MINOR PROJECT-1

SYNOPSIS

ON

Predicting Genetic Disorders in India Using Machine Learning and Geospatial Analysis

Submitted By

Mohit Tiwari
500121521

Vaachi Gupta
500119766

Under the guidance of

Dr. Neeraj Chugh
Associate Professor

Data Science Cluster

School of Computer Science

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

Dehradun

January-May, 2025

Index

Chapter 1: Abstract

This project seeks to create an AI-driven web application that estimates genetic disorder risk in a person and maps disease prevalence in India through geospatial analysis. Through the use of machine learning models and interactive heatmaps, the system will enable users not only to determine their risk levels but also to see the geographical spread of genetic disorders. The initial dataset will be used for model training, with future plans to switch to RTI-purchased official health records for greater accuracy. The main challenges of the project are data availability, model reliability, and regulatory compliance.

Chapter 2: Introduction

The healthcare system of India faces substantial challenges from genetic disorders because the nation lacks adequate genetic testing as well as early diagnosis options. This project establishes a connection between genomic information assessment and public health legislation implementation by combining artificial intelligence methods with location-based research. The final product is a web-based platform that provides:

- A heatmap of India identifying disorder prevalence.
- A data-entry-based forecasting system that estimates one's risk from information given.

The coupling of real-time updates of data with geospatial mapping of illness makes this system an incredibly valuable tool for the healthcare community and the general public.

Chapter 3: Problem Statement

India faces multiple challenges in predicting and managing genetic disorders, including:

1. Limited access to genetic testing services, particularly in rural areas.
2. Lack of well-structured, large-scale data for India's varied population.
3. Lack of availability and infrastructure for early genetic risk assessment.

This project aims to develop a scalable, user-focused AI model that leverages data to facilitate personalized risk analysis and geospatial monitoring of diseases.

Chapter 4: Literature

The nation of India confronts several obstacles when it comes to predicting as well as managing genetic disorders.

- Limited accessibility to genetic testing, particularly in rural areas.
- Lack of standardized large-scale datasets matched to India's diverse populations.
- Inadequate infrastructure and awareness for early genetic risk assessment.

This project will create an AI model which performs scalable individual risk evaluation combined with geospatial disease monitoring through user-friendly data-driven technology.

Chapter 5: Objective

The primary objectives of this project are:

1. Develop a machine learning model to assess genetic disorder risks.
2. Integrate geospatial analysis to map the prevalence of genetic diseases.
3. Ensure model transparency and usability for healthcare professionals and users.
4. Transition from an initial dataset to RTI-based official health records for better accuracy.
5. Deploy an interactive web application with a heatmap and a personalized risk assessment tool.

Chapter 6: Methodology

1. Transitioning from Initial Dataset to RTI-Based Data

The project begins with an available dataset on genetic disorders. In parallel, an RTI request is filed to acquire official health records. Once received, the model will be retrained with the newly acquired dataset, ensuring improved accuracy and relevance.

2. Development of a Machine Learning Model for Disease Prediction

A supervised learning approach will be used to train the model, leveraging algorithms such as Random Forest, CNN, and LSTM. The model will predict genetic disorder risks based on user input, incorporating both genetic and environmental factors.

3. Implementation of Geospatial Analysis for Heatmap Generation

Geospatial clustering techniques will be applied to visualize disease prevalence across India. Data from demographic and health records will be processed to create an interactive heatmap, enabling pattern recognition in high-risk regions.

4. Integration of a Web-Based Interface for User Interaction

A frontend UI will be developed to provide users with two key features:

- Interactive Heatmap: Displays disease prevalence data across India.
- Form-Based Risk Assessment: Allows users to input personal health details and receive a predicted risk assessment based on ML model predictions.

5. Validation and Testing of Model Accuracy

The model's performance will be evaluated using metrics such as accuracy, recall, precision, and F1-score. Further, real-world user testing will be conducted to ensure the interpretability and usability of both the heatmap insights and risk assessment module.

6. Deployment and Future Enhancements

The system will first be tested locally, followed by a potential public deployment if time permits. Future enhancements will include:

- Integration of more datasets for improved model accuracy.
- Enhancing UI/UX for better accessibility.
- Refining model predictions based on new user feedback.

