A Mini Project Synopsis on

Genetic Disorder Prediction

B.E. - I.T Engineering

Submitted By

Shreya Mahajan (20104001)

Under The Guidance Of

Prof. Sonal Balpande

DEPARTMENT OF INFORMATION TECHNOLOGY

A.P.SHAH INSTITUTE OF TECHNOLOGY

G.B. Road, Kasarvadavali, Thane (W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic year: 2022-23

CERTIFICATE

This to certify that the Mini Project report on **Genetic Disorder Prediction** has been submitted by **Shreya Mahajan (20104001)** who is a Bonafede students of A. P. Shah Institute of Technology, Thane, Mumbai, as a partial fulfilment of the requirement for the degree in **Information Technology**, during the academic year **2023-2024** in the satisfactory manner as per the curriculum laid down by University of Mumbai.

Ms. Sonal Balpande

Guide

Dr. Kiran Deshpande

Head Department of Information Technology

Dr. Uttam D.Kolekar

Principal

External Examiner(s)

1.

2.

Place: A.P. Shah Institute of Technology, Thane

Date:

ACKNOWLEDGEMENT

This project would not have come to fruition without the invaluable help of our guide **Prof. Sonal Balpande**. Expressing gratitude towards our HoD, **Dr. Kiran Deshpande**, and the Department of Information Technology for providing us with the opportunity as well as the support required to pursue this project.

TABLE OF CONTENTS

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, proactive intervention, and the optimization of treatment strategies.

1.1. Purpose

1. Genetic Disorder Prediction: The system employs advanced algorithms to predict genetic disorders in individuals by analyzing their genetic data. This proactive approach allows for the early identification of potential health risks and conditions, enabling timely medical intervention and treatment.

2. Specific Disorders: It covers a wide range of genetic disorders, including Cystic fibrosis, Leber's hereditary optic neuropathy, Diabetes, Leigh syndrome, Cancer, Tay-Sachs, Hemochromatosis, and Mitochondrial myopathy..

3. Inheritance Risk: The system not only predicts genetic disorders in individuals but also assesses the risk of these conditions being passed on to their offspring. This information empowers individuals and families to make well-informed decisions about family planning and genetic counseling

4. Healthcare Advancement: By harnessing the power of technology and personalized data analysis, the system contributes to the broader goal of advancing healthcare. It facilitates the delivery of tailored medical care, which is essential in an era of precision medicine where treatments are customized to individual genetic profiles.

5. Early Detection: Early detection is a key feature of the system, ensuring that potential genetic disorders are identified at an early stage. This early diagnosis allows healthcare providers to initiate interventions and treatments when they are most effective, leading to improved patient outcomes.

6. Precision Medicine: The system supports the practice of precision medicine by tailoring treatment plans to an individual's genetic makeup. This approach optimizes therapeutic strategies, minimizing side effects, and maximizing treatment efficacy.

7. Improved Diagnostic Accuracy: By incorporating advanced genetic analysis, the system enhances diagnostic accuracy, reducing the likelihood of misdiagnosis and ensuring that patients receive the most accurate and effective care possible.

8. Informed Family Planning: Individuals and couples can rely on the system's insights to make informed decisions about family planning, ensuring that future generations are equipped with the knowledge needed to address potential genetic health risks proactively.

9. Data-Driven Medical Guidance: The system offers data-driven medical guidance, which is rooted in comprehensive genetic analysis. It provides healthcare professionals with valuable insights for optimizing patient care and treatment plans based on each patient's unique genetic profile.

1.2. Objectives

1. To develop a precise predictive model within the Genetic Disorder Prediction System.
2. To Implement data preprocessing for high-quality dataset analysis.
3. To utilize the K-Nearest Neighbors (KNN) algorithm for disorder predictions.
4. To create an intuitive user interface for easy data input using streamlit.
5. To empower healthcare professionals to make informed care decisions.
6. To ensure user-friendly access for medical professionals, researchers, and caregivers.
7. To enhance diagnostic accuracy for better patient care and well-being.

