MJCIS Vol.16 No.2 Dec 2020

Mansoura Journal For Computers and Information Sciences

Gene-Disease Association

E.E.Abdelbadeea, M.A.El-Dosuky, M.Z.Rashad

Faculty of Computers and Information, Computer Science Dept. Mansoura University, Egypt
, Esraa2020 @yahoo.com
Faculty of Computers and Information, Computer Science Dept. Mansoura University, Egypt
, mouh_sal_010@mans.edu.eg
Faculty of Computers and Information, Computer Science Dept. Mansoura University, Egypt
, magdi_z2011@yahoo.com

ABSTRACT

Disease susceptibility prediction is defined as follows. Given training set S and a test case t∉S as a tuple (known as SNP, unknown disease), trying predicting the unknown disease with maximum accuracy. DisGeNET is a proponent dataset in disease susceptibility research. This paper reviews DisGeNET comprehensive information, before introducing a proposed system operating atop it. First, vetting the dataset by consolidation, and removing genes with effects beyond a certain threshold. Second, computing the empirical cumulative distribution function, using it for plotting and printing gene associations for many diseases such as, and not limited to, Alzheimer, Anemia, and Brain, breast cancer proposed methods such as applying C4.5 & naïve Bayes give better accuracy then previous works

Keywords

DNA analysis, epidemiological, DisGeNET, DNA Disease susceptibility, and disease susceptibility prediction.

1. INTRODUCTION

Computational DNA analysis in epidemiological studies **[1]** is carried out in three phases: finding useful single nucleotide polymorphisms (SNPs), search for SNPs-to-diseases associations, and Disease Susceptibility Prediction (DSP).

This paper focuses on DSP using previously prepared standard dataset. Data acquisition is achieved by using DisGeNET data set **[2].** It contains 17,381Genes and 15,093Diseases as shows in table1

In this study, we used the disease susceptibility prediction defined as follows. Given training set S and a test case t∉S as a tuple (known as SNP, unknown disease), trying predicting the unknown disease with maximum accuracy to determine the two most important genes in the disease **[3].**

The operation of the proposed system is clarified in view of the general operational framework, by combining C4.5 and decision tree. The result of the accuracy measure is 81.7% for Crohn disease for instance, compared to support vector machine (SVM).

Table 1. Statistic of database [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]

Database	Disease	Gene
Curated	1857	825
UniProt	606	125
Lhgdn	163	182
Gad	133	2168
Befree	8296	4819
Literature	7416	9240
CTD Human	3090	5269
Predicted	133	16
M.musculus	1021	871
R.Norvergicus	664	1035

Table 1. Displayed means, the curated has 1857 diseases, 825 genes. The UNIPORT has 606 diseases, 125 genes. The LHGDN has 163 diseases, 182 genes. The GAD has 133 diseases, 2168 genes. The Befree has 8296 diseases, 4819 genes. The LITERATURE has 7416 diseases, 9240 genes. The CTD human has 3090 diseases, 5269 genes. PREDICTED has 133diseases, 16 genes. The M.musculus has 1021 diseases, 871 genes. The R.Norvergicus has 664 diseases, 1035 genes.

2. RELATED WORK

There were many researchers who have taken different ways to identify accurate methods for diagnosis of diseases such as Alzheimer's, anemia, brain and breast cancer.

Main algorithms for Protein sequence-sequence alignment are PAM matrix (construct a score matrix for guide protein sequence alignment) **[14],** BLOSUM: (Most often-used score matrix for protein sequence alignment) **[15],** Needleman-Wunsch :(A dynamic programming algorithm for sequence alignments) **[16].**

The General Method for Sequence Comparison is Smith-Waterman: (An extension of Needleman-Wunsch algorithm) **[17]** A solution to asymmetric gap penalty by recursion **[18],** FASTA (A heuristic arrangement speedier than flow programming) **[19],** BLAST (The most often-used heuristic

alignment) **[20]** and Dumas look-up table for identifying common words from a sequence database **[21].**

Multiple sequence alignments algorithms include PSI-BLAST (The most often-used algorithm for sequence-profile alignment) **[22],** ClustalW (An algorithm to alignment multiple sequences) **[23],** Neighbor-joining method (for constructing phylogenetic tree) **[24],** protein sequence profile **[25],** and Hidden Markov Model **[26],** Profile-profile alignment algorithms include **[27], [28]** and **[29].** Algorithms for Protein structure comparison are TM-score (propose TM-score which is more delicate to topology than RMSD) **[30],** Dali (a generally utilized program for protein structure arrangement) **[31].**

CE (a broadly utilized program for Protein structure arrangement) **[32]**, and TM-adjust (a fast and productive calculation for protein structure arrangements) **[33].**

Algorithms for Protein secondary structures **[34]** are **[35], [36]** and PSI-PRED: the most popular software for protein Secondary structure prediction. **[37]** Main algorithms for Protein structure prediction are protein folds recognition (threading) **[38],** HMM-based threading. **[39].**

Profile-profile based threading method **[40],** Rosetta: a free modeling algorithm based on fragment assembly **[41],** TASSER: a composite technique for protein structure expectation **[42], [43]** and **[44].**

The first paper introducing the replica-exchange Monte Carlo simulation was **[45]** while the first paper introducing the simulated annealing method was **[46].**

Cell type-selective Disease-association of genes under high regulatory load **[47],** DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. **[48],** DisGeNET: a Comprehensive platform coordinating data on human disease-associated genes and variants **[49],** the relationship of PLA2G2A single nucleotide polymorphisms with type II a secretory Phospholipase A2 level yet not its activity in patients with stable coronary heart disease **[50].**

3. Proposed Methods

Based on the code, the site shows that I have a disease in the data decreased and left 150 diseases called disease-ontology.

3.1 Preprocessing

Consolidation Table 2 shows a sample of data consolidated by doid name, and gene id. Then select the score max, score mean, and sort consolidated data frequency.

