A Study on Benchmark Datasets for Human Disease Genes

-The more we Learn, The more we Serve

Group Members :

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

Supervised By:
Dr. S.M.A. Al-Mamun

Objective & Goal

- Continuous Research and Collaboration
- Personalized Medicine
- Clinical Applications
- Limitations of Models
- Enrich the Benchmarking Datasets.
- Risk Assessment
- Early Diagnosis

Human Gene-Disease datasets

Raw Datasets

- **Availability:** Scattered over literature as free-text data, Unorganized, Not recorded properly
- **Usefulness:** Not compatible for computational analysis
- **Work compatible:** Not suitable and slow process

Benchmark Datasets

- **Availability:** OMIM, DisGeNet, GAD, GWAS
- **Usefulness:** Cytoscape plugin and Computational analysis
- **Work compatible:** Representable as GDAs graph

State of art Genomics Datasets

- OMIM
- DisGeNET
- CTD
- UniPort
- GAD
- GWAS Catalog
- Orphanet
- ClinVer
- dbGaP
- GeneCards
- GWAS Central

Resources On Genotype and Phenotype

Name	URL	Scope	Organism	Current Statistics	Original Reference	Current Reference
DisGeNET	DisGeNET[1]	Gene-disease, and variant-disease associations	Human	1134942 associations, between 21,671 genes and 30170 diseases, 46589 SNPs	2010 [33]	[7]
Comparative Toxicogenomics Database(CTD)	CTD[2]	Chemicals, genes, and disease associations	Human and animal models	1127498 associations between 20,027 genes and 1504 diseases	2003 [34]	[6]
Online Mendelian Inheritance in Man(OMIM)	OMIM[3]	Mendelian diseases and their genes	Human	121512 associations between 29,596diseases and 20,790 genes	1998 [35]	[5]
Genetic Association Database(GAD)	GAD[4]	Genes, variants, and complex diseases and traits	Human	74928 associations between 12,774 diseases and 10,697 genes	2004 [11]	[11]
UniProt Knowledgebase	UniPort[5]	Proteins	Human	566467 associations between 1545 diseases and 19,368 genes	2004 [36]	[9]
The NHGRI-EBI Catalog of published GWAS (GWAS Catalog)	GWAS[6]	GWAS studies	Human	30,148 associations between 2,743 diseases and 21,449 genes (18,666 variants)	2009 [36]	[12]

Tools for Genomic Data Extraction

BeFree

BioBERT

PubTator

MetaMap

DisGeNet- A Dataset of Gene-Disease Associations

Released On

September, 2010

Raw Data Source

Omim, Unipport, CTD, CURATED

Data Extraction

Text mining tools

Repository

Open Access

Curated Data Source

Mendelian, Curated repositories, GWAS catalogues, Scientific literature

Access DisGeNET

Web Interface, SQLite database, Cytoscape, Semantic web

Informations

GDAs & VDAs

DisGeNet Versions

V4.0 (June,2016)
- 429, 036 GDAs
- 17, 381 genes
- 15, 093 diseases

V6.0 (June, 2019)
- 628, 625 GDAs
- 17, 549 genes
- 24, 166 diseases

V7.0 (June, 2020)
- 1, 134, 942 GDAs
- 21, 671 genes
- 30, 170 diseases

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

Detailed Statistics of the DisGeNET's Current Version: v7.0

OMIM(Online Online Mendelian Inheritance in Man)

