A Study On Benchmark Datasets For Human Disease Genes

A thesis
Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Mahin Hossain	**190204061**
MD Shihabul Islam Shovo	**190204075**
MD Fardin Jaman Aranyak	**190204093**
Md. Symum Hossain	**190204105**

Supervised by

Dr. S.M.A. Al-Mamun

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

Dhaka, Bangladesh

April 2024

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 Dr. S.M.A. Al-Mamun, Professor, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work is spread over two final year courses, CSE4100: Project and Thesis I and CSE4250: 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 project nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Mahin Hossain
190204061

MD Shihabul Islam Shovo
190204075

MD Fardin Jaman Aranyak
190204093

Md. Symum Hossain
190204105

CERTIFICATION

This project titled, **"A Study On Benchmark Datasets For Human Disease Genes"**, submitted by the group as mentioned below has been accepted as satisfactory in fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in April 2024.

Group Members:

Mahin Hossain	**190204061**
MD Shihabul Islam Shovo	**190204075**
MD Fardin Jaman Aranyak	**190204093**
Md. Symum Hossain	**190204105**

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

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

A Journey of Gratitude

In this labyrinth of efforts, We stand on the shoulders of giants, grateful for the unwavering support, wisdom, and inspiration that guided us through the journey. At the forefront, our heartfelt appreciation extends to our guiding light, Dr. S.M.A. Al-Mamun, whose mentorship has been the compass navigating the uncharted waters of research. His profound expertise and unwavering encouragement have shaped the very essence of this work. A symphony of gratitude resonates for Dr. Md. Shahriar Mahbub, Head of the Department of Computer Science and Engineering at Ahsanullah University of Science and Technology. His leadership and encouragement have been instrumental in amplifying the impact of this endeavor. To our family, the bedrock of my strength, your unwavering support and understanding have been the wind beneath my wings. Your sacrifices and encouragement are etched into the very fabric of this journey. A nod of gratitude to the academic community, the silent architects of knowledge, and the tireless library staff who provided the scaffolding for my intellectual exploration. In the grand tapestry of journey, each thread is a note of thanks to those who, directly or indirectly, contributed to the symphony of this thesis.

Dhaka
April 2024

Mahin Hossain

MD Shihabul Islam Shovo

MD Fardin Jaman Aranyak

Md. Symum Hossain

ABSTRACT

Human genetic diseases are one of the leading causes of death worldwide. The main reasons could be inheritance, changes in environmental conditions, or mutations in certain genes that cause genetic diseases. These genes are not negligible; on the contrary, a wide range of genes are involved in the development and progression of diseases. The information about these genes is crucial for advancing genetic research in the biomedical domain. In this report, we are going to explore the association of genes and study them into different association classes and relevant datasets. We emphasize the significance of using benchmark datasets that link genes to human diseases. Access to these datasets is vital for better understanding the causes of various illnesses and how they relate to other genetic factors. This report begins with a general introduction to various accessible genetic datasets and literature resources, such as DisGeNet, OMIM, CTD, GWAS, UniProt, and HuGE Literature Finder. This is followed by a brief overview of the uses, implications, and importance of these databases. This report also discusses genomic data extraction tools, such as PubTator, BeFree, BioBERT, and MetaMap. Moreover, it explores the performance of machine learning models, including Decision Trees, Random Forests, SVMs, and KNN, when applied to available datasets containing genetic data. Our research aims to play a crucial role for guiding and advancing further research in the field of gene-disease associations.

Keywords: Benchmarking, Gene-Disease Association, Datasets, OMIM, DisGeNET, CTD, GWAS, BeFree, MetaMap.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

This report aims to study benchmark datasets for human disease genes, exploiting the link between the human genome and diseases. By analyzing vast genetic data from various sources e.g DisGeNET, HuGE Navigator the research seeks to help in developing new treatments and fostering improved patient care. The standardized datasets will serve as a foundation for comprehensive investigations in this vast research field and promote collaboration among researchers. The collection of the datasets will provide significant insights to the researchers. Ultimately, this work holds promise for advancing disease management and personalized medicine, benefiting society as a whole.

1.2 Problem Statement

The problem of finding benchmark datasets for human disease genes revolves around the scarcity of readily available and standardized datasets. Existing resources such as OMIM and DisGeNet contain datasets that often lack standardization, hindering their immediate use for benchmarking purposes. The credibility of datasets claiming to be benchmarked needs verification to ensure their reliability. The fragmented distribution of benchmark datasets across various platforms further complicates the process of locating relevant information. Addressing this problem requires the development and evaluation of benchmark datasets to provide researchers with reliable and accessible resources for their studies.

1.3 Objective

Our thesis aims to study benchmark datasets for Human disease genes which involves analyzing large and complex datasets including genetic and clinical data. The objective is to explore various methods and tools used to create and evaluate the benchmark datasets in the state of art literature.

1.4 Document Summary

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

Chapter 2: This chapter includes a background study on the problem domain and discussion on relevant papers of studies previously conducted.

Chapter 3: This chapter includes discussion of the benchmarking process.

Chapter 4: This chapter includes the discussion of the experimentation on an available dataset with machine learning models.

Chapter 5: This chapter provides discussion about future work and conclusion.

Chapter 2

Background & Literature Review

2.1 Project Management

The Gantt chart displays a detailed plan for our thesis tasks, running from Jan 18, 2023, to April 3, 2024. It outlines our entire research process.

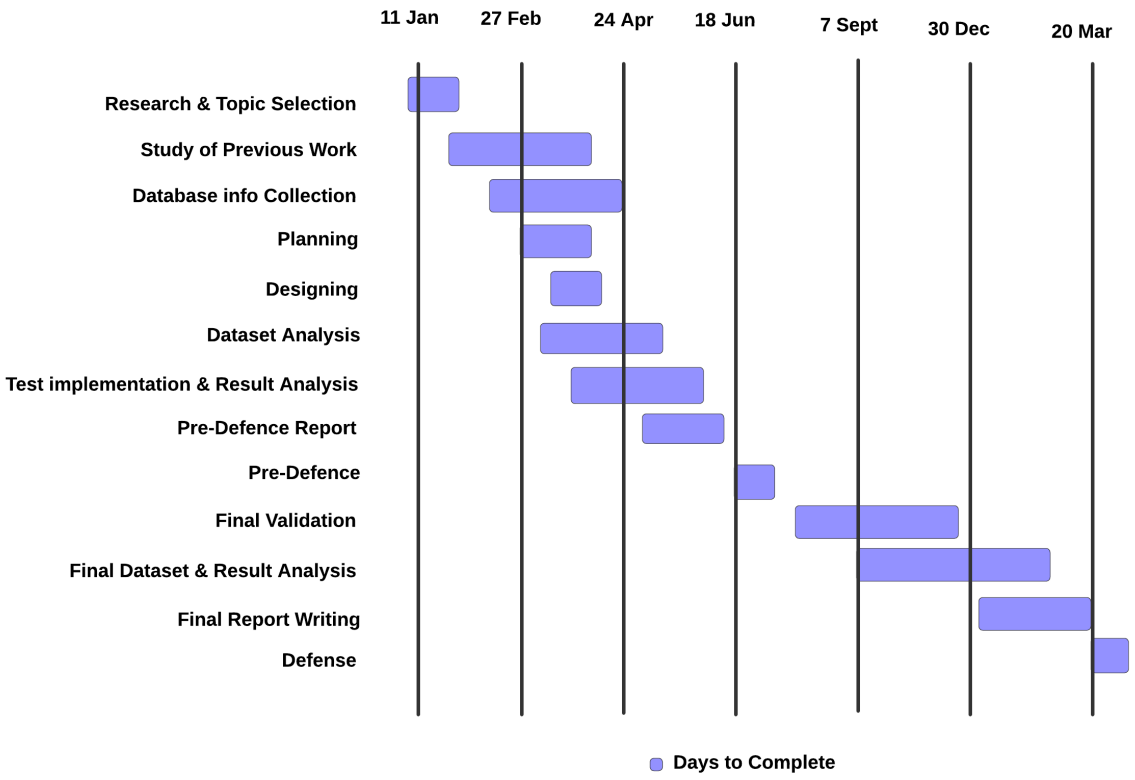

Figure 2.1: Gantt Chart.

2.2 Genomics

2.2.1 Gene

A gene is a basic unit of heredity that carries the instructions for making a specific protein. A gene is made up of DNA, which is a molecule that contains the genetic code. The genetic code is a set of instructions that tells cells how to make proteins. Proteins are the building blocks of cells and they carry out many important functions in the body. Genes are passed down from parents to offspring, and they determine many of our physical traits.

2.2.2 Gene Disease Association

Gene-disease associations [7] serve as foundational pillars upon which advancements in clinical diagnostic methodologies can be built. By identifying the specific genetic signatures associated with particular diseases, researchers can develop targeted diagnostic tests that enable early detection, accurate diagnosis, and risk stratification. This information empowers clinicians to tailor treatment plans to individual patients, optimizing therapeutic efficacy and minimizing adverse effects.

2.3 Human Gene-Disease Datasets

2.3.1 Raw Dataset

The raw dataset would encompass the unrefined, original data gathered for the study. This data collection process might entail extracting information on gene-disease associations, genetic variations, and other pertinent attributes that are crucial for understanding the genetic underpinnings of human diseases.

Raw datasets often have missing information, either because of mistakes when collecting the data, or because the information was not available or not recorded completely. If this missing information is not fixed, it can make the analyses less complete and might lead to unfair results. Raw datasets might also have mistakes, like wrong numbers or different ways of collecting the data. These mistakes can make the analyses less accurate and make it harder to get good results.

Genetic data are scattered across different domains and databases (i.e. OMIM [8], CTD [9], DisGeNET [10], ClinVar [11], Uniprot [3], Orphanet [12], GAD [13], GWAS Catalog [14], dbGaP [15]).The identification and prioritization of the relevant information from

that scattered and vast quantity of data is often a challenging task for the end user. Another difficulty that arises is that the data is only available as free text in scientific publications and this data is not compatible for computational analysis. Working with these types of data is not suitable and a slow process.

2.3.2 DisGeNet

The DisGeNET which was first released on September, 2010 is a database which integrates information of human gene-disease associations (GDAs) and variant-disease associations (VDAs) from various repositories including Mendelian, complex and environmental diseases.DisGeNET integrates data from expert curated repositories, GWAS catalogues, animal models and the scientific literature. These data are complemented with information extracted from the scientific literature using NLP-based text-mining tools. A distinctive feature of DisGeNET is its unique collection of GDAs and VDAs extracted by text mining the scientific literature. [1]

DisGeNET seamlessly integrates meticulously curated databases with text-mined data, encompassing a comprehensive spectrum of information on both Mendelian and complex diseases. Accessible through multiple platforms, including a web interface, a Cytoscape plugin, and as a Semantic Web resource, *scripts in several programming languages and an R package,* DisGeNET stands as an open-access repository, representing one of the most extensive and comprehensive collections of associations between human genes and diseases.Here, the data were combined from curated datasets called “OMIM, UNIPROT, PHARMGKB, CTD, CURATED”. As a result, DisGeNET is a coherent tool for easy analysis and interpretation of human gene–disease networks.

In DisGeNet we can see the continuous update, As DisGeNET 4.0 (June, 2016) [1] contained 429 036 gene-disease associations (GDAs), linking 17 381 genes to 15 093 diseases which gradually improved in 08 January 2020. In that time, the database(v6.0) [16] contained 628 685 gene-disease associations (GDAs), involving 17 549 genes and 24 166 diseases, and 210 498 variant-disease associations (VDAs), including 117 337 variants and 10 358 diseases. The latest updates of DisGeNET is shown below in the table:

Category	Clinical Concepts	Associated Genes	Associated Variants
Disease	21838	20163	139004
Disease Group	962	15474	22477
Phenotype	7493	16854	62686

Table 2.1: The current version of DisGeNET (v7.0) [1]

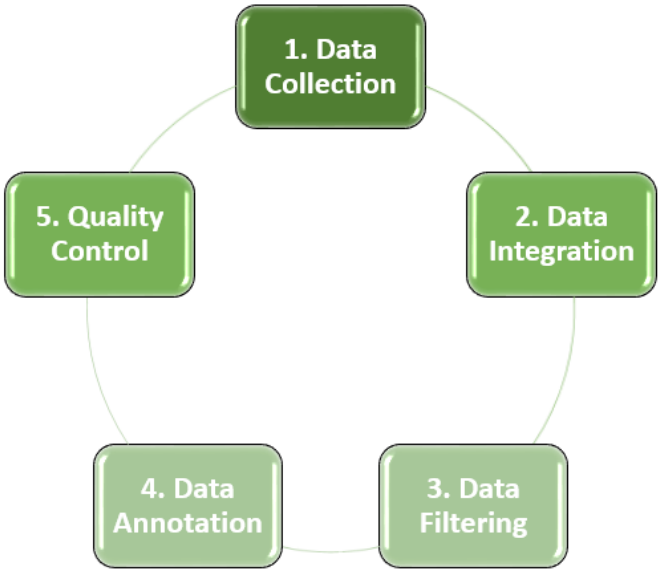

Figure 2.2: DisGeNET Flow Diagram [1]

2.3.3 OMIM

An Online Catalog of Human Genes and Genetic Disorders,OMIM was first introduced in the 60s. Dr. McKusick, the initiator of the Online Mendelian Inheritance in Man (OMIM) [2] database, started collecting information about genes and their association to diseases first as a book and later as a database. OMIM has become a highly popular source in medical genetics.OMIM entries have a structured free-text format between genes and genetic phenotypes that provides complex relationships in an efficient manner. [17]

OMIM is a well-established database with a long history of expert curation, making it a reliable source for accurate and comprehensive information on Mendelian disorders. It also provides detailed literature curation for each gene-disease association, offering in-depth supporting evidence. OMIM is recommended for research focused on Mendelian disorders and requiring in-depth literature curation.

