MACHINE LEARNING APPROACH FOR DETECTING GENE-DISEASE ASSOCIATION

Project & Thesis-I

CSE 4100

A thesis Report

Submitted in partial fulfillment of the requirements for the Degree of

Bachelor of Science in Computer Science and Engineering

Submitted by

Md. Al-Amin	**190104001**
Md. Shafayat Jamil	**190104022**
Arittra Das	**190104083**
Swarnajit Saha	**190104086**

Supervised by

Prof. Dr. S.M.A. Al-Mamun

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

May 16, 2023

ABSTRACT

Gene-disease association is a vital and critical area of research that has the potential to improve our understanding of the underlying causes of many diseases. It is very important to find the underlying gene of a disease for prevention, diagnosis, and therapy. In our study, we propose a machine learning approach to predict the score between genes and diseases based on association type and the number of PubMed articles. Specifically, we consider five different types of gene-disease associations: AlteredExpression, GeneticVariation, PostTranslationalModification, Therapeutic, and Biomarker. We then use a range of machine learning algorithms to predict the association score. Our models can be used to identify genes that are likely to be involved in a specific disease based on the available literature. In addition, our approach can help identify new associations that have not yet been reported in the literature, providing new insights into the underlying causes of diseases. Our approach is aimed at providing a valuable tool for researchers working in the field of gene-disease association, as it can help identify relevant associations and prioritize further research efforts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The goal of this thesis is to develop a machine-learning approach for predicting gene-disease associations based on association type and the number of related PubMed articles. Specifically, we consider five different types of gene-disease associations: Biomarker, AlteredExpression, GeneticVariation, PostTranslationalModification, and Therapeutic. We use a range of machine learning algorithms to predict the association score between genes and diseases, which represents the strength of the association. We collected a dataset of gene-disease associations from publicly available databases and preprocessed the data to extract relevant features. We then trained our machine-learning models using a range of techniques, including logistic regression and decision tree regression. We evaluate our approach using Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R2) Score. We also discuss the possible reasons for which the appropriate outcome gets restricted.

1.2 Objective

The primary objective of this research is to evaluate methods of machine learning for predicting gene-disease relationships. Identification and association of genes with the disease require time-consuming and expensive experimentations of a great number of potential candidate genes. so, we are trying to emphasize less expensive and time-consuming methods. our core intention is to use the powerful computational capability of machine learning to identify the complex relationship between the gene and disease.

1.3 Document Structure

The present introductory chapter gives a contextualization for the problem at hand and introduces the main objectives and contributions of this dissertation. The remaining five chapters are organized as follows:

- Chapter 2 defines and explains the basic concepts vital for the understanding of the problem itself namely Genomics, Gene disease association, and machine learning.
- Chapter 3 presents an overview of the methodology developed with a description of the main tasks.
- Chapter 4 describes the model and evaluates the corresponding result.
- Chapter 5 presents the discussion and conclusion regarding our study.

Chapter 2

Background Study

2.1 Genomics

Genomics refers to the analysis and interpretation of the structure, function, and evolution of the complete set of genes, as well as their interactions with each other and with the environment. Every cell of the human body contains approximately 3 billion complete copies of DNA base pairs. [8]

On the other side, a gene is the basic physical and functional unit of heredity. A gene refers to a unit of DNA that carries the instructions for making a specific protein or set of proteins. Humans have between 20,000 and 25,000 genes. [9] Proteins make up body structure by controlling the chemical reaction and carrying signals between cells. If a cell's DNA is mutated an unusual protein may be produced, which can interrupt the body's usual processes and lead to many diseases like diabetes, cancer, high blood pressure, etc. Here are some causes behind the genetic variations. [8]

2.1.1 Mutations

A mutation represents the permanent alteration of a DNA sequence. Mutation occurs when there is an error during DNA replication and that error is not corrected by DNA repair enzymes. It is only once the error is copied by DNA replication, and fixed in the DNA that it is considered to be a mutation. mutation can occur due to many external or internal reasons. The mutation may be harmful or harmless to an organism. They can lead to local changes in tissues [10]. A common process of mutation has been depicted in Figure 2.1 below.