3.2 ECDF

ECDF stands for Empirical Cumulative Distribution Function. It is a preprocessing phase before actual classification.

The Fig.1 shows the steps of Empirical Cumulative Distribution Function. In, python, ECDF is calculated as arrange (1, Len(x) +1) / Len(x).Frist, the disease name is set, then projection is done, then selection of data related to the disease. The ECDF is calculated, then threshold cut is performed, then the LOOP iterates to operate on the next disease.

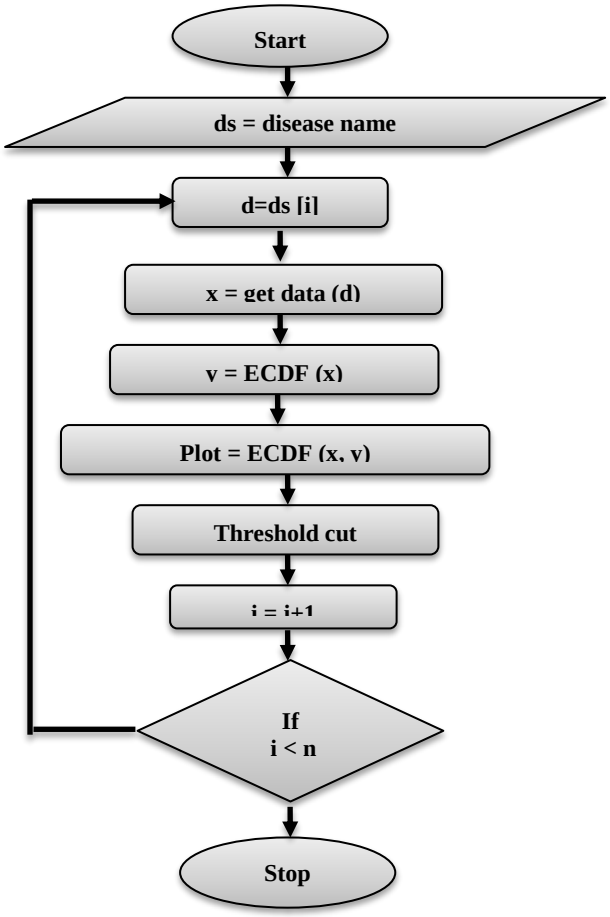

Fig 1: ECDF Flow Chart

Table2.Sample of Dataset

Do id code	Do id name	Gene id	Gene symbol	count	Pub meds-max	Source max	Source mean	Association Type
DOID: 2377	multiple sclerosis	3123	HLA-DRB1	3	232	0.362439	0.124658	Biomarker |Genetic Variation
DOID: 2377	multiple sclerosis	348	APOE	2	68	0.302967	0.259084	Biomarker |Genetic Variation
DOID: 2377	multiple sclerosis	3119	HLA-DQB1	2	63	0.289955	0.147295	Altered Expression| Biomarker |Genetic Variation
DOID: 2377	multiple sclerosis	3575	IL7R	1	38	0.274986	0.274986	Biomarker |Genetic Variation
DOID: 2377	multiple sclerosis	3559	IL2RA	3	46	0.266881	0.089244	Biomarker |Genetic Variation
DOID: 2377	multiple sclerosis	3456	IFNB1	3	149	0.253425	0.087027	Biomarker |Genetic Variation |Therapeutic
DOID: 2377	multiple sclerosis	23274	CLEC16A	1	20	0.248286	0.248286	Biomarker |Genetic Variation
DOID: 2377	multiple sclerosis	3553	IL1B	1	21	0.239379	0.239379	Biomarker |Genetic Variation |Therapeutic