First Introduced
In the 60s

Text Data
Expert Curation

Data Format
Text but structured

Traditional focus
Mendelian diseases

Recent focus
Complex diseases

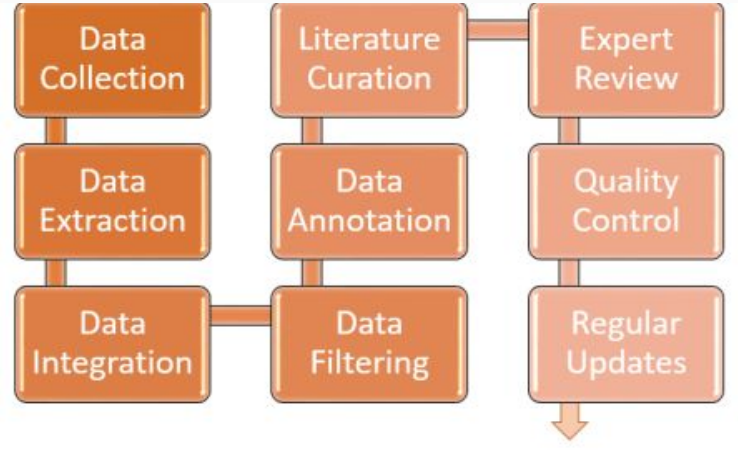
Access OMIM
Website, API

Category	GDAs	Genes	Diseases
Total	121, 512	29, 596	20, 790

Statistics of the OMIM Current Version

Workflow Diagram

Disgenet



OMIM

Type Of Machine Learning Model

- **Supervised Learning**
  - Classification
  - Regression
- **Unsupervised Learning**
  - Clustering
  - Dimensionality Reduction
  - Anomaly detection
- **Reinforcement learning**
- **Semi-supervised learning**

Machine Learning Model for Gene Disease Association

- Decision Tree (DT)
- Support Vector Machines (SVMs)
- Random Forests (RF)
- Neural Networks (NN)

Decision Tree (DT)

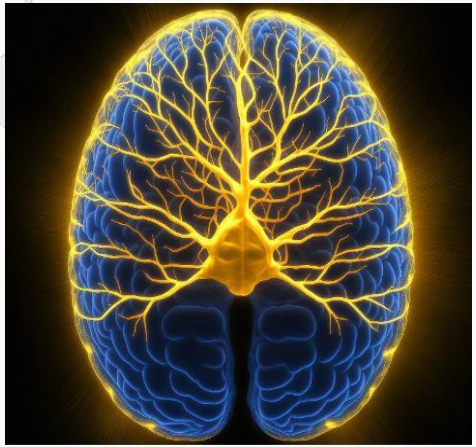
- Recursively Partitioning
- Splitting Until Pure
- Root to Leaf
- Non- Parametric
- Overfitting
- HD not
- Medical D., Fraud Detection, Customer Segmentation, Risk Assessment.

Support Vector Machines (SVMs)

- Optimal Hyperplane
- HD yes
- Well-suited for Classification
- Discrete Category
- Image classification, Text classification, Bioinformatics, Anomaly detection

Random Forest (RF)

- Combines Multiple Decision Trees
- Concept of Bagging
- Randomness
- Voting
- HD yes
- Medical D., Fraud Detection, Customer Segmentation, Risk Assessment.



NEURAL NETWORK (NN)

- Human Brain
- Adaptability
- Both (C & R)
- Very Effective & Expensive
- P, MLP, CNN and RNN
- Image Recognition, NLP, Medical D., Speech R., Recommender Systems, Predictive Analytic

Comparative Analysis Between Models

Feature	SVM	Random Forest	Neural Networks	Decision Tree
Classification	Binary and multi-class	Binary and multi-class	Binary and multi-class	Binary and multi-class
Regression	Yes	Yes	Yes	No
Interpretability	High	Low	Low	Medium
Computational complexity	Medium	High	High	Low
Robustness to outliers	High	High	High	Low
Feature scaling	No	No	Yes	No

