MACHINE LEARNING APPROACH FOR DETECTING GENE-DISEASE ASSOCIATION

A thesis
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

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

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

November 2023

CANDIDATES' DECLARATION

We, hereby, declare that the project presented in this report is the outcome of the investigation performed by us under the supervision of Professor Dr. S.M.A. Al-Mamun, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work was spread over two final year courses, CSE 4100: Project and Thesis I and CSE 4250: Project and Thesis II, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md. Al-Amin
190104001

Md. Shafayat Jamil
190104022

Arittra Das
190104083

Swarnajit Saha
190104086

CERTIFICATION

This thesis titled, **"MACHINE LEARNING APPROACH FOR DETECTING GENE-DISEASE ASSOCIATION"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in November 2023.

Group Members:

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

Professor Dr. S.M.A. Al-Mamun
Professor & Supervisor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Professor Dr. Md. Shahriar Mahbub
Professor & Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to Dr. S.M.A. Al-Mamun, our respected Professor in the Department of Computer Science and Engineering at Ahsanullah University of Science and Technology. His inspiration and unwavering support made us capable of working with such a complicated topic as genes and disease. This research is the outcome of this continuous supervision.

We extend our heartfelt thanks to the faculty members of Ahsanullah University of Science and Technology for their thoughtful feedback and recommendations throughout the course of this project.

At last, we acknowledge the Department of Computer Science and Engineering at Ahsanullah University of Science and Technology for fostering an environment conducive to academic growth and research excellence. The department's support, scholarly environment, and dedication to academic excellence greatly aided the completion of this thesis.

Dhaka
November 2023

Md. Al-Amin

Md. Shafayat Jamil

Arittra Das

Swarnajit Saha

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-based 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 Motivation

The investigation of gene-disease connections is a driving force in the dynamic field of biomedical research, providing light on the complex mechanisms underlying a wide range of health disorders. This effort is motivated by the need to improve our understanding of the genetic foundations of illness, with the ultimate goal of paving the way for transformative breakthroughs in medical science.

There is great promise in the study of the relationship between genes and diseases, as any findings could fundamentally change the medical industry. Recognising the intricate relationships that exist between genes and diseases has the potential to revolutionise approaches to diagnosis, treatment, and prevention. By eliminating significant information gaps, the ultimate purpose of this research is to enhance patient outcomes. A breakthrough new era in healthcare is represented by precision medicine, where individual treatment plans are created based on the specific genetic profile of each patient. Precision medicine relies on a deep understanding of the connections between genes and diseases, which allows doctors to treat patients with more targeted, less damaging, and more effective medicines. For any disease to be effectively treated, a timely and precise diagnosis is necessary. Our objective is to investigate the connections between genes and diseases in order to facilitate the development of cutting-edge diagnostic tools. Identifying genetic markers that can serve as early indicators of illness susceptibility is one aspect of this that will allow for the development of precise and timely treatments. Novel therapeutic techniques are made possible by the identification of specific genes associated with disorders. Our research paves the way for ground-breaking treatments that address the root causes of disease rather than just its symptoms, spurred by the prospect of discovering new therapeutic targets. The knowledge gleaned from studying the connections between genes and illnesses has ramifications for public health that extend beyond individual health. Comprehending the genetic basis of diseases informs public health strategies, guides preventive measures, and facilitates the creation of laws meant to lessen the overall effect of diseases on the population. The rapid progress in computational biology and machine learning offers unmatched opportunities for the analysis of large genomic datasets. This research combines cutting-edge technologies with biological research to unleash the potential to extract meaningful patterns from complex genomic data.

In short, the research is driven by an uncompromising dedication to improving patient care, expanding the field of medicine, and influencing the revolutionary field of precision medicine. Our goal is to bring in a new era of healthcare innovation with significant consequences for public health as well as individual well-being by unraveling the complex web of gene-disease relationships.

1.2 Problem Statement

Advancements in genomic research, buoyed by the capabilities of high-throughput technologies, have led to the generation of expansive datasets. This surge in data volume underscores the pressing need for sophisticated analytical tools to navigate and distill meaningful insights from the complex interplay between genes and diseases. In response to this imperative, the present study seeks to leverage the potency of machine learning methodologies.

The dataset under scrutiny comprises pivotal features, each offering a unique perspective on the genomic landscape. These features include, but are not limited to, GeneSymbol, GeneId, DiseaseID, AlteredExpression, Biomarker, GeneticVariation, PostTranslationalModification, Therapeutic, and NumberOfPubmeds. Each of these features represents distinct facets of the genetic and molecular landscape, collectively forming a comprehensive basis for predictive modeling. The ultimate objective of this study is to create a machine-learning model that can predict the relationships between genes and diseases. The numerical association score, a metric intended to measure the degree of correlation between a gene and a particular disease, captures the essence of this predictive modeling endeavor. In order to provide a more nuanced understanding of the complex relationships within the genetic framework, this score acts as a quantitative representation. of the intricate relationships within the genomic framework. By unraveling these complex associations, the study aspires to contribute to the broader understanding of genetic mechanisms underlying diseases.

1.3 Objective

The primary objective of this study is to evaluate the effectiveness of machine learning techniques in predicting associations between genes and diseases. In the traditional process of identifying links between genes and illnesses, testing numerous candidate genes is both time-consuming and resource-intensive. This necessitates a more efficient and cost-effective approach, which we aim to address through the application of machine learning.

The conventional method of testing candidate genes involves a significant investment of both time and financial resources. Our study endeavors to emphasize an alternative, less costly, and time-consuming approach. By harnessing the computational prowess of machine learning, we seek to streamline and expedite the identification and association of genes with specific illnesses. Our core intention is to leverage the advanced computational capabilities of machine learning algorithms to unravel the intricate relationships between genes and diseases. By doing so, we aim to not only reduce the economic and temporal burdens associated with traditional methods but also enhance the efficiency and accuracy of the gene-disease association identification process.

In essence, this study aligns with the overarching goal of optimizing the gene-disease association discovery process, emphasizing the potential of machine learning as a transformative tool in biomedical research.

1.4 Document Structure

This chapter provides background information on the issue at hand as well as an overview of the primary goals and contributions of the dissertation. The structure of the final five chapters is as follows:

- Chapter 2 establishes and explains the basic concepts vital for the understanding of the problem itself namely Genomics, Gene disease association, and machine learning.
- Chapter 3 demonstrates a general structure of the technique and a description of the primary 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

The gene, acknowledged as the fundamental unit of heredity with both physical and functional significance, is constituted by deoxyribonucleic acid (DNA). Particular genes assume a pivotal role in molecular processes, serving as instructive entities for protein synthesis. Comprising diverse sizes, these genetic elements are structured from DNA, presenting notable heterogeneity, with the human genome encompassing an estimated 20,000 to 25,000 genes [1]. This genetic diversity is manifested in varying sizes, ranging from a few hundred to exceeding 2 million DNA bases in the human genomic context [2].

Inheritance dictates that each individual possesses two copies of every gene, inherited from each parent. While the majority of genes are conserved across individuals, a minor fraction (less than 1 percent) [1] exhibits slight variations, known as alleles, contributing to unique physical characteristics in each person. Scientific nomenclature assigns genes distinctive names, often abbreviated through symbols composed of letters and numbers for brevity. For instance, the cystic fibrosis transmembrane conductance regulator, associated with cystic fibrosis, is located on chromosome. Genomics, as a discipline encompassing the exhaustive analysis of an organism's genes, delves into the intricate structures, functions, and evolutionary aspects of genetic components. Each human cell harbors an impressive three billion complete copies of DNA base pairs, emphasizing the complexity of the genetic makeup. Proteins, orchestrated by the 20,000 to 25,000 genes [1], govern structural formations, regulate chemical reactions, and facilitate intercellular communication. Genes transcend isolated functions, interacting internally and with the environment. Deviations, such as mutations in the DNA sequence, can yield anomalous proteins, disrupting physiological processes and paving the way for diseases like diabetes, cancer, hypertension, and many more. Unraveling the mysteries of genomics involves understanding the causes behind genetic variations, with mutation being a common catalyst. Through genomics, researchers and medical professionals aspire to decipher the complexities of genetic information, providing profound insights into health, disease, and the intricate web of molecular interactions, forming a cornerstone in scientific and medical advancements.

2.1.1 Mutations

Mutation is a permanent alteration in the genetic material of a cell or virus, with the potential for transmission to descendants [3]. It serves as a crucial source of genetic variation, forming the basis for evolution by natural selection. Mutations can occur in body cells (somatic mutations) or in egg or sperm cells (germinal mutations), impacting descendant cells or entire organisms. Causes include accidents during DNA transactions and exposure to radiation or reactive chemicals [3]. While mutations are often considered random and mostly deleterious, some can be beneficial in specific environments. Understanding how a single change in the DNA nucleotide sequence can lead to mutation is vital. It may result in the production of an incorrect amino acid, affecting protein structure or function. The genome, consisting of long DNA molecules, is susceptible to mutations anywhere on these molecules, with severe changes often occurring in genes. Mutations encompass various types, including point mutations, base-pair substitutions, and deletions or insertions of single base pairs. Chromosomal mutations, spanning multiple genes, may involve deletions, duplications, inversions, and translocations, influencing gene balance and potentially causing abnormalities. Aneuploidy, characterized by the loss or gain of whole chromosomes, can lead to conditions like Down syndrome. Polyploidy, involving the gain of whole chromosome sets, contributes to the evolution of new species [3]. Mutation is also responsible for a variety of diseases, underscoring its significance in human health. Mobile DNA elements within genomes can contribute to mutation by moving between locations. This movement may affect genes or promote large-scale chromosome mutations through recombination. Comprehending the intricacies of mutation is essential for understanding genetic diversity, evolution, and potential impacts on health and development [3].

A common process of mutation has been depicted in Figure 2.1 below.

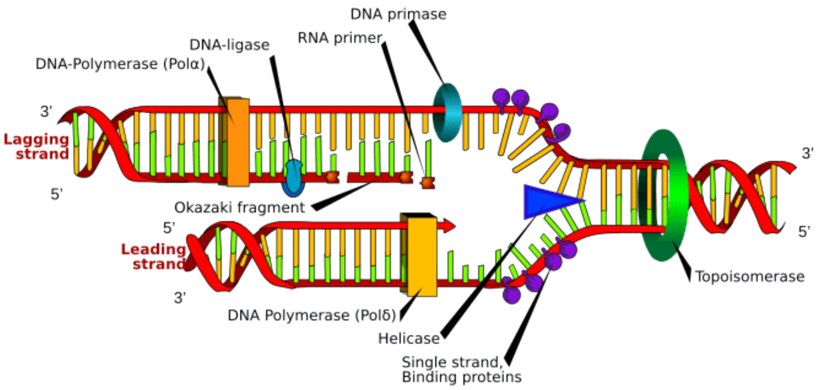

Figure 2.1: Mutation during DNA replication [4].

2.1.2 Recombination

Recombination is a fundamental genetic process crucial for introducing variability within populations. It occurs during meiosis, where maternal and paternal genes undergo rearrangement in the formation of gametes. This process is inherently random during meiosis but is notably augmented by crossing over, disrupting linkage groups and allowing the exchange of segments between paired chromosomes [3]. Consequently, each daughter cell produced in meiosis is haploid, possessing unique genetic content and ensuring constant variability among daughter cells and compared to the parent cell.

In the realm of genetic research, recombination has played a significant role in advancing our understanding of genetic mechanisms. It enables the mapping of chromosomes, identification of linkage groups, investigation into the causes of genetic anomalies, and the manipulation of recombination through gene transplantation. Despite these contributions, experimental recombination presents potential risks, particularly concerning human health and the creation of exclusive disease-resistant varieties. Homologous recombination, involving the exchange of genetic material between DNA strands with similar base sequences, is a natural occurrence in various organisms, including eukaryotes, bacteria, and certain viruses. In eukaryotes, homologous recombination is integral during meiosis, contributing to DNA repair and augmenting genetic diversity through chromosomal crossover. Viruses leverage homologous recombination to shape their evolution [3].

