Machine Learning Project 2
Smoker Status Prediction Using Bio-Signals

Kartikey Dubey (MT2025061) Jeevesh Rai (MT2025054)

December 12, 2025

Abstract

This report presents a comparative study of three supervised learning algorithms (Logistic Regression, Support Vector Machine, and Multilayer Perceptron) on the Smoker Status Prediction using Bio-Signals dataset from Kaggle as well as Forest Cover Type Dataset from Kaggle. The full pipeline includes data exploration, pre-processing, model training, hyperparameter tuning, and evaluation using cross-validation. The results show that on the smoker status data set, non-linear models outperform the linear baseline, with the tuned Multilayer Perceptron achieving the best overall performance and on the forest cover data set deep learning approaches significantly improved classification accuracy over linear methods. The complete source code and datasets are available at: https://github.com/kartikeyblaze/Machine_learning_project_2.

1 Introduction

In this project, we were given two data sets (Smoker Status Prediction (Binary Classification), and Forest Cover Type (Multiclass Classification).The goal was to build and compare multiple machine learning models taught in the class (Logistic Regression, SVM, MLP) using a consistent, reproducible pipeline.

2 Dataset Details

2.1 Smoker Status Dataset

- Source: Kaggle – Smoker Status Prediction using Bio-Signals(link).
- Task: Binary classification of `smoking` (0 = non-smoker, 1 = smoker).
- Number of instances and features - 389848 rows and 22 columns.
- Example features: age, height, weight, waist circumference, blood pressure, cholesterol, triglyceride, HDL, LDL, GTP, etc.

The smoker status dataset contains 38,984 anonymised health-check records with 22 numeric biosignal and laboratory features. Each instance represents a single individual, and the target variable smoking indicates whether the person is a smoker (1) or non-smoker (0). The input features include demographic attributes (such as age), anthropometric measures (height, weight, waist), blood pressure, blood chemistry (cholesterol, triglycerides, HDL, LDL), and liver function markers (AST, ALT, GTP), which together provide a rich set of routine clinical signals for predicting smoking behaviour In the given data set, around 63% were non-smokers. This indicated that the data is slightly skewed towards non-smokers, but it's still usable. In addition to this, the average age was 44 years, weight around 66 kgs.

2.2 Forest Cover Type Dataset

- Source: Kaggle – Forest Cover Type Dataset(link).
- Task: Can you build a model that predicts what types of trees grow in an area based on the surrounding characteristics?.
- Number of instances and features - 581012 rows and 55 columns.
- Example features: Elevation, Aspect, Slope, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology etc.
- Target variable : **Cover_Type** (The target variable Cover_Type is encoded as integers 1–7, corresponding respectively to Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, and Krummholz.)
- Types of features: Continuous (elevation, aspect, slope, distances, hillshade) as well as Binary(4 wilderness areas, 40 soil types).

This dataset contains tree observations from four areas of the Roosevelt National Forest in Colorado. All observations are cartographic variables (no remote sensing) from 30 meter x 30 meter sections of forest. There are over half a million measurements total! This dataset includes information on tree type, shadow coverage, distance to nearby landmarks (roads etcetera), soil type, and local topography.

3 Exploratory Data Analysis

3.1 Smoker Status Dataset

The key insights from EDA are as follows:

- Target distribution: Around 37% smokers vs 63% non-smokers.
- Uni variate plots :

