy and science, and therefore the genetic risk for decision taking between individuals and families should be assessed as accurately as possible [11]. All relevant information gained from demographic research and lineage and genetic testing improves the accuracy of the genetic risk assessment for a person. Oh *et al* [8] used the quintile method for normalization of background correction with “normexp”. They have employed SVM; LDA algorithms with the Hierarchical based Euclidian distance method. They have obtained the accuracy as 93.8%, 68.8% for SVM and LDA respectively. They have limited the testing space by selecting two genes for classification. They have also experimented the model with Artificial Neural Network (ANN) and obtained 93% accuracy. They have used S-Coonan method as well, which yielded 91% accuracy. Alshamlan *et al.* [1] presented a hybrid algorithm by using the Ant Bee Colony algorithm with Genetic Algorithm. They have selected the distinguished genes from the pool of genes. They have employed SVM algorithm for classification of the identified dataset. They have obtained 100% accuracy in Genetic Bee Colony – SVM model. They have used PSO and KNN model to minimize the time computation. It is yielded 97.05% accuracy. Kalaiselvi *et al.* [5] designed a fuzzy based model for classification. They have employed entropy to identify the i-genes. They also used featured selection algorithms. They have obtained 100% accuracy using locally weighted learning algorithms. They also used many ML algorithms for their experiments. Their model has taken additional computations due to overhead operations.

Hammed *et al.* [2] have invented an algorithm using particle optimization techniques. They have experimented with the many statistical methods. They have used SVM, GPSO algorithms. Their model yielded 92.1% accuracy in identifying the risk factor genes. Vinita *et al.* [13] have introduced hybrid algorithm using ANN, SVM, and KNN. They have employed “Leave One Out” method for cross validation. They have experimented with the colon cancer datasets. Their model yielded 95.98% accuracy in classifications.

The genetic testing techniques still evolve, and there's no doubt that genetic risk assessment will become fully incorporated into all areas of medical aid. Advances in gene identification and characterization, studies of associations with polymorphism, classification of diseases, and so on, still provide rapidly new and clinically relevant information which will contribute to raise detection, evaluation, prevention and follow-up of human disease

approaches [12]. Additionally, prospective clinical trials must be performed to work out the simplest use of current management techniques, establish risk assessment methods that integrate additional risk-factor knowledge, and analyze populations that established risk assessment approaches don't yet exist. A concise overview of possible issue genetic diseases and related to problems which will occur during sedation or anesthesia is given. Recommendations are issued for permeation assessment and checklist items which will have an impact on healthcare delivery for these patients. This analysis isn't meant to be a scientific list of all possible complicated genetic diseases. Several of those disorders are fairly normal while those with numerous congenital defects that impact health care access are uncommon [14]. It is becoming increasingly popular for "healthy" people to pursue genetic predisposition testing, and physicians will integrate genetic risk assessment and management into their regular screening and appointments for health maintenance. The prevailing limits, risks, legal and psychological consequences of the genetic risk evaluation. As mentioned above, the incorporation of genetic data from an in depth relative are often extremely useful in assessing a person's genetic risk. These earlier methods are still of use today. For over 30 years, chromosomal analysis has been wont to diagnose defects within the number or sort of chromosomes which may contribute to genetic abnormalities and disease [15].

## Proposed Genetic Disease Analyzer

The aim of this research is typically to identify the genetic variations that are occurring in DNA which may directly or indirectly lead to the increased risk of diseases. R programming provides a collection of built-in libraries that help with minimal code and flexibility to create visualizations. Data Visualization is an art of turning data into insights easily interpretable. Now that we have an overview of the dataset, and the variables, we have to define the interest variables. Domain awareness and the correlation among variables help to select these variables. Explain what matters when choosing data analysis tools, and give them experience when making such choices.

