e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 5, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.54

International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |

Disease Gene Association Analysis Using Machine Learning

Sannareddy. Hemalatha , Sai Harshitha, Anbumani.A

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

Assistant Professor, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT: To recognize the basis of disease, it is essential to determine its underlying genes. Understanding the association between underlying genes and genetic disease is a fundamental problem regarding human health. Identification and association of genes with the disease require time consuming and expensive experimentations of a great number of potential candidate genes. Therefore, inexpensive and rapid computational methods have been proposed that can identify the candidate gene associated with a disease.In this study, we propose and analyze some novel computational methods for the identification of genes associated with diseases. Some advanced topological and biological features that are overlooked currently are introduced for identifying candidate genes. We evaluate different computational methods on disease-gene association data from DisGeNET based on TP rate, FP rate, precision, recall, F-measure, and ROC curve evaluation parameters. The results reveal that various computational methods with advanced feature sets outperform previous state-of-the-art techniques by achieving precision up to 93.8%, recall up to 93.1%, and F- measure up to 92.9%. Significantly, we apply our methods to study three major disease types: Group, Disease and Phenotype. Simulation results show that the proposed Extreme Gradient Boosting Algorithm (XGBoost) gives more accurate results as compared to previously published approaches.

I. INTRODUCTION

A gene is the basic physical and functional unit of heredity that is responsible for different biological processes in an organism. The mutation in a single gene sequence may mutate a biological process and lead to a certain disease. The genes in the human body are not isolated, they interact with one another, therefore, the mutation in a single gene may affect its interacting gene which may also play a part in the mutation of different biological processes and cause different diseases.

Correctly predicting new gene-disease associations has long been an important goal in computational biology. One very successful strategy has been the so-called guilt-by-association (GBA) approach, in which new candidate genes are found through their association with genes already known to be involved in the condition studied. This association can in practice be derived from many different types of data. Goh et al construct a network where genes are connected if they are associated with the same disease, whereas Tian et al. combine protein interactions, genetic interactions, and gene expression correlation, and Ulitsky and Shamir combine interactions from published networks and yeast two-hybrid experiments.

Therefore, consideration of biological mechanisms and based on these mechanisms discovering the relationship between the diseases and genes is a serious challenge in modern biology and medicine. Understanding the association between casual genes and their genetic disease is a fundamental problem regarding human health. Technology is involved in the detection and monitoring of various human diseases such as Parkinson. Also, the Internet of Medical Things (IoMT) is in focus for addressing human health. Different experimental methods have been proposed to associate genes with a disease but these methods are expensive in terms of cost and time.

II. LITERATURE REVIEW

C. K. Saket Navlakha, “The power of protein interaction networks for associating genes with diseases,” Bioinformatics, vol. 26, no. 8, p. 1057–1063, 2010.

Understanding the association between genetic diseases and their causal genes is an important problem concerning human health. With the recent influx of high-throughput data describing interactions between gene products, scientists have been provided a new avenue through which these associations can be inferred. Despite the recent interest in this

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |

problem, however, there is little understanding of the relative benefits and drawbacks underlying the proposed techniques.

O. M. E. R. T. S. R. S. Oron Vanunu, "Associating Genes and Protein Complexes with Disease via Network Propagation," PLoS Computational Biology, vol. 6, no. 1, pp. 1-9, 2010.

A fundamental challenge in human health is the identification of disease-causing genes. Recently, several studies have tackled this challenge via a network-based approach, motivated by the observation that genes causing the same or similar diseases tend to lie close to one another in a network of protein-protein or functional interactions. However, most of these approaches use only local network information in the inference process and are restricted to inferring single gene associations. Here, we provide a global, network-based method for prioritizing disease genes and inferring protein complex associations, which we call PRINCE. The method is based on formulating constraints on the prioritization function that relate to its smoothness over the network and usage of prior information. We exploit this function to predict not only genes but also protein complex associations with a disease of interest. We test our method on gene-disease association data, evaluating both the prioritization achieved and the protein complexes inferred.

