PREDICTING GENETIC VARIANTS PATHOGENECITY

Prof. Syed Arbeena Kausar[1], Venkatesh I[2], Vijay H[3], Yamuna P[4], Mahesh C R[5]

[1] *Assistant Professor, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[2] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[3] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[4] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[5] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*

ABSTRACT

Single Nucleotide Polymorphism (SNP) detection plays a pivotal role in understanding the intricacies of genetic inheritance and its influence on phenotypic traits. This project focuses on developing a predictive model that analyzes and determines the parental origin of specific SNPs in a child's DNA. By leveraging genetic data from both parents, we aim to gain deeper insights into hereditary patterns and how they contribute to observable characteristics.

To accomplish this, we implemented a logistic regression model that integrates not only genetic information but also lifestyle and environmental factors such as smoking, alcohol consumption, radiation exposure, and mutation scores. These factors were included to examine their potential impact on SNP expression and inheritance. The model was trained and tested on a dataset composed of these variables, resulting in an overall accuracy of 58%, an ROC AUC of 0.604, a recall of 63.27%, and a precision of 56.36%. These metrics suggest the model has moderate predictive capabilities but also emphasize the need for more sophisticated feature selection and machine learning algorithms to enhance performance.

In conclusion, this project underlines the multifaceted nature of genetic prediction and the influence of environmental variables on genetic traits. While the logistic regression model provides a foundational approach to SNP inheritance analysis, future work will explore advanced models and deeper biological data integration to improve accuracy and practical applicability in genetic research and personalized medicine.

Keyword: - *Single Nucleotide Polymorphism (SNP), Phenotypic Traits, Predictive Modeling, Logistic Regression, Parental Origin, Environmental Factors, Mutation Score, Genetic Data Analysis*

1. INTRODUCTION

Genetic variations, particularly Single Nucleotide Polymorphisms (SNPs), play a crucial role in understanding hereditary traits and disease susceptibility. SNPs, which occur when a single nucleotide in the genome differs between individuals, serve as biological markers that can help trace genetic inheritance, disease predisposition, and gene-environment interactions. Recent advancements in genome-wide association studies (GWAS) and machine learning techniques have enabled researchers to investigate SNP patterns and their correlation with complex diseases such as diabetes, asthma, osteoporosis, and migraine [1,2].

However, a significant challenge remains in accurately predicting SNP mutations and their impact on health, particularly when considering environmental factors like smoking, drinking, and radiation exposure [3].

This study focuses on SNP detection by comparing parental DNA with a child's DNA while integrating environmental factors that may contribute to genetic mutations. By leveraging logistic regression, the research aims to predict the probability of SNP variations based on inherited genetic data and lifestyle influences. Logistic regression, widely used in medical and genetic studies, provides a robust statistical approach to model the relationship between independent variables (parental DNA, environmental exposure) and dependent outcomes (child's SNP mutations). While previous studies have applied logistic regression in SNP-based disease prediction [4,5], challenges such as non-linearity in genetic interactions and limited prediction accuracy necessitate further model refinement.

This research builds upon ten key studies in SNP analysis, machine learning applications in genetics, and environmental influences on DNA variation. Studies such as [6,7] have explored different SNP detection techniques, including whole-genome sequencing and PCR-based methods, providing foundational methodologies for genetic comparison. Additionally, SNP diversity studies [8] and machine learning-based susceptibility analyses [9] contribute to identifying significant mutation patterns. While genetic inheritance remains the primary factor in SNP variation, environmental factors can introduce epigenetic modifications, increasing the complexity of genetic modeling [10].

By integrating SNP analysis with environmental data and optimizing logistic regression with feature engineering techniques, this study aims to enhance the prediction of SNP mutations across generations. The proposed methodology will not only improve the accuracy of SNP-based disease susceptibility models but also contribute to personalized medicine by identifying high-risk individuals based on both genetic and environmental factors. Through a systematic evaluation of genetic inheritance and external influences, this research seeks to provide novel insights into how lifestyle choices may shape genetic evolution and disease risk over time.

2. METHODOLOGY

The predicting genetic variants pathogenecity system involves several stages as shown in Figure 1:

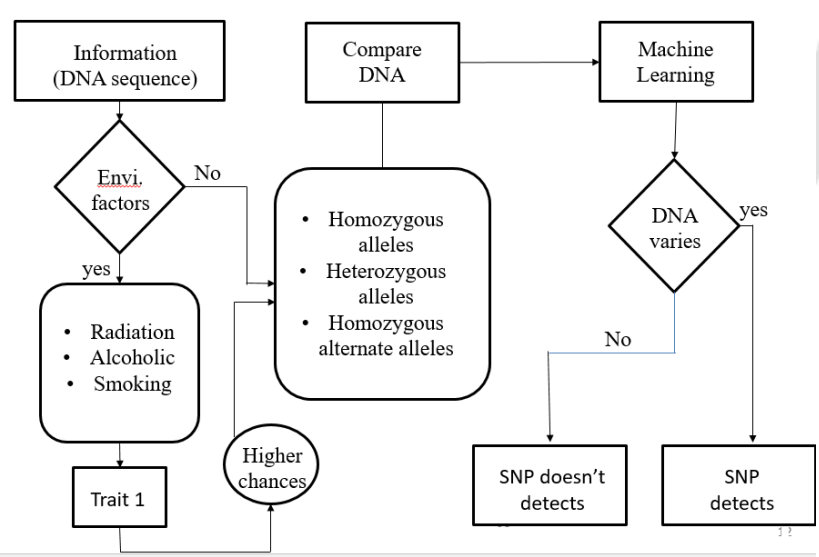

Fig -1: Methodology of Predicting Genetic Variants Pathogenecity System

i. Data Collection and Input

The system accepts genetic data in the form of DNA sequences (consisting of characters A, T, C, G) from three family members: the mother, the father, and the child. In addition to genetic information, users can input lifestyle

and environmental factors such as:

- Smoking habits
- Alcohol consumption
- Industrial or factory exposure

The interface also allows users to select diseases or traits known to be inherited genetically, such as cystic fibrosis or sickle cell anemia

ii. DNA Sequence Validation and Preprocessing

Each DNA input is validated to contain only valid nucleotide characters (A, T, C, G). Real-time validation and auto-correction ensure clean input. Once validated, DNA sequences are preprocessed to:

- Normalize the sequence length for comparison
- Remove invalid characters
- Calculate similarity and mutation scores

iii. Mutation Score and Similarity Calculation

The system uses string matching and comparison techniques to determine:

- The percentage similarity between each parent's DNA and the child's DNA
- The mutation score, which is computed as the normalized ratio of differing positions in DNA sequences

iv. Result Visualization and Reporting

Once the prediction is made, the results are displayed through an interactive dashboard that includes:

- Parent-wise similarity and mutation metrics
- Inferred trait probability
- Environmental influence summary
- Genetic diseases reported by parents

3. PROPOSED SYSTEM

Understanding the impact of environmental factors on DNA sequences is a crucial aspect of genetics and bioinformatics. This project aims to design a system that predicts the possibility of mutation in a given DNA sequence when subjected to certain environmental conditions, using Logistic Regression, a supervised machine learning technique well-suited for binary classification problems.

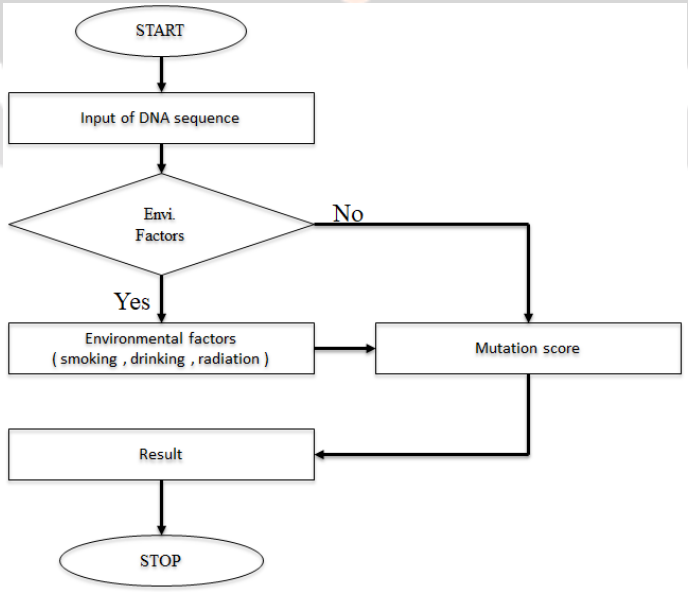

Fig -2: Workflow of Predicting Genetic Variants Pathogenecity System

3.1 System Architecture Overview

i. The system starts with a DNA sequence input, consisting of bases A, T, G, and C.
ii. This sequence can be entered manually, uploaded, or retrieved from a biological database.
iii. Environmental factors like radiation, chemicals, and temperature are then collected.
iv. These factors are represented numerically or categorically and require preprocessing.
v. Logistic Regression is applied to model mutation probability under the given conditions.
vi. The logistic function outputs a probability score between 0 and 1.
vii. A threshold (e.g., 0.5) determines whether mutation is likely or not.
viii. If mutation is predicted, a mutation score quantifies its potential severity.
ix. The system categorizes mutation risks as mild, moderate, or severe.
x. Results are generated and can be visualized or exported for biological analysis.

