IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ADVANCE GENOME DISORDER PREDICTION MODEL EMPOWERED WITH MACHINE LEARNING

Rathod Sai Vamshi Krishna[1], Balakrishna Maruthiram[2]

[1]PG Scholar, Department of IT, [2]Asst Professor,

University College Of Engineering Science And Technology, Jawaharlal Nehru Technological University Hyderabad

ABSTRACT

Predicting genomic disorder is a significant and crucial problem in biomedical research. Multivariate diseases with high global death rates, such as cancer, dementia, diabetes, cystic fibrosis, Leigh syndrome, etc., are brought on by genome defects. Historically, theoretical and explanatory Methods for predicting genetic disorders were introduced. As technology advanced, genetic data were expanded to nearly encompass the entire genome and proteins. Subsequently, machine learning and deep learning techniques were employed to forecast disorders related to the genome. Deep learning and machine learning techniques were introduced in parallel. Numerous studies on the prediction of genomic disorders have been carried out in the past utilizing supervised, unsupervised, and semi-supervised learning techniques; the majority of these studies used genetic sequence data to predict binary problems. These techniques' prediction findings were unclear because to due to their reduced accuracy rate and binary class prediction methods that use genome sequence data but do not use the medical history of patients with genome disorders. Consequently, an advanced genomic disorder prediction model (AGDPM) was developed in this study using a huge quantity of data by utilizing XGBoost and SVM, an efficient Machine Learning architecture. When compared to the pre-trained XGBoost model, AGDPM produces the greatest results, with training and testing accuracy rates respectively. Therefore, the advanced genome disease prediction model demonstrates the capacity to process a substantial quantity of data and forecast genomic disorders effectively. genetic disease data from patients using a multi-class prediction technique. According to a number of statistical performance metrics, AGDPM has demonstrated its ability to predict single gene inheritance disorders, mitochondrial gene inheritance disorders, and multifactorial gene inheritance disorders. Therefore, biomedical research will be enhanced with the aid of AGDPM in order to forecast genetic illnesses and control excessive death rates.

1. INTRODUCTION

Each syndrome has unique phenotypic characteristics that are the biological expressions of the underlying genes, and each one varies somewhat from the others. This facilitates phenotype-gene identification.

Interactions are an essential biological process that help scientists and clinicians comprehend the pathogenetic mechanisms underlying the disorders. Finding the genes causing particular illnesses makes patient diagnosis easier and sheds light on the workings of the network of linkages and mutation. Put differently, by examining the causative mutant genotypes throughout the disease gene discovery procedure, a possible hereditary ailment can be identified. Single nucleotide changes, additions or deletions of a single nucleotide, complete gene loss, and other genetic disorders can all be found in these disease-causing genes. Traditionally, positional cloning, linkage analysis, and mutation analysis have been utilized to find the genes causing the disease. Using linkage analysis on human pedigrees, the vulnerable chromosomal interval—roughly where the disease-associated candidate genes are located—is first discovered. Second, a discussion is had regarding the sequencing of a set of putative genes in the field using positional cloning. This technique includes both transcription mapping and physical mapping. A human genetic disorder is a prenatal genetic condition brought on by anomalies in the genes or chromosomes. There are two categories of genetic disorders: complicated disorders and single-gene illnesses. A single mutation in the structure of deoxyribonucleic acid causes a single gene illness, which has serious consequences due to a single fundamental deficiency. These diseases are easily inherited by subsequent generations.

When compared to the XGBoost produces the greatest results, with training and testing accuracy rates respectively. Therefore, the advanced genome disease prediction model demonstrates the capacity to process a substantial quantity of data and forecast genomic disorders effectively. genetic disease data from patients using a multi-class prediction technique. According to a number of statistical performance metrics, XGBoost has demonstrated its ability to predict single gene inheritance disorders, mitochondrial gene inheritance disorders, and multifactorial gene inheritance disorders. Therefore, biomedical research will be enhanced with the aid of XGBoost in order to forecast genetic illnesses and control excessive death rates.

