Smoker Status Analysis and Prediction

Vansh Doshi

Roll No: **MT2025729**

Department of Artificial Intelligence and Data Science

Dholu Vishnu

Roll No: **MT2025044**

Department of Computer Science and Engineering

IIIT Bangalore

December 12, 2025

Contents

Abstract

Smoking remains one of the leading preventable causes of global morbidity and mortality. This project develops a complete machine learning pipeline to classify individuals as **Smokers** or **Non-Smokers** using physiological and biochemical health indicators. A dataset of 33,467 unique samples containing 23 biometric features was analyzed and preprocessed through duplicate removal, log-transformation for skewed variables, feature scaling using `StandardScaler`, and class balancing using the **SMOTE** oversampling technique.

Three models were trained and evaluated: Logistic Regression, Support Vector Machines (SVM), and a Deep Neural Network (DNN) incorporating batch normalization and dropout regularization. Experimental results demonstrate that the **Deep Neural Network (75.63% accuracy)** outperformed both the SVM (72.05%) and Logistic Regression (72.57%). The analysis highlights Hemoglobin and GTP levels as highly discriminative features, reinforcing medical evidence that smoking alters hematological and liver-function biomarkers.

1 Introduction

1.1 Background and Motivation

Smoking is a major contributor to chronic illnesses such as cardiovascular diseases, pulmonary disorders, and many cancers. Traditional survey-based methods to identify smoking habits often suffer from bias and underreporting, highlighting the need for robust, data-driven screening tools that rely on objective biomarkers.

Advances in machine learning enable automated extraction of complex patterns from biometric data, making it possible to infer smoking status in a scalable and non-invasive manner.

1.2 Objectives

This project aims to:

- Build a predictive model to classify individuals as smokers or non-smokers.
- Compare classical machine learning models with a modern Deep Neural Network.
- Identify physiological markers most strongly associated with smoking behavior.

1.3 Dataset Description

The dataset contains 23 features and one binary target variable (`smoking`). It includes:

- **Demographics:** Age, Height, Weight, Waist.
- **Cardiovascular:** Systolic and Diastolic Blood Pressure.
- **Biochemical:** Hemoglobin, Triglycerides, Cholesterol, Creatinine.
- **Liver Enzymes:** AST, ALT, GTP.

After removing duplicates, 33,467 samples were retained.

2 Exploratory Data Analysis (EDA)

2.1 Correlation Analysis

A correlation heatmap revealed strong relationships such as Waist–Weight (0.82) and Systolic–Diastolic Blood Pressure (0.76), indicating multicollinearity among physical and cardiovascular factors.

