Disease-Gene Prediction: A Machine Learning Perspective

Mrs. Vasepalli Kamakshamma[1], Gajula Jayanth Babu[2], Budavarapu Likitha[3], Golla Navya Teja[4], Dubbala Mounika[5]

[1]Assistant Professor, Dept. of CSE(AI&ML), Srinivasa Ramanujan Institute of Technology, Anantapur, India;
[2,3,4,5]. Students, Dept. of CSE (Data Science), Srinivasa Ramanujan Institute of Technology, Anantapur, India;

Email: kamakshammav.cse@srit.ac.in

Abstract

The paper, "Disease-Gene Prediction: A Machine Learning Perspective" aims at analyzing and predicting the associations between genes and diseases by advanced techniques of machine learning. Due to the fast-increasing availability of genetic data, there has been an increasing need to understand the correlation of specific genes with a disease in biomedical research. This study employs a comprehensive data set, which includes gene-specific information such as DSI and DPI, as well as several disease features including semantic type and classification. It applies four ml algorithms, namely XGBoost, Random Forest, LightGBM and K-Nearest Neighbors (KNN), to predict the three significant output classes-Disease, Group, and Phenotype. It was found that the best model for this purpose came out to be that of Random Forest with 97. 81% accuracy. The same model was implemented using Flask as a framework to gain real-time predictions. Preprocessing mainly involved filling missing values, label encoding, and even clustering into diseases. There are chances of using KMeans clustering for organizing diseases into broader categories based on their similarities for a stronger prediction. The paper demonstrates the potential of machine learning in advancing genomic research by providing insights into gene-disease associations. It offers a practical tool for researchers to explore genetic links to diseases efficiently.

Keywords: gene-disease association, machine learning, clustering, Flask, Random Forest, disease classification, phenotype prediction, gene prediction, bioinformatics, data preprocessing.

1. Introduction

The past couple of years has been very fruitful as far as the genomics field is concerned and has seen quite a number of breakthroughs that led scientists to a better comprehension in the genetic base of various diseases. Research into the genetics of disease has provided scientists with clear avenues for understanding how particular genes may be responsible for certain diseases. However, Such relationship between a few genes and diseases is exceedingly more complex and remains unresolved. This paper, in this case, is "Analysis for Disease Gene Association Using Machine Learning," which will address this challenge by applying machine learning for the prediction of associations between genes and diseases.
Machine learning provides powerful tools to make sense of large datasets, helping us uncover patterns that might not be obvious with traditional methods. By using these models, This paper focuses on analyzing gene-disease associations to predict three key output categories: Disease, Group, and Phenotype. The use of advanced algorithms, including Random Forest, XGBoost, LightGBM, and K-Nearest Neighbors (KNN), provides a multi-model approach to improve prediction accuracy. Among these models, Random Forest achieved the highest accuracy, making it the primary model for final predictions.

For this paper, I worked with a dataset that includes detailed information on gene-disease associations, covering aspects like the Disease Specificity Index (DSI), Disease Pleiotropy Index (DPI), disease names, and the strength of supporting evidence. Techniques of imputation were used to deal with missing values, and categorical variables were preprocessed by assigning label encoding to the data so that it can be used in concurrence with the algorithm implemented for the process of machine learning. The genetic associations of diseases were further categorized into broad groups using KMeans clustering to create useful concepts for analysis.
This paper was developed using the Flask framework, enabling the integration of the machine learning model into a web-based application for real-time predictions. This makes the system accessible to researchers, allowing them to input specific gene or disease data and receive predictions on the likelihood of their association. Overall, This paper demonstrates the potential of machine learning to enhance the understanding of gene-disease relationships. By automating the prediction process and offering real-time insights, this system can contribute to advancements in medical research, personalized treatment, and early disease diagnosis.

