PREDICTION OF HEREDITARY DISEASES THROUGH DNA USING MACHINE LEARNING ALGORITHMS

Pyla Jyothi [1], Rajapudi Priyusha [2], Varri Rupavathi [3], Singampalli Harsha Sri[4], Tebelu Prameela Rani[5], Pureddy Bindhu Priyanka[6]

Department of Computer Science and Engineering, Vignan’s Institute of Engineering for Women, Visakhapatnam, AndhraPradesh, India.

ABSTRACT

The modern individual's lifestyle has undergone several modifications. This generation is more vulnerable to diseases than the older generation. The primary causes of illnesses' attacks are also poor eating habits and a lack of exercise. Genes play a role in many common diseases, including cancer, heart disease, diabetes, and high blood pressure. Some illnesses are regarded as hereditary, meaning they are transmitted from parents to offspring. Hereditary Disorders are caused by genetic mutations. The genes are to blame for this. Diseases can be anticipated early on or much earlier before their assault by comparing the DNA (Deoxyribonucleic Acid) sequence of children with the DNA (Deoxyribonucleic Acid) sequence of their parents. All diseases, to some extent, are influenced by genetics. Disease processes are influenced by variations in our DNA and variances in how that DNA operates (individually or in combination), as well as by the environment (which includes lifestyle). In this review, the genetic basis of human disease including single gene disorders, chromosomal abnormalities, epigenetics, cancer, and complex disorders is examined. It also considers how scientific knowledge and technological advancements can be used to provide patients with the best possible diagnosis, treatment, and care.

Keywords: DNA (deoxyribonucleic acid), Diseases, Genes, Hereditary.

INTRODUCTION

Heredity is the sum of all biological processes by which particular characteristics are transmitted from parents to their children. The concept of heredity is explained in two ways, species from generation to generation and the variation among individuals within a species. Every member of a species has a fixed number of genes unique to that species. It is this set of genes that gives the fidelity of the species. Among the individuals within a species, however, variations can occur in the form each gene takes, providing the genetic basis for the truth that no individuals (besides the same twins) are precisely identical. The set of genes that children inherit from both parents, a combination of the genetic material of each is called the organism’s genotype. The genotype is contrasted to the phenotype, that is, the organism’s outward appearance and the final development results of its genes. The phenotype consists of an organism’s physical structures, physiological processes, and behaviors. Although the genotype determines the extensive limits of the capabilities an organism can develop, the capabilities that without a doubt develop, i.e., the phenotype, depend upon complicated interactions between genes and their environment. The genotype remains constant throughout an organism’s lifetime. However, due to the fact the organism’s inner and outside environments change continuously, so does its phenotype. In carrying out genetic studies, it's far more critical to find out the degree to which the observable trait is due to the sample of genes inside the cells. Machine learning plays a major role in identifying hereditary diseases from genes and it provides efficient results. The pressure on man’s health is increasing based on various factors like the environment, lifestyle, etc. Nowadays, hereditary diseases are very common, and predicting them before and changing our lifestyle will give better results. This ensures that our future generations will not suffer that particular disease. We consider the medical history of our past generations and predict diseases like heart attack, diabetes, nerve weakness, hemophilia, hair issues, and cancer. This system predicts whether the disease will pass on or not. Based on some conditions, some people might not get affected, but some people might get affected. The main objective is to efficiently predict hereditary diseases from the dataset. Use numerous ml algorithms to construct prediction models, examine the accuracy and overall performance of those models. To increase the accuracy of the classification results.

The human body contains 46 chromosomes in total. The human body contains 46 chromosomes in total. These chromosomes are a combination of X and Y chromosomes and are found in the human body. Out of these chromosomes, half come from the mother and half from the father. One chromosome is stated to contain one DNA (Deoxyribonucleic Acid), which is composed of the molecular building blocks A, C, T, and G. Entire is a combination of the above 4 blocks in various combinations. The complete DNA, which is dispersed throughout all chromosomes, is used to create the human genome. The DNA that is present in the genome is neatly arranged into the units that are referred to as the genes. Human bodies are made of proteins, which are produced by genes. Human blood, hair, the heart, and skin are all made of protein. Blood sugar and heartbeat rate are two other things that protein regulates. The proteins help to maintain the body's metabolism. Approximately 6000 diseases have been found by scientists as being caused by incorrectly spelled genes. genetic sequences are collected from normal human beings and also from the human beings who are suffering with the genetic related diseases which are identified by the scientists.