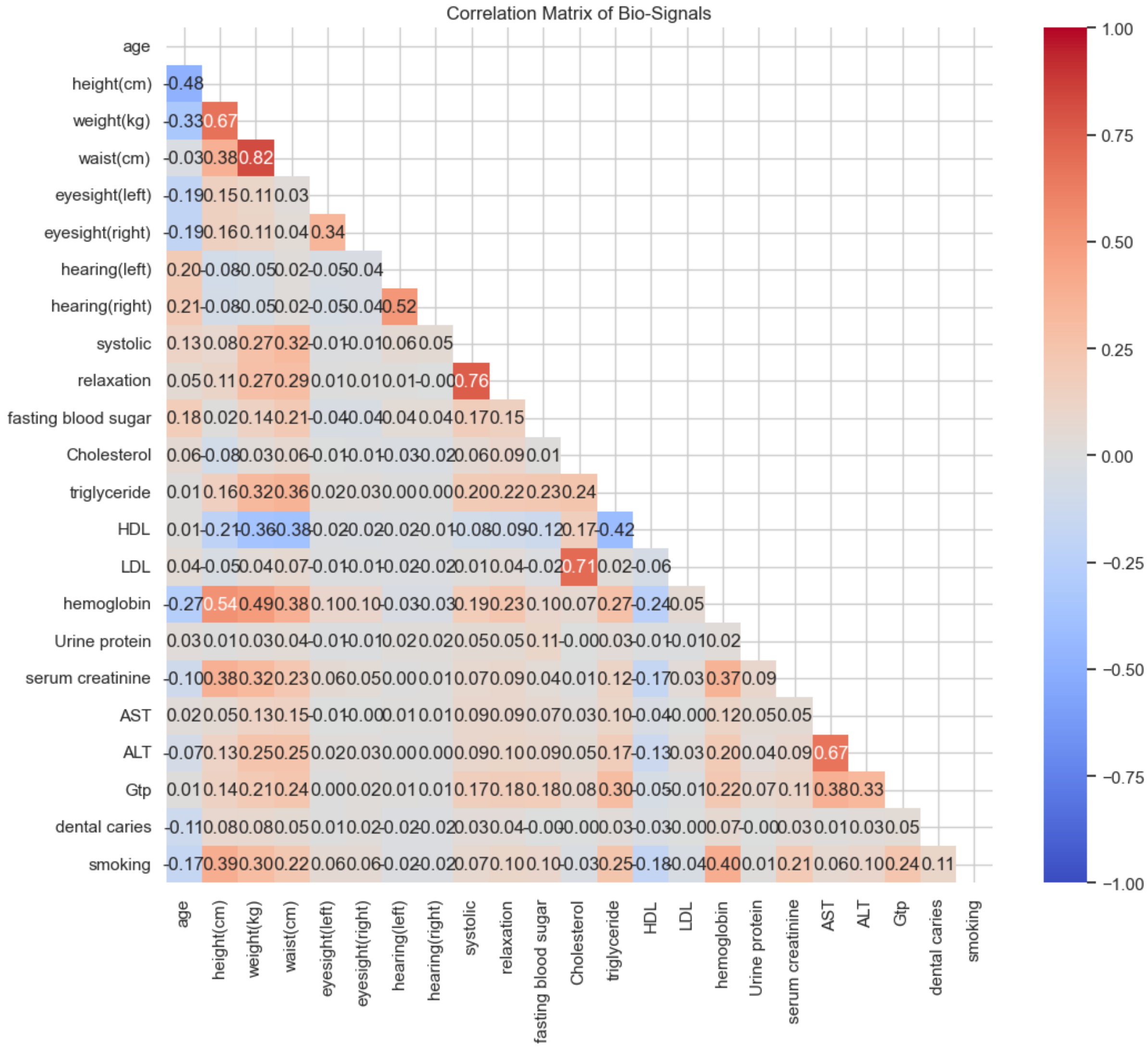

Figure 1: Correlation Matrix of Bio-Signals

2.2 Outlier and Skewness Analysis

Certain biochemical features (e.g., GTP, Triglycerides) exhibited heavy right skewness. Logarithmic transformation was applied to stabilize variance and improve learning.