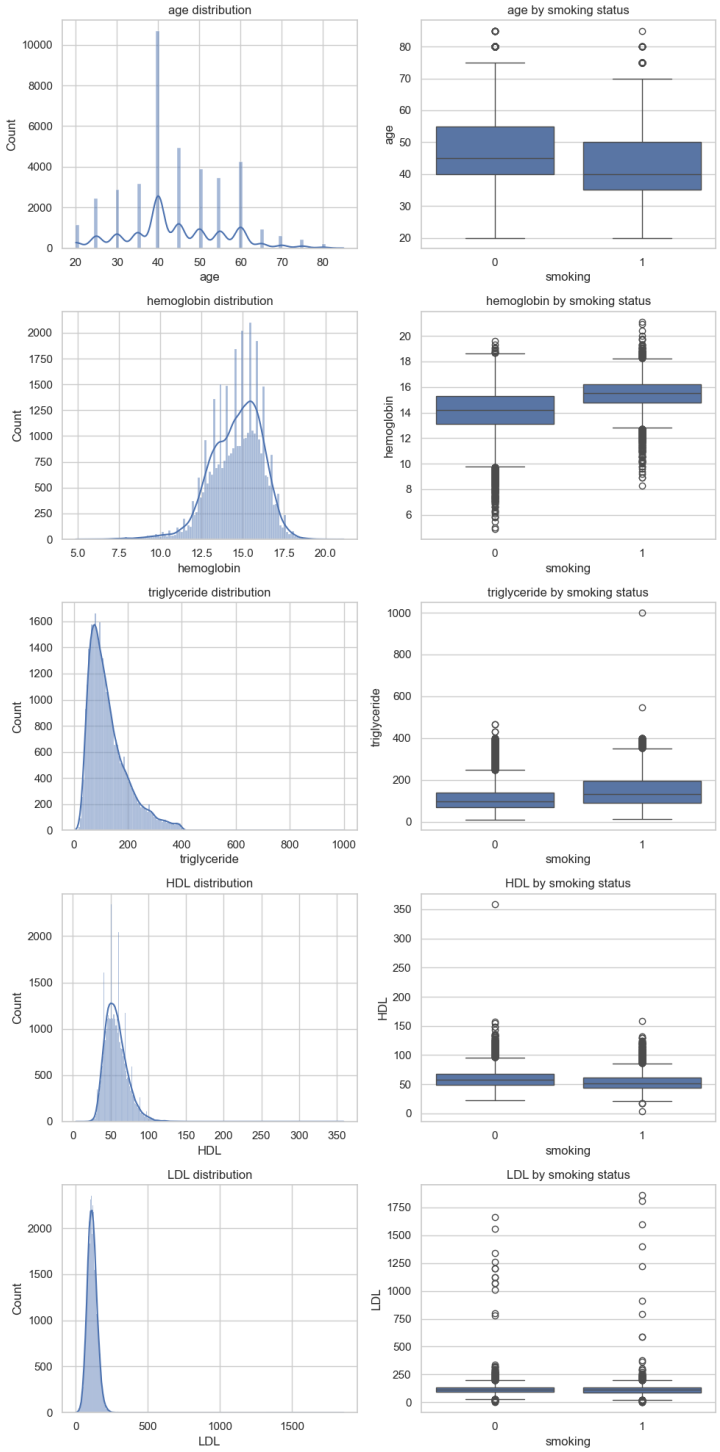
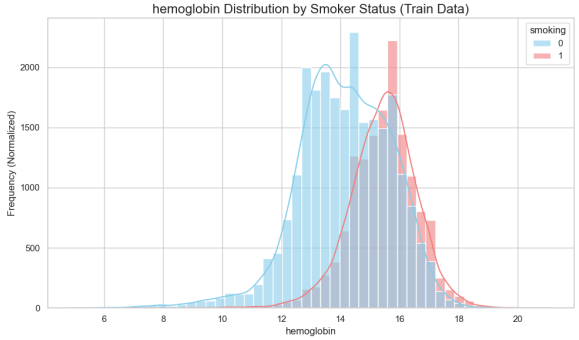


