Supplementary Material

Title:

Machine-Learning-guided recognition of α and β cells from label-free infrared micrographs of living human islets of Langerhans

Fabio Azzarello[a,*], Francesco Carli[b], Valentina De Lorenzi[a], Marta Tesi[c], Piero Marchetti[c], Fabio Beltram[a], Francesco Raimondi[b,*], Francesco Cardarelli[a,*]

Affiliations:

[a] NEST Laboratory - Scuola Normale Superiore, Piazza San Silvestro 12, Pisa, Italy.

[b] Laboratorio di Biologia Bio@SNS, Scuola Normale Superiore, 56126, Pisa, Italy

[c] Department of Clinical and Experimental Medicine, Islet Cell Laboratory, University of Pisa, Pisa, Italy.

* to whom correspondence should be addressed: fabio.azzarello@sns.it; francesco.raimondi@sns.it; francesco.cardarelli@sns.it

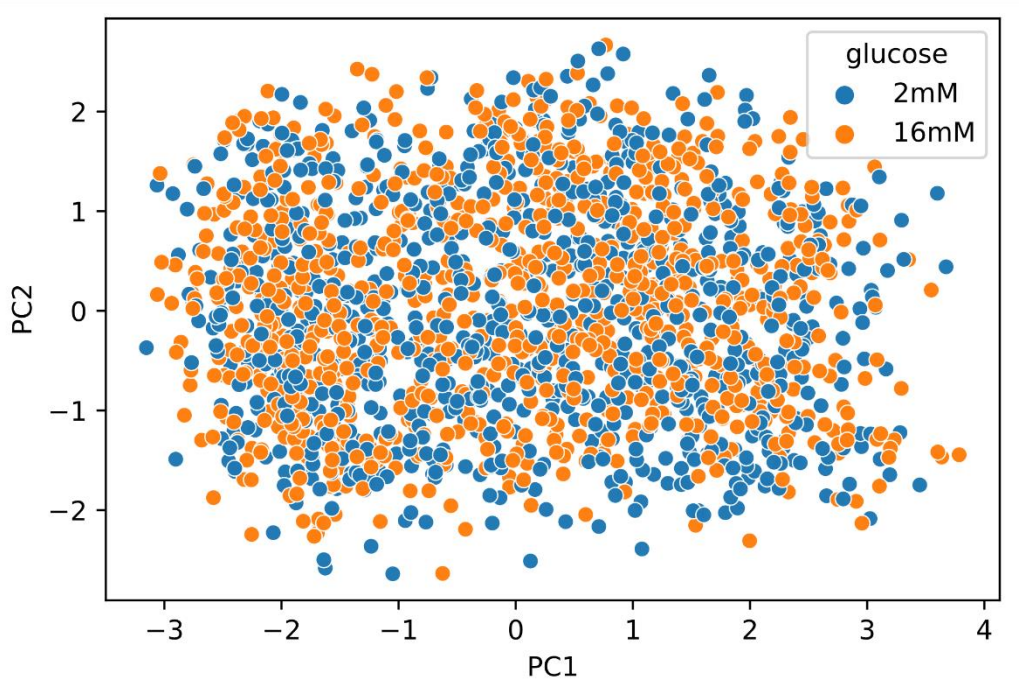

Figure S1: Effect of Experimental Glucose Concentration on Acquired Data. PCA (Principal Component Analysis) is a dimensionality reduction algorithm used to reduce the dataset dimensionality from 151 to 2, enabling it to be plotted in a graphical form. Using the experimental glucose concentration during image acquisition, it becomes evident that glucose concentration does not significantly affect cell classification. This implies that glucose concentration has low classification power, but the classification model will be able to classify cells independently of this experimental condition.