Table 2.4: Differences Between Models

Summarizing the Analytic Discussion

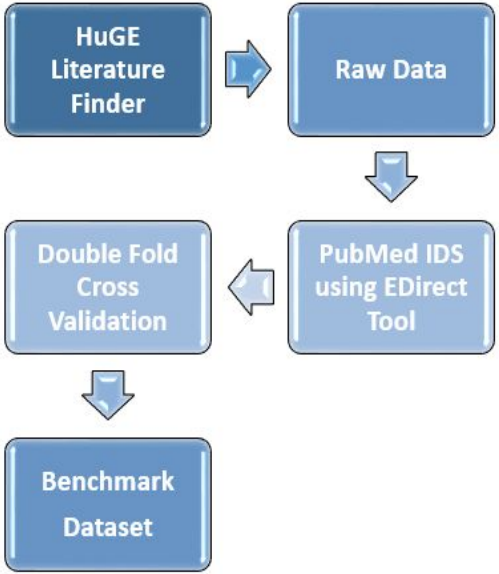
Model	Description	Advantageous	Disadvantageous
SVM	Supervised	HD data, Robustness, Interpretation	Computational Complexity, tuning
RF	Ensamble	High accuracy, No scaling, Robustness,	Complexity, Hyperparameter
NN	Human Brain	High Accuracy, adaptability	Interoperability, Hyperparameter
DT	Tree Like Structure	Simplicity, Efficiency	Overfitting, Noise S.

Table 2.5: Summary of the Models

Study of Existing Papers

Title: Benchmark data set for breast cancer associated genes.[1]

- 12565 records were processed
- Raw data were collected From HuGE Literature Finder
- EDirect Tool was used to separate PubMeds
- Double Cross Validation were applied
- The benchmark Dataset was achieved



Title: Benchmarking network propagation methods for disease gene identification.[2]

- Identify genes for targeted drug treatment
- Data Source is OpenTargets database
- 12 specialized computer algorithms were used across 22 common diseases
- Six different measures used to evaluate algorithm efficacy.
- Cross-validation were employed to ensure reliability of results

Title: Genomic variant benchmark: if you cannot measure it, you cannot improve it.[3]

- Crucial for evaluating variant caller accuracy in genomics research
- Dataset Created using diverse sequencing technologies
- Establishes pipelines, fosters new sequencing approaches
- Short-read technologies widely used but have limitations
- Long-read sequencing used for detecting previously undetected SVs

