SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY

(AUTONOMOUS)

Department Of Computer Science & Engineering (Data Science)

Disease-Gene Prediction: A Machine Learning Perspective

Academic Year : 2024-2025

Year & Sem : IV-II

ABSTRACT

Understanding the intricate relationship between genes and genetic diseases is pivotal for advancing human health. This project explores computational methods as a cost-effective alternative to traditional experimental approaches for identifying disease-associated genes. By integrating advanced topological and biological features, the study enhances gene-disease association predictions using DisGeNET data. Performance metrics include true positive rate, precision, recall, accuracy, F-measure, and ROC curve analysis. Models like XGBoost and Random Forest are expected to excel, with Random Forest achieving a remarkable 97.81% accuracy, especially in major disease classifications (Group, Disease, and Phenotype). These findings aim to advance computational genetics, surpassing current methodologies.

EXISTING SYSTEM METHODS

- k-Nearest Neighbors
- Graph Neural Networks
- LightBGM

DISADVANTAGES OF EXISTING SYSTEM

- High Complexity
- Scalability Issues
- Data Dependency

PROPOSED SYSTEM

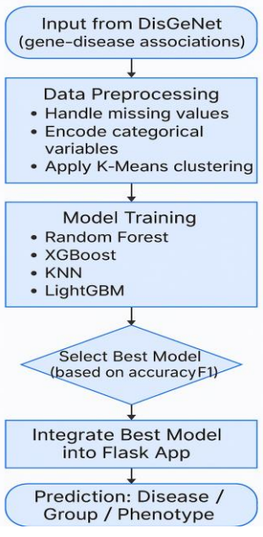
In this project, we propose a novel approach that integrates advanced machine learning models like Random Forest, XGBoost, LightGBM, and KNN with robust preprocessing techniques to tackle the task of gene-disease association prediction.

The system combines machine learning models like Random Forest, XGBoost, LightGBM, and KNN with preprocessing techniques like KMeans clustering and normalization. Built using Flask, it enables real-time predictions with high accuracy and improved gene-disease association insights.

ADVANTAGES OF PROPOSED SYSTEM

1. Cost-Effective and Efficient
2. User-Friendly Interface
3. Scalable

FLOW CHART



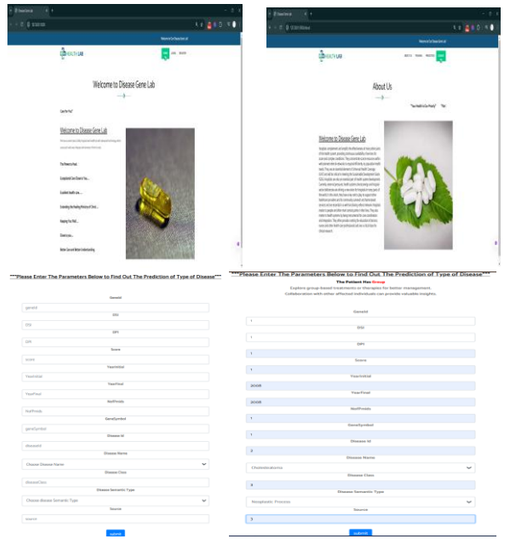
MODULES

1. Data Preprocessing and Clustering
2. Machine Learning Model Training
3. Real-Time Prediction

MODULES DESCRIPTION

- Data preprocessing and clustering ensure clean data preparation, encoding variables, and clustering diseases using KMeans.
- Machine learning models are trained on processed data to predict gene-disease associations with high accuracy.
- Real-time prediction deploys the best model via Flask to classify diseases, groups, and phenotypes efficiently.

OUTPUTS



Quality Classification

CONCLUSION

In summary, our project integrates advanced preprocessing techniques and machine learning models to enhance the accuracy of gene-disease association predictions. By combining robust preprocessing methods with the predictive strengths of Random Forest, XGBoost, LightGBM, and KNN, we address the limitations of traditional approaches. Fine-tuning the models and leveraging clustering techniques ensure adaptable and generalizable outcomes, validated by high accuracy and performance metrics. Deploying this system as a Flask web application provides real-time predictions, revolutionizing computational genetics and supporting early diagnosis. This project highlights the transformative role of machine learning in advancing healthcare and genetic research for improved outcomes

BATCH NO: A – 07

UNDER THE GUIDANCE OF

Mrs. V. Kamakshamma,

M. Tech., (Ph.D)

TEAM MEMBERS

Jayanth Babu G	**214G1A3233**
Likitha B	**214G1A3246**
Navya Teja G	**214G1A3263**
Mounika D	**214G1A3255**