SCOPE

To anticipate genetic illnesses by using a streamlit web framework for category type detection and to reduce high death rates by taking preventative measures. With a multi-class prediction method, the advanced genome disorder prediction model can process a significant amount of patient data related to genome disorders and demonstrates the ability to predict genome disorders quickly. According to a number of statistical performance metrics, AGDPM has demonstrated its ability to predict single gene inheritance disorders, mitochondrial gene inheritance disorders, multifactorial gene inheritance disorders, and Disorder Subclass.

2. OBJECTIVE

Utilizing a number of statistical performance metrics, the XGBoost forecasted the multifactorial gene inheritance disease simulation findings. Genetic illnesses can also be multifactorial, meaning that only a fraction of the phenotypes linked to the disorders are caused by genetic abnormalities. These complicated diseases represent the harmful effects of a combination of genetic abnormalities, lifestyle choices, and environmental factors. A genetic mutation in a single gene is the cause of a single gene condition. Single-gene disorders are extremely diverse and can affect many elements of functioning because they can occur in any gene. All single-gene disorders share a common biological basis, can be passed on to progeny, and require crucial genetic and counseling treatments, regardless of their clinical differences. There are five to ten circular deoxyribonucleic acid segments in each mitochondrial genome. When they fertilize, their organelles remain in the eggs. This means that the sickness is always transferred from mother to child. Ocular abnormalities, lactic acidosis, mitochondrial encephalopathy, and stroke-like events are caused by the genetic mitochondrial disease. Genetic disorders have several causes, including abnormalities in multiple genes. These illnesses are usually the consequence of complex interactions between environmental and nutritional factors. Another name for it is polygenic or complex illness. A complex hereditary illness is the root cause of diabetes, Alzheimer's disease, and cancer. Machine learning is a genetic prediction technique that is distinct from previous approaches. Advances in machine learning, together with growing data sets and processing power, have increased its appeal. The growing amount of data sets and computing power, along with the breakthroughs in machine learning, have increased its appeal.

2.1 EXISTING SYSTEM

Complex disorders involving several genes, such as single gene inheritance disorder (SGID), mitochondrial gene inheritance disorder (MGID), and multifactorial genome disorder (MGD), may exhibit a wide range of symptoms.

More precise genetic data collecting has been made possible by recent developments in genomic technology. Numerous individuals with anomalies have been uncovered by extensive genetic research, such as those conducted on SGID and MGD [4], [5]. Even with the massive amount of data generated by this study, identifying the exact genes that cause disease has proven to be a difficult process [6]. Similar symptoms are often caused by different abnormalities within the same disorder module, which has led to the suggestion that genetic data can be very informative [7]. Furthermore, knowledge-based techniques were applied to phenomenon networks, where genes are linked to endpoints if they demonstrate forecast gene-disease relationships. Genes that co-occur within known gene-disease association data are used to obtain gene-gene mutual information. By embedding the heterogeneous network made up of genes and disorders, as well as their unique features, Li et al. [16] created a novel technique. To further study the impact of the proposed Checkerboard Dropout on the generalization of object localization and small object recognition, the gradient class activation mapping (Grad-CAM) from ResNet-50 baseline, ResNet-50 with Drop Block and ResNet-50 with the proposed
Checkerboard Dropout are visualized.

Disadvantage of Existing System

- The proposes Checkerboard Dropout to deal with the over fitting problem.
- The Checkerboard Dropout is an efficient structured dropout technique to alleviate the problems of randomness and spatial correlation while improving the generalization of the mode.
- It makes an uncertain prediction rate.

2.2 PROPOSED SYSTEM

A wide range of symptoms can be present in complex illnesses, such as Multifactorial Genome Disorder (MGD), Mitochondrial Gene Inheritance Disorder (MGID), and Single Gene Inheritance Disorder (SGID), which involve many genes. Numerous people with these illnesses have been identified in extensive genetic studies as a result of recent advances in genomic technology that have made it easier to collect genetic data with greater precision. Even with the massive amount of data that these studies have produced, it is still difficult to identify the precise genes that cause the disorders. For instance, the disease is always transferred from mother to child in cases of mitochondrial abnormalities, which are transmitted maternally since the organelles are preserved during fertilization. The XGBoost and SVM technique is used to address the difficulty of gene identification by utilizing its potent machine learning capabilities. the ability to evaluate the vast amount of genetic data. By increasing the precision of identifying disease-causing genes, our method seeks to shed more light on the genetic foundations of these intricate illnesses. method for prioritizing disease genes using graph convolutional networks, or PGCN. Yang et al. [17] have created a unique deep neural network model based on genotype and phenotype variables.

