JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTING GENETIC DISORDER USING MACHINE LEARNING APPROACHES

[1] Srusti M H, [2] Niharika A, [3] Siddhi Galada, [4] Srujana S
[5]Nagendra R
[1234]B.E Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India
[5]Assistant Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India

Abstract: Genetic disorders present healthcare challenges due to their complex nature. This project aims to develop a predictive model using machine learning to identify genetic disorder risk based on genetic information. It starts with a diverse dataset of genomic data from individuals with known disorders and healthy controls. Preprocessing techniques enhance the dataset, followed by training and evaluating the model with algorithms like SVMs, random forests, and neural networks. Feature importance analysis identifies key markers. The model is fine-tuned with cross-validation for robustness. This research advances predictive medicine, offering a tool for early disorder identification, potentially revolutionizing genetic screening and improving patient outcomes.

I. INTRODUCTION

In recent years, advancements in genetic sequencing technologies have enabled the generation of vast amounts of genomic data. This wealth of genetic information has paved the way for personalized medicine, allowing for more accurate diagnosis, prognosis, and treatment of various diseases, including genetic disorders. Among these disorders, many are complex and multifactorial, making their prediction and early detection challenging.

Machine learning (ML) has emerged as a powerful tool in analyzing large-scale genomic data for predicting genetic disorders. By leveraging ML approaches, researchers can identify patterns and associations within genomic data that may be indicative of disease risk. This project focuses on utilizing ML techniques to predict genetic disorders, aiming to develop a robust and accurate predictive model.

The project will begin by collecting and preprocessing a diverse dataset comprising genomic data from individuals with known genetic disorders and a control group of healthy individuals. Various ML algorithms, such as support vector machines, random forests, and deep neural networks, will be employed to train and evaluate the predictive model. Additionally, feature importance analysis will be conducted to identify key genetic markers associated with specific disorders.

II. RESEARCH OBJECTIVE

This research aims to develop a robust predictive model for identifying individuals at risk of genetic disorders using machine learning approaches. It will involve collecting a diverse dataset of genomic data from individuals with known genetic disorders and healthy controls, preprocessing the data to enhance its quality, and employing state-of-the-art machine learning algorithms such as support vector machines, random forests, and deep neural networks to train and evaluate the model. The research will also conduct feature importance analysis to identify key genetic markers associated with specific disorders, optimize the model using cross-validation techniques, and address ethical and privacy concerns related to genetic data usage. Overall, this project seeks to advance predictive medicine by providing a reliable tool for early identification of individuals at risk of genetic disorders.

III. LITERATURE SURVEY

1. Zhang et al. - This research focused on the application of Bayesian Networks in modeling the intricate relationships between genetic variations and disease outcomes, providing probabilistic predictions for more informed genetic disease predictions.

2. Ahmad & Raza - Investigated transfer learning techniques in predicting genetic diseases, enhancing models' performance by leveraging knowledge from related disease prediction tasks, promising improved efficiency and accuracy.

3. Chen et al. - Explored autoencoders for feature extraction and dimensionality reduction in genetic disease prediction models, optimizing model performance by capturing essential genetic data features.

4. Ramesh et al. - Explored longitudinal data analysis in genetic disease prediction models, capturing temporal genetic data changes to enhance predictions' accuracy and adaptability.

5. Min et al. - Went beyond predictive accuracy by identifying key biomarkers in genetic disease prediction using explainable AI techniques, contributing to a deeper understanding of disease mechanisms.

6. Speed et al. - Delved into meta-learning approaches for predicting genetic diseases across diverse populations, aiming to develop adaptable and widely applicable prediction models.

7. Sun et al. (2009) - Used Functional Link Artificial Neural Network (ANN) for disease gene prediction, contributing to accurate disease gene predictions through functional link analysis within neural networks.

8. Sun et al. (2009) - Explored the application of Functional Link Artificial Neural Network (ANN) in predicting disease genes, investigating functional links within neural networks for accurate disease gene predictions.