Chapter 7: System Requirements

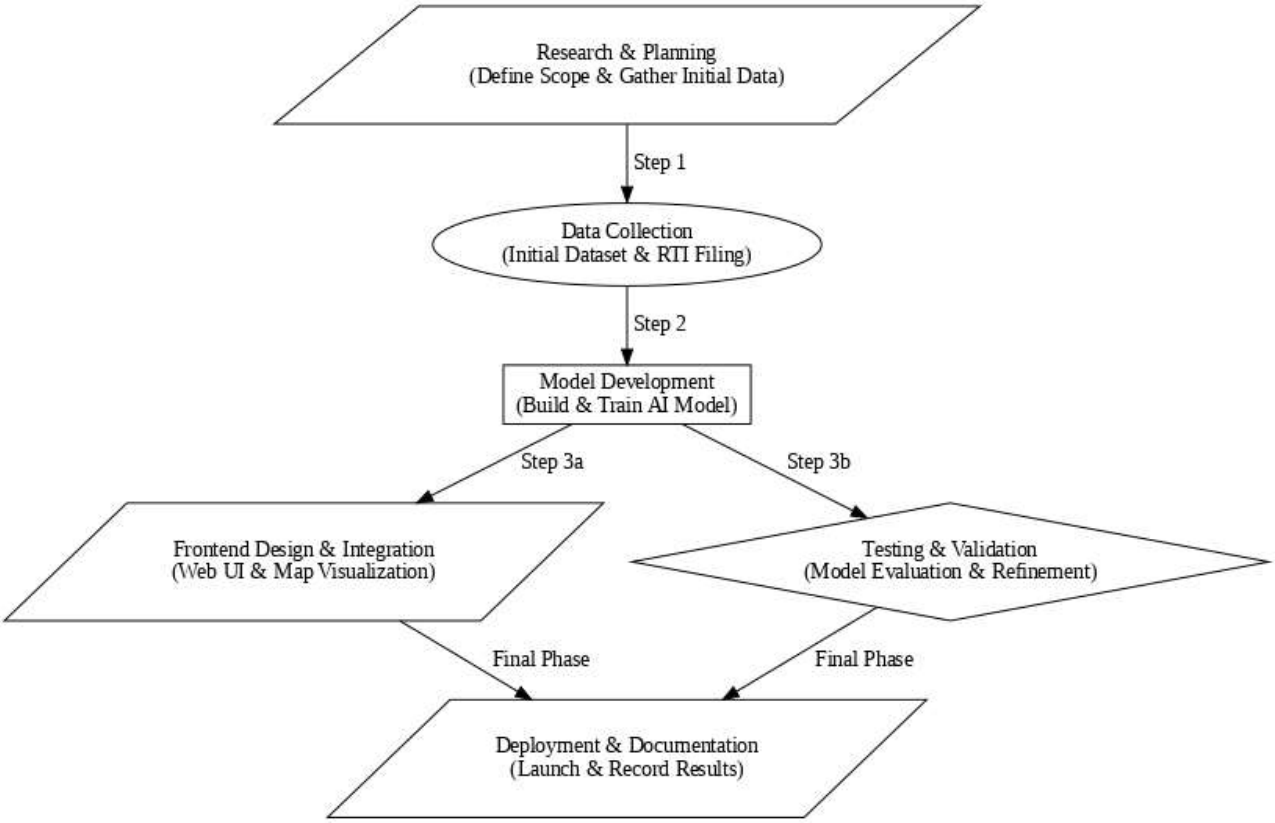
Software Requirements:

- **Programming Languages:** Python, R
- **ML Frameworks:** TensorFlow, Scikit-learn
- **Geospatial Tools:** ArcGIS, QGIS
- **Database:** PostgreSQL with PostGIS extension
- **Cloud Services:** AWS/GCP for scalable computation

Hardware Requirements:

- High-performance GPUs for deep learning training.
- Secure servers for storing sensitive health data.

Chapter 8: Pert Chart



Chapter 9 :SWOT ANALYSIS

STRENGTHS	WEAKNESSES
• **Early Detection of Genetic Disorders:** Machine learning enables rapid and accurate predictions, allowing for timely medical intervention. • **Data-Driven Insights:** The model uses its ability to analyze data-driven insights which traditional methods cannot detect in genomic data. • **Scalability:** The predictive model shows better performance when more genomic data becomes accessible which increases its accuracy level.	• **Data Quality and Availability:** Prediction accuracy depends directly on the quality along with the size of genomic datasets. • **Computational Requirements:** Large genomic data processing needs robust computational resources because of its size. • **Interpretability Issues:** Machine learning models can be complex and difficult for medical professionals to interpret.

OPPURTUNITIES	THREATS
• **Advancements in Genomic Research:** Ongoing research can enhance model accuracy and expand its applications. • **Personalized Medicine:** The model can help create tailored treatment plans based on an individual's genetic profile. • **Collaboration with Healthcare Institutions:** Partnerships between medical establishments enable institutions to gain access to authentic patient information which supports validation efforts.	• **Data Privacy Risks:** Unauthorized access or misuse of genetic data will produce ethical and legal obstacles that could endanger data privacy. • **Regulatory Challenges:** The process of adopting medical and data protection regulations becomes a substantial obstacle for adoption. • **Resistance to AI in Healthcare:** Some medical professionals may be hesitant to rely on AI-based predictions.

Chapter 10: References

[1] B. Dhanalaxmi, K. Anirudh, G. Nikhitha, and R. Jyothi, "A Survey on Analysis of Genetic Diseases Using Machine Learning Techniques."

[2] S. Naik, D. Nevare, A. Panchal, and C. Pawar, "Prediction of Genetic Disorders using Machine Learning."

[3] N. Chaplot, D. Pandey, Y. Kumar, and P. S. Sisodia, "A Comprehensive Analysis of Artificial Intelligence Techniques for the Prediction and Prognosis of Genetic Disorders Using Various Gene Disorders.