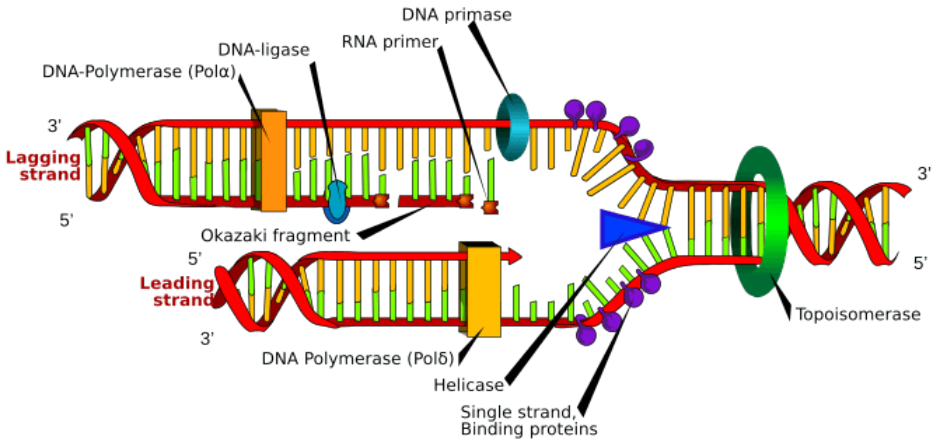

Figure 2.1: Mutation during DNA replication. [1]

2.1.2 Recombination

Recombination, which occurs when homologous DNA strands align and cross over, is a significant contributor to genetic diversity. Through recombination, each individual receives a unique mixture of genetic material from their parents, resulting in new combinations of variants in the daughter germ cells. This process effectively 'shuffles' maternal and paternal DNA, leading to the creation of diverse genetic profiles in each individual [10].

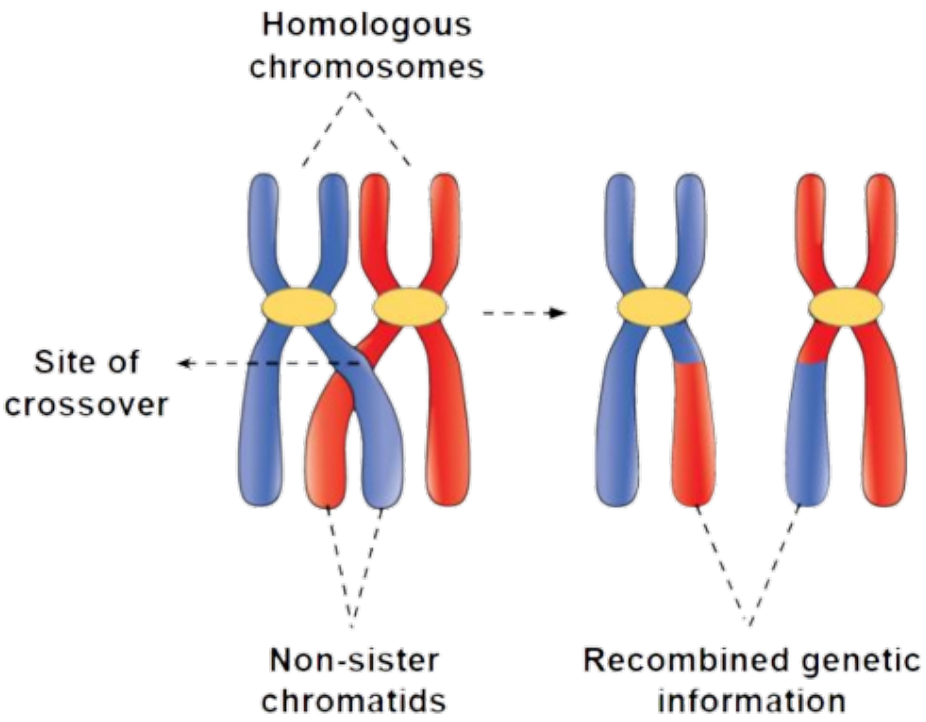

Figure 2.2: Recombination through crossover. [2]

2.1.3 Epigenetic modifications

There are chemical changes to DNA and associated proteins that do not alter the DNA sequence itself but can affect gene expression. Epigenetic modifications can be influenced by environmental factors, such as diet and stress, and can be passed down from one generation to the next.

2.1.4 Natural selection

Natural selection is the process by which organisms with traits that are better adapted to their environment are more likely to survive and reproduce. Natural selection can lead to changes in the frequency of certain alleles in a population over time, resulting in evolution.

2.2 Gene Disease Association

The association between a specific gene or group of genes and a particular disease is known as gene-disease association. In humans, numerous genes exist in the genome. Genetic variations in these genes have been found to play a crucial role in the pathogenesis of many diseases including diabetes, cancer, high blood pressure, etc. Various mechanisms can link genes to disease, with one of the most prevalent being genetic variations in the DNA sequence. These variations can take many forms, such as single-nucleotide polymorphisms (SNPs), insertions, and deletions, and can impact the function of genes or the proteins they produce, thereby altering cellular processes and potentially causing disease. Another way that genes can be associated with disease is through alterations in gene expression. Gene expression refers to the process by which genes are transcribed into messenger RNA and ultimately translated into proteins. Changes in gene expression can result from alterations in DNA sequence, as well as from environmental factors such as diet and exposure to toxins. Finding the underlying relationship between genes and diseases is the key to the development of effective prevention and treatment strategies. Advances in genomics and bioinformatics have made it possible to identify genetic variants associated with a wide range of diseases. Applying machine learning along with these processes, will make it faster and less expensive. By better understanding the underlying genetic factors contributing to disease, we can develop more targeted and personalized approaches to diagnosis, treatment, and prevention.