Within genetic engineering, homologous recombination serves as a valuable tool for gene targeting. Engineered mutations can be introduced into specific genes, facilitating the investigation of their functions. This process involves the introduction of foreign DNA with a sequence akin to the target gene, flanked by identical sequences upstream and downstream of the target gene's location. Recognition of these identical flanking sequences by the cell leads to the exchange of target gene DNA with the foreign DNA during replication. Known as gene knockout, this method is applied to study human genetic diseases, offering insights into their underlying mechanisms [3].

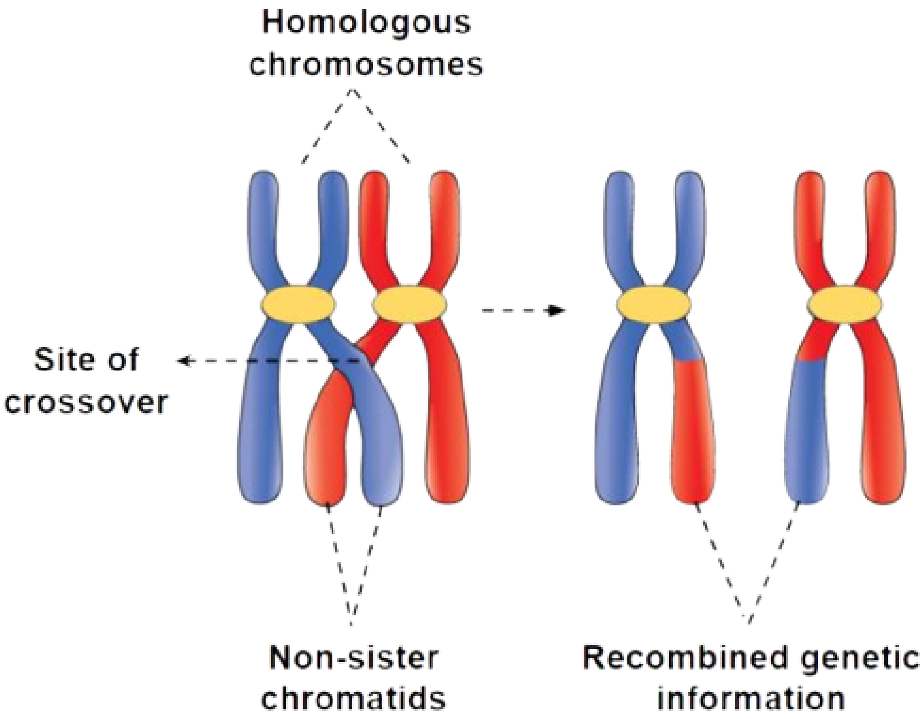

Figure 2.2: Recombination through crossover [5].

2.1.3 Epigenetic modifications

Epigenetics, which investigates modifications influencing gene activity without altering DNA, encompasses the epigenome, characterized by DNA methylation and histone modification. DNA methylation involves the addition of methyl groups to silence genes, while histone modification regulates gene activation or repression by altering histone protein structure. The dynamic nature of the epigenome results in variations among individuals, tissues, and cells, influenced by environmental factors. Crucially, these modifications are heritable, persisting through cell division. Errors in this complex process, such as mismodifications of genes, lead to abnormal gene activity, contributing significantly to genetic disorders like cancers and metabolic diseases. Current research is dedicated to comprehending the intricate relationship between the genome and these modifications, with a specific focus on their impact on gene function, protein production, and overall human health. This exploration is crucial for unraveling the mechanisms underlying genetic disorders, offering potential avenues for targeted therapeutic interventions [6].

2.1.4 Gene Environment Interaction

Gene–environment interaction is a phenomenon characterized by the dynamic interplay between genetic factors and the physical and social environment, thereby exerting a profound influence on the manifestation of phenotypes [7]. Human traits and diseases are intricately shaped by the intricate relationships formed through the complex interplay of one or more genes with environmental elements, including chemicals, nutritional factors, radiation, and societal contexts. Such interactions result in nuanced disease risks contingent upon an individual's specific genotype or environmental exposure. The examination of gene–environment interactions serves as a valuable avenue for discerning the underlying biological mechanisms of diseases, carrying significant implications for public health. A paradigmatic illustration of this concept is observed in the context of the NAT2 gene, where tobacco smoking acts as the environmental factor influencing the risk of bladder cancer. Individuals with distinct variants in the NAT2 gene exhibit an elevated risk of bladder cancer when exposed to tobacco smoke [7], thereby underscoring the pivotal role played by the combined effects of genetic and environmental factors in determining disease susceptibility.

2.1.5 Polygenic Inheritance

Polygenic inheritance, also referred to as quantitative inheritance, pertains to the transmission of a single phenotypic trait influenced by multiple genes. Unlike Mendelian inheritance patterns, polygenic traits are characterized by the involvement of two or more genes, leading to a quantitative measurement of the trait. The resulting phenotypes often exhibit a continuous distribution represented by a bell curve [8]. This inheritance mechanism arises when a particular characteristic is under the control of numerous genes, each exerting a minor influence. While the quantity of genes involved may be substantial, their individual effects are relatively small. The expression of polygenic traits is marked by incomplete dominance, resulting in offspring phenotypes that represent a blend of parental traits. Polygenic traits encompass a spectrum of potential phenotypes, shaped by the intricate interactions between multiple genes. Physical characteristics subject to polygenic inheritance, such as hair color, height, and skin color, along with non-visible traits like blood pressure, intelligence, autism, and longevity, exhibit continuous gradients with various quantifiable increments [8]. This mode of inheritance is not exclusive to humans; other organisms, such as Drosophila, also manifest polygenic traits like wing morphology and bristle count. In summary, polygenic inheritance underscores the complexity of traits, elucidating the multifactorial nature of their genetic control across diverse species [8].

2.2 Gene Disease Association

Gene-disease relationships provide a framework for understanding the genetic foundations of illness presentation. Many genes work together to coordinate physiological activities in the human genome. Genetic variants affecting these genes play a major role in the pathophysiology of a number of diseases, including diabetes, cancer, and hypertension.

The relationship between genes and diseases is regulated by multiple pathways, with genetic differences in the DNA sequence being the main cause. These changes take many different forms, such as insertions, deletions, and single-nucleotide polymorphisms (SNPs). Their capacity to disrupt gene or protein activity has ramifications for several cellular functions and may even initiate disease states.

Changes in gene expression provide a different pathway that connects genes to illnesses. Gene expression describes the complex process by which genetic information is translated into messenger RNA, which leads to the synthesis of proteins. Changes in the sequence of DNA can cause disruptions in the expression of genes, in addition to extrinsic factors like exposure to toxins and changes in food.

Understanding the complex interactions between genes and illnesses is essential to creating effective preventative and therapeutic strategies. Revolutionary developments in the fields of genomics and bioinformatics have made it possible to identify genetic variants that are closely associated with a wide range of disorders. These approaches' incorporation of machine learning speeds up and streamlines the identification process.

A thorough understanding of the underlying genetic factors that contribute to the pathophysiology of disease enables the development of focused and individualised methods for diagnosis, treatment, and prevention. The convergence of state-of-the-art genomic, bioinformatics, and machine learning technology heralds a promising era in the quest for rapid and affordable advances in our understanding and management of diseases, as we traverse the complexities of gene-disease relationships.

2.3 Machine Learning

Machine learning is a complex discipline integral to computer science, providing computational systems with the capability to independently acquire knowledge and make informed decisions. It denotes the scientific exploration of computer programs that employ algorithms and statistical models to glean insights from data through inference and pattern identification, eliminating the requirement for direct programming by humans [9].

The distinctive attribute of machine learning lies in its capacity to enable machines to enhance their performance autonomously over time, devoid of direct human intervention or instructional input. This autonomy is achieved through the assimilation of knowledge acquired from extensive datasets, which are systematically processed by a collection of algorithms during a training phase. Subsequently, these algorithms construct a model based on the accrued knowledge, leveraging it to execute specific tasks.

In machine learning, supervised learning involves training algorithms on labeled datasets, effective for tasks like image recognition and language translation. Unsupervised learning processes unlabeled data to reveal patterns, seen in applications like clustering. Additionally, machine learning contributes significantly to human disease detection, utilizing algorithms to analyze medical data for diagnostic purposes and predicting patient outcomes [9]. Reinforcement learning stands as another noteworthy facet of machine learning, wherein an autonomous agent acquires decision-making capabilities by engaging with an environment. The agent garners feedback in the form of rewards or penalties, refining its behavior towards optimal outcomes. This paradigm finds application in areas such as game theory, robotics, and the development of autonomous systems [9]. The integration of machine learning and deep learning, focusing on multi-layered neural networks like CNNs and RNNs, has greatly enhanced capabilities in areas such as image and speech recognition, natural language processing, and strategic gaming. Furthermore, this synergy plays a vital role in gene-disease association detection, where machine-learning algorithms analyze genetic data to identify links between specific genes and diseases [9].

In healthcare, machine learning facilitates the analysis of medical imaging for diagnostic purposes, prediction of patient outcomes, and expedited drug discovery. The financial sector leverages machine learning for tasks such as fraud detection, risk assessment, and the implementation of algorithmic trading strategies.

2.3.1 Types of Machine Learning

Machine learning can be divided into four primary categories according on how it approaches and learns.

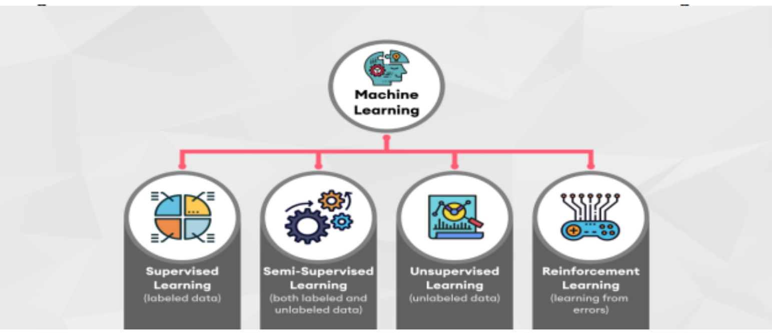
Machine Learning
Supervised Learning (labeled data)
Semi-Supervised Learning (both labeled and unlabeled data)
Unsupervised Learning (unlabeled data)
Reinforcement Learning (learning from errors)

Figure 2.3: Classification of machine learning approaches [10].

Supervised Learning

Supervised learning is a category of machine learning where machines are educated using data that is accurately labeled. This data serves as the foundation for the machines to forecast outcomes. Labeled data implies that some input data has been appropriately marked with the correct output.

The training data in supervised learning serves as a guide, instructing the machines on how to accurately predict outcomes. This is akin to a student's learning process under a teacher's guidance. The process of supervised learning involves supplying the machine learning model with both input data and the correct output data. The supervised learning algorithm's objective is to discover a mapping function that connects the input variable (x) with the output variable (y). Models in supervised learning are educated using a labeled dataset, allowing the model to learn about each type of data. Following the completion of the training process, the model is evaluated using test data (a subset of the training set), after which it forecasts the output. Supervised learning enables the model to predict outcomes based on previous experiences. Supervised learning allows us to gain a detailed understanding of object classes. It aids us in addressing various real-world issues such as detecting fraud and filtering spam. However, these models are not equipped to handle complex tasks. If the test data varies from the training dataset, supervised learning will not be able to accurately predict the outcome. The training process requires a considerable amount of computational time. In supervised learning, a thorough understanding of object classes is necessary [11].