Steps of benchmarking

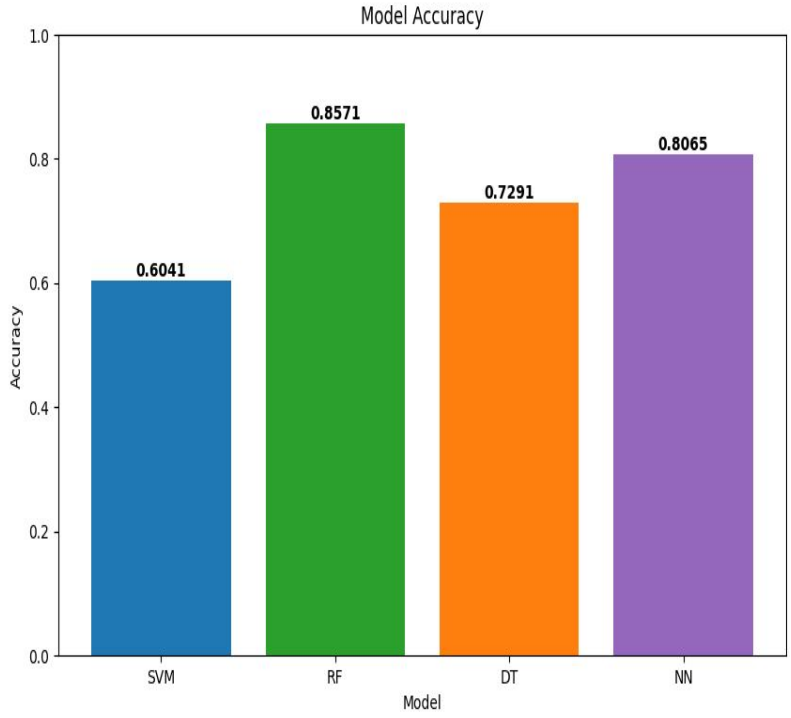
Methodology

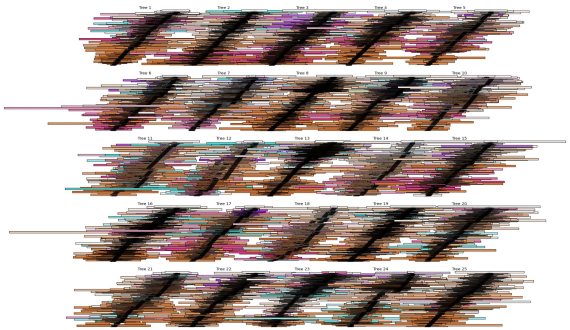
S.NO: Serial number
DB_ID: Database identifier
DIS_CLASS: Disease class
GENE: Gene information
PUBMED.ID: PubMed identifier
LACKASSO: Lack of association indicator
TITLE: Title of the publication
YEAR: Year of publication
CONCLUSION: Conclusion of the publication
REF_SENTENCE: Reference sentence
ASSOCIATION_CLASS: Association class
REF_GENE: Reference gene
GENE_NEW: New gene information
WEIGHT: Weight value
Dimension of the dataset: 14

It doesn't directly provide numerical values for the Interpretability, Robustness to noise, Handling of missing data, and Computational complexity for any ML model

Method Name	Accuracy
Neural Network	80.65%
SVM	60.41%
Decision Tree	72.91%
Random Forest	85.71%

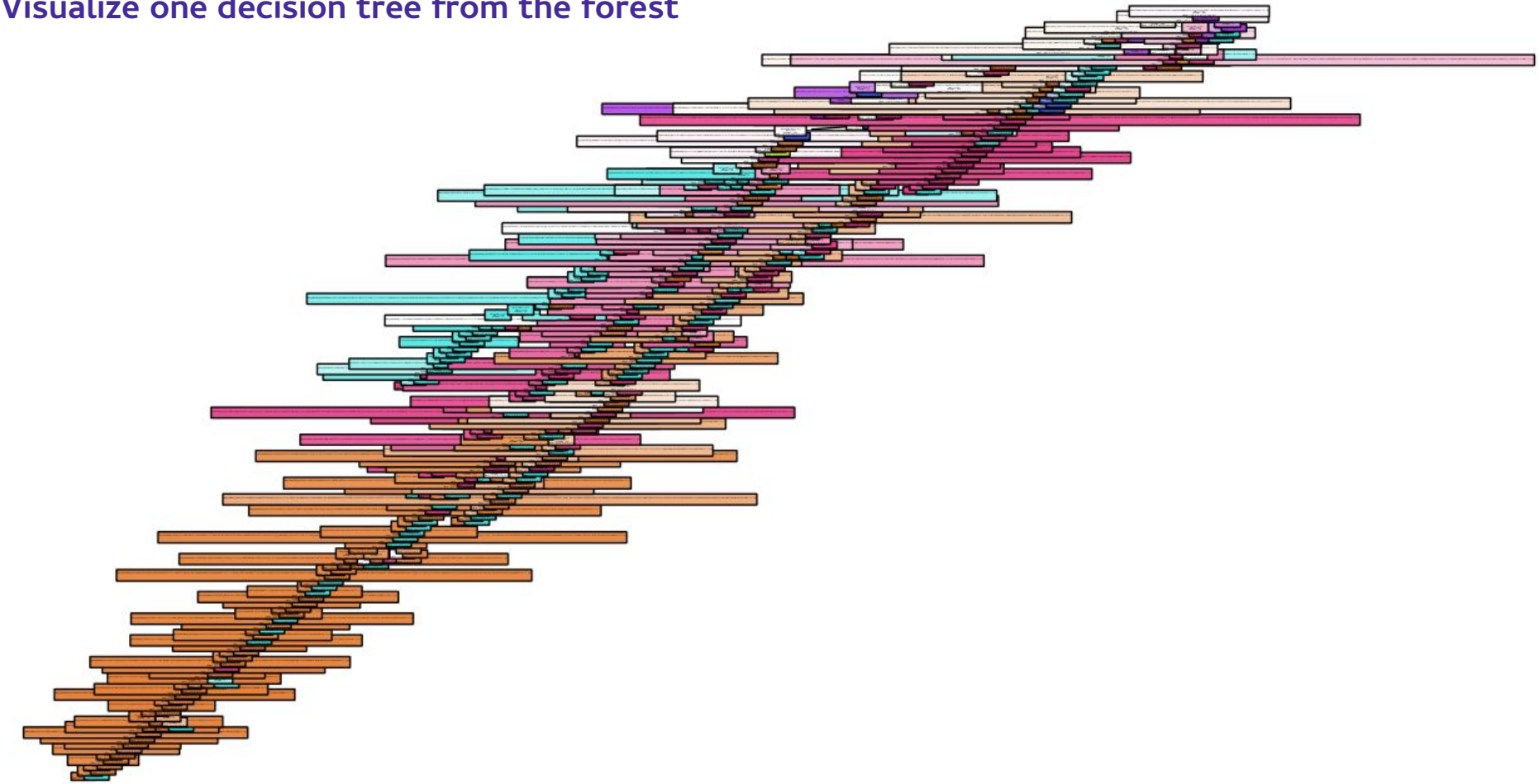
Table 4.1: Perfomance on Different Method