2.3 Machine Learning

Machine learning refers to the scientific study of computer programs that utilize algorithms and statistical models to learn from data through inference and identification of patterns, without the need for explicit programming by humans. [11] This unique feature enables machines to improve their performance over time, without explicit human intervention or instruction. Machine learning is characterized by a collection of algorithms that operate on large datasets, which are fed into the algorithms for training purposes. Subsequently, the algorithms build a model based on the acquired knowledge and use it to carry out specific tasks.

2.3.1 Types of Machine Learning

Machine learning can be categorized into four main types based on their learning methods and approaches.

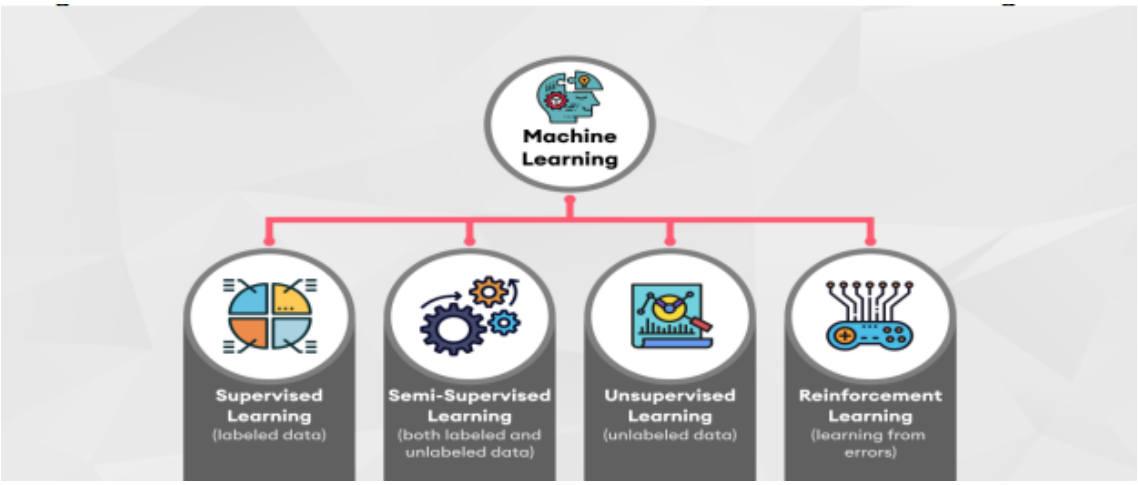

Figure 2.3: Classification of machine learning approaches. [3]

Supervised Learning

Supervised machine learning relies on supervision, which involves training machines using labeled datasets, allowing them to make predictions based on their training. In this technique, labeled data indicates that some inputs have already been mapped to their corresponding output. First, the machine is trained using both the input and its corresponding output, and then it is asked to predict the output based on a test dataset. There are two types of supervised learning based on labels Classification and Regression. Classification algorithm used when the output is “Yes” or “No”, “Male” or “Female” etc. On the other hand, the Regression algorithm is used when the output is a continuous value. [5]

Unsupervised Learning

Unsupervised machine learning involves training machines using unlabeled datasets, without any supervision or pre-defined output. The primary objective of unsupervised learning algorithms is to categorize or group unsorted data based on their similarities, patterns, and differences. Machines are designed to uncover hidden patterns in the input dataset. Unsupervised learning can be classified into two categories, namely clustering and association. The clustering technique is used when the objective is to identify inherent groups within the data. In contrast, association learning algorithms aim to identify the interdependency between different data items and map them accordingly. [5]

Semi-supervised Learning

Semi-supervised learning is a type of machine learning algorithm that falls between supervised and unsupervised machine learning approaches. It leverages a combination of labeled and unlabeled datasets during the training process. Similar data is clustered using an unsupervised learning algorithm, and this helps to label the previously unlabeled data. The primary objective of semi-supervised learning is to utilize all available data effectively, rather than solely relying on labeled data as in supervised learning. The concept of semi-supervised learning was introduced to address the limitations of both supervised and unsupervised learning algorithms. [5]

Reinforcement Learning

Reinforcement learning is a type of machine learning that relies on a feedback-based process. An AI agent, which is a software component, automatically explores its surroundings by trial and error, taking actions, learning from experiences, and improving its performance. The agent receives a reward for each good action and is punished for each bad action. Therefore, the goal of a reinforcement learning agent is to maximize rewards. Unlike supervised learning, there is no labeled data in reinforcement learning, and the agents learn solely from their experiences. Reinforcement learning can be categorized into two types: positive reinforcement learning and negative reinforcement learning. [5]

2.3.2 Machine Learning Models for Gene-Disease Association

Linear Regression

The goal of regression problems is to predict the value of one variable based on the values of other variables. In the case of Linear Regression, we assume that there is a linear relationship between the given input features and the target label, and we are trying to find the exact form of that relationship. If the value of independent variables will change (increase or decrease), the value of the dependent variable will also change accordingly (increase or decrease). There are two types of linear regression: Simple Linear Regression and Multiple Linear Regression. Simple Linear Regression predicts a response using a single feature of the dataset. Multiple Linear Regression predicts a response using two or more features [12]. In the case of our study, we need to use Multiple Linear Regression as we need to calculate the score depending on five association types and PubMed values.