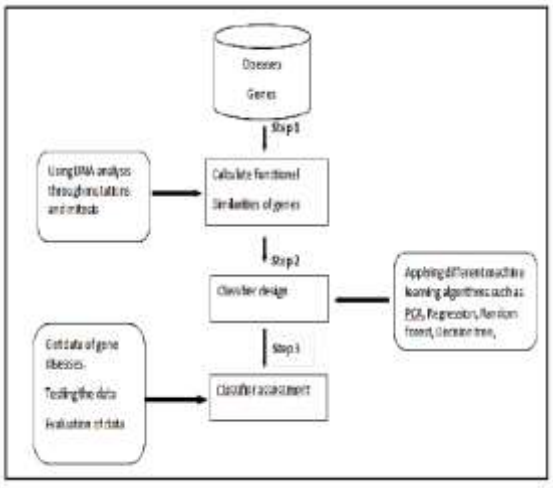


**Figure 1: Proposed Genetic Disease Analyzer Architecture**

The basic idea of our classification is to suggest pathways as identification criteria for a genetic disorder. The idea of classifying human disease using pathway-based approaches to the molecular and cellular processes had marked the

start of the post-genomic era. The basic criteria of the project is to analyze the DNA structure through mutations in DNA to find the chances of a person to be vulnerable to the disease .The prediction of the mutations to occur the DNA helps to early swapping of the DNA cells and prevent the disease even though the DNA mutation occurs .This projects helps to find various DNA mutations and mitosis which helps the future generation of the family to overcome at least a few genetic diseases to occur. The proposed GDA is designed as Hybrid method using PCA, Regression, Random Forest, Decision tress algorithms as shown in Figure 1. The classification is carried out as shown in Figure 1. The classifier data set is used for reference. The machine learning techniques were used in this research. PCA, Random forest, Regression, when building machine learning algorithms, just one feature is employed at a time to separate the node and partition the info. By using the R studio from the info we discover the danger analysis. The cycle is repeated recursively on the pruned set before eventually hitting the optimal number of attributes to settle on from. If it's an abnormal effect on biological pathways within the cell, any alteration within the DNA sequence are often pathogenic. A correct diagnosis involves identification of the genetic basis of the disease. As a powerful and flexible procedure, very small quantities of DNA can be amplified. This method has many uses in various biological fields and was developed to diagnose the genetic diseases on the DNA level. Examining and analyzing DNA will reveal the changes occurred in genes which may cause various genetic diseases. The identification of mutations occurring in DNA is found because of the molecular diagnosis of genetic disorders. It's going to encourage the fine sub-classification, prognosis, and condition therapy. Since most inherited disorders of DNA affect people at the infant level, it's critical for pediatricians to be conversant in the methods of genetic testing also as clinical applications of those tests to realize an accurate diagnosis. The doctors should be ready to identify and analyze genetic diseases and patients suffering from a subtype of chromosomal or single gene diseases on the idea of symptoms and signs, in order to that, they will provide an appropriate genetic test for diagnosing that particular disease.

## Results and Discussion

The proposed GDA model is experimented in the defined testing environment. The GEO dataset is considered for testing environment. The Cityscape tool is used. The PCA, Random Forest algorithms also experimented in the same environment for evaluation. The Accuracy and Sensitivity are considered as performance evaluation parameters. The accuracy is defined as ratio of True Positive, True Negative, False Positive, and False Negative as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FT+FN}$$

The Sensitivity can be defined as ratio of True Positive to the sum of True Positive) and False Negative as

$$\text{Accuracy} = \frac{TP}{TP+FN}$$

The experiments were carried out in Cityscape tool for PCA, Random Forest and proposed GDA algorithms independently. A total of 20 trails experimented and the average of them is recorded as shown in Table 1.

**Table 1: Performance of PCA, RF, GDA algorithms**

| Performance Metrics | Algorithms | | |
|---|---|---|---|
| | PCA | Random Forest | Proposed GDA |
| Accuracy in % | 93.65 | 91.68 | 97.34 |
| Sensitivity in % | 94.32 | 93.78 | 96.45 |

The recorded results were plotted with the performance of PCA, Random Forest, Proposed GDA algorithms. The accuracy of algorithms is plotted as shown in Figure 2.