Advantages of Proposed System

- Its capacity to carry out feature engineering on its own. Its use of a gradient descent approach to minimize the loss when adding new models.
- Using patients' clinical features as its foundational data, the suggested model XGBoost Algorithm obtained 92.65% prediction accuracy.
- The suggested model had a perfect space and computational complexity, and it employed the ideal XGBoost Algorithm to predict this disease.
- It achieved a high degree of accuracy in forecasting results.

3. RELATED WORKS

complicated illnesses with a large gene count, such as mitochondrial gene inheritance disease (MGID), multifactorial genome disorder (MGD), and single gene inheritance disorder (SGID), can have a variety of symptoms. More accurate sources of genetic data have emerged from latest advancements in the field of genetic technology. Large-scale genetic studies, such those for SGID and MGD, have identified hundreds of individuals with abnormalities [4], [5]. Finding the precise genes that cause disease has proven to be a challenging task, despite the substantial amount of data produced by this investigation [6]. Given that distinct disruptions within a same disorder module usually result in comparable phenotypes [7], genetic data have been suggested to be very revealing. Additionally, phenomena networks—where genes are attached endpoints if they show related phenotypic states have a strong relationship with proteins. transcription factor networks and genome connections [8]. Moreover, distinct phenotypes are induced by anomalies discovered in the interactome's remote neighbors [6]. Numerous methods that take into account these various kinds of data have been developed for using genes to predict disease [9]. The data is combined using a number of algorithms into a single graph, which is then used for prediction. Fundamental research principles state that when genetic variations involved in intermediary variables are included in a dependent variable that also includes these intermediate components, the genetic variants will no longer have any significance. Genetic polymorphisms have the ability to improve illness prediction beyond established risk factors when they are involved in previously unidentified pathways or processes with detectable intermediate components.

4. METHODLOGY OF PROJECT

The high-level Python web framework Stream lit will be used in the project's implementation. Early genetic disease diagnosis is very helpful for doctors and the biomedical industry when it comes to recommending pharmaceuticals for therapy. In this work, we suggest using AGDPM to identify multi-class genetic anomalies early on. This study's progression using the XGBoost algorithm's trained model and AGDPM is shown. Because the prediction output includes single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder without the need for a doctor, this technique will employ a streamlit framework to make the user participate.

MODULE'S

1)Data Source:

The initial and most important step in developing a machine learning model is data collection. The model's effectiveness will increase with the amount and quality of data we collect during this critical phase, which will have a cascading effect on the model's effectiveness.

Many techniques, such as physical interventions, online scraping, and others, can be used to collect data.

2)Dataset:

The genome disorders dataset is available for public access on Kaggle [12]. The history of single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder for 22083 patients is contained in the genome disorder data. There are 43 independent variables in this set of data, 3 classes, and 1 dependent variable.

To obtain the optimal features, we so used a number of outliers, regression, and normalization techniques. The top 24 retrieved features for predicting genomic disorders are listed in Table 5. Thus, AGDPM trains and tests the model using this feature-based dataset. Patient Id – patient id which is containing Genetic Disorder