4. EQUATION:

In this project, we use a logistic regression model to estimate the probability (P) that a child inherits a specific SNP based on parental genetic data and several environmental/lifestyle factors. The generalized logistic regression hypothesis function is:

$P(y = 1 \mid x) = 1 / (1 + e^{(-z)})$ where $z = \theta_0 + \theta_1x_1 + \theta_2x_2 + \theta_3x_3 + \theta_4x_4 + \theta_5x_5 + \theta_6x_6 + \theta_7x_7$

Where:

- x_1 = Mother's allele (e.g., 0 or 1)
- x_2 = Father's allele (e.g., 0 or 1)
- x_3 = Smoking habit (0 = no, 1 = yes)
- x_4 = Alcohol consumption (0 = no, 1 = yes)
- x_5 = Radiation exposure level (numerical score)
- x_6 = Mutation score (numerical score)
- x_7 = Other environmental factor(s), if applicable
- θ_0 to θ_7 = Model parameters learned during training
- $y \in \{0, 1\}$ = Output variable (1 = SNP inherited, 0 = not inherited)

Final Prediction Rule:

If $P \geq 0.5 \rightarrow$ Predict SNP inherited, else If $P < 0.5 \rightarrow$ Predict SNP not inherited

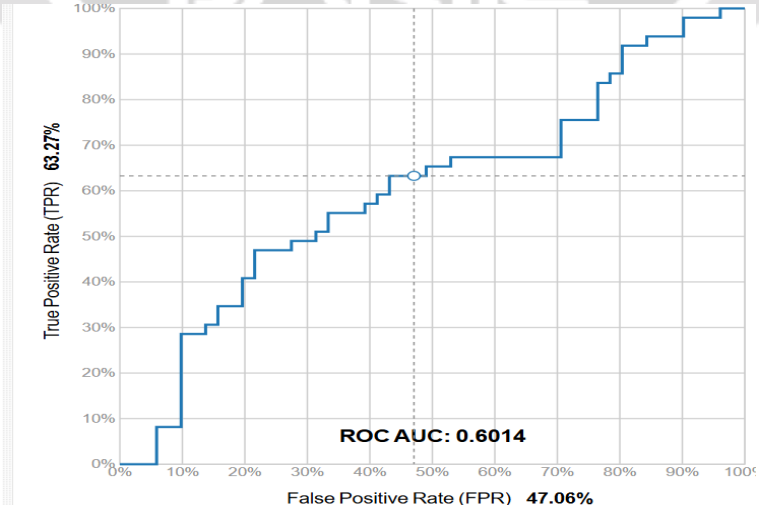

Fig – 3 : Graph of Logistic Regression (with metrics)

5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this project, a web-based system was successfully developed to predict the parental origin of Single Nucleotide Polymorphisms (SNPs) in a child's DNA by integrating genetic and environmental factors. The system uses DNA sequence similarity, mutation scores, lifestyle data (such as smoking, drinking, and radiation exposure), and a logistic regression model to deliver trait probability predictions.

The model achieved an accuracy of 58%, a recall of 63.27%, a precision of 56.36%, and a ROC AUC score of 0.604. These results indicate that the system can moderately distinguish between inherited traits from the mother or father. Additionally, the platform offers a real-time, user-friendly interface that allows users to input data, receive immediate predictions, and download detailed reports.

By incorporating environmental and lifestyle factors along with genetic information, the system presents a more realistic and practical approach to genetic prediction compared to traditional models based solely on DNA sequences. The use of modern web technologies such as Node.js, React, and TypeScript (TSX) ensures that the system is scalable, fast, and accessible through any modern browser without database dependency.

6. REFERENCES

[1] Hettiarachchi, G., & Komar, A. A. (2022). GWAS to Identify SNPs Associated with Common Diseases and Individual Risk: Genome Wide Association Studies (GWAS) to Identify SNPs Associated with Common Diseases and Individual Risk. https://doi.org/10.1007/978-3-031-05616-1_4

[2] Kaur, S., et al. (2019). Role of single nucleotide polymorphisms (SNPs) in common migraine. https://doi.org/10.1186/s41983-019-0093-8

[3] Gaudillo, J., et al. (2019). Machine learning approach to single nucleotide polymorphism-based asthma prediction. https://doi.org/10.1371/journal.pone.0225574

[4] Su, L., et al. (2015). Research on SNP Interaction Detection from a Network Perspective.

[5] Wakayu, E. G. (2021). Machine Learning Analysis of Single Nucleotide Polymorphism (SNP) Data to Predict Bone Mineral Density in African American Women. http://dx.doi.org/10.34917/28340371

[6] Kwok, P-Y., & Chen, X. (2003). Detection of Single Nucleotide Polymorphisms.

[7] Jian, Y., & Li, M. (2021). A narrative review of single-nucleotide polymorphism detection methods and their application in studies of Staphylococcus aureus. https://spj.science.org/doi/10.1097/JBR.0000000000000071

[8] Yirgu, M., et al. (2023). Single nucleotide polymorphism (SNP) markers for genetic diversity and population structure study in Ethiopian barley (Hordeum vulgare L.) germplasm. https://doi.org/10.1186/s12863-023-01109-6

[9] Silva, P. P., et al. (2022). A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci.

[10] López, B., et al. (2017). Single Nucleotide Polymorphism relevance learning with RandomForests for Type 2 diabetes risk prediction. www.elsevier.com/locate/aiim