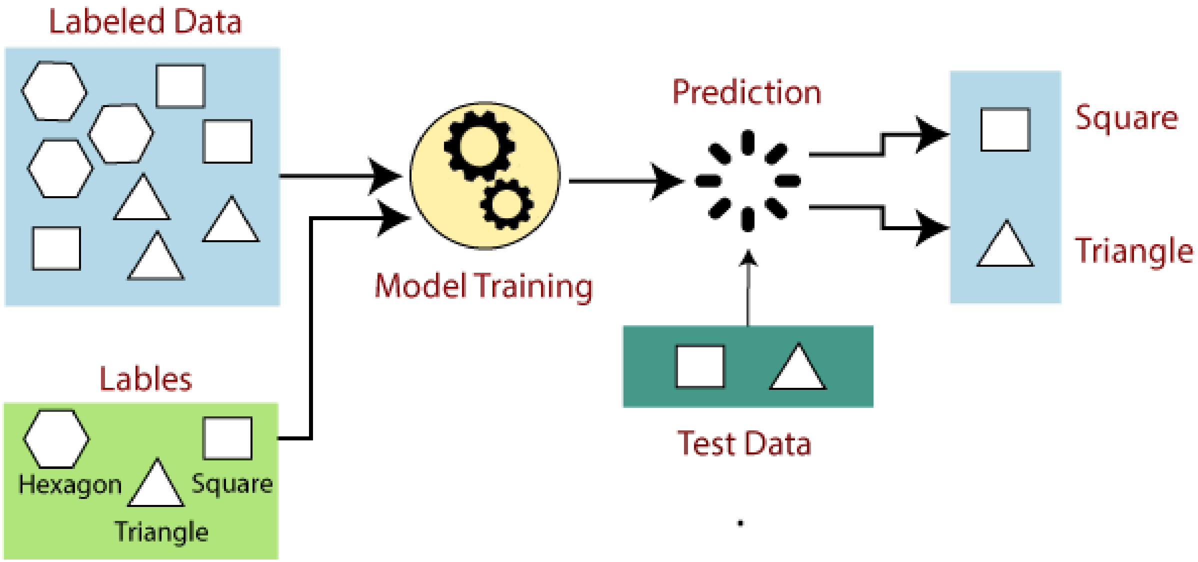

Figure 2.4: Supervised machine learning [11].

Unsupervised Learning

Unsupervised learning, as implied by its name, is a machine learning approach where models autonomously discern latent patterns and insights from provided data without explicit training. This process draws parallels to human cognition when encountering new information. This form of machine learning is characterized by training models on an unlabeled dataset, allowing them to operate without supervision. The primary goal is to uncover the inherent structure of a dataset, group data based on similarities, and represent it in a concise format. Unsupervised learning proves valuable for extracting insights from data, resembling how humans learn from experiences. It excels in handling unlabeled and uncategorized data, making it essential in real-world scenarios where corresponding output data may be unavailable [11]. In contrast to supervised learning, unsupervised learning deals with more complex tasks due to the absence of labeled input data. Obtaining unlabeled data is often more feasible, rendering unsupervised learning a practical choice.

However, the inherent challenge of unsupervised learning lies in its lack of predetermined outputs. The algorithms may produce less accurate results as they operate on data without explicit labels, making it a more intricate endeavor than supervised learning [11].

Semi-supervised Learning

Semi-supervised learning is a classification of machine learning algorithms that operates between the realms of supervised and unsupervised learning methodologies. This approach utilizes both labeled and unlabeled datasets during the training phase, serving as a vital bridge between the well-defined domains of supervised and unsupervised machine learning. The adoption of semi-supervised learning becomes particularly relevant when the acquisition of labeled data proves to be resource-intensive. In practical scenarios, obtaining labeled data can incur substantial costs, prompting a judicious use of labeled instances tailored to corporate objectives.

In the context of semi-supervised learning, the training dataset comprises a mixture of labeled and predominantly unlabeled data. It is noteworthy that the quantity of labeled data is significantly smaller than the abundance of unlabeled data, reflecting the inherent expense associated with labeling. To optimize resource utilization, the algorithm initially employs unsupervised learning techniques to cluster similar data. This clustering process facilitates the assignment of labels to previously unlabeled data, effectively converting it into labeled data.

It is imperative to recognize that labeled data acquisition is a relatively more costly endeavor than unlabeled data acquisition. Semi-supervised learning addresses this challenge by strategically incorporating labeled instances, thereby contributing to a more efficient and cost-effective learning process. Pseudo-labeling is a technique commonly employed in semi-supervised learning, allowing models to be trained with a reduced amount of labeled training data compared to conventional supervised learning methods. This method enhances the model's capacity to generalize patterns and insights from a limited labeled dataset, enabling informed predictions on larger, predominantly unlabeled datasets [11].

Reinforcement Learning

Reinforcement Learning constitutes a feedback-driven machine learning technique, where an agent learns to navigate an environment by executing actions and observing the subsequent outcomes. Positive feedback is received for favorable actions, while negative feedback or penalties are incurred for unfavorable ones. In contrast to supervised learning, Reinforcement Learning operates without the reliance on labeled data, necessitating the agent to learn solely from experiential insights. In the absence of labeled data, the agent's learning process is inherently empirical, particularly adept at solving problems involving sequential decision-making and long-term objectives, such as game-playing and robotics. The agent autonomously engages with and explores its environment, aiming to refine performance by maximizing positive rewards.

Reinforcement Learning's learning mechanism relies on a trial-and-error approach, with the agent incrementally honing its ability to perform tasks more adeptly through iterative experiences. This approach epitomizes a machine learning paradigm where an intelligent agent, devoid of pre-programmed instructions, interacts autonomously with its environment to refine decision-making capabilities.

This methodology is integral to Artificial Intelligence, serving as the foundation for AI agents. It eliminates the need for human intervention to pre-program the agent, allowing it to undergo self-guided evolution through its own experiences. Two principal forms of reinforcement exist within Reinforcement Learning: Positive Reinforcement and Negative Reinforcement. Positive Reinforcement involves amplifying the likelihood of desired behavior by introducing positive stimuli. While effective in instigating behavioral changes, an excess of positive reinforcement may lead to state overload, potentially diminishing its efficacy. Conversely, Negative Reinforcement mitigates undesirable behavior through the avoidance of negative conditions. This form of reinforcement can be highly effective, contingent on context and behavior, primarily serving to meet the minimum behavioral threshold [11].

2.3.2 Machine Learning Models for Gene-Disease Association

Linear Regression

Predicting the value of a single factor based on the values of other factors is the aim of regression problems. When using linear regression, we attempt to determine the exact form of a linear relationship that we think exists between the target label and the supplied input characteristics.The value of the dependent variable will also alter in accordance with any changes in the values of the independent variables (increase or decrease). Simple linear regression and multiple linear regression are the two different types of linear regression. A single dataset characteristic is used in simple linear regression to predict a response. 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.

We can describe it mathematically in the following way: Examine a dataset that has y as the dependent variable and n observations, or independent variables, and p characteristics, or dependent variables. This is how the regression line for p characteristics can be computed-

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + \varepsilon_i \tag{2.1}$$

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

One kind of supervised machine learning approach for continuous variable prediction is called decision tree regression. In order to provide meaningful continuous output, decision tree regression analyses aspects of a dataset and trains a model on the tree's structure to predict data in the future. The leaves of the tree represent the predicted output values for the corresponding input regions. A commonly used machine learning strategy, decision tree regression can handle both continuous and categorical input data and is easy to understand. Starting with the root node, also known as the parent node (dataset), a decision tree is built, splitting into left and right child nodes (subsets of the dataset). These kid nodes split off into their own children's nodes and become those nodes' parent nodes [11].

Random Forest

For problems with regression and classification, the widely used supervised machine learning algorithm Random Forest is effective. Its foundation is the idea of ensemble learning, in which several decision trees are joined to solve a challenging issue and enhance the functionality of the model. Using several subsets of the provided dataset, the algorithm creates a forest of decision trees, then averages them to increase prediction accuracy. The random forest forecasts the outcome based on the majority vote by combining predictions from numerous decision trees, in contrast to a single decision tree. By increasing the number of trees in the forest, overfitting is avoided and accuracy is improved. In comparison to other algorithms, it requires less training time. It operates effectively even with massive datasets and makes highly accurate output predictions [11].

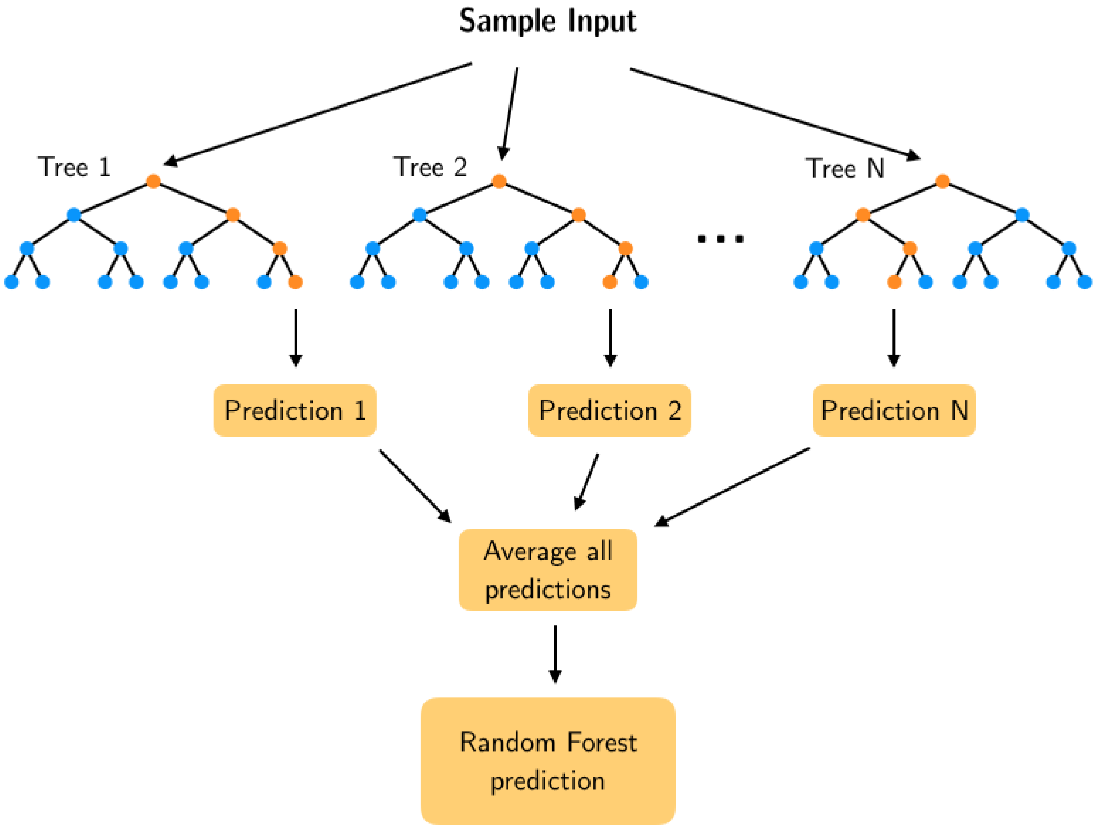

Figure 2.5: Random Forest [13].

Support Vector Regression

A well-liked machine learning approach for classification and regression issues is called Support Vector Machine (SVM). SVM seeks to identify the optimal border or line that divides n-dimensional space into distinct classes so that future data points can be classified with ease. We refer to this decision boundary as a hyperplane. Support vectors are the extreme points that SVM selects to aid in the creation of the hyperplane. These support vectors are the source of the algorithm's name. SVM comes in two types: non-linear and linear. When utilizing Linear Support Vector Machines (SVM), a dataset with linearly separable data can be divided into two classes using a single straight line. On the other hand, non-linear SVM use Non linearly separable data, where the data can't be classified by using a straight line [11].