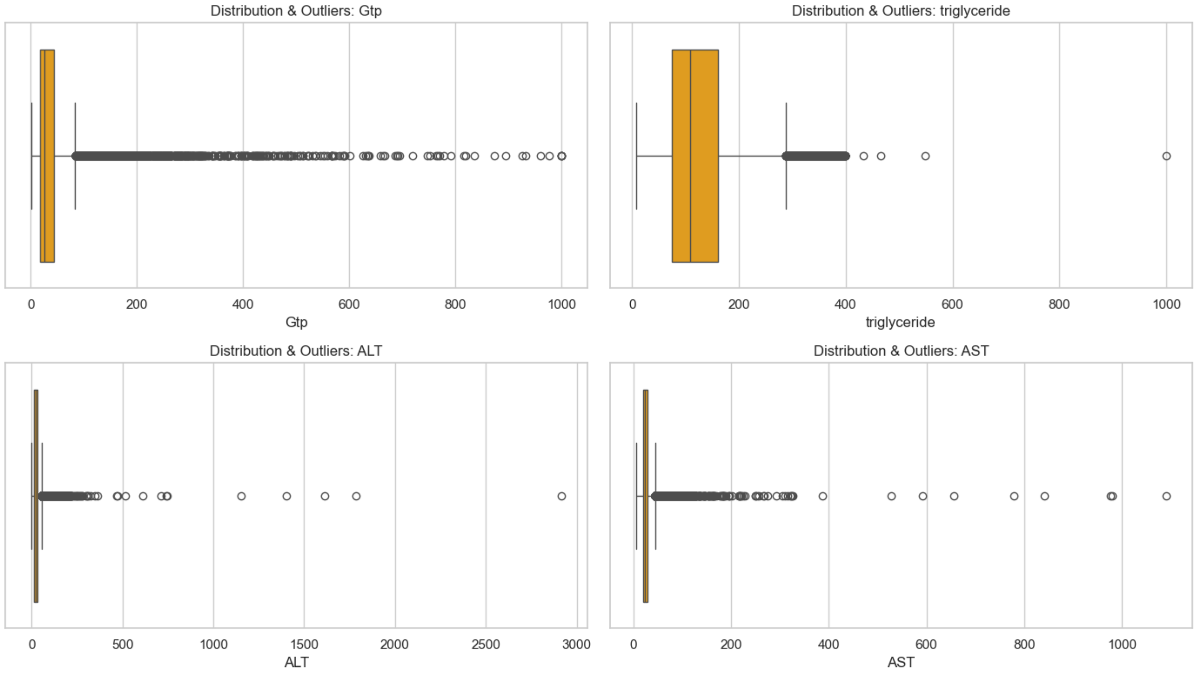

Figure 2: Enzyme Distributions Showing Skewness and Outliers

2.3 Class Imbalance

The dataset had a 63:37 imbalance between non-smokers and smokers. **SMOTE** was used to balance the classes and prevent model bias.

3 Methodology

3.1 Data Preprocessing

Steps include:

1. Removal of duplicate samples.
2. Log-transformation of skewed features.
3. Standardization of all numeric features.
4. Train-test split (80:20).
5. SMOTE applied before model training.

3.2 Model 1: Logistic Regression

Logistic Regression estimates:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Best Parameters:

$$C = 10, \quad \text{solver} = \texttt{liblinear}$$

Test Accuracy: 72.57%

3.3 Model 2: Support Vector Machine (SVM)

The RBF kernel models non-linear relationships:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

Best Parameters:

$$C = 50, \quad \gamma = \text{scale}$$

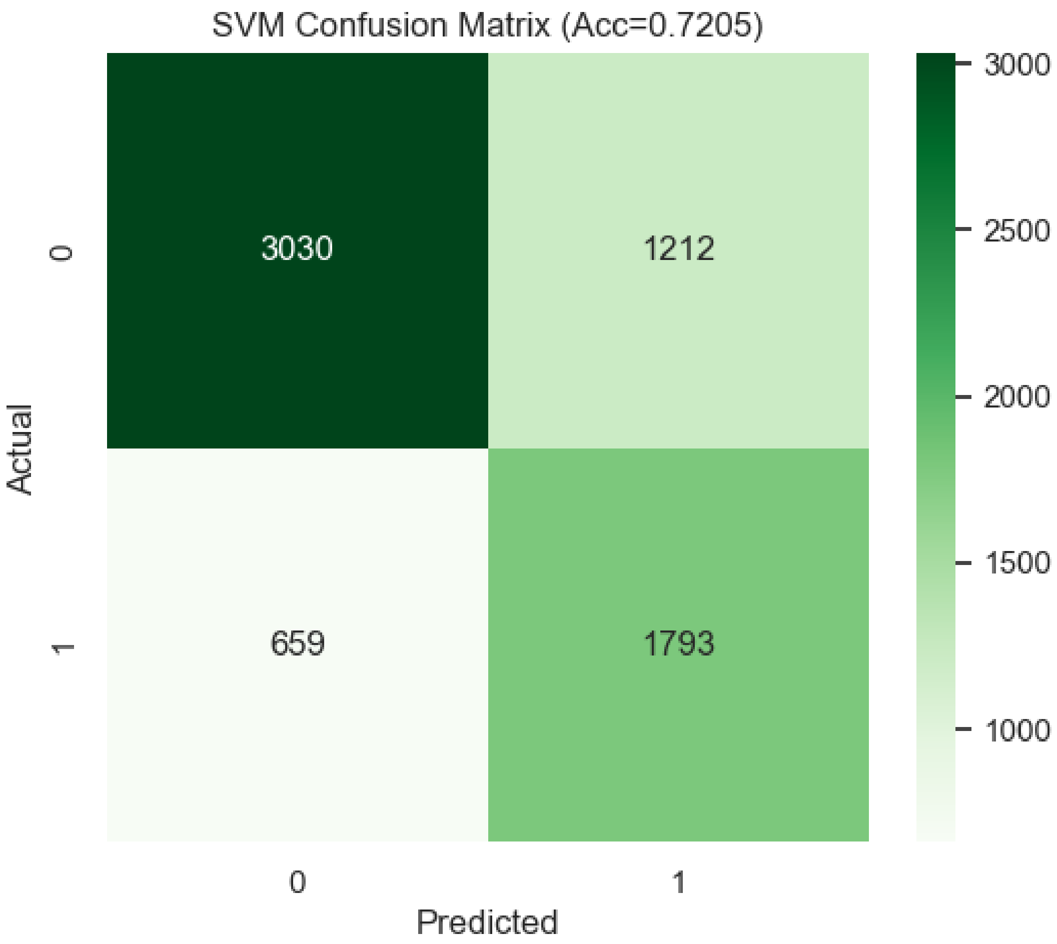

Figure 3: SVM Confusion Matrix

3.4 Model 3: Deep Neural Network (DNN)

Architecture

- Dense(128) → BatchNorm → ReLU → Dropout(0.3)
- Dense(64) → BatchNorm → ReLU → Dropout(0.3)
- Dense(32) → BatchNorm → ReLU
- Output: Dense(1, Sigmoid)

Training Setup

- Optimizer: Adam (lr = 0.001)
- Loss: Binary Crossentropy
- Callbacks: EarlyStopping, ReduceLROnPlateau

Performance

$$\textbf{Neural Network Accuracy} = \mathbf{75.63\%}$$

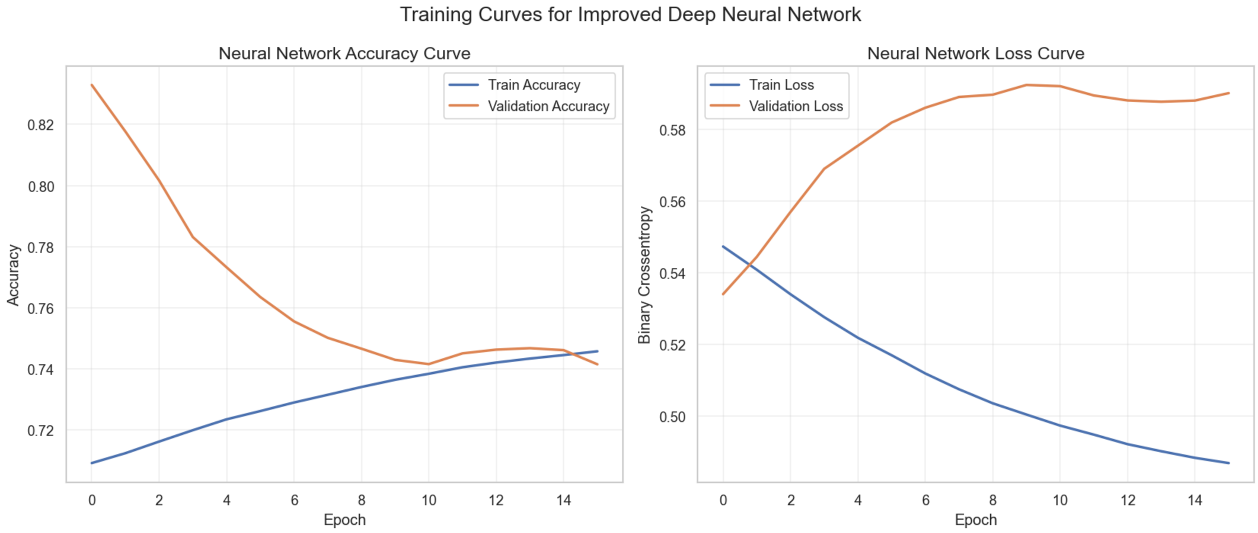

Figure 4: Neural Network Training Curves

4 Results and Comparative Analysis

4.1 Accuracy Comparison

Model	Test Accuracy	Rank
Logistic Regression	72.57%	3
Support Vector Machine	72.05%	2
Deep Neural Network	**75.63%**	**1**

4.2 Feature Insights

Hemoglobin and GTP showed strong discrimination ability, consistent with clinical observations.

