A Thesis Submitted to

MAHATMA GANDHI MISSION'S

Institute of Biosciences and Technology

N-6,CIDCO, Aurangabad.

Year: 2021-2022

PROJECT REPORT ON

"VARIANT PREDICTION SYSTEM DEVELOPED FOR THE NERVOUS SYSTEM DISEASES."

SUBMITTED BY

Mr. SAURABH SUNILRAO BOBADE

Research Project Guide**: Dr. Archana Panche**

Project Head**: Ms. Krutanjali Patil**

CANDIDATES DECLARATION

I wish to state that the work embodied in this project titled, **"VARIANT PREDICTION SYSTEM DEVELOPED FOR THE NERVOUS SYSTEM DISEASES."** forms our contribution to the research work carried out under the guidance of Dr. Archana Panche at the MGM's Institute of Biosciences and Technology, affiliated to MGM's University, Aurangabad. This work has not been submitted for any other degree of this or any other University. Wherever references have been made to previous works of others, it has been indicated as such and included in the References.

Signature of the Candidate

Mr. Saurabh Sunilrao Bobade

MBI202204

Certified By

Guide Name: **Dr. Archana Panche**

Affiliation: **MGM'S Institute of Biosciences & Technology.**

Date:

CERTIFICATE

The project work presented in this project has been carried out by **Mr. Saurabh Sunilrao Bobade** under my guidance and have completed as per the requirements of **MGM's University, Aurangabad** in partial fulfilment of the degree of **Master of Science (Bioinformatics)** for the academic year 2021-22. This constitutes their bonafide work. The project work done is original and has not been submitted for any other degree for this or another University. Further that they were regular students and have worked under my guidance at the Department of Bioinformatics until the submission of the thesis to the University.

Date:

Place: Aurangabad

Guide	**Examiner**	**Director**
(Dr. Archana Panche)		**(Dr. Sanjay Harke)**

ACKNOWLEDGEMENT

The success of any project depends on the team work and also encouragement and guidelines of many others. I take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

To start with, I would like very much thankful to then **Director of MGM's Institute of Biosciences & Technology, Aurangabad, Dr. Sanjay Harke** for granting us the opportunity to work in the laboratories of this institute and supporting me during whole curriculum.

I would like to express our sincere gratitude to our Project Guide **Dr. Archana Panche** & Project Head **Ms. Krutanjali Patil , Department at MGM's Institute of Biosciences & Technology, Aurangabad** who offered me guidance and support all along the completion of the project.

I am thankful to all members of internal monitoring committee for giving all guidance for completion of project. I express my sincere thanks to all academic staff and non-teaching staff of this college, for their kind help and support during the project work. I also grateful to the authors for past and present whose contribution were great help to undertaken this investigation.

Last but not least, I would like thanks to our parents and friends who have been a source of motivation and enthusiasm and without whom I would have never been able to reach this height.

INDEX

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

DNA – Deoxyribonucleic acid

SNP – Single Nucleotide Polymorphism

SNV – Single Nucleotide Variants

SVM – Support Vector Machine

AD – Alzheimer's Disease

ClinVar – Database for the clinically associated SNVs

dbSNP – Database for SNPs

OMIM – Online Mendelian Inheritance in Man

ABSTRACT

The present experiment was conducted during the year 2021-2022 in MGM's Institute of Biosciences and Technology, Aurangabad with a view to develop a dataset of all SNPs associated with the genetic nervous system diseases .

I have taken 36 different genetically affected nervous system diseases caused by single nucleotide polymorphism. Total numbers of SNVs collected during this study are 1,68,472. This dataset is useful for the recognition of clinical significance of SNPs and can be used for the further study and develop a tool for the prediction clinical significance and association of unknown SNPs.

Keyword: SNP , SNV, clinical significance.

CHAPTER – 1

INTRODUCTION

CHAPTER – 1
INTRODUCTION

The sequence of the human genome is providing us with the first holistic view of our genetic heritage. The 46 human chromosomes (22 pairs of autosomal and 2 sex chromosomes) between them almost 3 billion base pairs of DNA that contains about 30,000-40,000 protein-coding genes. The coding regions make up less than 5% of the genome (the function of remaining DNA is not clear).