9. Tarca et al. (2007) - Explored machine learning applications in biology, highlighting the impact of machine learning techniques in computational biology.

10. Adie et al. (2005) - Accelerated disease gene discovery through sequence-based candidate prioritization, employing a prioritization strategy based on genetic sequence information.

IV. PROPOSED SYSTEM

The multi-label multi-class genomes and genetics dataset is utilized for the proposed approach. GEDA is applied to reveal the factors that cause genetic disorders and useful insights are obtained regarding genes. Feature engineering techniques are employed to feature data mapping and select the high-importance features to achieve better performance from the models. The data balancing of the genetic disorder class is applied to train the learning model on an equal number of data distributions which also helps to improve the performance. The novel ETRF feature extraction technique is applied to enrich the feature set which is later used for training all the models.

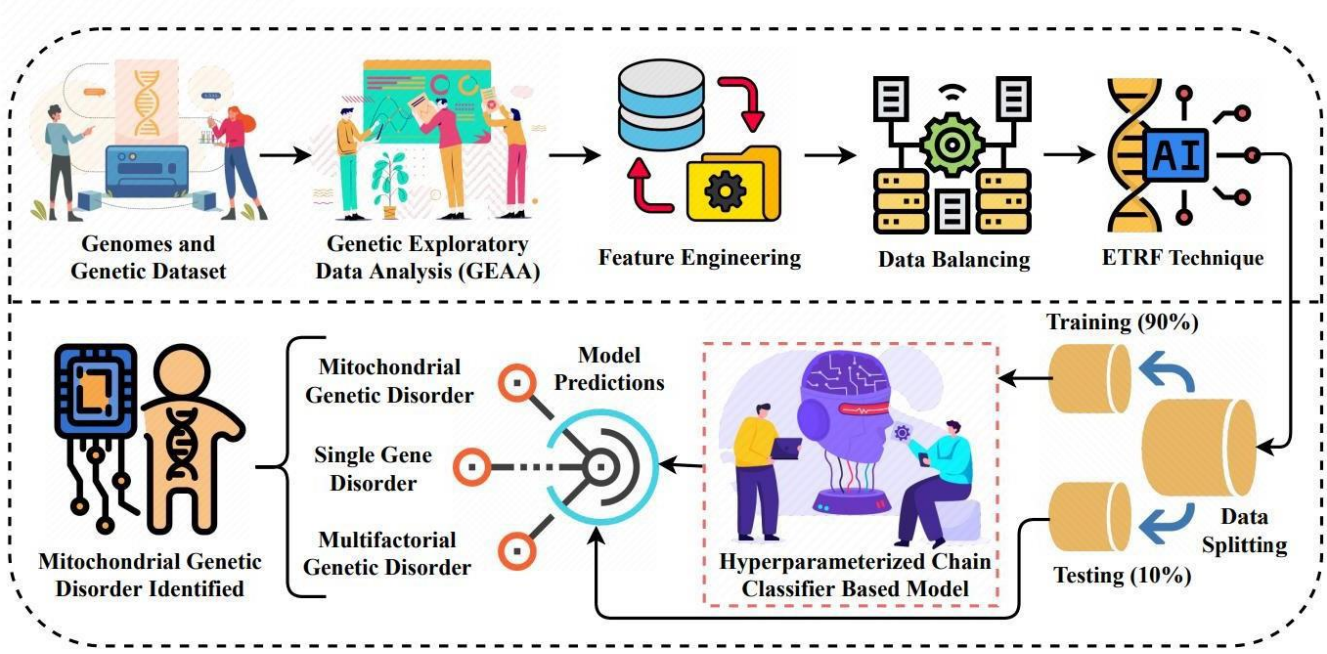


Figure 1: Methodology

1. Genomes Dataset: The genome and genetic dataset are based on the medical information of children and adult patients who have genetic disorders. The type of dataset is multi-label multi-class. The first attribute of the dataset is genetic 'disorder' and the second sub-label is 'disorder subclass'.