3.3 Classification

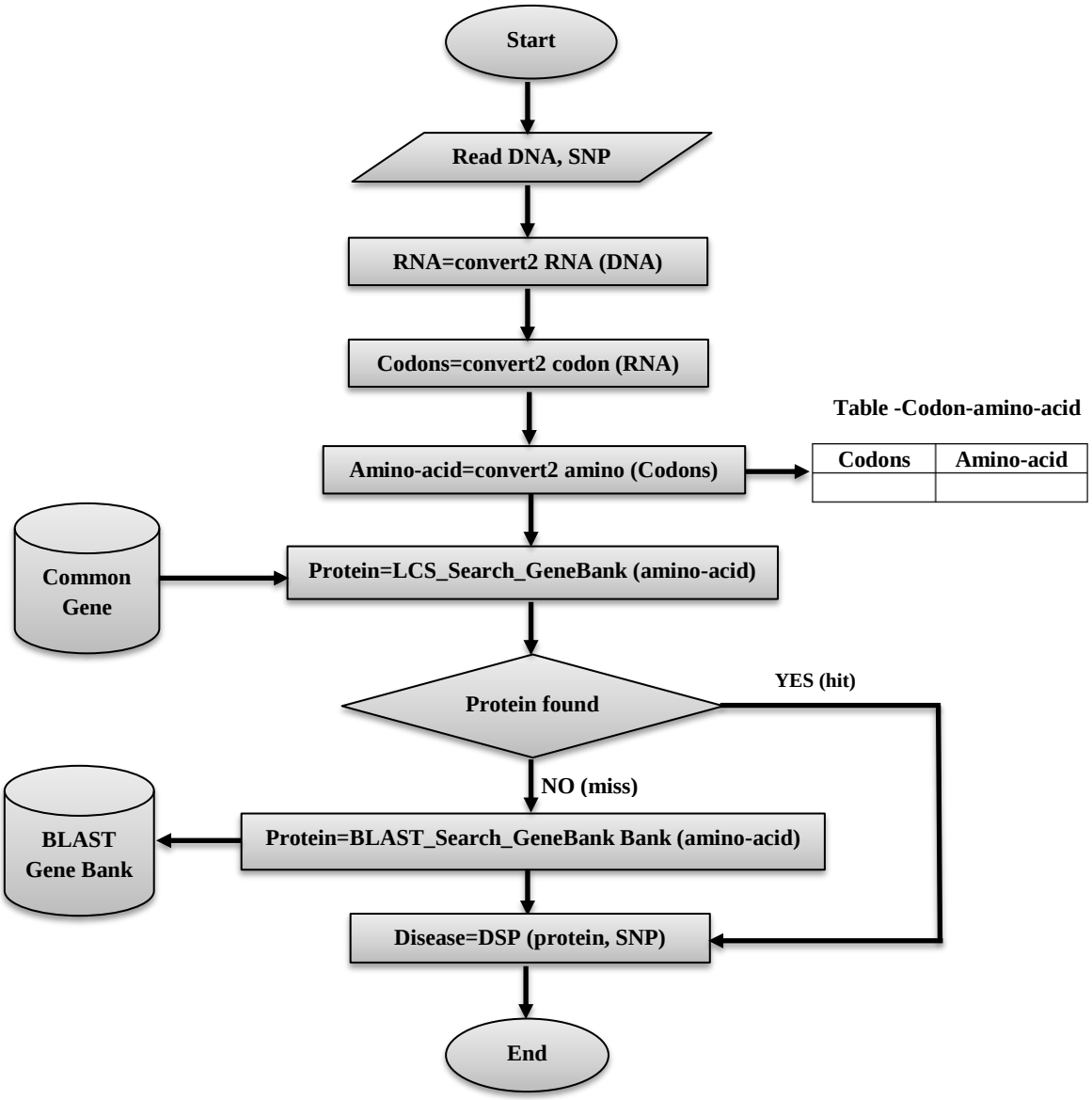

Fig 2: Proposed system flowchart

The previous flow chart is explained as follows:

1. Reading DNA

In this step, DNA sequence is acquired here is an example.

a	c	g	c	a	t	c	g	g	c	t	a	t	a	c	a	a	g

2. Convert2RNA

Are replaces T with U. The transcribed DNA into an mRNA sequence is

a	c	g	c	a	u	c	g	g	c	u	a	u	a	c	a	a	g

3. Convert 2 Codon

By splitting, sequence three by three letters,

ACG/ CAU/ CGG / CUA/ UAC / AAG

4. Convert 2 Amino

Is done using standard conversion table **[51, 52], so,** the **a**mino acid sequence is

T / H / E /L /Y /K

5. Search gene bank using Longest Common Subsequence **[53].** A subsequence of a sequence S is a set of types that appear in left-to-right direction, but not essentially successively. For **AAACCGTGAGTTATTCGTTCTAGAA** and **CACCCCTAAGGTACCTTTGGTTC**, LCS = is **ACCTAGTACTTTG**.

6. BLAST_Search_GeneBank operation is shown in figure 3.

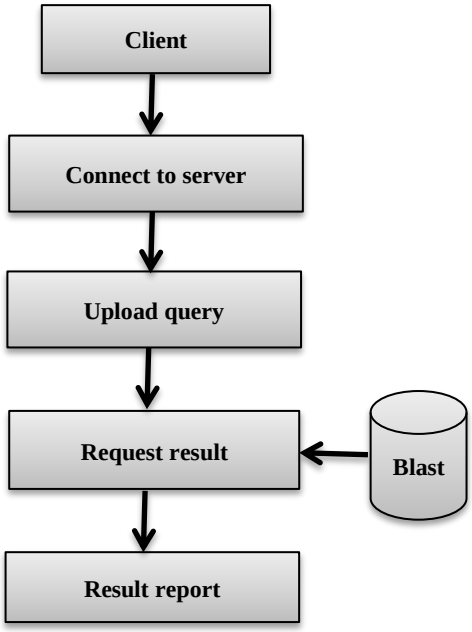

Figure 3:BLAST_Search_GeneBank

For the tutorial how to implement BLAST_Search_GeneBank operation, refer to **[52]** to show how to connect to **HTTP:// *www.ncbi.nlm.nih.gov/ blast/*** and get result report.

4. EVALUATION

4.1 ECDF

In, python, ECDF is calculated the proportion of more genes affecting the occurrence of diseases. As shown figures.

Fig 4. Shows ECDF for Alzheimer's disease, the major gene contribute with 0.7, the minor gene contribute with 0.1, and the confidence is 0.8

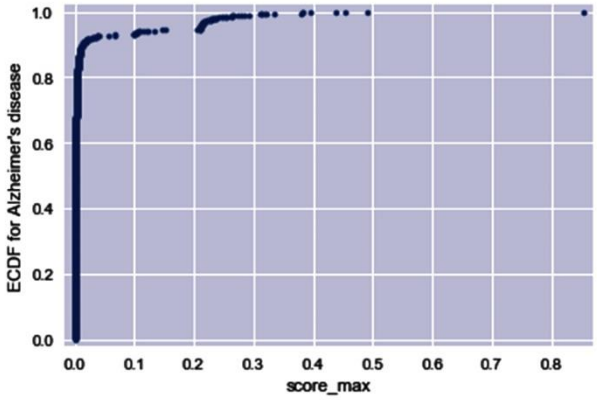

Fig 4: ECDF of Alzheimer's-disease

Fig 5. Shows ECDF for Grave's disease, the major gene contribute with 0.59, the minor gene contribute with 0.01, and the confidence is 0.6