Mathematically we can explain it as follows: Consider a dataset having n observations, p features, i.e. independent variables and y as one response, i.e. dependent variable. The regression line for p features can be calculated as follows –

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + \varepsilon_i$$

where:

- y_i is the dependent variable for observation i,
- b_0 is the intercept,
- $b_1, b_2, \ldots, b_p$ are the coefficients for the independent variables $x_{i1}, x_{i2}, \ldots, x_{ip}$ respectively,
- ε_i is the error term for observation i.

Decision Tree Regression

Decision Tree Regression is a type of supervised machine learning algorithms used for predicting continuous variables. Decision tree regression observes features of a dataset and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. The leaves of the tree represent the predicted output values for the corresponding input regions. Decision tree regression is a popular machine learning algorithm due to its simplicity, interpretability, and ability to handle both continuous and categorical input variables. A decision tree is constructed starting from the root node/parent node

(dataset), which splits into left and right child nodes (subsets of the dataset). These child nodes are further divided into their children node, and themselves become the parent node of those nodes [5].

Random Forest

Random Forest is a popular supervised machine learning algorithm that can be used for both Classification and Regression problems. It is based on the concept of ensemble learning, where multiple decision trees are combined to solve a complex problem and improve the model's performance. The algorithm generates a forest of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy. Unlike a single decision tree, the random forest combines predictions from multiple trees and predicts the final output based on the majority vote. Increasing the number of trees in the forest improves accuracy and prevents overfitting. It takes less training time as compared to other algorithms. It predicts output with high accuracy and even for a large dataset it runs efficiently [5].

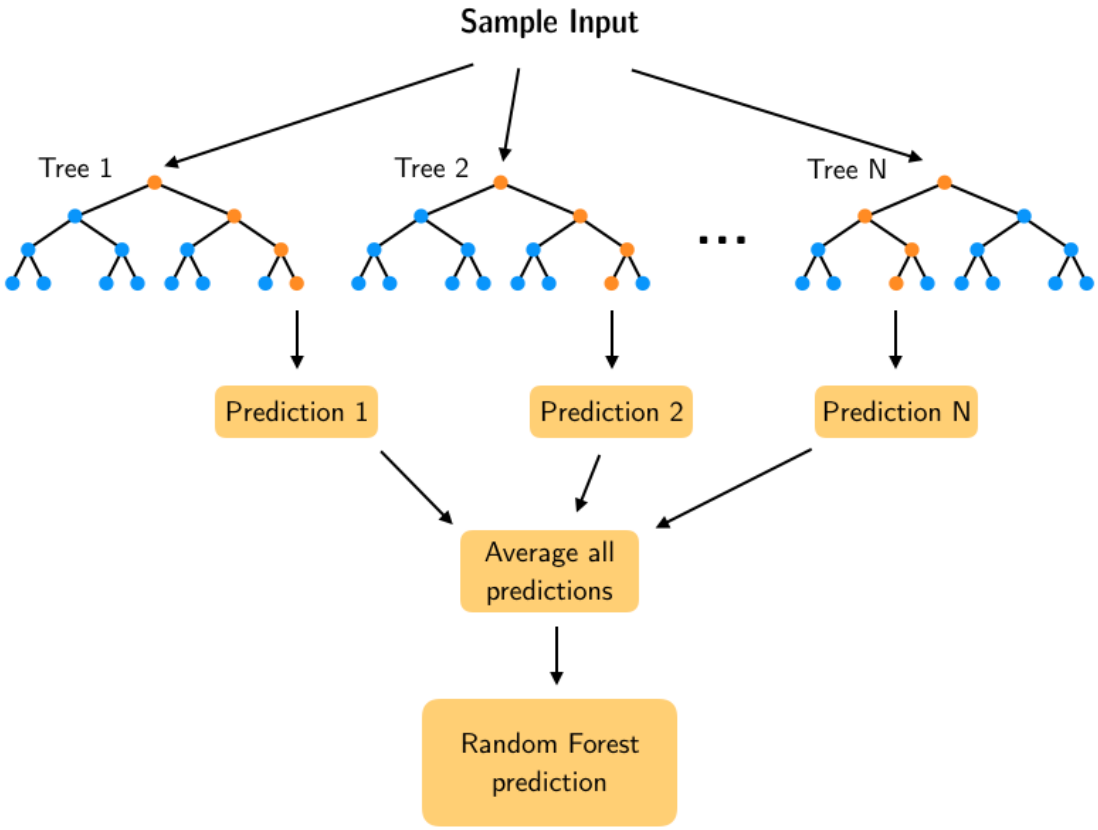

Figure 2.4: Random Forest. [4]