OMIM can be searched from its homepage or from any page [18]. Information in OMIM can be retrieved by queries on MIM number, disorder, gene name and/or symbol, or plain English. OMIM assigns unique identifiers (e.g. "#146300" = "HYPOPHOSPHATASIA, ADULT") to each phenotype and gene in a vast catalogue containing 6,583 disease outcome (phenotype) descriptions and 16,953 gene descriptions as of February 27, 2023 (Entry Statistics - OMIM, 2023) [19]. An OMIM entry includes the primary title and symbol, alternative titles and symbols, and 'included' titles. The limits function may be used to perform a restricted search of parts of a MIM entry (number, titles, references, etc.) and/or type of MIM entry (gene or phenotype).Each OMIM entry is assigned a unique six-digit number whose first

digit indicates whether its inheritance is autosomal, X-linked, Y-linked or mitochondrial: 1, autosomal loci or phenotypes (entries created before May 15, 1994); 2, autosomal loci or phenotypes (entries created before May 15, 1994); 3, X-linked loci or phenotypes; 4, Y-linked loci or phenotypes; 5, mitochondrial loci or phenotypes; and 6, autosomal loci or phenotypes (created after May 15, 1994).References within an OMIM entry are linked to the complete citation at the end of the entry. The PubMed ID at the end of the OMIM reference is linked to the PubMed abstract and in some instances to the full text of the article if the journal is online [20].

Selection criteria for GDAs in OMIM OMIM has standards in-place to establish a GDA. It uses specific prefixes (e.g. #, *, +, %) and symbols (e.g. brackets—[], braces—, question mark—?) and genotype-phenotype mapping key (denoted as integer number—1, 2, 3, 4) to denote the certainty and status of its entries.

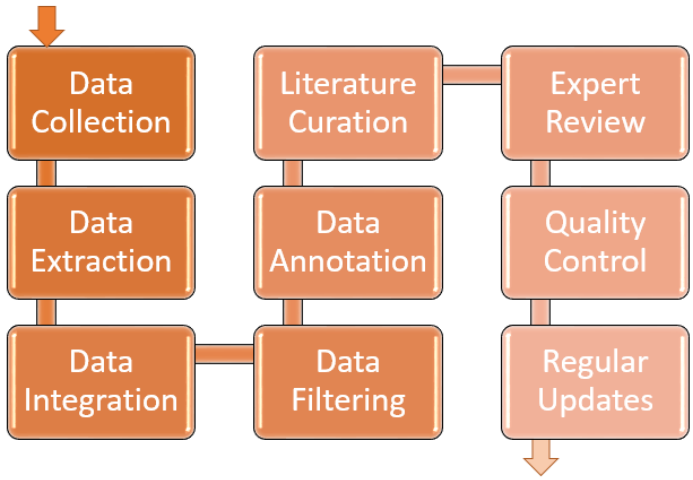

Figure 2.3: OMIM Work Process [2]

Rahit et al. [21] developed a natural language processing-based application(GPAD) to extract the gene-disease association discovery information from OMIM. GPAD utilizes these markers to initially filter out associations that are yet not confirmed. After filtration, they identified 5236 confirmed GDAs.

Jiang et al. [22] developed a medical genetics–based approach to identify potentially new indications for over 1,000 FDA-approved drugs. Integrating the data from 3 public databases for GDI data and 4 public databases for GDA data. For GDA data the OMIM, HuGE Navigator, PharmGKB, and CTD, dataabse were used. The OMIM contained 4,132 GDA pairs connecting 2,716 disease genes in 3,294 Mendelian diseases or disorders (December 2012). For each drug-disease–association pair, counted the number of genes that were associated with a given disease and/or bound by a specific drug. Duplicated pairs from different data sources were deleted. In total, 177,397 GDA pairs were obtained connecting 2,746 genes

with 2,298 unique disease terms, which were further used to build a global GDA network. Consequently, combining the 17,490 drug-gene pairs with 177,397 GDA pairs to identify a set of genes that were targeted by a given drug and associated with a specific disease using a statistical framework. The area under the receiver operating characteristic (ROC) curve (AUC) was calculated, and showed that the AUC was 0.747.

2.3.4 UniProt

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the Species (Proteomes) and the UniProt Archive (UniParc). [3]

UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). Across the three institutes more than 100 people are involved through different tasks such as database curation, software development and support. EMBL-EBI and SIB together used to produce Swiss-Prot and TrEMBL, while PIR produced the Protein Sequence Database (PIR-PSD). These two data sets coexisted with different protein sequence coverage and annotation priorities. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up. Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families. In 2002 the three institutes decided to pool their resources and expertise and formed the UniProt consortium.
UniProtKB 204,052 results (Humans)

COL NAMES:

- Entry [Unique and stable entry identifier.]
- Entry Name [Mnemonic identifier of a UniProtKB entry]
- Protein Names [Name(s) and synonym(s) of the protein]
- Gene Names [Name(s) of the gene(s) encoding the protein]
- Organism [Scientific name (and synonyms) of the source organism]
- Length [Length of the canonical sequence]

UniRef 1,889,105 results (Humans)

COL NAMES:

- Cluster ID [Unique and stable entry identifier.]
- Cluster Name [Protein name of the representative UniRef cluster member]
- Types
- Size [Number of cluster member(s)]
- Organisms [Scientific name (and synonyms) of the source organism]
- Length [Length Of the representative sequence]
- Identity [Identity threshold to the seed sequence (100%, 90% or 50%)]

UniParc 4,218,015 results (Humans)

COL NAMES:

- Entry [Unique and stable entry identifier]
- Organisms [Scientific name (and synonyms) of the source organism]
- Length [Sequence length]
- UniProtKB [UniProtKB entries describing this protein]
- First Seen [Date when source database entry was associated with this sequence for the first time]
- Last Seen [Date when source database entry was last confirmed to be associated with this sequence]

Proteomes 12,449 results (Humans)

COL NAMES:

- Entry [Unique proteome identifier]
- Organism [Scientific name (and synonyms) of the source organism]
- Organism ID [NCBI taxonomy identifier of the source organism (TaxId)]
- Protein Count [Number of protein entries associated with this proteome: UniProtKB entries for regular proteomes or UniParc entries for redundant proteomes]
- BUSCO [The Benchmarking Universal Single-Copy Ortholog (BUSCO) assessment tool is used, for eukaryotic and bacterial proteomes, to provide quantitative measures of UniProt proteome data completeness in terms of expected gene content.]

- CPD: Complete Proteome Detector is an algorithm which employs statistical evaluation of the completeness and quality of proteomes in UniProt, by looking at the sizes of taxonomically close proteomes. Possible values are 'Standard', 'Close to standard (high value)', 'Close to standard (low value)', 'Outlier (high value)', 'Outlier (low value)' or 'Unknown'. [3]

UniProt, short for Universal Protein Resource, is a comprehensive resource for protein sequence and functional information. Here's how UniProt is used in human gene-disease association studies and its significance:

- Database of Protein Information: UniProt provides a vast collection of protein sequences, structures, functions, and annotations. Researchers can access detailed information about proteins encoded by human genes, including their biological functions, pathways, and interactions.
- Gene Annotation: UniProt annotates genes with information about protein function, domains, post-translational modifications, subcellular localization, and protein-protein interactions. This rich annotation helps researchers understand the biological roles of genes and their potential involvement in disease processes.
- Disease-Associated Proteins: UniProt catalogs proteins known to be associated with human diseases. Researchers can explore curated information about disease-associated proteins, including their roles in disease pathogenesis, genetic variants linked to diseases, and functional consequences of mutations.
- Gene-Disease Association Analysis: UniProt data can be leveraged in gene-disease association studies to investigate the relationship between genetic variants, protein alterations, and disease phenotypes. Researchers can analyze protein sequences, structural features, and functional annotations to prioritize candidate genes for disease susceptibility or pathogenesis.
- Functional Analysis: UniProt facilitates functional analysis of genes and proteins implicated in human diseases. Researchers can explore protein functions, biological pathways, and molecular interactions to gain insights into disease mechanisms, drug targets, and therapeutic interventions.
- Integration with Other Databases: UniProt integrates data from various sources, including genomic databases, protein interaction databases, and disease databases. This integration allows researchers to cross-reference information and validate gene-disease associations using complementary datasets.

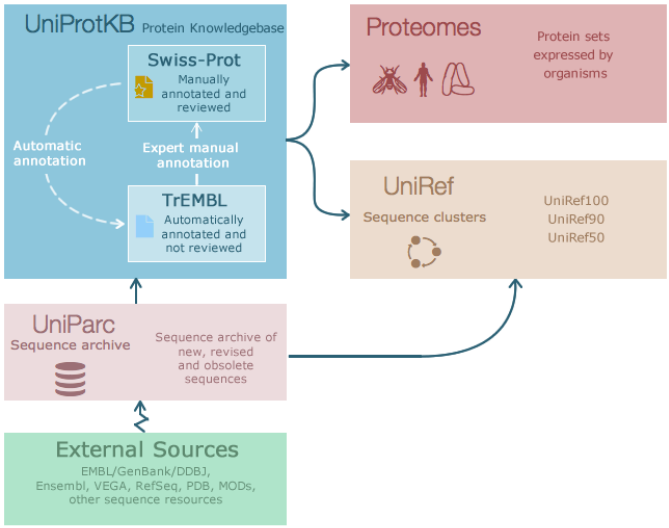

Figure 2.4: UniPort Work Flow [3]

Overall, UniProt plays a crucial role in human gene-disease association studies by providing comprehensive protein information, facilitating functional analysis, and aiding in the interpretation of genomic data. Its significance lies in its ability to empower researchers with the knowledge and resources needed to elucidate the molecular basis of human diseases and develop novel therapeutic strategies.

2.3.5 Orphanet

Orphanet is a unique resource, gathering and improving knowledge on rare diseases so as to improve the diagnosis, care and treatment of patients with rare diseases. Orphanet aims to provide high-quality information on rare diseases, and ensure equal access to knowledge for all stakeholders. Orphanet also maintains the Orphanet rare disease nomenclature (ORPHAcode), essential in improving the visibility of rare diseases in health and research information systems. [12]

Orphanet was established in France by the INSERM (French National Institute for Health and Medical Research) in 1997. This initiative became a European endeavor from 2000, supported by grants from the European Commission: Orphanet has gradually grown to a Consortium of 40 countries, within Europe and across the globe. [12]

Orphanet Report Series

Orphanet produces a series of highly-downloaded reports showcasing aggregated data covering topics relevant to all rare diseases. This series includes a list of rare diseases, reports on epidemiological data, list of orphan drugs, rare disease registries in Europe, list of research infrastructures useful to rare diseases in Europe, Orphanet's annual activity report,

and Orphanet's satisfaction surveys, as well as the list of experts having contributed to data in Orphanet.

Orphanet in numbers

- 6313 Diseases
- 4489 Genes
- 8392 Expert centers
- 41416 Diagnostic tests
- 29340 Professionals
- 2.8 M Pages viewed monthly [12]

Orphanet works towards meeting three main goals:

Improve the visibility of rare diseases in the fields of healthcare and research by maintaining the Orphanet rare disease nomenclature (ORPHAcode): providing a common language to understand each other across the rare disease field. In a global community, we need to understand each other, although we may not speak the same language. A stable nomenclature, cross-referenced with other international terminologies is therefore essential. In order to improve the visibility of rare diseases in information systems, Orphanet has developed, and maintains, a unique, multi-lingual nomenclature of rare diseases, around which the rest of our relational database is structured. Each disease is assigned a unique ORPHAcode: integrating this nomenclature in health and research information systems is essential in ensuring that rare diseases are visible. This nomenclature is aligned with other terminologies: OMIM, ICD, SNOMED-CT, MedDRA, UMLS, MeSH, GARD. This cross-referencing is a key step towards the interoperability of databases.

Provide high-quality information on rare diseases and expertise, ensuring equal access to knowledge for all stakeholders: orientating users and actors in the field in the mass of information online. Rare diseases patients are scattered across the globe, as are rare disease experts. Orphanet provides visibility to experts and for patients by providing access to a directory of expert services by disease, such as centres of expertise, laboratories and diagnostic tests, patient organisations, research projects and clinical trials. This data promotes networking, tackles isolation and helps foster appropriate referrals. Orphanet draws on the expertise of professionals from across the world to provide scientific data on rare diseases (gene-disease relationship, epidemiology, phenotypic features, functional consequences of the disease, etc.). In addition, Orphanet produces an encyclopaedia of rare diseases, progressively translated into the 7 languages of the database (English, French, Spanish, Italian,

German, Dutch, Portuguese) with texts also currently available in Polish, Greek, Slovak, Finnish and Russian, freely available online. Orphanet integrates and provides access to quality information produced around the world, such as clinical practice guidelines and information geared to the general public.