The Nervous System Diseases:

The brain and nervous system form an intricate network of electrical signals that are responsible for coordinating muscles, the senses, speech, memories, thought and emotions. Several neurodegenerative diseases that directly affect the nervous system have a genetic component: some are due to mutation in a single nucleotide. The pathogenesis of neurodegenerative disorders deepens, common themes begin to emerge: Alzheimer brain plaques and the inclusion bodies found in Parkinson disease contain at least one common component, while Huntington disease, fragile X syndrome and spinocerebellar atrophy are all 'dynamic mutation' diseases in which there is an expansion of a DNA repeat sequence. Apoptosis is emerging as one of the molecular mechanisms invoked in several neurodegenerative diseases, as are other, specific, intracellular signaling events. The biosynthesis of myelin and the regulation of cholesterol traffic also figure in Charcot-Marie-Tooth and Neimann-Pick disease, respectively.

Diseases :

1. Adrenoleukodystrophy
2. Alzheimer disease
3. Amyotrophic lateral sclerosis
4. Angelman syndrome
5. Ataxia telangiectasia
6. Charcot-Marie-Tooth syndrome
7. Cockayne syndrome
8. Deafness
9. Duchenne muscular dystrophy

10. Epilepsy
11. Essential tremor
12. Fragile X syndrome
13. Friedreich's ataxia
14. Gaucher disease
15. Huntington disease
16. Lesch-Nyhan syndrome
17. Maple syrup urine disease
18. Menkes syndrome
19. Myotonic dystrophy
20. Narcolepsy
21. Neurofibromatosis
22. Niemann-Pick disease
23. Parkinson disease
24. Phenylketonuria
25. Prader-Willi syndrome
26. Refsum disease
27. Rett syndrome
28. Spinal muscular atrophy
29. Spinocerebellar ataxia
30. Tangier disease
31. Tay-Sachs disease
32. Tuberous sclerosis
33. Von Hippel-Lindau syndrome
34. Williams syndrome
35. Wilson's disease
36. Zellweger syndrome

Single Nucleotide (variants) Polymorphism (SNPs):

In genetics, a single-nucleotide polymorphism is a germline substitution of single nucleotide at a specific position in the genome. A single nucleotide polymorphism, or SNP (pronounced "snip"), is a variation at a single position in a DNA sequence among individuals. Recall that the DNA sequence is formed from a chain of four nucleotide bases: A, C, G, and T. If more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP. If a SNP occurs within a gene, then the gene is described as having more than one allele. In these cases, SNPs may lead to variations in the amino acid sequence. SNPs, however, are not just associated with genes; they can also occur in noncoding regions of DNA.

Although a particular SNP may not cause a disorder, some SNPs are associated with certain diseases. These associations allow scientists to look for SNPs in order to evaluate an individual's genetic predisposition to develop a disease. In addition, if certain SNPs are known to be associated with a trait, then scientists may examine stretches of DNA near these SNPs in an attempt to identify the gene or genes responsible for the trait. The severity of illness and the way the body responds to treatments are also manifestations of genetic variations caused by SNPs. For example, a single-base mutation in the APOE (apolipoprotein E) gene is associated with a lower risk for Alzheimer's disease. A single-nucleotide variant (SNV) is a variation in a single nucleotide.

Machine Learning in Variant prediction:

Machine learning approaches adapt a set of sophisticated statistical and computational algorithms (e.g. Support vector machine (SVM) or Random Forest) to make predictions by mathematically mapping the complex associations between a set of risk SNPs to complex disease phenotypes. This methods use supervised or unsupervised approaches to map the associations with complex diseases. Machine learning variant prediction model will be generated by training the pre-set learning algorithms to map the relationship between individual sample genotype data and the associated diseases.

CHAPTER – 2

SCOPE OF PROJECT

CHAPTER – 2

SCOPE OF PROJECT

This project will be focusing on developing Dataset of SNPs respective to their neural diseases and algorithm development for detection of single nucleotide variants (SNVs).

This will help in the prediction of variants with respect to neural diseases with the help of machine learning algorithms such as logistic regression or support vector machine.

CHAPTER – 3

REVIEW OF LITERATURE

CHAPTER – 3

REVIEW OF LITERATURE

An attempted has been made in this chapter to review the research work done in paston the aspect of present by various scientists in India and Abroad.