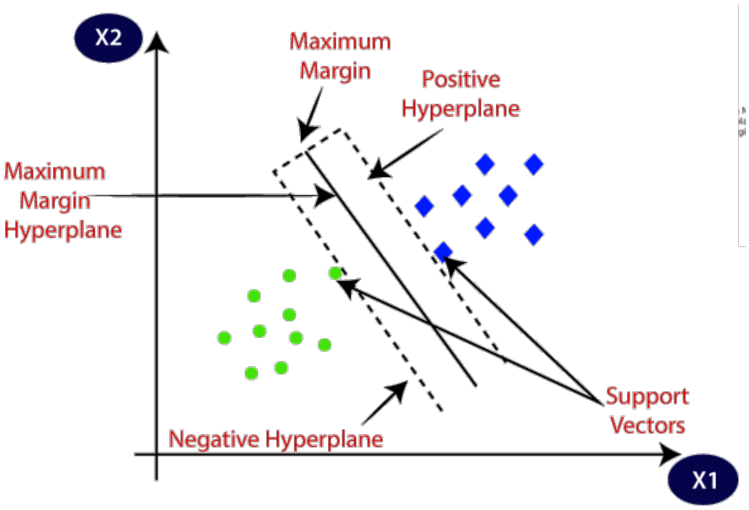

Figure 2.6: Support vector regression [11].

K-Nearest Neighbors (K-NN)

The K-Nearest Neighbors (K-NN) algorithm stands out as a fundamental component in the domain of Supervised Learning within Machine Learning. It operates on the premise of assessing the similarity between newly introduced cases or data and the existing dataset, ultimately assigning the new case to the category most closely aligned with those already established. The K-NN algorithm effectively manages and organizes all available data, facilitating the classification of a new data point by evaluating its likeness to the stored dataset.

K-NN is a versatile algorithm applicable to both Regression and Classification tasks, with a primary emphasis on Classification. It adopts a non-parametric stance, refraining from making assumptions about the underlying data distribution. This characteristic, coupled with its discerning approach, categorizes it as a 'lazy learner' algorithm. Unlike immediate learning from the training set, the K-NN algorithm retains the dataset during training and applies its classification methodology when presented with new data.

The algorithm's implementation is straightforward, contributing to its attractiveness. Its ability to withstand the influence of noisy training data is a noteworthy feature, and its efficacy tends to improve with larger training datasets. Nevertheless, challenges arise in determining the optimal value for K, the parameter denoting the number of neighbors considered during classification.

While K-NN boasts simplicity and adaptability, it is crucial to recognize certain drawbacks. The computational cost is a significant consideration due to the need to calculate distances between data points for all training samples. This computational overhead warrants careful consideration when evaluating the algorithm's suitability for specific applications. In summary, K-Nearest Neighbors offers a straightforward and adaptable approach, but careful attention to computational costs and parameter tuning is essential for prudent application in real-world scenarios.

K-Nearest Neighbors (KNN) classifies a new data point, x_1, by measuring its distances to the k-nearest neighbors in the feature space. The algorithm employs majority voting, assigning the class label most common among these neighbors to x_1. Between two classes, A and B, the class with the majority among the k-nearest neighbors determines the classification of x_1. Careful parameter selection, particularly for 'k', is crucial for optimal performance in the KNN algorithm.

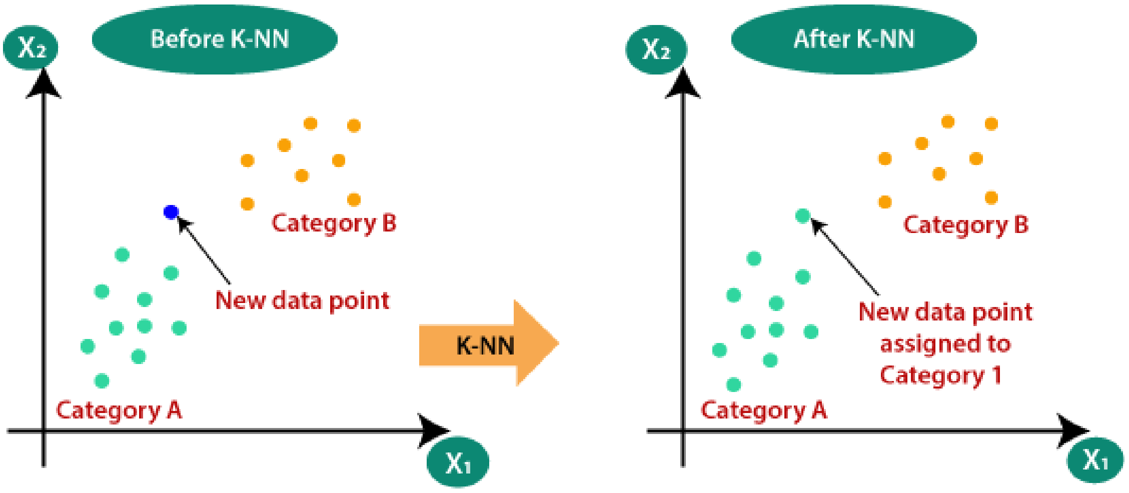

Figure 2.7: K-Nearest Neighbors [13].

Gradient Boosting

Gradient Boosting is a potent boosting method that turns multiple weak learners into strong learners. It uses gradient descent to train each new model to minimize the loss function, such as mean squared error or cross-entropy of the preceding model. The method calculates the gradient of the loss function with respect to the current ensemble's predictions in each iteration and then trains a new weak model to minimize this gradient. After that, the new model's predictions are included in the ensemble, and the process is continued until a stopping requirement is satisfied. XGBoosting and CatBoosting are two powerful variants of gradient boosting algorithms.

The ensemble includes a few trees. The first tree is trained using the feature matrix X and the labels y. The predictions of that tree are used to compute the training set residual errors or r1. After that, feature matrix X and r1 are used as labels to train the following tree. The expected result of this tree and r1 are then used to compute the residual r2. This continues until every tree in the group has received training [14].

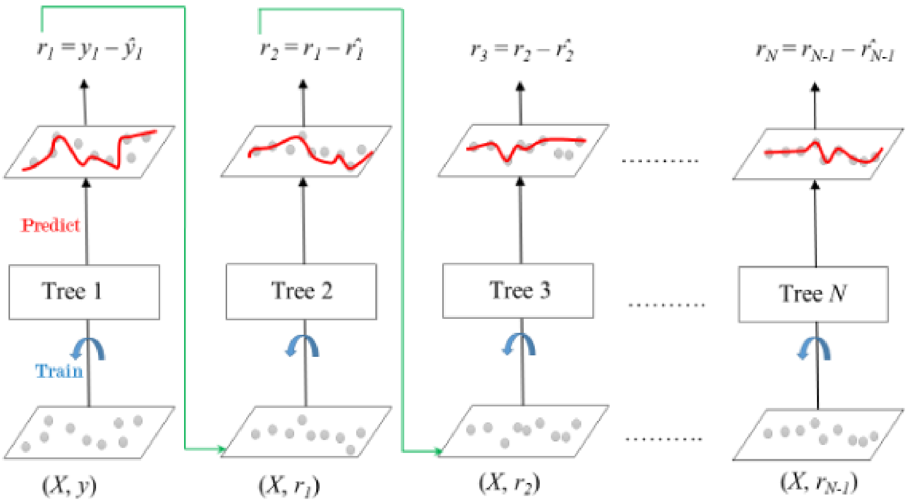

Figure 2.8: Gradient Boosted Trees [14].

XGBoost

One popular ensemble learning technique, called XGBoost, is well known for its ability to get around the drawbacks of using individual machine learning models. XGBoost is an example of ensemble learning, which methodically combines the predictive abilities of several learners to produce a unified model that combines information from several sources, regardless of whether the learning algorithms are the same or different. Boosting and bagging, two popular ensemble learning strategies, work very well in conjunction with decision trees.

XGBoost, which is a gradient-boosting implementation, adds various features that make it popular. To penalize complex models and avoid overfitting, it combines regularisation techniques, using both L1 and L2 regularisation [15]. A sparsity-aware split discovery technique that is adept at handling a variety of sparsity patterns resulting from missing values or data pretreatment operations like one-hot encoding demonstrates its robustness in handling sparse data. One of XGBoost's unique features is that it makes use of a weighted quantile sketch, which allows it to handle weighted data effectively [15]. This is something that other tree-based algorithms frequently lack. Through the use of several CPU cores, the system's block structure for parallel learning improves computational performance. This novel method uses in-memory units (blocks) to store sorted data, making it easier to reuse the data structure in later iterations and increasing productivity for tasks like split finding and column subsampling. Additionally, XGBoost exhibits cache awareness by carefully allocating internal buffers for non-continuous memory access in each thread while retrieving gradient statistics by row index. This design contributes to computational efficiency by making the best use of hardware resources [15].

The data science community has widely accepted XGBoost as an open-source library due to its high accuracy, scalability with large datasets, computational efficiency, versatility in handling different types of data and objectives, integration of regularisation techniques, provision of feature importance scores for interpretability, and more. The purpose of this study is to investigate XGBoost's potential and applications in a variety of data science contexts.

CatBoost

A popular open-source toolkit for classification, regression, and ranking problems, CatBoost is well-known for its effectiveness in gradient boosting on decision trees. Its ability to handle big datasets with categorical characteristics is especially impressive because it incorporates ordered boosting, random permutations, and gradient-based optimization [16]. CatBoost uses a different approach to create balanced trees with a symmetrical structure than other algorithms that are similar, such as XGBoost and LightGBM. Aside from being a kind of regularisation to avoid overfitting, this balanced design has benefits like fast CPU implementation and shortened prediction times [16]. CatBoost presents the idea of ordered boosting as a solution to overfitting problems in small or noisy datasets. This technique addresses target leakage and overfitting issues by training the model on one data subset and computing residuals on a different one. CatBoost uses a number of methods, including as ordered boosting, decision tree optimization, and feature engineering, to improve the efficiency and accuracy of gradient boosting. To update predictions, the technique computes the negative gradient of the loss function in each iteration. A line search approach is then used to find the scaling factor. CatBoost employs gradient-based optimization in decision tree construction, aligning trees with the loss function's negative gradient. With this focused method, trees can concentrate on feature space areas that have the biggest influence on the loss function, resulting in predictions that are more accurate [16]. CatBoost has included a noteworthy approach called ordered boosting, which uses a certain sequence of feature permutations to optimize the learning objective function. It is especially useful for datasets with a large number of features because this innovation improves model correctness and speeds up convergence. The overfitting detector in CatBoost, which is intended to stop training when overfitting is noticed, is essential to the system's effectiveness. The model performs better when this feature is applied, which strengthens its resistance to new data. CatBoost outperforms XGBoost and LightGBM in terms of prediction speed and accuracy because of its special features and methods [16]. CatBoost is a great option for big data applications because of its scalability on large datasets, which allows for distributed training across several computers and GPUs. This flexibility upholds CatBoost's reputation as a strong and adaptable instrument in machine learning applications.

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' [17], proposed some cutting-edge methods for computing 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%-99%, using biological features but with a low FP rate. They overcome this by using topological features to train their dataset. For their computational technique, they get 93.8% precision, 93.1% recall, and 92.9% F-measure. Random Forest performs better than any other experimental setup. Limited information about genes and diseases restricts performance. At the same time, weighted features of gene interaction networks can also be used to increase accuracy.

Le et al., in their study named 'Machine learning-based approaches for disease gene prediction ' [18], gives a wide knowledge of machine learning algorithms for finding the relation between genes and diseases. They discuss two unaries (one-class SVM and one-class Hempstalk), three semi-supervised learning approaches (Graph-based SSL, positive and unlabeled learning, and ANN), five conventional binary classification techniques (DT, k-NN, NB, SVM, and ANN), and three contemporary approaches (Ensemble Learning, Deep Learning).

