International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 3, May-June 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

HEREDITARY DISEASE PREDICTION USING MECHINE LEARNING

[1]Nallamudi Neeraja, [2]Mr. M.Buchibabu

[1]Student, Dept. of Master of Computer Applications, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.
[2]Asst.Prof, Dept. of Computer Science & Engineering, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.

ABSTRACT

This system utilises machine learning methods to forecast hereditary diseases. These conditions also arise in individuals coming from families with faulty recessive or dominant genes. These rare diseases rarely impact more than one person per every thousand or million individuals. The goal of this system is to use family history information to estimate the risk that a person may inherit a disease. These machine learning models — decision trees and random forests — are used to create forecasts. Metrics such as accuracy, precision, recall and F1 score are used to gauge how well the algorithms are performing. These outcomes allow us to identify the algorithm that produces the most accurate forecasts. People can utilise this information to develop strategies for preserving their good health. The main diseases addressed by this system include diabetes, cancer, cardiovascular problems, nervous system defects and hair conditions.

Keywords: *Hereditary, machine learning, decision tree, random forest, prediction.*

1. INTRODUCTION

The mechanism by which inherited characteristics are transmitted from one generation to the next accounts for both the stability of species and the diversity among individuals.[1] Genes convey the information that specifies how an organism's traits (phenotype) are expressed inherited from its parents (genotype The genotype remains the same which varies because of the impacts that the surroundings have on the organism.[2], [3]

Machine learning is now being used to help predict inherited diseases like cardiovascular disorders, diabetes, neuropathies and cancer.[4], [5] Many of these diseases share similar genes among family members.[8] Heart disease and diabetes can be more or less likely to develop in a child depending on whether the disease affects the parent's genes. Knowing the nature of the genetic links helps us anticipate and often avoid these illnesses before they develop.[5], [6]

We can forecast and prevent hereditary illnesses more effectively by analysing both people's genes and their lifestyles together.[5] Machine learning helps to accurately determine these risks, giving clinicians powerful resources to make more informed decisions and improve the lives of patients in the years to come.[5], [7], [8]

1.1 Problem Definition

Hereditary diseases are diseases caused by problematic genes that families pass down. Typically, these conditions must be managed constantly and there are few good treatments available.[5], [6], [9] Receiving a diagnosis for one association member may make it likely that others in the family have an increased risk.[10] Families deal with problems related to medical issues, emotions and reproduction, including those about inheritance, testing before birth and what type of treatment to choose. Though there are no easy answers, it's usually important to accept the situation.[5], [7], [11]

1.2 Solution For Problem Definition

There are different ways to predict hereditary diseases and one method is to analyze the historical genetic information passed down from a person's ancestors.[12]The objective of this solution is to create a model that can quantify the risk that a person may develop a specific illness due to their genes.[13], [14] The model is designed using decision trees and random forests which are machines learning algorithms.[5], [15]Information on genetic heritage will be used to train these algorithms so they can effectively forecast whether someone is likely to develop a hereditary illness.[5], [6], [16], [17]

1.3 Process Diagram

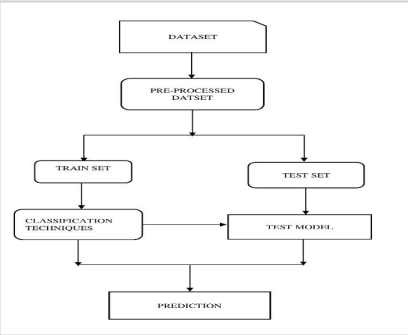

Fig1: Process Diagram

A series of steps are taken to determine what diseases might be passed on to an individual. Here's the outline:

I. **Data Collection**: A file containing aggregated information, known as the hereditary dataset, is downloaded from different websites in the.CSV format.[16]
II. **Data Pre-processing**: Dataset is pre-processed to suit the requirements of machine learning. This process becomes essential when working with real-world data which is not always organized or ready to be processed by machine learning algorithms.[18], [19]
III. **Missing Data Removal**: Missing data is filled using the techniques provided by the imputer library.[20], [21]
IV. **Encoding Categorical Data**: Categorical variables are transformed into numerical values using integer or one-hot encoding.[22]
V. **Data Splitting**: Data is separated into training and test sets which lead to better performance in predictive models.[23]
VI. **Model Training**: A combination of techniques such as Decision Tree and Random Forest are used to train the model on the training data.[24]
VII. **Prediction & Evaluation**: The prediction accuracy of each algorithm on the test data is compared to choose the best one.[17]