Contribute to generating knowledge on rare diseases: piecing together the parts of the puzzle to better understand rare diseases. To develop and curate the scientific data in the Orphanet database, Orphanet works with experts from around the globe, from health care professionals and researchers, to patient representatives and professionals from the medical -social sector. The wealth of data in Orphanet and the way this data is structured allows additional knowledge to be generated, helping to piece together data that at times can resemble pieces of an irresolvable puzzle. Integration of this data adds value and renders it interpretable. Orphanet provides standards for rare disease identification, notably via the Orphanet nomenclature, an essential key for interoperability. Orphanet provides integrated, re-usable data essential for research on the Orphadata platform and as a structured vocabulary for rare diseases, the Orphanet Ontology of Rare Diseases (ORDO). These resources contribute to improving the interoperability of data on rare diseases across the globe and across the fields of health care and research. They are being integrated in several bioinformatics projects and infrastructures around the world in order to improve diagnosis and treatment. Orphanet is committed to networking with partners across the globe order to help piece together the parts of this puzzle. [12]

Orphanet offers a range of freely accessible services:

- An inventory of rare diseases mapped with resources as OMIM, ICD10, MeSH, MedDRA, GARD and UMLS and a classification of diseases elaborated using existing published expert classifications. Diseases are also annotated with phenotypic features and frequency using HPO.
- An encyclopaedia of rare diseases in English, progressively translated into the other languages of the website.
- An inventory of orphan drugs at all stages of development.
- A directory of expert resouces, providing information on expert clinics, medical laboratories, ongoing research projects, clinical trials, registries, networks, technological platforms and patient organisations, in the field of rare diseases, in each of the countries in Orphanet's network.
- An assistance-to-diagnosis tool allowing users to search by signs and symptoms.

- An encyclopaedia of recommendations and guidelines for emergency medical care and anaesthesia.
- A fortnightly newsletter, OrphaNews which gives an overview of scientific and political current affairs in the field of rare diseases and orphan drugs, in English, French and Italian.
- A collection of thematic reports, the Orphanet Reports Series,focusing on overarching themes, directly downloadable from the website.
- A platform, Orphadata, providing high-quality datasets related to rare diseases and Orphan Drugs in a reusable and computable format.

The Orphanet Rare Disease Ontology (ORDO), a structured vocabulary for rare diseases derived from the Orphanet database, capturing relationships between diseases, genes and other relevant features. ORDO provides integrated, re-usable data for computational analysis. [12]

Orphanet is a unique resource dedicated to rare diseases and orphan drugs. Here's how Orphanet is used in human gene-disease association studies and its significance:

Comprehensive Rare Disease Information: Orphanet provides comprehensive information on rare diseases, including their epidemiology, clinical features, genetic causes, diagnostic criteria, and available treatments. Researchers can access detailed data on thousands of rare diseases and their associated genes.

Genetic Basis of Rare Diseases: Orphanet catalogs genetic information related to rare diseases, including genes known to be associated with specific disorders, inheritance patterns, and genetic variants implicated in disease pathogenesis. Researchers can explore curated data on gene-disease associations and genetic testing resources.

Phenotype-Genotype Correlation: Orphanet facilitates phenotype-genotype correlation studies by linking clinical features of rare diseases with underlying genetic abnormalities. Researchers can investigate the relationship between genetic variations, disease manifestations, and patient outcomes to better understand disease mechanisms and inform clinical management.

Diagnostic and Therapeutic Insights: Orphanet provides diagnostic guidelines, genetic testing recommendations, and therapeutic options for rare diseases. Researchers can utilize this information to improve diagnostic accuracy, develop targeted therapies, and optimize patient care for individuals affected by rare genetic disorders.

Rare Disease Research: Orphanet serves as a valuable resource for rare disease research by aggregating data from scientific literature, clinical databases, and expert contributions. Re-

searchers can access curated information on disease prevalence, natural history, genotype-phenotype correlations, and therapeutic interventions to advance rare disease research and drug development efforts.

Patient and Healthcare Provider Education: Orphanet offers educational resources for patients, families, and healthcare providers to raise awareness about rare diseases, facilitate early diagnosis, and improve access to specialized care. Its significance lies in its role as a central hub for rare disease information, fostering collaboration among stakeholders and promoting research and innovation in the field of rare diseases.

Overall, Orphanet plays a critical role in human gene-disease association studies by providing comprehensive rare disease information, supporting genotype-phenotype correlations, and facilitating diagnostic and therapeutic advances for individuals affected by rare genetic disorders. Its significance extends beyond research to encompass patient care, public health initiatives, and advocacy efforts aimed at addressing the unmet needs of the rare disease community.

2.3.6 GeneCards: The Human Gene Database

GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. The knowledgebase automatically integrates gene-centric data from 150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.

164,869 results (for humans)

COL NAMES:

- Symbol
- Description
- Category
- UniPort ID
- GIFtS
- GC id
- Score

GeneCards is a searchable, integrative database that provides comprehensive information about human genes. [23]

Here's how GeneCards is used in human gene-disease association studies and its significance:

- **Gene Information Aggregation:** GeneCards aggregates information from various sources, including genomic, transcriptomic, proteomic, and functional data, to create detailed gene profiles. Each gene profile contains annotations, summaries, and links to external resources, providing a comprehensive overview of gene characteristics and functions.
- **Gene-Disease Associations:** GeneCards catalogs known associations between genes and diseases, including genetic disorders, complex diseases, and susceptibility traits. Researchers can explore curated data on gene-disease relationships, genetic variants, and disease phenotypes to investigate the molecular basis of human diseases.
- **Variant Annotation and Analysis:** GeneCards provides annotations for genetic variants associated with human diseases, including single nucleotide polymorphisms (SNPs), insertions, deletions, and copy number variations (CNVs). Researchers can analyze variant data within the context of gene-disease associations, functional consequences, and clinical relevance.
- **Functional Annotation and Pathway Analysis:** GeneCards offers functional annotations for genes, including protein domains, pathways, biological processes, and molecular interactions. Researchers can explore gene functions and pathways implicated in disease pathogenesis, drug targets, and therapeutic interventions.
- I**ntegration with External Databases:** GeneCards integrates data from various external databases, including OMIM, ClinVar, GWAS Catalog, and PubMed. This integration allows researchers to cross-reference gene-disease associations, validate findings, and access additional information relevant to their research interests.
- **Data Visualization and Mining Tools:** GeneCards provides data visualization tools and mining capabilities to facilitate exploratory analysis and hypothesis generation. Researchers can visualize gene networks, expression patterns, and functional relationships to gain insights into gene-disease associations and biological mechanisms.
- **Significance:** GeneCards is a valuable resource for human gene-disease association studies due to its comprehensive coverage of gene information, curated annotations, and integration with external databases. Its significance lies in its ability to empower researchers with the knowledge and tools needed to investigate the genetic basis of human diseases, identify candidate disease genes, and elucidate disease mechanisms. By providing a centralized platform for accessing gene-related information, GeneCards

facilitates interdisciplinary research, fosters collaboration, and accelerates discoveries in the field of genomics and biomedicine.

GeneCards is a comprehensive database and search engine that provides detailed information about human genes. Developed by the Weizmann Institute of Science in Israel, GeneCards serves as a valuable resource for researchers, clinicians, and students interested in exploring the vast landscape of human genetics. Here are some key features and components of GeneCards:

- **Gene Summaries:** GeneCards contains summaries for over 40,000 human genes, offering concise descriptions of gene function, expression patterns, protein products, and associated diseases.
- **Multiple Data Sources:** GeneCards aggregates data from a wide range of sources, including scientific literature, genomic databases, protein resources, and bioinformatics tools. This integrative approach ensures comprehensive coverage and up-to-date information for each gene entry.
- **Gene-Disease Associations:** GeneCards catalogs known associations between genes and diseases, providing curated data on genetic disorders, complex diseases, susceptibility traits, and pharmacogenomics. Each gene entry includes information about disease associations, phenotypic traits, and clinical relevance.
- **Variant Annotations:** GeneCards annotates genetic variants associated with human diseases, including single nucleotide polymorphisms (SNPs), insertions, deletions, and copy number variations (CNVs). Researchers can explore variant data within the context of gene-disease associations, functional consequences, and clinical significance.
- **Functional Annotations:** GeneCards offers functional annotations for genes, including protein domains, pathways, biological processes, and molecular interactions. Researchers can investigate gene functions, pathway involvement, and protein interactions to gain insights into disease mechanisms and drug targets.
- **Integration with External Databases:** GeneCards integrates data from various external databases, such as OMIM, ClinVar, GWAS Catalog, PubMed, and UniProt. This integration allows users to cross-reference gene information, validate findings, and access additional resources relevant to their research interests.
- **Data Visualization Tools:** GeneCards provides data visualization tools and interactive features to facilitate exploratory analysis and hypothesis generation. Users can visualize gene networks, expression patterns, and functional relationships to uncover new insights into gene-disease associations and biological processes.

- **User-Friendly Interface:** GeneCards features a user-friendly interface with intuitive search capabilities, customizable filters, and interactive visualizations. Users can easily navigate through gene entries, browse related information, and access external resources with just a few clicks.

Overall, GeneCards serves as a valuable platform for exploring human gene information, understanding gene-disease associations, and uncovering new insights into the genetic basis of human health and disease. Its comprehensive coverage, integrative approach, and user-friendly interface make it an indispensable tool for genomics research, clinical genetics, and biomedical education.

2.3.7 dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans. [15]

The Database of Genotypes and Phenotypes (dbGaP) is a public repository maintained by the National Center for Biotechnology Information (NCBI) that houses genetic and phenotypic data collected from research studies involving human participants. Here's how dbGaP is used in human gene-disease association studies and its significance:

- **Data Sharing:** dbGaP facilitates the sharing of large-scale genomic and phenotypic data generated from studies such as genome-wide association studies (GWAS), sequencing projects, and other genetic research initiatives. Researchers can deposit their data in dbGaP to make it accessible to the broader scientific community for secondary analysis and exploration.

- **Genetic Data Access:** dbGaP provides access to a wide range of genetic data, including genotyping arrays, next-generation sequencing data, and imputed genotype data. Researchers can query and download genetic datasets to investigate associations between genetic variants and human diseases.

- **Phenotypic Data Access:** In addition to genetic data, dbGaP hosts phenotypic information collected from study participants, including clinical characteristics, disease diagnoses, demographic information, and environmental exposures. This rich phenotypic data enables researchers to perform phenotype-genotype correlation analyses and identify genetic factors contributing to disease susceptibility, severity, or progression.

- **Genome-Wide Association Studies (GWAS):** dbGaP hosts a large collection of GWAS datasets, which have been instrumental in identifying genetic variants associated with various human diseases and traits. Researchers can access GWAS summary statistics, individual-level genotype data, and associated phenotype data to conduct gene-disease association analyses and replication studies.

- **Complex Disease Research:** dbGaP supports research on complex diseases by providing access to datasets related to common disorders, rare diseases, and multifactorial traits. Researchers can explore genetic contributions to disease risk, explore gene-environment interactions, and investigate underlying biological mechanisms using dbGaP data.

- **Privacy and Data Security:** dbGaP implements stringent data access policies and safeguards to protect the privacy and confidentiality of study participants. Access to sensitive genetic and phenotypic data is restricted to authorized researchers who agree to adhere to ethical and legal guidelines for data usage and confidentiality.

- **Collaborative Research:** dbGaP fosters collaboration and data sharing among researchers from diverse disciplines and institutions. By providing a centralized platform for accessing and analyzing large-scale genetic and phenotypic datasets, dbGaP promotes interdisciplinary research efforts aimed at advancing our understanding of human genetics and disease.

Overall, dbGaP plays a crucial role in human gene-disease association studies by providing access to high-quality genetic and phenotypic data, facilitating collaborative research, and accelerating discoveries in the field of genomics and personalized medicine. Its significance lies in its contribution to advancing our knowledge of the genetic basis of human diseases and informing precision medicine approaches for disease prevention, diagnosis, and treatment.

The Database of Genotypes and Phenotypes (dbGaP) is a publicly accessible repository maintained by the National Center for Biotechnology Information (NCBI), part of the United States National Institutes of Health (NIH). It serves as a central resource for storing and sharing genomic and phenotypic data collected from human research studies. Here are some key aspects of dbGaP:

- **Data Repository:** dbGaP houses a wide variety of genomic and phenotypic data obtained from studies involving human participants. This includes data from genome-wide association studies (GWAS), sequencing projects, clinical trials, epidemiological studies, and other research initiatives.

- **Genetic Data:** The repository contains genetic data such as genotyping arrays, next-generation sequencing (NGS) data, imputed genotype data, and genetic variant annotations. Researchers can access individual-level genotype data or summary statistics from GWAS studies to investigate genetic associations with diseases, traits, and other phenotypes.

- **Phenotypic Data:** In addition to genetic data, dbGaP hosts detailed phenotypic information about study participants, including demographic characteristics, clinical diagnoses, medical history, laboratory measurements, and environmental exposures. Phenotypic data are linked to genetic data, enabling researchers to perform genotype-phenotype association analyses and explore complex traits.

- **Study Accession:** Data deposited in dbGaP are associated with specific research studies, each of which is assigned a unique Study Accession identifier. This identifier allows researchers to easily locate and access data from specific studies of interest.

- **Data Access Policies:** Access to dbGaP data is governed by strict data access policies designed to protect the privacy and confidentiality of study participants. Researchers must submit data access requests and provide justification for accessing specific datasets. Access is granted to authorized researchers who agree to abide by ethical and legal guidelines for data usage and confidentiality.

- **Collaborative Research:** dbGaP facilitates collaborative research by enabling data sharing and secondary analysis of large-scale genomic and phenotypic datasets. Researchers from different institutions and disciplines can access and analyze data deposited in dbGaP, fostering collaboration and accelerating scientific discoveries.

- **Data Security and Privacy:** dbGaP implements robust security measures to safeguard sensitive genetic and phenotypic data. This includes encryption of data during transmission and storage, access controls, user authentication mechanisms, and auditing of data access activities to ensure compliance with privacy regulations.

- **Community Resources:** In addition to data access, dbGaP provides resources and tools for data analysis, metadata exploration, and study documentation. Researchers can access documentation, tutorials, and FAQs to aid in navigating the repository and utilizing its features effectively.