A training set was created in this work to test several classification techniques for illness gene prediction. To acquire known disease proteins, known disease genes were gathered from OMIM and linked to the human PPI network. An unlabeled set (U) comprising unknown illness proteins and non-disease proteins was created using the remaining proteins. For binary classification algorithms, proteins from the unlabeled collection were selected at random to create a negative training set (N). In order to construct the training set for the binary SSL-based approach, neighbors of recognized illness proteins in the human PPI network were also gathered. Finally, PU-based approaches, such as the multi-level weighted SVM-based method employed in PUDI, were trained using both the positive (P) and negative (N) training sets.

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%, and one-class SVM was the worst of the overall competitive methods.

Asif et al., in their research titled 'Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology' [19] 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 performed better than previously published ASD classifiers were trained and assessed on functional similarity between ASD and non-ASD genes. For this work, they used a range of machine learning classification approaches, such as the decision tree-based Random Forest (RF) method, Naive Bayes (NB), and linear and radial SVM. For the proposed methodology, genes with evidence of participation in ASD were gathered from the Simons Foundation Autism Research Initiative (SFARI) gene database (N = 990). Genes in the SFARI database are categorised into seven groups according to the quality of the evidence that is currently available. During performance testing, the RF classifier operated at its best when the number of trees was set at 500. The maximum accuracy was 80%, and the classifier correctly identified 73 ASD genes out of 554 genes.

Luo et al., in their work titled 'Enhancing the prediction of disease-gene associations with multimodal deep learning' [20] suggest using multimodal DBN (dgMDL) to predict disease-gene relationships. More precisely, two DBNs independently pick up concepts from gene ontology and learn hidden representations of protein-protein interaction networks. Next, a joint DBN is used to train cross-modality representations from the two sub-models using the joint output of their latent representations as the multimodal input. 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” [21] 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. After that, they merge biological information from various sources to find any shared characteristics that might characterize each class. Degree, eccentricity, betweenness centrality, closeness centrality, authority, hub, modularity class, page rank, component ID, clustering coefficient, number of triangles, and eigenvector centrality are the topological traits that they extract. 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.

Mukherjee et al., in their study named "Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network" [22], suggested using a random forest classifier called DiGePred to find potential gene pairs associated with digenic diseases. Its capacity to identify digenic gene pairs of relevance drawn from the literature perfectly illustrates the usefulness of this unique classifier, which has exhibited noteworthy results in terms of high precision and recall. Carefully training the DiGePred classifier on a dataset comprising gene pairs from disease-free individuals as well as those known to cause disease is necessary. The classifier gains the capacity to differentiate between gene pairs that may be harmful and those that are not as a result of this training. Looking more closely reveals that DiGePred is particularly good at finding correlations between gene pairs that are important for the start and progression of different diseases. This is particularly useful when discussing uncommon illnesses where several genes are involved.

Barman et al., in their research titled "Identification of Infectious Disease-Associated Host Genes Using Machine Learning Techniques" [23], employed computational models to identify genes associated with infectious diseases. Their approach yielded favorable results, achieving an accuracy rate of 86.33%. Furthermore, the model demonstrated effectiveness in identifying common genes associated with infectious diseases, cancer, and metabolic disorders, even when applied to new data. The authors trained their model using a large dataset containing information about genes and proteins, and they explored various methods, including Support Vector Machines (SVM), Naive Bayes (NB), Random Forest (RF), and Deep Neural Networks (DNN). Notably, the DNN method performed exceptionally well in reliably identifying disease-related genes. When the authors applied the DNN model to proteins that had undergone rigorous review, it successfully predicted additional disease-related genes, which were subsequently validated through experiments. This discovery holds significant promise for enhancing our understanding of diseases and facilitating the development of novel treatments.

Opap et al., in their article titled ‘Recent advances in predicting gene–disease associations’ [24] provide a comprehensive review of recent computational methods for predicting gene disease associations, categorizing approaches into genome variation, text mining, crowdsourcing, and networks. The genome variation section discusses the significance of GWAS and genetic linkage studies, emphasizing databases like GWAS Catalogue and ClinVar. Tools such as Exomiser and algorithms like SIFT, PolyPhen-2, and CADD are highlighted for prioritizing gene–disease associations based on genomic data. The text mining segment underscores the growing need for automated extraction of gene–disease associations from the scientific literature. Tools like tmVar, DNorm, and GNormPlus, along with resources like PubTator, are discussed as valuable text-mining assets. Crowdsourcing, defined as delegating tasks to a large group, is explored in the context of gene–disease associations, with examples such as Dizeez and the hybrid method of Burger et al. employing platforms like Crowdflower and Amazon Mechanical Turk. Network-based algorithms, relying on functional relationships between genes, are examined through examples like HeteSim and recommendation-based approaches such as that of Natarajan and Dhillon. The article emphasizes the challenges of scaling up gene–disease associations due to the exponential growth of biomedical databases, leading to innovative solutions like modular tool development and the integration of diverse data sources. Challenges related to standardization and data harmonization across tools are addressed, and the importance of ontologies for data standardization is highlighted. In summary, the review highlights the strengths and challenges of computational methods in predicting gene–disease associations, showcasing innovative approaches that leverage genomic data, text mining, crowdsourcing, and network-based algorithms to advance our understanding of disease etiology on a larger scale.

Wang et al., in their study named ‘Predicting gene-disease associations from the heterogeneous network using graph embedding’ [25] introduces a novel approach, the Heterogeneous Network-based Representation Learning Ensemble Method (HNEEM), designed to enhance gene-disease association prediction. Leveraging graph embedding and ensemble learning, HNEEM utilizes a heterogeneous network constructed with gene-disease, gene-chemical, and disease-chemical associations. The network incorporates genes, diseases, and chemicals as nodes, interconnected by their associations. Representation learning methods extract vectors of nodes in the heterogeneous network, merging feature vectors of genes and diseases to represent gene-disease pairs. A random forest classification engine is employed to build prediction models based on these pairs. The study evaluates six representative graph embedding methods—LE, GF, HOPE, DeepWalk, node2vec, and SDNE—comparing their performances in gene-disease association prediction. The results demonstrate the efficacy of graph embedding methods, with further improvements achieved through the integration of different embedding methods. In computational experiments, HNEEM outperforms existing gene-disease prediction methods, showcasing robustness and effectiveness across varying data richness. Case studies validate the utility of HNEEM.

In conclusion, HNEEM emerges as a promising method for predicting gene-disease associations, offering a comprehensive and innovative approach that harnesses the potential of graph embedding and ensemble learning.

Chapter 3

Methodology

3.1 Overview

This study aims to develop a robust machine-learning methodology for predicting relationships between genes and diseases, leveraging association types and the count of related PubMed articles as key features. The research centers on five different categories of gene-disease associations: therapeutic, post-translational modification, altered expression, genetic variation, and biomarker. Our primary objective is to predict the association score between genes and diseases, representing the strength of their correlation.

To achieve this goal, we assembled a comprehensive dataset of gene-disease associations from publicly available databases. The dataset underwent meticulous preprocessing to extract relevant features, ensuring that the subsequent machine-learning models could effectively capture the nuances of the relationships under consideration. We employed a diverse array of machine-learning algorithms to train our models. Notably, different machine learning algorithms were chosen for their distinct strengths in handling the complexities inherent in gene-disease association prediction. The selection of these algorithms was driven by their applicability to the specific nature of our dataset and the task at hand.

The features considered in our models encompass association types and the number of related PubMed articles. This deliberate choice of features is rooted in the significance of understanding the diverse molecular mechanisms (as denoted by association types) and the wealth of biological evidence available in the scientific literature (reflected in PubMed articles) to gauge gene-disease relationships comprehensively. The performance of our approach is rigorously assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) Score. These metrics provide quantitative insights into the accuracy and predictive capabilities of our machine-learning models, allowing for a comprehensive evaluation of their effectiveness.

Throughout the study, we engage in a discussion of potential factors that may constrain or influence the outcomes of our approach. By critically examining the results, we aim to provide a nuanced understanding of the challenges and opportunities inherent in predicting gene-disease associations. This thesis represents a dedicated effort to contribute to the field of bioinformatics by introducing a machine-learning framework tailored for predicting gene-disease relationships. The utilization of diverse algorithms, careful feature selection, and thorough evaluation metrics collectively underscore our commitment to developing a robust and interpretable methodology. Through this research, we aspire not only to advance predictive capabilities but also to inspire future exploration and innovation in the dynamic realm of computational biology.

3.2 Data Description

Our dataset was gathered from DisGeNET [26], 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: Collected Dataset [26].

geneId	**geneSymbol**	**geneName**	**diseaseId**	**diseaseName**	NumberOf Pubmeds	**score**	association Type
540	ATP7B	ATPase	umls:C0019202	Hepatolenticular Degeneration	200	0.97260683	AlteredExpression, Biomarker, GeneticVariation
4160	MC4R	melanocortin 4 receptor	umls:C0028754	Obesity	264	0.94	Biomarker, GeneticVariation
5621	PRNP	prion protein	umls:C0022336	Creutzfeldt-Jakob Syndrome	272	0.884360623	AlteredExpression, Biomarker, GeneticVariation
1756	DMD	dystrophin	umls:C0013264	Muscular Dystrophy	510	0.865413064	AlteredExpression, Biomarker, GeneticVariation
2200	FBN1	fibrillin 1	umls:C0024796	Marfan Syndrome	289	0.858256352	AlteredExpression, Biomarker, GeneticVariation

GeneID is a numerical or alphanumeric code assigned to a gene to provide a systematic and standardized way of referencing genes across databases and scientific literature.

GeneSymbol is a unique identifier for a gene. It is typically a short alphanumeric code that is based on the gene name. For example, the BRCA1 gene, which is associated with breast cancer, has the gene symbol BRCA1.

Gene name refers to the unique identifier assigned to a specific gene, typically based on its function or characteristics. It serves as a label for a particular segment of DNA that encodes information for the synthesis of functional molecules, such as proteins or RNA.

Disease ID represents a unique identifier associated with a specific medical condition or disease in biomedical databases. These identifiers, often linked to standardized vocabularies like the Unified Medical Language System (UMLS), facilitate the organization and retrieval of information related to diseases. Disease IDs are instrumental in connecting genetic information with specific health conditions, aiding researchers and healthcare professionals in understanding the genetic basis of diseases and developing targeted treatments.

DiseaseName is the name of a disorder or medical condition. Disease names can be based on symptoms, causes, or affected organs.

AssociationType is the nature of relationships between genes and diseases.There is five association type in the dataset Altered Expression, Biomarker, Genetic Variation, Post-Translational Modification, and Therapeutic. The strength of the connections between genes and illnesses is defined by these association types in this dataset.

NumberOfPubmeds is a quantitative measure of the strength of the connection between a certain gene and a disease. The number of PubMed is a measure of how many times a gene and a disease have been mentioned together in scientific literature. A higher number of PubMed suggests a stronger connection between the gene and the disease.

Score is the measure of the strength of the evidence supporting the association between a gene and a disease. The score is based on a number of factors, including the number and type of sources, the number of publications, and the quality of the evidence. A higher score suggests stronger evidence supporting the association between the gene and the disease.

In this dataset, it is noteworthy that the depiction of the relationship between genes and diseases is characterized by a nuanced complexity. Specifically, the dataset reflects the reality that the causative factors for a given disease can emanate from either a singular gene or the interplay of multiple genes. This characteristic underscores the intricacies inherent in the dataset, providing a comprehensive portrayal of the diverse genetic influences on disease manifestation. Here is a straightforward scenario for the collaboration.