In modern medicine, understanding the relationship between specific genes and diseases is essential but challenging due to the complexity of genetic data. Existing manual methods for gene-disease association analysis are time-consuming and prone to errors. This paper addresses the problem by automating the process using machine learning techniques to predict associations between genes and diseases efficiently. It further classifies diseases into groups and phenotypes, offering a streamlined approach for biomedical researchers to explore genetic links. The main goal of this paper is to develop a machine learning-based framework that can predict gene-disease associations using a large dataset. System will categorize diseases into three different output classes, namely Disease, Group, and Phenotype using a set of machine learning models like Random Forest, XGBoost, LightGBM, and K-Nearest Neighbors. The system, intended to bring preprocessing of

data by dealing with missing values and clustering the diseases into more general categories to improve the accuracy rate in predictions. System will use the Flask web application to be deployed in taking genetic or disease-related data input from users and providing back real-time predictions. This therefore aims at improving the ability to analyze gene-disease association, and with this, making it easy for researchers to access the tool comfortably in pursuit of discovering the diseases' genetic basis more easily.

2. Related work

Prediction and Validation of Gene-Disease Associations by Methods Developed from Social Network Analyses (2013): Singh-Blom et al. developed an innovative methodology relying on social network analysis techniques, providing a significant accuracy in prediction. [1]

A Deep Learning Framework Using Graph Augmentation and Functional Modules to Predict Disease-Gene Associations (2024): ModulePred is a deep learning system that uses graph augmentation on protein-protein interaction networks to predict disease-gene connections and demonstrate enhanced predictive performance, as proposed by Jia et al. [2]

Xie et al. (2020) explored network-based techniques for predicting disease-related genes, assessing different computational methods and their effectiveness in identifying gene-disease associations. [3]

Alashwal et al. (2019) utilized supervised machine learning techniques to predict disease-gene associations, highlighting their effectiveness in biomedical research. [4]

A 2023 study presented an interpretable deep learning model for predicting disease-gene associations, offering valuable insights into the biological mechanisms involved. [5]

Chang et al. (2024) introduced a framework that leverages large language models to automate the identification of gene-disease associations, improving the efficiency of literature-based discovery. [6]

Unveiling New Disease, Pathway, and Gene Associations via Multi-Scale Neural Networks (2019): Gaudelet et al. utilized multi-scale neural networks to uncover new associations between diseases, pathways, and genes, providing a comprehensive view of disease mechanisms. [7]

Singh and Lio' (2019) investigated probabilistic generative models combined with graph neural networks for disease-gene prediction, showcasing their ability to capture complex biological relationships. [8]

A 2018 study introduced a text-mining approach to predict potential gene-disease associations by analyzing biomedical literature, assessing its effectiveness in discovering new connections. [9]

A 2022 study introduced GediNET, a sophisticated network-based approach designed to uncover gene associations across various diseases. By integrating machine learning techniques, the model effectively analyzes complex biological networks, providing deeper insights into gene-disease relationships and their underlying mechanisms. [10]

Recent studies have harnessed the power of machine learning to predict gene-disease associations with high accuracy. Techniques like Random Forest, Support Vector Machines, and Gradient Boosting are applied to capture such complex relationships between genes and diseases. For example, PLOS ONE published a study that introduces methods based on social network analysis utilizing Katz and Positive-Unlabeled learning to predict gene-phenotype interactions using the network walk. [11]

Another novel approach is the use of cross-species phenotype networks. By integrating human and model organism data, researchers can form bipartite graphs between phenotypes and human genes. This cross-species analysis has yielded accurate predictions of gene-disease associations by leveraging evolutionary-conserved gene functions and phenotype similarities across species. [12]

Network-based methods provide a robust framework for disease-gene association predictions. In these methods, diseases and genes are treated as nodes in a bipartite graph. Using random walks, network propagation, and kernel-based approaches, researchers have been able to efficiently predict potential disease-gene links. For example, one study shows that incorporating multiple kernel learning (MKL) outperforms single-kernel methods when identifying associations in a gene-disease bipartite network. [13]

3. Dataset

A. Description

The dataset is a collection of several features through which the analysis of genetic contribution to disease susceptibility, progression, and treatment potential is possible. These key features present in the dataset are as follows:

S. No.	Column Name	Description
1.	geneId	A unique identifier for every gene.
2.	geneSymbol	A symbolic name for every gene.
3.	DSI-Disease Specificity Index	Measures how specific a gene is to a disease.
4.	DPI-Disease Pleiotropy Index	Describes the breadth of gene-disease connections.

