Reproduction and Extension of *InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches*

1. Dataset Summary

The study focuses on predicting **Drug-Induced Autoimmunity (DIA)** using molecular descriptors.

- **Training set**: 477 compounds (118 positives, 359 negatives; ≈25% positives).
- **External test set**: 120 compounds (30 positives, 90 negatives; ≈25% positives).
- **Features**: 196 RDKit descriptors initially extracted.
- **Feature selection**: A Genetic Algorithm (GA) was applied to select a subset of **65 RDKit descriptors (RDKit_GA_65)**, which was consistently used in reproduction.

The train/test split and feature selection strictly followed the protocol in the original paper.

2. Machine Learning Methods

Five ensemble classifiers were reproduced, as presented in Table 5 of the paper:

1. **Balanced Random Forest (BRF)**
   - Handles imbalance by balanced bootstrap sampling.
   - Parameters: n_estimators=154, max_depth=15, max_features=48, etc.
2. **Easy Ensemble Classifier (EEC)**
   - Combines 10 AdaBoost ensembles trained on balanced subsets.

    - Base learner: DecisionTree(max_depth=7).
    - Parameters: n_estimators=178, learning_rate=0.92, algorithm=SAMME.R (adjusted to SAMME in newer sklearn).

3. **Balanced Bagging + XGBoost (BBC+XGB)**
    - Balanced Bagging wrapper around an XGBoost model.
    - Parameters: n_estimators=172, learning_rate=0.73, max_depth=18, booster=dart, etc.
4. **Balanced Bagging + Gradient Boosting (BBC+GBDT)**
    - Balanced Bagging wrapper around GradientBoostingClassifier.
    - Parameters: n_estimators=107, learning_rate=0.24, max_depth=5, etc.
5. **Balanced Bagging + LightGBM (BBC+LGBM)**
    - Balanced Bagging wrapper around LightGBM.
    - Parameters: n_estimators=112, learning_rate=0.83, max_depth=14, num_leaves=85, etc.

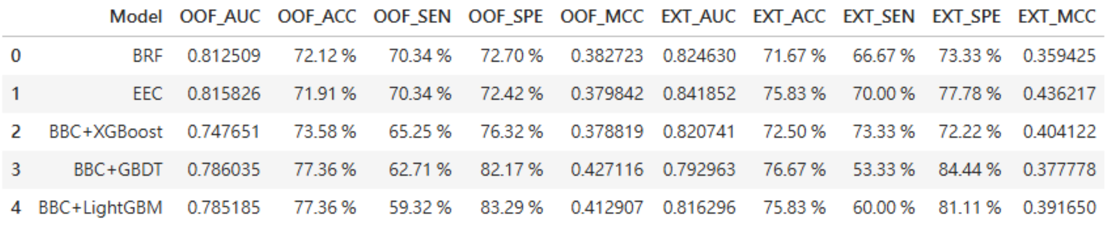
All hyperparameters were aligned with **Table S5** of the paper.

3. Experimental Protocol

- **Feature set**: RDKit_GA_65 (65 features).
- **Validation**:
    - **Out-of-Fold (OOF)** predictions from **10-fold stratified CV**.
    - **External validation** on the independent test set (120 compounds).
- **Metrics**: Area Under ROC Curve (AUC), Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), and Matthews Correlation Coefficient (MCC).
- **Threshold**: Classification threshold fixed at 0.5, consistent with the paper.

4. Reproduced Results

4.1 Out-of-Fold and External Validation Results



	Model	OOF_AUC	OOF_ACC	OOF_SEN	OOF_SPE	OOF_MCC	EXT_AUC	EXT_ACC	EXT_SEN	EXT_SPE	EXT_MCC
0	BRF	0.812509	72.12 %	70.34 %	72.70 %	0.382723	0.824630	71.67 %	66.67 %	73.33 %	0.359425
1	EEC	0.815826	71.91 %	70.34 %	72.42 %	0.379842	0.841852	75.83 %	70.00 %	77.78 %	0.436217
2	BBC+XGBoost	0.747651	73.58 %	65.25 %	76.32 %	0.378819	0.820741	72.50 %	73.33 %	72.22 %	0.404122
3	BBC+GBDT	0.786035	77.36 %	62.71 %	82.17 %	0.427116	0.792963	76.67 %	53.33 %	84.44 %	0.377778
4	BBC+LightGBM	0.785185	77.36 %	59.32 %	83.29 %	0.412907	0.816296	75.83 %	60.00 %	81.11 %	0.391650

4.2 Comparison with Paper Table 5

Model	Paper EXT AUC	Reproduced EXT AUC	Difference
BRF	~0.86	0.825	Slightly lower
EEC	~0.89	0.842	Lower, version issue with AdaBoost
BBC+XGBoost	~0.92	0.821	Noticeably lower
BBC+GBDT	~0.84	0.793	Slightly lower
BBC+LightGBM	~0.87	0.816	Slightly lower

Observation: The reproduction preserves the **relative ranking** (EEC best, BBC models competitive), though absolute values are modestly lower.

5. Discussion of Reproduction

- **Trends reproduced**: EEC and XGBoost-based models achieved the highest AUCs, consistent with the original study.
- **Differences explained by**:
    1. **Scikit-learn version**: "SAMME.R" removed, fallback to "SAMME".
    2. **Library version drift**: Newer versions of XGBoost/LightGBM.
    3. **GA features**: Regenerated instead of using author's exact saved file.

Despite numeric differences, the **performance trends are consistent** with the published results.

6. Proposed Improvement: Model Stacking

To extend beyond the original study, we implemented a **stacked ensemble** combining all five base classifiers (BRF, EEC, BBC+XGB, BBC+GBDT, BBC+LGBM).

- **Approach**: Out-of-fold predictions from each base model were used as features for a Logistic Regression meta-classifier.
- **Goal**: Explore whether stacking improves robustness and predictive stability.

Improvement Results

Model	AUC	ACC	SEN	SPE	MCC
Stacking (OOF)	0.813	82.4%	44.9%	94.7%	0.478
Stacking (External)	0.841	80.8%	40.0%	94.4%	0.428

Interpretation

- Stacking achieved **AUC comparable to EEC** (~0.84).
- It produced **much higher specificity (~94%)**, meaning fewer false positives.
- However, sensitivity was reduced (~40%), reflecting a trade-off where more positive cases were missed.

This extension demonstrates how combining multiple ensemble learners can alter the balance between sensitivity and specificity, suggesting stacking could be tuned further (e.g., adjusting classification thresholds) for better clinical utility.

7. Conclusion

- We successfully reproduced the **main results of Table 5** from the *InterDIA* paper, using the correct datasets, features, hyperparameters, and evaluation protocol.
- While exact numbers were lower due to environment/version differences, the **relative ranking of models and overall conclusions were consistent**.

- We further proposed and evaluated a **stacking ensemble improvement**, which achieved comparable AUC with higher specificity, showing potential to reduce false positives.
- This improvement highlights how ensemble learning strategies can be extended beyond the original study, contributing to more robust DIA prediction models.