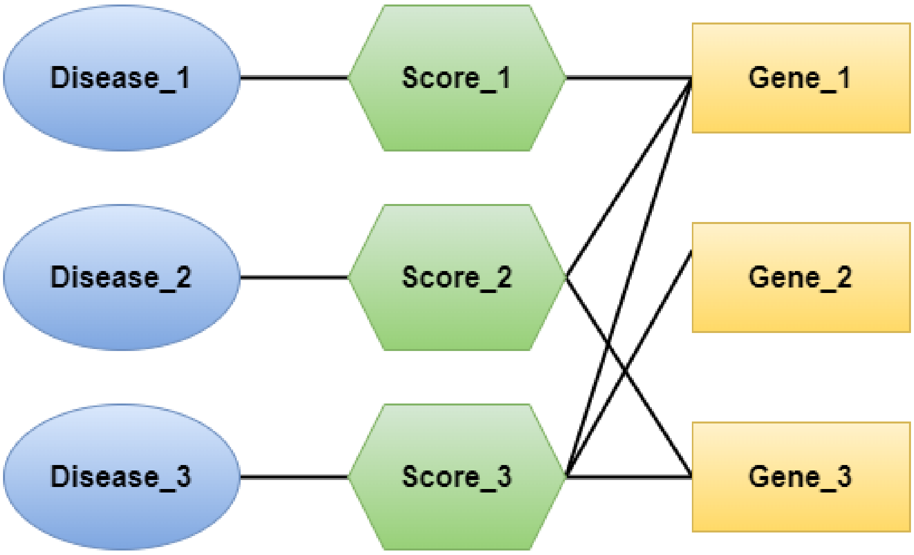

Figure 3.1: Genes-Diseases Correlation.

3.3 Data Preprocessing

3.3.1 Data Segmentation

As part of our extensive data pretreatment workflow, we carefully separated the cases according to the selected variable. In order to properly identify cases where the association score between genes and diseases met or above predetermined threshold values, such as 0.25, 0.30, 0.35, and 0.40, this segmentation approach was designed. This structured approach is essential in forming the input data that our machine learning models receive. We want to give the models a more sophisticated grip on the data by examining a variety of relationship score thresholds, highlighting varying degrees of link between genes and diseases. This complex viewpoint is essential for the models to identify and assimilate patterns associated with different degrees of relationships.

3.3.2 Association Type Representation

In the preprocessing phase, a meticulous approach was undertaken to represent the diverse association types within the dataset effectively. Specifically, the five association types—Altered Expression, Biomarker, Genetic Variation, Post-Translational Modification, and Therapeutic—were manually encoded using a binary representation (0 and 1). This strategic encoding methodology aimed to imbue the dataset with a structured format, enabling the machine learning models to interpret and leverage the nuances associated with each association type during the subsequent training process.

Table 3.2: Seperation of Association Type.

geneId	geneSymbol	diseaseId	Altered Expression	Biomarker	Genetic Variation	Post Translational Modification	Therapeutic	Number Of Pubmeds	score
540	ATP7B	19202	1	1	1	0	0	200	0.97260683
4160	MC4R	28754	0	1	1	0	0	264	0.94
5621	PRNP	22336	1	1	1	0	0	272	0.884360623
1756	DMD	13264	1	1	1	0	0	510	0.865413064
2200	FBN1	4796	1	1	1	0	0	289	0.858256352

3.3.3 Label Encoding

Building upon the segmented dataset, label encoding was applied to transform string-based categorical variables into numerical representations. This deliberate conversion facilitated a more seamless interpretation of the data by the machine learning models. By numerically encoding categorical variables, the models could effectively navigate and process this information during training, contributing to their overall interpretability and predictive accuracy.

Table 3.3: Final form of input data.

geneId	geneSymbol	diseaseId	Altered Expression	Biomarker	Genetic Variation	Post Translational Modification	Therapeutic	Number Of Pubmeds	score
540	0.082264822	19202	1	1	1	0	0	200	0.97260683
4160	0.53738733	28754	0	1	1	0	0	264	0.94
5621	0.461470103	22336	1	1	1	0	0	272	0.884360623
1756	0.164529643	13264	1	1	1	0	0	510	0.865413064
2200	0.542211502	4796	1	1	1	0	0	289	0.858256352

3.3.4 Train-Test Split

Before training the model, the most important phase was dividing the dataset using a "train-test split." Of the total data, thirty percent (30%) was set aside for testing and seventy percent (70%) was used to train the machine learning models. This methodical separation guarantees that the models are assessed on cases that have not yet been observed, offering a strong indicator of their capacity to extrapolate from the training set and boosting the accuracy of their forecasts.

3.4 Model Description

3.4.1 Model Overview

In this section, we present a comprehensive overview of the machine learning models employed in our study to predict the relationship scores between genes and diseases. Before delving into the details of each model, it is imperative to highlight the critical steps taken during the data preprocessing stage. These steps were designed to enhance the quality and suitability of the dataset, ensuring that the models could effectively capture meaningful patterns. The dataset underwent segmentation based on the target variable, focusing on instances where the relationship score between genes and diseases exceeded 0.25. Subsequently, label encoding was applied to transform categorical variables into numerical representations, facilitating the interpretability of the models. Furthermore, a meticulous manual representation of association types was introduced, assigning binary values (0 and 1) to signify the presence or absence of each type. With the preprocessed data in hand, the following section provides succinct descriptions of each machine-learning model employed in our study. These models were chosen for their diverse capabilities and suitability to handle the nuanced characteristics of the dataset, resulting in a robust and interpretable machine-learning pipeline. **Linear Regression** utilizes a linear relationship, incorporating the segmented and encoded data to interpret patterns effectively. **Random Forest** is an ensemble method adept at handling both numerical and encoded categorical features, including the representation of association types. **Support Vector Regression** constructs a hyperplane for effective regression, considering the segmented data and encoded association types. **K-Nearest Neighbors (KNN) Regression** leverages the proximity of data points, with improved interpretability due to label encoding and association type representation. **Decision Tree** capitalizes on hierarchical decision-making, particularly valuable for discerning patterns within the segmented and encoded dataset. **Gradient Boosting** sequentially builds trees to correct errors, with enhanced performance on the segmented and encoded data. **XGBoost** Optimized for efficiency, handles labeled categorical features and association types well. **CatBoost** Specifically designed to handle categorical features and association types, ensuring robust performance on the preprocessed dataset.

Finally, the presented machine learning models, augmented by meticulous data preprocessing, offer a promising avenue for predicting the relationship scores between genes and diseases. The segmentation of data based on the target variable, label encoding, and manual representation of association types have collectively contributed to the models' enhanced interpretability and efficacy. The diverse ensemble of models, including Linear Regression, Random Forest, Support Vector Regression (SVR), K-Nearest Neighbors (KNN) Regression, Decision Tree, Gradient Boosting, XGBoost, and CatBoost, underscores a comprehensive approach to capturing intricate patterns within the dataset. As we reflect on the outcomes and insights gleaned from this study, it is essential to acknowledge that the field of machine learning is dynamic and ever-evolving. Future research endeavors could explore the integration of additional features, refine existing methodologies, or experiment with advanced model architectures to further enhance predictive performance.

In light of the accomplishments achieved in this study, it is our hope that these findings contribute not only to the understanding of gene-disease relationships but also inspire continued exploration and innovation in the realm of bioinformatics and computational biology. With the foundation laid by this study, the journey toward unraveling the complexities of gene-disease associations continues, fueled by the intersection of cutting-edge technology and biological inquiry.

3.4.2 Input

In pursuit of our research objectives, our model is furnished with requisite gene and disease-related data. This encompasses the provision of pertinent information such as the gene ID, gene symbol, disease ID, the count of PubMed references, and the specification of five distinct association types: Altered Expression, Biomarker, Genetic Variation, Post-Translational Modification, and Therapeutic. Among these inputs, the quintet of association types assumes paramount significance as they serve as pivotal features in delineating the intricate relationships between genes and diseases. The model then takes these inputs and interprets them to determine and understand the different degrees of correlation between genes and illnesses. The subtle support that the five association types discussed above provide determines the kind and degree of this association. By means of its learning mechanism, the model skillfully classifies and deciphers the meaning of every type of association, making it possible to create strong predictive relationships that differentiate between high and low associations based on a sophisticated comprehension of the underlying biological connections. This framework of analysis highlights the significant contribution that the model may make to clarifying the links between genes and diseases.

3.4.3 Output

Upon considering all input features within the dataset, the machine learning model undergoes a training process to acquire the capability to predict a numerical score. This training involves exposing the model to a set of labeled data, allowing it to discern patterns and relationships within the input features. Following the training phase, the model is then poised to predict scores when presented with new, unseen data during the testing phase. The model receives the input features of the test dataset at the start of the testing phase and produces predictions for each associated instance. Following that, these anticipated scores are regarded as the model's output, signifying its approximation of the numerical result connected to the specified collection of input features. In order to evaluate the model's accuracy and generalizability, its prediction performance is frequently evaluated using a variety of evaluation measures. The model's ability to produce insightful predictions based on the patterns it has learned from the training data is embodied in this methodical process, which includes training and testing phases.

3.5 Project Management

The Gantt chart displays a detailed plan for our thesis tasks, running from November 1, 2022, to November 21, 2023. It outlines our entire research process.

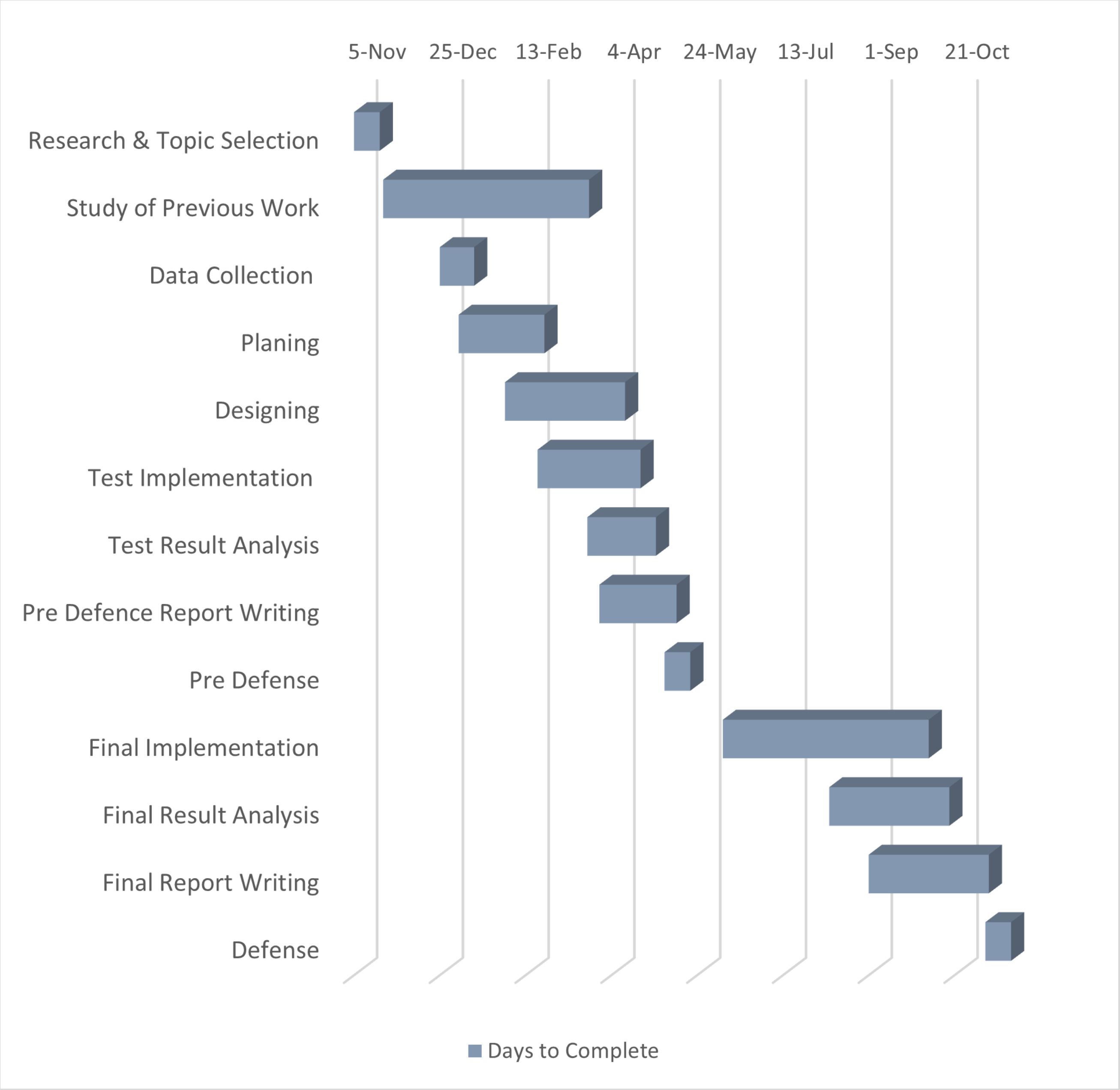

Figure 3.2: Gantt Chart.