2. LITERATURE SURVEY

2.1 Epidemiology and Risk Profile of Heart Failure

Authors: *Anh L. Bui, Tamara B. Horwich, Gregg C. Fonarow*
Methodology:
The paper focuses on identifying the distribution and risk factors of HF in different populations, considering the differences between reduced and preserved ejection fraction types.[9], [10], [11] Main contributors to HF are ischemic heart disease, high blood pressure, being overweight and diabetes.[5]

Advantages:
Heart failure is a serious public health problem that requires effective disease prevention strategies.

Disadvantages:
The high cost of HF management presents a significant challenge to healthcare systems.

2.2 Knowledge Discovery in Traditional Medicine

Author: *Int. J. Recent Technol. Eng*
Methodology:
An in-depth review of 502 studies published between 2000 and 2017 analyses how machine learning is applied in traditional medicine, synthesising findings by topic, methodology, benefits, drawbacks and results.[5], [9], [19]

Advantages:
Advocates deeper researches with high level methods including complex network analysis, genetic algorithm, etc.[5]

Disadvantages:
Certain potentially relevant studies were not searched and reflected literature gaps.[15]

2.3 Hybrid System for Diabetes and Heart Diseases

Authors: *Humar Kahramanli, Novruz Allahverdi*
Methodology:
This paper aims to design a hybrid learning system where ANN and FNN are integrated, and are used to classify the data of diabetes and heart disease. The accuracy of the method was 84.24% for diabetes and 86.8% for heart disease.[20], [22], [24]

Advantages:
The combination system enhances the accuracy and robustness of the classification.[20]

Disadvantages:
It exhibits a slowness and memory space deficiency compared with traditional multilayer perceptron networks.[15], [17]

3. SYSTEM ANALYSIS

3.1 Existing System

The genetic diseases are transmitted between generations, from parent to child, by faulty genes. Chromosomes are important as they carry the genetic traits from parents to their children in humans. In the current system, analysis is conducted on data that relates to particular hereditary diseases like hereditary heart disease or hereditary diabetes.[22], [24] These data sets are conventional processed one at a time by using machine learning methods. But this kind of method also has some limitations, such as low prediction accuracy and then poor practical performance, which can be used for predicting only one kind of disease. [15], [17], [21]

Although machine learning (ML) algorithms harbor great promise for use in conventional medical care, to date, there has been no systematic review and classification of the

applications in traditional medicine.[15] A systematic review was performed based on the Kitchenham method and the study analyzed data from the five databases dating from 2000 to 2017.[22] A total of 502 studies were considered to be relevant to the application of machine learning in conventional medicine. 42 of these papers were included and classified into four categories on:

1. The application domain of data mining techniques for traditional medicine
2. The most frequently applied data mining techniques in TM
3. (ADLS) are far from being guaranteed by these data mining methods.
4. Performance assessment approaches of DM techniques in TM

3.2 Proposed System

The suggested method is meant to bypass some of the present approach's drawbacks in regards to discovering genetic diseases that are handed down through generations.[15]The genetic data will be evaluated more accurately thanks to the implementation of machine learning techniques like Decsion Trees and Random Forest. Unlike the current model, ours does not only predict one disease, but instead looks at how different hereditary diseases could be passed on from each family to the next.[8], [10], [12]

Using this approach, doctors hope to identify hereditary conditions early so that prevention can be possible. Greater efficiency and accuracy in predicting outcomes are achieved when the classification process of the system improves.[2] In addition, the system effectively deals with data, leading to more dependable and faster results. The success of the predictions will be measured to find which machine learning algorithm works best.[6], [7], [15]

3.3 Analysis Model

Each phase in the Waterfall Model builds on the last, with all being completed before the process moves ahead. It moves step by step, where the results from one step are used as guidance for the following one.[10]

Phases:

1. **Requirement Gathering & Analysis:** Capture and document system requirements.
2. **System Design:** Make sure to understand what hardware, software and system architecture are.
3. **Implementation:** Make system units and test each part of the system.
4. **Integration & Testing:** Unite different parts and test the system from start to finish.
5. **Deployment:** After testing the system successfully, let it go into the real world.
6. **Maintenance:** Release new updates and manage any problems that crop up after deploying the software.