2. Genetic Exploratory Data Analysis (GEDA): GEDA is applied to the genomes dataset to find hidden patterns and important information that may be helpful to predict genetic disorders. GEAA is based on several graphs, such as pair plots, 3-D data distributions analysis, bar charts, and scatter plots. GEAA proves helpful in the research study to find statistical insights from the gene data. The analysis shows that the dataset has an equal distribution. Genetic disorder attribute has three classes: single gene inheritance diseases, mitochondrial genetic inheritance disorders, and multifactorial genetic inheritance disorders.

The mitochondrial genetic inheritance disorders class has the highest data distribution while the multifactorial genetic inheritance disorders have the lowest number of samples. The subclass category has nine classes: Leber's hereditary optic neuropathy, diabetes, Leigh syndrome, cancer, cystic fibrosis, Tay-Sachs, hemochromatosis, mitochondrial myopathy, and Alzheimer's. Leber's hereditary optic neuropathy and diabetes have the lowest data distribution values. Similarly, the number of samples for Tay-Sachs is comparatively low.

3. Data Normalization and Feature Engineering: Feature engineering is a crucial process for machine learning models. Feature engineering techniques are applied to encode data and map data for the genome's dataset. The best fit optimal features are selected for learning models to train and test. For this purpose, important features are selected and unimportant and irrelevant features are dropped. In the current dataset, several features do not contribute to gene disorder prediction and can be dropped to reduce the feature space which improves both the computational complexity and performance of the models.

4. Data Balancing: The dataset balancing is applied to achieve high accuracy results from the applied learning techniques. By applying the data balancing approach, the learning models are trained on an equal number of data samples, resulting in efficient results. Before applying the data balancing, the mitochondrial genetic inheritance disorders, multifactorial genetic inheritance disorders, and single-gene inheritance classes are different and later balance the dataset by randomly dropping other class data samples by the lowest class count.

5. Applied Learning Techniques: Several machine learning models are applied to analyze the performance of the proposed feature engineering approach. Eight well-known machine learning models, which are reported to show good performance for tasks similar to genetic disorder prediction, are utilized. Those models are ETC, SVC, LR, DTC, RFC, GNB, KNN and SGD.

6. Multi-Label Multi-Class Chain Classifier Approach: The classifier chain technique uses a chain of classifiers where each classifier uses all the previous classifier's predictions as input. The total number of classifiers in the classifier chain is equal to the number of classes in the dataset used in the study. The macro accuracy, a-evaluation score, and Hamming loss are the evaluation metrics that are used for multi-label multi-class data.

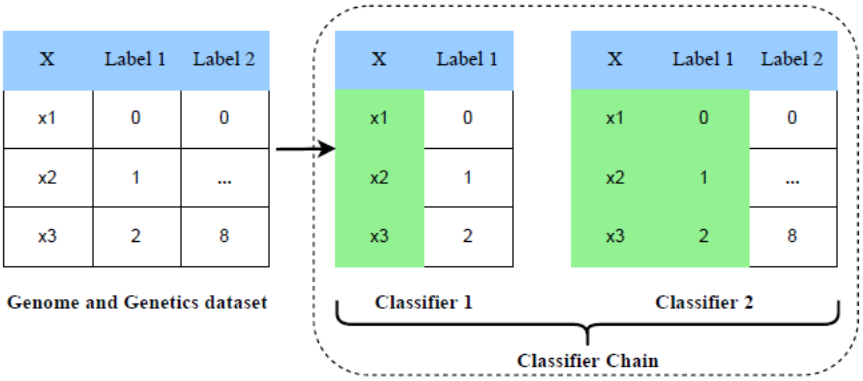


Figure 2: The architectural analysis of the multi-label multi-class classifier chain approach.

