**GENETIC VARIANT CLASSIFICATION**

***Project synopsis submitted to***

***Guru Jambheshwar University of Science and Technology, Hisar***

***for the partial award of the degree***

***of***

**Bachelor of Technology**

**in Computer Science and Engineering**

***by***

**Ashish Kumar (200010130020) Dr. Yogesh Chaba**

**Arvind (200010130018) Professor**

****

**Department of Computer Science & Engineering**

**GURU JAMBHESHWAR UNIVERSITY OF SCIENCE AND TECHNOLOGY, HISAR**

**January, 2024**

**CONTENT**

**1. Introduction ………………………………………………………………. 3**

**2. Background Details and Literature Review ……………………………. 4**

**3. Problem Formulation and Objectives …………………………………….. 5**

**4. Methodology and tools to be used …………………………………………. 6**

**5. References/Bibliography …………………………………………………… 7**

**Introduction**

Genetic variant classification is a critical process in the field of genetics, particularly in a diagnostic setting. It forms the basis for clinical decision making and is crucial for managing patients and realizing the best possible outcomes.

The classification of genetic variants involves assigning categories to these variants based on their potential impact on health or disease. The process entails the collation and evaluation of various sources of evidence to determine the clinical significance of variants identified through diagnostic testing for a disease with a suspected underlying genetic cause.

Most genetic diagnostics laboratories develop and use their own in-house variant classification system, with many following the recommendations of the American College of Medical Genetics and Genomics (ACMG) guidelines. However, there can be differences between the classification systems, which can sometimes result in different classifications of the same variant between companies.

Blueprint Genetics, for example, has developed a variant classification system intended to classify variants in dominant monogenic disorders, which are rare diseases caused by single variants in single genes. Their system closely follows the guidelines and interpretation criteria established by the ACMG.

The classification system typically involves a 5-tiered scheme that describes the quantity and quality of evidence needed to classify a genetic variant as pathogenic, likely pathogenic, a variant of uncertain significance (VUS), likely benign, or benign. This systematic, clear, and sensible variant evaluation criteria is crucial for making confident diagnostic decisions.

**Background Details**

Genetic variant classification is a critical process in the field of genetics, particularly in a diagnostic setting. It forms the basis for clinical decision making and is crucial for managing patients and realizing the best possible outcomes.

The classification of genetic variants involves assigning categories to these variants based on their potential impact on health or disease. The process entails the collation and evaluation of various sources of evidence to determine the clinical significance of variants identified through diagnostic testing for a disease with a suspected underlying genetic cause.

Most genetic diagnostics laboratories develop and use their own in-house variant classification system, with many following the recommendations of the American College of Medical Genetics and Genomics (ACMG) guidelines. However, there can be differences between the classification systems, which can sometimes result in different classifications of the same variant between companies.

Blueprint Genetics, for example, has developed a variant classification system intended to classify variants in dominant monogenic disorders, which are rare diseases caused by single variants in single genes. Their system closely follows the guidelines and interpretation criteria established by the ACMG.

The classification system typically involves a 5-tiered scheme that describes the quantity and quality of evidence needed to classify a genetic variant as pathogenic, likely pathogenic, a variant of uncertain significance (VUS), likely benign, or benign. This systematic, clear, and sensible variant evaluation criteria is crucial for making confident diagnostic decisions.

**Literature Review**

**Oscar Campuzano**

Our cohort included 104 cases diagnosed with inherited arrhythmogenic syndromes and 17 post-mortem cases in which inherited arrhythmogenic syndromes was cause of death. 71.87% of variants change their classification. While 65.62% of variants were classified as likely pathogenic in 2010, after reanalysis, only 17.96% remain as likely pathogenic. In 2010, 18.75% of variants were classified as uncertain role but nowadays 60.15% of variants are classified of unknown significance.

**Lynn B Jorde & Stephen P Wooding**

while genetic variation shows geographical structure and ancestral inference may be useful in certain contexts, the concept of race as discrete, biologically meaningful categories is scientifically unfounded. Modern human genetics underscores the shared nature of genetic variation, highlighting its potential to render race largely irrelevant in clinical decision-making as individual genetic assessments become more prevalent.

