ST443 Group Project — Task 1: Hyperspectral Land Classification

Ethan, Matthias, Till, and Tommy
MSc Data Science, London School of Economics

Abstract

This report presents the analysis and modelling results for task 1 of the ST443 Group Project. Firstly, to understand the data, we performed comprehensive exploratory data analysis (EDA) of the provided hyperspectral dataset. It documents each of the 215,604 pixels forming a satellite image of an alpine region in Tyrol, Austria; detailing each pixel's surface-reflectance measurements for 218 reflectance bands and its spatial coordinates (p_x, p_y). Following our EDA, we proceed to train and evaluate seven types of supervised classifiers. Each classifier is trained twice, once on the raw features, then separately after reducing the feature set to 10 components via Principal Component Analysis (PCA). We evaluated models using accuracy, misclassification error, macro balanced accuracy, macro F1, and macro AUC. Finally, we apply our understanding of the data to train three classifiers to conduct a binary classification experiment: identifying glacier ice pixels versus all other classes, evaluated with the F1 metric.

1 Data Overview

The dataset is comprised of 215,604 observations and 221 feature columns (218 spectral bands and two spatial coordinates (p_x, p_y)), each representing a pixel of a satellite image of an alpine region in Tyrol, Austria. The target variable `land_type` represents eight land cover classes. Each of the bands' reflectance values are continuous and have class-means constrained between 0 and 1.

Before EDA, we coerced all features to numeric format, and checked for missing values (there were none). We stored class labels as categorical factors for interpretability, then examined class counts discovering an imbalance. Vegetation and soil categories (e.g., alpine meadow, valley floor / meadow) are more common, whilst glacier and rock classes contain fewer samples (e.g., snow / ice, dark rock). The observed class imbalance supported our use of macro-averaged metrics (like macro F1 or macro balanced accuracy) for evaluation, since relying on overall accuracy alone could allow models to achieve high scores by simply predicting the majority class(es) and neglecting the minority one(s).

2 Exploratory Data Analysis (Task 1.1)

2.1 Class Distribution

Figure 1 depicts the distribution of observations across land types (classes). The largest classes represent vegetation and soil surfaces (e.g., $alpinemeadow$, $alpinetundra$, and $valleyfloor/medow$). Meanwhile, glacier and rock surfaces (e.g., $darkrock$, $snow/ice$, and $scree/sunlitrock$) are underrepresented, further emphasizing the need for robust evaluation measures beyond raw accuracy.

Preprint. Under review.

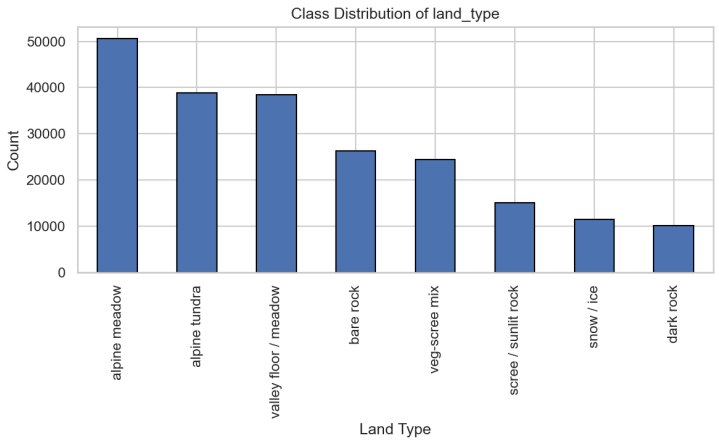

Figure 1: Distribution of observations by class. Vegetation and soil types dominate, whereas glacier and rock are rare.

2.2 Reflectance Characteristics

Our analysis of per-band descriptive statistics revealed that reflectance values span a broadly consistent numerical range, with nearly all minima near zero and maxima typically below 1.6. Figure 2 shows histograms for six representative bands, illustrating the diversity of spectral responses across wavelengths.

At lower wavelengths (e.g., bands 1 and 44), reflectance distributions are sharply right-skewed, dominated by low reflectance values with a long upper tail. In contrast, mid- to long-wavelength bands (e.g., bands 87, 131, 174, and 218) have broader or even bimodal shapes. These patterns suggest that while some pixels cluster within a narrow reflectance range, certain bands capture more distinct surface contrasts or brightness variations.