Support Vector Regression

Support Vector Machine (SVM) is a popular machine learning algorithm used for both classification and regression problems. The goal of SVM is to find the best line or decision boundary that separates n-dimensional space into different classes, making it easy to classify new data points in the future. This decision boundary is called a hyperplane. SVM chooses the extreme points called support vectors to help create the hyperplane. The algorithm is named after these support vectors. There are two types of SVM: Linear SVM and Non-linear SVM. In Linear SVM a dataset can be classified into two classes by using a single straight line, where the data are linearly separable. On the other hand, non-linear SVM use Non linearly separable data, where the data can't be classified by using a straight line [5].

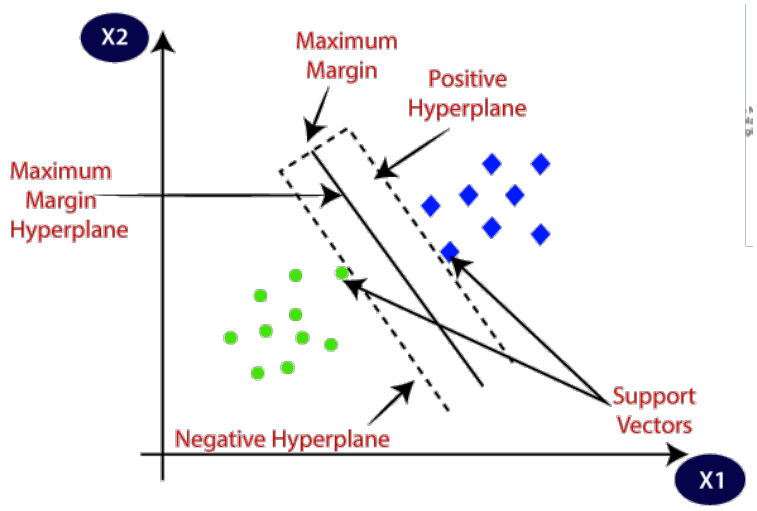

Figure 2.5: Support vector regression. [5]

2.3.3 Gradient Boosting

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

There are some trees in the ensemble. The feature matrix X and the labels y are used to train the first tree. The training set residual errors r1 are calculated using the predictions of that tree. Then, the next tree is trained using r1 and feature matrix X as labels. The residual r2 is then calculated using the anticipated outcome of this tree and r1. This goes on until all trees in the ensemble are trained [6].

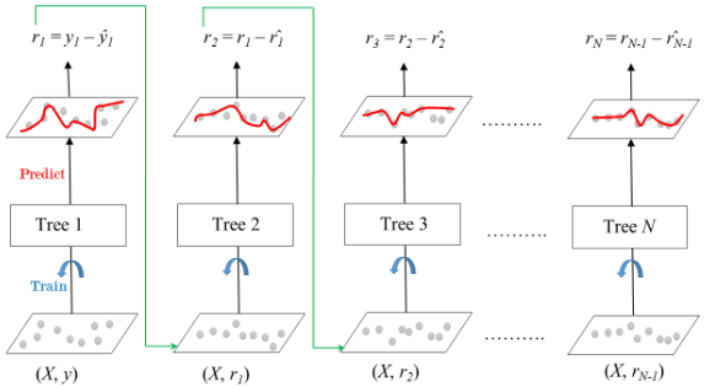

Figure 2.6: Gradient Boosted Trees. [6]

2.4 Study of Related Works

Numerous research papers are available that attempt the application of various machine learning models and their combinations for finding the association of different diseases to genes with improved accuracy.

Sikandar et al., in their paper titled 'Analysis for disease gene association using machine learning' [13], proposed some novel computational methods for identifying gene-disease relations. Their proposed methods were based on advanced biological and topological features. Where the biological features were calculated from gene sequences and the topological features were calculated from protein complexes. They applied their computational strategy to four diseases, which were Thalassemia, Diabetes, Malaria, and Asthma, and found 874 related genes. Several publicly accessible databases are used to download the diseases and their corresponding genes. The UniProt and HPRD databases are used to download gene sequences related to specific disorders. Data mining is done using Weka, which is an online, accessible resource. They achieved the highest accuracy, i.e., up to 90

Le, in his review paper titled 'Machine learning-based approaches for disease gene prediction' [14], gives a wide knowledge of machine learning algorithms for finding the relation between genes and diseases. They describe five traditional binary classification techniques (i.e., DT, k-NN, NB, SVM, and ANN), two unaries (one-class SVM and one-class Hempstalk), three semi-supervised learning methods (i.e., Graph-based SSL, positive and unlabeled learning), and three modern methods (Ensemble Learning, Deep Learning).