Hofmann et al. (2015), issued a review paper "Bioinformatics mining and modeling for identification of disease mechanism in neurodegenerative disorders". They describe a panel of bioinformatics and modelling approaches that have recently been developed to identify candidate mechanisms of neurodegenerative diseases based on publicly available data and knowledge. They identify two complementary strategies—data mining techniques using genetic data as a starting point to be further enriched using other data-types, or alternatively to encode prior knowledge about disease mechanisms in a model based framework supporting reasoning and enrichment analysis. They conclude that progress would be accelerated by increasing efforts on performing systematic collection of multiple data-types over time from each individual suffering from neurodegenerative disease.

Ho et al. (2019) published a review article "Machine learning SNP based prediction for precision medicine". They provide an overview of polygenic risk scoring and machine learning in complex disease risk prediction. They discuss how the future application of machine learning prediction models might help manage complex disease by providing tissue-specific targets for customized, preventive interventions.

Mishra and Li (2020) proposed that in addition to conventional statistical methods for the processing of genetic data of Alzheimer's disease (AD), artificial intelligence (AI) technology shows obvious advantages in analyzing such complex projects, in their review paper "The application of artificial intelligence in the genetic study of Alzheimer's Disease". This article briefly revives the application of AI technology in medicine and the current status of genetic research in AD.

Rangaswamy et at. (2020), in their paper “VEPAD - Predicting the effect of variants associated with Alzheimer's disease using machine learning” stated that Next generation Sequencing (NGS) techniques are widely used for developing high-throughput screening methods to identify biomarkers and variants , which help early diagnosis and treatments. They developed a classification model using machine learning for predicting the deleterious effect of variants with respect to AD.

Monk et al. (2021), described that there is hope that genomic information will assist prediction, treatment and understanding of Alzheimer’s disease (AD) in “A machine learning method to identify genetic variants potentially associated with Alzheimer’s disease ”. They used exome data from ~10,000 individuals , and explored machine learning neural network(NN) methods to estimate the impact of SNPs(i.e. genetic variants) on AD risk . They developed NN-based method (netSNP) that identifies hundreds of novel potentially protective or at-risk AD-associated SNPs (along with an effect measure); the majority with frequency under 0.01.

CHAPTER - 4

OBJECTIVES

CHAPTER - 4
OBJECTIVES

The project research work will be conducted with following objectives-

- Retrieval of SNVs from SNP databases or ClinVar Database respective to the nervous system diseases.
- Data processing of SNVs to create a training data.
- Development of Machine learning algorithm for variant prediction.

CHAPTER – 5

MATERIALS AND METHODOLOGY

CHAPTER – 5

MATERIALS AND METHODS

The details of various material and methods will be conducting during the course of present investigation are narrated in this chapter under suitable sub-heads.

Database required :

- ClinVar
- dbSNP
- OMIM

To create dataset :

- MS Excel
- MySQL

Programming languages :

- SQL
- Machine learning in python

Collection Of Data :

For the Collection of Data , I had used ClinVar Database . Through the ClinVar database we can extract the SNP data for genetic diseases which are associated with the mutation in DNA. The SNP data is extracted and then converted into excel format for better view and for further analysis. Total 1,68,472 SNPs are extracted from ClinVar for 36 diseases. The dataset contains of several columns such as Name of SNV , Gene , Protein Change , Position of SNV , Chromosome, Condition , Clinical significance , Type of mutation, etc.

Creation of training dataset and analysis of Data :

As the Dataset has huge number of SNPs , I have created a training Dataset with 50 SNVs of each diseases both pathogenic and benign variants. The training dataset has total number of 1450 SNVs for all 36 diseases . This training dataset is curated and I have excluded some columns .