The per-band summary in Table 1 reveals a mean minimum reflectance of approximately 0.002 ± 0.009, confirming that lower bounds across bands remain close to zero. The mean maximum reflectance is around 0.958 ± 0.444, indicating moderate variation in overall brightness across bands. This is further emphasised by the interquartile spread of maximum values (0.53–1.51).

Table 1: Summary statistics for per-band minimum and maximum values (range_df).

Statistic	Min	Max
Count	218	218
Mean	0.00195	0.95786
Std	0.00852	0.44396
Min	-0.09120	0.41120
25%	-0.00030	0.53165
50%	0.00355	0.79470
75%	0.00638	1.51405
Max	0.01070	1.61380

In summation, we found that whilst the reflectance ranges are comparable across wavelengths, the distributional shapes vary appreciably. Based on the observed skewness and occasional bimodality across bands we expect that models assuming multivariate normality and equal covariance, such as LDA, may be less well suited for this dataset. We would expect approaches that relax these assumptions, like QDA, or that estimate class boundaries directly without covariance modelling, like Logistic Regression, to perform more robustly. Furthermore, non-parametric or tree-based methods might better capture complex spectral relationships, though at the cost of computational efficiency.

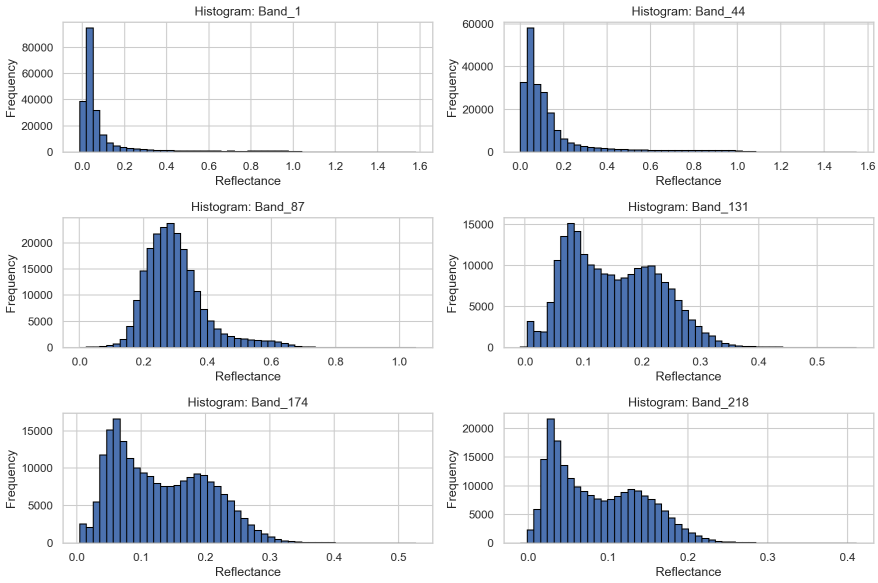

Figure 2: Example histograms of six spectral bands. Reflectance values span a similar range across wavelengths but vary in shape, from sharply right-skewed at low wavelengths to broader or bimodal at higher ones.

2.3 Spectral Signatures

Next, we examined the mean spectral signatures for the eight classes (Figure 3), revealing clear, wavelength-dependent differences in reflectance.

The "snow / ice" class shows the highest reflectance at the beginning of the spectrum but then declines sharply shortly after band 100, remaining the lowest thereafter. This steep drop indicates a strong change in how light interacts with the surface at higher band indices. The rock-related classes ("dark rock," "scree / sunlit rock," and "bare rock") generally maintain moderate reflectance throughout, suggesting relatively limited variation in reflectance across wavelengths. In contrast, the vegetation- and soil-related classes ("veg-scree mix," "alpine tundra," "alpine meadow," and "valley floor / meadow") start with low reflectance, then peak suddenly roughly between bands 50 to 130, before gradually decreasing again.

From this analysis we infer that class separability arises both from differences in overall reflectance and the specific shape of each class's reflectance curve. Accordingly, we would expect the "snow / ice" class to be easily distinguishable from vegetation, soil, or rock related surfaces, whereas distinguishing between vegetation types will likely be more challenging as their curves share similarities in shape and reflectance levels across bands.

