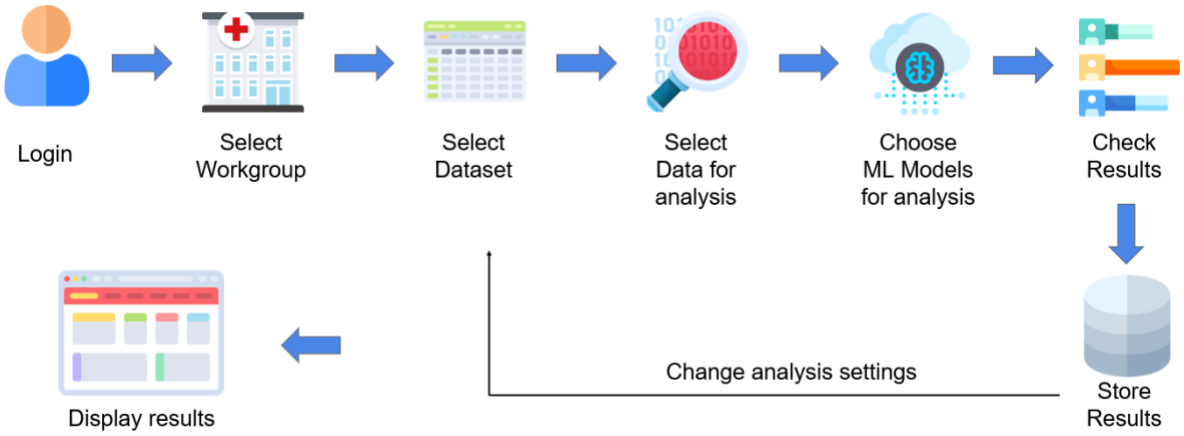
# Supplementary Material

## Personalis: Medical Software for Autoimmune Diseases

Patrick Meier, Jan Kruta, Nicolas Bopp, Thierry Martin, Raphael Carapito, Marta Rizzi, Reinhard E. Voll, Andreas Schwarting, Anne-Sophie Korganow, Erik Schkommodau, Enkelejda Miho

Corresponding author: Prof. Dr. Enkelejda Miho (enkelejda.miho@fhnw.ch)




**Supplementary Figure S1.** Illustration of the prediction workflow performed on the data utilizing the Personalis web platform.

| Component | Role |
|---|---|
| MongoDB | Handles persistent data storage for the MEAN stack. |
| Express.js | Manages back-end logic and facilitates web application development. |
| Angular | Enables the development of dynamic front-end interfaces. |
| Node.js | Executes back-end logic and serves as the server-side engine. |

**Supplementary Table S1.** Depicting the MEAN stack components and their respective roles within the software application.

| Section | Actions |
|---|---|
| Users | Modify user settings, allocate workgroups and roles. |
| Workgroups | Adjust workgroup settings, assign datasets. |
| Datasets | Edit datasets, assign ML models. |
| ML Models | Create and edit machine learning models. |

**Supplementary Table S2.** Description of the interface and functionalities of the administration page, designed to facilitate efficient management and control of system parameters and settings.

| Category | Dataset / model | Evaluation setting |
|---|---|---|
| Dataset | LBBR | Target: discoidLupus; N=1,165; 146 predictors; 5 outer × 5 inner folds |
| Dataset | RAPP | Target: RHEUMATOLOGE_Erkrankung_RA_PsA_SpA_kein; N=799; 18 predictors; 5 outer × 5 inner folds |
| Dataset | Cytokine | Target: Discoid_Lupus; N=18; 233 predictors; 3 outer × 2 inner folds |
| Classifier | Random Forest | class_weight="balanced"; random_state=42 |
| Classifier | Logistic Regression | class_weight="balanced"; random_state=42; max_iter=5000 |
| Classifier | MLPClassifier | random_state=42; 1 hidden layer (100 neurons) |
| Classifier | SGDClassifier | class_weight="balanced"; random_state=42; max_iter=2000; sigmoid probability calibration |
| Classifier | SVM | class_weight="balanced"; probability=True; random_state=42 |

**Supplementary Table S3.** Settings used for the standardized internal machine-learning evaluation. The table summarizes the representative prediction task and nested cross-validation design for each dataset, together with the fixed settings of the five evaluated classifier families. The prediction target, administrative identifiers, and predefined direct target-derived or diagnostic-proxy variables were excluded from the predictor sets.