In this study, a training set was constructed for different classification methods used for disease gene prediction. Known disease genes were collected from OMIM and mapped to the human PPI network to obtain known disease proteins. The remaining proteins were used to construct an unlabeled set (U) containing unknown disease proteins and non-disease proteins. A negative training set (N) for binary classification algorithms was constructed by

randomly choosing proteins from the unlabeled set.. Neighbors of known disease proteins in the human PPI network were also collected to build the training set for the binary SSL-based method. Finally, both the positive (P) and negative (N) training sets were used for training PU-based methods, including the multi-level weighted SVM-based method used in PUDI.

ANN was the best classifier among the traditional classifiers. TSVM was the best among the SSL-based methods and DNN has the best performance, and NB has the lowest. One-class Hempstalk has the best ACC (accuracy) of about 85

Asif, et al. in their paper titled 'Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology' [15] guide the use of machine learning to predict complex disease and gene association based on gene functional similarity based on gene ontology. In order to predict disease genes, they designed a supervised machine-learning approach using Autism Spectrum Disorder (ASD) as a case study. Using several approaches, they compare the similarities between the genes associated with ASD and those that are not. The classifiers beat previously reported ASD classifiers when they were trained and tested on ASD and non-ASD gene functional similarity. They employed a variety of machine learning classification techniques for this work, including Naive Bayes (NB), linear and radial SVM, and the Random Forest (RF) method, which is based on decision trees. The Simons Foundation Autism Research Initiative (SFARI) gene database (N = 990) was used to collect the genes with evidence of involvement in ASD for the suggested methodology. Based on the strength of the available evidence, genes in the SFARI database are rated in seven different categories. When the number of trees was set to 500, the RF classifier performed at its peak level during performance testing. The maximum accuracy was 80

Luo et al.in their paper titled 'Enhancing the prediction of disease–gene associations with multimodal deep learning' [16] suggest using multimodal DBN (dgMDL) to predict disease-gene relationships. More specifically, two DBNs independently learn hidden representations of protein-protein interaction networks and gene ontology concepts. Then, using the combined output of the two sub-models latent representations as the multimodal input, a joint DBN is utilized to train cross-modality representations from the two sub-models. Finally, using the learned cross-modality representations, disease-gene relationships are predicted. The Online Mendelian Inheritance in Man (OMIM) is where they gather their data. There are about sixty-one hundred disorders, thirty-nine hundred genes, and over seventy-five hundred entries arranged alphabetically by disease names. They compare the output of their model using two recently created algorithms, PBCF and Know-GENE. They discover that the AUC for dgMDL, Know-GENE, and PCFM in the ROC curve is 0.969, 0.941, and 0.791, respectively. dgMDL performs superiorly to rival algorithms.

Hanna et al. in their paper titled “Gene-disease association through topological and biological feature integration” [17] consider a classification-based technology that mixes biological data gathered from multiple data sources with the topology properties of protein-protein interaction networks. In order to identify discrete gene placements, they examine the topology of the relevant PPI network. They then combine biological data from other sources to identify potential similarities that might define each class. The topological features that they extract are degree, eccentricity, closeness centrality, betweenness centrality, authority, hub, modularity class, page rank, component ID, clustering coefficient, number of triangles, and eigenvector centrality. Sequence length, gene ontology (GO), topological domains, chain, domain, protein family, and pathway were the biological characteristics taken into consideration in the study. Out of the 9228 genes in their final learning dataset, 839 of them are linked to illnesses. The technique yields an AUC result of 0.941. Breast cancer and Type II diabetes mellitus are additional diseases to which they apply this classification technique. Out of the 23 genes contained in the database, the proposed model was able to identify 16 of them as being the cause of "Type II Diabetes Mellitus". However, out of the 23 genes contained in the dataset for "Breast Cancer", the model can detect 13 of them.

Chapter 3

Methodology

3.1 Introduction

The score of the gene-disease connection can be determined using a wide range of machine-learning algorithms. In this part, we offer our approach for predicting the gene-disease association score.