From a modelling perspective, the curved and non-parallel nature of the class mean signatures could present a challenge for models relying on linear decision boundaries. LDA, under its equal covariance and linear separability assumptions, might therefore underperform. Meanwhile, QDA can accommodate for the differing curve shapes by estimating separate covariance structures per class, whilst Logistic Regression, though linear in its basic form, may still capture key distinctions through feature scaling or interaction terms as it models class boundaries directly rather than relying on covariance estimation. Lastly, non-parametric or tree-based classifiers remain attractive for their ability to capture complex, non-linear relationships between bands, though this flexibility comes at the cost of increased computational demand.

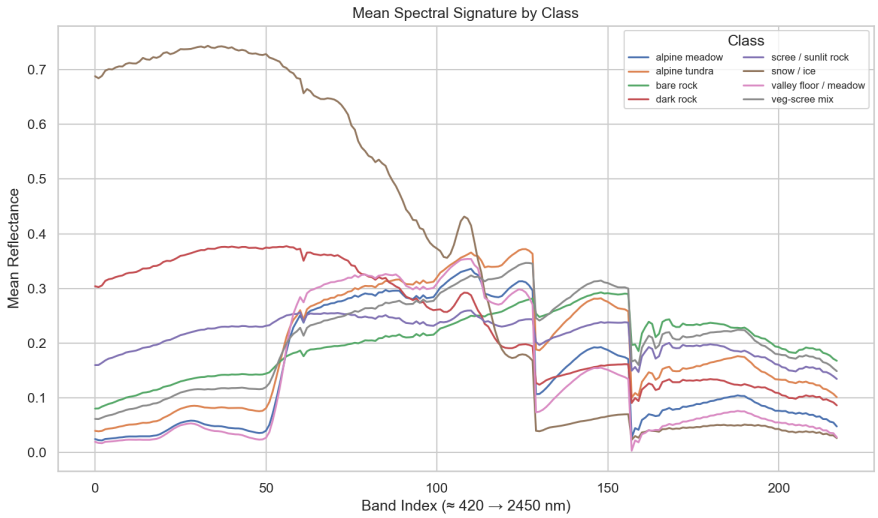

Figure 3: Mean spectral signature by land type. Each line represents the average reflectance curve across all spectral bands for one class.

2.4 Inter-band Correlation and Dimensionality

To assess redundancy across bands before dimensionality reduction, we examined inter-band correlations. The subset correlation matrix (Figure 4) shows strong correlations between neighbouring bands (often above 0.85). We interpret these blocks as smooth transitions in reflectance between neighbouring wavelengths, rather than abrupt spectral changes, implying that consecutive bands carry overlapping information. Accordingly, Principal Component Analysis (PCA) could effectively reduce dimensions, capturing the main sources of spectral variation while removing noise and redundancy. We found that the first ten principal components account for approximately 99.1% of the total variance, further supporting the use of PCA(10) as a compact yet information-preserving representation for subsequent modelling.

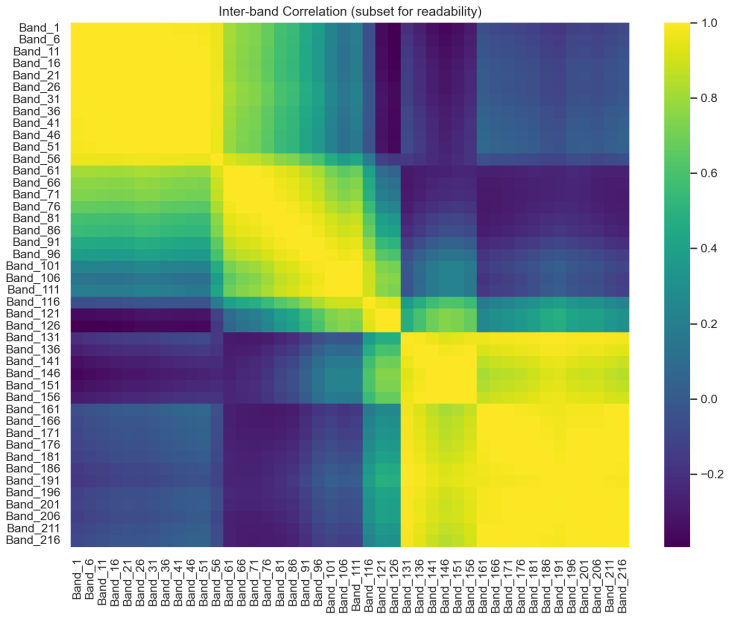

Figure 4: Inter-band correlation among every 5th band. Strong adjacent correlation supports PCA.