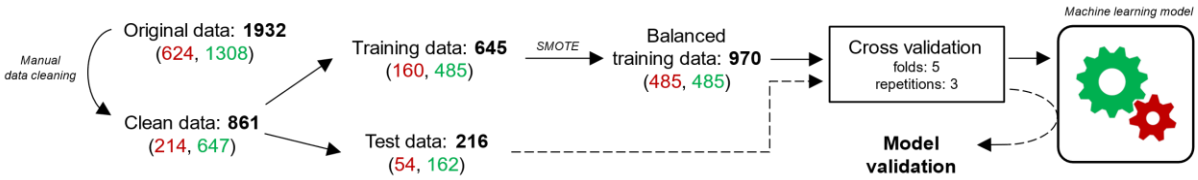

Figure S2: Dataset Handling Through the Training and Testing Workflow. The original dataset consists of a 1932x151 matrix, where the rows represent a single cell, and the columns are the computed features, with a total of 624 alpha cells (red) and 1308 beta cells (green). Before any analysis, the dataset has been cleaned by excluding cells with uncertain identity, which was done by manually cross-checking FLIM, autofluorescence, and immunofluorescence images. In total, 1071 cells were excluded from the analysis, leaving 861 cells. The dataset has been split into training and test sets with a 4:1 proportion to avoid overestimation of performance, resulting in 645 cells for training and 216 for testing. For optimal training, it is beneficial to have balanced classes, meaning the same number of alpha and beta cells. Class balance was achieved by generating synthetic data for alpha cells using the Synthetic Minority Oversampling Technique (SMOTE), which resulted in 970 cells in total with perfectly balanced classes (i.e., 485 cells for both alpha and beta cells). A repeated stratified 5-fold cross-validation was implemented for the training dataset, where the dataset is split into 5 folds, and each fold is used independently for training the algorithm. This process was repeated 3 times, resulting in 15 training scores. The final training score is the average of all the training scores. The testing set, which is an agnostic dataset used to test the model's performance, was treated in the same way, resulting in a final testing score used to validate the model's performance without overestimations.