5.	diseaseId	Identification number for each disease.
6.	Disease Name	Common name of the disease.
7.	Disease Type	Categorizes disease (e.g., genetic, infectious).
8.	Disease Class	Groups diseases into related categories.
9.	diseaseSemanticType	Defines disease significance in medical context.
10.	Score	Measures the intensity of gene-disease association.
11.	EI (Evidence Index)	Measures confidence in gene-disease association.
12.	YearInitial/YearFinal	First and last reported years of association.
13.	NofPmids	Number of PubMed publications available.
14.	NofSnps	Number of Single Nucleotide Polymorphisms linked.

Table 1: Features in DisGeNet dataset

This dataset (as shown in Table.1) establishes a robust basis for studying gene-to-disease correlations and is a really useful resource available for researchers intending to identify unknown therapeutic targets for different diseases as well as provide enhanced disease predictors.

4. Architecture Details

A. Random Forest Classifer

Random Forest is a machine learning method that builds several decision trees on randomly chosen subsets of data and aggregates their predictions to enhance accuracy as well as reliability. The method is effective in reducing the risk of overfitting, is suitable for categorical and numerical variables, and deals with missing data effectively, making it particularly valuable in high-dimensional data such as that of genomics. The approach operates by training each tree on randomized subsets of data to bring variability, selecting a random subset of features at each node to reduce correlation, and building decision trees to predict the target variable. In classification, the final prediction is done using majority voting of the trees, whereas in regression, the predictions are calculated as an average. Random Forest also provides feature importance scores that aid in the identification of important variables.

B. Extreme Gradient Boosting (XGBoost) Classifier

Extreme Gradient Boosting, is a speed-optimized and scalable gradient boosting. It grows trees sequentially, and each subsequent tree learns from the mistakes of previous trees using gradient descent to minimize error for accuracy optimization. XGBoost is renowned for its speed and efficiency, particularly on big data, and efficiently manages missing values while improving base gradient boosting with parallelization and regularization to speed up training and avoid overfitting. The algorithm starts with a weak decision tree, computes residual errors, and recursively adds trees in a way to minimize them. It avoids overfitting by pruning trees, eliminating nodes with low predictive power, and uses regularization to limit tree complexity in order to generalize better.

C. LightGBM Classifier

LightGBM (Light Gradient Boosting Machine) is a fast gradient boosting framework that is more efficient in terms of speed and memory, which makes it more appropriate for larger datasets. Similar to XGBoost, LightGBM builds the model sequentially in order to improve predictions of the test set with lower errors from the previous iterations. The way in which LightGBM builds trees in a leaf-wise manner and allows maximum depth, while also implementing a split based on the maximizing gain from features, makes LightGBM an efficient and computationally-effective gradient boosting method. In addition, LightGBM is capable of pruning features that are not as important to the dataset, and minimizing total residual error in a sequential way through adding trees. The algorithm for building trees using LightGBM turns continuous features into binned ranges and creates a histogram-based approach to speeding up training time and decreasing the amount of memory. Overall, LightGBM is faster and memory-efficient for replacing features than traditional boosting methods. In simply other words, LightGBM is equipment in genomics and is more fast than traditional boosting methods.

D. K-Nearest Neighbors(KNN) Classifier

The K-Nearest Neighbors, abbreviated as KNN, is a very straightforward, but generalizable, instance-based learning algorithm which assigns a class for a data point based on the class of a proportion of the k nearest neighbors. KNN is suited for classification tasks, such as gene-disease associations. KNN is transparent, however it can perform poorly with data that is high-dimensional. KNN is non-parametric, which means it makes no assumptions about the distribution of the data, and relies on the idea that similar instances will be close on the feature space. The typical KNN workflow is to compute the distance, often Euclidean distance, between the query data and all of the training examples, choose the k nearest neighbors, and then classify

the query data based on the majority vote or average the value if for regression. Choosing the optimal k value is often critical, because small k are often prone to overfitting and larger k would underfit the data.