Random Forest Tree Snapshot

Visualize one decision tree from the forest



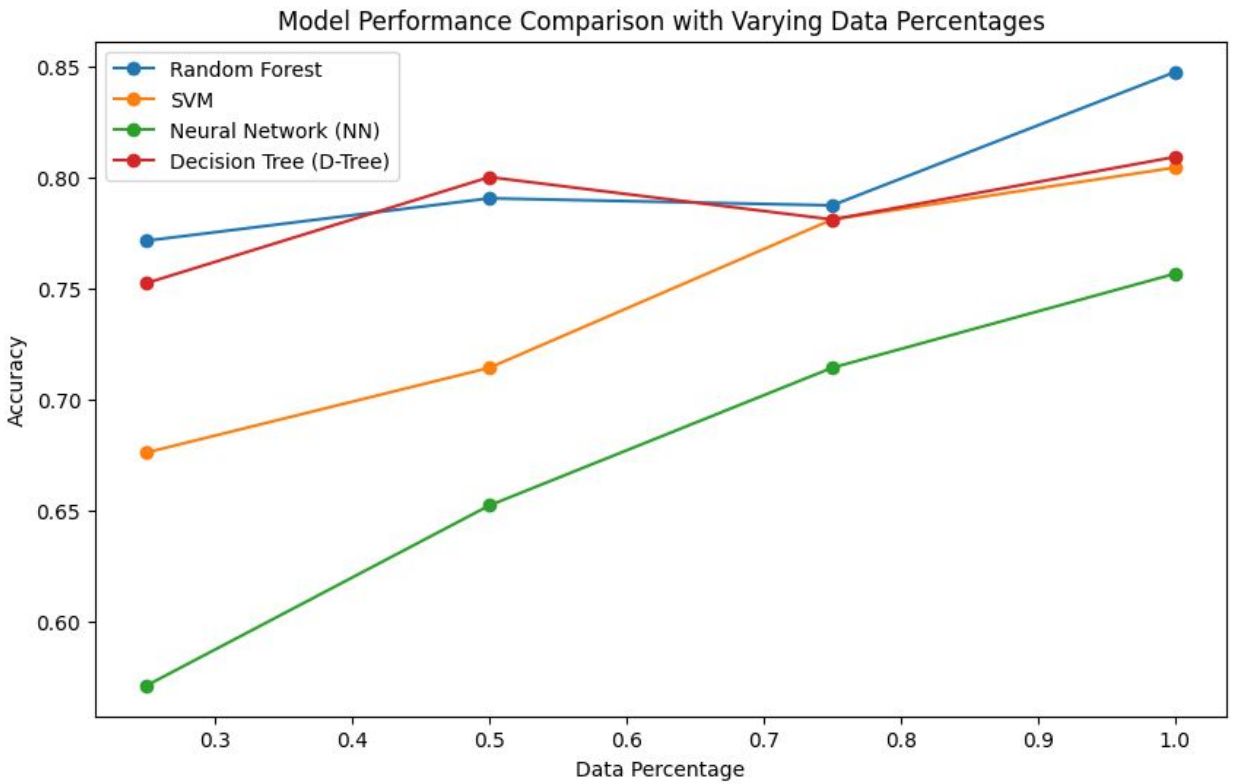
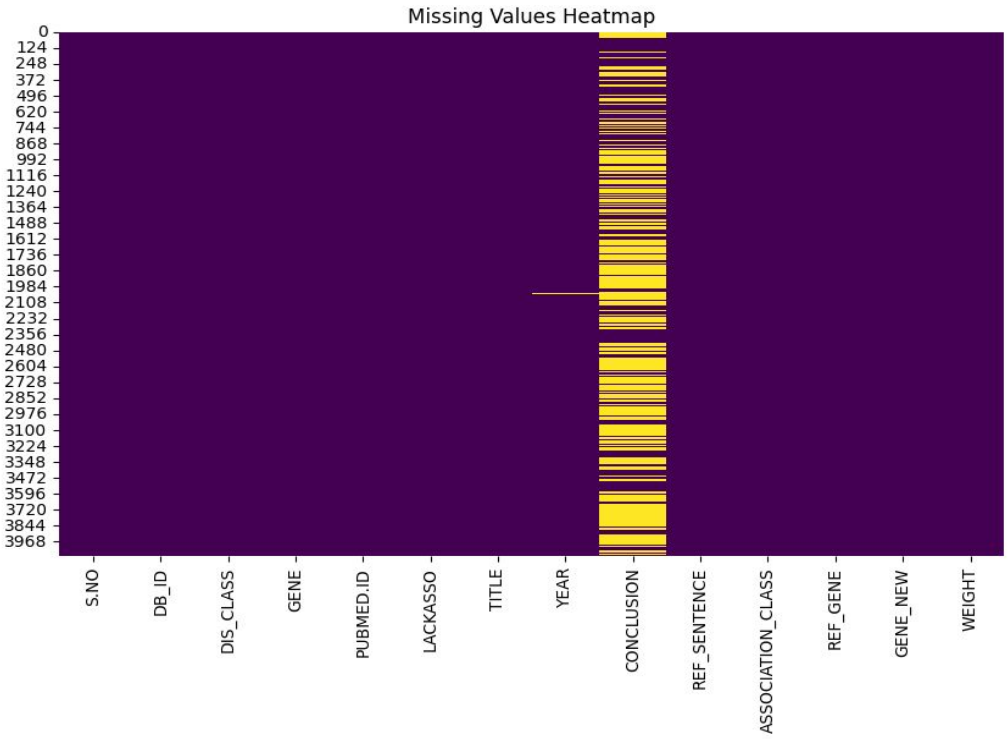