Overall, dbGaP serves as a vital resource for the research community by providing access to high-quality genomic and phenotypic data, fostering collaborative research, and facilitating discoveries in the fields of genetics, genomics, and personalized medicine. Its role in advancing our understanding of human health and disease through large-scale data sharing and analysis is of significant importance in biomedical research.

2.3.8 ClinVar

ClinVar is a freely accessible, public database developed and maintained by the National Center for Biotechnology Information (NCBI), a part of the National Institutes of Health (NIH) in the United States. It serves as a comprehensive repository of information about variations in the human genome and their relationship to observed health status. The primary objective of ClinVar is to enhance our understanding of the genetic underpinnings of human diseases and to facilitate genomic medicine by providing a centralized resource where researchers, clinicians, and the general public can access information about the relationships between genetic variants and human health. [11]

Key Features of ClinVar:

- **Variants and Their Interpretations:** ClinVar aggregates information about genetic variants and their effects on health, including pathogenic (disease-causing), likely pathogenic, uncertain significance, likely benign, and benign interpretations based on current scientific evidence.
- **Evidence-Based Submissions:** The data in ClinVar is submitted by clinical testing laboratories, research groups, and expert panels worldwide. Submissions include detailed evidence supporting the interpretation of each variant, which can encompass clinical data, functional studies, computational predictions, and literature references.
- **Standardized Reporting:** To ensure consistency and reliability, ClinVar encourages submitters to follow standard guidelines, such as those from the American College of Medical Genetics and Genomics (ACMG) for interpreting genetic variants. This standardization facilitates the comparison and understanding of genetic information across different studies and clinical reports.
- **Integration with Other Databases:** ClinVar is closely integrated with other NCBI databases and resources, such as GenBank, PubMed, and dbSNP, providing users with a rich context for understanding genetic variants, including their frequency in populations, associated phenotypes, and underlying research articles.
- **Accessibility and Collaboration:** ClinVar is designed to be accessible to a wide range of users, from researchers and healthcare professionals to patients and the general public. It supports transparency in genetic variant interpretation and encourages collaboration and data sharing among the scientific and clinical communities to refine our understanding of genetic contributions to disease.

- **Regular Updates and Refinements:**The database is regularly updated to reflect new scientific discoveries and consensus opinions on variant interpretations. This ongoing curation effort ensures that ClinVar remains a current and reliable resource for genomic information.

Importance of ClinVar:

ClinVar plays a crucial role in advancing precision medicine and genomic research. By providing a centralized source of well-curated and standardized genetic variant information, it supports the diagnosis and management of genetic disorders, facilitates the development of new therapies, and enables researchers and clinicians to make informed decisions based on the latest genetic insights. Ultimately, ClinVar is instrumental in bridging the gap between genomic research and clinical application, improving patient care and outcomes in the era of personalized medicine.

The uses of Clinvar:

To illustrate the utility and functionality of ClinVar, let's consider a specific example involving the BRCA1 gene, which has been extensively studied for its role in hereditary breast and ovarian cancer syndrome.

The BRCA1 Gene and Its Significance

The BRCA1 gene (Breast Cancer 1) is a gene located on chromosome 17, and it produces a protein that plays a crucial role in repairing damaged DNA. Inherited mutations in the BRCA1 gene can lead to a significantly increased risk of breast and ovarian cancer, among other cancers. Variants in the BRCA1 gene can be classified into different categories based on their impact on gene function and their association with cancer risk.

Utility for Different Stakeholders

Healthcare Professionals: Clinicians and genetic counselors can use the information provided by ClinVar to inform patients about their risks associated with this variant, discuss genetic testing options, and guide management strategies for individuals carrying the mutation.

Researchers: Scientists studying hereditary breast and ovarian cancer can use ClinVar to explore known genetic variants, investigate their biological impact, and identify areas where further research is needed.

Patients and the Public: Individuals who have undergone genetic testing and found to carry the c.68_69delAG mutation can use ClinVar to understand the implications for their health and to find resources for support and further information.

This example underscores ClinVar's role as a critical resource for consolidating and dis-

seminating genetic variant information, facilitating informed decisions in both clinical and research contexts, and supporting the broader goals of personalized medicine and public health.

2.3.9 CTD:

The Comparative Toxicogenomics Database (CTD) is a robust, publicly available resource designed to enhance our understanding of the impact of environmental chemicals on human health. Developed and maintained by the Mount Desert Island Biological Laboratory in Salisbury Cove, Maine, the CTD provides a comprehensive compilation of curated data that connects chemicals, genes, and diseases. By integrating information on chemical-gene interactions, chemical-disease, and gene-disease relationships, the CTD offers insights into the molecular mechanisms underlying various environmental diseases and conditions. [9]

Key Features of the CTD:

Chemical-Gene Interactions: The CTD includes detailed information about how chemicals interact with genes, affecting gene expression and leading to potential changes in biological pathways. This data helps researchers understand the molecular basis of diseases.

- Chemical-Disease Associations: It also provides information on the linkage between chemicals and diseases, highlighting potential toxic effects and exposures that may lead to health issues. This aspect is crucial for public health and epidemiology.
- Gene-Disease Connections: By documenting how specific genetic variations or mutations can influence disease susceptibility, especially in the context of environmental exposures, the CTD bridges the gap between genetics and environmental health. Inferred Data Through
- Computational Approaches: One of the unique features of the CTD is its use of advanced computational methods to predict potential relationships between chemicals, genes, and diseases based on existing curated data. These inferred relationships can guide future experimental research.
- Integrated Approach: The database integrates data from various sources, including scientific literature and other reputable databases, to provide a comprehensive resource for researchers, educators, and policymakers.
- Research and Education Tool: The CTD serves as a valuable tool for researchers in toxicology, environmental health, genetics, and related fields. It also serves as an educational resource for teaching these concepts in academic settings.

Applications of the CTD:

- Research: Scientists use the CTD to discover potential toxic effects of chemicals, understand mechanisms of action, and identify biomarkers for environmental diseases.
- Public health officials and regulatory agencies can use the data to assess risks associated with chemical exposures and to develop guidelines to protect human health.
- Development: Pharmaceutical researchers might use the CTD to identify potential adverse effects of drug candidates or to discover new therapeutic targets by understanding disease mechanisms. Education: Educators utilize the CTD as a teaching tool to illustrate the complex interactions between genes, diseases, and the environment in health and disease.

Accessibility:

The CTD is freely accessible online, providing a user-friendly interface that allows users to search for specific chemicals, genes, or diseases, as well as to explore the connections between these entities. The database is regularly updated, ensuring that it remains a relevant and authoritative resource for the latest research in toxicogenomics and environmental health.

By offering a comprehensive and integrative approach to understanding the molecular mechanisms by which environmental exposures affect human health, the CTD plays a crucial role in advancing the fields of environmental health, toxicology, and genomics.

Interactions between the chemical Bisphenol A (BPA), a common industrial chemical used in plastics, and its potential health impacts.

Example for the uses of CTD: Bisphenol A (BPA)

- Chemical: Bisphenol A (BPA): Description: BPA is a synthetic compound found in many consumer products, including water bottles, food containers, and thermal paper receipts. There is concern about its effects on health due to its estrogen-like activity.
- Searching for BPA in the CTD: When you enter “Bisphenol A” into the CTD search bar, the database would likely return a comprehensive profile that includes its chemical properties, synonyms, and external identifiers. It would also list gene interactions, related diseases, and references to scientific studies.
- Chemical-Gene Interactions: The CTD might show that BPA interacts with estrogen receptor 1 (ESR1), a gene that codes for a hormone receptor involved in various biological processes. BPA’s binding to this receptor could mimic or interfere with estrogen’s natural effects, leading to potential health issues.

- Chemical-Disease Associations: The database could reveal associations between BPA exposure and several health conditions, such as reproductive disorders, obesity, and diabetes. Each condition would be linked to scientific studies that have investigated these relationships.
- Gene-Disease Connections: For the gene ESR1, the CTD would likely list diseases associated with mutations or alterations in this gene, providing insight into how BPA's interaction with ESR1 might contribute to disease risk.
- References and Further Reading: The profile for BPA would include references to scientific papers that have studied its effects, offering pathways for researchers to explore the topic further.

Utilizing the Information:

For Researchers: This information can spur further studies to explore the mechanisms by which BPA affects gene expression and contributes to disease. Researchers might also investigate potential interventions to mitigate BPA's harmful effects.

For Public Health Officials: Understanding the connections between BPA exposure and health outcomes can inform policies and guidelines to reduce exposure, such as regulations on the use of BPA in consumer products.

For Educators and Students: The example of BPA provides a real-world case study to understand the complexities of environmental health and the importance of integrating genetic, biochemical, and epidemiological data to assess risk and inform public health decisions.

This hypothetical example illustrates how the CTD serves as a critical resource for exploring the intricate relationships between chemicals, genes, and diseases, supporting a wide range of scientific inquiries and public health initiatives.

2.3.10 GWAS Catalog

The GWAS Catalog serves as a comprehensive repository of Genome-Wide Association Studies (GWAS), facilitating the exploration of genetic associations with various traits and diseases. Developed to enhance the understanding of complex genetic factors underlying human traits and diseases, the catalog provides invaluable insights into the genomic landscape of health and disease. [24]

Curation Criteria: The GWAS Catalog follows strict inclusion and exclusion criteria to ensure the quality and relevance of the data. Included studies must adhere to primary GWAS analysis standards, employing array-based genotyping with a focus on genome-wide coverage. Additionally, studies incorporating published GWAS data or utilizing imputation tech-

niques are eligible under certain conditions. Exclusion criteria encompass limitations such as language, candidate gene focus, and the absence of new GWAS data.

Curation Process: Data extracted from literature encompasses publication details, study cohort specifics (e.g., size, recruitment country, subject ancestry), and SNP-disease association details (e.g., SNP identifier, p-value, gene, risk allele). Each study is associated with a trait reflecting the phenotype under investigation. Multiple traits analyzed in a study lead to either multiple entries or individual SNPs annotated with specific traits. Expert scientists conduct data extraction and curation, supported by a web-based tracking and data entry system. The process undergoes two levels of curation to ensure accuracy and consistency, with additional validation provided by an automated pipeline.

Curation involves three key steps: Firstly, identifying relevant studies through systematic searches, leveraging machine learning-assisted systems for recent studies, and employing query-based methods for older ones. Secondly, extracting ancestry data in both free text and structured formats, utilizing controlled terms for consistent representation and referring to detailed framework documentation for clarity. Lastly, ensuring quality control and SNP mapping through an automated pipeline, which adds specific SNP information, conducts consistency checks, and validates missing data, utilizing resources like the Ensembl API and NCBI for SNP retrieval.

Contribution in Gene Disease Association: The GWAS Catalog plays a crucial role in advancing our understanding of gene-disease associations. By systematically curating associations between genetic variants and traits/diseases, the catalog provides researchers with a valuable resource for exploring the genetic basis of complex diseases. The inclusion of statistically significant associations and the availability of detailed information on study cohorts and SNP-trait associations enhance the utility of the catalog for genetic research.

The GWAS Catalog serves as an indispensable resource for researchers investigating the genetic underpinnings of human traits and diseases. Through meticulous curation processes and stringent criteria, the catalog provides reliable and comprehensive data on GWAS findings. Its contribution to advancing our understanding of gene-disease associations underscores its significance in genomic research and personalized medicine efforts.

2.3.11 GWAS Central

GWAS Central stands as a pioneering database devoted to collating and curating data from Genome-Wide Association Studies (GWAS). Established in 2006, its primary objective is to serve as a pivotal resource for researchers delving into the genetic underpinnings of various traits and diseases [25]. Drawing from a multitude of sources including published literature, supplementary materials, direct submissions from researchers, and other GWAS

databases, GWAS Central offers a comprehensive repository of GWAS-related information. the world's most extensive openly accessible repository of summary-level GWAS association data, it encompasses over 72.5 million P-values for over 5000 studies testing over 7.4 million unique genetic markers investigating over 1700 unique phenotypes [26]. GWAS Central offers a comprehensive toolkit for seamless access and visualization of an unparalleled compilation of genome-wide association study (GWAS) data. The database amalgamates direct submissions from GWAS authors and consortia with meticulously curated datasets from diverse public sources. Users can explore GWAS data from various perspectives, including genetic markers, genes, genome regions, or phenotypes, facilitated by graphical visualizations and downloadable data reports. The integrated genome browser enables the visual interrogation of tested genetic markers and pertinent genomic features across up to sixteen association datasets simultaneously. Through the semantic standardization of phenotype descriptions utilizing Medical Subject Headings and the Human Phenotype Ontology, GWAS Central enables precise identification of genetic variants linked to diseases, phenotypes, and traits of interest. Furthermore, harmonization of phenotype descriptions across multiple GWAS-related resources enhances cross-database study discovery by leveraging a range of ontologies. [27]

Data Accessibility and Visualization: GWAS Central champions accessibility and user-friendliness by offering intuitive tools for data retrieval and visualization. Researchers can seamlessly access curated GWAS data, facilitating meta-analyses, replication studies, and in-depth investigations into the genetic architecture of complex traits and diseases. Through graphical visualizations and detailed downloadable data reports, users can explore genetic markers, genes, genome regions, and phenotypes of interest.

Semantic Standardization and Phenotype Search: A key strength of GWAS Central lies in its commitment to semantic standardization, achieved through the use of standardized phenotypic descriptions employing resources such as Medical Subject Headings and the Human Phenotype Ontology. This meticulous approach ensures precise identification of genetic variants associated with diseases, phenotypes, and traits. Furthermore, the harmonization of phenotype descriptions across GWAS-related resources enhances cross-database study discovery, enriching the overall research landscape.