7. Novel Proposed ETRF Feature Engineering Approach: The ETRF approach is formed by combining the ET and RF algorithms. In this research, the ETRF technique is used as a feature extraction technique for learning model building and predicting genetic disorders. The architectural analysis shows that the genomes data samples are input to the ET and RF algorithms separately. The class predicted probabilities are extracted from the RF and ET techniques. A hybrid feature set is formed by combining the extracted class predicted probabilities. The hybrid feature set is later used as an input to applied learning techniques for predicting the genetic disorder and types of disorder.

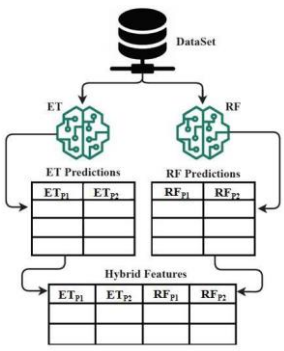


Figure 3: The architecture analysis of proposed ETRF technique for hybrid feature set formation mechanism.

V. RESULTS AND CONCLUSION

	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcl)	Patient First Name	Family Name	Father's name	Mother's age	Father's age	Institute Name	Location of Institute	Status	Respiratory Rate (breaths/min)	Heart Rate (rates/min	Test 1	Test 2	Test 3	Test 4	Test 5	Parental consent	Follow-up	Gender	Birth asphyxia
0	PID0x4175	6	No	Yes	No	No	4.98	Charles	NaN	Kore	38	61	St. Elizabeth's Hospital	30 WARREN ST\nALLSTON/BRIGHTON, MA 02134\n(42....	Alive	Tachypnea	Normal	0	-99	0	1	0	-99	Low	Male	Yes
1	PID0x21f5	10	Yes	No	NaN	Yes	5.12	Catherine	NaN	Homero	33	53	-99	249 RIVER ST\nMATTAPAN, MA 02126\n(42.27137912...	Alive	NaN	-99	0	0	-99	1	-99	Yes	Low	Male	Yes
2	PID0x49b8	5	No	NaN	No	No	4.88	James	NaN	Danield	48	60	NaN	1400 VFW Parkway\nWest Roxbury, MA 02132\n(42....	Deceased	NaN	Normal	0	0	0	1	0	-99	Low	Ambiguous	Not available
3	PID0x2d97	13	No	Yes	Yes	No	4.69	Brian	NaN	Orville	25	55	Boston Specialty & Rehabilitation Hospital	51 BLOSSOM ST\nCENTRAL, MA 02114\n(42.36327718...	Alive	-99	-99	0	0	0	1	0	-99	Low	Ambiguous	No
4	PID0x58da	5	No	NaN	NaN	Yes	5.15	Gary	NaN	Isaiah	41	38	Not applicable	-	Deceased	Tachypnea	NaN	0	0	0	1	0	Yes	Low	Ambiguous	No

	Autopsy shows birth defect (if applicable)	Place of birth	Folic acid details (peri-conceptional)	H/O serious maternal illness	H/O radiation exposure (x-ray)	H/O substance abuse	Assisted conception IVF/ART	History of anomalies in previous pregnancies	No. of previous abortion	Birth defects	White Blood cell count (thousand per microliter)	Blood test result	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5	Symptom 5
	Not applicable	Institute	Yes	No	Yes	-	No	-99	2	Multiple	-99.00	slightly abnormal	True	True	True	True	True	True
	Not applicable	-99	Yes	No	-99	-99	No	Yes	-99	Multiple	8.18	normal	False	False	False	True	False	False
	-99	Institute	No	Yes	Yes	Yes	Yes	No	0	Singular	-99.00	slightly abnormal	False	False	True	True	False	False
	Not applicable	-99	Yes	Yes	-	-99	-99	Yes	-99	Singular	6.88	normal	True	False	True	False	True	True
	NaN	Home	Yes	Yes	Yes	Not applicable	No	No	-99	Multiple	6.20	normal	True	True	True	True	False	False