The final training dataset contains columns as follows :

1. Name of SNV
2. Gene
3. Condition
4. Clinical Significance
5. Chromosome as per GRCH38
6. Location of SNV as per GRCH38
7. Type of Mutation

Name	Gene(s)	Condition(s)	Clinical_significance	GRCh38Chromosome	GRCh38Location	Canonical SPDI	Type of Mutation
NM_000033.4(ABCD1):c.1438C>A (p.Pro480Thr)	ABCD1	Adrenoleukodystrophy	Benign	X	153737201	NC_000023.11:153737200:C:A	Inversion
NM_000484.4(APP):c.663-7T>C	APP	Alzheimer disease	Benign	21	26022049	NC_000021.9:26022048:A:G	Inversion
NM_020919.4(ALS2):c.4123-64G>A	ALS2	Amyotrophic lateral sclerosis type 2	Benign	2	201710102	NC_000002.12:201710101:C:T	Inversion
NM_001323289.2(CDKL5):c.463+8A>G	CDKL5	Angelman syndrome-like	Benign	X	18581958	NC_000023.11:18581957:A:G	Inversion
NM_001195248.2(APTX):c.874+84G>A	APTX	Ataxia-oculomotor apraxia type 1	Benign	9	32974374	NC_000009.12:32974373:C:T	Inversion
NM_001605.3(AARS1):c.1009G>A (p.Glu337Lys)	AARS1	Charcot-Marie-Tooth disease	Pathogenic	16	70268333	NC_000016.10:70268332:C:T	Inversion
NM_005236.3(ERCC4):c.1698G>A (p.Leu566=)	ERCC4	Cockayne syndrome	Benign	16	13935630	NC_000016.10:13935629:G:A	Inversion
NM_001199107.2(TBC1D24):c.965+1G>A	TBC1D24	Deafness	Pathogenic	16	2497114	NC_000016.10:2497113:G:A	Inversion
NM_004006.3(DMD):c.9975-1G>T	DMD	Duchenne muscular dystrophy	Pathogenic	X	31180482	NC_000023.11:31180481:C:A	Inversion
NM_001182.5(ALDH7A1):c.130G>T (p.Glu44Ter)	ALDH7A1	Pyridoxine-dependent epilepsy	Pathogenic	5	126595069	NC_000005.10:126595068:C:A	Inversion
NM_000796.6(DRD3):c.112G>A (p.Ala38Thr)	DRD3	Hereditary essential tremor	Benign	3	114171881	NC_000003.12:114171880:C:T	Inversion

NM_002025.4(AFF2):c.1016A>T (p.Lys339Met)	AFF2	Fragile X syndrome	Benign	X	148662743	NC_000023.11:148662742:A:T	Inversion
NM_000144.5(FXN):c.517T>G (p.Trp173Gly)	FXN	Friedreich ataxia	Pathogenic	9	69072646	NC_000009.12:69072645:T:G	Inversion
NM_000157.4(GBA):c.898G>A (p.Ala300Thr)	GBA|LOC106627981	Gaucher disease	Pathogenic	1	155237442	NC_000001.11:155237441:C:T	Inversion
NM_000447.3(PSEN2):c.448G>A (p.Val150Met)	PSEN2	Huntington disease	Pathogenic	1	226885629	NC_000001.11:226885628:G:A	Inversion
NM_000194.3(HPRT1):c.486-11G>A	HPRT1	Lesch-Nyhan syndrome	benign	X	134498379	NC_000023.11:134498378:G:A	Inversion
NM_000709.4(BCKDHA):c.884C>A (p.Ser295Ter)	BCKDHA	Maple syrup urine disease	Pathogenic	19	41422659	NC_000019.10:41422658:C:A	Inversion
NM_000052.7(ATP7A):c.3943G>A (p.Gly1315Arg)	ATP7A	Menkes kinky-hair syndrome	Pathogenic	X	78042726	NC_000023.11:78042725:G:A	Inversion
NM_003418.5(CNBP):c.61A>G (p.Thr21Ala)	CNBP	Myotonic dystrophy	Pathogenic	3	129171697	NC_000003.12:129171696:T:C	Inversion
NM_001130823.3(DNMT1):c.1814G>C (p.Gly605Ala)	DNMT1	Narcolepsy	Pathogenic	19	10154604	NC_000019.10:10154603:C:G	Inversion
NM_001042492.3(NF1):c.-272G>A	LOC111811965|NF1	Neurofibromatosis, type 1	Pathogenic	17	31095038	NC_000017.11:31095037:G:A	Inversion
NM_000271.5(NPC1):c.2300C>G (p.Ala767Gly)	NPC1	Niemann-Pick disease type C1	Benign	18	23541379	NC_000018.10:23541378:G:C	Inversion
NM_001256864.2(DNAJC6):c.1831G>A (p.Ala611Thr)	DNAJC6	Parkinson disease 19a, juvenile-onset	Pathogenic	1	65392793	NC_000001.11:65392792:G:A	Inversion
NM_000277.3(PAH):c.544G>A (p.Glu182Lys)	PAH	Phenylketonuria	Pathogenic	12	102855298	NC_000012.12:102855297:C:T	Inversion
NM_004667.6(HERC2):c.5546A>G (p.Lys1849Arg)	HERC2	Prader-Willi syndrome	Pathogenic	15	28222134	NC_000015.10:28222133:T:C	Inversion
NM_006214.4(PHYH):c.135-1G>C	PHYH	Refsum disease, adult, 1	Pathogenic	10	13295607	NC_000010.11:13295606:C:G	Inversion
NM_005249.5(FOXG1):c.173C>T (p.Pro58Leu)	FOXG1	Rett syndrome	Pathogenic	14	28767452	NC_000014.9:28767451:C:T	Inversion
NM_001003800.2(BICD2):c.404C>T (p.Thr135Met)	BICD2	Spinal muscular atrophy	Benign	9	92729073	NC_000009.12:92729072:G:A	Inversion