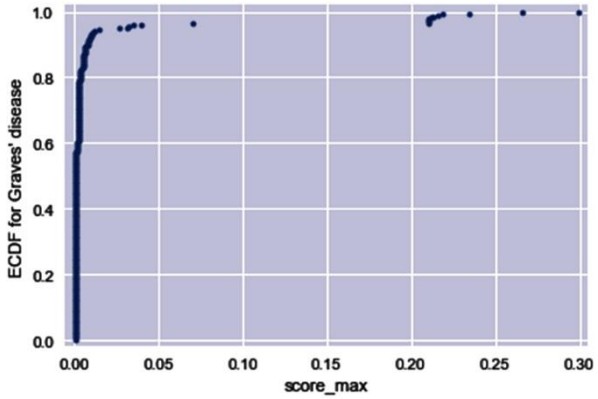

Fig 5: ECDF of Grave's Disease

Fig 6. Shows ECDF for brain cancer, the major gene contribute with 0.68, the minor gene contribute with 0.22, and the confidence is 0.9

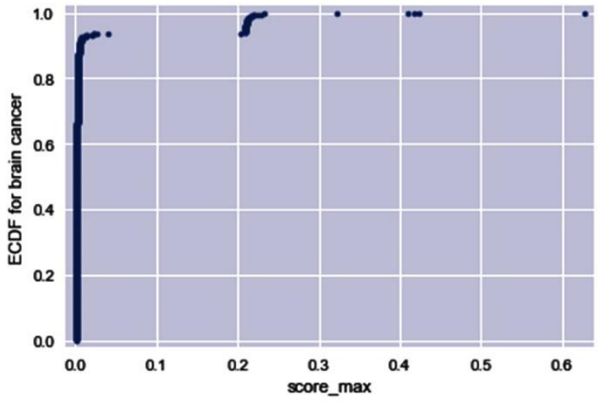

Fig 6: ECDF of brain cancer

Fig 7. Shows ECDF for breast cancer, the major gene contribute with 0.58, the minor gene contribute with 0.12, and the confidence is 0.7

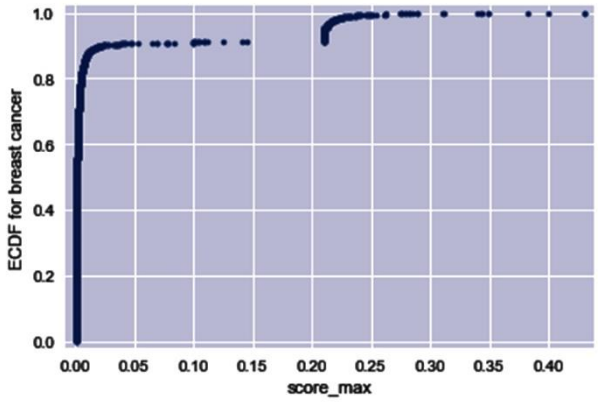

Fig 7: ECDF breast cancer

Fig 8. Shows ECDF for anemia, the major gene contribute with 0.65, the minor gene contribute with 0.15, and the confidence is 0.8

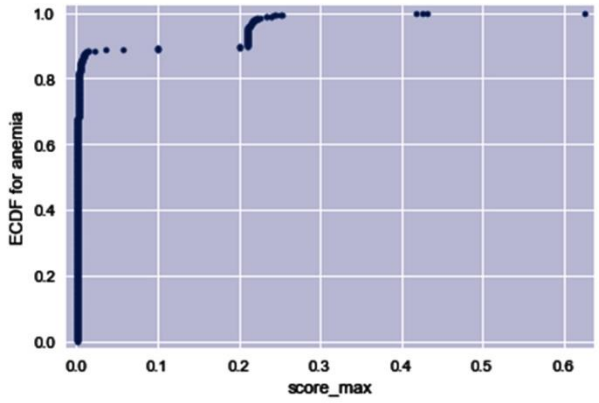

Fig 8: ECDF of anemia

Fig 9. Shows ECDF for Behcet's disease, the major gene contribute with 0.44, the minor gene contribute with 0.35, and the confidence is 0.79

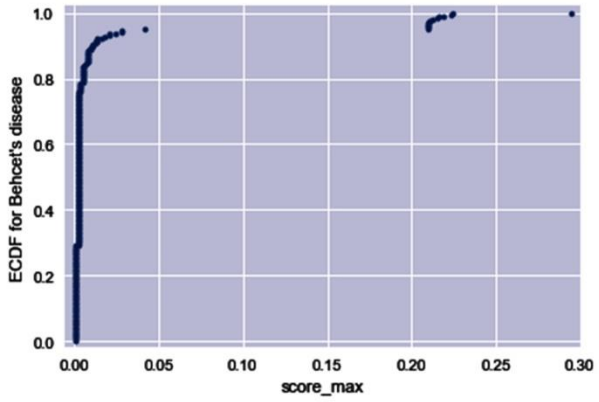

Fig 9: ECDF of Behcet's disease

Fig 10. Shows ECDF for celiac disease, the major gene contribute with 0.39, the minor gene contribute with 0.01, and the confidence is 0.4

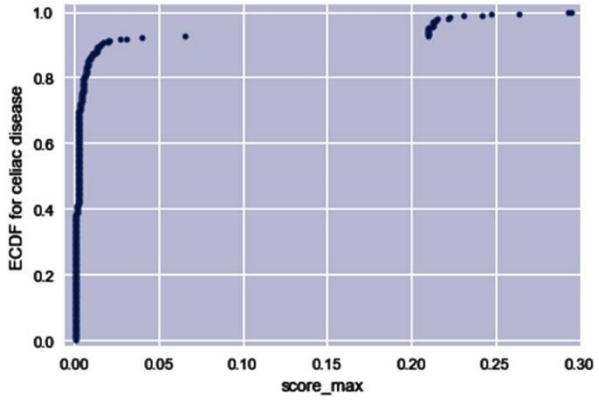

Fig 10: celiac disease

Fig 11. Shows ECDF for chronic kidney failure, the major gene contribute with 0.39, the minor gene contribute with 0.37, and the confidence is 0.76