M. S. Mabrouk, "A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques," American Journal of Biomedical Engineering, vol. 2, no. 2, pp. 17-22, 2012

This paper presents a hybrid approach based on digital band pass filtering with non-parametric estimation techniques for the analysis of deoxyribonucleic acid (DNA) sequences. These spectral estimation techniques improve the analysis of DNA sequences and enable the extraction of some desirable information about them. The electron-ion interaction pseudo potential (EIIP) numerical representation method is used to convert a DNA sequence to numerical values through a mapping function. Also, mathematical modeling is used to create closed formulas for the represented DNA data sequences with different studied methods. The importance of this process is that the mathematical models can be used for any further processing or identification when applied to DNA sequences. The metrics used for performance evaluation are root mean square error (RMSE) and correlation coefficient(R) metrics. Also, the objective of this paper is investigating and predicting the location of the coding region (exon) in DNA sequences using the proposed approach. The results of gene prediction from DNA sequences for the original and modeled DNA sequences coincide and ensure the success of the proposed sum-of-sinusoids method for modeling of DNA sequences

D. M. Z.-H. D. Adarsh Jose, "A gene selection method for classifying cancer samples using 1D discrete wavelet transform," International Journal of Computational Biology and Drug Design, vol. 2, no. 4, pp. 398-411, 2009.

Selecting a set of discriminant genes for biological samples is an important task for designing highly efficient classifiers using DNA microarray data. The wavelet transform is a very common tool in signal-processing applications, but its potential in the analysis of microarray gene expression data is yet to be explored fully. In this paper, we present a wavelet-based feature selection method that assigns scores to genes for differentiating samples between two classes. The gene expression signal is decomposed using several levels of the wavelet transform. The genes with the highest scores are selected to form a feature set for sample classification. In this study, the feature sets were coupled with k-nearest neighbor(kNN) classifiers. The classification accuracies were assessed using several real data sets. Their performances were compared with several commonly used feature selection methods. The results demonstrate that 1D wavelet analysis is a valuable tool for studying gene expression patterns.