5. Proposed methodology

A. Data Preprocessing

Data preprocessing, in any paper with machine learning algorithms, usually plays vital role, especially in massive and complex datasets, such as genomic data. The dataset used in this paper consists of several fields that contain numerical and categorical values and, typically, some are missing. Thus, the next preprocessing steps to make the algorithms at their best performance include:

Handling Missing Values: The DSI, DPI and EI fields, in the dataset, have missing values. To impute missing values, the dataset uses the median value of respective fields. This will ensure a balanced dataset and algorithms are not predisposed to biased results because the data is incomplete.

Encoding Categorical Variables: The majority of the columns are categorical variables; two examples are diseaseClass and diseaseSemanticType. Label encoding is applied in order to code the categorical into an appropriate format that could be utilized by the machine learning algorithms. All categories are converted to numeric values, hence algorithms can process them well.

Feature Scaling and Normalization: For some algorithms, machine learning requires that the features be on the same scale. Min-Max scaling is used as a technique to normalize so that all features contribute equally to model predictions.

B. Clustering for Disease Categorization

To enhance the interpretability and organization of the data, clustering techniques are applied. KMeans Clustering is used to group diseases based on their genetic associations, creating broader disease categories that can simplify the prediction process. KMeans is an unsupervised learning algorithm. This algorithm is utilized to group data into clusters. In This paper, the KMeans algorithm is implemented on fields diseaseName and diseaseSemanticType to group diseases under categories. Every disease is assigned to a cluster based upon its features, and the clusters provide a high-level view on how diseases The algorithm performs at the following steps

6. Results

The performance of different machine learning models—KNN, LightGBM, XGBoost, and Random Forest—was assessed in terms of their classification accuracy and respective misclassification rates. Out of the models, KNN had the highest misclassification rate, particularly in discriminating between Class 1 and Class 2, which were most commonly misclassified as Class 0. The observation indicates that KNN is highly incapable of discriminating between different classes of diseases and respective phenotypic features.

Class Label	Description	Example
0	Defines the disease exactly	Arthritis
1	Defines the disease group	Diabetes Mellitus
2	Represents the phenotype presented	Exanthema

Table 2: Defining the classes for output

The classification system, as mentioned in Table 2, classifies diseases into three different classes: Class 0, which very specifically classifies diseases (e.g., Arthritis), Class 1, which comprises more generalized categories of diseases (e.g., Diabetes Mellitus), and Class 2, which is classified on the basis of phenotypically accessible features (e.g., Exanthema).

LightGBM was better than KNN; however, it was still making heavy misclassifications, especially in its classification of Class 1. The model is struggling to classify diseases under this specific category well. Nevertheless, it far surpassed KNN with respect to accuracy, precision, and recall.

Of all the models, XGBoost performed better than LightGBM and KNN but slightly more misclassifying than Random Forest. Specifically, it struggled to classify between Class 1 and Class 2 with more Class 1 misclassified as Class 0 than Random Forest. Nevertheless, XGBoost had extremely high precision and recall and, therefore, is still a suitable model for classification.

Algorithm	Precision	Recall	Accuracy
KNN	0.57	0.37	0.78
LightBGM	0.94	0.77	0.95
XGBoost	0.98	0.94	0.97
RandomForest	0.99	0.96	0.98

Table 3: Classification Report for all models

Random Forest was the best, with the least classification errors. Although some samples of Class 1 and Class 2 were misclassified as Class 0, the misclassification rate was much less than the other models. It had extremely good predictive capability with nearly zero errors in all the classes. The quantitative performance comparison of the models is displayed in the classification report in Table 3. KNN was the worst-performing with precision, recall, and accuracy measures of 0.57, 0.37, and 0.78, respectively. LightGBM performed significantly better with precision, recall, and accuracy measures of 0.94, 0.77, and 0.95, respectively. XGBoost performed even better with precision, recall, and accuracy measures of 0.98, 0.94, and 0.97, respectively. Random Forest, however, performed the best among the models with the best precision (0.99), recall

(0.96), and accuracy (0.98). The findings indicate the supremacy of ensemble-based models, specifically XGBoost and Random Forest, in classification.