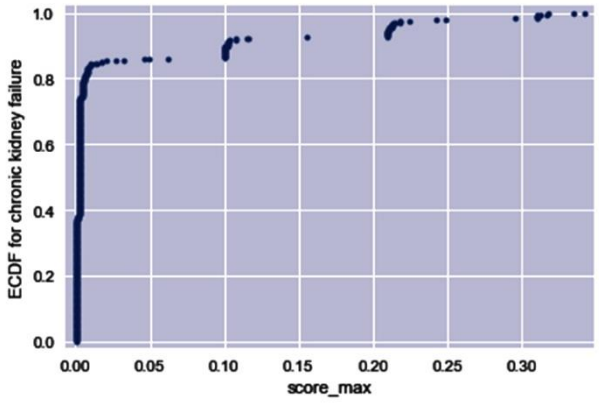

Fig 11: chronic kidney failure

4.2 Classification

For DSP, we use Weka **[53]** on data set from **[54]** containing 435 Human chromosome sequences.

There are too matters for implement are explained as follow:

4.2.1 C4.5 - based classifiers

Are a PC program for actuating characterization administers as choice trees from an arrangement of given examples and a product augmentation of the fundamental ID3 calculation Designed by Quinlan.

Disentangled Algorithm: Let T be the arrangement of preparing examples, choose a characteristic that best separates the occasions contained in T (C4.5 utilizes the Gain Ratio to decide), create a tree hub whose esteem is the picked property, create tyke joins from this hub where each connection speaks to a one of a kind incentive for the picked trait, use the kid interface qualities to additionally subdivide the cases into Subclasses **[55], [56].**

Data is pipelined into a preprocessing phase, in which many operations are performed such as data wrangling then selection and projection of relevant pieces of information.

$$\Pi_{\text{snpId,geneSymbo,disease Name}}$$

This means that Projection selecting columns.

$$\sigma_{\text{disease Name IN(ObsessiveCompulsiveDisorder,Obesity,InflammatoryBowelDiseases,CrohnDisease namely)}}$$

For a matter of simplicity, this paper is scrutinizing only four diseases: Obsessive Compulsive Disorder, Obesity, Inflammatory Bowel Diseases, and Crohn Disease namely.

Selection means selecting certain rows.

Table 3. Accuracy measure

Class	F-Measure	Recall	Precision	FP Rate	TP Rate
Obsessive Compulsive Disorder	0.766	0.692	0.857	0.002	0.692
Obesity	0.924	0.979	0.875	0.185	0.979
Inflammatory Bowel Diseases	0.353	0.25	0.603	0.022	0.25
Crohn Disease	0.809	0.817	0.801	0.084	0.817

The SNPID “1926065” is associated with both diseases Crohn and Inflammatory Bowel but is correlated with Crohn Disease are shown in figure 12.

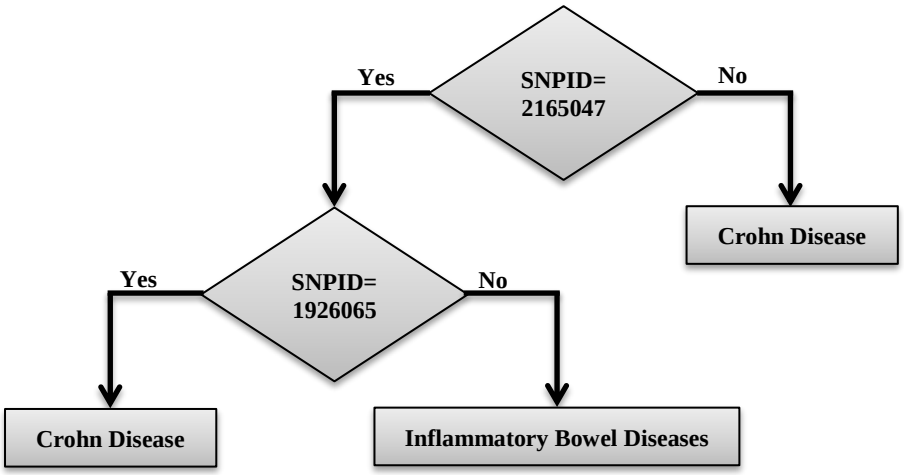

Fig 12: Decision Tree for SNPID= 2165047 and associated diseases

The SNPID “10618418” is associated with both diseases Crohn and Obesity but is correlated with Obesity shown in figure 13.

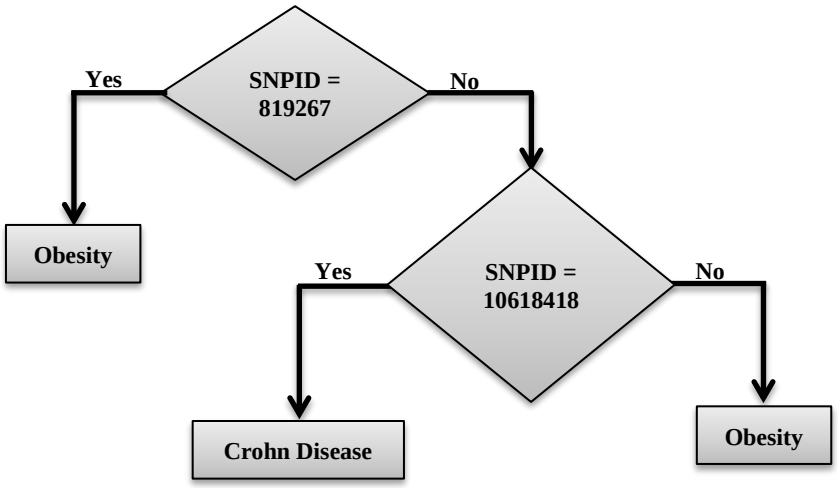

Fig13: Decision Tree for SNPID= 8192678 and associated diseases

The SNPID “4988235” is associated with both diseases Crohn and Obesity but is correlated with Crohn are Shown in figure 14.

