Knowledge Graph Embedding and Graph Convolutional Network based approach to Gene-Disease Association Prediction

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

Muhammad Hossain	**190104060**
Md. Monim Hossain	**190104062**
Syed Mushfiqur Rahman	**190104068**
Md. Mahfuzur Rahman	**190104071**

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

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Many diseases such as cancer, diabetes, and cardiovascular diseases, are caused by gene mutations [4]. Thus, investigating the link between these causal genes, and the diseases they produce is a significant area of study in biological research.

Bioinformatics employs computational techniques for biological research. It plays a crucial role in deciphering the process of evolution, inheritance, disease, and the nature of life. Apart from understanding the biological mechanisms, bioinformatics also focuses on understanding the genetic architecture of human diseases [5].

Genome-wide association study (GWAS) [6] is a field of bioinformatics that is used to identify genes likely to be involved in human diseases. It is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait. The method involves surveying the genomes of many people, looking for genomic variants that occur more frequently in those with a specific disease or trait compared to those without the disease or trait. Once such genomic variants are identified, they are typically used to search for nearby variants that contribute directly to the disease or trait. The GWAS results in massive amounts of candidate genes, which make the validation of candidate disease genes time-consuming and expensive. To aid the discovery of disease genes, computational methods can play a vital role to speed up the process.

The last few years have seen an enormous growth in the amount of available data and knowledge, which has led to a surge in the utilization of Machine Learning and Deep Learning techniques. Deep learning algorithms like GCN can learn complex patterns and relationships between genes and diseases, which are difficult to identify using traditional statistical methods. This can improve the accuracy of gene-disease association predictions.

1.2 Problem Statement

Identifying the genetic basis of complex diseases is a challenging task, and predicting disease related genes plays a critical role in understanding the pathology of various genetic disorders. Despite the significant progress in this field, existing prediction methods still suffer from limitations such as low prediction performance and poor generalizability. Thus, the problem addressed in this thesis is to develop an effective and interpretable computational approach for predicting disease-gene associations, which can contribute to our understanding of disease mechanisms and facilitate the development of new therapies.

1.3 Objective

Our thesis aims to implement a Graph Convolutional Neural Network model for predicting gene-disease associations which involves analyzing large and complex datasets including genetic and clinical data. The primary goal is to train our model to recognize complex patterns and relationships between genes and diseases, with the ultimate objective of making precise predictions about gene-disease associations.

1.4 Document Summary

Here, the overall structure of our thesis report is briefly described. The current introduction chapter contains our motivation, main goals and the problem definition. The remaining chapters have been organized as follows:

- **Chapter 2:** This chapter includes a background study on the problem domain, graph neural network model and performance metrics.
- **Chapter 3:** This chapter includes discussion on relevant papers of studies previously conducted.
- **Chapter 4:** In this chapter, each element of our proposed method is elaborated in details.
- **Chapter 5:** This chapter provides discussion about future work and conclusion.

Chapter 2

Background

2.1 Genomics

The field of genomics has revolutionized our understanding of human biology, offering unprecedented insights into the complexity of the human genome and its role in health and disease. For many diseases including cancer, genomics research has revealed key genetic mutations that contribute to the development and progression of the disease.

The entire set of information in an organism's DNA is called its genome [7] as shown in Figure 2.1 and genomics is a branch of biological study which focuses on the structure, function and inheritance of an organism's genome. To identify the combined impact of genes on the growth and development of an organism, genomics examines all genes and their interactions with each other. Functional genomics, which involves the analysis of genes at the functional level, is a primary application of genomics. This can potentially result in novel approaches for disease diagnosis, treatment, and prevention [8]. The Human Genome Project (HGP) was a ground-breaking international research effort that sequenced the entire human genome and provided the foundation for modern genomics research [9].

2.1.1 Gene

A gene is a unit of DNA that holds the instructions for creating a particular protein or set of proteins. There are about 20,000 to 25,000 genes in the human genome [10]. Genes are located on 23 pairs of chromosomes within the nucleus of a human cell. Enzymes copy gene information from DNA to messenger RNA, which then moves from the nucleus into the cytoplasm of the cell. There, ribosomes read the mRNA, use the information to link amino acids together in the proper sequence, and create specific proteins. This gene expression process is illustrated in Figure 2.2.