1. Patient Age – age of the patient/ user
2. Genes in mother's side – genes which are present from mother side or not
3. Inherited from father – father/ Parents pass on traits or characteristics, such as eye colour and blood type, to their children through their genes.
4. Maternal gene - Maternal genes are from RNA
5. Paternal gene - Paternal inheritance refers to the transmission of any attribute from a father to his offspring.
6. Blood cell count (mcL) - A measure of the number of platlets in the blood.
7. Patient First Name – surname of the patient
8. Family Name – father's name
9.Father's name – mother's name
10. Mother's age – mother's age
11. Father's age – father's age
12. Institute Name – institute or hospital name
13. Location of Institute - institute or hospital location
14. Status – is that person/ patient alive or dead
15. Respiratory Rate (breaths/min) - The respiratory rate is the rate at which breathing occurs; it is set and controlled by the respiratory centre of the brain.
16. Heart Rate (rates/min) - Heart rate (or pulse rate) is of the heartbeat calculated by the number of contractions of the heart per minute (beats per minute, or bpm).
17.Test 1 – test1 is done or not
18. Test 2 – test2 is done or not
19. Test 3 – test3 is done or not
20. Test 4 – test4 is done or not
21. Test 5 – test5 is done or not
22. Parental consent - Parental consent laws (also known as parental involvement laws) in some countries require that one or more parents consent to or be notified before their minor child can legally engage in certain activities.
23. Follow-up – fellow-up is high or low
24. Gender - Female or Male or Ambiguous
25. Birth asphyxia - Asphyxia or asphyxiation is a condition of deficient supply of oxygen to the body which arises from abnormal breathing. Asphyxia at birth
26. Autopsy shows birth defect (if applicable) - An autopsy (post-mortem examination, obduction, necropsy, or autopsia cadaverum) is a surgical procedure that consists of a thorough examination of a corpse by dissection to determine the cause, mode, and manner of death or to evaluate any disease or injury that may be present for research or educational purposes.
27. Place of birth – the place of birth
28. Folic acid details (peri-conceptional) - Folic acid is a B vitamin. It helps the body make healthy new cells.
29. 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
30. H/O radiation exposure (x-ray) - Represents whether a patient has any radiation exposure history
31. H/O substance abuse - Represents whether a parent has a history of drug addiction
32. Assisted conception IVF/ART - Represents the type of treatment used for infertility
33. History of anomalies in previous pregnancies – any history of unknown things in previous pregnancies yes or no
34. No. of previous abortion – number of previous abortions
35. Birth defects - Represents whether a patient has birth defects
36. White Blood cell count (thousand per microliter) – number of White Blood cell count
37. Blood test result - Blood test result normal, slightly abnormal, inconclusive and abnormal
38. Symptom 1 - Symptom 1 yes or no
39. Symptom 2 - Symptom 2 yes or no
40. Symptom 3 - Symptom 3 yes or no
41. Symptom 4 - Symptom 4 yes or no
42. Symptom 5 - Symptom 5 yes or no
43. Genetic Disorder – Genetic Disorder Detection from professional doctor
44. Disorder Subclass – disorder type

3) Data Preparation:

To reduce the influence of the particular order in which the data was collected and/or produced, it should be randomized. To perform further exploratory research, find relevant relationships between variables, use data visualization.

Separated into groups for evaluation and training.

4) Model Selection:

We implemented the Extreme Gradient Boosting Algorithm and also the Support Vector Machine algorithms after achieving 98% and 80% accuracy on the train set.

5) Analyze and Prediction:
In the actual dataset, we chose only 2 features:
1. Description - detailed values of the health.
2. Outcome - indicates which type of the Genome Disorder for that person/ patient contains.

6) Accuracy on test set:
We achieved 92.65% accuracy on test dataset.

5. DATA FLOW DIAGRAM

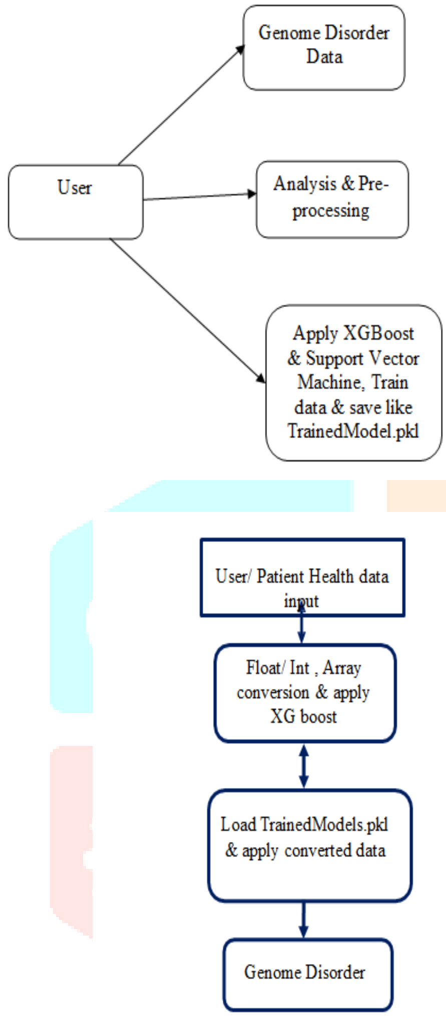

Genome Disorder Data
User
Analysis & Pre-processing
Apply XGBoost & Support Vector Machine, Train data & save like TrainedModel.pkl
User/ Patient Health data input
Float/ Int , Array conversion & apply XG boost
Load TrainedModels.pkl & apply converted data
Genome Disorder