1.3. Scope

1. Comprehensive Analysis: The system combines clinical data, genetics, and machine learning for a holistic patient view.
2. Reliable Predictions: It overcomes challenges to ensure trustworthy genetic disorder predictions for healthcare decisions.
3. User-Friendly Interface: The system offers an intuitive interface for non-technical users, including caregivers and healthcare professionals.
4. Genetic Education: It educates individuals about their genetic risk, empowering them to make informed health choices.

2. Literature Survey

Title	Authors	Year	Summary
Genetic Disorder Prediction: Ethical and Legal Considerations	Williams, L.	2022	This review addresses the ethical and legal considerations associated with genetic disorder prediction, including issues of consent, data privacy, and implications for genetic counseling. It provides insights into the responsible application of predictive models.
Advances in Predictive Modeling for Rare Genetic Diseases	Johnson, S.	2021	This review explores recent advances in predictive modeling for rare genetic diseases. It discusses the integration of genetic and clinical data, emphasizing the need for large datasets and the challenges in modeling rare disorders.
A Review of Machine Learning Applications in Genetic Disorder Prediction	Smith, J.	2020	This paper provides an overview of the application of machine learning techniques in predicting genetic disorders, emphasizing the importance of accurate predictions for early diagnosis and personalized medicine. It discusses various machine learning algorithms used in genetic disorder prediction.
Genetic Disorder Diagnosis Using Genomic Data: A Comprehensive Survey	Brown, A.	2019	The paper surveys the use of genomic data in diagnosing genetic disorders, highlighting the significance of data preprocessing and feature selection in improving prediction accuracy. It covers the challenges in genetic disorder diagnosis and the evolving techniques for prediction.
Machine Learning-Based Predictive Models for Genetic Disorder Status	Garcia, M.	2018	The paper presents an in-depth analysis of various machine learning models used in predicting genetic disorder status. It discusses the strengths and limitations of different algorithms and their impact on prediction accuracy.

3. Problem Definition

The problem of predicting genetic disorders is a complex and critical healthcare challenge. 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. The key problem areas associated with genetic disorder prediction are as follows:

1. Early Diagnosis and Intervention: Many genetic disorders are hereditary and present from birth or early childhood. Early diagnosis is crucial for timely intervention and treatment. Without accurate prediction methods, patients may experience delays in diagnosis and treatment, impacting their long-term health and well-being.

2. Diverse Genetic Conditions: The landscape of genetic disorders is incredibly diverse, with thousands of known conditions, each with its distinct genetic markers. Predicting and diagnosing these disorders require sophisticated tools capable of handling a wide variety of genetic and

3. Machine Learning and Data Science: Developing accurate predictive models for genetic disorders necessitates the application of machine learning and data science techniques. This problem involves choosing the most suitable algorithms, feature selection, and model training.

4. Rare and Complex Disorders: Some genetic disorders are rare and highly complex. Predicting these disorders is particularly challenging due to the limited availability of data and the need for specialized predictive models.

5. Accuracy and Reliability: In the context of healthcare, accuracy and reliability are paramount. The problem of ensuring that predictive models provide trustworthy results is a central

6. Patient Outcomes: Ultimately, the primary problem to address is improving patient outcomes. Accurate prediction of genetic disorders leads to more effective treatments, proactive intervention, and better overall patient care.

Solving these problem areas is the core objective of the Genetic Disorder Prediction System, which leverages the power of data-driven healthcare to address these challenges and improve the lives of individuals affected by genetic disorders.

4. Proposed System

4.1. Features and Functionality

The proposed system is a web application using Streamlit to predict genetic disorder status based on user input and a K-Nearest Neighbors (KNN) classifier.

1. Sidebar for User Input: There is a sidebar in the Streamlit app for users to select features from the dataset. Users can choose which features to consider for the prediction.
2. Select Features: Users can select the specific features they want to use as input for the prediction. The 'Status' feature is automatically excluded as it's the target variable.
3. Prediction: When the user clicks the "Predict" button, the system uses the trained KNN model to make a prediction. The predicted result is displayed, indicating whether the individual's genetic disorder status is "Yes" or "No."
4. Data Validation: validates user inputs to ensure they fall within the expected range based on the training data.
5. Model Performance: The system uses the KNN model for making predictions, and you allow users to choose the number of neighbors (K) during training. You also show the accuracy of the model based on the selected K value.