LITERATURE SURVEY

[1] Z. Elyazghi, L. E. Yazouli, K. Sadki and F. Radouani, "ABI Base Recall: Automatic DNASequence Correction and End Trimming" [1] is the title of the paper. The goal of this paper is to create a program that can correct ambiguities in a DNA sequence automatically. It is possible to achieve ABI-based recall,good correction,and high accuracy.

[2] Chengye Zou , Xiaopeng Wei, Qiang Zhang , Chanjuan Liu, and Yuan Liu Analog DNA Strand Displacement Circuits for Equation Solving" [2].In this paper, the displacement of DNA strands is used to construct complex functional circuits, which are used in the solution of linear, quadratic, and simultaneous equations.

[3] S. Brignac,"Sequencing Support System: A Robotic System for DNA SampleProcessing," S.Brignac [3]. This paper's robotics system includes cycle sequencing, detection, and data collection and analysis.

[4] B.YimwadsanaandP.Artiwet,"OnOptimizingDNASequenceDesignforDNALogicANDCircuit. Each DNA strand must connect to at least one other strand in order to form a composite structure. To avoid the formation of unexpected structures, each DNA strand must be meticulously crafted.

[5] A.Bekkanti,V.S.K.P.Gunde,S.Itnal,G.ParasaandC.M.A.K.Z.Basha,"ComputerBasedClassification of Diseased Fruit Using K-Means and Support Vector Machine," [5].The primary goal ofthis project is to distinguish between diseased and non-diseased fruits using algorithms such as BPNN(Back Propagation Neural Network),SVM(Support Vector Machine),and PNN.

[6] Y. Yao, J. Ren, R. Bi and Q. Liu, "Computer Based Classification of Diseased Fruit UsingK-Means and Support Vector Machine," [5]. The primary goal of this project is to distinguish between diseased and non-diseased fruits using algorithms such as BPNN (Back Propagation Neural Network),SVM(Support Vector Machine),and PNN.

PROPOSED SYSTEM

The objective of our project is to predict hereditary diseases at an early stage or much earlier before their attack and can take precautions accordingly. Hereditary diseases, also known as inherited diseases or genetic disorders, are defined and categorized as being a set of genetic diseases that are caused by changes in one's genetic material (DNA). These diseases are then transmitted from generation to generation, or in other words, they are inherited from parents to their children. We have used some algorithms to check the accuracy. Those algorithms are Logistic Regression, K-Neighbor Classifier, Decision Tree Classifier, Random Forest Classifier.

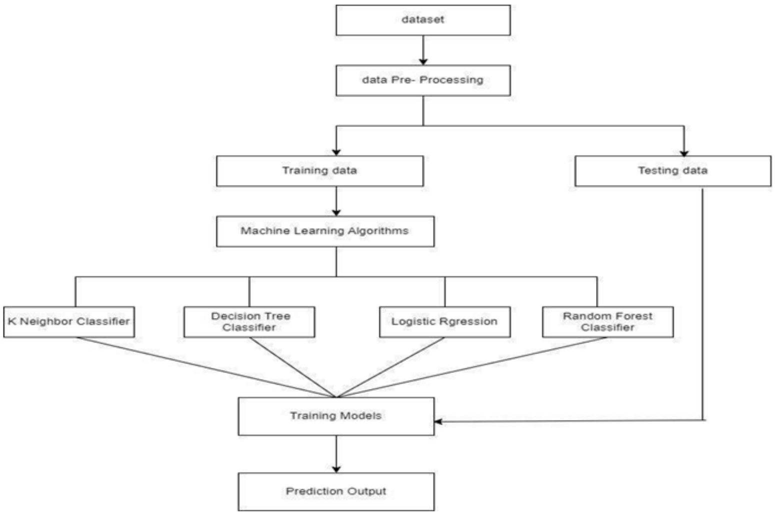


Fig 1: Proposed System Architecture

Machine learning models:

K-Nearest Neighbor Algorithm:

The k-nearest neighbors' classifier (KNN) is a non-parametric supervised machine learning algorithm. It's distance-based: it classifies objects based on their proximate neighbors' classes. KNN is most often used for classification, but can be applied to regression problems as well. The parameter k in KNN refers to the number of labeled points (neighbors) considered for classification. The value of k indicates the number of these points used to determine the result.

Decision Tree Classification Algorithm:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

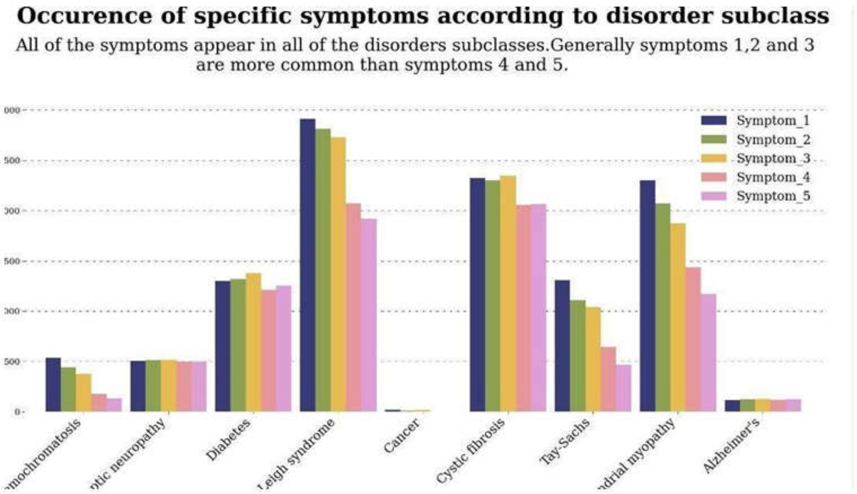
Logistic Regression Algorithm:

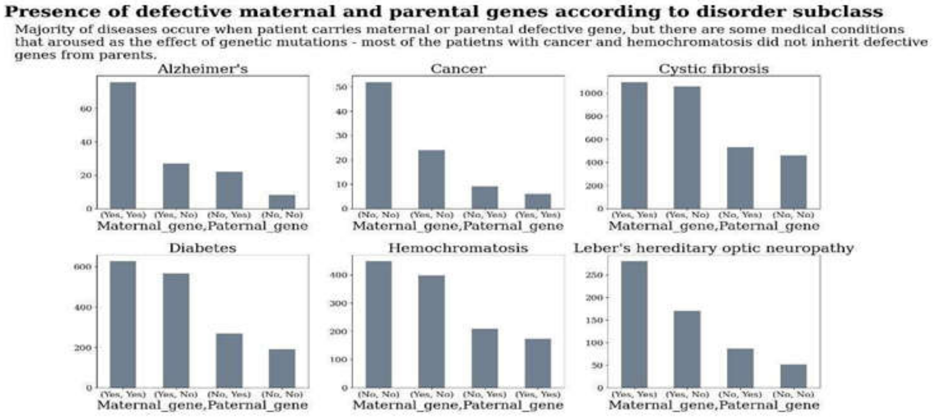
Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. Logistic Regression is much similar to Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

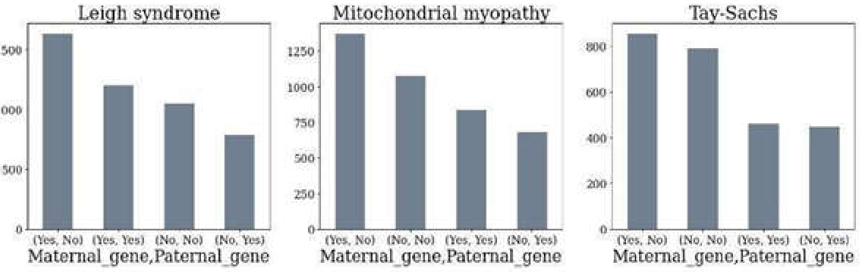
Random Forest Algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. Instead of relying on one decision tree the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output i.e which combines the output of multiple decision tree stor each a single result.