30

| Dataset | Model | Accuracy (95% CI) | Precision (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | F1-score (95% CI) | ROC-AUC (95% CI) |
|---|---|---|---|---|---|---|---|
| **LBBR** | RF | 0.800 (0.785–0.815) | 0.515 (0.424–0.604) | 0.216 (0.165–0.267) | 0.948 (0.933–0.961) | 0.304 (0.241–0.365) | 0.671 (0.630–0.710) |
| | LR | 0.726 (0.701–0.749) | 0.348 (0.300–0.394) | 0.403 (0.339–0.466) | 0.808 (0.781–0.834) | 0.373 (0.320–0.421) | 0.634 (0.591–0.676) |
| | MLP | 0.794 (0.785–0.803) | 0.441 (0.278–0.606) | 0.064 (0.034–0.093) | 0.980 (0.970–0.988) | 0.111 (0.062–0.162) | 0.664 (0.624–0.703) |
| | SGD | 0.797 (0.795–0.797) | 0.000 (0.000–0.000) | 0.000 (0.000–0.000) | 0.999 (0.997–1.000) | 0.000 (0.000–0.000) | 0.615 (0.575–0.655) |
| | SVM | 0.773 (0.751–0.793) | 0.426 (0.363–0.485) | 0.352 (0.288–0.411) | 0.879 (0.858–0.900) | 0.385 (0.325–0.439) | 0.680 (0.638–0.719) |
| | Dummy baseline | 0.797 (0.797–0.797) | 0.000 (0.000–0.000) | 0.000 (0.000–0.000) | 1.000 (1.000–1.000) | 0.000 (0.000–0.000) | 0.498 (0.470–0.527) |
| **RAPP** | RF | 0.538 (0.504–0.572) | 0.486 (0.447–0.529) | 0.470 (0.431–0.509) | 0.819 (0.806–0.833) | 0.477 (0.437–0.517) | 0.756 (0.730–0.781) |
| | LR | 0.474 (0.442–0.511) | 0.427 (0.391–0.465) | 0.439 (0.400–0.482) | 0.802 (0.789–0.816) | 0.427 (0.391–0.466) | 0.716 (0.688–0.744) |
| | MLP | 0.512 (0.482–0.541) | 0.466 (0.406–0.532) | 0.363 (0.331–0.396) | 0.789 (0.777–0.801) | 0.375 (0.333–0.416) | 0.702 (0.674–0.729) |
| | SGD | 0.496 (0.481–0.512) | 0.281 (0.245–0.318) | 0.273 (0.261–0.287) | 0.760 (0.754–0.767) | 0.216 (0.196–0.236) | 0.691 (0.662–0.719) |
| | SVM | 0.499 (0.472–0.528) | 0.462 (0.411–0.518) | 0.378 (0.342–0.414) | 0.787 (0.776–0.799) | 0.373 (0.335–0.412) | 0.657 (0.630–0.684) |
| | Dummy baseline | 0.481 (0.481–0.481) | 0.120 (0.120–0.120) | 0.250 (0.250–0.250) | 0.750 (0.750–0.750) | 0.162 (0.162–0.162) | 0.495 (0.468–0.521) |
| **Cytokine** | RF | 0.611 (0.500–0.667) | 0.000 (0.000–0.000) | 0.000 (0.000–0.000) | 0.917 (0.750–1.000) | 0.000 (0.000–0.000) | 0.597 (0.319–0.847) |

| Dataset | Model | Accuracy (95% CI) | Precision (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | F1-score (95% CI) | ROC-AUC (95% CI) |
|---|---|---|---|---|---|---|---|
| | LR | 0.611 (0.500–0.667) | 0.000 (0.000–0.000) | 0.000 (0.000–0.000) | 0.917 (0.750–1.000) | 0.000 (0.000–0.000) | 0.444 (0.222–0.667) |
| | MLP | 0.556 (0.389–0.722) | 0.250 (0.000–0.751) | 0.167 (0.000–0.500) | 0.750 (0.500–1.000) | 0.200 (0.000–0.545) | 0.278 (0.042–0.542) |
| | SGD | 0.611 (0.500–0.667) | 0.000 (0.000–0.000) | 0.000 (0.000–0.000) | 0.917 (0.750–1.000) | 0.000 (0.000–0.000) | 0.472 (0.194–0.764) |
| | SVM | 0.667 (0.444–0.889) | 0.500 (0.199–1.000) | 0.500 (0.167–0.833) | 0.750 (0.500–1.000) | 0.500 (0.181–0.800) | 0.639 (0.375–0.875) |
| | Dummy baseline | 0.667 (0.667–0.667) | 0.000 (0.000–0.000) | 0.000 (0.000–0.000) | 1.000 (1.000–1.000) | 0.000 (0.000–0.000) | 0.500 (0.500–0.500) |

**Supplementary Table S4.** Internal cross-validation performance of the machine-learning classifiers evaluated in Personalis. One representative prediction task was selected for each dataset: discoid lupus for LBBR (N=1,165; 929 No, 236 Yes), disease class for RAPP (N=799; 384 None, 201 RA, 141 PsA, 73 SpA), and discoid lupus for the Cytokine dataset (N=18; 12 No, 6 Yes). Model performance was assessed using nested cross-validation, in which each sample was evaluated using a model that had not been trained on that sample. Values are reported as performance estimates with 95% confidence intervals. For the multiclass RAPP task, precision, recall, specificity, F1-score, and ROC-AUC were calculated as macro-averaged metrics, giving equal weight to each disease class. For comparison, a DummyClassifier function (Dummy baseline) based on the observed class distribution where it predicts the frequency of the labels was used as a baseline reference for the ML models' accuracy predictions.