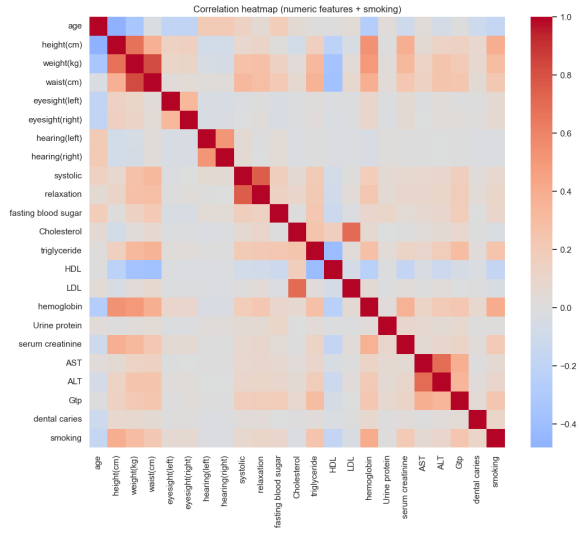
Figure 1: Distribution of a selected feature for smokers and non-smokers.

- Correlation analysis: The features which correlate most with `smoking` are shown by this

correlation heatmap in the last row/column. The ones which are most correlated to the target variable smoking are : hemoglobin (0.40), height (0.39), weight (0.29) and triglyceride (0.25).



(a) Hemoglobin distribution by smoker status.



(b) Correlation heatmap of numeric features.

Figure 2: Exploratory data analysis plots for the smoker dataset.

- PCA visualisation: In the 2D PCA space, the two classes are heavily mixed, suggesting that the data is not linearly separable.

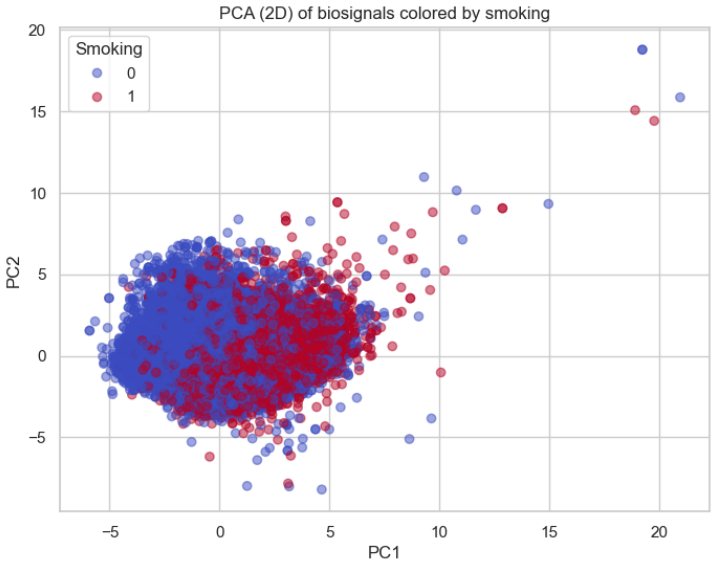


Figure 3: PCA projection coloured by smoking status.
No separate clusters formed (which is indicative of linear separability)

As we can see, from the plots and PCA, that the data is not linearly separable, so this observation motivates the use of non-linear models such as RBF-SVM and MLP, using Logistic regression as a base for feature importance only.

3.2 Forest Cover Dataset

The key insights from EDA are as follows:

- Target class distribution: Two cover types(1 & 2) form most of the data, some rare classes(4) $< 2\%$.