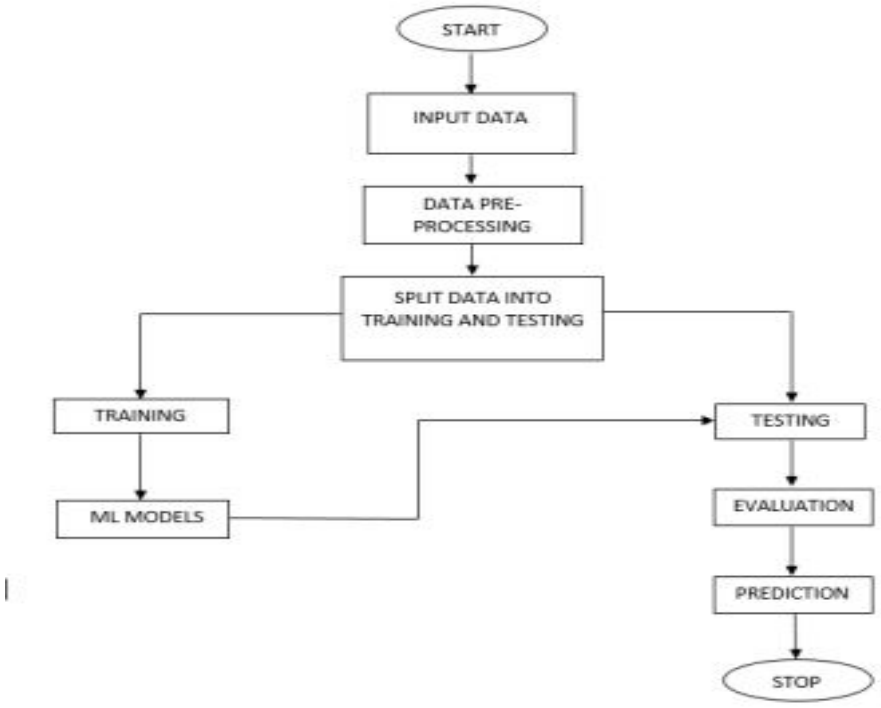
IV. PROPOSED METHOD

In the proposed system, we implement supervised machine learning algorithm Extreme Gradient Boosting (XGBoost Classifier), for detection of the Disease Gene Association. Our Proposed model outperforms existing methods and gives best results.

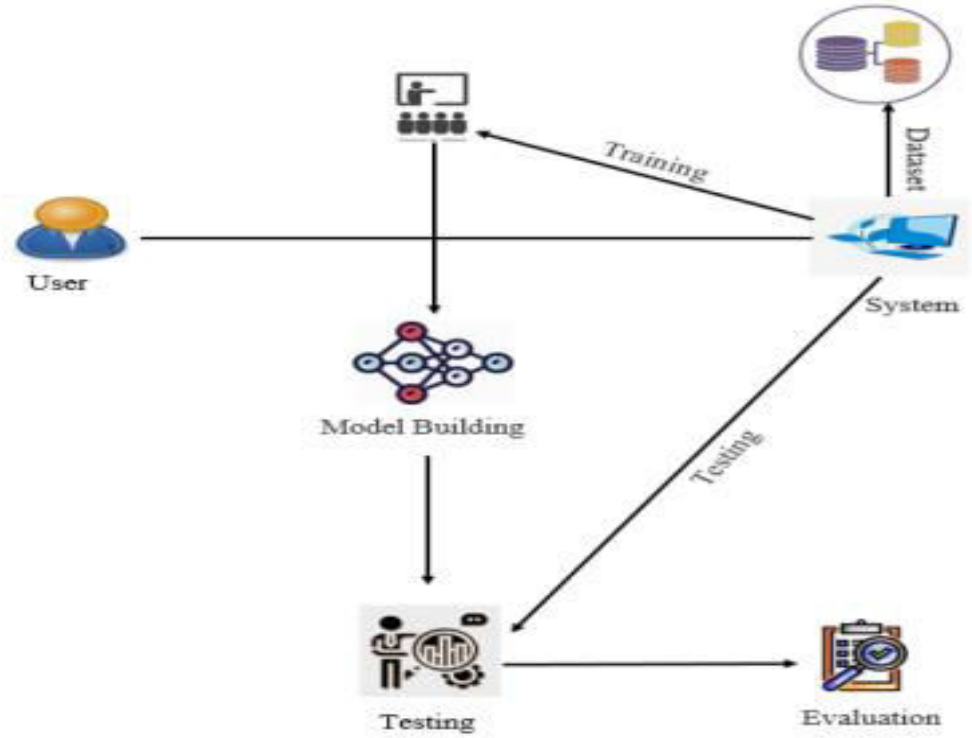
| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |



ARCHITECTURE DIAGRAM



| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |

Module Description:

XGBoost:XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Bagging: Now imagine instead of a single interviewer, now there is an interview panel where each interviewer has a vote. Bagging or bootstrap aggregating involves combining inputs from all interviewers for the final decision through a democratic voting process.

Random Forest:

First, the Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

There are two stages in the Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage.

K Nearest Neighbor:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

Support Vector Machine:

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

So you're working on a text classification problem. You're refining your training data, and maybe you've even tried stuff out using Naive Bayes. But now you're feeling confident in your dataset, and want to take it one step further. Enter Support Vector Machines (SVM): a fast and dependable classification algorithm that performs very well with a limited amount of data to analyze

Light GBM:

LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage.