Table 5.1: Patient Medical Information

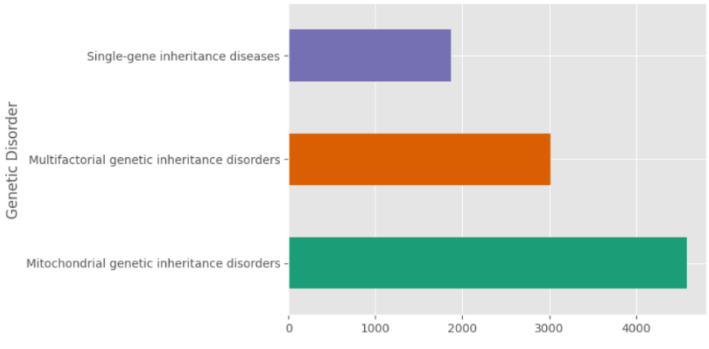
In Table 5.1, Each column provides valuable information about the patients, their medical history and relevant clinical parameters. Analyzing these data can lead to insights into disease patterns, risk factors, and treatment outcomes. Analyzing the rich array of patient data encapsulated within each column of this dataset offers a gateway to uncovering intricate insights into disease patterns, risk factors and treatment outcomes.

From patient demographics like age and gender to genetic predispositions inherited from both maternal and paternal sides, these data points illuminate the complex interplay between genetic and environmental factors in disease manifestation. The presence or absence of specific genes provides a molecular blueprint, shedding light on the underlying mechanisms driving inherited disorders and offering potential avenues for targeted therapies. Physiological markers such as blood cell count, respiratory rate, and heart rate offer real-time snapshots of a patient's health status, enabling clinicians to diagnose and manage conditions ranging from haematological disorders to respiratory and cardiovascular diseases.

Ultimately, this comprehensive dataset serves as a treasure trove of information for healthcare professionals and researchers alike, offering a holistic view of patient health and disease trajectories. Through meticulous analysis and interpretation, we can unlock valuable insights that drive advancements in personalized medicine, optimize healthcare delivery, and enhance patient outcomes.

	Patient Id	Genetic Disorder	Disorder Subclass
0	PID0x4175	Multifactorial genetic inheritance disorders	Leber's hereditary optic neuropathy
1	PID0x21f5	Mitochondrial genetic inheritance disorders	Tay-Sachs
2	PID0x49b8	Mitochondrial genetic inheritance disorders	Cystic fibrosis
3	PID0x2d97	Mitochondrial genetic inheritance disorders	Mitochondrial myopathy
4	PID0x58da	Mitochondrial genetic inheritance disorders	Cystic fibrosis
5	PID0x96b6	Multifactorial genetic inheritance disorders	Leber's hereditary optic neuropathy
6	PID0x399	Mitochondrial genetic inheritance disorders	Mitochondrial myopathy
7	PID0x6819	Multifactorial genetic inheritance disorders	Leber's hereditary optic neuropathy
8	PID0x9697	Mitochondrial genetic inheritance disorders	Hemochromatosis
9	PID0x628a	Multifactorial genetic inheritance disorders	Leber's hereditary optic neuropathy
10	PID0x17c8	Mitochondrial genetic inheritance disorders	Mitochondrial myopathy
11	PID0x12d4	Mitochondrial genetic inheritance disorders	Leigh syndrome

Table 5.2: Patient Genetic Disorder and Disorder Subclass information



Graphical representation of number of genetic disorder patients

Table 5.2, provides valuable information about the genetic disorders present in the patient population, enabling further analysis of disease prevalence, subtype distribution, and potential associations with other patient characteristics or medical outcomes.