NM_006796.3(AFG3L2):c.753-55T>C	AFG3L2	Spinocerebellar ataxia type 28	Benign	18	12358998	NC_000018.10:12358997:A:G	Inversion
NM_005502.4(ABCA1):c.2660G>T (p.Cys887Phe)	ABCA1	Tangier disease	Pathogenic	9	104822664	NC_000009.1:104822663:C:A	Inversion
NM_000405.5(GM2A):c.*227A>G	GM2A	Tay-Sachs disease, variant AB	Benign	5	151267678	NC_000005.10:151267677:A:G	Inversion
NM_000368.5(TSC1):c.2691G>T (p.Gln897His)	TSC1	Tuberous sclerosis 1	Benign	9	132897545	NC_000009.12:132897544:C:A	Inversion
NM_000551.4(VHL):c.340+821C>G	LOC107303340|VHL	Von Hippel-Lindau syndrome	Benign	3	10143008	NC_000003.12:10143007:C:G	Inversion
NM_000501.4(ELN):c.1675G>A (p.Val559Ile)	ELN|ELN-AS1	Williams syndrome	Benign	7	74060429	NC_000007.14:74060428:G:A	Inversion
NM_000053.4(ATP7B):c.3971A>G (p.Asn1324Ser)	ATP7B	Wilson disease	Pathogenic	13	51937326	NC_000013.11:51937325:T:C	Inversion
NM_000466.3(PEX1):c.3439-16A>G	GATAD1|PEX1	Zellweger syndrome	Benign	7	92489927	NC_000007.14:92489926:T:C	Inversion

Table no 1 – Sample of training dataset

Methodology:

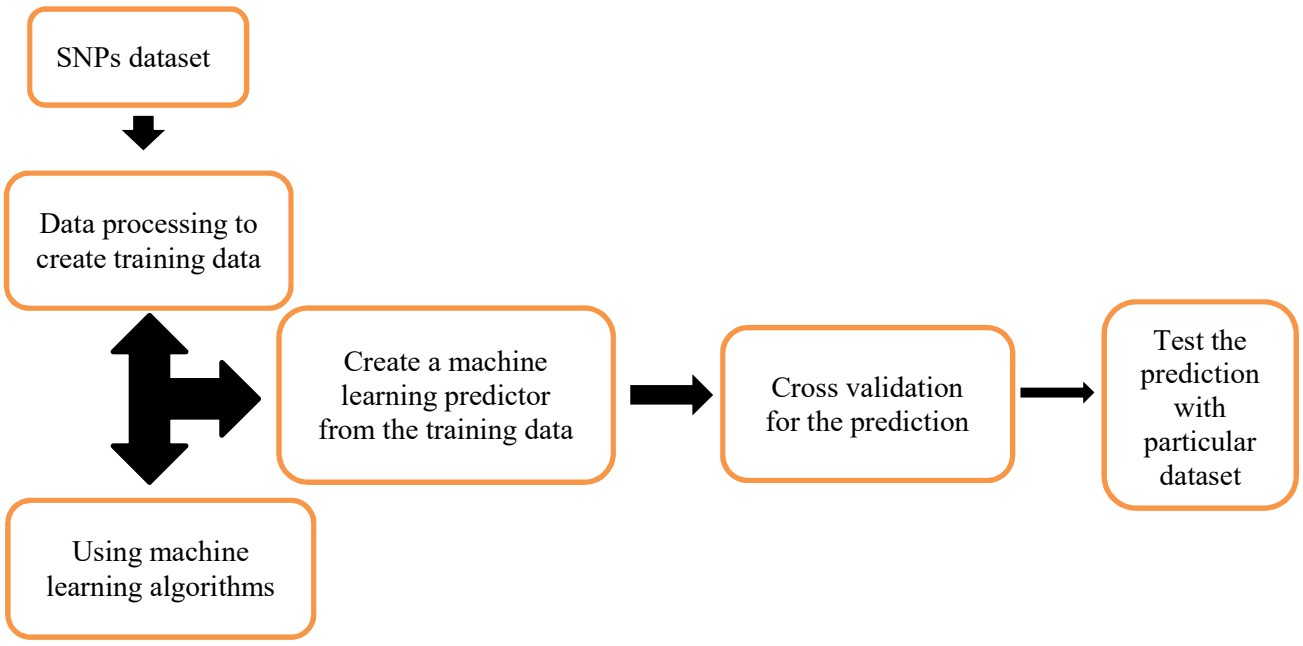

Fig 1 : Work flow for creating a machine learning model from a SNPs (Variants) dataset

CHAPTER – 6
RESULT AND DISCUSSION

CHAPTER – 6

RESULT

After the Collection of SNP data , I have prepared a training dataset of 50 variants from each disease and proceed further procedure on this dataset. This dataset contains of both pathogenic and benign variants of diseases. In Excel , I have used the analysis method in which we can graph for the present data. This method is used for the analysis and displaying the sorted way as pathogenic and benign significance , which can be helpful for the further procedure . Through the analysis of training dataset using the MS Excel we get the output as a histogram defining the numbers of pathogenic and numbers of benign variants present as per every 36 diseases.

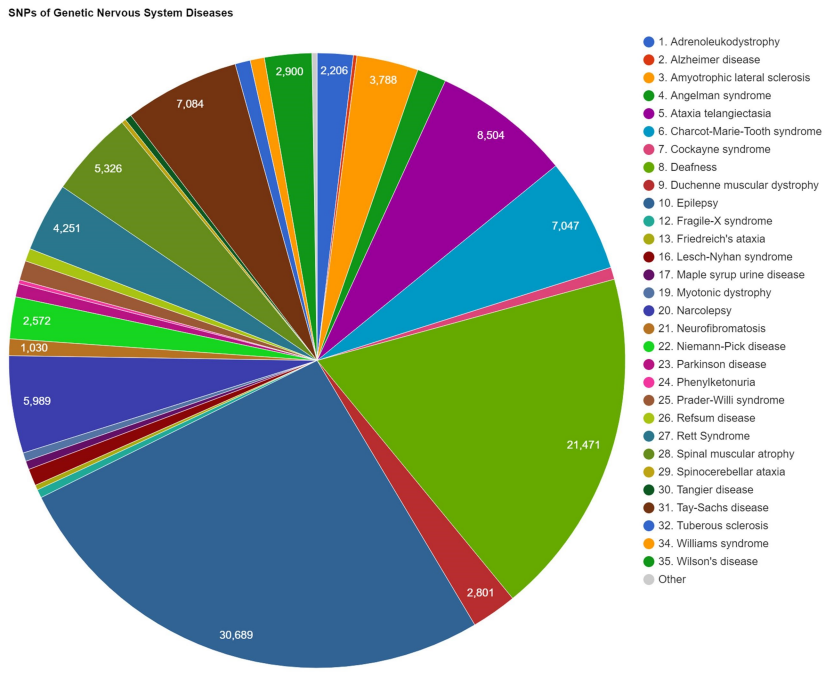

Fig 2 - SNPs collected from ClinVar database for diseases.