**Laura Valle, Eduardo Vilar, Sean V Tavtigian, Elena M Stoffel**

early identification of genetic predispositions to colorectal cancer not only allows for precision treatment and prevention but also underscores the significance of ongoing advancements in screening and management strategies to reduce associated morbidity and mortality.

**Gunnar Houge**

The results emphasize the need for cautious interpretation, as discrepancies between ACMG and ABC classifications, conducted by the same individual, underscore the importance of planned follow-up between-lab comparison studies to assess robustness and reproducibility.

**Problem Formulation**

Elucidate the relationship between the pathogenicity of genetic modifications and the category of sickness that results from them.

The toxicity of the resulting genetic anomalies has a significant impact on the classification of ensuing disorders; many mutations result in benign ailments or have no discernible effects, while others create serious or potentially deadly issues. The classification of genes and the degree of severity and clinical manifestation of the disorders they correspond with are often based on the functional implications of mutations, such as nonsense mutations that produce truncated, non-functioning proteins or missense mutations that alter the structure of proteins. Understanding this connection is essential to approaches including targeted and tailored therapy.

**Objectives**

1. Functional Impact on Protein Structure: Highlight the crucial role of genetic modifications in altering protein structure, emphasizing the distinction between benign mutations and those with serious implications. Clarify that missense mutations may lead to changes in protein structure, while nonsense mutations result in the production of non-functional proteins, influencing the severity of associated disorders.

2. Impact on Severity and Clinical Manifestation: Emphasize the direct correlation between the toxicity of genetic anomalies and the classification of resulting disorders. Illustrate how the degree of severity and clinical manifestation of diseases is often determined by the nature of mutations, providing a spectrum from benign conditions to potentially life-threatening issues.

3. Importance for Targeted and Tailored Therapy: Stress the significance of understanding the relationship between pathogenicity and disease classification for the development of targeted and tailored therapeutic interventions. Highlight that insights into the functional implications of genetic mutations enable more precise and effective approaches in personalized medicine, enhancing the potential for successful treatment strategies.

**Methodology**

Data Collection:

Genetic Variant Classification Dataset will be downloaded from www.kaggle.com.

EDA:

Exploratory Data Analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. In EDA we first clean the data by removing null values and outliers.

Feature Engineering:

We build a simple data model to join all datasets. After that we will perform feature engineering to choose effective feature sets for the different fraud patterns.

Model Selection and Training:

Following algorithms and models are selected and trained on the preprocessed data.

1. Chi-Squared

2. PCA Reduction

3. Lasso Regression

4. Recursive Feature Elimination

5. Logistic Regression

6. Random Forest

7. XGBoost

Model Evaluation:

Each model's predictions on the test set are evaluated using various metrics, including accuracy, precision, recall, F1 score, ROC AUC, and confusion matrix.

**Tools Required**

**Hardware requirements**:

The following are the hardware required for the project:

* PC with Pentium II Processor,450MHz (Recommended Pentium III Processor,800MHz)
* 4GB RAM
* Minimum 1.2 GB magnetic disk space.
* PC should be connected with Network .
* CD-ROM (48 X or higher recommended).
* Mouse or Similar Pointing device.
* A Printer is require to take out Reports.

**Software requirements**:

The following are the Software required for the project:

* Microsoft Windows/Linux/Mac OS.
* Google Chrome
* Anaconda Navigator
* Jupyter Notebook
* Pandas
* Numpy
* Scipy
* Matplotlib
* Scikit-learn
* XGBoost

**Bibliography Reference**

* Dataset: [Genetic Variant Classifications (kaggle.com)](https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting)
* https://www.sciencedirect.com/science/article/pii/S2352396420301079
* https://www.nature.com/articles/ng1435#Abs1
* https://pathsocjournals.onlinelibrary.wiley.com/doi/full/10.1002/path.5229
* https://www.nature.com/articles/s41431-021-00903-z#Sec3