Chapter 4

Implementation of Models and Result Analysis

4.1 Models Implementation

The research conducted to investigate the correlation between genes and diseases involved the utilization of various machine-learning algorithms on a comprehensive dataset.

The initial step in the analysis was data filtering, where emphasis was placed on a high relationship score that serves as an indicator of the gene's association with specific diseases. This step aimed to focus the investigation on instances where the genetic connection with a disease was pronounced. Following the initial filtering process, the dataset was further refined to include instances with elevated relationship scores, ensuring that the subsequent analyses would concentrate on genes exhibiting a strong connection with diseases.

To facilitate the application of machine learning algorithms, a crucial aspect of the methodology involved the systematic conversion of all string data into numerical form. This transformation was applied consistently across the dataset, encompassing non-numerical representations such as geneSymbol. The conversion process was integral to enhancing the compatibility of the data for machine learning analyses, as many algorithms operate more effectively on numerical data. The goal of the study hinged upon several key variables, including GeneID, Gene Symbol, Disease Name, Altered Expression, Biomarker, Genetic Variation, Post Translational Modification, Therapeutic, and the Number of Pubmeds. These variables were identified as crucial elements that could provide valuable insights into the intricate interplay between genes and diseases. The comprehensive analysis of these variables was undertaken to derive meaningful and nuanced insights. GeneID and Gene Symbol served as identifiers, while Disease Names pinpointed the specific diseases under consideration. Altered Expression, Biomarkers, Genetic Variation, and Post Translational Modification represented different aspects of gene behavior and characteristics that could be influential in disease manifestation. The inclusion of Therapeutic and the Number of Pubmeds provided additional dimensions, considering the potential therapeutic implications and the extent of scientific literature support for the gene-disease associations.

In order to break down the complex relationships between genes and diseases and ultimately derive important insights for the scientific understanding of these intricate interplays, the study used a methodical and multifaceted approach that integrated machine learning algorithms, data filtering based on relationship scores, and comprehensive variable analysis.

Here is the proposed workflow of our study.

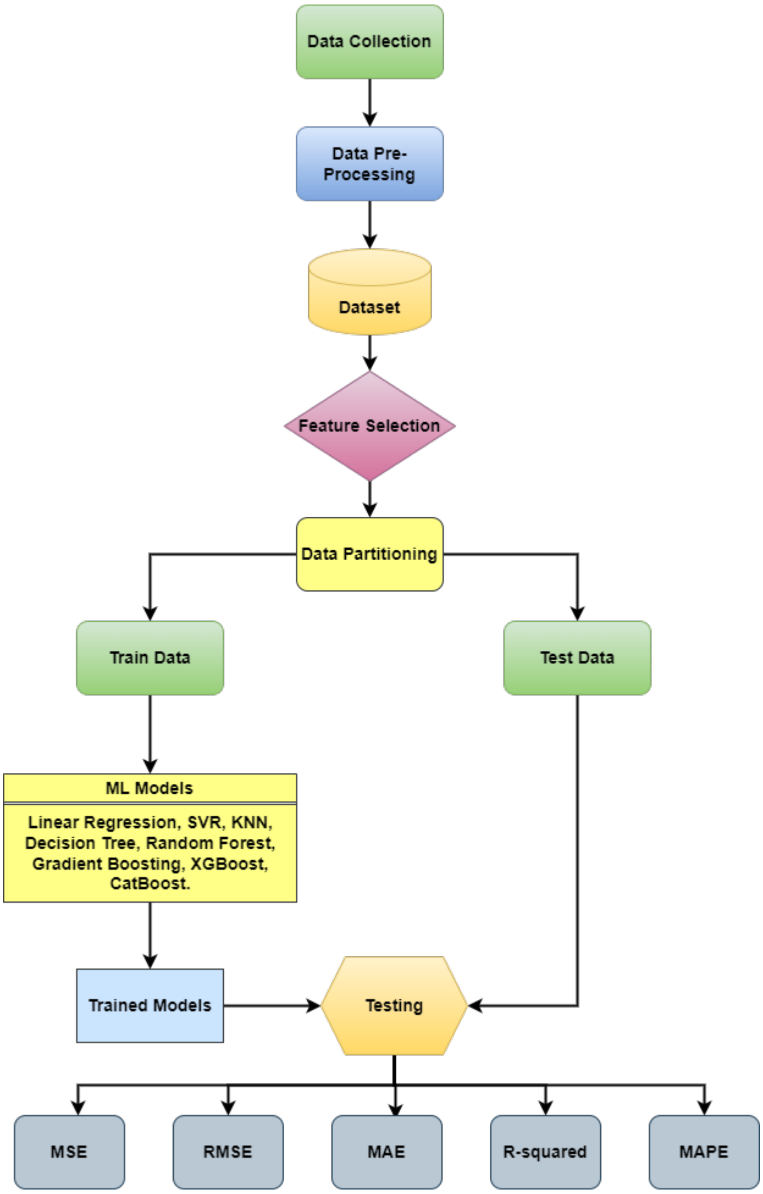

Figure 4.1: Process-Flow Diagram of the Presented Work.

4.2 Results

Table 4.1: Performance Metrics of Regression Models with score ≥ 0.25.

Model	MSE	RMSE	MAE	R-squared	MAPE
Linear Regression	0.0088	0.094	0.156	0.373	19.3
Decision Tree	0.008	0.094	0.0895	0.376	10.5
Random Forest	0.003	0.07	0.031	0.81	9.31
SVR	0.01	0.1	0.08	0.3	35.76
KNN	0.006	0.081	0.135	0.537	10.71
Gradient Boosting	0.003	0.07	0.036	0.74	11.1
XGBoost	0.005	0.0721	0.0498	0.633	10.24
CatBoost	0.005	0.0722	0.0474	0.63	10.51

The Random Forest Regressor and XGBoost models exhibit the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), signifying superior overall predictive accuracy by minimizing errors. Additionally, both models demonstrate elevated R-squared values, with Random Forest Regressor achieving 0.81 and XGBoost attaining 0.633. These values indicate a robust fit to the data and the ability to explain a substantial portion of the variance in the target variable. Notably, Random Forest and XGBoost also present relatively low Mean Absolute Percentage Error (MAPE) values of 9.31 and 10.24, respectively, underscoring favorable performance by representing error as a percentage relative to actual values.

The Decision Tree Regression model performs commendably, showcasing a comparatively low RMSE and a respectable R-squared value. Linear Regression, Gradient Boosting Regressor, and CatBoost also yield satisfactory results, with decent R-squared values indicative of their capacity to capture the variability in the data. In contrast, Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) models manifest higher errors, reflected in elevated MSE, RMSE, and MAPE values, coupled with lower R-squared values. This implies that SVR and KNN may not be as well-suited to the specific requirements of the regression task at hand.

Table 4.2: Performance Metrics of Regression Models with score ≥ 0.30.

Model	MSE	RMSE	MAE	R-squared	MAPE
Linear Regression	0.014	0.118	0.095	0.30	20.10
Decision Tree	0.021	0.144	0.092	0.033	18.84
Random Forest	0.012	0.107	0.076	0.429	15.60
Gradient Boosting	0.011	0.107	0.082	0.432	17.08
SVR	0.015	0.122	0.1	0.267	21.47
KNN	0.014	0.12	0.085	0.291	17.48
XGBoost	0.011	0.107	0.079	0.437	16.43
CatBoost	0.011	0.106	0.080	0.443	16.70

On a dataset with more than or equivalent to a 0.30 association score, the regression models were assessed, and the performance measures provide intriguing new information. Simpler models like Linear Regression and Decision Tree are routinely outperformed by ensemble techniques like Random Forest, Gradient Boosting, XGBoost, and CatBoost. The ensemble models' R-squared values show a superior fit to the data with the lowest RMSE, MAP, and MAPE of all of them.

Table 4.3: Performance Metrics of Regression Models with score ≥ 0.35.

Model	MSE	RMSE	MAE	R-squared	MAPE
Linear Regression	0.013	0.115	0.095	0.16	18.01
Decision Tree	0.021	0.147	0.1	0.368	18.32
Random Forest	0.012	0.109	0.083	0.247	15.58
Gradient Boosting	0.012	0.109	0.088	0.251	16.74
SVR	0.015	0.122	0.105	0.059	19.20
KNN	0.015	0.121	0.092	0.070	17.26
XGBoost	0.012	0.110	0.087	0.223	16.63
CatBoost	0.012	0.110	0.088	0.226	16.77

Ensemble techniques Random Forest, Gradient Boosting, XGBoost, and CatBoost consistently outperform Linear Regression with smaller prediction errors (MSE, RMSE, MAP) when evaluating regression models on a dataset containing the relationship score above or equal to 0.35. The performance of Decision Tree Regression is adequate. Mixed results are obtained using Support Vector Regression (SVR) and K-Nearest Neighbours (KNN).

Table 4.4: Performance Metrics of Regression Models with score ≥ 0.40.

Model	MSE	RMSE	MAE	R-squared	MAPE
Linear Regression	0.011	0.107	0.088	0.229	16.98
Decision Tree	0.019	0.139	0.094	0.31	17.47
Random Forest	0.011	0.105	0.08	0.247	15.24
Gradient Boosting	0.011	0.103	0.083	0.282	16.02
SVR	0.013	0.115	0.103	0.104	19.56
KNN	0.013	0.113	0.087	0.132	16.63
XGBoost	0.011	0.106	0.084	0.244	16.05
CatBoost	0.011	0.106	0.084	0.241	16.20

Regarding the dataset with an association score of at least 0.40 With smaller prediction errors and greater R-squared values, Random Forest, Gradient Boosting, XGBoost, and CatBoost routinely outperform Linear Regression and Decision Tree Regression. With the lowest MSE and RMSE, Random Forest shines out. K-Nearest Neighbours (KNN) and Support Vector Regression (SVR) perform moderately well.

4.3 Result Analysis

The computational table for the dataset having more or equal **0.25(Table 4.1)** relationship score, displays the R-squared value of the linear regression, which is 0.3735 and roughly equal to 37.35%. R-squared is a measure of the regression line's degree of data fit.The R-squared of 0.373 indicates that 37% of the data points are within a given distance of the regression line. The Decision Tree Regression value is 0.376, or about 37%. It is similar to how linear regression works. Nonetheless, we note an improvement in both Random Forest Regression and Gradient Boosting Regression. R-squared values from Gradient Boosting and Random Forest contribute to 81% and 74% of the output, respectively, which is excellent enough. For the two models, the regression line seems to fit the data pretty well.As we can also see, the results for XGBoost and CatBoost are both 63%, while the KNN R-squared result is 53%, which is below identical levels.