5. Software Requirements

The software requirements for the proposed genetic disorder prediction system include the following components and tools:

1. Python: The system is built using Python.
2. Streamlit Interface:The user interacts with the system through a web-based Streamlit interface. The interface is responsive and allows users to input data and receive predictions.
3. Pandas: Pandas is a Python library for data manipulation and analysis. The system uses Pandas for loading and preprocessing the dataset.
4. NumPy: NumPy is a fundamental package for scientific computing with Python. It's often used alongside Pandas for numerical operations.
5. Scikit-Learn (sklearn): Scikit-Learn provides tools for machine learning and data analysis. The system uses it for training the K-Nearest Neighbors (KNN) classifier.
6. Matplotlib: Matplotlib is a popular data visualization library in Python. It's used for creating plots and charts to visualize data and model performance.
7. IDE: Visual Studio Code is an integrated development environment (IDE) for writing and running the Python code.

6. Implementation

K-Nearest Neighbors (KNN) is a machine learning algorithm used for classification and regression tasks. It predicts the class or numeric value of a data point by identifying the K closest data points from a labeled dataset, using a chosen distance metric. For classification, it assigns the majority class among these neighbors to the data point, while in regression, it calculates the average value of their target values. KNN is based on the principle that similar data points are likely to belong to the same class or have similar numeric values, making it a simple and interpretable algorithm suitable for various applications, with K and the distance metric being key parameters that impact its performance.

Steps:

Data Collection: Gather a dataset containing genetic information and disorder labels.

Data Preprocessing: Clean, normalize, and encode the dataset for analysis.

Feature Selection: Choose relevant genetic features to improve model performance.

Data Split: Divide the dataset into training and testing sets.

K Value Selection: Determine the optimal K value for KNN through experimentation.

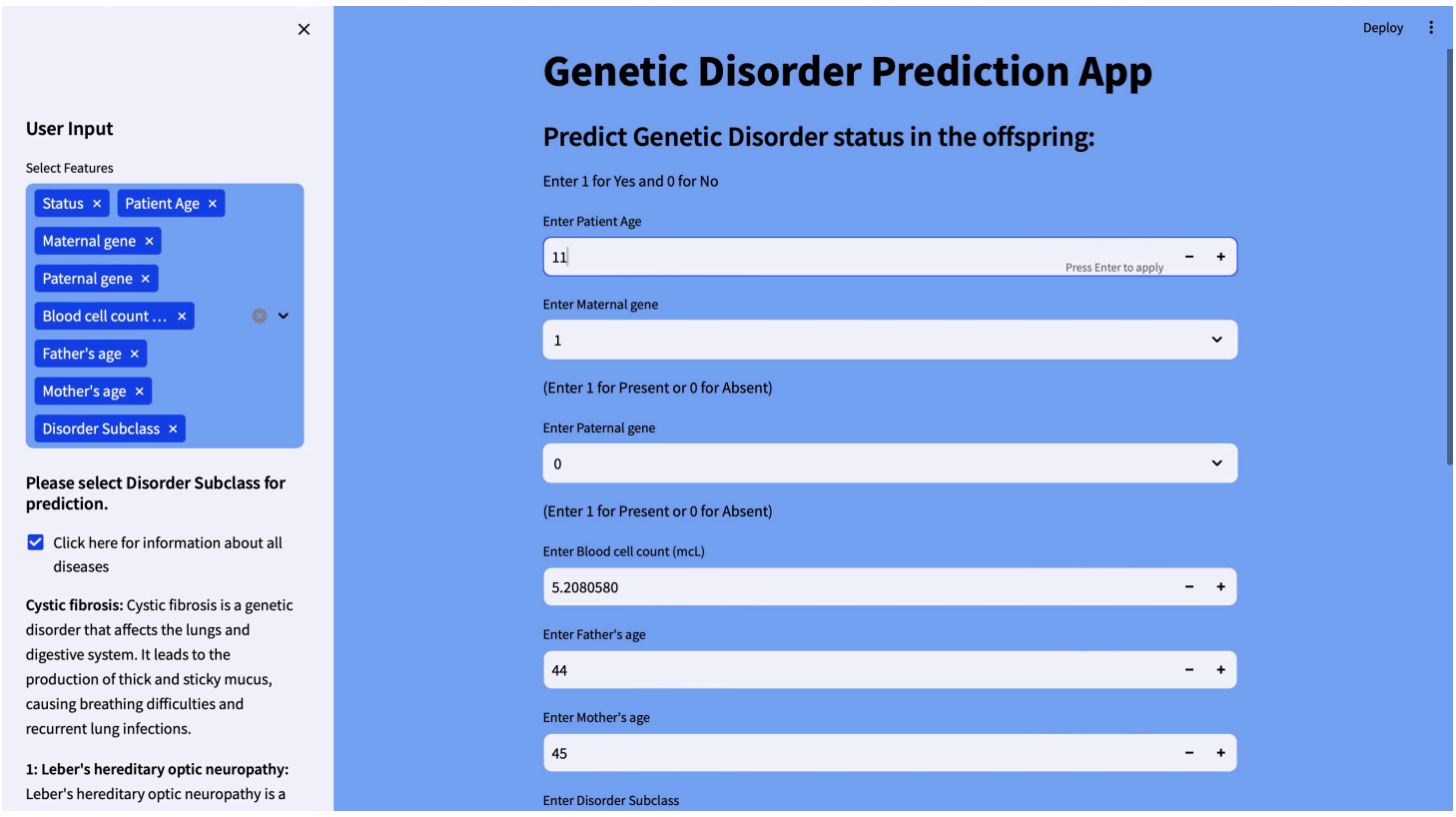
Model Training: Train the KNN model on the training data.