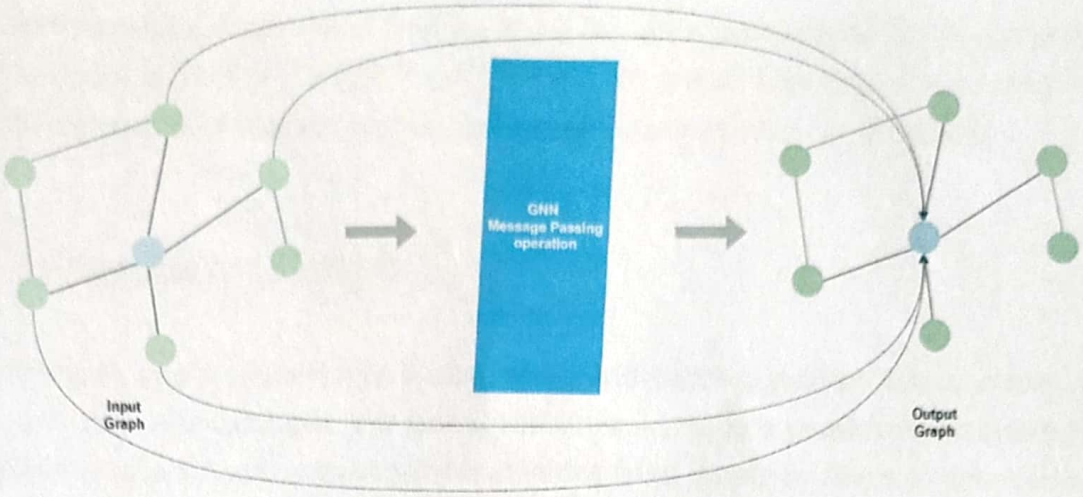

Figure 2.6: Graph Neural Network

sociation network, relationship between humans from Human-Interaction network etc.

Molecule's property detection from molecular structure, image classification etc. are examples of graph-level tasks.

2.2.2 Graph Convolutional Network

Graph Convolutional Network (GCN) is a type of GNN where message passing is done by matrix multiplication of a normalized adjacency matrix $\hat{A}$ with the feature matrix X. The following formulation of GCN was proposed by *T. N. Kipf* and *Max Welling* in their 2016 paper titled *Semi-Supervised Classification with Graph Convolutional Networks* [11]

$$X^{(l)} = Act(\hat{A}X^{(l)}W^{(l)})$$

$$X^{(0)} = X$$

where,

X = Feature matrix representing the vertices i.e., Genes and Diseases

$X^{(l)}$ = Output of Graph Convolutional Layer (GCL) l i.e., new representation of the nodes after applying layer l

$W^{(l)}$ = kernel i.e., weight matrix in layer l

Act = Some activation function like $ReLU$, $Sigmoid$ etc.

$\hat{A} = \check{D}^{-1/2} \times \check{A} \times \check{D}^{-1/2}$

$\check{A} = A + I$

$\check{D}$ = Diagonal node degree matrix from $\check{A}$

A = Adjacency matrix of the knowledge graph

I = Identity matrix

It is a measure of the performance of a binary classification model. The ROC curve (Figure 2.7) is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at different classification thresholds. The AUC-ROC is the area under the ROC curve and is a commonly used metric for evaluating the predictive accuracy of a binary classifier.

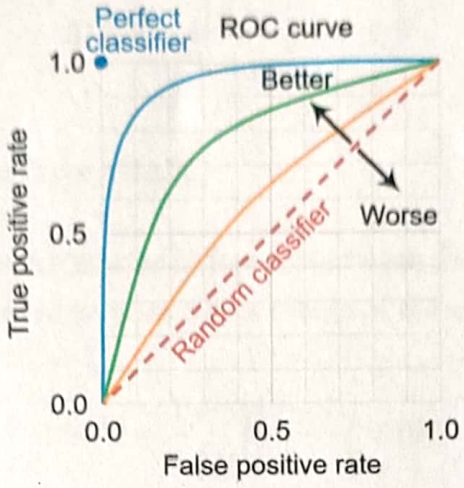

Figure 2.7: ROC Curve [3]

The AUC-ROC ranges from 0 to 1, where a value of 0.5 indicates a random classifier and a value of 1 indicates a perfect classifier. An AUC-ROC of 0.7 or higher is generally considered to be a good classifier.