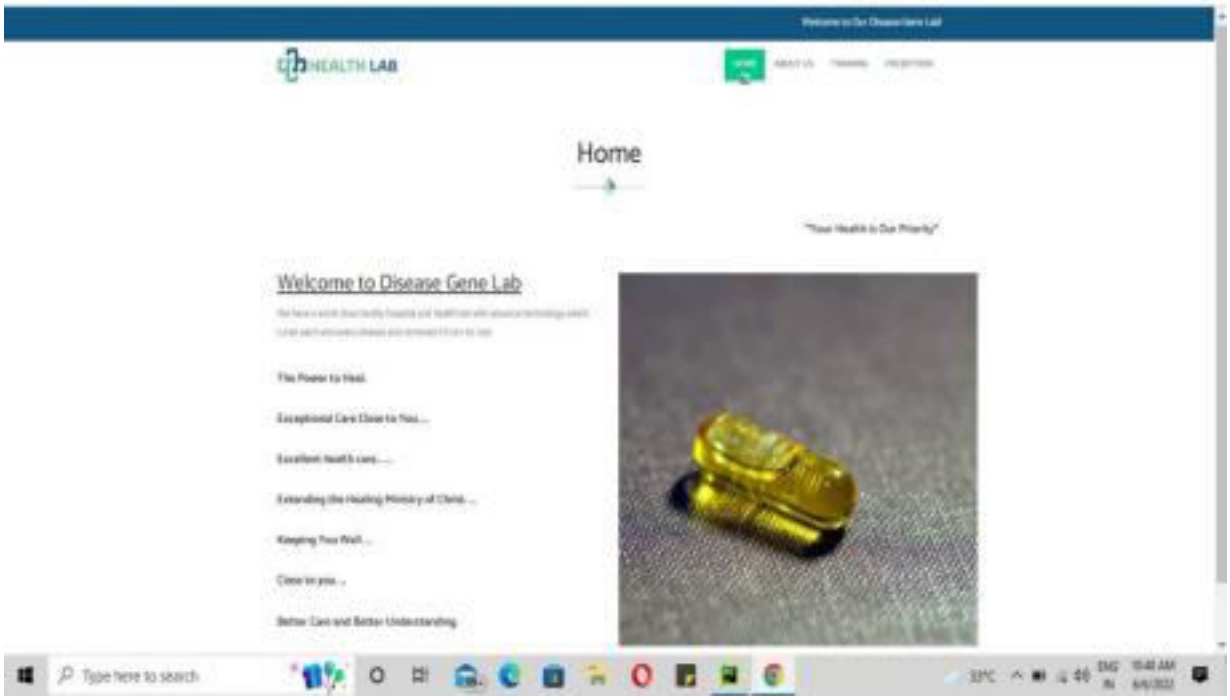
It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks..