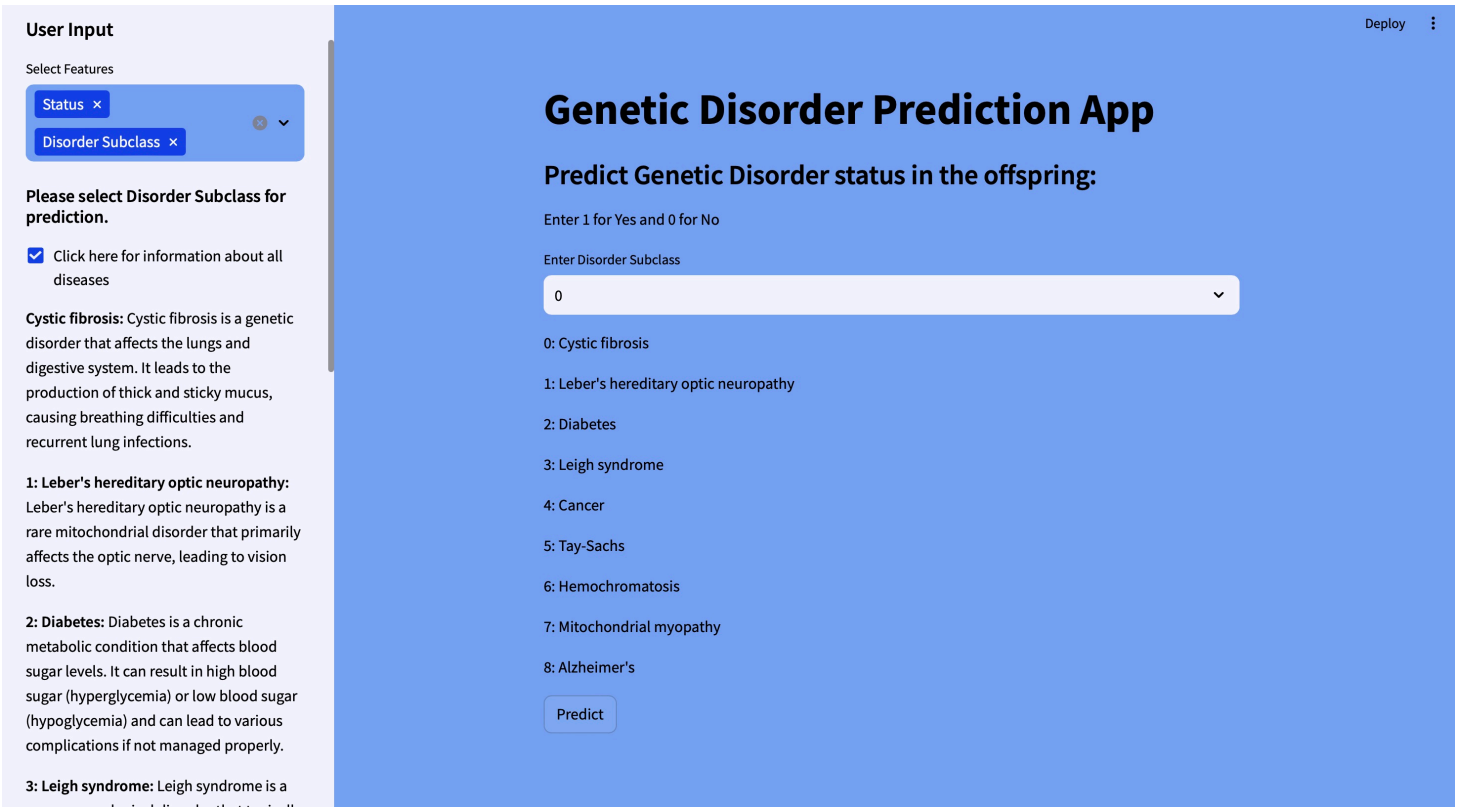
Prediction: Use the model to predict genetic disorders in the testing set.

