International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Genetic Disorder Prediction Using K-Nearest Neighbors Algorithm

Shreya Mahajan[1], Saylee Mahajan[2]

[1]Student, P. Shah Institute of Technology
[2]Doctor

Abstract:
There are many websites available which can be used to calculate the chances of certain diseases occurring in a person such as diabetes, Atherosclerotic Cardiovascular Disease and many others. People use these kind of websites to gain knowledge and rough insights about their health by inputting their body statistics. This project mainly focuses on the genetic aspect of the person's health. This paper will discuss the making of a website which states if the offspring carries the genetic disorder or not.

Keywords: Genetic Disorder, Web application, Predictor, Machine Learning, Healthcare, KNN

1. INTRODUCTION

Genetic disorders have long been a subject of significant concern within the realm of healthcare. These complex, multifaceted conditions have a profound impact on the lives of individuals and their families. Identifying, diagnosing, and predicting the status of genetic disorders is a crucial area of study, as it can lead to early intervention, more effective treatments, and improved patient outcomes. To address these challenges and harness the power of data- driven healthcare, the Genetic Disorder Prediction System has been developed. Genetic disorders encompass a wide range of conditions, each with its unique genetic and clinical characteristics. These disorders can manifest in various ways, from subtle symptoms to severe and life-threatening complications. Predicting the status of individuals affected by genetic disorders is a multifaceted task that requires the integration of clinical data, genetic information, and advanced computational techniques. The Genetic Disorder Prediction System has been designed to meet these challenges and provide a platform for accurate predictions. The primary purpose of this system is to leverage the power of machine learning and data preprocessing to predict the status of individuals with genetic disorders. By doing so, it contributes to the advancement of personalized medicine, where treatment plans can be tailored to an individual's specific genetic profile. This has the potential to revolutionise healthcare by enabling early diagnosis of multiple diseases, proactive intervention, and the optimization of treatment strategies.

A. *Diseases*

1. ***Cystic fibrosis:*** Cystic Fibrosis is a rare genetic disorder which affects the lungs, pancreas, kidneys and intestine. It affects multi-organ systems. It can be inherited in an autosomal recessive manner. CF is caused by a mutation in the *CF transmembrane conductance regulator (CFTR) gene.* When mutations occur in one or both copies of the gene, ion (sodium and chloride) transport is altered, and results in a buildup of thick mucus leading to defective mucociliary clearance. This defective

mucociliary clearance and altered ion transport leads to systemic obstruction and complications. Cystic fibrosis has no cure but a range of treatments can be used to inhibit the disease progression.

2. ***Leber's hereditary optic neuropathy (LHON):*** LHON is a rare disorder with mitochondrial pattern of inheritance (maternal inheritance). Mutations in the mitochondrial DNA (mtDNA) affect proteins located in mitochondrial membranes that are involved in cellular respiration through the process of oxidative phosphorylation. LHON affects young males with progressive visual loss due to optic neuropathy. It eventually leads to bilateral central vision loss. Pharmacologic and gene therapy treatment approaches have been shown to be safe and possibly effective in promoting some visual recovery.
3. ***Diabetes Mellitus:*** Diabetes Mellitus is a chronic disorder characterized by high blood glucose level due to insulin deficiency/resistance. Genetic (HLA- DR3, DR4, DQ) and environmental factors play a huge role in the pathogenesis of Diabetes Mellitus. DM can cause damage to various organ systems, leading to the development of disabling and life-threatening health complications, most prominent of which are microvascular (retinopathy, nephropathy, and neuropathy) and macrovascular complications leading to an increased risk of cardiovascular diseases. Symptoms develop gradually and include increased urination, increased thirst, fatigue, vision and nerve complications, and impaired wound healing. Type 1 DM has a strong genetic predisposition than Type 2 DM.
4. ***Leigh syndrome:*** Leigh syndrome is a rare genetic disorder which affects the nervous system (brain, nerves and spinal cord). In most cases, Leigh syndrome is inherited as an autosomal recessive trait. However, X-linked recessive and maternal inheritance, due to a mitochondrial DNA mutation, are additional modes of inheritance. It appears in infancy, causing developmental delays, hypotonia, poor reflexes, dementia, seizures et al. There is no cure, and treatments aim to manage symptoms. Prognosis is often poor, impacting a child's lifespan.
5. ***Cancer:*** Cancer refers to a group of diseases characterized by the uncontrollable growth of abnormal cells. It can occur anywhere in the body and is often caused by genetic mutations (BRCA gene mutation for breast carcinoma and APC gene mutation for colorectal carcinoma) or environmental factors. Treatments vary and may include surgery, chemotherapy, radiotherapy, immunotherapy, or a combination depending on the type and staging of cancer.
6. ***Tay-Sachs:*** Tay-Sachs is a rare genetic lysosomal storage disorder that primarily affects the nervous system It is caused by mutations in the genes encoding for β-hexosaminidase enzyme. This results in a accumulation of GM2 ganglioside in the brain and other nerve cells, leading to progressive damage. Symptoms often appear in infancy and include developmental regression, loss of motor skills, seizures, and vision and hearing problems. Unfortunately, there's no cure, and treatment focuses on managing symptoms.
7. ***Hemochromatosis:*** Hemochromatosis is a genetic disorder (HFE gene mutation-C282Y and H63D) characterized by excessive absorption of dietary iron, leading to an accumulation of iron in various organs, particularly the liver, skin, heart, and pancreas. Over time, this excess iron can cause damage to these organs leading to various complications. Symptoms might include fatigue, joint pain, abdominal pain, bronze/grey skin colour and weakness. Treatment typically involves regular removal of blood (phlebotomy) to reduce iron levels and manage symptoms.
8. ***Mitochondrial myopathy:*** Mitochondrial myopathy refers to a group of neuromuscular diseases. There are nine main forms of mitochondrial myopathy : Mitochondrial encephalopathy, lactic acidosis and stroke-like episodes (MELAS) syndrome, Kearns-Sayre syndrome (KSS), Leigh syndrome,