GWAS Central remains committed to its mission of curating and updating GWAS summary-level association data and study metadata. Through ongoing collaborations with other GWAS databases and stakeholders, such as the NHGRI-EBI GWAS Catalog, GWAS Central strives to ensure the database remains a reliable and up-to-date resource for the scientific community. [25]

The insights garnered from GWAS have far-reaching implications, from predicting disease

risk and unraveling the genetic architecture of phenotypes to estimating heritability and exploring rare variants through cutting-edge sequencing technologies. By facilitating the exploration of genetic associations, GWAS Central plays a pivotal role in advancing our understanding of human genetics and personalized medicine.

A difference between OMIM and Disgenet is illustrated in the Table 2.2.

Feature	OMIM	DisGeNET
Research Focus	Mendelian disorders	Broad range of gene-disease associations
Data Quality	Expert curation	Automated and manual curation
Data Coverage	Limited to Mendelian disorders	Broader coverage, including complex and non-Mendelian disorders
Data Updates	Regular, but may lag behind DisGeNET	Very frequent updates
Genomic Data Integration	Limited	Integrated with genetic variants and expression profiles
Community Involvement	Limited	Open-source, fostering community contributions

Table 2.2: Difference between OMIM and DisGeNET

The field of genetics relies heavily on standardized datasets for accurate analysis and interpretation of genetic information. Here, we have conducted a comparative analysis of various genetic standard datasets to assess their utility, accuracy, and applicability in genetic research in Table 2.3.

Name	**Scope**	**Organism**	**Current Statistics**	**First Published**
DisGeNET [1]	Gene-disease, and variant-disease associations	Human	1134942 associations, between 21,671 genes and 30170 diseases, 46589 SNPs	2010
Comparative Toxicogenomics Database(CTD) [9]	Chemicals, genes, and disease associations	Human and animal models	1127498 associations between 20,027 genes and 1504 diseases	2003
Online Mendelian Inheritance in Man(OMIM) [2]	Mendelian diseases and their genes	Human	121512 associations between 29,596diseases and 20,790 genes	1998
Genetic Association Database(GAD) [13]	Genes, variants, and complex diseases and traits	Human	74928 associations between 12,774 diseases and 10,697 genes	2004
UniProt Knowledgebase [3]	Proteins	Human	566467 associations between 1545 diseases and 19,368 genes	2004
GWAS Catalog [24]	GWAS studies	Human	30,148 associations between 2,743 diseases and 21,449 genes (18,666 variants)	2009

Table 2.3: Comparative Study Between Different Genetic Standard Datasets

2.4 Tools Employed for Genomic Data Extraction

Creating benchmark datasets uses different sequencing technologies [28]. This sequencing technologies benefited in the development of genomic, transcriptomic, and epigenomic studies. Gene sequencing technologies focus on decoding the genetic information encoded in DNA and RNA. These sequencing technologies gather a large amount of data but the text mining tools play a crucial role in extracting valuable information from the vast amount of scientific literature and textual data associated with genomics. In recent years, the rapid progress of artificial intelligence has significantly advanced natural language processing (NLP), merging linguistic principles, computer science, and mathematics [29]. NLP facilitates human-computer communication. Text classification within NLP is a fundamental task, systematically assigning a given text to predefined categories based on specific characteristics [30]. The process involves three key steps: text preprocessing, vector representation extraction, and training a classifier for effective categorization [31]. Text classification is divided into single-label and multi label types, depending on whether each text is associated with one or more categories [32]. Single-label assigns one category per text, while multilabel allows for multiple category assignments.

In the past, most efforts in text mining of relationships have been devoted to the identification of interactions between proteins, both due to the availability of corpora and the push driven by specific text mining challenges [33]. But In recent years, many text mining tools have been developed for the purpose of classification genomics data. BeFree is a text mining tool used to recognize biomedical entities that identifies genes, diseases, and chemicals from scientific texts using rule-based and dictionary-based approaches. By exploiting morpho-syntactic information of the text, BeFree is able to identify gene-disease, drug-disease and drug-target associations with state-of-the-art performance. Named Entity Recognition (NER) is a crucial task in natural language processing (NLP) that involves identifying and classifying entities. NER operates by recognizing and categorizing specific terms or phrases that represent entities, contributing to the overall comprehension of textual data. The goal of NER is to identify, within a collection of text, all of the instances of a name for a specific type of thing: for example, all of the drug names within a collection of journal articles, or all of the gene names and symbols within a collection of MEDLINE abstracts. [34]

BERT [35] is a contextualized word representation model that is based on a masked language model and pre-trained using bidirectional transformers. BERT to classify genetic mutations based on text evidence from an annotated database [36]. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), is a domain-specific language representation model pre-trained on large-scale biomedical corpora. BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. [37]

There are other tools available in the literature i.e. PubTator [38], MetaMap [39].

These tools play a vital role in knowledge discovery, offering efficient and speedy extraction of relevant information from diverse textual sources. The integration of heterogeneous data types, including genetic and clinical information, benefits from automated extraction processes. Text mining facilitates literature reviews and summarization, enabling researchers to quickly grasp key insights and trends. Overall, these tools are indispensable for navigating and interpreting the vast landscape of genomics literature, accelerating knowledge discovery and advancing genomic research. BeFree is a text mining tool designed to identify biomedical entities such as genes, diseases, and chemicals in scientific literature. Some tools here are discussed below:

2.4.1 BeFree

A text mining system to extract relations between genes, diseases and drugs for translational research. Current biomedical research needs to leverage the large amount of information reported in publications. Text mining approaches aimed at finding relationships between entities are key for identification of actionable knowledge from free text repositories. The development of the BeFree system is designed to identify relationships between biomedical entities with a special focus on genes and their associated diseases. BeFree, by exploiting morpho-syntactic information of the text, performs competitively not only for the identification of gene-disease relationships, but also for drug-disease and drug-target associations. The application of BeFree to a real-case scenario shows its potentiality in extracting relevant information for translational research. BeFree (Biomedical Evidence Retrieval for Genetic Association Studies) is a text-mining tool designed to extract gene-disease associations from scientific literature. It aims to automate the process of retrieving evidence from biomedical articles, particularly in the context of genetic association studies. Here's how BeFree is used and its significance in human gene-disease association studies: [40]

Text Mining: BeFree employs advanced text-mining algorithms to analyze biomedical literature and identify mentions of genes, diseases, and their associations. It utilizes natural language processing (NLP) techniques to parse text, recognize relevant entities, and extract semantic relationships between genes and diseases mentioned in the literature.

Association Extraction: BeFree extracts gene-disease associations from the text by identifying sentences or passages that mention both a gene and a disease in proximity. It analyzes the context surrounding gene and disease mentions to infer potential associations, taking into account syntactic structures, semantic cues, and linguistic patterns indicative of relationships between genes and diseases.

Evidence Retrieval: BeFree retrieves evidence supporting extracted gene-disease associations from the literature, including relevant citations, article metadata, and contextual information. It helps researchers identify articles containing relevant evidence for specific gene-disease associations, facilitating literature review and evidence synthesis in genetic association studies.

Scoring and Ranking: BeFree assigns scores or ranks to extracted gene-disease associations based on various criteria, such as the frequency of occurrence in the literature, the strength of supporting evidence, and the reliability of text-mining predictions. These scores help researchers prioritize and assess the significance of identified associations, guiding further investigation and validation efforts.

Integration with Databases: BeFree integrates with existing databases and resources containing curated gene-disease associations, such as OMIM, ClinVar, and GWAS Catalog. By cross-referencing extracted associations with curated databases, BeFree provides additional validation and context for identified associations, enhancing their reliability and utility for research.

Customization and Adaptability: BeFree allows for customization and adaptation to specific research contexts and domains. Researchers can fine-tune the tool's parameters, adjust search strategies, and incorporate domain-specific knowledge to improve the accuracy and relevance of extracted gene-disease associations.

Significance in Research: BeFree plays a significant role in human gene-disease association studies by automating the extraction of evidence from biomedical literature, reducing the time and effort required for literature review, and facilitating data-driven research in genetics and genomics. Its ability to systematically analyze large volumes of literature helps researchers uncover novel gene-disease associations, validate existing hypotheses, and generate new insights into the genetic basis of human diseases. Overall, BeFree contributes to advancing our understanding of gene-disease relationships and accelerating discoveries in biomedical research.

2.4.2 BioBERT

BioBERT is a state-of-the-art pre-trained language representation model specifically designed for biomedical text processing tasks. It is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, which has been fine-tuned on large-scale biomedical text corpora. BioBERT is trained on PubMed abstracts and full-text articles, making it particularly well-suited for tasks related to biomedical research, including gene-disease association (GDA) studies. [35]

Named Entity Recognition (NER): BioBERT can perform named entity recognition to identify and extract gene and disease mentions from biomedical texts. By accurately recognizing gene and disease entities, BioBERT enables researchers to focus on relevant information for GDA studies.

Relation Extraction: BioBERT can extract relationships between gene and disease entities mentioned in biomedical texts. Using its contextual understanding of language, BioBERT can infer associations between genes and diseases based on their co-occurrence patterns and syntactic structures in the text.

Association Analysis: Researchers analyze the extracted gene-disease associations to identify patterns, trends, and potential causal relationships. They use BioBERT's semantic understanding of language to interpret the associations and prioritize those with high confidence scores or supporting evidence.

Semantic Understanding: BioBERT encodes the semantics of biomedical texts into dense vector representations, capturing nuanced relationships between genes and diseases. This semantic understanding allows BioBERT to effectively capture the context and meaning of gene-disease associations, even in complex and ambiguous text.s

Validation and Interpretation: Researchers validate the extracted associations by comparing them with existing knowledge from biomedical databases, literature reviews, and experimental studies. They interpret the associations in the context of known biological mechanisms and pathways associated with breast cancer and BRCA1.

Integration with External Data: Researchers integrate the extracted gene-disease associations with other sources of genomic and clinical data, such as gene expression profiles, patient cohorts, and genetic variant databases. This integrated analysis provides a comprehensive understanding of the genetic basis of breast cancer and the role of BRCA1 in disease pathogenesis.

Hypothesis Generation and Experimental Design: Based on the findings from the GDA study, researchers generate hypotheses about the functional significance of BRCA1 in breast cancer and potential therapeutic targets. They design experimental studies to validate these hypotheses in cell lines, animal models, or clinical trials.

Transfer Learning: BioBERT benefits from transfer learning, where knowledge learned from pre-training on large-scale biomedical corpora is fine-tuned on specific downstream tasks, such as GDA studies. This transfer learning process enhances the model's performance and adaptability to the target domain, improving the accuracy of extracted gene-disease associations.

Interpretability and Explainability: BioBERT provides insights into the underlying patterns and features learned during training, enabling researchers to interpret and explain

the extracted gene-disease associations. By understanding how BioBERT makes predictions, researchers can gain confidence in the reliability and validity of the extracted associations

Scalability and Efficiency: BioBERT's efficient architecture and pre-trained parameters enable scalable processing of large volumes of biomedical texts, making it suitable for analyzing extensive literature databases and repositories. This scalability facilitates comprehensive and systematic exploration of gene-disease associations across diverse biomedical domains.

2.4.3 PubTator

PubTator is a suite of text-mining tools developed by the National Center for Biotechnology Information (NCBI) that aims to assist researchers in extracting and annotating biological entities and their relationships from biomedical literature. Specifically, PubTator focuses on identifying gene-disease associations from scientific articles indexed in PubMed. To ensure high quality of automatically processed results, we used tools that have been extensively evaluated for superlative performance in various text-mining competition events. [38]

Automated Text Mining: PubTator employs advanced text-mining algorithms to automatically identify and extract gene-disease associations from PubMed articles. This automated process accelerates the literature review process, allowing researchers to efficiently identify relevant information without manually scanning through numerous articles.

Comprehensive Coverage: PubTator provides access to a wide range of biomedical literature indexed in PubMed, encompassing research articles, reviews, and clinical studies. This comprehensive coverage ensures that researchers have access to a diverse array of evidence when exploring gene-disease associations.

- **Entity Recognition:** PubTator employs natural language processing (NLP) techniques to automatically recognize and annotate various biological entities mentioned in biomedical texts, including genes, diseases, proteins, chemicals, and mutations. For gene-disease association studies, PubTator specifically identifies gene and disease mentions within PubMed articles.
- **Relation Extraction:** In addition to entity recognition, PubTator extracts relationships between recognized entities, such as gene-disease associations. By analyzing the text context surrounding gene and disease mentions, PubTator identifies instances where genes are mentioned in relation to specific diseases, indicating potential associations between the two.
- **Annotation Enrichment:** PubTator enriches entity annotations with additional information, such as gene identifiers, disease names, and associated metadata. This

enrichment process enhances the usability of extracted gene-disease associations by providing standardized identifiers and additional context for the recognized entities.

- **Scoring and Confidence Assessment:** PubTator assigns confidence scores to extracted gene-disease associations based on various factors, including the frequency of occurrence in the literature, the strength of supporting evidence, and the reliability of text mining predictions. These scores help researchers prioritize and assess the significance of extracted associations.

Let's consider a scenario where PubTator extracts a gene-disease association from a scientific article in PubMed. The extracted association is between the gene "BRCA1" and the disease "Breast Cancer."

- Frequency of Occurrence: The frequency of occurrence refers to how often the gene-disease association appears in the literature. If multiple articles report the same association between BRCA1 and breast cancer, it suggests a higher level of consensus and strengthens the confidence in the association. Example: If the association between BRCA1 and breast cancer is reported in 50 different articles, it indicates a high frequency of occurrence and increases the confidence in the association.

- Strength of Supporting Evidence: The strength of supporting evidence assesses the quality and reliability of the evidence supporting the gene-disease association. Strong evidence might include experimental studies, clinical trials, meta-analyses, or systematic reviews. Example: If several well-designed studies with large sample sizes consistently demonstrate an association between BRCA1 mutations and increased breast cancer risk, it provides strong supporting evidence for the association.