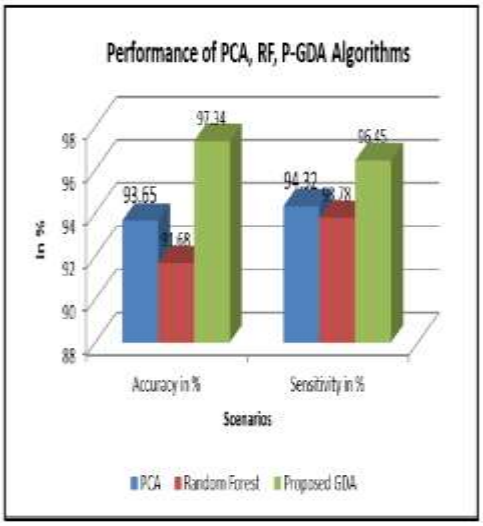


**Figure 2: Performance of PCA, RF, P-GDA algorithms**

The proposed GDA algorithm is experimented along with the PCA and Random Forest algorithm. The recorded values are analyzed and plotted as shown in Figure 2. The P-GDA algorithm is yielded 97.34% accuracy and 96.45% Sensitivity. It is outperformed than PCA as 3.9%, 2.2% in accuracy and sensitivity respectively, the proposed GDA algorithm is outperformed than Random Forest as 6.17% and 2.84% in accuracy and sensitivity respectively.

## Conclusion

An examination of genetic illnesses and machine learning has been conducted. The hybrid model has been suggested for the study of hereditary illnesses. The suggested GDA model was developed using a combination of the principal component analysis, regression, random forest, and decision tree methods. Different analytic procedures, including

principal component analysis and random forest, were used to the GDA. When applied to the GEO data set, the P-GDA model is said to be 97.34 percent accurate and 96.4 percent sensitive. P-GDA outperforms PCA and Random Forest in accuracy by 3.9% and 6.17%, respectively. Furthermore, the sensitivity is improved by 2.2% compared to PCA and 2.8% compared to Random Forest. In the next years, we'll be working to perfect a multi-dimensional algorithm for identifying and predicting hereditary disorders.

## References

*1. Alshamlan, H.M., Bard, G.H. and Aloha, Y.A., 2015. Genetic Bee Colony (GBC) algorithm, Computational Biology and Chemistry, 56, pp.49-60.*

*2. Hammed, 2017. Use of a combination of statistical filters and a GBPSO-SVM algorithm. PloS, 12(11), p.e0187371.*

*3. D. Gong, Z. Zhang, Q. Shi, A. van den Hengel, C. Shen and Y. Zhang, "Learning Deep Gradient Descent Optimization for Image Deconvolution," in IEEE Transactions on Neural Networks and Learning Systems, doe: 10.1109/TNNLS.2020.2968289.*

*4. Longman Yu, Jeonghwan Gawk, Seeing Lee and Mongo Jean, "An incremental learning approach for restricted Boltzmann machines," 2015 International Conference on Control, Automation and Information Sciences (ICCAIS), Changsha, 2015, pp. 113-117, doe: 10.1109/ICCAIS.2015.7338643.*

*5. Kalaiselvi, N. and Inbarani, H.H., 2013. Fuzzy soft set based classification for gene expression data. Arrive preprint arXiv: 1301.1502.*

*6. M. A. Kuzhippallil, C. Joseph and K. A, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 778-782, doi: 10.1109/ICACCS48705.2020.9074368.*

*7. M. Kim, J. H. Won, J. Hong, J. Kwon, H. Park and L. Shen, "Deep Network-Based Feature Selection for Imaging Genetics: Application to Identifying Biomarkers for Parkinson's Disease," 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 1920-1923, doi: 10.1109/ISBI45749.2020.9098471.*

*8. Oh, D.H., Kim, I.B., Kim, S.H. and Ahn, D.H., 2017. Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning. Clinical Psychopharmacology and Neuroscience, 15(1), p.47.*

*9. Pandiaraj, S., Sudalai Muthu, T., Prioritization of replica for replica replacement in data grid, International Journal of Recent Technology and Engineering, 2019, Vol: 7, Issue: 5, pp. 245-248.*

*10. Ranjana, P., Lakshmi Sridevi, S., Sudalai Muthu, T., Vikram Gnanaraj, V., Machine Learning Algorithm in Two wheelers fuel Prediction, Proceedings of 1st International Conference on Innovations in Information and Communication Technology, ICIICT 2019, 2019.*