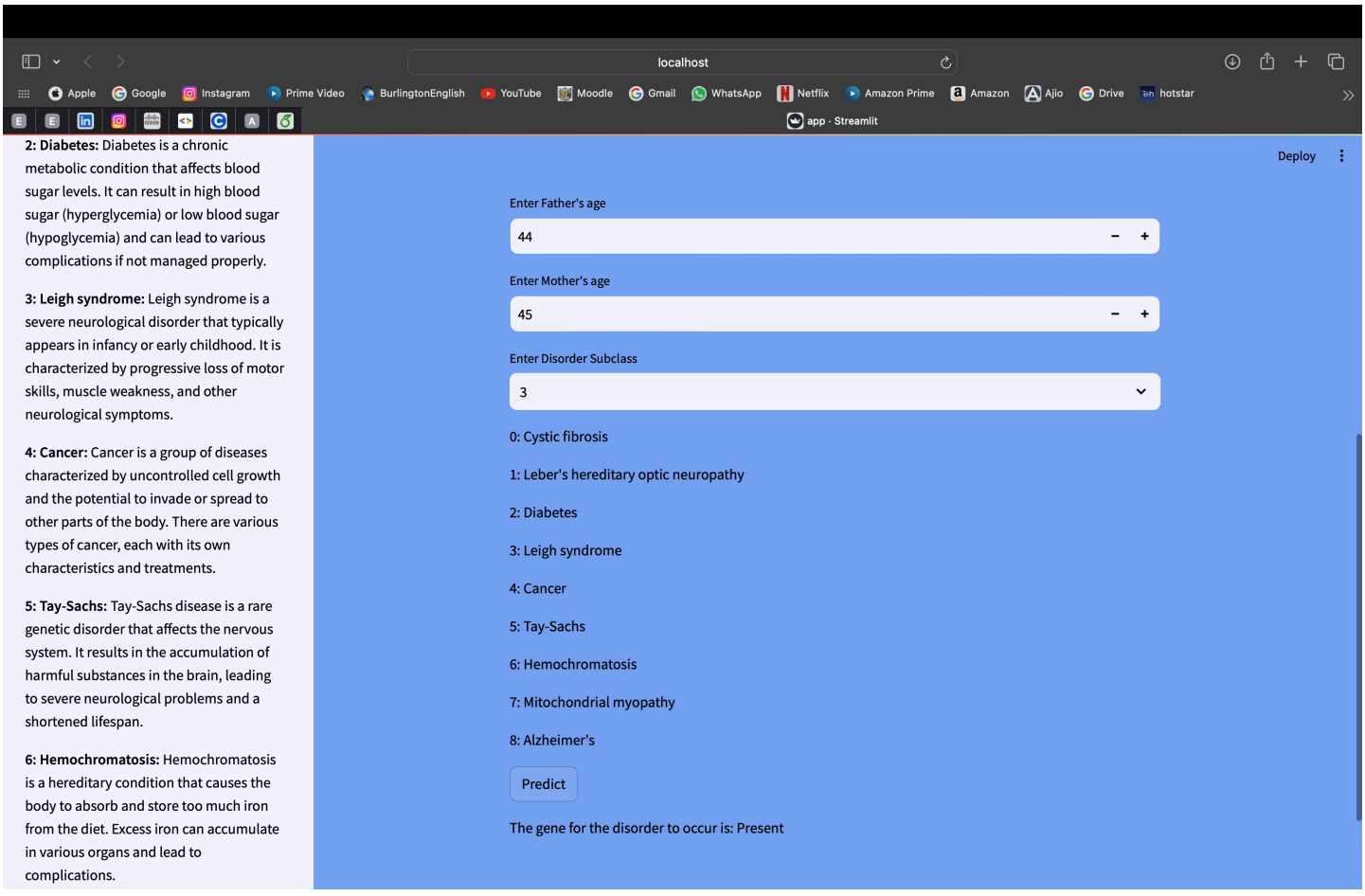
Evaluation: Assess the model's performance using appropriate evaluation metrics.