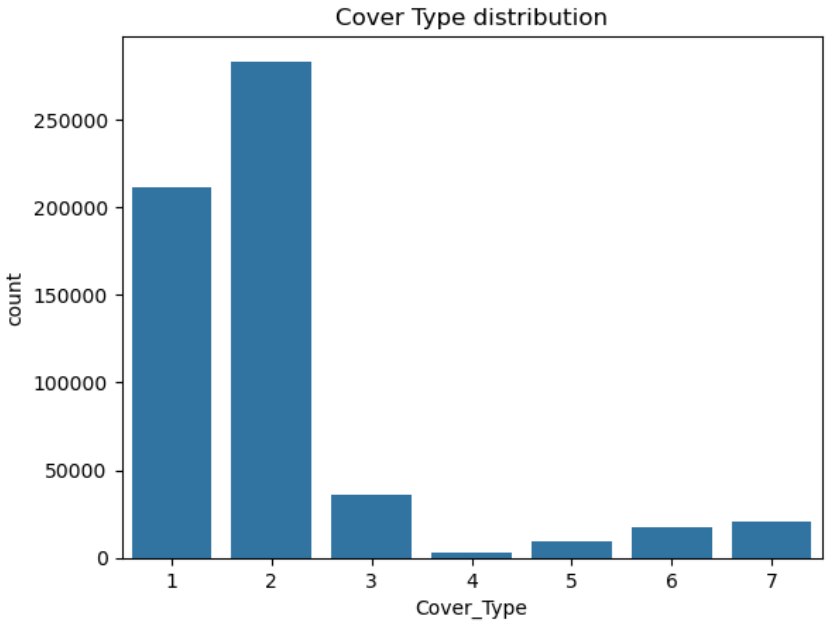


Figure 4: Target class distribution showing imbalance.

- Univariate Plots:

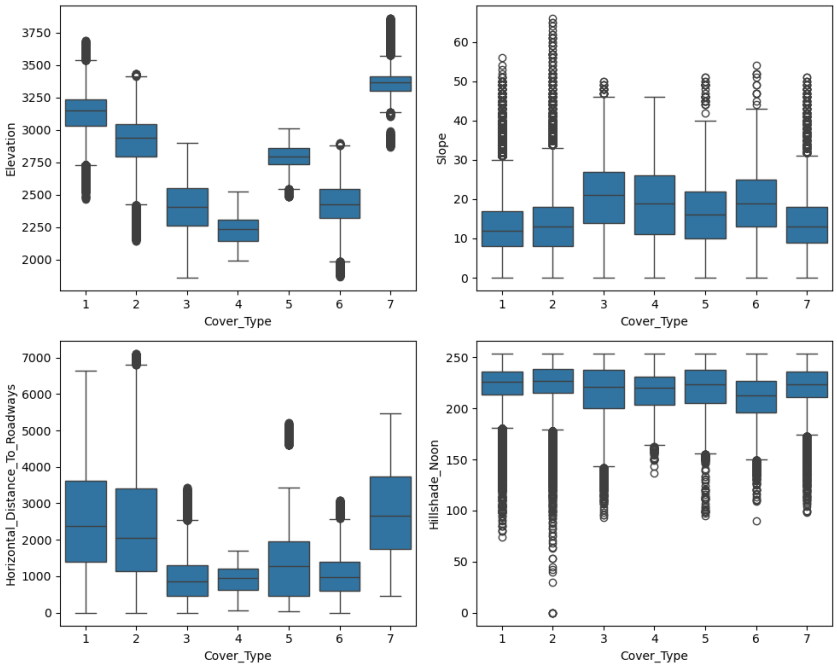
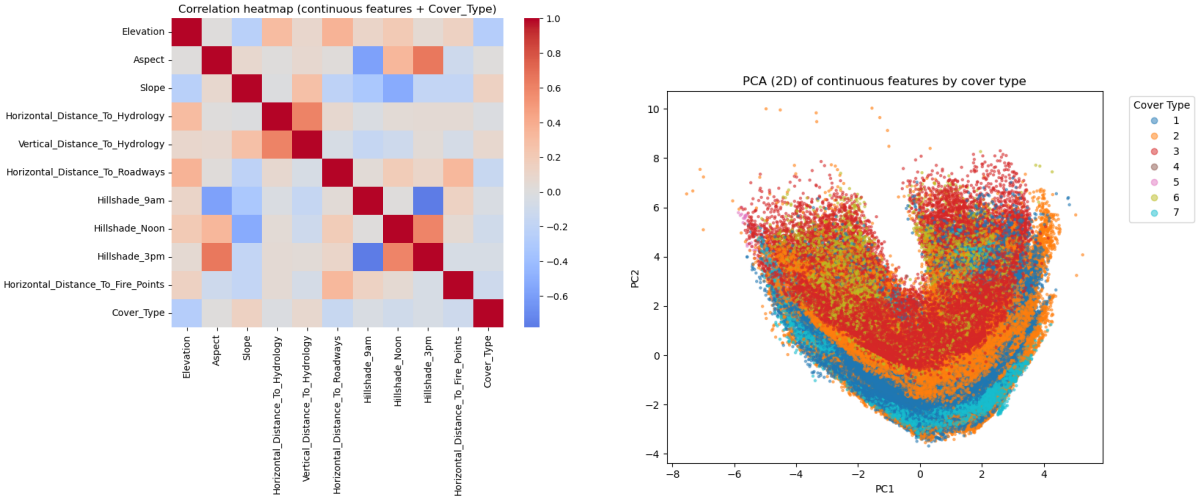