Fig: Flow Diagrams of Modules

Impediments :
In the fields of genomics and medical research, predicting genome disorders is an essential task. While deep neural networks (DNNs) have shown a lot of promise in solving this issue, their performance and generalization are limited by their propensity for overfitting. In Convolutional Neural Networks (CNNs), dropout—a widely used technique to counter overfitting—has its limitations because of the enhanced spatial correlation of zeroed-out values in the output feature maps. In order to address the spatial correlation issue and enhance generalization and performance, the current system suggests the Checkerboard Dropout as an effective structured dropout strategy. But even with all of its benefits, there can be issues with the Checkerboard Dropout that need to be resolved.

6. ALGORITHM USED IN PROJECT

XGBoost Algorithm: Extreme Gradient Boosting, or XGBoost, is a technique that minimizes a regularized objective function by combining complexity with a convex loss function based on the difference between the predicted and target outputs. Iteratively, the training process adds new trees that fore cast the residuals or errors of earlier trees, which are merged with earlier trees to get the final prediction.

XGBoost is widely utilized in several field such as text mining, and recommender system, due to its accuracy, speed, and scalability.

Data on genetic disorders are used as input features by AGDPM

SVM ALGORITHM :
For classification, the Support-Vector-Machine(SVM) is a supervised learning technique.

1. **Hyper Plane :** To divide the data points in to different groups, SVM search for the most suited hyperplane. In two dimensions, this is a line; in higher dimensions, it is a plane or huperplane.
2. **Margin :** The separation between the closest data points from each class and the hyperplane.To guarantee the optimal degree of separation between classes, SVM seeks to maximize this margin.
3. **Support Vectors:** The nearest data points to the hyperplane, which affect its orientation and location. Since they establish the gap, these points are crucial.
4. **Kernel Trick:** SVM can effectively handle non-linear data by converting it into a higher- dimensional space where a linear function can be applied. There is a separator available. Radial basis function(RBF), polynomial, and linear kernels are examples of common kernels.

Fig: Support Vector Machine

7. SYSTEM ARCHITECTURE

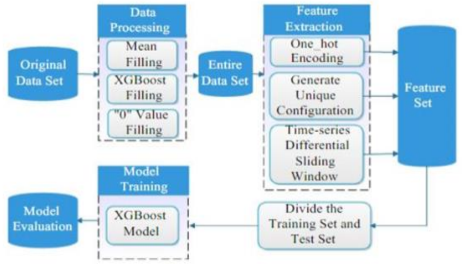
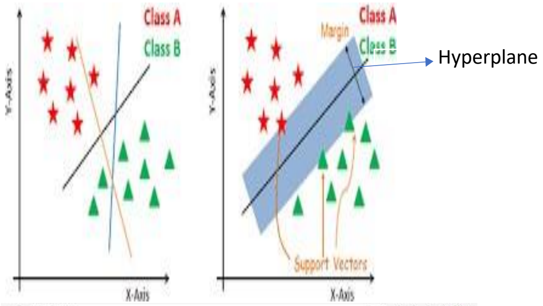
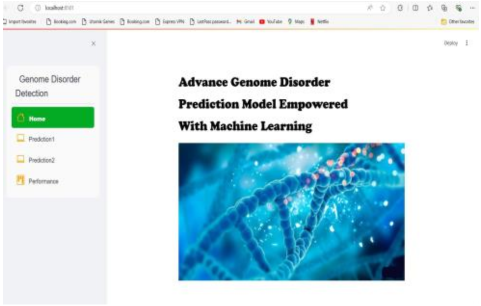