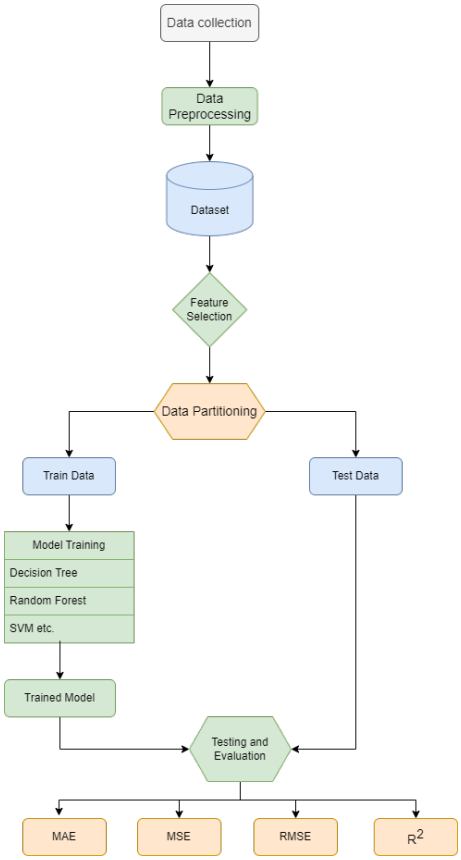

Figure 3.1: Process-Flow Diagram of the Presented Work.

3.2 Data

3.2.1 Dataset

Our dataset was gathered from DisGeNET [18], one of the most reliable and informative websites in gene-disease association studies. The latest version of DisGeNET contains 1,134,942 gene-disease associations (GDAs), between 21,671 genes and 30,170 diseases, disorders, traits, and clinical or abnormal human phenotypes. In the human-to-animal model expert-curated databases, the DisGeNet database integrates approximately 400,000 relationships where genes are between 17,000 and 14,000, and diseases are between 14,000 and 16,000.

From the dataset, we got 26523 combinations of gene-disease relations. Here we got 5 related features which are gene symbol, disease name, association type, NumberOfPubmeds, and score.

Table 3.1: Sample Dataset [7]

Gene Symbol	Disease Name	Association Type	Number Of Pubmeds	score
ATP7B	Hepatolenticular Degeneration	AlteredExpression, Biomarker, GeneticVariation	200	0.97260683
MC4R	Obesity	Biomarker, GeneticVariation	264	0.94
IRS1	Diabetes Mellitus, Type 2	Biomarker, GeneticVariation	112	0.907216439
CFTR	Cystic Fibrosis	AlteredExpression, Biomarker, GeneticVariation	1074	0.9
MECP2	Rett Syndrome	AlteredExpression, Biomarker, GeneticVariation, PostTranslationalModification	498	0.9

GeneSymbol is a particular name given to a particular gene.
DiseaseName is a disorder caused by a certain gene.
AssociationType is the nature of relationships between genes and diseases.
NumberOfPubmeds Quantitative measure of the strength of the connection between a certain gene and a disease.
Score accounts for the number and type of sources and the number of publications that support the association.

3.2.2 Data Preprocessing

Table 3.2: Seperation of Association Type.

Gene Symbol	Disease Name	Altered Expression	Biomarker	Genetic Variation	Post Translational Modification	Therapeutic	Number Of Pubmeds	score
ATP7B	Hepatolenticular Degeneration	1	1	1	0	0	200	0.97260683
MC4R	Obesity	0	1	1	0	0	264	0.94
IRS1	Diabetes Mellitus, Type 2	0	1	1	0	0	112	0.907216439
CFTR	Cystic Fibrosis	1	1	1	0	0	1074	0.9
MECP2	Rett Syndrome	1	1	1	1	0	498	0.9

In our study, we want to predict the association's score, which depends on the associationType and NumberOfPubmeds. The feature association type is a multivalue one. In order to express this association type by 0 and 1. We arrange the value of this feature in separate columns. AlteredExpression, Biomarker, GeneticVariation, PostTranslationalModification, and Therapeutic are the five association types that exist in the database.

In this dataset, there are some diseases with multiple names, for which, we need to select the most popular one. To ensure the success of our approach, we must additionally translate the text data of gene symbols and disease names into numerical values.To convert the strings values of gene id and disease name we used scikit-learn's OrdinalEncoder or LabelEncoder from preprocessing module.

Table 3.3: Final form of input data.

Gene Symbol	**Disease Name**	**Altered Expression**	**Biomarker**	**Genetic Variation**	**Post Translational Modification**	**Therapeutic**	**Number Of Pubmeds**	**score**
0.082264822	0.439266164	1	1	1	0	0	200	0.97260683
0.53738733	0.721556443	0	1	1	0	0	264	0.94
0.461470103	0.291315283	0	1	1	0	0	112	0.907216439
0.164529643	0.249593135	1	1	1	0	0	1074	0.9
0.542211502	0.844799527	1	1	1	1	0	498	0.9

Chapter 4

Implementation of Models

4.1 Results of Test Run

4.1.1 Linear regression

The mean squared error (MSE) score of this model is 0.29, which is a very large number.However, there is still room for improvement in the model to better predict the target variable.

Our R^2 score is 0.23. This means that the model is able to explain about 23% of the variance in the dependent variable.

The regression score is 0.60, which means that our model is 60% accurate in predicting the score of gene and disease association.