7. Conclusion

The Examination for Infection Quality Affiliation Using AI paper viably addresses the test of recognizing quality disease affiliations utilizing advanced AI calculations like Random forest, XGBoost, LightGBM, and KNN. By integrating key preprocessing techniques, feature engineering, and clustering, This paper demonstrates a robust framework capable of handling large-scale, complex genomic data. The high accuracy achieved, particularly with the Random Forest model, highlights the potential of these methods in providing valuable insights for genetic research and disease prediction. The deployment of the model using Flask ensures practical and accurate results, benefiting both researchers and healthcare professionals. This system can contribute significantly to the ongoing efforts in personalized medicine, genetic research, and disease diagnosis, marking a step toward more efficient and scalable solutions in bioinformatics.

8. References

1. M. Asif, H. F. M. C. M. Martiniano, A. M. Vicente, and F. M. Couto, “Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology,” *PLoS One*, vol. 13, no. 12, p. e0208626, Dec. 2018, doi: 10.1371/JOURNAL.PONE.0208626.

2. X. Jia *et al.*, “A deep learning framework for predicting disease-gene associations with functional modules and graph augmentation,” *BMC Bioinformatics*, vol. 25, no. 1, pp. 1–14, Dec. 2024, doi: 10.1186/S12859-024-05841-3/TABLES/2.

3. S. K. Ata, M. Wu, Y. Fang, L. Ou-Yang, C. K. Kwoh, and X.-L. Li, “Recent Advances in Network-based Methods for Disease Gene Prediction,” *Brief Bioinform*, vol. 22, no. 4, Jul. 2020, doi: 10.1093/bib/bbaa303.

4. U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, “Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses,” *PLoS One*, vol. 8, no. 5, p. e58977, May 2013, doi: 10.1371/JOURNAL.PONE.0058977.

5. Y. Li, Z. Guo, K. Wang, X. Gao, and G. Wang, “End-to-end interpretable disease–gene association prediction,” *Brief Bioinform*, vol. 24, no. 3, pp. 1–9, May 2023, doi: 10.1093/BIB/BBAD118.

6. J. Chang, S. Wang, C. Ling, Z. Qin, and L. Zhao, “Gene-associated Disease Discovery Powered by Large Language Models,” Jan. 2024, Accessed: Jan. 29, 2025. [Online]. Available: https://arxiv.org/abs/2401.09490v1

7. T. Gaudelet, N. Malod-Dognin, J. Sanchez-Valle, V. Pancaldi, A. Valencia, and N. Przulj, “Unveiling new disease, pathway, and gene associations via multi-scale neural networks,” *PLoS One*, vol. 15, no. 4, Jan. 2019, doi: 10.1371/journal.pone.0231059.

8. V. Singh and P. Lio’, “Towards Probabilistic Generative Models Harnessing Graph Neural Networks for Disease-Gene Prediction,” Jul. 2019, Accessed: Jan. 29, 2025. [Online]. Available: https://arxiv.org/abs/1907.05628v1

9. J. Zhou and B. quan Fu, “The research on gene-disease association based on text-mining of PubMed,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–8, Feb. 2018, doi: 10.1186/S12859-018-2048-Y/FIGURES/6.

10. E. Qumsiyeh, L. Showe, and M. Yousef, “GediNET for discovering gene associations across diseases using knowledge based machine learning approach,” *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–17, Nov. 2022, doi: 10.1038/s41598-022-24421-0.

11. H. Yang, Y. Ding, J. Tang, and F. Guo, "Identifying potential association on gene-disease network via dual hypergraph regularized least squares," BMC Genomics, vol. 22, no. 1, pp. 1–16, Dec. 2021, doi: 10.1186/S12864-021-07864-Z/TABLES/9.

12. Xie, L., et al. (2020). "Recent Advances in Network-Based Methods for Disease Gene Prediction." *Journal of Biomedical Informatics*, 106, 103432. doi:10.1016/j.jbi.2020.103432.

13. Zhang, L., et al. (2021). "deepDGA: Biomedical Heterogeneous Network-based Deep Learning Framework for Disease-Gene Association Predictions." IEEE Access, 9, 17392-17401. doi:10.1109/ACCESS.2021.3056842.