Mitochondrial DNA (mtDNA) depletion syndrome, Mitochondrial neurogastrointestinal encephalopathy (MNGIE), Myoclonic epilepsy with ragged red fibers (MERFF), Neuropathy, ataxia and retinitis pigmentosa (NARP) syndrome and Pearson syndrome. It is caused by genetic mutations in either nuclear DNA (nDNA) or mitochondrial DNA (mtDNA). Symptoms of mitochondrial myopathies vary widely by type and from case to case. These disorders can lead to muscle weakness, exercise intolerance, fatigue, developmental regression, ophthalmoplegia, neurological problems like seizures et al. There is no specific cure, but treatments focus on managing symptoms and may involve physiotherapy, medications and lifestyle adjustments.

2. DEVELOPMENT OF WEB PAGES

Crafting web pages for an optimal user experience holds immense significance, as it captivates users immediately. Even in seemingly static pages primarily displaying data, the choice of frontend technology significantly impacts user engagement. Opting for a robust frontend technology like Streamlit offers numerous advantages, including user-friendliness and enhanced speed. Our utilization of Streamlit underscores our commitment to providing an interface that's intuitive, swift, and user-oriented.

3. LITERATURE SURVEY

3.1 Advance Genome Disorder Prediction Model Empowered With Deep Learning

The Advance Genome Disorder Prediction Model (AGDPM) represents a significant leap in biomedical research, empowering the accurate prediction of various genetic disorders using a wealth of patient data. This model, fortified with deep learning capabilities and drawing from the vast landscape of genetic information, stands out for its ability to predict single-gene inheritance disorders, mitochondrial gene inheritance disorders, and multifactorial gene inheritance disorders. With an emphasis on multi-class prediction, AGDPM harnesses convolutional neural network architecture, specifically AlexNet, to process a substantial volume of patient data. Validated through rigorous testing, AGDPM showcases remarkable accuracy rates of 89.89% in training and 81.25% in testing, marking a substantial improvement over previous methods. By enabling efficient prediction and processing of genome disorder data, AGDPM promises to significantly enhance biomedical research efforts aimed at predicting genetic disorders and curbing high mortality rates associated with such conditions.

3.2 Multiple Genetic Syndromes Recognition Based on a Deep Learning Framework and Cross-Loss Training

This study pioneers BioFace, a deep learning model tailored for genetic disease recognition via facial features analysis despite limited datasets. Utilizing Resnet as its foundation and integrating Squeeze-and-Excitation blocks for feature weighting, coupled with a cross-loss training strategy rooted in transfer learning, BioFace achieved a groundbreaking 93.5% accuracy in recognizing 10 syndromes. This framework's potential spans wider clinical applications, promising assistance in genetic diagnosis with small datasets and bolstering genetic research and clinical practice.

3.3 Computerized Prediction of Hereditary Diseases Through DNA Sequence Using Support Vector Machine (SVM)

This paper introduces a Support Vector Machine (SVM) based system for predicting hereditary diseases by analyzing DNA sequences. With lifestyle changes impacting disease susceptibility, common illnesses

like cancer, cardiovascular diseases, diabetes, and hypertension are increasingly influenced by genetic factors. Hereditary diseases, passed from parents to children due to genetic mutations, can now be anticipated earlier by comparing the DNA sequences of parents and children. This predictive model aims to forecast diseases in their early stages, empowering proactive measures and precautions based on genetic information.