- Consistency Across Studies: Consistency across studies refers to the degree of agreement among different research findings regarding the gene-disease association. Associations that are consistently observed across independent studies are considered more reliable. Example: If multiple independent research groups from different countries or populations replicate the association between BRCA1 mutations and breast cancer risk, it adds to the consistency of the evidence and increases confidence in the association.

- Biological Plausibility: Biological plausibility assesses whether the gene-disease association is supported by known biological mechanisms or pathways. Associations that align with current understanding of molecular biology and disease pathogenesis are considered more biologically plausible. Example: Since BRCA1 is involved in DNA repair and maintenance of genomic stability, it is biologically plausible that mutations in BRCA1 could predispose individuals to breast cancer by impairing DNA repair mechanisms.

Based on these criteria, a scoring and confidence assessment system could assign a numerical score or confidence level to the gene-disease association between BRCA1 and breast cancer. Associations with higher scores or confidence levels indicate stronger evidence, greater consistency, and higher biological plausibility, increasing confidence in the validity of the association.

- **Visualization and Exploration:** PubTator offers visualization tools and interactive interfaces to facilitate the exploration and interpretation of extracted gene-disease associations. Researchers can visualize gene-disease networks, browse annotated articles, and explore supporting evidence from PubMed literature.
- **Integration with External Resources:** PubTator integrates with external databases and resources to enhance the utility of extracted gene-disease associations. By linking annotations to curated databases such as GeneCards, OMIM, and ClinVar, PubTator provides additional context and validation for identified associations.
- **Data Accessibility and Availability:** PubTator provides public access to its annotated datasets and text-mining tools, enabling researchers to leverage its capabilities for their own studies and analyses. The availability of PubTator data fosters collaboration, reproducibility, and innovation in gene-disease association research.

2.4.4 MetaMap

MetaMap, developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM), is a highly configurable software designed to link biomedical text with concepts in the UMLS Metathesaurus, an extensive repository of biomedical terminology. Originally developed to enhance retrieval of bibliographic content such as MEDLINE citations. MetaMap utilizes a knowledge-intensive approach grounded in symbolic, natural-language processing (NLP), and computational-linguistic techniques. Its versatility extends to information retrieval (IR) and data-mining applications, and it forms a cornerstone of NLM's Medical Text Indexer (MTI), facilitating both semi-automatic and fully automatic indexing of biomedical literature

MetaMap offers three distinct data models catering to varying levels of filtration: the Strict Model for precise semantic processing, employing comprehensive filtering; the Moderate Model, which applies manual, lexical, and type-based filtering but excludes syntactic filtration, suitable for holistic term processing; and the Relaxed Model, featuring only manual and lexical filtering, providing access to the vast majority of Metathesaurus strings and ideal for browsing purposes. These models enable users to tailor their MetaMap usage based on their specific needs, balancing accuracy, consideration of input text as a whole, and access to a broad array of Metathesaurus strings. [5]

- **Usage:** MetaMap is employed to map biomedical text, including genomic literature, to concepts in the UMLS Metathesaurus. It utilizes a knowledge-intensive approach based on natural language processing (NLP) and computational-linguistic techniques to identify and extract relevant information from textual data. Researchers use MetaMap to analyze large volumes of genomic literature, extracting associations between genetic variants (e.g., SNPs) and specific diseases or traits.
- **Significance in Human Gene-Disease Association:** MetaMap enables researchers to identify and extract gene-disease associations from diverse sources of genomic literature, including scientific publications and databases. By accurately mapping text to concepts in the UMLS Metathesaurus, MetaMap facilitates the discovery of relevant genetic variants and their associations with specific diseases or traits. The tool's ability to process large amounts of text and extract meaningful associations enhances the efficiency and effectiveness of research into human gene-disease associations. MetaMap's contribution to genomic data extraction and analysis aids in advancing our understanding of the genetic basis of diseases and traits, leading to potential insights into disease mechanisms, biomarker discovery, and therapeutic targets.

In summary, MetaMap serves as a valuable tool for genomic data extraction, playing a significant role in identifying and elucidating human gene-disease associations by efficiently processing and analyzing vast amounts of genomic literature.

2.5 Study of Existing Papers

Raj et al. [6] explored the association of breast cancer genes by classifying them into different association classes: positive, negative, and neutral. The researchers obtained raw data from the HuGE Literature Finder via Finder HuGe [41] , which provided PubMed IDs related to breast cancer. To extract abstract text data from these PubMed IDs, the researchers used EDirect, a tool developed by NCBI for data retrieval from PubMed. A total of 12,565 records were processed to eliminate anomalies, duplicate entries, and redundancy. The researchers successfully extracted 12,565 reference sentences containing at least one disease and gene term from 7,073 abstracts. The method employed for processing the raw data was double-fold cross-validation, which helped discard false predictions and generate a processed file used to calculate the weight of individual gene association classes. The researchers used benchmark gene reference data for breast cancer, which can be found here [4]. In summary, the study focused on the association of breast cancer genes, utilizing a robust methodology involving data from the HuGE Literature Finder, EDirect, and a double-fold cross-validation process. The benchmark gene reference data served as a comparison for the classification

of gene association classes.

Armada et al. [5] focused on evaluating and comparing various computational techniques for predicting or prioritizing genes associated with specific diseases. The code and datasets supporting the conclusions can be found on the GitHub repository [4]. The benchmarking process involves assessing and comparing the performance of different network propagation methods to identify the most effective one for disease gene identification. The key steps in the benchmarking process include Data Preparation, Method Selection, Cross-validation, Comparison, Statistical Analysis, and Visualization. The primary objective of the study was to identify genes that could be targeted by drugs for treating common diseases. The researchers utilized data from the OpenTargets database, which provides information about genes and diseases.The study employed 12 different computer algorithms proficient in analyzing biological networks. These algorithms were tested on data related to 22 common diseases. Two types of data were considered: genetic information related to diseases and data about how genes interact with each other and with proteins.To ensure the reliability of their results, the researchers used cross-validation, a special testing method for algorithms. The majority of the project was coded in R, with some Matlab code required to incorporate state-of-the-art approaches. In summary, the article presents a comprehensive evaluation and comparison of computational techniques for predicting disease-associated genes. The study involves the use of diverse algorithms, data types, and the OpenTargets database, and the results are supported by thorough cross-validation.

Majidian et al. [28] provides Genomic variant benchamrk (GVB) which helps to improve variant detection methods. Benchmark datasets should include information about genomic regions with variants. Creating benchmark datasets uses different sequencing technologies. These include long-reads, short-reads, and linked-reads. It is hard to get perfect accuracy and sensitivity, but the processes of sequencing, read alignment, genome assembly, and variant calling help to make better pipelines and technologies. Benchmark dataset curation uses different sequencing technologies to overcome limitations and errors. Groups like the Genome in a Bottle Consortium (GIAB) and Platinum Genome are working to make or improve benchmark datasets. The CMRGs benchmark dataset looks at challenging genes that were not fully studied in other benchmark datasets. These Clinically Meaningful Reference Genes (CMRGs) have been studied a lot and are linked to one or more diseases. Short-read technologies, like Illumina's exome sequencing, are common because they are cheap and accurate. But they do have limitations. Long-read sequencing technology may be able to find structural variants (SVs) that cause diseases that short-read sequencing misses. HiFi long-reads are being used to make a group of genes from different databases. This shows that scientists are working to fix problems and improve benchmark datasets.

Chapter 3

Benchmarking Datasets

3.1 Steps of Benchmarking

Benchmarking is an essential tool for evaluating the performance of various methods. It involves a systematic process comprising several crucial steps:

Step 1: Data Preparation

The initial phase of benchmarking, data preparation, plays a pivotal role in ensuring the integrity and consistency of the data upon which subsequent analysis will be conducted. This step encompasses meticulously gathering, cleaning, and refining the data, ensuring its compatibility for the intended analysis. Key activities in this step include:

- Duplicate Removal: Identifying and eliminating duplicate data entries to maintain data accuracy and prevent potential biases.
- Missing Value Handling: Addressing missing values appropriately, either through imputation techniques or exclusion, to ensure data completeness without introducing distortions.
- Data Format Structuring: Structuring the data into a suitable format that aligns with the chosen analysis tools and methods.

Step 2: Method Selection

The second step in benchmarking involves carefully selecting the methodologies or algorithms that will be evaluated. These methods, which may encompass machine learning models, statistical techniques, or optimization algorithms, are chosen based on their suitability for the specific dataset and problem at hand. The selection process often involves considering factors such as the type of data, the desired outcome, and the computational requirements of each method.

Step 3: Cross-validation

Cross-validation, a cornerstone of benchmarking, serves as a rigorous technique to assess the generalizability of the selected methods. This process involves dividing the dataset into multiple subsets, known as folds. The chosen methods are then trained on one subset and subsequently evaluated on the remaining subsets. The iterative nature of cross-validation allows for a comprehensive assessment of how well the methods perform across diverse data segments, ensuring their ability to generalize effectively to unseen data.

Step 4: Performance Comparison

Following cross-validation, the fourth step entails compiling and comparing performance metrics for each method. These metrics, which may include accuracy, precision, recall, F1-score, or other relevant measures, are tailored to the specific problem and the nature of the dataset. The comparison of performance metrics allows for the identification of the method that consistently outperforms its counterparts, thereby determining the optimal method for the given task.

Step 5: Statistical Analysis

The fifth step introduces statistical analysis to ascertain whether observed performance differences between methods are statistically significant or could potentially result from random chance. This analytical step ensures that conclusions drawn from the benchmarking process are grounded in robust statistical evidence, providing confidence in the selection of the optimal method.

Step 6: Visualization

The final step in benchmarking involves employing visual aids such as charts, graphs, and plots to effectively communicate the findings to a diverse audience. Visualization serves to present the results in a clear, concise, and easily understandable manner, making the benchmarking process accessible to both technical experts and laypeople alike.

In conclusion, benchmarking is an indispensable tool for evaluating the effectiveness of various methods. By adhering to a systematic approach encompassing the six crucial steps outlined above, researchers and practitioners can make informed decisions regarding the selection of optimal methods for their specific tasks.

3.2 Examples of Benchmarking Process

Following the process figure 3.1, Raj et al. [6], developed a benchamark dataset for breast cancer. This benchmark dataset for breast cancer-associated genes provides a comprehensive summary of gene-disease associations. The benchmark gene reference data for Breast

Cancer [4] accessible here. By following the process of benchmarking the breast cancer dataset, other gene-disease association datasets can also be benchmarked.

Process flow:

Figure 3.1: Dataset Benchmarking Process Flow for Breast Cancer Associated Genes [4]

How data were acquired :

Raw data was downloaded from HuGE Literature Finder [41]. Human Genome Epidemiology (HuGE) Navigator gives access to an up-to-date knowledge database that comprises information on population prevalence of genetic variants, gene-disease associations, gene-gene and gene- environment interactions, and evaluation of genetic tests. Since the HuGE literature finder returns only the PubMed IDS related to a specific disease, we further used EDirect database to extract abstract text data of these PubMed IDs. Edirect is a tool developed by NCBI used to retrieve data from Pubmed. [6]

Data format:

Raw, Processed, and Analyzed

Parameters for data collection:

Raw data collected contains disease name, disease class, gene, PubMed id, association class, title of article, year of publication and reference sentence supports the association.

Description of data collection:

Raw data collected from an archive of published genetic association studies that contains information on molecular and clinical parameters.

HuGE Literature Finder:

- The HuGE Literature Finder is a component of the HuGE Navigator, providing access to an extensive knowledge database focused on genetic epidemiology.
- It allows researchers to search for literature specifically related to human genetic

epidemiology, including studies on population prevalence of genetic variants, gene-disease associations, gene-environment interactions, and evaluation of genetic tests.

- Researchers can use the HuGE Literature Finder to identify relevant scientific literature, particularly PubMed IDs (identifiers assigned to articles in PubMed), related to specific diseases or genetic variants of interest.

EDirect Database:

- EDirect is a tool developed by the National Center for Biotechnology Information (NCBI) that facilitates the retrieval of data from PubMed, a comprehensive database of biomedical literature maintained by the NCBI.
- In the context of the study on benchmark datasets for human disease genes, researchers use EDirect to extract abstract text data from PubMed articles identified through the HuGE Literature Finder.
- By leveraging EDirect, researchers can automate the process of retrieving abstracts or full-text articles associated with specific PubMed IDs obtained from the HuGE Literature Finder.
- This allows researchers to access detailed information from relevant scientific literature, enabling them to analyze gene-disease associations, genetic variant prevalence, and other epidemiological data for inclusion in benchmark datasets or further analysis.

The false positives and false negatives from the raw data were processed, the double fold manual validation were applied for each record.

"**Double fold manual validation**" is a method used to validate data or findings by independently verifying them through two separate manual validation processes. This approach is commonly employed in situations where accuracy and reliability are crucial, such as in scientific research, data analysis, or quality assurance processes.

Here's how double fold manual validation typically works:

- First Validation Round:
  - In the initial validation round, one set of validators or reviewers independently examines the data or findings.
  - Validators carefully assess the data against predefined criteria, guidelines, or standards to determine their accuracy, completeness, and consistency.

  - They may perform manual checks, comparisons, or analyses to validate the data, ensuring that it meets the required quality standards.

- Second Validation Round:

  - In the second validation round, a separate set of validators or reviewers independently repeats the validation process.
  - Similar to the first round, validators meticulously examine the data using the same criteria, guidelines, or standards.
  - They conduct their own manual checks, comparisons, or analyses to verify the accuracy and reliability of the data, without relying on the findings of the first validation round.