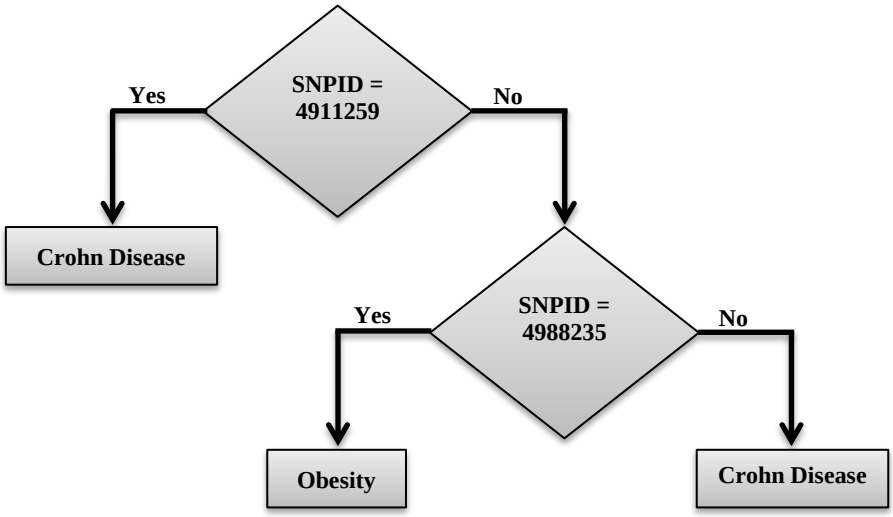

Fig 14: Decision Tree for SNPID= 4911259 and associated diseases

4.2.2 Naive Bayes based classifier

The Bayesian is Classification speaks to a managed learning technique and additionally a factual strategy for arrangement. It Accept a fundamental probabilistic model and it enables us to catch instability about the model principled by deciding probabilities of the results. It can solve diagnostic and predictive problems **[57].**

Table 4. Accuracy measure

Class	F-Measure	Recall	Precision	FP Rate	TP Rate
Obsessive Compulsive Disorder	0.109	0.058	1	0	0.058
Obesity	0.9	0.923	0.879	0.169	0.923
Inflammatory Bowel Diseases	0.185	0.106	0.735	0.005	0.106
Crohn Disease	0.764	0.882	0.674	0.176	0.882

The table 4**.** Shows Accuracy measure. Main trend of accuracy is Precision in Obsessive Compulsive Disorder, and low tendency is recall in Obsessive Compulsive Disorder. Obsessive declines sharply as in Precision 1 and in recall 0.058 and Obesity is trend increases gradually as in Precision 0.879 and recall 0.923. Inflammatory Bowel Diseases decreases sharply as 0.735 in Precision, 0.106 in recall, and Crohn Disease increases as 0.674 in Precision, 0.882 in recall.

The margin curve of the NB tree classifer with the increase of X, Y stcady linearly increases, until X reach the value -1. 2858 quadratically increase are Shown in figure 15.

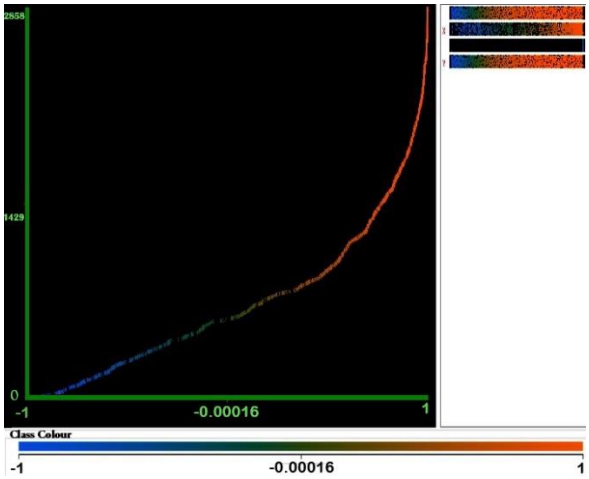

Fig 15: Margin curve of the Naive Bayes classifier

Table 5. Comparison between methods

Methods	TP rate
SVM[50]	63.6%
LP	69.5%
RF	66.1%
CGSP	67.3%
CSP	76.1%
Proposed 1(C4.5-based)	81.7 %
Proposed 2(Naive Bayes-based)	67.4 %

The table 5. Shows Comparison of performance between methods Naïve Bayes, Decision Tree, SVM, LP, RF, CGSP and CSP for Crohn disease where it turns out that the result of the two methods the proposal is better than the previous work.

5. CONCLUSION AND FUTURE WORK

As shown, the system is able to answer the core genes affecting certain diseases. What are the genes associated to Alzheimer Disease? What are the genes that support the association? Future direction may consider proteins networks. Proposed methods such as applying C4.5 & naïve Bayes give better accuracy then previous works

One future direction may be scrutinizing specific disorders such as Copper Related Disorders **[58, 59]** and DNA-Based Nano biosensors as an Emerging Platform for Detection of Disease **[60].** Another future work direction is trying different techniques such as protein network analysis.

Table 6. The ratio between Major, Minor, Confidence for diseases name

Diseases Name	Major	Minor	Confidence
Alzheimer's-disease	0.7	0.1	0.8
Grave's Disease	0.59	0.01	0.6
Behcet's disease	0.44	0.35	0.79
brain cancer	0.68	0.22	0.9
breast cancer	0.58	0.12	0.7
celiac disease	0.39	0.01	0.4
Chronic kidney failure	0.39	0.37	0.76

Fig 16. Shows stacked area. It shows major &minor denes contributed to each disease

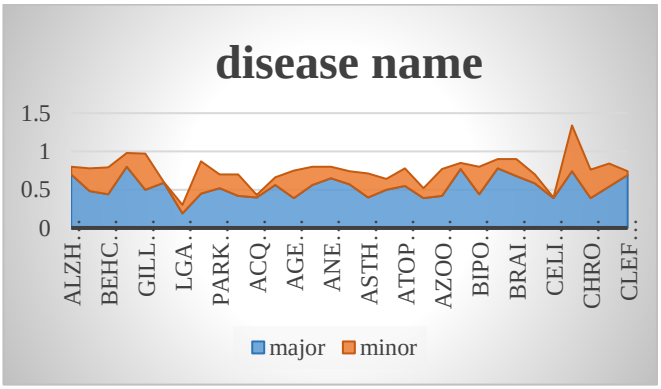

Fig 16: Stacked area of the ratio between major, minor for diseases name.