Figure 5: It can be seen that Cover type 7 (Krummholz) distinctly grows at higher elevation than the rest of the group.

- Correlation matrix and PCA:



(a) Correlation heatmap of numeric features.

(b) PCA in 2D

Figure 6: (a) Elevation strongly correlated with cover type; (b) some classes separate partially in PCA, but many overlap, so expect non-linear boundaries and multi-class complexity

4 Preprocessing Steps

4.1 Smoker Status Dataset

The preprocessing pipeline applied before each model follows this sequence:

- Separation of features and target: X (all numeric features), y (`smoking`).
- Handling of missing values - None detected in this case.
- Standardization of numeric features using `StandardScaler`. This is implemented inside scikit-learn pipelines for all models.
- Use of stratified cross-validation (e.g., 3-fold) to preserve class balance in each fold.
- All experiments use fixed random seeds (e.g. `random_state = 42` in cross-validation and MLP initialization) to ensure that model training and cross-validation splits are reproducible.

4.2 Forest Cover Dataset

- **Feature Separation:** The target variable `Cover_Type` (7 classes) was separated from the predictors. The `Id` column was removed as it carries no predictive information.
- **Continuous Features:** The 10 continuous features (Elevation, Aspect, Slope, Distances, Hillshade metrics) were standardized using `StandardScaler` to have zero mean and unit variance. This is critical for distance-based models like SVM and gradient-based models like MLP and Logistic Regression.
- **Binary Features:** The 44 binary indicator columns (4 Wilderness Areas, 40 Soil Types) were passed through the pipeline unchanged (`passthrough`), as they are already on a 0/1 scale and standardizing sparse binary data destroys its structure.
- **Stratified Splitting:** Due to the class imbalance in the target variable, all model evaluations used `StratifiedKFold` cross-validation ($k = 3$) to ensure that the proportion of each forest cover type was preserved in every training and validation fold.

5 Models and Hyperparameters

5.1 Smoker Status Dataset

5.1.1 Logistic Regression

Describe the model and tuning:

- Pipeline: StandardScaler + LogisticRegression.
- Penalty: L2, solver (e.g., `liblinear` or `lbfgs`).
- Hyperparameter search over $C \in \{0.01, 0.1, 1, 10, 100\}$ using grid search with ROC–AUC as the scoring metric.
- Best LR C was found to be = 100 .

5.1.2 Support Vector Machine (RBF)

- Pipeline: StandardScaler + SVC with RBF kernel.
- Hyperparameters tuned: $C \in \{1, 10\}$, $\gamma \in \{\texttt{scale}, 0.1\}$.
- Grid search configuration: stratified 3-fold CV, scoring by ROC–AUC.

Table 1: Top SVM hyperparameter combinations based on 3-fold CV ROC–AUC.

C	γ	ROC–AUC (mean)	ROC–AUC (std)
1	scale	0.8271	0.0042
1	0.1	0.8266	0.0033
10	scale	0.8255	0.0029
10	0.1	0.8200	0.0015

- A baseline RBF–SVM with default hyper-parameters (C=1, γ=scale) achieved a mean ROC–AUC of 0.8271 (±0.0042) over 3-fold CV. After grid search over C∈1,10 and γ∈scale,0.1, the best configuration remained C=1, γ=scale, with the same ROC–AUC (0.8271 ± 0.0042), indicating that the default SVM hyper- parameters are already near-optimal for this dataset.