Fig: System Architecture of Project

8. RESULTS




Genome Disorder Detection

Home

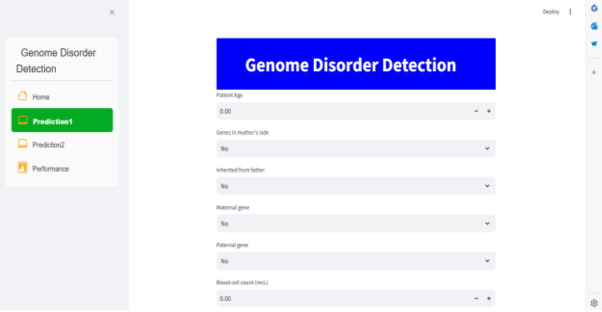
Prediction1

Prediction2

Performance

Advance Genome Disorder Prediction Model Empowered With Machine Learning

Fig: Home page



Genome Disorder Detection

Patient Age: 0.00

Genes in mother's side: No

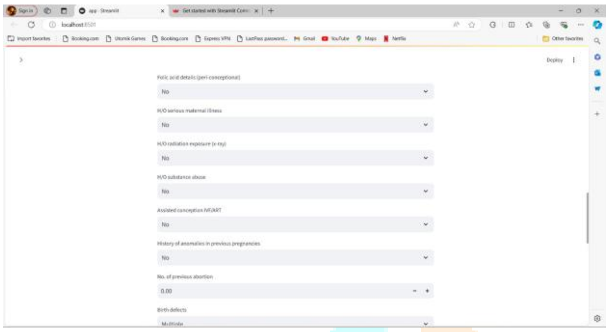
Inherited from father: No

Maternal gene: No

Paternal gene: No

Blood cell count (mcL): 0.00

Fig: Prediction 1



Folic acid details (peri-conceptional): No

H/O serious maternal illness: No

H/O radiation exposure (x-ray): No

H/O substance abuse: No

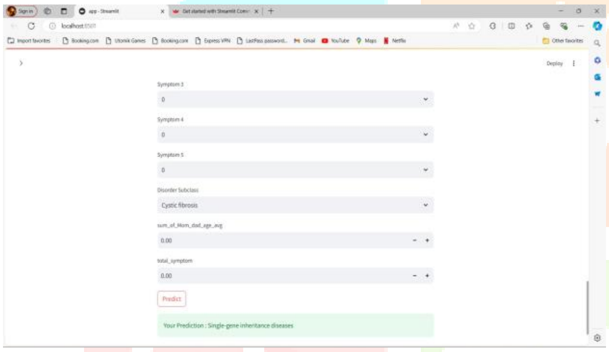
Assisted conception IVF/ART: No

History of anomalies in previous pregnancies: No

No. of previous abortion: 0.00

Birth defects: Multiple

Fig: Input Values



Symptom 3: 0

Symptom 4: 0

Symptom 5: 0

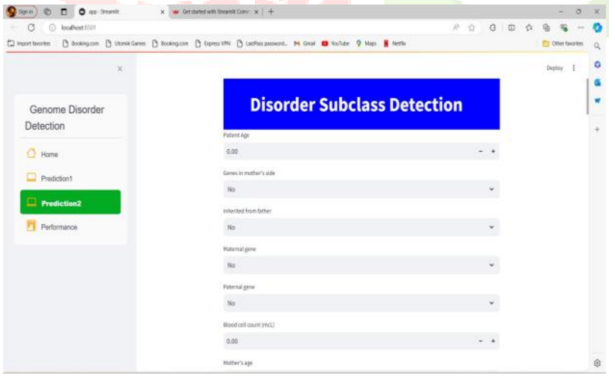
Disorder Subclass: Cystic fibrosis

sum_of_Mom_dad_age_avg: 0.00

total_symptom: 0.00

Your Prediction : Single-gene inheritance diseases

Fig: Predicting of Genome Disorder



Disorder Subclass Detection

Patient Age: 0.00

Genes in mother's side: No

Inherited from father: No

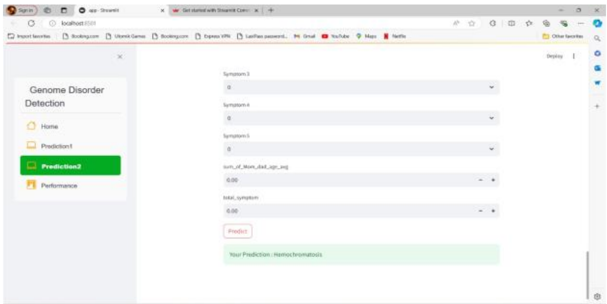
Maternal gene: No

Paternal gene: No

Blood cell count (mcL): 0.00

Mother's age

Fig: prediction 2



Symptom 3: 0

Symptom 4: 0

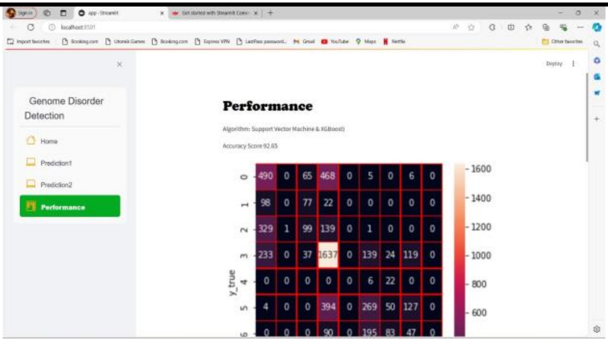
Symptom 5: 0

sum_of_Mom_dad_age_avg: 0.00

total_symptom: 0.00

Your Prediction : Hemochromatosis

Fig: Predicting of sub-class Detection



Performance

Algorithm: Support Vector Machine & XGBoost)

Accuracy Score 92.65

0	490	0	65	468	0	5	0	6	0
1	98	0	77	22	0	0	0	0	0
2	329	1	99	139	0	1	0	0	0
3	233	0	37	1637	0	139	24	119	0
4	0	0	0	0	0	6	22	0	0
5	4	0	0	394	0	269	50	127	0
6	0	0	0	90	0	195	83	47	0

Fig: Performance Analysis

9. FUTURE ENHANCEMENT

The Automated Genetic Disorder Prediction Model, or AGDPM, has demonstrated its efficacy in predicting single gene inheritance disorders, mitochondrial gene inheritance disorders, multifactorial gene inheritance disorders, and Disorder Subclass, through validation against several statistical performance indicators. This model can be extended in the future to include chromosomal abnormalities and polygenic disorders, among other genetic conditions. By combining many prediction