Fig 17. Shows Clustered bar chart. It shows major &minor denes contributed to each disease

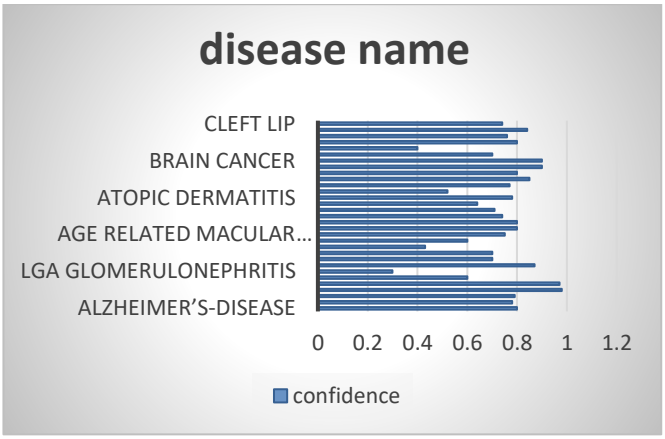

Fig 17: Clustered bar chart of the ratio confidence for diseases name

6. REFERENCES

[1] DumitruBrinza, JINGWU He, and Alexander Zelikovsky, 2008. Optimization Methods for Genotype Data Analysis in Epidemiological Studies, Bioinformatics Algorithms: Techniques and Applications, Edited by Ion I. M, andoiu and Alexander Zelikovsky, JOHN Wiley & Sons, Inc, pp 395.

[2] Janet Piñero, NúriaQueralt-Rosinach, Àlex Bravo, JordiDeu-Pons, Anna Bauer-Mehren, Martin Baron, FerranSanz, and Laura I. Furlong, 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes

[3] Brinza D, Zelikovsky A. 2006. Combinatorial Methods for Disease Association Search and Susceptibility Prediction. Proceedings of the Sixth Workshop on Algorithms in Bioinformatics (WABI 2006); pp. 286–297.

[4] Davis, A.P., Murphy, C.G., Johnson, R. et al. 2013. The comparative toxic genomics database: update 2013. Nucleic Acids Res., 41, D1104–D1114.

[5] Becker, K.G., Barnes, K.C., Bright, T.J. et al. 2004. The genetic association database. Nat. Genet. 36, 431–432.

[6] Bauer-Mehren, A., Rautschka, M., Sanz, F. et al. 2010. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics, 26, 2924–2926.

[7] Bauer-Mehren, A., Bundschus, M., Rautschka, M. et al. 2011. Gene-disease network analysis reveals functional modules in Mendelian, complex and environmental diseases. PloS One, 6, e20284.

[8] Bravo, A., Pinero, J., Queralt, N. et al. 2015. Extraction of Relations between Genes and Diseases from Text and Large-scale Data Analysis: Implications for Translational Research. Cold Spring Harbor Labs Journals, 007443. BMC Bioinformatics, 16:55. Doi: 10.1186/s12859-015-0472-9.

[9] The UniProt Consortium. 2014. Activities at the universal protein resource (UniProt). Nucleic Acids Res. 42, D191–D198.

[10] Laulederkind, S.J.F., Hayman, G.T., Wang, S.-J. et al. 2013. The Rat Genome Database 2013—data, tools and users. Brief. Bioinform, 14, 520–526.

[11] Blake, J.A., Bult, C.J., Eppig, J.T. et al. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. 42, D810–D817.

[12] Mitchell, J.A., Aronson, A.R., Mork, J.G. et al. 2003. Gene indexing characterization and analysis of NLM's GeneRIFs. AMIA Annu. Symp. Proc. 460–464.

[13] Bundschus, M., Dejori, M., Stetter, M. et al.2008 "Extraction of semantic biomedical relations from text using conditional random fields". BMC Bioinformatics, 9, 207.

[14] M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt..1978 .A model of evolutionary change in proteins. In Atlas of protein sequence and structure, (M. Dayhoff, ed.). National Biomedical Research Foundation, Washington, D.C., Volume 5 Pages 345-352.

[15] S. Henikoff, J. G. Henikoff.1992. Amino acid substitution matrices from protein blocks, ProcNatlAcadSci U S A, 89(22):10915-9.

[16] S. B. Needleman and C. D. Wunsch.1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. vol.48, pp.443-453.

[17] T. F., M. S. Waterman.2008. Identification of common molecular subsequences, J MolBiol 147, 195-197.10.

[18] Gotoh.1982. an Improved Algorithm for Matching Biological Sequences, J MolBiol 162, 705-8.

[19] W. R. Pearson, D. J. Lipman.1998. , Improved tools for biological sequence comparison, ProcNatlAcadSci U S A, 85: 2444-2448.

[20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman.1990. "Basic local alignment search tool". J MolBiol, 215 (3): 403.

[21] J P Dumas and J Ninio. 1982 "Efficient algorithms for folding and comparing nucleic acid sequences". Nucleic Acids Res 10: 197-206.

[22] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25(17):3389-402.

[23] J. D. Thompson, D. G. Higgins, T. J. Gibson. . 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucl Acid Res, 4673-4680.