- Comparison and Resolution:

  - Once both validation rounds are completed, the results from the two rounds are compared and discrepancies are identified.
  - Any inconsistencies or discrepancies between the findings of the two validation rounds are carefully reviewed and resolved.
  - Validators may engage in discussions, consultations, or additional analyses to address discrepancies and arrive at a consensus on the final validated data.

- Final Validation Outcome:

  - The validated data is considered reliable and accurate if it successfully passes both validation rounds with consistent results.
  - The final validated data is then used for further analysis, reporting, or decision-making with confidence in its quality and integrity.

The paper published by Armada et al. [5], investigates the efficacy of network propagation algorithms in identifying potential drug targets through genetic data. They systematically test 12 different algorithms based on network propagation using gene-disease data from 22 common non-cancerous diseases. The study considers two biological networks, six performance metrics, and compares two types of input gene-disease association scores. They quantify the impact of design factors on performance using explanatory models. To obtain realistic estimates, the paper introduces two novel protein complex-aware cross-validation schemes. It is emphasized that leveraging genetic association information is a promising approach for identifying drug targets. Three fundamental types of data are used: data-driven

networks, interactions curated from literature, and interactions extracted from literature using text mining approaches. Network propagation approaches are highlighted as a key family of algorithms for extracting useful information from biological networks. These methods have been applied in various contexts including disease gene identification. The study aims to systematically evaluate the usefulness of network propagation methods for prioritizing novel drug targets, using various networks and validation schemes to reflect realistic drug development scenarios.

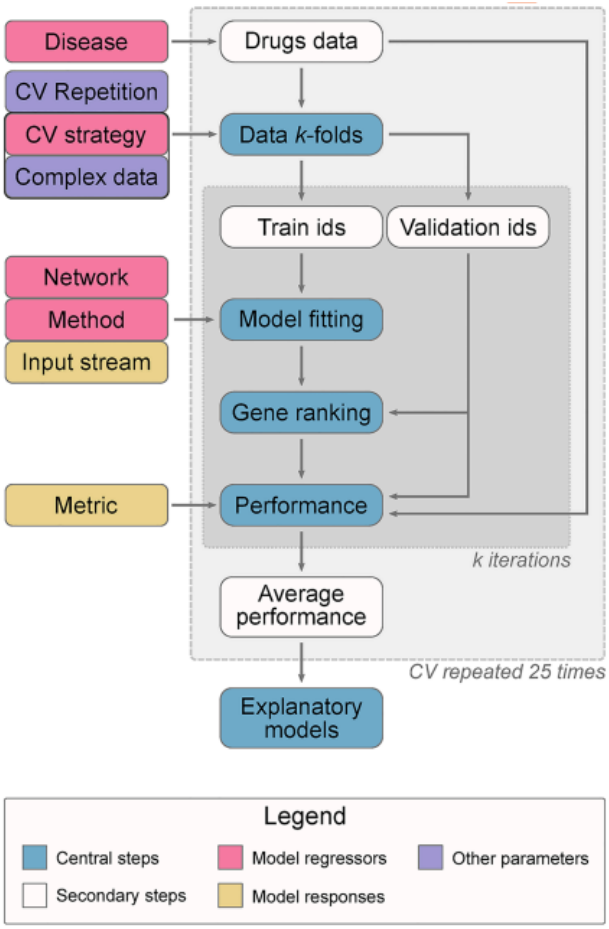

Figure 3.2: Method for Disease Gene Identification [5]

The benchmarking process for network propagation methods for disease gene identification involves comparing various algorithms to determine the most effective method. Key steps include:

- **Disease and Drugs Data:** Information on the diseases and drugs under study.
- **CV Repetition:** Cross-validation repetition process to assess model generalizability.
- **CV Strategy and Data k-folds:** Splitting data into three folds for 3-fold cross-validation.
- **Complex Data:** Data used in the complex validation scheme, one of three CV strategies.

- **Network:** Biological network representing gene or protein relationships.
- **Method:** Network propagation method being compared (15 methods in this study).
- **Model Fitting:** Training the network propagation method on the data.
- **Input Stream:** Two types of input data streams used: genetic association and drug-based genes.
- **Gene Ranking:** Ranking genes by their likelihood of association with the disease.
- **Metric:** Performance metric used to evaluate methods (six metrics used).
- **Performance:** Method's performance on a specific metric.
- **Explanatory Models:** Models used to explain performance variation across methods, diseases, networks, and CV strategies.

Overall, this rigorous benchmarking process helps identify the most effective network propagation methods for disease gene identification, with a focus on practical applications in drug discovery and development.

Chapter 4

Experimentation on a Dataset with Machine Learning Models

4.1 Work Flowchart

Figure 4.1: Benchmarking Work-Flow

- Data preparation ensures data integrity through activities like removing duplicates,

handling missing values, and structuring data.

- Carefully select appropriate methodologies for the data and problem, considering data type, outcomes, and computational needs.
- Cross-validation is a way to check how well a model performs on new data.This is done by dividing the data into multiple subsets. The model is then trained on one subset and evaluated on another. This process is repeated multiple times, each time using a different subset for training and evaluation.
- Compare different methods using performance metrics like accuracy and precision to find the best approach for the task.
- Use statistics to make sure the differences in method performance are real and not just random.
- Show the results of the benchmarking using charts and graphs so everyone can understand them.

4.2 Comparative Analysis of Machine Learning Models for Gene Disease Association

4.2.1 Breast Cancer Associated Gene Dataset

Here are some sample breast cancer dataset entities shown below:

S.NO	DB_ID	DIS_CLASS	GENE	PUBMED.ID	LACKASSO	TITLE
1	bc_id1	CANCER	MYCL1	1345822	Y	Association
2	bc_id2	CANCER	HRAS	2086347	Y	Analysis
3	bc_id3	CANCER	ERBB2	3664511	Y	Association
4	bc_id4	CANCER	TP53	7524772	B	Quantitative
5	bc_id5	CANCER	RB1	7615356	Y	Association

Table 4.1: Breast Cancer Dataset Entities - Part 1

YEAR	CONCLUSION	REF_SENTENCE	CLASS	REF_GENE	GENE_NEW	WEIGHT
1992		No differences	N	L-myc	MYCL	1
1990		No absolute	N	Ha-ras	HRAS	1
		Amplification	Y	c-erbB-2	ERBB2	0.285714286
1994		Breast tumors	Y	p53	TP53	0.367816092
		No differences	Y	RB	RB1	0.5

Table 4.2: Breast Cancer Dataset Entities - Part 2

Here are some sample columns shown in the one hot encoding process below:

GENE	GENE Encoded	DIS CLASS	DIS CLASS Encoded	A CLASS	A CLASS Encoded
MYCL1	543	CANCER	0	N	3
HRAS	346	CANCER	0	N	3
ERBB2	232	CANCER	0	Y	7
TP53	830	CANCER	0	Y	7
RB1	696	CANCER	0	Y	7

Table 4.3: One Hot Encoding Sample

The benchmark dataset for breast cancer was created by processing 12,565 data entries [4]. To ensure the accuracy of the dataset, a manual validation process was conducted, aiming to identify and remove false positives and false negatives. In this context, false positives refer to instances where the raw data suggests a positive association between a gene and the disease but no actual association exists. On the other hand, false negatives occur when the raw data indicates a negative association, but there is an actual association. To address this, a double-fold manual validation process was implemented for each record. In instances where references lacked clear associations between the disease (breast cancer) and the gene, they were categorized into various subcategories outlined in the "ASSOCIATION.CLASS" field, as detailed in Table 4.4. This categorization process allowed for a more nuanced understanding of the gene-disease associations within the dataset, providing valuable insights for future research and analysis

Gene association classification describes the attributes of a Gene Disease Association dataset. The followed table describes,

Different machine learning models use various evaluation metrics to assess their performance on a given task. The choice of evaluation metric depends on the nature of the problem (classification, regression, clustering, etc.) and the specific goals of the model.

SVM (Support Vector Machine):
Accuracy: **81%** SVM is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points belonging to different classes.

RF (Random Forest):
Accuracy: **84%** Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

DT (Decision Tree):
Accuracy: **88%** A Decision Tree is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node

Serial No.	Sub-Category	Description
1	Y	Gene is positively associated with Breast cancer disease
2	N	Gene is not associated with breast cancer disease
3	A	Gene is ambiguous in nature. Word "may" or "may not" is used to define the association of genes with Breast cancer disease
4	MC	Mis Classified. Multiple entries for which PubMed ID, gene symbol, reference sentences are the same, but classification are different.
5	ANF	Abstract Not Found. For the given PubMed ID abstract is not found in the reference article.
6	NR	Not Related. Abstract is not related to Alzheimer
7	PNF	PubMed ID Not Found
8	X	Abstract does not contain any information about the association between mentioned gene in GENE field and Breast cancer

Table 4.4: Gene Association Classification [6]

represents the outcome. Decision trees are used for both classification and regression.

KNN (K-Nearest Neighbors):

Accuracy: **77%** KNN is a simple supervised learning algorithm used for classification and regression tasks. It classifies objects based on the majority vote of their nearest neighbors. It's easy to implement and understand, making it popular for beginners and as a baseline model for comparison.

These accuracy values represent the performance of each model on a specific task or dataset. Higher accuracy generally indicates better performance, but it's essential to consider other metrics and factors depending on the specific requirements of your machine learning application. The comparative bar diagram of all four method is given with it's accuracy (%),

A comprehensive comparative study analysis of the four machine learning models, namely k-Nearest Neighbors, Support Vector Machines (SVM), Decision Trees, and Random Forests, reveals their distinct performances and applicability in genomic data analysis.

k-Nearest Neighbors (k-NN) is a non-parametric, instance-based learning algorithm that classifies new cases based on similarity measures. k-NN is particularly useful for genomic

Class	Support	SVM			RF		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
A	248	0.79	0.82	0.80	0.76	0.94	0.84
ANF	10	1.00	1.00	1.00	1.00	1.00	1.00
MC	1	1.00	1.00	1.00	1.00	1.00	1.00
N	133	0.66	0.59	0.62	0.86	0.41	0.56
NR	16	1.00	1.00	1.00	1.00	1.00	1.00
PNF	1	0.00	0.00	0.00	0.00	0.00	0.00
X	248	1.00	1.00	1.00	1.00	1.00	1.00
Y	159	0.62	0.64	0.63	0.72	0.77	0.74
Class	**Support**	**DT**			**KNN**		
		Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
A	248	0.86	0.91	0.88	0.68	0.85	0.76
ANF	10	1.00	1.00	1.00	1.00	1.00	1.00
MC	1	0.50	1.00	0.67	0.50	1.00	0.67
N	133	0.75	0.71	0.73	0.66	0.53	0.59
NR	16	1.00	1.00	1.00	1.00	1.00	1.00
PNF	1	0.00	0.00	0.00	0.00	0.00	0.00
X	248	1.00	1.00	1.00	0.96	1.00	0.98
Y	159	0.81	0.78	0.79	0.65	0.47	0.54

Table 4.5: Merged Classification Reports

Method	Weighted Avg				Macro Avg		
	Acc	Prec	Rec	F1	Prec	Rec	F1
SVM	0.81	0.80	0.81	0.80	0.76	0.76	0.76
RF	0.84	0.85	0.84	0.83	0.79	0.77	0.77
DT	0.88	0.88	0.88	0.88	0.74	0.80	0.76
KNN	0.77	0.77	0.77	0.76	0.68	0.73	0.69

Table 4.6: Additional Metrics Results

data analysis when the underlying structure of the data is not well understood and when there are no clear boundaries between classes. However, k-NN can be computationally expensive, especially with large datasets, as it requires storing and searching through the entire training dataset for each prediction.

Support Vector Machines (SVM), known for their robustness to high-dimensional data and outliers, offer a balanced trade-off between accuracy and interpretability. SVMs excel in tasks requiring classification of genomic data, especially when dealing with non-linear relationships. Yet, SVMs require careful parameter tuning, and their computational complexity can become prohibitive for large-scale datasets.

Decision Trees provide an intuitive and interpretable framework for gene-disease association analysis. With their simplicity and transparency, decision trees offer valuable insights into the underlying decision-making process. Despite their susceptibility to overfitting, particularly in noisy datasets, decision trees remain a popular choice for exploratory analysis

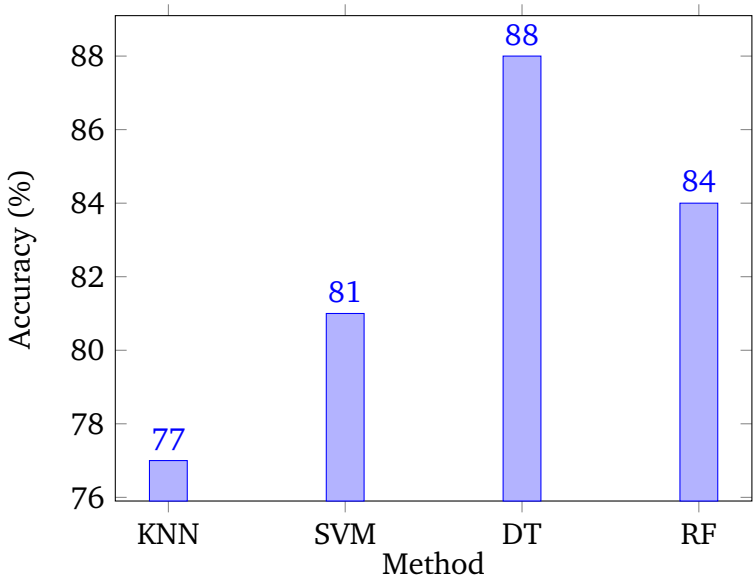

Figure 4.2: Performance of Different Methods

and initial model prototyping due to their ease of implementation and quick computation.