models with stacking, accuracy can also be increased. Epigenetic information and advanced feature engineering methods can be used to further enhance predictions. Customizing genetic profiles for each patient and integrating it with practical Decision Support Systems (CDSS) can improve the model's practical usefulness. Gaining expertise in making predictions in real time, collaborating with genetic research organizations, and verifying the model through Long-term cohort studies will encourage ongoing development. Widespread adoption will depend on ensuring regulatory compliance and points out those pertaining to patient data protection and informed permission. XGBoost can develop into a comprehensive analysis for monitoring and forecasting genetic illnesses by concluding which will ultimately advance personalized medicine and enhance patient outcomes.

10. CONCLUSION

The field of bio-medical research has been widely impacted by the development of artificial intelligence. We used both the machine learning model and the newly created AGDPM model in this investigation. The AGDPM was developed using the XGBoost model. A range of statistical performance indicators were used to evaluate the model's performance after genome disorder data was collected from an internet repository. With 92.65% prediction accuracy, AGDPM outperforms ResNet-50 in the prediction of single-gene inheritance disorder, mitochondrial gene inheritance disorder, and multifactorial gene inheritance disorder. The AGDPM will have a significant positive impact on scientific research aimed at predicting genetic disorders. To get more accurate and better prediction results, this research can be expanded to include more hereditary disorders and many forecasts.

REFERENCES:

[1] McKusick-Nathans Institute of Genetic Medicine. Online Mendelian Inheritance in Man Johns Hopkins University School of Medicine. Accessed: Nov. 1, 2021. [Online]. Available: www.ncbinlmnih.gov/omim

[2] B. Irom, ‘‘Genetic disorders: A literature review,’’ Genet. Mol. Biol. Res., vol. 4, no. 2, p. 30, 2020.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ Commun. ACM, vol. 60, no. 2,
pp. 84–90, Jun. 2012.

[4] S. J. Sanders, ‘‘First glimpses of the neurobiology of autism spectrum disorder,’’ Current Opinion Genet. Develop., vol. 33, pp. 80–92, Aug. 2015.