RESULTS AND DISCUSSIONS







COMPARISON OF THE MODELS

MODEL	ACCURACYSCORE
K Neighbor Classifier	0.3192982456140351
Decision Tree Classifier	0.3610341643582641
Logistic Regression	0.41865189289012006
Random Forest Classifier	0.38264081255771004

CONCLUSION

In this study, heredity is predicted using machine learning classifiers. The Pre-processing method isapplied to the hereditary data, which is used as input data. When using the pre- processing method, the dataset will be cleaned up and the label encoding will be applied. The dataset is divided into a training dataset and a testing dataset before being processed using a feature selection method. Finally, the hereditary in human gene data is predicted using classification method machine learning algorithm, and the results are found based on accuracy. With four algorithms, we have got more accuracy to logistic regression, and thescoreis0.41865189289012006.

FUTURE SCOPE

It is possible to provide extensions or modifications to the proposed clustering and classification algorithms in the future to achieve even higher performance. This project will be improved in the future by including many more hereditary diseases. It is also possible to use other machine learning algorithms and create simple real- world application for the user experience so that they can take precautions against hereditary diseases.

REFERENCES

1. Z. Elyazghi, L. E. Yazouli, K. Sadki and F. Radouani, "ABI BasERecall: Automatic Correctionand Ends Trimming of DNA Sequences," in IEEE Transactions on NanoBioscience, vol.16, no.8,pp.682-686,Dec.2017,doi:10.1109/TNB.2017.2755004.
2. Chengye Zou, Xiaopeng Wei, Qiang Zhang, Chanjuan Liu, and Yuan Liu,"Solution of EquationsBasedonAnalogDNAStrandDisplacementCircuits",IEEETRANSACTIONSONNANOBIOSCIENCE,VOL.18,NO.2,APRIL2019.
3. S. Brignac, "Sequencing Support System: A robotic system for processing DNA samples," inIEEE Engineering in Medicine and Biology Magazine, vol. 16, no. 2, pp. 92-93, March-April1997,doi:10.1109/51.582189.
4. B. Yimwadsana and P. Artiwet, "On Optimizing DNA Sequence Design for DNA Logic ANDCircuit,"TENCON2018-2018IEEERegion10Conference,2018,pp.1828-1833,doi:10.1109/TENCON.2018.8650528.
5. A. Bekkanti, V. S. R. K. P. Gunde, S. Itnal, G. Parasa and C. M. K. Z. Basha, "Computer BasedClassificationofDiseasedFruitusingK-MeansandSupportVectorMachine,"2020ThirdInternational Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1227-1232,doi:10.1109/ICSSIT48917.2020.9214177.
6. Y. Yao, J. Ren, R. Bi and Q. Liu, "Bacterial Foraging Algorithm Based on Activity of BacteriaforDNAComputing Sequence Design," in IEEE Access, vol. 9, pp. 2110-2124, 2021, doi:10.1109/ACCESS.2020.3047469.
7. K.A.S.ImminkandK.Cai,"PropertiesandconstructionsofconstrainedcodesforDNA-baseddatastorage,"IEEEAccess,vol.8,pp.49523–49531,March2020.
8. C. Y. Zou, X. Wei, Q. Zhang, C. Liu, and Y. Liu, "Solution of equations based on analog DNAstranddisplacementcircuits,"IEEETrans.Nanobiosci.,vol.18,no.2,pp.191–204,Apr.2019.
9. Q. Zhang, B. Wang, X. Wei, X. Fang, and C. Zhou, ''DNA word set design based on minimum freeenergy,''IEEETrans.Nanobiosci.,vol.9,no.4,pp.273–277,Dec.2010,doi:10.1109/TNB.2010.2069570.