Random Forests, leveraging ensemble learning techniques, enhance the robustness and generalizability of decision trees by aggregating multiple models. Random Forests are well-suited for handling high-dimensional genomic data and mitigating issues like overfitting. However, their interpretability diminishes compared to individual decision trees due to the complexity introduced by ensemble methods.

In summary, the selection of a suitable machine learning model for genomic data analysis necessitates a careful consideration of factors such as dataset characteristics, computational resources, interpretability requirements, and the trade-off between predictive accuracy and model transparency. Each model presents unique advantages and limitations, and researchers must weigh these factors judiciously to choose the most appropriate model for their specific research objectives and constraints.

From the study above, it is clearly seen that the Decision Tree gives the best accuracy on the dataset of breast cancer.

Chapter 5

Discussion

5.1 Conclusion

In conclusion, our thesis has presented a comprehensive exploration of the process involved in creating a benchmark dataset for human disease genes. Through meticulous investigation, we have provided detailed explanations of various existing datasets, shedding light on their strengths, limitations, and suitability for benchmarking purposes. Additionally, we have conducted experiments using a selection of available datasets, employing machine learning models to analyze and evaluate their performance.

Our findings underscore the importance of robust benchmark datasets in advancing research on human disease genes. By meticulously documenting the characteristics of existing datasets and rigorously testing their utility through machine learning experiments, we have contributed valuable insights to the field. Furthermore, our analysis has identified areas for improvement and future directions in dataset creation and utilization.

Moving forward, the benchmark datasets studied here serves as valuable insights for researchers and practitioners in genetics, genomics, and bioinformatics. It provides a standardized framework for evaluating the efficacy of machine learning models in predicting gene-disease associations and offers a foundation for advancing our understanding of the genetic data of human diseases.

In summary, our report represents a significant contribution to the field of human gene-disease association studies. By elucidating the process of benchmark dataset creation and conducting empirical evaluations with machine learning models, we have laid the groundwork for future research endeavors aimed at unraveling the complexities of human genetics and improving disease diagnosis, prognosis, and treatment.

5.2 Future Work:

Our future research endeavors will be directed towards the advancement of a benchmark dataset by integrating multi-omics data, temporal information, and validation on larger datasets. We aim to delve deeper into several key areas of machine learning for genomic data analysis, focusing on advancements and in that can further enhance the efficacy and applicability of our research. Specifically, our future work will center on the following aspects:

- Integration of Multiple Databases: Explore the integration of multiple gene-disease benchmark databases to create a comprehensive and curated resource. This integrated database could provide a more comprehensive view of gene-disease associations, leveraging the strengths of each individual database while addressing their limitations.
- Advanced Machine Learning Models: Extend your analysis by exploring more advanced machine learning models beyond decision trees, random forests, neural networks, and SVMs. Consider models such as gradient boosting machines, and deep learning architectures, which may offer improved performance and predictive accuracy for gene-disease association prediction tasks.
- Ensemble Learning Approaches: Investigate ensemble learning techniques that combine predictions from multiple models to improve overall performance. Ensemble methods such as stacking, boosting, and bagging can harness the complementary strengths of different models and enhance predictive accuracy.
- Validation and Benchmarking: Conduct rigorous validation and benchmarking studies to evaluate the performance of different models and tools on diverse datasets. This could involve cross-validation, external validation on independent datasets, and comparison against established benchmarks to assess predictive accuracy, robustness, and generalizability.
- Interactive Visualization and Data Exploration: Design interactive visualization tools and data exploration interfaces for gene-disease benchmark databases. Incorporate features such as network visualization, interactive filtering, and data summarization to empower researchers with intuitive and informative ways to explore and analyze complex biomedical data.

By pursuing these avenues for further research, you can advance the field of gene-disease association analysis, enhance the utility of benchmark databases and tools, and contribute to the development of more accurate, interpretable, and clinically relevant predictive models for precision medicine.

References

[1] Pinero, Janet, Bravo, Alex, Queralt-Rosinach, Nuria, Gutiérrez-Sacristan, Alba, Deu-Pons, Jordi, Centeno, Emilio, Garcia-Garcia, Javier, Sanz, Ferran, Furlong, and L. I., "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, pp. D833–D839, 10 2016.

[2] McKusick and V.A., "Mendelian inheritance in man: a catalog of human genes and genetic disorders, jhu press," 1998.

[3] Coudert, Elisabeth, Gehant, Sebastien, de Castro, Edouard, Pozzato, Monica, Baratin, Delphine, Neto, Teresa, Sigrist, C. J. A, Redaschi, Nicole, Bridge, Alan, and T. U. Consortium, "Annotation of biologically relevant ligands in UniProtKB using ChEBI," *Bioinformatics*, vol. 39, p. btac793, 12 2022.

[4] "Benchmark gene reference data for breast cancer, lastvisited: 31 feb' 2024, url: https://data.mendeley.com/datasets/xdkvk75ns7/2,"

[5] S. P. Armada, S. J. Barrett, D. R. Wille, A. Perera-Lluna, A. Gutteridge, and B. H. Dessailly", "Benchmarking network propagation methods for disease gene identification," 2019.

[6] U. Raj, A. P. Anil, A. Shukla, K. Anoosha, and A. Srivastava", "Benchmark gene reference data for breast cancer, mendeley data, v2,"

[7] S. Raj, A. P. Anil, A. Shukla, K. Anoosha, and A. Srivastava, "Benchmark data set for breast cancer associated genes," *Data in Brief*, vol. 45, p. 108583, 2022.

[8] Ambergera, J. S, Bocchini, C. A, Scott, A. F, Hamosh, and Ada, "OMIM.org: leveraging knowledge across phenotype–gene relationships," *Nucleic Acids Research*, vol. 47, pp. D1038–D1043, 11 2018.

[9] Davis, A. Peter, Wiegers, T. C, Johnson, R. J, Sciaky, Daniela, Wiegers, Jolene, Mattingly, and Carolyn.J, "Comparative Toxicogenomics Database (CTD): update 2023," *Nucleic Acids Research*, vol. 51, pp. D1257–D1262, 09 2022.

[10] Pinero, Janet, Ramírez-Anguita, J. Manuel, Saüch-Pitarch, Josep, F. Ronzano, Centeno, Emilio, Sanz, Ferran, Furlong, and L. I, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, pp. D845–D855, 11 2019.

[11] Landrum, M. J, Lee, J. M, Benson, Mark, Brown, G. R, Chao, Chen, Chitipiralla, Shanmuga, and Gu, "ClinVar: improving access to variant interpretations and supporting evidence," *Nucleic Acids Research*, vol. 46, pp. D1062–D1067, 11 2017.

[12] Orphanet, "Orphanet: an online database of rare diseases and orphan drugs," 2023.

[13] T. J. H. University, T. J. H. Hospital, and J. H. H. System, "Generalized anxiety disorder (gad)," 2023.

[14] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[15] Mailman and M. D. et al., "The ncbi dbgap database of genotypes and phenotypes." nature genetics vol. 39,10," 2007.

[16] Pinero, Janet, Ramírez-Anguita, J. Manuel, Sauch-Pitarch, Josep, Ronzano, Francesco, Centeno, Emilio, Sanz, Ferran, Furlong, and L. I, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, pp. D845–D855, 11 2019.

[17] Amberger, J. S, Bocchini, C. A, Schiettecatte, François, Scott, A. F, Hamosh, and Ada, "OMIM.org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 43, pp. D789–98, Jan. 2015.

[18] "Online mendelian inheritance in man, accessed on 25 feb 2023,"

[19] Keane and D.A., "Integrating OMIM and IntAct Data for the Analysis of Gene-Phenotype Interactions in Complex Diseases: A Linux-Based Computational Tool for Network Analysis ," 2023.

[20] Hamosh, Ada, Scott, A. F., Amberger, J. S., Bocchini, C. A., McKusick, and V. A., "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 01 2005.

[21] Rahit, K.T.H., Avramovic, V., Chong, J.X., Tarailo-Graovac, and M., "GPAD: a natural language processing-based application to extract the gene-disease association discovery information from OMIM ," 2024.

[22] Iang, X., Lu, W., Shen, X., Wang, and Q., "Repurposing sertraline sensitizes non–small cell lung cancer cells to erlotinib by inducing autophagy," 2018.

[23] GeneCards, ""https://www.genecards.org/"," 15 February 2024.

[24] G. Catalog, ""https://www.ebi.ac.uk/gwas/"," 18 February 2024.

[25] G. Caentral, ""https://www.gwascentral.org/"," 20 February 2024.

[26] B. T. R. T, S. T, and B. AJ", "Gwas central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies.," vol. GWAS Central: , Nucleic Acids Research, 6 January 2023.

[27] A. J. B. Tim Beck, Tom Shorter, "A comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies," *SMBM 2014*, vol. GWAS Central: , Nucleic Acids Research, 08 January 2020.

[28] "Majidian, Agustinho, Chin, and C. et al.", "Genomic variant benchmark: if you cannot measure it, you cannot improve it. genome biol 24, 221," 2023.

[29] "Indurkhya, . Damerau, and F.J.", "Handbook of natural language processing (2nd ed.). chapman and hall/crc.," 2010.

[30] "Kao, A., Poteet, and S.R.", "Natural language processing and text mining. springer, berlin.," 2007.

[31] A. Moschitti and R. Basili, "European conference on information retrieval, springer," 2007.

[32] H. Amazal, M. Ramdani, and M. Kissi", "International conference on smart applications and data analysis, springer," 2020.

[33] "Rebholz-Schuhmann, Dietrich, Oellrich, Anika, Hoehndorf, and Robert", "Text-mining solutions for biomedical research: enabling integrative biology," *Nature Reviews Genetics*, vol. 13, no. 12, pp. 829–839, 2012.

[34] "Cohen, A. M., Hersh, and W. R.", "A survey of current work in biomedical text mining," *Briefings in Bioinformatics*, vol. 6, pp. 57–71, 03 2005.

[35] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, and Kristina, "BERT: Pre-training of deep bidirectional transformers for language understanding," (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, 2019.

[36] Su, Yuhan, Xiang, Hongxin, Xie, Haotian, Yu, Yong, Dong, Shiyan, Yang, Zhaogang, Zhao, and Na, "Application of BERT to enable gene classification based on clinical evidence," *Biomed Res. Int.*, vol. 2020, p. 5491963, Oct. 2020.

[37] Lee, Jinhyuk, Yoon, Wonjin, K. adn Sungdong, Kim, Donghyeon, Kim, Sunkyu, C. Ho, and K. adn Jaewoo, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 09 2019.

[38] Wei, Chih-Hsuan, Kao, Hung-Yu, Lu, and Zhiyong, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic Acids Research*, vol. 41, pp. W518–W522, 05 2013.

[39] MetaMap, "Metamap brings greater clarity to identity verification,"

[40] Bravo, Janet, Queralt, Rautschka, Michael, Furlong, and L. I, "Befree: a text mining system to extract relations between genes, diseases and drugs for translational research," *SMBM 2014*, vol. 79, 2014.

[41] "Public health genomics and precision health knowledge base (v9.0), lastvisited: 5 mar' 2024, url: shttps://phgkb.cdc.gov/phgkb/startpagepublit.action,"

Appendix

Appendix A

Codes for SVM

Listing A.1: Python code of Decision Tree

```
# Define the features (X) and the target variable (y)
# Assuming 'ASSOCIATION_CLASS' is a categorical column
X = pd.get_dummies(data[['YEAR', 'WEIGHT','GENE','REF_GENE',
            'REF_SENTENCE','TITLE','DIS_CLASS','CONCLUSION']])

# Continue with the rest of your code
# Adjust the feature columns as needed
y = data['ASSOCIATION_CLASS']  # Replace 'CONCLUSION' with the actual target

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the SVM classifier
svm_classifier = SVC(kernel='linear')

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

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

Appendix B

Codes for Random Forest

Listing B.1: Python code of Random Forest

```
# Define the features (X) and the target variable (y)
# Assuming 'ASSOCIATION_CLASS' is a categorical column
X = pd.get_dummies(data[['YEAR', 'WEIGHT','GENE','REF_GENE',
            'REF_SENTENCE','TITLE','DIS_CLASS','CONCLUSION']])

# Continue with the rest of your code
# Adjust the feature columns as needed
y = data['ASSOCIATION_CLASS']  # Replace 'CONCLUSION' with the actual target

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)

# Define the Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)

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

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

Appendix C

Codes for Decision Tree

Listing C.1: Python code of Decision Tree

```
# Define the features (X) and the target variable (y)
# Assuming 'ASSOCIATION_CLASS' is a categorical column
X = pd.get_dummies(data[['YEAR', 'WEIGHT','GENE','REF_GENE',
'REF_SENTENCE','TITLE','DIS_CLASS','CONCLUSION']])

# Continue with the rest of your code
# Adjust the feature columns as needed
y = data['ASSOCIATION_CLASS']  # Replace 'CONCLUSION' with the actual target

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the Decision Tree classifier
decision_tree_classifier = DecisionTreeClassifier()

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

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

Appendix D

Codes for KNN

Listing D.1: Python code of KNN

```
# Define the features (X) and the target variable (y)
# Assuming 'ASSOCIATION_CLASS' is a categorical column
X = pd.get_dummies(data[['YEAR', 'WEIGHT','GENE','REF_GENE',
        'REF_SENTENCE','TITLE','DIS_CLASS','CONCLUSION']])

# Define the target variable
y = data['ASSOCIATION_CLASS']  # Replace 'CONCLUSION' with the actual target

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the KNN classifier
k = 5  # Number of neighbors to consider (you can adjust this value)
knn_classifier = KNeighborsClassifier(n_neighbors=k)

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

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