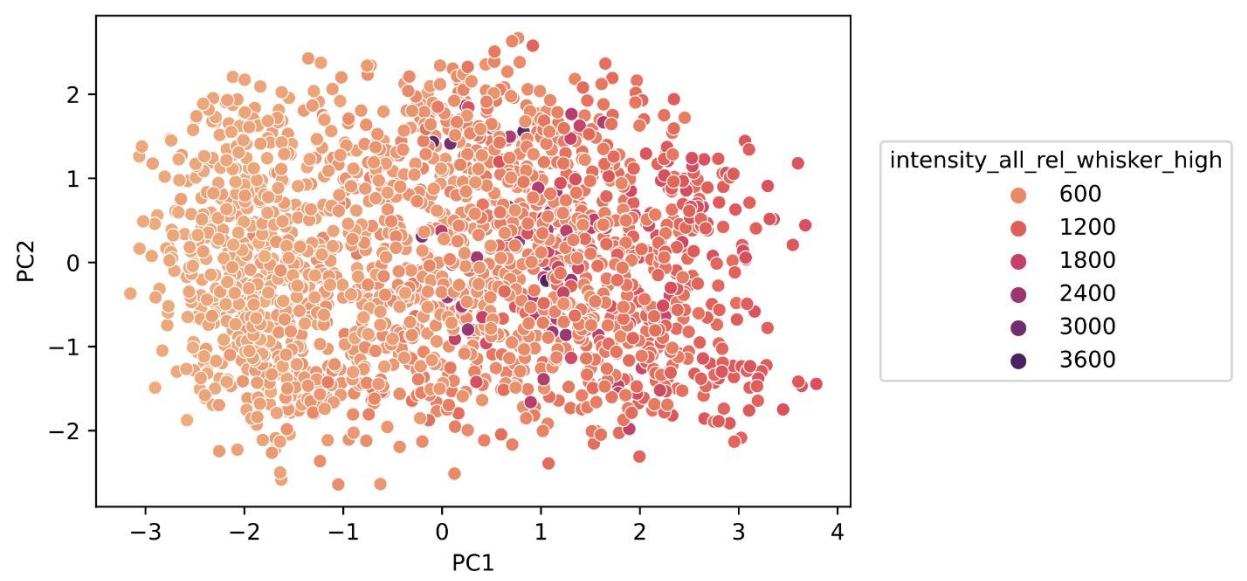

Figure S3: Effect of the most important feature on XGBoost classification power. After XGBoost optimization with Optuna, the most important features have been extracted using the XGboost *.feature_importances_* method. The most important feature was “intensity_all_whisker_high”, which is the higher whisker coordinate of a boxplot of the fluorescence intensity in the image. As evidenced by the colomap, the least intense cells are located on the left side of the plot and the most intense ones on the right side, following the cell type distribution of KDE plots in **Figure 4**.

Alpha cells

a training

number of features	precision	recall	F1 score	roc_auc
151	1.00	1.00	1.00	1.00
149	1.00	1.00	1.00	1.00
116	1.00	1.00	1.00	1.00
48	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00

b test

number of features	precision	recall	F1 score	roc_auc
151	0.73	0.76	0.75	0.83
149	0.75	0.80	0.77	0.85
116	0.75	0.81	0.78	0.86
48	0.75	0.78	0.76	0.85
24	0.66	0.74	0.70	0.81

Beta cells

c training

number of features	precision	recall	F1 score	roc_auc
151	1.00	1.00	1.00	1.00
149	1.00	1.00	1.00	1.00
116	1.00	1.00	1.00	1.00
48	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00

d test

number of features	precision	recall	F1 score	roc_auc
151	0.92	0.91	0.91	0.83
149	0.93	0.91	0.92	0.85
116	0.94	0.91	0.92	0.86
48	0.92	0.91	0.92	0.85
24	0.91	0.87	0.89	0.81

Table S1: Optuna-Optimized Scores of XGBoost at Different Feature Selection Cutoff. XGBoost has been trained and optimized using Optuna with varying feature selection cutoffs. **a)** For alpha cells during the training phase, the algorithm displayed the highest score of 1.0 for every computed metric. **b)** During the test phase, the highest-performing trial has been obtained using 116 features out of 151, with a ROC_AUC of 0.86. The remaining scores ranged from 0.75 to 0.81. Beta cells displayed similar results, but with higher scores. **c)** For beta cells during the training phase, the algorithm displayed a perfect score of 1.0 for every metric. **d)** During the test phase, the highest-performing trial was again achieved with the 116 most important features, resulting in a ROC_AUC of 0.86, with all scores higher than 0.90.

a 10-fold cross validation

	alpha	beta
precision	(1.0), 0.70	(1.0), 0.94
recall	(1.0), 0.83	(1.0), 0.88
F1 score	(1.0), 0.76	(1.0), 0.91
roc_auc	(1.0), 0.86	(1.0), 0.86

b Salzberg test

	alpha	beta
precision	(0.53), 0.27	(0.53), 0.77
recall	(0.53), 0.52	(0.53), 0.54
F1 score	(0.53), 0.36	(0.53), 0.64
roc_auc	(0.53), 0.53	(0.53), 0.53

c Excluded data

	alpha	beta
precision	0.54	0.73
recall	0.59	0.69
F1 score	0.57	0.71
roc_auc	0.64	0.64

Table S2: Assessment of Optuna-Optimized XGBoost Stability. We performed various assessments to evaluate the stability of the Optuna-optimized XGBoost model. **a)** In a 10-fold repeated stratified cross-validation on the dataset, XGBoost displayed similar scores to the 5-fold cross-validation during the training phase, reinforcing the hypothesis that overfitting is absent. **b)** The Salzberg test involved shuffling the training target vector to make XGBoost learn from noise. The results showed a drop in both training and testing scores, as expected. This indicates that the original XGBoost was effectively learning from the data. **c)** In a final test, XGBoost was allowed to predict the outcomes for the 1071 out of 1932 cells that were discarded from the analysis. It displayed scores lower than those during the testing phase. This could be a clue of either overfitting or erroneous manual labeling. Based on the previous stability assessments, it is more likely that the latter hypothesis is accurate.