[5] Europe PMC Funders Group, ‘‘Biological insights from 108 schizophrenia-associated genetic loci,’’ Nature, vol. 511, no. 7510, pp. 421–427, Jul. 2014.

[6] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabasi, ‘‘Uncovering disease-disease relationships through the incomplete interactome,’’ Science, vol. 347, no. 6224, Feb. 2015, Art. no. 1257601.

[7] A. L. Barabási, N. Gulbahce, and J. Loscalzo, ‘‘Network medicine: A network-based approach to human disease,’’ Nature Rev. Genet., vol. 12, pp. 56–68, Oct. 2011.

[8] M. Vidal, M. E. Cusick, and A.-L. Barabási, ‘‘Interactome networks and human disease,’’ Cell, vol. 144, no. 6, pp. 986–998, Mar. 2011.

[9] X. Wang, N. Gulbahce, and H. Yu, ‘‘Network-based methods for human disease gene prediction,’’ Briefings Funct. Genomics, vol. 10, no. 5, pp. 280–293, 2011.

[10] T.-P. Nguyen and T.-B. Ho, ‘‘Detecting disease genes based on semisupervised learning and protein–protein interaction networks,’’ Artif. Intell. Med., vol. 54, no. 1, pp. 63–71, Jan. 2012.

[11] P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh, and S. K. Ng, ‘‘Positive-unlabeled learning for disease gene identification,’’ Bioinformatics, vol. 28, no. 20, pp. 2640–2647, 2012.

[12] A. Rishabh. Of Genomes and Genetics HackerEarth Machine Learning Challenge. Kaggle. Accessed: Oct. 27, 2021. [Online]. Available:
https://www.kaggle.com/aryarishabh/of-genomes-and-geneticshackerearth-ml-challenge

[13] P. Han, P. Yang, P. Zhao, S. Shang, Y. Liu, J. Zhou, X. Gao, and P. Kalnis, ‘‘GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization,’’ in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2019, pp. 705–713.

[14] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, ‘‘Prediction and validation of disease genes using HeteSim scores,’’ IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 14, no. 3, pp. 687–695, May 2017.

[15] H. Zhou and J. Skolnick, ‘‘A knowledge-based approach for predicting gene–disease associations,’’ Bioinformatics, vol. 32, no. 18, pp. 2831–2838, Sep. 2016.

[16] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, ‘‘PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks,’’ bioRxiv, vol. 2019, Jan. 2019, Art. no. 532226, doi: 10.1101/532226.

[17] K. Yang, Y. Zheng, K. Lu, K. Chang, N. Wang, Z. Shu, J. Yu, B. Liu, Z. Gao, and X. Zhou, ‘‘PDGNet: Predicting disease genes using a deep neural network with multi-view features,’’ IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 1, pp. 575–584, Jan. 2022, doi: 10.1109/TCBB.2020.3002771.

[18] M. Alshahrani and R. Hoehndorf, ‘‘Semantic disease gene embeddings (SmuDGE): Phenotype-based disease gene prioritization without phenotypes,’’ Bioinformatics, vol. 34, no. 17, pp. i901–i907, Sep. 2018.

[19] K. Yang, R. Wang, G. Liu, Z. Shu, N. Wang, R. Zhang, J. Yu, J. Chen, X. Li, and X. Zhou, ‘‘HerGePred: Heterogeneous network embedding representation for disease gene prediction,’’ IEEE J. Biomed. Health Informat., vol. 23, no. 4, pp. 1805–1815, Jul. 2019.

[20] K. Yang, N. Wang, G. Liu, R. Wang, J. Yu, R. Zhang, J. Chen, and X. Zhou, ‘‘Heterogeneous network embedding for identifying symptom candidate genes,’’ J. Amer. Med. Inform. Assoc., vol. 25, no. 11, pp. 1452–1459, Nov. 2018.

[21] Y. Liu, H.-Q. Qu, X. Chang, L. Tian, J. Qu, J. Glessner, P. M. A. Sleiman, and H. Hakonarson, ‘‘Machine learning reduced gene/non-coding RNA features that classify schizophrenia patients accurately and highlight insightful gene clusters,’’ Int. J. Mol. Sci., vol. 22, no. 7, p. 3364, Mar. 2021.