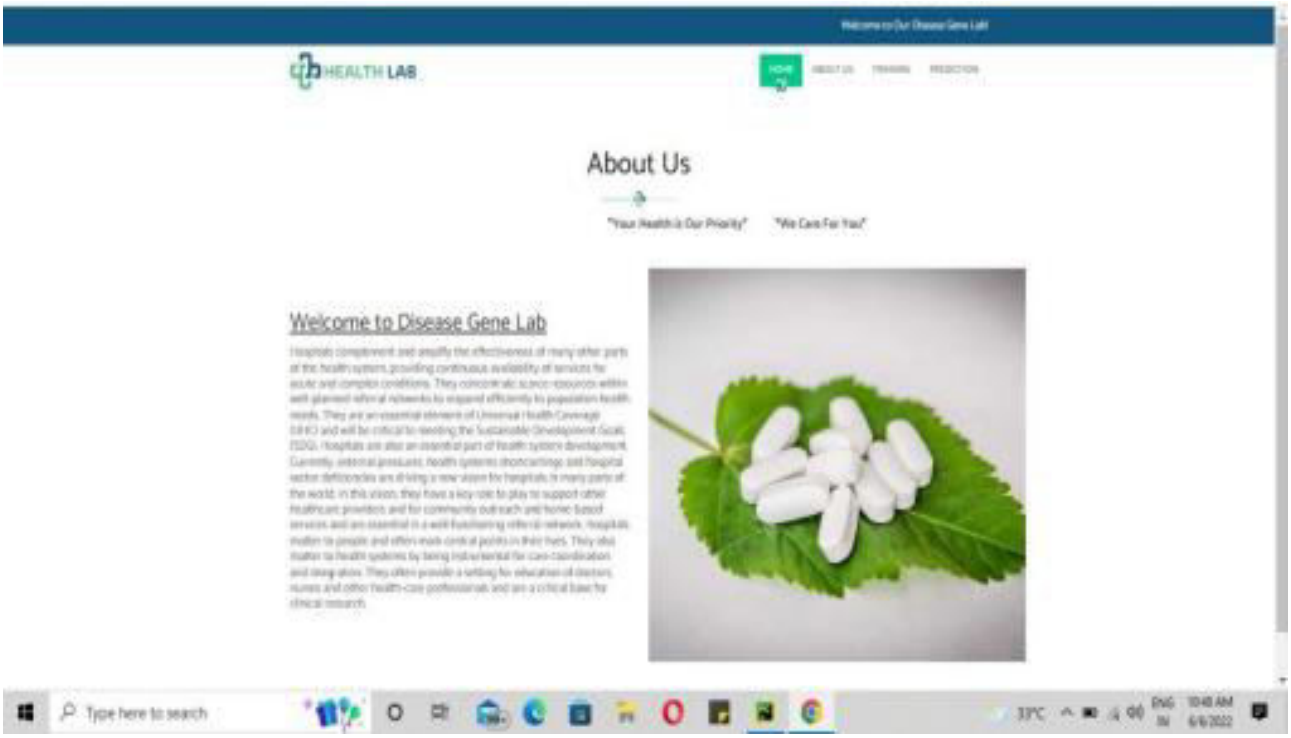
| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |

V. RESULT

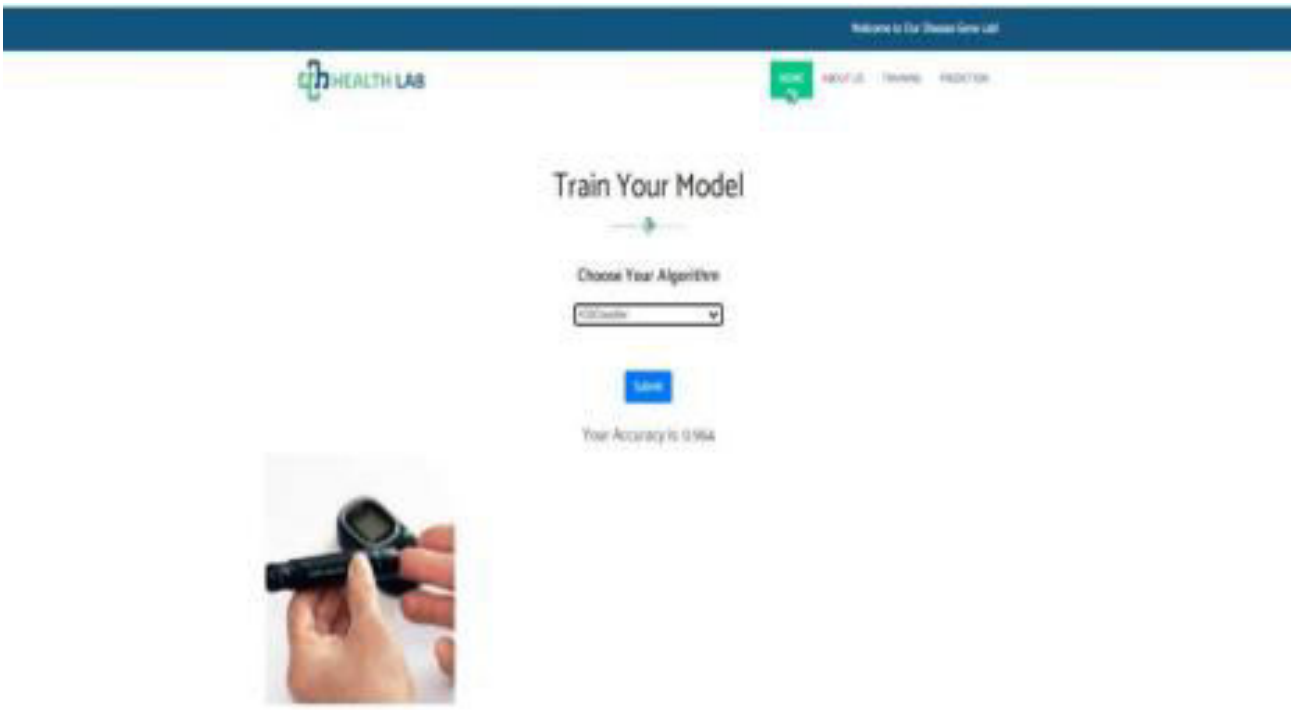




| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

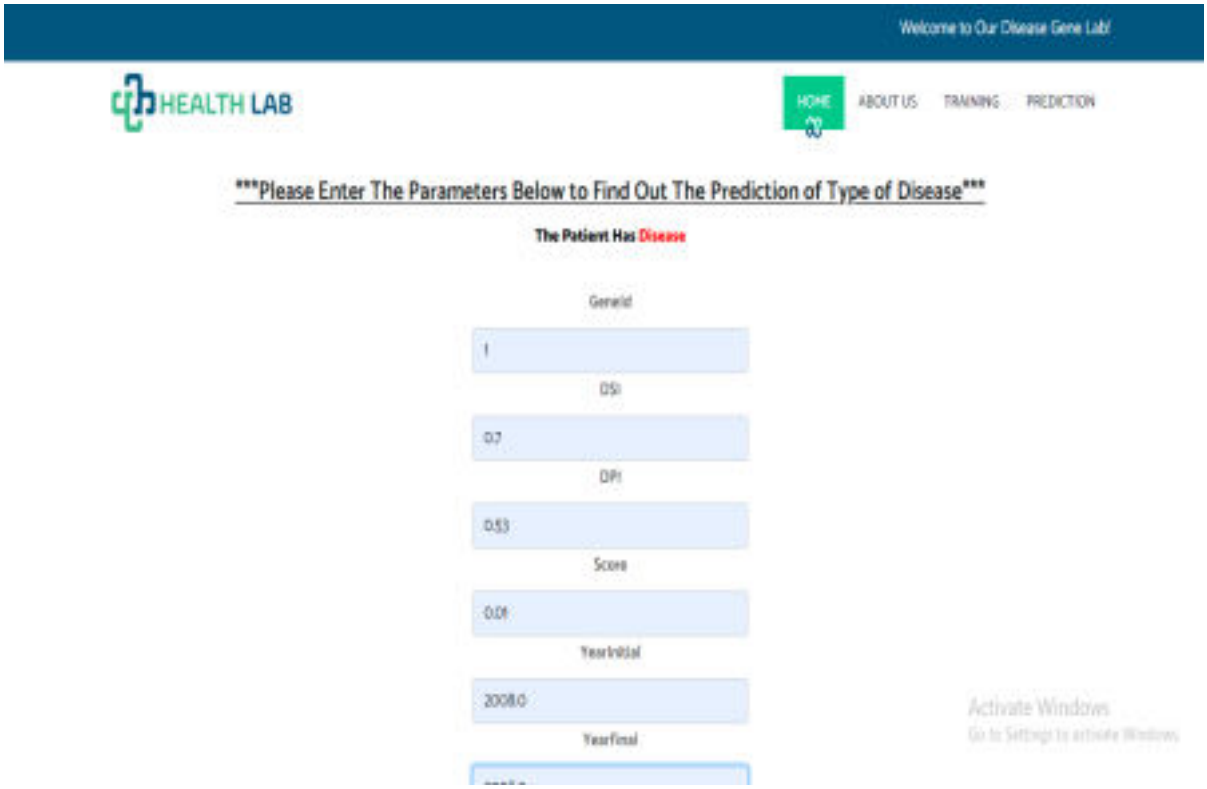
| DOI:10.15680/IJMRSET.2022.0506084 |



| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |



VI. CONCLUSION

In this application, we have successfully created machine learning models to predict the type of Disease Gene. This is developed in pycharm. We noticed that out of the XGBoost Classifier, Random Forest Classifier, Light GBM, K-Nearest Neighbor and Support Vector Classifier, XGBoost Classifier performs well with better accuracy.

FUTURE SCOPE:

This system can be extended to build the neural network. Additionally, clustering may help to Cluster them into groups

REFERENCES

[1] C. K. Saket Navlakha, "The power of protein interaction networks for associating genes with diseases," Bioinformatics, vol. 26, no. 8, p. 1057–1063, 2010.
[2] O. M. E. R. T. S. R. S. Oron Vanunu, "Associating Genes and Protein Complexes with Disease via Network Propagation," PLoS Computational Biology, vol. 6, no. 1, pp. 1-9, 2010.
[3] M. S. Mabrouk, "A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques," American Journal of Biomedical Engineering, vol. 2, no. 2, pp. 17-22, 2012.
[4] G. M. M. T. P. T. A. P. C. Y. J. F. F. R. S. V. R. B. T. D. M. P.-Y. K. C. A. M. F. P. R. Rahul C Deo, "Prioritizing causal disease genes using unbiased genomic features," Genome Biology, vol. 15, no. 12, pp. 1-19, 2014.
[5] D. M. Z.-H. D. Adarsh Jose, "A gene selection method for classifying cancer samples using 1D discrete wavelet transform," International Journal of Computational Biology and Drug Design, vol. 2, no. 4, pp. 398-411, 2009..