3.4 A Survey on Analysis of Genetic Diseases using Machine Learning Techniques

In 2021, B. Dhanalaxmi, K. Anirudh, G. Nikhitha and R. Jyothi published a research paper on a project which was a Genetic Disease Analyzer (GDA). This study explores the impact of technological advancements on genetic disease treatment through the utilization of machine learning algorithms. In the era following genomics, identifying genes responsible for complex diseases presents a significant challenge due to their heterogeneity. Biological markers are pivotal but challenging to pinpoint. Machine learning algorithms play a crucial role in defining these markers, their success reliant on data quality and scope. The study introduces a supervised machine learning approach to predict disease-causing genes, experimenting with algorithms like PCA, Random Forest, Naive Bayes, and Decision Trees. The resulting Genetic Disease Analyzer (GDA) boasts impressive accuracy and sensitivity rates of 98.79% and 98.67%, respectively, using the GEO dataset. The research delves into the practical applications of these machine learning approaches in studying genetic and genomic data, indicating promising avenues for genetic disease analysis.

NAME OF THE PAPER	AUTHORS	YEAR	FEATURES
Advance Genome Disorder Prediction Model Empowered With Deep Learning	Amir Mosavi, Atta-Ur-Rahman, Muhammad Umar Nasir, Mohammed Gollapalli, Muhammad Zubair, Muhammad Aamer Saleem, Shahid Mehmood, Muhammad Adnan Khan	2022	Utilizes machine learning for multi-class genome disorder prediction, introducing AGDPM, handling complex gene interactions, and offering potential for biomedical advancements.
Multiple Genetic Syndromes Recognition Based on a Deep Learning Framework and Cross-Loss Training	Jianfeng Wang, Bo Liang, Lijun Zhao, Yuanfang Chen, Wen Fu, Peiji Yu, Hongbing Chen, Hongying Wang, Guojie Xie, Ting Wu, Muhammad Alam, Haitao Lv, and Lin He.	2022	BioFace, a robust deep learning model, utilizes facial feature analysis, Resnet architecture, SE blocks, and transfer learning for high-accuracy recognition of multiple genetic syndromes.

Computerized Prediction of Hereditary Diseases Through DNA Sequence Using Support Vector Machine (SVM)	Pyla Jyothi; P. Ajitha	2021	DNA-based disease prediction, genetic susceptibility, lifestyle impact, hereditary diseases, genetic mutations, early detection, Support Vector Machine (SVM), proactive disease management.
A Survey on Analysis of Genetic Diseases Using Machine Learning Techniques	B. Dhanalaxmi; K. Anirudh; G. Nikhitha; R. Jyothi	2021	Study for a Genetic Disease Analyzer (GDA)

4. PROPOSED SYSTEM

There are many technologies available for building the frontend of the website. We can make the frontend using HTML, CSS, JS and PHP MySql but we have used Streamlit to create interactive and customizable web apps with minimal code.

4.1 Frontend development

Streamlit is a powerful Python library used to create web applications for data science and machine learning projects. It simplifies the process of building dynamic and customizable web interfaces directly from Python scripts. With Streamlit, we can quickly create frontend interfaces by writing simple and intuitive Python code. It offers a range of widgets for user input, data display, and visualization, enabling easy integration with machine learning models and data analysis pipelines. This framework's simplicity and efficiency make it an excellent choice for rapid prototyping, data exploration, and showcasing machine learning models to a broader audience through a user-friendly web interface.

4.2 Dataset

The dataset comprises various columns detailing genetic information, family history, medical records, and parental details of patients. These columns include genes inherited from the mother and father, respiratory and heart rates, parental consent for treatment, follow-up risk level, folic acid supplementation, maternal illness history, radiation exposure, substance abuse in parents, assisted conception method, and presence of birth defects. Notably, it excludes test and symptom columns due to insufficient information. The key target variables for analysis are the Genetic Disorder and its Disease Subclass.

4.3 Code Editor

Visual Studio Code, developed by Microsoft, stands as a robust integrated development environment (IDE) catering to Windows, Linux, and macOS users. Its versatile toolset encompasses debugging support, syntax highlighting, intelligent code completion, efficient snippets, code refactoring capabilities, and embedded Git functionalities. Leveraging its user-friendly interface and rich suite of features, we employed VS Code as the code editor for this website, benefiting from its ease of use, extensive features, and convenient shortcuts. Furthermore, a myriad of extensions significantly contributed to enhancing our workflow and productivity.