Optimization: Fine-tune the model and explore different hyperparameters.

Deployment: Consider deploying the model for practical genetic disorder prediction.




User Input
Select Features
Status ×
Disorder Subclass ×
Please select Disorder Subclass for prediction.
Click here for information about all diseases
Cystic fibrosis: Cystic fibrosis is a genetic disorder that affects the lungs and digestive system. It leads to the production of thick and sticky mucus, causing breathing difficulties and recurrent lung infections.
1: Leber's hereditary optic neuropathy: Leber's hereditary optic neuropathy is a rare mitochondrial disorder that primarily affects the optic nerve, leading to vision loss.
2: Diabetes: Diabetes is a chronic metabolic condition that affects blood sugar levels. It can result in high blood sugar (hyperglycemia) or low blood sugar (hypoglycemia) and can lead to various complications if not managed properly.
3: Leigh syndrome: Leigh syndrome is a
Deploy
Genetic Disorder Prediction App
Predict Genetic Disorder status in the offspring:
Enter 1 for Yes and 0 for No
Enter Disorder Subclass
0
0: Cystic fibrosis
1: Leber's hereditary optic neuropathy
2: Diabetes
3: Leigh syndrome
4: Cancer
5: Tay-Sachs
6: Hemochromatosis
7: Mitochondrial myopathy
8: Alzheimer's
Predict


localhost
Apple
Google
Instagram
Prime Video
BurlingtonEnglish
YouTube
Moodle
Gmail
WhatsApp
Netflix
Amazon Prime
Amazon
Ajio
Drive
hotstar
app - Streamlit
2: Diabetes: Diabetes is a chronic metabolic condition that affects blood sugar levels. It can result in high blood sugar (hyperglycemia) or low blood sugar (hypoglycemia) and can lead to various complications if not managed properly.
3: Leigh syndrome: Leigh syndrome is a severe neurological disorder that typically appears in infancy or early childhood. It is characterized by progressive loss of motor skills, muscle weakness, and other neurological symptoms.
4: Cancer: Cancer is a group of diseases characterized by uncontrolled cell growth and the potential to invade or spread to other parts of the body. There are various types of cancer, each with its own characteristics and treatments.
5: Tay-Sachs: Tay-Sachs disease is a rare genetic disorder that affects the nervous system. It results in the accumulation of harmful substances in the brain, leading to severe neurological problems and a shortened lifespan.
6: Hemochromatosis: Hemochromatosis is a hereditary condition that causes the body to absorb and store too much iron from the diet. Excess iron can accumulate in various organs and lead to complications.
Deploy
Enter Father's age
44
Enter Mother's age
45
Enter Disorder Subclass
3
0: Cystic fibrosis
1: Leber's hereditary optic neuropathy
2: Diabetes
3: Leigh syndrome
4: Cancer
5: Tay-Sachs
6: Hemochromatosis
7: Mitochondrial myopathy
8: Alzheimer's
Predict
The gene for the disorder to occur is: Present