Key Characteristics:

- **Sequential Process**: No step is taken in this work until the previous one is completed.
- **Linear Approach**: Advancement moves smoothly from the top to the bottom, just as a waterfall.
- **Defined Goals**: Removal from the phase is possible only when its goals have been accomplished.

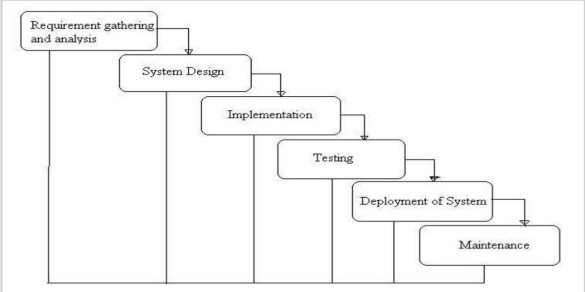

Fig2: Waterfal-model

3.4 MODULES:

A. **Data Selection and Loading**
The data is read from the file and loaded into the program by using pandas `read_csv()`.[22]

B. **Data Preprocessing**
The process includes getting rid of missing values and converting categorical data into one-hot codes.

C. **Splitting Dataset**
To ensure the model is well-fitted, the data is split into sets used for training and sets used for testing.[20]

D. **Classification**

- **Decision Tree**: A tree-based model for classification.
- **Random Forest**: Every decision is made using a set of decision trees for better precision.

E. **Prediction**
Predicts hereditary diseases, evaluated by accuracy, precision, recall, and F1 score.[20]

4. FEASIBILITY STUDY

A feasibility study offers main information about the project and deals with some important questions like: What is the problem? Is there a feasible solution? Is it worth solving? It allows us to figure out if the system can be put into

operation based on its technical, operational and financial expectations. The study includes:

- **Technical Feasibility**
- **Economic Feasibility**
- **Operational Feasibility**

4.1 Technical Feasibility

It involves finding out if the necessary technology and skills are on hand. The system is feasible if:

- Technological tools are available and can be used by students.
- It can process larger amounts of data and remains trustworthy.
- It ensures that prices are exact and simple to use.

4.2 Economic Feasibility

The step is about assessing if the benefits of the system are more than its costs. The cost of the project falls within the financial means when:

- The improvements made outweigh what it costs to build and support the system.

Since it does not cost much to set up and will earn a good profit, the project is considered worthwhile.

4.3 Operational Feasibility

It allows the team to determine if the system can be put into action using resources that are currently available. It is feasible if:

- The work can be finished by the designated time and number of workers.
- The system and its features satisfy what the users and clients require.
- The project can be carried out without much investment and matches what the users are looking for which makes it suitable.

5. System Requirement Specification

5.1 Introduction

The SRS is used to show what the system is supposed to do and includes both important functions and non-functional characteristics. Functional systems tell us how the system will operate, whereas non-functional ones specify issues like performance and quality. The business or systems analyst is tasked with creating the SRS which lists the following three requirement types:

- **Business Requirements**: What must be delivered.
- **Product Requirements**: The features and abilities of the system to satisfy the company's requirements.
- **Process Requirements**: Techniques and tools that must be implemented for the project.

5.2 Functional Requirements

What the system is required to do, along with calculations and data processing, is defined by the functional requirements. Essential functions that the system should have:

- The model can estimate how outputs will turn out using the training data set.
- Link hereditary conditions to the different diseases.