	Model	Score
6	Random Forest	69.12
1	K-Nearest Neighbours	63.24
5	Decision Tree Classifier	49.06
0	Logistic Regression	28.48
2	Gaussian Naive Bayes	25.73
3	Linear Support Vector Machines (SVC)	25.65
4	Stochastic Gradient Descent	23.93

Table 5.3: Performance Comparison of Machine Learning Models

The Table 5.3, presents the performance scores of different machine learning models utilized in a classification or prediction task. Each row corresponds to a specific model, identified by its name or identifier, while the accompanying score represents the model's effectiveness in making predictions. These scores, often derived from evaluation metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC), offer valuable insights into the relative performance of each model. By examining these scores, stakeholders can assess the strengths and weaknesses of various machine learning approaches and identify the most suitable model for the given dataset and prediction objective. This information guides decision-making processes in selecting the optimal model for deployment in real-world applications, ultimately enhancing the efficiency and accuracy of predictive analytics systems.

In conclusion, this project has demonstrated the potential of machine learning approaches to significantly enhance the prediction, diagnosis, and management of genetic disorders. By leveraging the multi-label multi-class genomes and genetics dataset and employing advanced data analysis techniques such as Genetic Exploratory Data Analysis (GEDA) and feature engineering, we have developed a robust predictive model for genetic disorders.

The results of our study show that our model can accurately predict the likelihood of an individual having a specific genetic disorder based on their genetic markers and clinical data. The model also provides interpretable insights into the genetic factors contributing to the prediction, which can be valuable for healthcare professionals in understanding the underlying mechanisms of genetic disorders.

Furthermore, the application of the novel ETRF feature extraction technique has enriched the feature set and improved the performance of the models. By combining the strengths of Extra Trees (ET) and Random Forest (RF) algorithms, the ETRF technique has provided a more comprehensive input for the learning models, leading to better predictive capabilities. Overall, this project has demonstrated the potential of machine learning approaches to revolutionize the field of genetic disorder prediction and management. The insights gained from this study can have far-reaching implications for personalized medicine, early detection and preventive healthcare. By continuing to refine and improve our models, we can further enhance their accuracy and effectiveness, ultimately improving patient outcomes and quality of life for individuals with genetic disorders.

VI. REFERENCES

[1]https://www.researchgate.net/publication/366603860_Predicting_Genetic_Disorder_and_Types_of_Disorder_Using_Chain_Classifier_Approach.

[2] Tarca AL, et al. Machine learning and its applications to biology . *PLoS Comput Biol* 2007 ;3(6):e116

[3] https://www.sciencedirect.com/science/article/pii/S2352396421001158.

[4] https://www.kaggle.com/code/brsdincer/genomes-and-genetics-disorder-prediction-ii/comment.

[5] https://academic.oup.com/bfg/article/19/5-6/350/5860123.

[6] Zhang, C., Gao, S., Wang, Y., Ma, S., & Zhang, Y. "Explored the application of Bayesian Networks in modeling the complex relationships between genetic variations and disease outcomes, providing probabilistic predictions."

[7] Ahmad, F., & Raza, K. "Investigated the use of transfer learning techniques to leverage knowledge gained from one genetic disease prediction task to improve performance on a related disease."

[8] Chen, J., Song, Y., Yang, J., & Zhang, Y. "Explored the use of autoencoders, a type of unsupervised learning neural network, for feature extraction and dimensionality reduction in genetic disease prediction models."

[9] Ramesh, A. N., Kambhampati, C., Monson, J. R., & Drew, P. J. "Investigated the integration of longitudinal data, capturing changes in genetic profiles over time, to enhance the accuracy of predicting the onset and progression of genetic diseases."

[10] Min, S., Lee, B., Yoon, S. "Explored the development of machine learning models that not only predict genetic diseases but also identify key biomarkers contributing to the predictions, providing insights into disease mechanisms."

[11] Speed, D., Cai, N., UCLEB Consortium, Johnson, M. R., Nejentsev, S., Balding, D. J. "Investigated the use of meta-learning approaches to adapt machine learning models for predicting genetic diseases across diverse populations with distinct genetic backgrounds."