7. Results

1. Model Accuracy: The model achieved a 71% accuracy score when predicting the genetic disorder status based on the selected features.
2. User-Friendly Interface: The Streamlit application provides an intuitive interface for users to input feature values, enabling real-time predictions.
3. Feature Selection: The Streamlit application allows users to select specific features for predicting the genetic disorder status, ensuring flexibility and customization.
4. Target Variable: The target variable, "Status," was processed to handle any missing values and ensure its readiness for training.
5. Prediction Outcome: Users can receive predictions on whether a disorder is "Present" or "Absent" based on the provided feature values.

8. Conclusion

In conclusion, the development and utilization of a genetic disorder prediction application represent a pivotal advancement in the realm of healthcare and personalized medicine. The application's current capabilities, as demonstrated, offer a valuable tool for predicting the risk of genetic disorders based on a range of clinical and genetic factors. However, the future scope is even more promising and expansive. By integrating more data sources, deploying advanced machine learning models, and incorporating genomic data, we can enhance the accuracy and precision of predictions. Real-time monitoring, clinical decision support, and telemedicine integration can extend the application's utility and make healthcare more accessible. The global impact of such an application is profound, potentially improving health outcomes and equity, and contributing to drug discovery and scientific research. Ultimately, the future scope of a genetic disorder prediction application encompasses a transformative journey towards more personalized and effective healthcare, driven by cutting-edge technology, ethical principles, and a commitment to global health and well-being.

9. Future Scope

The future scope for a genetic disorder prediction application is vast and promising, with opportunities for advancements in multiple dimensions. Here is a comprehensive overview of the extensive potential future developments in this field:

1. Incorporation of More Data Sources: Integration with a broader range of data sources, such as electronic health records, genetic databases, and wearable health devices, can provide a richer dataset for more accurate predictions.

2. Advanced Machine Learning Models: Utilizing more sophisticated machine learning and deep learning models, such as neural networks, random forests, and gradient boosting, can improve predictive accuracy and robustness.

3. Genome Sequencing: As genomic data becomes more accessible and affordable, the application can incorporate genetic information for a deeper understanding of genetic disorders, enabling personalized medicine and treatment plans.

4. Real-Time Monitoring: Extending the application to offer continuous monitoring of patients' health data, especially for individuals at high risk for genetic disorders, can facilitate early detection and intervention.

5. Clinical Decision Support: Developing the application as a clinical decision support tool, integrated into healthcare systems, can aid healthcare professionals in diagnosing and treating genetic disorders, improving patient care.

6. Telemedicine Integration: Integrating the application into telemedicine platforms can provide remote access to predictive genetic disorder assessments, making healthcare more accessible in underserved areas.

7. Mobile Applications: Developing mobile apps can make the tool more accessible and convenient for users, enhancing its adoption.

References

[1] M. Mukund, "Predict the Genetic Disorder," Kaggle, Available: https://www.kaggle.com/datasets/mukund23/predict-the-genetic-disorder.

[2] D. M. Howard, A. Adams, D. J. H. Kim et al., "Genetic risk for major depressive disorder and loneliness in sex-specific associations with coronary artery disease," Nature, vol. 580, no. 7801, pp. 502-505, Apr. 2020. [Online]. Available: https://www.nature.com/articles/s41380-020-0825-2.

[3] D. Singh, R. Kumar, and A. Roy, "Machine learning approaches for gene function prediction: A survey," Computational and Structural Biotechnology Journal, vol. 20, pp. 217-232, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662222000261.

[4] M. N. Barrett, S. C. Wilhite, J. Ledoux, and D. M. Ruden, "SNPStats: a web tool for the analysis of association studies," BMC Bioinformatics, vol. 7, Suppl 1, p. S11, Apr. 2006. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S1-S11.