For additional performance measures, The average of the squared differences between the actual and anticipated values is what Mean Squared Error (MSE), RMSE MAE, and MAPE measure. In this case, the performance of linear regression and decision tree regression is comparable. The mean squared error is less for both the Random Forest and Gradient Boosting regressors. Moreover, the MSE value of the Catboost and XGBoost models is less, at roughly 0.005. SVR displays 0.010, but the KNN displays 0.006.

The Mean Absolute Percentage Error, or MAPE, is a statistical measure of the average percentage difference between expected and actual data. The SVR model exhibits the highest mean absolute percentage error among all the models, measuring at 35.76. Random Forest Regressor and Gradient Boosting Regressor have lower MAPE values. KNN and linear regression also display moderate MAPE values. Additionally, two models were applied to our dataset; the results of XGBoost and CatBoost are 10.24 and 10.51, respectively.

A more comprehensible way to measure error is to use the square root of the mean square error or RMSE. The Random Forest and Gradient Boosting regressions with the lowest RMSE values exhibit the best fits. Less accurate fits are indicated by SVR and KNN's significantly higher RMSE values. XGBoost and CatBoost, however, provide much better results. Both Linear Regression and Decision Tree Regression show intermediate RMSE values.

The Mean Absolute Error (MAE) is utilized in the computation of the average absolute difference. Similar to RMSE are the algorithms with the lowest MAE values and somewhat better fits: Random Forest and Gradient Boosting. SVR, XgBoost, CatBoost, and KNN have somewhat elevated MAE values, signifying diminished accuracy, whilst Decision Tree and Linear Regression display moderate levels.

In summary, Random Forest Regressor and Gradient Boosting are the top-performing models in terms of predictive accuracy, as evidenced by their low RMSE and high R-squared values.XGBoost and CatBoost also perform well which will be second in position. These models can explain a significant portion of the variance in the target variable and provide accurate predictions. Decision Tree Regression and Linear Regression also perform well.

Using data with an association score of at least **0.30(Table 4.2)**, Random Forest, Gradient Boosting, XGBoost, and CatBoost regression consistently outperform other models in terms of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. For these models, the Mean Absolute Percentage Error (MAPE) values are rather small. To be more precise, the XGBoost regressor generates an R-squared value of 0.437, whereas Random Forest displays an R-squared value of 0.429. With their respective MAPE values of 15.60 and 16.43, Random Forest and XGBoost have the lowest values among these models. Notably, K-Nearest Neighbours (KNN) and Support Vector Regression (SVR) have the lowest R-squared values and perform less well. While they don't achieve the highest level of accuracy in their predictions, Linear Regression and Decision Trees do better than SVM and KNN. As a result, Random Forest and XGBoost stand out as the best models, constantly exhibiting great data fit and low prediction errors by every criterion that is assessed.

From our experiment, we found that all models show low MSE, RMSE, and MAE when using data with an association score of **0.35(Table 4.3)** or higher. All models, however, display extremely low R-squared values in spite of the modest errors. This shows that the models' ability to predict the score was compromised by their insufficient learning from the dataset's features. The inadequate amount of data is one of the main causes of this failure.Despite having an R-squared of only 0.37%, Decision Tree Regression turned out to be the best-performing model of all. It surprised everyone by outperforming the top two models, which showed relatively low R-squared values: Random Forest Regressor R-squared: 24% and Gradient Boosting Regressor R-squared: 25%. Overall, the models struggled to demonstrate good results due to the limited size of the dataset.

Given the dataset with the highest threshold value **0.40(Table 4.4)** or above, we find that the models struggled to produce results that met our expectations because of the small amount of data. All of the models show extremely low error rates, but they are unable to achieve a respectable R-square value. Looking at the result table, we see that the gradient boosting model has an R-squared value of 0.247, or 24.7%, while the decision tree model has the greatest R-squared value at 0.31, or 31%. The gradient boosting model has a comparatively lower MAPE value when compared to the other model. In light of this, gradient boosting functions effectively overall.

In conclusion, boosting algorithms regularly show better performance in all cases. One thing that stands out in particular is how good random forests are when there is a lot of data available. On the other hand, although certain models in the scenarios presented demonstrate low error rates, they are unable to achieve acceptable R-square values, suggesting that they are limited in their capacity to represent the underlying relationships in the data. This highlights how important it is to use boosting algorithms because of their resilience as well as their ability to work well in situations when there is a lack of data or complicated relationships. The results highlight how crucial it is to take error rates and R-square values into account when assessing model performance in various scenarios.

Chapter 5

Discussion

5.1 Limitation and Future Work

A primary limitation of our study revolves around dataset imbalances. The uneven distribution of information across different scores poses a challenge, notably with an overwhelming abundance of low-score data. This imbalance has implications for both model training and the resultant study outcomes, potentially skewing the predictive accuracy. Furthermore, the dataset's limitation extends to the scarcity of comprehensive representations for gene-disease associations. The absence of a dataset capturing diverse relationships between genes and diseases restricts the breadth of our analysis and may limit the generalizability of our findings

In addressing these limitations, a critical avenue for future work involves the development of a balanced dataset. Striving for an equitable representation of different score categories would mitigate the impact of imbalance, fostering a more robust and unbiased model training process. The enhancement of predictive capabilities can be achieved through the deployment of more powerful machine learning algorithms. By exploring and implementing advanced algorithms, we can potentially improve the model's ability to discern nuanced patterns within the dataset and enhance the accuracy of gene-disease association predictions.The application of deep learning algorithms represents another promising area for future exploration. Leveraging the capabilities of deep learning models on a refined dataset can unveil intricate relationships between genes and diseases, potentially leading to more accurate and nuanced predictions. In summary, future efforts should prioritize the development of a balanced dataset, the deployment of more potent machine learning algorithms, and the exploration of deep learning techniques. These endeavors aim to overcome current limitations, ensuring the study's findings are both comprehensive and applicable to a broader spectrum of gene-disease associations.

5.2 Conclusion

The exploration of gene-disease associations stands as a pivotal domain in research, holding immense significance for unraveling the underlying causes of diseases and pioneering innovative approaches in treatment, diagnosis, therapy, and prevention. A profound comprehension of these relationships is instrumental, potentially leading to life-saving interventions. Harnessing the speed and accuracy of machine learning, our study aspires to contribute to this crucial field. Building upon the findings of prior research, as expounded in the literature review, the demonstrated efficacy of machine learning in detecting gene-disease associations has laid a solid foundation for our endeavors. In alignment with established guidelines and methodologies from these studies, we have employed a diverse set of machine learning algorithms. In conclusion, we are optimistic that our machine-learning models will serve as valuable tools in uncovering novel gene-disease correlations. By leveraging the computational prowess of these algorithms and assimilating vast genomic data, our study strives to contribute to the expanding landscape of knowledge concerning the genetic underpinnings of diseases. The insights gained from this research not only advance our understanding but also set the stage for further exploration in this critical area of study. As we traverse the intersection of computer science and genomics, the outcomes of this study offer promise for continued advancements, fostering a deeper comprehension of the intricate relationships between genes and diseases.

References

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

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

[3] "Griffiths, a. (2019). mutation | definition, causes, types, facts. in: Encyclopædia britannica." https://www.britannica.com/science/mutation-genetics. Accessed on November 12, 2023.

[4] "Mutation. wikipedia." https://simple.wikipedia.org/wiki/Mutation. Accessed on May 15, 2023.

[5] "Homologous recombination in eukaryotes, bacteria and viruses. byjus." https://byjus.com/biology/homologous-recombination/. Accessed on May 15, 2023.

[6] "Medlineplus (2021). epigenetics: Medlineplus genetics." https://medlineplus.gov/genetics/understanding/howgeneswork/epigenome/. Accessed on November 12, 2023.

[7] "Gene environment interaction." https://www.genome.gov/genetics-glossary/Gene-Environment-Interaction. Accessed on November 12, 2023.

[8] "Polygenic inheritance." https://byjus.com/neet/polygenic-inheritance/. Accessed on November 12, 2023.

[9] "Advani, v. (2020). what is machine learning? how machine learning works and future of it?greatlearning." https://www.mygreatlearning.com/blog/what-is-machine-learning/. Accessed on May 15, 2023.

[10] "What is machine learning? definition, types, applications, and trends for 2022. spiceworks.." https://www.spiceworks.com/tech/

artificial-intelligence/articles/what-is-ml/. Accessed on May 15, 2023.

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

[12] "Machine learning with python tutorial." https://www.tutorialspoint.com/machine_learning_with_python/. Accessed on May 12, 2023.

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

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

[15] "Xgboost algorithm." https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboo Accessed on November 12, 2023.

[16] "Catboost | built in." https://builtin.com/machine-learning/catboost. Accessed on November 12, 2023.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] S. Mukherjee, J. D. Cogan, J. H. Newman, J. A. Phillips, R. Hamid, J. Meiler, and J. A. Capra, "Identifying digenic disease genes via machine learning in the undiagnosed diseases network," *The American Journal of Human Genetics*, vol. 108, no. 10, pp. 1946–1963, 2021.

[23] R. K. Barman, A. Mukhopadhyay, U. Maulik, and S. Das, "Identification of infectious disease-associated host genes using machine learning techniques," *BMC bioinformatics*, vol. 20, pp. 1–12, 2019.

[24] K. Opap and N. Mulder, "Recent advances in predicting gene–disease associations," *F1000Research*, vol. 6, 2017.

[25] X. Wang, Y. Gong, J. Yi, and W. Zhang, "Predicting gene-disease associations from the heterogeneous network using graph embedding," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 504–511, 2019.

[26] "A database of gene-disease associations." DisGeNET, 2010, https://www.disgenet.org/. Accessed on November 10, 2023.

Appendices

Appendix A

Codes for Data Filtering

Listing A.1: Selecting the data having sore more then or equal 0.25.

```
original_dataset_path = '/content/drive/MyDrive/Thesis/Dataset
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣/gene_disease_associations.csv'

df = pd.read_csv(original_dataset_path)

# Filter the dataset based on the condition (score > 0.2)
filtered_df = df[df['score'] > 0.25]

# Convert DataFrame to CSV format
filtered_csv = filtered_df.to_csv(index=False)

# Specify the filename for the filtered dataset
filtered_dataset_filename = 'filtered_dataset.csv'

with open(filtered_dataset_filename, 'w') as file:
    file.write(filtered_csv)

files.download(filtered_dataset_filename)
```

Appendix B

Codes for Data Conversion and Normalization

Listing B.1: Converting the non numerical data and normalize the numerical data.

```
label_encoder = LabelEncoder()

df[’geneSymbol_encoded’] = label_encoder.fit_transform(df[’geneSymbol’])
print(df)

numerical_features = [’NumberOfPubmeds’, ’geneId’, ’diseaseId’,
                      ’geneSymbol_encoded’]
scaler = MinMaxScaler()

# Fit the scaler to the numerical features and transform them
X_scaled = scaler.fit_transform(df[numerical_features])

shuffled_data = df.sample(frac=1, random_state=42)

# Reset the index of the shuffled dataset
shuffled_data.reset_index(drop=True, inplace=True)

# Display the shuffled dataset
print(shuffled_data.head())
```

Appendix C

Codes for Random Forest

Listing C.1: Random Forest.

```
model = RandomForestRegressor()

# Train the model on the training data
model.fit(X_train, y_train)

# Make predictions on the testing data
y_pred = model.predict(X_test)

# Make predictions on the testing data
y_pred = model.predict(X_test)
```

Appendix D

Codes for XGBoost

Listing D.1: XGBoosting.

```
import xgboost as xgb
from sklearn.metrics import mean_squared_error,mean_absolute_error,r2_score

# Create an XGBoost regression model
model = xgb.XGBRegressor(learning_rate=0.1, n_estimators=100,
                         max_depth=6)

# Train the model on the training data
model.fit(X_train, y_train)

# Make predictions on the testing data
y_pred = model.predict(X_test)
```

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This thesis was generated on Sunday 26th November, 2023 at 5:30pm.