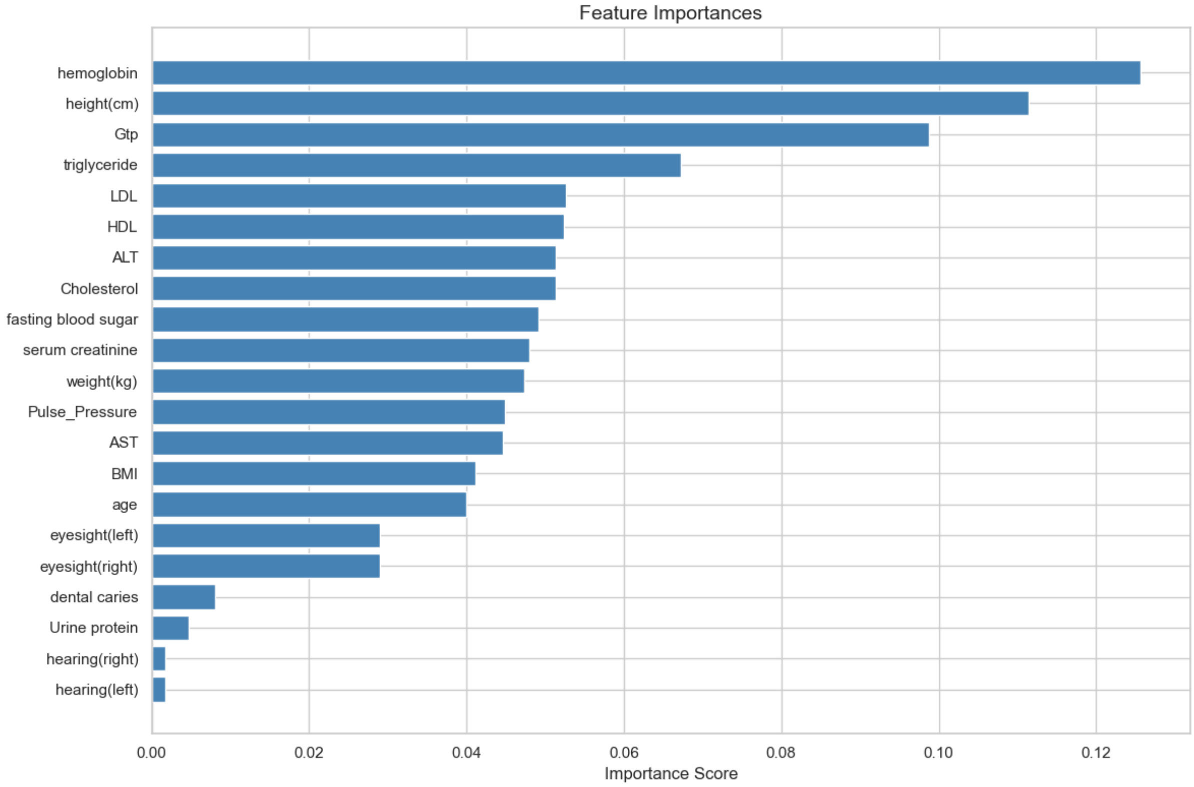

Figure 5: Distribution of Key Features by Smoking Status

5 Conclusion

This study demonstrates that smoking status can be predicted effectively using physiological and biochemical biomarkers. The Deep Neural Network achieved the highest accuracy (75.63%) due to its ability to learn complex non-linear patterns.

Key Takeaways

- Log transformation and SMOTE significantly improve model performance.
- Hemoglobin and GTP are reliable indicators of smoking behavior.
- Deep Neural Networks outperform classical ML models for this task.

Future Enhancements

- Apply SHAP or LIME for model explainability.
- Evaluate ensemble models such as XGBoost or CatBoost.
- Deploy the trained model as a web application for real-time prediction.