The histogram for the training dataset is as follows :

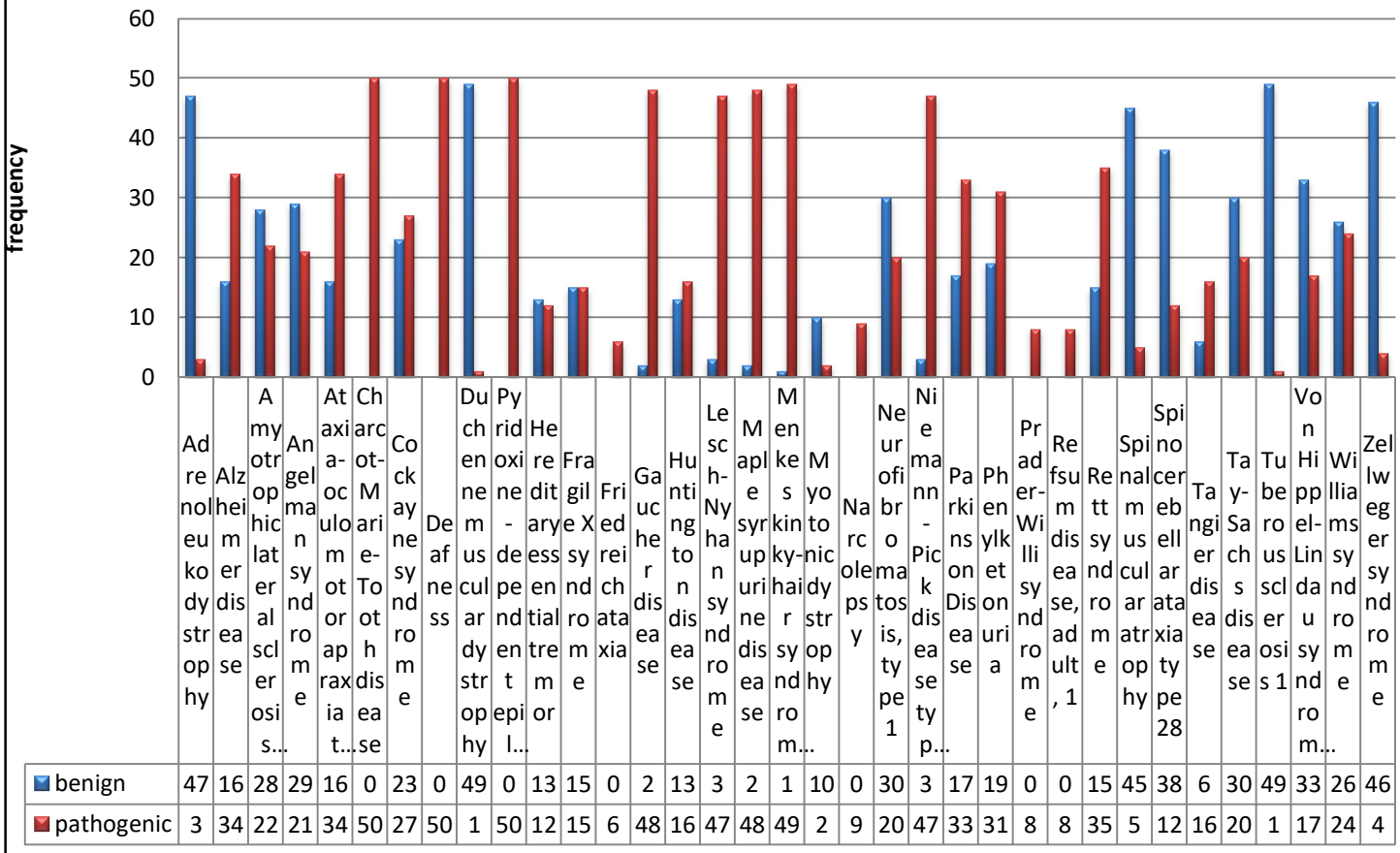

Fig 3 : Analysis of the training dataset

DISCUSSION

The project carried on **"VARIANT PREDICTION SYSTEM DEVELOPED FOR THE NERVOUS SYSTEM DISEASES"** is really helpful for noticing the current single nucleotide variants associated with the genetic nervous system diseases takes in these study.

In the further aspect of this stud we can develop a machine learning algorithm such as logistic regression or support vector machine to predict the clinical significance of any SNP on the basis of this dataset.