2.4.2 Performance Metrics for Regression

Mean Absolute Error (MAE)

MAE measures the average absolute difference between the predicted and actual values. It provides a measure of the average magnitude of errors.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

MSE calculates the average squared difference between the predicted and actual values. It penalizes larger errors more compared to MAE, making it sensitive to outliers.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and provides a metric in the same units as the target variable. It gives a measure of the average magnitude of errors and is more interpretable than MSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Percentage Error (MAPE)

MAPE calculates the average percentage difference between the predicted and actual values. It measures the average relative error as a percentage of the actual values.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Chapter 4

Methodology

4.1 Overview

In our study, we approach the problem of predicting association between diseases and genes as a link prediction problem (Figure 4.1). In the subsequent sections, we elaborate each element of our proposed method in details.

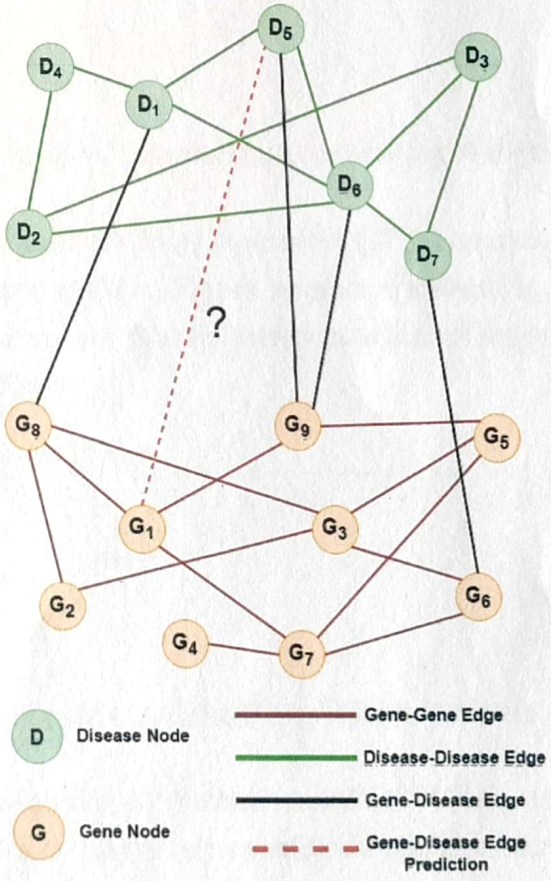

Figure 4.1: Gene Disease Heterogeneous Graph

Chapter 5

Future Work and Conclusion

In conclusion, our study demonstrates the effectiveness of using GCN for graph embedding in predicting the association between genes and diseases. Through our experiments, we observed that our model achieved promising results in terms of minimizing the loss over epochs. But it is not up to the mark yet. There is still room for improvements.

While our study provides valuable insights into the application of GCN for gene-disease association prediction, there are several avenues for future work. Firstly, our method has to be further optimized by exploring different graph convolutional network architectures and tuning hyper-parameters. Secondly, the integration of additional types of biological data, such as gene expression profiles and protein-protein interactions, can enhance the performance of our method and provide a more comprehensive understanding of the molecular mechanisms underlying gene-disease associations. We can also use word2vec or some other word embedding techniques for disease feature representation from text data.

Furthermore, the performance of our model can be compared to state-of-the-art methods on benchmark datasets.

Overall, we believe that the use of GCN for graph embedding has great potential for advancing the field of computational biology and facilitating the discovery of novel gene-disease associations.

References

[1] "A brief guide to genomics." https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics. Accessed: 2023-05-10.

[2] "gene." https://www.britannica.com/science/gene. Accessed: 2023-05-10.

[3] W. Commons, "File:roc-draft-xkcd-style.svg — wikimedia commons, the free media repository," 2021. Accessed: 2023-05-10.

[4] E. Seyyedrazzagi and N. J. Navimipour, "Disease genes prioritizing mechanisms: a comprehensive and systematic literature review," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 6, pp. 1–15, 2017.

[5] N. Gill, S. Singh, and T. C. Aseri, "Computational disease gene prioritization: an appraisal," *Journal of Computational Biology*, vol. 21, no. 6, pp. 456–465, 2014.

[6] "Genome-wide association studies (gwas)." https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies. Accessed: 2023-05-07.

[7] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. New York: Garland Science, 2002.

[8] "genomics." https://www.britannica.com/science/genomics. Accessed: 2023-05-10.

[9] "The human genome project (hgp)." https://web.ornl.gov/sci/techresources/Human_Genome/index.shtml. Accessed: 2023-05-10.

