Hepatitis C Prediction Model: Data Analysis and Neural Network Implementation - Practice 2

Igor Vons[a], Wassim Bouzarhoun[a] and Endika Aguirre[a]

1. Dataset Overview

The hepatitis C dataset contains 615 samples with the following features:

- Laboratory measurements: ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT
- Demographic information: Age, Sex
- Target variable: Disease category (Healthy vs Hepatitis C)
- An index column (Patient ID)

2. Data Exploration

2.1. Data Quality Assessment

The dataset quality analysis reveals several important characteristics:

2.1.1. Missing Values Analysis

The dataset contains missing values in several laboratory measurements, as shown in 1.

*

September 28, 2025

Table 1: Missing values by feature

Feature	Missing Count
ALB	1
ALP	18
ALT	1
CHOL	10
PROT	1

2.1.2. Target Distribution

The distribution of disease categories shows a significant class imbalance, as detailed in

The analysis shows that the dataset is heavily imbalanced, with healthy blood donors comprising 87.8% of the samples, while the three disease categories make up only 12.2% of the data. This imbalance must be addressed during model training.

2.2. Feature Distributions

The violin plots in 1 show the distribution of each laboratory measurement.

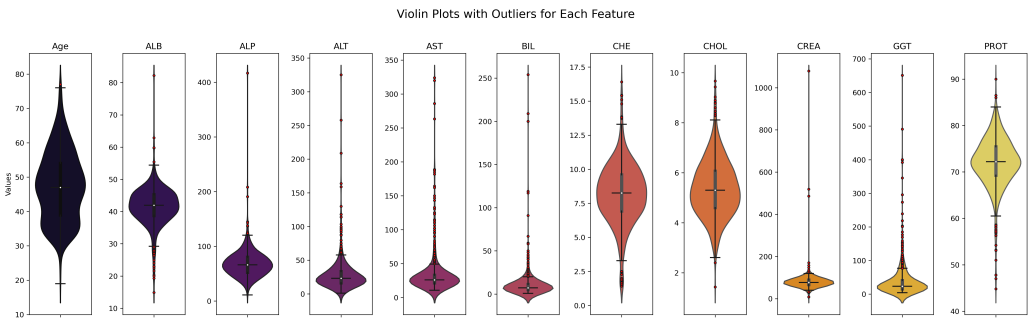

Figure 1: Distribution of laboratory measurements with outliers highlighted

2.3. Correlation Analysis

Reviewing the correlation matrix, we don't find much strong dependeces. The most notable correlations are: PROT and ALB (0.56 positive), GGT and AST (0.49 positive), GGT and ALP (0.45 positive), CHOL and CHE (0.43 positive), Age and Patient ID (0.42 positive)

3. Data Preprocessing

The data preprocessing steps included:
1. Handling missing values using median imputation.
2. Encoding the categorical variable "Sex" using one-hot encoding.
3. Scaling numerical features using StandardScaler.
4. Splitting the data into training and testing sets (80/20 split).

4. Model Development

The neural network model was developed using PyTorch and trained on the preprocessed dataset. The architecture shape is (12-128-64-32-2) consists of an input layer, two hidden layers with ReLU activation, and an output layer with a sigmoid activation function for binary classification.

5. Model Evaluation

5.1. Prediction Performance

Currently, the model achieves a great accuracy for the No Hepatitis C class, but struggles with the Hepatitis C class due to the significant class imbalance in the dataset. The confusion matrix in 2 illustrates this performance.

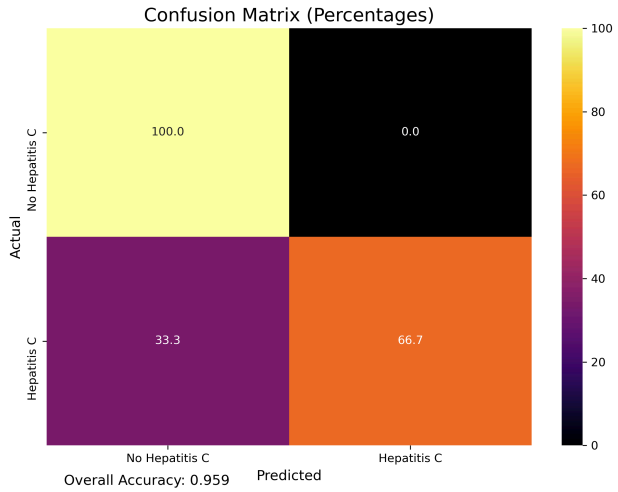

Figure 2: Confusion Matrix of the Neural Network Model

5.2. Calibration Analysis

The calibration curve in 3 indicates that the model's predicted probabilities are not well-calibrated, particularly around 0.5. This suggests that the model tends to be overconfident in its predictions of No Hepatitis C cases.

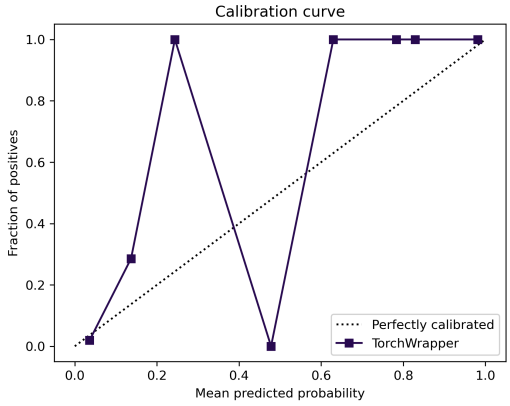

Figure 3: Calibration Curve of the Neural Network Model