COPD Risk Prediction Using Machine Learning

Model Development, Evaluation, and Tuning

Vishnu Sai

November 27, 2025

1 Introduction

The goal of this project is to build and compare multiple machine learning models to predict whether a patient is at risk of Chronic Obstructive Pulmonary Disease (COPD). The dataset contains clinical, demographic, and biochemical variables for over 44,000 patients. The final predictions were submitted to Kaggle, and the F1-score was used as the primary evaluation metric due to class imbalance.

2 Exploratory Data Analysis and Preprocessing

The dataset consists of 44,553 rows and 27 columns. Initial exploration revealed:

- No missing values in any feature.
- No duplicate patient IDs.
- Three categorical variables: `sex`, `oral_health_status`, and `tartar_presence`.
- `oral_health_status` contained only the value "Y" and was removed.
- `patient_id` was unique and removed because it carries no predictive information.

After converting categorical columns to numerical form (`Y=1`, `N=0`), all remaining features were numeric.

2.1 Distribution Analysis

Histograms were generated for all numeric variables. The distributions were found to be medically realistic:

- Metabolic and enzyme-related variables (e.g., triglycerides, GGT, AST) exhibited strong right-skewness, which is expected in biochemical data.

- Vision and hearing metrics showed high skewness due to clinical measurement scales.
- Anthropometric variables (height, weight, waist circumference) and blood pressure values were approximately symmetric.

Outliers: Boxplots and IQR-based detection methods identified numerous values outside typical IQR ranges. However, these outliers represent genuine biological variability rather than data-entry errors, so no removal or capping was applied.

2.2 Scatterplots and Class Separability

Scatterplots and pairplots showed:

- Considerable overlap between the COPD-positive and COPD-negative classes across individual features.
- No single feature provided linear class separability.
- Visual patterns suggested multivariate or nonlinear decision boundaries.

This justified the use of models capable of nonlinear decision functions, such as SVMs with RBF kernels and Neural Networks.

2.3 Correlation Heatmap

The correlation matrix indicated:

- Physiologically coherent clusters such as:
  - Blood pressure cluster: systolic and diastolic.
  - Lipid profile cluster: cholesterol, HDL, LDL, triglycerides.
  - Body measurement cluster: height, weight, waist circumference.
- No severe multicollinearity.
- Only mild correlations between features and the COPD label.

2.4 Skewness Analysis

Skewness computations showed:

- Several biomarkers exhibited substantial right-skewness (skew > 5).
- Anthropometric and blood pressure variables exhibited low skewness (< 1).

Because the skewness reflects natural biological distributions and because models such as SVMs and Neural Networks are robust to non-normal features, no transformations (e.g., log or Box–Cox) were applied.

2.5 Feature Scaling

All numeric features were standardized using `StandardScaler`. The scaler was fit on the training split only to avoid data leakage. Scaling was necessary for SVMs and Neural Networks.

2.6 Train–Validation Split

An 80–20 stratified split was used to preserve the class balance. All tuning and evaluation were performed on this consistent split.

Overall, the dataset was clean, biologically realistic, and suitable for machine learning without requiring aggressive preprocessing.

3 Models Implemented

The following models were developed:

1. Logistic Regression
2. Linear SVM (`LinearSVC`)
3. RBF Kernel SVM (attempted but computationally infeasible)
4. Neural Network (`MLPClassifier`)
5. TensorFlow-based Neural Network (lightweight architecture)

F1-score was used as the main metric to handle class imbalance.

4 Model Results Before Tuning

4.1 Logistic Regression (Baseline)

- Train F1: ≈ 0.71
- Validation F1: ≈ 0.66
- Kaggle score: 0.706

4.2 Linear SVM (Baseline)

- Train F1: ≈ 0.72
- Validation F1: ≈ 0.67
- Kaggle score: 0.709

4.3 Neural Network (Baseline)

- Train F1: ≈ 0.71
- Validation F1: ≈ 0.69
- Kaggle score: 0.706

4.4 RBF SVM

RBF SVM training and tuning were attempted, but due to dataset size (44k samples), each model required 30–60 seconds to train. Full tuning was not computationally feasible and was discontinued.

5 Hyperparameter Tuning Results

5.1 Tuned Logistic Regression

Best parameter:

$$C = 5$$

Performance:

- Train F1: 0.6678
- Validation F1: 0.6670

5.2 Tuned Linear SVM

Best parameter:

$$C = 0.01$$

Performance:

- Train F1: 0.6750
- Validation F1: 0.6739

Linear SVM performance remained close to Logistic Regression.

5.3 Tuned Neural Network (Best Model)

Using `RandomizedSearchCV`, the best configuration found was:

$$\text{hidden\_layer\_sizes} = (128, 64, 32), \quad \text{activation} = \text{tanh},$$

$$\alpha = 0.001, \quad \text{learning\_rate\_init} = 0.001, \quad \text{max\_iter} = 400.$$

Performance:

- Train F1: 0.7344
- Validation F1: 0.6810
- Kaggle Public F1: **0.724** (highest score)

This model significantly outperformed all others.

6 TensorFlow Neural Network

A shallow TensorFlow neural network was trained for comparison. It achieved:

- Validation F1: ≈ 0.676
- Kaggle score: 0.687

Due to time constraints and lack of tuning, it did not outperform the sklearn-based tuned neural network.

7 Conclusion

This project implemented and compared several ML models for COPD risk prediction. Logistic Regression and Linear SVM performed reasonably well but lacked the ability to capture nonlinearity. RBF SVM training was too computationally expensive for reliable tuning.

The tuned Neural Network with `tanh` activation and architecture $(128, 64, 32)$ achieved the highest Kaggle Public F1-score of **0.724**, making it the best-performing model.

Future improvements could include:

- Feature engineering (BMI, ratios, log transforms),
- Threshold tuning,
- More advanced Neural Network architectures,
- GPU-accelerated SVM methods.