[24] N. Saitou and M. Nei. 1987 .The neighbor-joining Method: A new method for reconstructing hylogenetic tree. MolBiolEvol. 4: 406-425.

[25] M. Gribskov, A. D. McLachlan, D. Eisenberg. 1987. Profile analysis: detection of distantly related proteins, ProcNatlAcadSci U S A. Jul; 84(13):4355-8.

[26] L. R. Rabiner, 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. Of the IEEE, Vol.77, No.2, pp.257—286.

[27] A Panchenko, 2003. Finding weak similarities between proteins by sequence profile comparison, Nucleic Acids Research, Vol. 31, No. 2 683-689.

[28] R. Edgar, K Sjolander.2004. A comparison of scoring functions for protein sequence profile alignment, Bioinformatics 20, 1301-1308.

[29] G Wang, R. Dunbrack, Jr., 2004 scoring profile-to-profile sequence alignments, Protein Sci. 13: 1612-1626.

[30] Y. Zhang, J. Skolnick. 2004. A scoring function for the automated assessment of protein structure template quality, Proteins, vol 57, 702.

[31] L. Holm, C. Sander. 1993 Protein structure comparison by alignment of distance matrices, J MolBiol, 233: 123-28 24.

[32] I.N. Shindyalov, P. E.Bourne, 1998. Protein structure alignment by incremental combinatorial extension (CE) of optimal path, ProtEng, 11 739-747.

[33] Y. Zhang, J. Skolnick. 2005.TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Research, vol 33, 2302.

[34] W. Kabsch, S. 1983 Sander: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577-637.

[35] P. Y. Chou, G. Fasman, 1974. Prediction of protein conformation, Biochemistry 13:222-245.

[36] B. Rost, C. Sander, 1993. Prediction of protein secondary structure at better than 70% Accuracy, Journal of Molecular Biology, 232, 584-599.

[37] D. Jones.1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol, 292: 195-202.

[38] J. U. Bowie, R. Luthy, D. Eisenberg. 1991. A method to identify protein sequences that fold into a known three-dimensional structure, Science. 253:164-170.

[39] J Soding. 2005. Protein homology detection by HMM-HMM comparison. 21: 951-960.

[40] S Wu, Y Zhang. 2008. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Structure, Function, and Bioinformatics, 72: 547-556.

[41] K. T. Simons, C. Kooperberg, E. Huang, D. Baker. 1997 "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions", J Mol Biol. 268(1):209-25.

[42] Y. Zhang, J. Skolnick. , 2004. Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of USA vol 101, 7594.

[43] Y. Zhang. 2014. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10 Proteins. 82: 175-187.

[44] Kryshtafovych A1, Fidelis K, Moult J. 2004 CASP10 results compared to those of previous CASP experiments. Proteins. 82:164-74.

[45] R. H. Swendsen, J. S. Wang. 1986. Replica Monte Carlo Simulation of Spin-Glasses. Phys. Rev. Lett. 57: 2607-2609.

[46] S. Kirkpatrick, C. D. GelattJr, M. P. Vecchi. 1983 .Optimization by Simulated Annealing. Science, 220: 671–680.

[47] Mafalda Galhardo1, Philipp Berninger2, Thanh-Phuong Nguyen1, Thomas Sauter1 and Lasse Sinkkonen1.2015. Nucleic Acids Research Advance Access.

[48] Matthaei, J.H., Jones, O.W., Martin, R.G. and Nirenberg, M.W. 1962 "Characteristics and composition of RNA coding units", Proc. NatlAcadSci U. S. A, 48, 666-677.

[49] Nirenberg, M.W. 1963, "The Genetic Code: II. The Molecular Basis of Life - An Introduction to Molecular Biology", pp. 206-216.

[50] Gusfield, Dan. 1999, "Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology". USA: Cambridge University Press.

[51] DANIEL S. HIRSCHBERG, 1977, "Journal of the Association for Computing Machinery", Vol 24, No 4.

[52] Tom Madden, 2001, "The BLAST Sequence Analysis Tool, The NCBI Handbook", Chapter 16

[53] Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. 2004, "Data mining in bioinformatics using Weka". Bioinformatics 20,2479 –248

[54] ShangweiNing†, Zuxianglan Zhao†, Jingrun Ye†, Peng Wang, HuiZhi, Ronghong Li, Tingting Wang and Xia Li, 2014," LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs", Ning et al. BMC Bioinformatics, 15:152.

[55] J. Han and M. Kamber, 2001, Data Mining Concepts and Techniques. Morgan-Kaufmann Publishers, San Francisco.

[56] O. Maimon and L. Rokach, Data Mining and Knowledge Discovery. Springer Science and Business Media, 2005.

[57] D. Xhemali, C. J. Hinde, and R. G. Stone, 2009, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," International Journal of Computer Science Issue, Vol. 4(1).

[58] Janet Pin˜ ero, A` lex Bravo, Nuria Queralt-Rosinach, Alba Gutierrez-Sacristan, Jordi-Pons, Emilio Centeno, Javier Garc´ıa-Garc´ıa, FerranSanz and Laura I. Furlong*,2016. Nucleic Acids Research Advance Access, doi: 10.1093/nar/gkw943.

[59] Yulia A. Shuvalovaa, , , Zukhra B. Khasanovab, Violetta I. Kaminnayaa, Elena V. Samoilovac, Alexandra A. Korotaevac, Alexander V. Rubanovichd, Alexander I. Kaminnyia, 2015. The association of PLA2G2A single nucleotide polymorphisms with type IIa secretory phospholipase A2 level but not its activity in patients with stable coronary heart disease, Gene, Volume 564, Issue 1, Pages 29-34 .

[60] Dumontier.M. Baker, C.J., Baran.J. et al. 2014. The Semantic science Integrated Ontology (SIO) for biomedical research and knowledge discovery. J. Biomed. Semantics, 5, 14.