4.1.2 Decision tree regression

Our mean squared error (MSE) score is 0.3. An MSE score of 0.3 indicates that, on average, the difference between the predicted values and the actual values of the target variable is quite large.

Our R^2 score is 0.37. This means that the model is able to explain about 37% of the variance in the dependent variable.

The regression score is 0.68, which means that our model is 68% accurate in predicting the score of gene and disease associations.

4.2 Result Evaluation

Here we can see the result was not good in all the cases. There can be several reasons behind this such as

- The values of the train and test datasets were not properly distributed. As the dataset was sorted in descending order,the model trained by the high-scored attribute.
- The method used to convert a string to a numerical value might not be appropriate.

Chapter 5

Discussion

Understanding the relationships between genes and diseases is an important and critical field of research because it can help us understand the causes of disease and create innovative methods for treatment, diagnosis, therapy, and prevention. A proper and timely understanding of the root causes of any disease can be life-saving. Machine learning has the capability of speedy and accurate identification. In our study, we are trying to use this capability.

The result of previous studies, which we reviewed in the literature review section, has proven the effectiveness of machine learning in detecting gene-disease associations. By following their guideline and methodology we are trying to use a diverse set of machine learning algorithms like linear regression, decision tree regression, support vector machines (SVM), random forests, etc.

Concluding, we are hopeful that machine learning models will provide a useful tool for discovering new gene-disease correlations by utilizing the strength of computer algorithms and massive genomic data. The results of this study improve our knowledge of the genetic causes of diseases and lay the foundations for more research in this area

References

[1] "Wikipedia." https://simple.wikipedia.org/wiki/Mutation#/media/File:DNA_replication_en.svg. Accessed on May 15, 2023.

[2] "Byju's." https://byjus.com/biology/homologous-recombination/. Accessed on May 15, 2023.

[3] "spiceworks." https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/. Accessed on May 15, 2023.

[4] https://lewtun.github.io/dslectures/lesson05_random-forest-deep-dive/. Accessed on May 15, 2023.

[5] "Javapoint." https://www.javatpoint.com/types-of-machine-learning. Accessed on May 12, 2023.

[6] "geeksforgeeks." https://www.geeksforgeeks.org/ml-gradient-boosting/. Accessed on May 12, 2023.

[7] https://github.com/dhimmel/disgenet. Accessed on May 12, 2023.

[8] "National human genome research institute." https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics. Accessed on May 12, 2023.

[9] "Medlineplus." https://medlineplus.gov/genetics/understanding/basics/gene/. Accessed on May 15, 2023.

[10] "Human genetic variation." https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/. Accessed on May 12, 2023.

[11] "Great learning." https://www.mygreatlearning.com/blog/what-is-machine-learning/. Accessed on May 15, 2023.

[12] "tutorialspoint." https://www.tutorialspoint.com/machine_learning_with_python/. Accessed on May 12, 2023.

[13] M. Sikandar, R. Sohail, Y. Saeed, A. Zeb, M. Zareei, M. A. Khan, A. Khan, A. Aldosary, and E. M. Mohamed, "Analysis for disease gene association using machine learning," *IEEE Access*, vol. 8, pp. 160616–160626, 2020.

[14] D.-H. Le, "Machine learning-based approaches for disease gene prediction," *Briefings in functional genomics*, vol. 19, no. 5-6, pp. 350–363, 2020.

[15] M. Asif, H. F. 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, 2018.

[16] P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu, "Enhancing the prediction of disease–gene associations with multimodal deep learning," *Bioinformatics*, vol. 35, no. 19, pp. 3735–3742, 2019.

[17] E. M. Hanna and N. M. Zaki, "Gene-disease association through topological and biological feature integration," in *2015 11th international conference on innovations in information technology (IIT)*, pp. 225–229, IEEE, 2015.

[18] "Disgenet." https://www.disgenet.org/. Accessed on May 12, 2023.

Appendix A

Data Normalization Using LabelEncoder

Listing A.1: Gene and Disease Name Conversion

```
le = LabelEncoder()
scaler = MinMaxScaler()

data['geneSymbol'] = le.fit_transform(data['geneSymbol'])
column_to_normalize = 'geneSymbol'
data[column_to_normalize] = scaler.fit_transform(data[column_to_normalize]
                                                 .values.reshape(-1, 1))


data['diseaseName'] = le.fit_transform(data['diseaseName'])
column_to_normalize = 'diseaseName'
data[column_to_normalize] = scaler.fit_transform(data[column_to_normalize]
                                                 .values.reshape(-1, 1))
```

Appendix B

Codes for Linear Regression

Listing B.1: Implementation of Linear Regression

```
model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

Appendix C

Codes for Decision Tree Regression

Listing C.1: Implementation of Decision Tree Regression

```
regressor = DecisionTreeRegressor(random_state=42)
regressor.fit(X_train, y_train)

y_pred = regressor.predict(X_test)
```