5.1.3 Multilayer Perceptron

- Pipeline: StandardScaler + MLPClassifier.
- Activation: ReLU, solver: Adam, `max_iter` = 500.
- Hyperparameter grid:
  - Hidden layers: $(50), (100), (50, 50), (100, 50)$.
  - Regularization: $\alpha \in \{0.0001, 0.001, 0.01\}$.
  - Learning rate: `learning_rate_init` $\in \{0.001, 0.01\}$.
- The best combination we found was : One hidden layer of size 50, $\alpha = 0.01$, `learning_rate_init` = 0.01 .

5.2 Forest Cover Dataset

5.2.1 Logistic Regression

A Logistic Regression model was implemented as a linear baseline to benchmark performance on the multiclass forest cover problem.

- **Pipeline:** The preprocessed features were fed into a `LogisticRegression` classifier.
- **Configuration:** The model used the `lbfgs` solver, which supports multinomial loss natively, allowing it to handle the 7-class target directly without needing a One-vs-Rest wrapper.
- **Performance:** The baseline model achieved the following performance metrics via 3-fold stratified cross-validation:
  - **Accuracy:** 0.7246 ± 0.0002
  - **Macro F1-Score:** 0.5317 ± 0.0014
- **Key Insight:** Feature importance analysis (based on average absolute coefficients) revealed that `Elevation` is the single most discriminative feature, followed by distance to roadways and fire points. While the model provides a reasonable baseline, the modest accuracy suggests that the relationship between cartographic features and forest cover types is highly non-linear, motivating the use of more complex models like SVM and MLP.
- The top features found in LR are:

Feature	Importance
Elevation	3.3954
Wilderness_Area1	2.9270
Soil_Type4	1.5623

Table 2: Top 3 Features by Coefficient

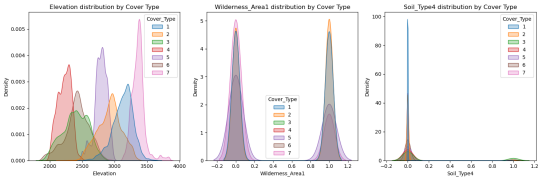


Figure 7: Visual comparison of importance

5.2.2 Support Vector Mahcine

- Two distinct SVM strategies were implemented to address the large scale of the dataset ($N \approx 581,000$):
  - Linear SVM (LinearSVC): To assess linear separability efficiently, we utilized the LinearSVC implementation with dual=False, which optimizes the primal objective function ($O(N)$ complexity) rather than the dual. This configuration is critical for minimizing training time on high-dimensional datasets where the number of samples significantly exceeds the number of features.
  - Non-Linear SVM (RBF Kernel): To capture complex decision boundaries, a standard SVC with a Radial Basis Function (RBF) kernel was employed. Due to the quadratic complexity ($O(N^2)$) of the RBF solver, training was restricted to a stratified subset of 25,000 samples. Key hyperparameters included C=10 (regularization) and gamma='scale' (kernel coefficient).

5.2.3 Multi Layer Perceptron

- A deep feedforward neural network was designed to model the non-linear interactions between geological features. The architecture and optimization parameters were tuned as follows:
  - Architecture: Two hidden layers with 128 and 64 neurons respectively (128, 64), using ReLU activation functions to introduce non-linearity.

* Optimization: The Adam solver was selected for its adaptive learning rate capabilities, which is superior for non-convex loss landscapes compared to stochastic gradient descent.
* L2 Regularization: alpha=0.0001 to prevent overfitting.
* Early Stopping: Enabled to terminate training when validation loss failed to improve for 10 consecutive epochs, preventing wasteful computation.
* Batch Size: Set to 256 to stabilize gradient updates and improve training throughput.

6 Evaluation Metrics

6.1 Smoker Status Dataset

All models are evaluated with stratified k-fold cross-validation ($k = 3$), using the following metrics:

- Accuracy.
- ROC–AUC.
- F1-score (for the positive class).