[10] "human genome." https://www.britannica.com/science/human-genome. Accessed: 2023-05-10.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[12] L. Li, X. Zhang, Y. Ma, C. Gao, J. Wang, Y. Yu, Z. Yuan, and Q. Ma, "A knowledge graph completion model based on contrastive learning and relation enhancement method," *Knowledge-Based Systems*, vol. 256, p. 109889, 2022.

[13] L. Zhu, Z. Hong, and H. Zheng, "Predicting gene-disease associations via graph embedding and graph convolutional networks," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 382–389, 2019.

[14] J. Piñero, Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, pp. D833–D839, 10 2016.

[15] C. Y. Kim, S. Baek, J. Cha, S. Yang, E. Kim, E. Marcotte, T. Hart, and I. Lee, "HumanNet v3: an improved database of human gene networks for disease research," *Nucleic Acids Research*, vol. 50, pp. D632–D639, 11 2021.

[16] "Medical subject headings (mesh." https://www.nlm.nih.gov/mesh/meshhome.html. Accessed: 2023-04-28.

[17] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "Pgcn: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks," *bioRxiv*, 2019.

[18] W. Choi and H. Lee, "Identifying disease-gene associations using a convolutional neural network-based model by embedding a biological knowledge graph with entity descriptions," *PLOS ONE*, vol. 16, pp. 1–27, 10 2021.

[19] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[20] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, and D. Phung, "A novel embedding model for knowledge base completion based on convolutional neural network," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, (New Orleans, Louisiana), pp. 327–333, Association for Computational Linguistics, June 2018.

[21] S. Espe, "Malacards: The Human Disease Database," *J Med Libr Assoc*, vol. 106, p. 140–141, 1 2018.

[22] A. P. Heath, V. Ferretti, S. Agrawal, M. An, J. C. Angelakos, R. Arya, R. Bajari, B. Baqar, J. H. B. Barnowski, J. Burt, A. Catton, B. F. Chan, F. Chu, K. Cullion, T. Davidsen, P.-M. Do, C. Dompierre, M. L. Ferguson, M. S. Fitzsimons, M. Ford, M. Fukuma, S. Gaheen, G. L. Ganji, T. I. Garcia, S. S. George, D. S. Gerhard, F. Gerthoffert, F. Gomez, K. Han, K. M. Hernandez, B. Issac, R. Jackson, M. A. Jensen, S. Joshi, A. Kadam, A. Khurana, K. M. J. Kim, V. E. Kraft, S. Li, T. M. Lichtenberg, J. Lodato, L. Lolla, P. Martinov, J. A. Mazzone, D. P. Miller, I. Miller, J. S. Miller, K. Miyauchi, M. W. Murphy, T. Nullet, R. O. Ogwara, F. M. Ortuño, J. Pedrosa, P. L. Pham, M. Y. Popov, J. J. Porter, R. Powell, K. Rademacher, C. P. Reid, S. Rich, B. Rogel, H. Sahni, J. H. Savage, K. A. Schmitt, T. J. Simmons, J. Sislow, J. Spring, L. Stein, S. Sullivan, Y. Tang, M. Thiagarajan, H. D. Troyer, C. Wang, Z. Wang, B. L. West, A. Wilmer, S. Wilson, K. Wu, W. P. Wysocki, L. Xiang, J. T. Yamada, L. Yang, C. Yu, C. K. Yung, J. C. Zenklusen, J. Zhang, Z. Zhang, Y. Zhao, A. Zubair, L. M. Staudt, and R. L. Grossman, "The NCI Genomic Data Commons," *Nature Genetics*, vol. 53, p. 257–262, 2021.

[23] L.-C. Li, H. Zhao, H. Shiina, C. J. Kane, and R. Dahiya, "PGDB: a curated and integrated database of genes related to the prostate," *Nucleic Acids Res.*, 1 2003.

[24] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, pp. i60–i68, 06 2014.

[25] "Pgcn: Disease gene prioritization by disease and gene embedding through gcn." https://github.com/liyu95/Disease_gene_prioritization_GCN. Accessed: 2023-04-28.

[26] "Omim." https://www.omim.org/. Accessed: 2023-04-28.

[27] "Humannet." https://staging2.inetbio.org/humannetv3/. Accessed: 2023-04-28.

[28] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, p. 535–542, 2006.

[29] "Biogps." http://biogps.org/#goto=welcome. Accessed: 2023-04-28.

[30] "Connectivity map." https://www.broadinstitute.org/connectivity-map-cmap. Accessed: 2023-04-28.

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

List of Figures

Contents

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.