Model Performance Comparison with Varying Data Percentages
Random Forest
SVM
Neural Network (NN)
Decision Tree (D-Tree)
Accuracy
0.85
0.80
0.75
0.70
0.65
0.60
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Data Percentage

Table 1: Missing Values in the Dataset

Column	Missing Values
S.NO	0
DB_ID	0
DIS_CLASS	0
GENE	0
PUBMED.ID	0
LACKASSO	0
TITLE	1
YEAR	8
CONCLUSION	2403
REF_SENTENCE	0
ASSOCIATION_CLASS	0
REF_GENE	0
GENE_NEW	0
WEIGHT	0



Conclusion & Future Work

- Vast field of Research
- Human Helpful Innovation
- Medical Milestone
- Need to Walk More
- Hybrid Model
- Feature Selection
- Optimization
- Interpretability
- Applications

References

1. S. A. A. P. S. A. A. K. S. Raj and Alok, ““benchmark gene reference data for breast cancer”, mendeley data, v2, doi: 10.17632/xdkvk75ns7.2,”

2. S. P.-A. S. J. B. D. R. W. A. P.-L. A. G. B. H. Dessailly, “Benchmarking network propagation methods for disease gene identification,” 2019.

3. S. A. D. Majidian and C. C. et al., “Genomic variant benchmark: if you cannot measure it, you cannot improve it. genome biol 24, 221,” 2023.

Thank You!

Do you have any questions?