The final comparison table for the smoker dataset:

Table 3: Cross-validated performance on the smoker status dataset.

Model	Accuracy	ROC–AUC	F1-score
Logistic Regression	0.7231	0.8070	0.6006
SVM (RBF)	0.7516	0.8271	0.6563
MLP	0.7509	0.8306	0.6632

LR (80.6%) < SVM (82.71%) confirms the PCA insight - the data needs non-linear modeling. MLP has best ROC-AUC and best F1, even if accuracy is almost tied with SVM. Differences between SVM and MLP are small; pushing MLP harder (more layers, larger grids) will cost time and may overfit, with only tiny gains. Data is non-linearly separable, so non-linear models (SVM, MLP) clearly beat Logistic Regression.Among non-linear models, MLP is slightly better overall, especially on F1 and ROC-AUC, it is the best model for this dataset.

6.2 Forest Cover Dataset

- The quantitative performance of the three classification architectures (Logistic Regression, Support Vector Machines, and Multi-layer Perceptron) is summarized below. The models were evaluated based on Accuracy, F1-Macro Score (to account for class imbalance), and Total Computational Time (training plus validation).

Model Architecture	**Accuracy**	**F1-Macro**	**Time (s)**
MLP (Optimized)	**0.9088**	**0.8641**	332.31
SVM (RBF - Subset)	0.8135	0.7144	126.56
Logistic Regression	0.7249	0.5328	211.07
SVM (Linear)	0.7129	0.4614	**51.19**

Table 4: Performance metrics for Forest Cover Type classification. The MLP model achieves the highest predictive performance, while the Linear SVM offers the fastest execution time.

- The Optimized MLP achieves a dominant accuracy of 90.88%, outperforming the linear baselines by nearly 20 percentage points. This significant gap confirms that the decision boundaries between forest cover types are highly non-linear and cannot be effectively captured by simple hyperplanes (Linear SVM/Logistic Regression).The MLP's high F1-Macro score of 0.8641 is particularly notable. While the Linear SVM achieved 71% accuracy, its low F1-Macro score (0.4614) suggests it biased heavily toward the majority classes (Spruce/Fir and Lodgepole Pine) and failed to correctly identify rare classes (Cottonwood/Willow). The MLP maintained precision across all classes.

- Linear SVM was the fastest model (51.19s), making it suitable only for rapid prototyping where accuracy is secondary.RBF SVM provided a middle ground in accuracy (81.35%) but required a massive data reduction (trained on only 5% of samples) to remain computationally feasible.The MLP, despite a longer training time of 332.31s, offers the best trade-off for deployment, providing superior accuracy with manageable training costs on modern hardware.

7 Comparative Analysis

Discuss the relative performance of the models:

7.1 Smoker Data Set

- Logistic Regression performs worst, which is consistent with the non-linearly separable structure observed in PCA space.
- SVM with RBF kernel improves all metrics, showing the benefit of non-linear decision boundaries.
- The tuned MLP achieves very similar accuracy to SVM but slightly higher ROC–AUC and F1-score, indicating better positive-class discrimination i.e across all thresholds, MLP ranks smokers vs non-smokers slightly better and better balance between precision and recall for the smoker class, which matters more than raw accuracy when classes are not perfectly balanced.
- Although SVM and MLP achieve almost identical accuracy, the MLP attains higher ROC-AUC and F1-score, indicating better discrimination between smokers and non-smokers, especially near the decision boundary. Given the non-linear structure of the biosignal data, the MLP's learned hidden representations appear to capture complex feature interactions slightly better than the fixed RBF kernel of the SVM, making it the preferred model for this task.

