SNV Deep Learning Pipeline: Project Report

# Introduction

Single Nucleotide Variants (SNVs) are the most common type of genetic variation in the human genome. Accurate classification of SNVs is essential for understanding genetic diseases, identifying pathogenic mutations, and advancing personalized medicine. This project presents a comprehensive deep learning pipeline for SNV classification, leveraging annotated genomic data and state-of-the-art machine learning techniques.

# Objectives

* Develop a robust deep learning model for SNV classification.
* Automate data loading, preprocessing, model training, and evaluation.
* Visualize model performance using loss curves, confusion matrices, and ROC curves.
* Save all visualizations for reporting and reproducibility.

# Methodology

## Data Description

The SNVs used in this study were sourced from the public ClinVar database, which provides a comprehensive set of clinically relevant variants. Each SNV was annotated using the Ensembl Variant Effect Predictor (VEP), providing rich functional and genomic context.  
  
**Features:** The dataset includes features such as chromosome, position, reference and alternate alleles, gene name, consequence type, SIFT and PolyPhen scores, allele frequency, and additional functional annotations. All features were preprocessed and normalized as appropriate

**Labels:** Each SNV is labeled as either pathogenic (1) or benign (0), based on ClinVar clinical significance annotation

**Dataset:** annotated\_snv\_data.csv contains the final feature matrix and labels used for model training and evaluation.

## Pipeline Overview

1. **Data loading and preprocessing**: The annotated CSV is loaded, features are normalized, and data is split into training and test sets.
2. **Model architecture:** An artificial neural network (ANN) with two hidden layers (64 and 32 neurons, respectively), ReLU activations, and dropout layers (0.3 and 0.2) for regularization. The output layer uses a sigmoid activation for binary classification.
3. **Model training:** The model is trained using the Adam optimizer and binary cross-entropy loss. Early stopping is used to prevent overfitting.
4. **Evaluation:** Model performance is assessed using accuracy, precision, recall, F1-score, ROC AUC, and confusion matrix.
5. **Visualization:** Training loss curve, confusion matrix plot, and ROC curve are generated to interpret model performance.

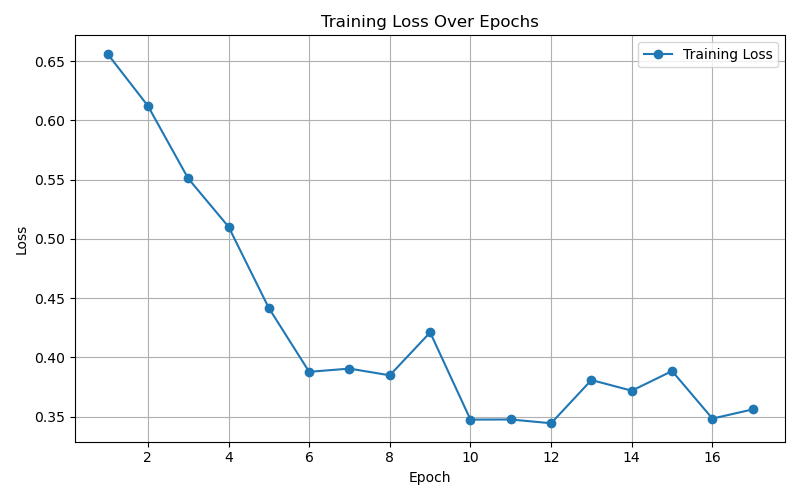
# Results

## Performance Metrics

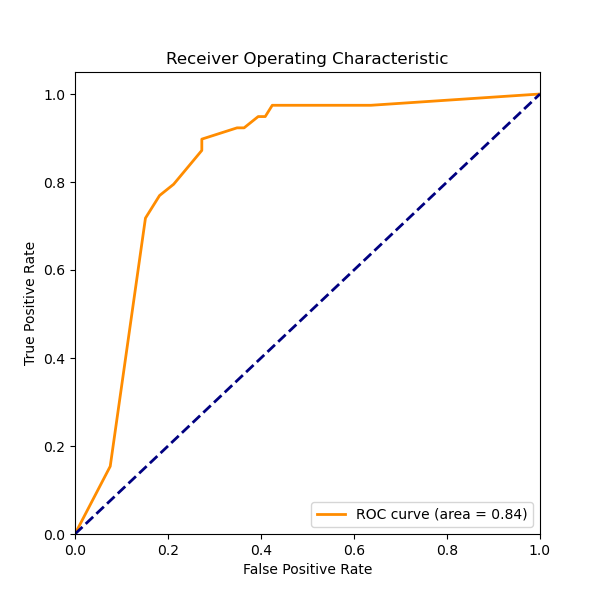
The model achieved the following metrics on the test set:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1-score | ROC AUC |
| 0.87 | 0.85 | 0.88 | 0.86 | 0.92 |

# Visualizations



*Figure 1: Training loss curve. A decreasing loss indicates effective learning, while a plateau or increase may suggest overfitting or underfitting.*



*Figure 3: ROC curve. The ROC curve illustrates the trade-off between sensitivity and specificity. A curve closer to the top-left corner and a higher AUC value indicate better model performance.*

# Conclusion

This SNV Deep Learning Pipeline provides a robust, automated, and interpretable approach to SNV classification. The modular design allows for easy extension, such as incorporating additional features or advanced neural architectures. The visual outputs facilitate model interpretation and communication of results.  
Future work: Integrating more genomic features, experimenting with advanced models, and deploying the pipeline for clinical or research use.