CHAPTER – 7

SUMMARY AND CONCLUSION

CHAPTER – 7
SUMMARY AND CONCLUSION

SUMMARY

The summary of the project conducted on "VARIANT PREDICTION SYSTEM DEVELOPED FOR THE NERVOUS SYSTEM DISEASES" is that , there are several diseases which are caused by the mutation in single nucleotide and can be pathogenic in the future for the human beings. The all SNPs collected in this project as a dataset can be useful for he further study for recognizing the clinical significance of those SNPs .

CONCLUSION

The summary of the project conducted on "VARIANT PREDICTION SYSTEM DEVELOPED FOR THE NERVOUS SYSTEM DISEASES" is stated as below :

1. There are vast number of single nucleotide variants presents for this genetic nervous system diseases.
2. These all SNPs collected in this project as a dataset can be useful for he further study for recognizing the clinical significance of those SNPs , either those are pathogenic or benign for certain human being.

CHAPTER – 8

RERFERENCES

CHAPTER – 8
RERFERENCES

Gracia-Fonseca, A. , Martin-Jimenez, C. , Barreto, G.E. , Pachon A.F.A. & Gonzalez, J. (2021). The Emerging Role of Long Non-Coding RNAs and MicroRNAs in Neurodegenerative Diseases: A Perspective of machine learning , *Biomolecules , 11(8)* , 1132 .
https://doi.org/10.3390/biom11081132

Ho, D.S., Schierding, W.S., Wake, M., Saffery, R. & O'Sullivan, J.M. (2019). Machine learning SNP based prediction for precision medicine , *Frontiers in Genetics , 10* , 267 .
https://doi.org/10.3389/fgene.2019.00267

Hofmann-Apitus, M. , Ball, G. , Gebel, S. , Bagewadi, S. , Bono, B. , Schenider, R. , Page, M. , Kodamullil, A.T. , Yonesi, E. , Ebeling, C. , Tenger, J. & Canard, L. (2015). Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders , *International Journal of Molecular Science , 16(12)* , 29179-29206.
https://doi.org/10.3390/ijms161226148

Hung, C., Chen, Y., Hsieh, W., Chiou, S. & Kao, C. (2010). Ageing and neurodegenerative diseases , *Ageing Research Revives. , 9*, S36-S46.
https://doi.org/10.1016/j.arr.2010.08.006

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype, *Nucleic acids research*, *42*, D980–D985.
https://doi.org/10.1093/nar/gkt1113

Mishra, R., & Li, B. (2020). The Application of Artificial Intelligence in the Genetic Study of Alzheimer's disease. *Aging and disease*, *11*(6), 1567–1584.
https://doi.org/10.14336/AD.2020.0312

Monk, B. , Rajkovic, A. , Petrus, S. , Rajkovic, A. , Gaasterland, T. & Malinow, R. (2021), A machine learning method to identify genetic variants potentially associated with alzheimer's disease, *Frontiers in Genetics ,12* , 647436 .
https://doi.org/10.3389/fgene.2021.647436

Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M.B., Neil, D., Saffari, A., Mead, R., Hautbergue, G.M., Holbrook, J.D. & Ferraiulo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases . *Nature Reviews Neurology, 16,* 440–456.
https://doi.org/10.1038/s41582-020-0377-8

Pihlstrom, L., Wiethoff, S. & Houlden, H., (2017). Genetics of neurodegenerative diseases: an overview, *Handbook of Clinical Neurology, 145,* 309-323.
https://doi.org/10.1016/B978-0-12-802395-2.00022-5

Rangaswamy, U., Dharshini, S., Yesudhas, D., & Gromiha, M. M. (2020). VEPAD - Predicting the effect of variants associated with Alzheimer's disease using machine learning. *Computers in biology and medicine*, *124,* 103933.
https://doi.org/10.1016/j.compbiomed.2020.103933

Varma, M., Paskov, K.M., Jung, J., Chrisman, B.S., Stockham, N.T., Washington, P.Y. & Wall, D.P. (2019). Outgroup machine learning approach identifies single nucleotide variants in noncoding DNA associated with autism spectrum disorder*, Pacific Symposium on Biocomputing, 24* , 260-271.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417813

https://www.nature.com/scitable/definition/snp/ (13/05/2022)

https://www.ncbi.nlm.nih.gov/books/NBK22197/ (21/10/2021)