5.3 Non - Functional Requirements

They refer to aspects such as the performance of the system and contain:

- **Availability**: It is always available and changes are applied with minimum disruption.
- **Flexibility**: The system can add new modules without disturbing existing ones.
- **Portability**: Operates on different platforms.
- **Scalability**: Handles increasing loads and hardware expansion.
- **Usability**: Users can figure out how to use the site with little effort.

5.4 SYSTEM REQUIREMENTS

5.4.1 Hardware Requirements

Component	Requirement
Processor	Minimum i3
RAM	Minimum 4GB
Hard Disk	Minimum 25GB

TABLE1:HAREWARE REQ

5.4.2 Software Requirements

Component	Requirement
Operating System	Windows or Linux
Modules	Matplotlib, Numpy, Pandas
Programming Language	Python

TABLE2:SOFTWARE REQ

6. CONCLUSION

We carried out two methods for comparison in this study and the results were positive. From our results, it is clear that machine learning algorithms performed better than other traditional techniques. To check how well each method performed, we measured them by using the confusion matrix, precision, recall and F1 score. Out of all the features in the dataset, Random Forest achieved the best results once data was preprocessed.

The purpose of the study was to predict hereditary traits using machine learning. In the data preparation stage, we cleaned the data and encoded labels using label encoding. After that, the data was reduced to important variables through feature selection and the data was then divided into a training set and a test set. With the rise in technology, machine learning is predicted to take on a bigger role in financial analysis.

Although more people are now aware of health problems and are practicing yoga and dancing, the rise of technology continues to make it difficult to fight inactive lifestyles. All in all, machine learning approaches do well at predicting human genetic traits and their performance is measured mainly by the accuracy they achieve.

ACKNOWLEDGMENT

I am thankful to the Management of Amrita Sai Institute of Science and Technology for giving me an opportunity to work with his project.

I would like to the thank **Dr. M. Sasidhar**, Principal, Amrita Sai institute of science and technology, for his constant encouragement and support during the progress of this work.

I am deeply grateful to **Dr. P. Chiranjeevi**, Professor and Head of the Department, for his valuable guidance and consistent support during the course of the project.

A special note of thanks to my internal guide, **Mr. M.Buchibabu**, for his exceptional guidance, constant motivation, and continuous encouragement, which played a crucial role in the successful completion of this project.

NALLAMUDI NEERAJA

REFERENCES

[1] P. Kalra, "A COMPREHENSIVE REVIEW ON GENETICS AND ITS IMPACT ON HUMAN DISEASES," International Journal of Engineering Applied Sciences and Technology, vol. 6, no. 5. IJEAST, Sep. 01, 2021. doi: 10.33564/ijeast.2021.v06i05.038.

[2] M. Jackson, L. Marks, G. May, and J. B. Wilson, "The genetic basis of disease," Essays in Biochemistry, vol. 62, no. 5. Portland Press, p. 643, Dec. 03, 2018. doi: 10.1042/ebc20170053.

[3] S. Okser, T. Pahikkala, and T. Aittokallio, "Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives," BioData Mining, vol. 6, no. 1, Mar. 2013, doi: 10.1186/1756-0381-6-5.

[4] T. Guo and X. Li, "Machine learning for predicting phenotype from genotype and environment," Current Opinion in Biotechnology, vol. 79. Elsevier BV, p. 102853, Dec. 02, 2022. doi: 10.1016/j.copbio.2022.102853.

[5] D. Ho, W. Schierding, M. Wake, R. Saffery, and J. M. O'Sullivan, "Machine Learning SNP Based Prediction for Precision Medicine," Frontiers in Genetics, vol. 10. Frontiers Media, Mar. 27, 2019. doi: 10.3389/fgene.2019.00267.

[6] D. Le, "Machine learning-based approaches for disease gene prediction," Briefings in Functional Genomics, vol. 19, p. 350, May 2020, doi: 10.1093/bfgp/elaa013.

[7] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," Healthcare, vol. 11, no. 12. Multidisciplinary Digital Publishing Institute, p. 1808, Jun. 20, 2023. doi: 10.3390/healthcare11121808.