2.5 2D PCA Visualization

Motivated by the strong inter-band correlations observed earlier, we visualized the dataset in a reduced two-dimensional PCA space to visually explore how much class separability is retained when compressing the 218 spectral bands into just two principal components. Figure 5 presents a sampled view of this projection, where each point represents one pixel colored by its land-type class.

The "snow / ice" class forms the most clearly distinct cluster on the left, separated from the rest of the data along the first principal component. This confirms its notably different spectral behavior identified in the mean reflectance analysis above. In comparison, the rock, vegetation, and soil-related classes occupy largely distinguishable but overlapping regions, arranged in a continuous gradient rather than discrete clusters. From this pattern we infer that whilst the first two components capture the dominant variation, there remain substantial similarities and gradual transitions amongst rock, vegetation, and soil covered surfaces.

Through this visualization we conclude that PCA effectively compresses the dataset while preserving key spectral distinctions. However, we also illustrate the inherent challenge of classifying land types with subtle spectral differences. Even after dimensionality reduction, the data form smooth transitions rather than sharply separable groups, reinforcing the need for classification methods capable of modelling complex, non-linear boundaries between classes.

Figure 5: Projection of 30,000 sampled observations onto the first two principal components. Each point represents one pixel, colored by its land-type class.

3 Methodology (Task 1.2)

Following our exploratory analysis, we applied the seven required classifiers to land-type classification (covering linear, distance-based, and ensemble approaches):

- **Linear Discriminant Analysis (LDA):** Identifies linear combinations of spectral features that best separate the classes. Automatic shrinkage was applied to regularize covariance estimates, improving numerical stability in the presence of the strong inter-band correlations observed during EDA.
- **Logistic Regression:** A multinomial model with L2 (ridge) regularization, providing a simple yet effective linear baseline. Its interpretability and efficiency make it a strong reference model when class boundaries are approximately linear after feature scaling.
- **Quadratic Discriminant Analysis (QDA):** Extends LDA by estimating a separate covariance matrix for each class, allowing curved decision boundaries. A small regularization parameter (0.1) was used to avoid instability when fitting high-dimensional covariances.

- **k-Nearest Neighbors (k=5):** A non-parametric, instance-based method that classifies each pixel according to the majority label among its five closest neighbors. Distance weighting was applied so that nearer samples exert greater influence, reducing sensitivity to local noise.

- **Gradient Boosted Decision Trees (GBDT):** An ensemble method that builds trees sequentially, where each new tree corrects the errors of previous ones. This approach captures complex, nonlinear interactions between spectral features.

- **Random Forest:** Another ensemble method that trains many independent trees on random subsets of data and features. It is robust to outliers, noise, and multicollinearity, making it a reliable benchmark for structured, high-dimensional data.

- **Support Vector Machine (SVM, RBF):** A margin-based classifier that finds optimal separating boundaries in a nonlinear feature space. Probability calibration was enabled so the model could output class probabilities for metric evaluation.

We divided all data into training (80%) and validation (20%) sets using stratified sampling in order to preserve the proportion of each land-type class in both splits.

For algorithms that depend on feature scaling (LDA, Logistic Regression, QDA, k-NN, and SVM), we standardized the features using a `StandardScaler` so that each band had mean 0 and standard deviation 1. We trained Tree-based models (GBDT and Random Forest) on unscaled inputs because they are inherently scale-invariant.

To explore the impact of dimensionality reduction, we also trained each model on PCA-reduced features (10 principal components) computed after scaling. As established in Section 4, these components capture approximately 99.1% of total variance while removing redundancy from the original 218 bands.

We evaluated model performance on the validation set using five complementary metrics:

- **Accuracy:** overall proportion of correct predictions.

- **Misclassification Error:** complement of accuracy, included for completeness.

- **Macro Balanced Accuracy:** averages the TPR of each class to treat all classes equally, regardless of size.

- **Macro F1 Score:** evaluates how well the model labels and identifies samples for each class, then averages those results so all classes contribute equally. A high score means the model performs consistently across both common and rare classes.

- **Macro AUC:** average area under the ROC curve across one vs rest comparisons.

Together, these metrics provide a balanced evaluation of classifier performance, especially given the dataset's moderate class imbalance and varied reflectance distributions.