Logistic regression works well with linearly separable data set, as it just tries to find a line which divides the two classes of data, here it was not so simple so as expected it performed worse than the other two. SVM without Kernel also works best on linearly separable data, so using Kernel SVM was the wiser choice, which resulted in better performance than Logistic regression, but the huge downside was the runtime of the algorithm, while LR finished in under 2 mins, RBF-SVM took around 15-20 minutes, showing that it requires much more computation. Surprisingly, the optimal parameters found by RBF-SVM were the same as the default one it uses. Finally, Multi layer Perceptron was used, which took less time than SVM but produced arguably better results(similar accuracy, better F1 score and ROC-AUC).

7.2 Forest Cover Data Set

- Final verification on the strictly held-out test set confirmed the robustness of the Optimized MLP, achieving a final accuracy of 91.37% and an F1-Macro score of 0.8694. The slight improvement over the cross-validation mean (90.88%) demonstrates that the model generalizes effectively to unseen data without overfitting.
- Although Logistic Regression provided a stable baseline with 71% accuracy, its performance is limited by its linear decision boundaries. The feature importance analysis suggests that variables like Elevation and Soil Type interact in complex ways that a simple linear equation cannot fully capture. To address this underfitting and capture non-linear spatial patterns in the forest terrain, we proceed to investigate Support Vector Machines (SVM) with non-linear kernels.

- Both Logistic Regression (72.49%) and Linear SVM (71.29%) hit a "performance ceiling" around 72%. This strongly suggests that the decision boundaries between forest cover types are non-linear. No amount of hyperparameter tuning on linear models would likely bridge the 18% gap to the MLP, as the underlying geometry of the data (elevation, soil type interactions) is inherently complex.
- The RBF SVM is theoretically capable of modeling non-linear data, and it improved accuracy to 81.35%. However, to make training feasible, it was trained on a subset of only 25,000 samples. Despite seeing only 5% of the data, it beat the linear models. However, it scales poorly; prediction times are slow because the model must calculate distances to thousands of support vectors, making it less practical for deployment than the MLP.
- The Optimized MLP (Acc: 90.88%) outperformed the next best model by nearly 10 percentage points. The F1-Macro score of 0.8641 indicates that the MLP did not just predict the dominant classes well, but also maintained high precision and recall across the rarer forest cover types (Classes 4 and 5). The training time of 5.5 minutes (332s) was a reasonable trade-off for this massive accuracy gain.
- The MLP's high F1-Macro score of 0.8641 is particularly notable. While the Linear SVM achieved 71% accuracy, its low F1-Macro score (0.4614) suggests it biased heavily toward the majority classes (Spruce/Fir and Lodgepole Pine) and failed to correctly identify rare classes (Cottonwood/Willow). The MLP maintained precision across all classes.

8 Final Conclusions

8.1 Smoker Dataset

- Non-linear models (SVM and MLP) clearly outperform the linear baseline on this biosignal dataset.
- Among the tested models, the Multilayer Perceptron achieved the best overall performance and is recommended as the final model for smoker status prediction.
- Logistic regression works well with linearly separable data set, as it just tries to find a line which divides the two classes of data, here it was not so simple so as expected it performed worse than the other two. SVM without Kernel also works best on linearly separable data, so using Kernel SVM was the wiser choice, which resulted in better performance than Logistic regression, but the huge downside was the runtime of the algorithm, while LR finished in under 2 mins, RBF-SVM took around 15-20 minutes, showing that it requires much more computation. Surprisingly, the optimal parameters found by RBF-SVM were the same as the default one it uses. Finally, Multi layer Perceptron was used, which took less time than SVM but produced arguably better results(similar accuracy, better F1 score and ROC-AUC).

8.2 Forest Cover Dataset

- For the Forest Cover Type dataset, Deep Learning (MLP) is the definitive choice. It provides the highest accuracy and robustness across all classes. While Linear SVMs offer speed, they sacrifice too much predictive power (nearly 20% accuracy drop). Future improvements could focus on further deepening the MLP architecture or exploring ensemble methods (Random Forests/XGBoost), but for the scope of this comparison, the MLP is the clear winner.

The project highlights the importance of feature scaling, appropriate model choice for non-linear data, and modest hyperparameter tuning for improving predictive performance.