[6] R. J. M. Q. Z. S. L. Xuebing Wu, "Network-based global inference of human disease genes," Molecular Systems Biology, vol. 4, pp. 1- 11, 2008.
[7] S. B. T. M. M. H. S. Yu Qian, "Identifying disease associated genes by network propagation," BMC Systems Biology, vol. 8, no. 1, pp. 1- 7, 2014.
[8] W.A.U.I.X.W.M.S.L.Y.Z.J.Z.C. Aisha Sikandar, "Decision Tree Based Approaches for Detecting Protein Complex in Protein Protein Interaction Network (PPI) via Link and Sequence Analysis," IEEE Access, vol. 6, pp. 22108-22120, 2018.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506084 |

[9] Y. S. S. N. F. A. T. A. N. T. Qura-Tul-Ein, "DNA Pattern Analysis using Finite Automata," International Research Journal of Computer Science, vol. 1, no. 2, pp. 1-4, 2014.
[10] Y. L. Jianzhen Xu, "Discovering disease-genes by topological features in human protein-protein interaction network," Bioinformatics, vol. 22, no. 22, pp. 2800-2805, 2006.

BIOGRAPHY

Sannareddy.Hemalatha is a B.E, Final Year student in the department of Computer Science and Engineering from Velammal Institute of Technology, Panchetti.Her current research focuses on disease gene classification by extreme gradient boosting classifier.

E.V Sai Harshitha is a B.E, Final Year student in the department of Computer Science and Engineering from Velammal Institute of Technology, Panchetti.Her current research focuses on disease gene classification by extreme gradient boosting classifier.

Mr.Anbumani.A, M.Tech is an assistant professor of computer Science and Engineering Department in Velammal Institute of Technology,Panchetti.

Impact Factor
7.54

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com