5. METHODOLOGY

***Algorithm*:** K-Nearest Neighbors (KNN) stands as a foundational machine learning algorithm employed extensively for both classification and regression tasks. Its operation involves predicting the class or numerical value of a data point by identifying the K closest data points from a labeled dataset, relying on a chosen distance metric. In classification, it attributes the majority class among these neighbors to the data point, while in regression, it calculates the average value of their target values. The algorithm's core principle revolves around the idea that similar data points share common class membership or have analogous numeric values, making it an intuitive and adaptable approach across a spectrum of applications. Key factors, such as the choice of K and the distance metric, significantly impact its predictive performance.

A. *Steps performed on the dataset*

1. ***Data Collection:*** We gathered a comprehensive dataset containing genetic information paired with disorder labels, forming the backbone of our predictive model.
2. ***Data Preprocessing:*** The dataset underwent thorough cleansing to rectify inconsistencies, normalization to ensure uniformity, and encoding of categorical variables for numerical analysis.
3. ***Feature Selection:*** Pertinent genetic features were identified and selected from the dataset to boost the model's predictive power, enhancing its accuracy and efficiency.
4. ***Data Split:*** We partitioned the dataset into distinct subsets: a larger portion for training the model and a smaller one for testing its predictive capability.
5. ***K Value Selection:*** Various values of K (the number of nearest neighbors) were experimented with to determine the optimal value maximizing the model's accuracy.
6. ***Model Training:*** The training dataset was used to train the KNN model, allowing it to learn underlying data patterns and relationships.
7. ***Prediction:*** We used the trained model to predict genetic disorders within the testing dataset, evaluating its ability to generalize to new, unseen data.
8. ***Evaluation:*** The model's performance was assessed using relevant evaluation metrics like accuracy, precision, recall, or F1-score to gauge its effectiveness.
9. ***Optimization:*** We fine-tuned the model by exploring different hyperparameters, tweaking feature selection strategies, and employing advanced techniques to enhance predictive capabilities further.
10. ***Deployment:*** Consideration was given to practically deploying the optimized model for genetic disorder prediction, integrating it into an application or platform for providing predictions based on new genetic data. The model was then deployed using Streamlit Community.

Once trained, the model is utilized to predict genetic disorders in the testing set. Evaluation metrics are employed to assess the model's performance, facilitating iterative model refinement by exploring diverse hyperparameters. Ultimately, there's consideration for deploying the optimized model for practical genetic disorder prediction.

6. DISCUSSION

The development of this website revolves around the integration of the specified technologies. This project aims to cater to provide caution and helps in early diagnosis of a disorder which will enable the patient to undergo necessary precautions and treatment.

REFERENCES

1. "Leber hereditary optic neuropathy." MedlinePlus Genetics. [Online]. Available: https://medlineplus.gov/genetics/condition/leber-hereditary-optic-neuropathy/.
2. A. Mosavi, A.-U.-R. Rahman, M. U. Nasir, M. Gollapalli, M. Zubair, M. A. Saleem, S. Mehmood, and M. A. Khan, "Advance Genome Disorder Prediction Model Empowered With Deep Learning," 2022.
3. J. Wang, B. Liang, L. Zhao, Y. Chen, W. Fu, P. Yu, H. Chen, H. Wang, G. Xie, T. Wu, M. Alam, H. Lv, and L. He,"Multiple Genetic Syndromes Recognition Based on a Deep Learning Framework and Cross-Loss Training," 2022.
4. P. Jyothi and P. Ajitha, "Computerized Prediction of Hereditary Diseases Through DNA Sequence Using Support Vector Machine (SVM)," 2021.
5. B. Dhanalaxmi, K. Anirudh, G. Nikhitha, and R. Jyothi, "A Survey on Analysis of Genetic Diseases Using Machine Learning Techniques," 2021.
6. Muscular Dystrophy Association (MDA), Mitochondrial Myopathies. [Online]. Available: https://www.mda.org/disease/mitochondrial-myopathies
7. American Academy of Ophthalmology (AAO), Leber Hereditary Optic Neuropathy. [Online]. Available: https://www.aao.org/eyenet/article/leber-hereditary-optic-neuropathy-6
8. LWW Journals. (2017, May). Keep Them Breathing: Cystic Fibrosis. *Journal of the American Academy of Physician Assistants*. [Online]. Available: https://journals.lww.com/jaapa/fulltext/2017/05000/keep_them_breathing__cystic_fibrosis.4.aspx