[8] P. Dworzyński et al., "Nationwide prediction of type 2 diabetes comorbidities," Scientific Reports, vol. 10, no. 1, Feb. 2020, doi: 10.1038/s41598-020-58601-7.

[9] A. Dhankhar and S. M. Jain, "Prediction of Disease Using Machine Learning Algorithms." p. 115, Mar. 24, 2021. doi: 10.1002/9781119752134.ch8.

[10] S. Bijarnia-Mahay et al., "The changing scenario in prenatal diagnosis of genetic disorders: Genetics to genomics," Current Medicine Research and Practice, vol. 8, no. 6, p. 203, Nov. 2018, doi: 10.1016/j.cmrp.2018.11.004.

[11] D. Schlegel, E. Cunningham, Z. Xing-hai, Y. Abdulhak, A. DeOrio, and T. Jayasundera, "Inheritance Pattern Prediction of Retinal Dystrophies: A Machine-Learning Model," Investigative Ophthalmology & Visual Science, vol. 58, no. 8, p. 5586, Jun. 2017, Accessed: Apr. 2025. [Online]. Available:

https://iovs.arvojournals.org/article.aspx?articleid=2642384
[12] S. Gallati, "Disease-modifying genes and monogenic disorders: experience in cystic fibrosis," The Application of Clinical Genetics. Dove Medical Press, p. 133, Jul. 01, 2014. doi: 10.2147/tacg.s18675.
[13] L. Iddamalgoda, P. Das, A. Aponso, V. S. Sundararajan, P. Suravajhala, and J. Valadi, "Data Mining and Pattern Recognition Models for Identifying Inherited Diseases: Challenges and Implications," Frontiers in Genetics, vol. 7. Frontiers Media, Aug. 10, 2016. doi: 10.3389/fgene.2016.00136.
[14] Z. Guan, G. Parmigiani, D. Braun, and L. Trippa, "Prediction of Hereditary Cancers Using Neural Networks," arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2106.13682.
[15] Z. Guan, G. Parmigiani, D. Braun, and L. Trippa, "Prediction of hereditary cancers using neural networks," The Annals of Applied Statistics, vol. 16, no. 1, Mar. 2022, doi: 10.1214/21-aoas1510.
[16] V. S. G., R. K., K. F. Mohammed, and R. R. D., "Disease Prediction using Machine Learning." Jan. 2021. Accessed: May 17, 2025. [Online]. Available: https://journals.bohrpub.com/index.php/bijcs/article/download/80/2295
[17] B. Ravi, "disease-prediction-model." May 2021. Accessed: May 17, 2025. [Online]. Available: https://github.com/BhuvaneshRavi/disease-prediction-model
[18] U. Verma, "Data Cleaning and Preprocessing." Nov. 2019. Accessed: May 17, 2025. [Online]. Available: https://medium.com/analytics-vidhya/data-cleaning-and-preprocessing-a4b751f4066f
[19] A. D'Agostino, "How To Prepare Data For Machine Learning." Apr. 2023. Accessed: May 17, 2025. [Online]. Available: https://towardsdatascience.com/how-to-prepare-data-for-machine-learning-eb9d9973832f?gi=51b266c25f31
[20] A. Pajankar and A. Joshi, "Preparing Data for Machine Learning," in Apress eBooks, 2022, p. 79. doi: 10.1007/978-1-4842-7921-2_6.
[21] S. Becker, R. Hug, W. Huebner, M. Arens, and B. Morris, "Handling Missing Observations with an RNN-based Prediction-Update Cycle," in Lecture notes in computer science, Springer Science+Business Media, 2021, p. 311. doi: 10.1007/978-3-030-89128-2_30.
[22] G. Truda, "Generating tabular datasets under differential privacy," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arXiv.2308.14784.
[23] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," Computational Statistics, vol. 37, no. 5, p. 2671, Mar. 2022, doi: 10.1007/s00180-022-01207-6.
[24] R. R. Picard and K. N. Berk, "Data Splitting," The American Statistician, vol. 44, no. 2, p. 140, May 1990, doi: 10.1080/00031305.1990.10475704.