4 Results and Discussion (Task 1.3)

4.1 Model Summary and Comparison

Table 2 reports accuracy, macro balanced accuracy, macro F1, and macro AUC for all seven models trained on both raw features and PCA(10). Performance is consistently strong across models, with the top raw-feature models (Logistic Regression and SVM with an RBF kernel) enjoying a slight edge over the PCA(10) versions (though the differences are marginal).

Table 2: Classifier performance on the validation set. Best-in-column values by variant are **bold**. The model we ultimately select is *Logistic Regression (PCA 10)*; the reasoning is given in Section 4.2.

Model	Variant	Accuracy	Misclass Err.	Bal. Acc.	F1 (macro)	AUC (macro)
SVM (RBF)	PCA(10)	**0.9896**	**0.0104**	**0.9900**	**0.9903**	**0.999934**
Logistic Regression	PCA(10)	0.9894	0.0106	0.9878	0.9880	0.999917
Random Forest	PCA(10)	0.9784	0.0216	0.9748	0.9753	0.999706
k-NN (k=5)	PCA(10)	0.9706	0.0294	0.9701	0.9705	0.998438
GBDT	PCA(10)	0.9725	0.0275	0.9684	0.9690	0.999449
QDA	PCA(10)	0.9539	0.0461	0.9518	0.9514	0.998050
LDA	PCA(10)	0.8398	0.1602	0.8301	0.8356	0.988208
SVM (RBF)	raw	0.9907	0.0093	**0.9907**	**0.9909**	0.999948
Logistic Regression	raw	**0.9918**	**0.0082**	0.9901	0.9903	**0.999949**
Random Forest	raw	0.9851	0.0149	0.9830	0.9832	0.999865
k-NN (k=5)	raw	0.9691	0.0309	0.9684	0.9690	0.998424
GBDT	raw	0.9616	0.0384	0.9633	0.9640	0.999156
QDA	raw	0.9533	0.0467	0.9500	0.9501	0.998056
LDA	raw	0.8640	0.1360	0.8597	0.8617	0.991076

4.2 Interpretation and Choice

Performance consistency. We were excited to see that performance across models was consistently strong, with Logistic Regression and SVM (RBF) emerging as the top performers on the raw features. Logistic Regression achieved the highest Accuracy and AUC, while SVM (RBF) narrowly led in Balanced Accuracy and Macro F1. The PCA(10) variants of both models negligibly trailed, confirming that dimensionality reduction preserved nearly all discriminative information whilst substantially simplifying the feature space.

Also worth addressing is that LDA performed notably worse, as anticipated from our exploratory analysis. Since classes exhibit distinct patterns of variability and strong inter-band correlations, LDA's assumption of equal covariance across classes is violated. Accordingly, its linear boundaries oversimplify the underlying spectral structure. QDA and Logistic Regression, which relax or bypass this assumption, produce more stable and accurate classifications.

Appropriateness to the data. From our EDA, we expected that models assuming simple linear boundaries, such as LDA and Logistic Regression, might struggle to fully capture class differences. The reflectance distributions showed skewness and occasional bimodality, the mean class signatures were curved and sometimes overlapping, and strong inter-band correlations suggested that much of the variation across wavelengths was redundant but not necessarily linear. From these observations we inferred that models capable of handling distinct covariance structures or moderate non-linearity would likely perform best.

The results only partially confirmed our expectations. LDA indeed performed worse than the more flexible models, consistent with its restrictive equal-covariance assumption. However, it still achieved stronger performance than we expected, suggesting that class boundaries are not as complex as we assumed from the initial reflectance patterns. Logistic Regression, while also linear, performed exceptionally well because it benefits from standardized features and does not rely on shared covariance across classes. Its success indicates that much of the separability arises from the smooth, monotonic changes in reflectance and the broad shape differences in mean spectral curves identified in EDA.

Lastly, the consistent performance of models trained on PCA(10) features confirms our suspicion that most discriminatory information is captured by a small number of orthogonal, low-dimensional components. Through PCA, we effectively reduced redundancy amongst correlated bands without sacrificing separability. Together, our findings suggest that the dataset's class distinctions can be represented accurately within an approximately linear, low-dimensional subspace.

Interpretability. Of the top-performing models, Logistic Regression is most easily interpretable. Its coefficients, particularly when expressed in PCA space, reveal exactly which combinations of

spectral components contribute most strongly to classification. In contrast, tree ensembles and RBF-based SVMs are high-performing "black boxes," offering limited interpretability into the specific spectral regions driving their predictions.

Computational efficiency. Through PCA(10), we compressed 218 correlated bands into 10 components with negligible loss of accuracy across models. As a result, training and inference were substantially faster and more stable on PCA-reduced inputs.

Decision. In selecting our "best" model, we considered performance, interpretability, efficiency, and alignment with the data structure, ultimately selecting *Logistic Regression (PCA 10)*. While *SVM (RBF, PCA 10)* achieves slightly higher scores (by roughly 0.02 to 0.23 percentage points), these gains are marginal relative to its greater computational cost and parameter sensitivity. Because the data are already close to linearly separable after scaling and PCA compression, the non-linear flexibility of the RBF kernel offers little practical advantage. In contrast, Logistic Regression is efficient, stable, and easy to deploy, with coefficients that directly reflect how spectral features influence class membership. We therefore believe it represents the most balanced and operationally suitable choice, maintaining near-optimal performance (≈ 0.9894 accuracy, ≈ 0.9878 macro F1, ≈ 0.9999 AUC) while remaining highly interpretable and computationally efficient.

4.3 Error Analysis

To better understand remaining misclassifications, we examined per-class F1 scores and the confusion matrix for our selected model, Logistic Regression (PCA 10). Errors were concentrated among classes with similar spectral and physical characteristics (e.g., $alpinemeadow$, $alpinetundra$, and $scree/sunlitrock$). As seen in the mean spectral signatures (Figure 3), these classes share comparable reflectance levels and curve shapes across large portions of the 218 bands, making them inherently harder to distinguish. This pattern aligns with our expectations from EDA, where smooth transitions and partial overlap among vegetation- and soil-dominated surfaces were evident. In comparison, $snow/ice$ achieved near-perfect TPR and F1, consistent with its isolated cluster in the PCA projection (Figure 5). Its distinct brightness and limited spectral overlap with other classes make it the most easily separable surface type. Overall, we conclude that misclassifications correspond to natural spectral gradients rather than systematic modelling errors.

5 Task 1.4: Glacier Ice (Binary) Experiment

Setup. For this task, we isolated the *snow / ice* class and reframed the problem as a binary classification, where *snow / ice* was treated as the positive class and all other land types were grouped as negative. To compare both linear and non-linear approaches, we evaluated four classifiers: LDA (with automatic shrinkage), Logistic Regression (balanced class weights), Random Forest (balanced), and SVM (RBF kernel, balanced). Since glacier pixels are relatively rare in our dataset, we relied on the positive-class F1 score as the main evaluation metric, as it balances correct detection (TPR) against false alarms (PPV).

Table 3: Binary glacier detection results (positive = snow / ice).

Classifier	F1 (glacier = pos)
LDA (auto shrinkage)	0.934
Logistic (balanced)	0.988
Random Forest (balanced)	**0.989**
SVM (RBF, balanced)	0.977

Results and interpretation. All four classifiers achieved strong performance with F1 scores above 0.93. The Random Forest and Logistic Regression models reach nearly identical results (0.989 and 0.988 F1, respectively), demonstrating that both linear and ensemble methods can distinguish snow/ice pixels accurately. This strong performance directly reflects our observations from EDA: snow/ice exhibits a uniquely high reflectance at shorter wavelengths and a sharp decline after roughly band 100, forming a clear and isolated spectral signature (Figure 3) and a distinct cluster in PCA

space (Figure 5). These characteristics make the glacier class linearly separable from the darker, more variable surfaces.

Notably, the SVM (RBF) performs similarly (0.977 F1) but does not outperform the simpler Logistic Regression baseline, suggesting that the additional non-linear flexibility offers little benefit once the main linear trend in reflectance is captured. LDA, although weaker overall (0.934 F1), performs better than initially expected.

Discussion. These results reinforce our earlier deductions from EDA that the $snow/ice$ class is spectrally distinct and largely linearly separable from other surfaces (explaining why even relatively simple models achieve near-perfect TPR). Ensemble methods like Random Forest provide a minor edge by capturing small non-linearities, but the balanced Logistic Regression model attains comparable accuracy while remaining faster, more stable, and more interpretable.

Acknowledgements