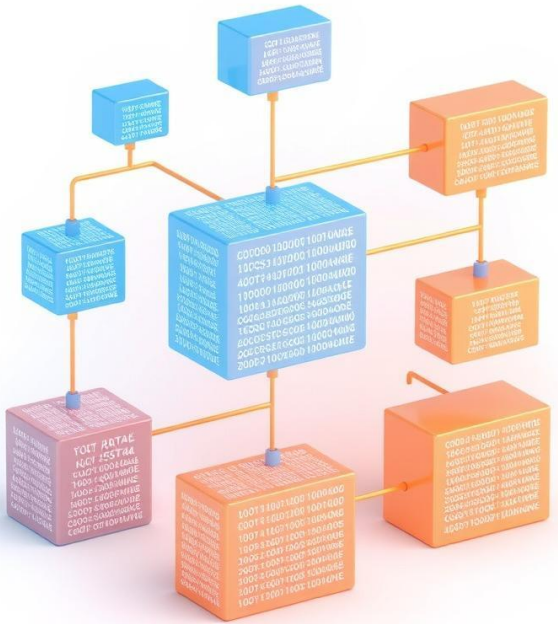
RING Classification of Residue-Residue Contacts in Protein Structures

Anahita Soltantouyeh (2131455)
Mikael Poli (2080422)
Biddut Bhowmik (2141802)

Data Preprocessing

Cleaning

- We merged the provided .tsv files into a single dataframe.
- Any instances labeled as "NA" were assigned to a new category labeled "Unclassified".
- We excluded all examples containing missing feature values from the dataset. We used all 22 features for training.



Data Preprocessing

Train-Test-Validation Split

Step 1

First, we set aside 10% of the data as the test set. The remaining 90% was used for training and validation.

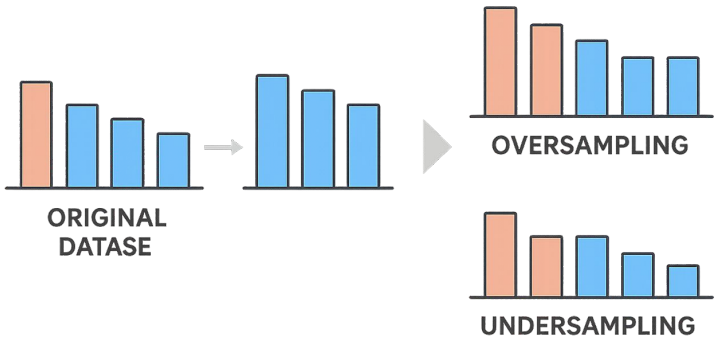
Step 2

We further divided the 90% into 80% for training and 20% for validation.

Data Preprocessing

Resampling

- We merged the provided .tsv files into a single dataframe.
- Any instances labeled as "NA" were assigned to a new category labeled "Unclassified".
- We excluded all examples containing missing feature values from the dataset. We used all 22 features for training.



Data Preprocessing

Interaction	Train (O.)	Train (R.)	Val.	Test
Unclassified	764,634	100,000	191,159	106,199
HBOND	739,818	100,000	184,954	102,753
VDW	516,198	100,000	129,050	71,694
PIPISTACK	26,888	30,000	6,722	3,734
IONIC	24,700	30,000	6,175	3,431
PICATION	6,174	30,000	1,544	858
SSBOND	1,453	30,000	363	202
PIHBOND	1,242	30,000	310	172
Tot.	2,081,107	450,000	520,277	289,043

Table 2. Dataset composition by interaction type across training (O. = original, R. = resampled) training, validation, and test sets.

Model

Training

We trained an ensemble of eight binary XGBoost classifiers in a One-vs-Rest (OvR) setup, with one model dedicated to each RING interaction type. Training took in total ≈ 20 seconds (wall time) using Google Colab's T4 GPU.

```
"model_parameters": {
    "device": "cuda",
    "objective": "binary:logistic",
    "eval_metric": "logloss",
    "max_depth": 10,
    "learning_rate": 0.2
},
"early_stopping_rounds": 15,
"num_boost_round": 2000,
```

Prediction

Step 1: Compute Probabilities

We use all eight models to compute class probabilities for each test example.

Step 2: Assign Label

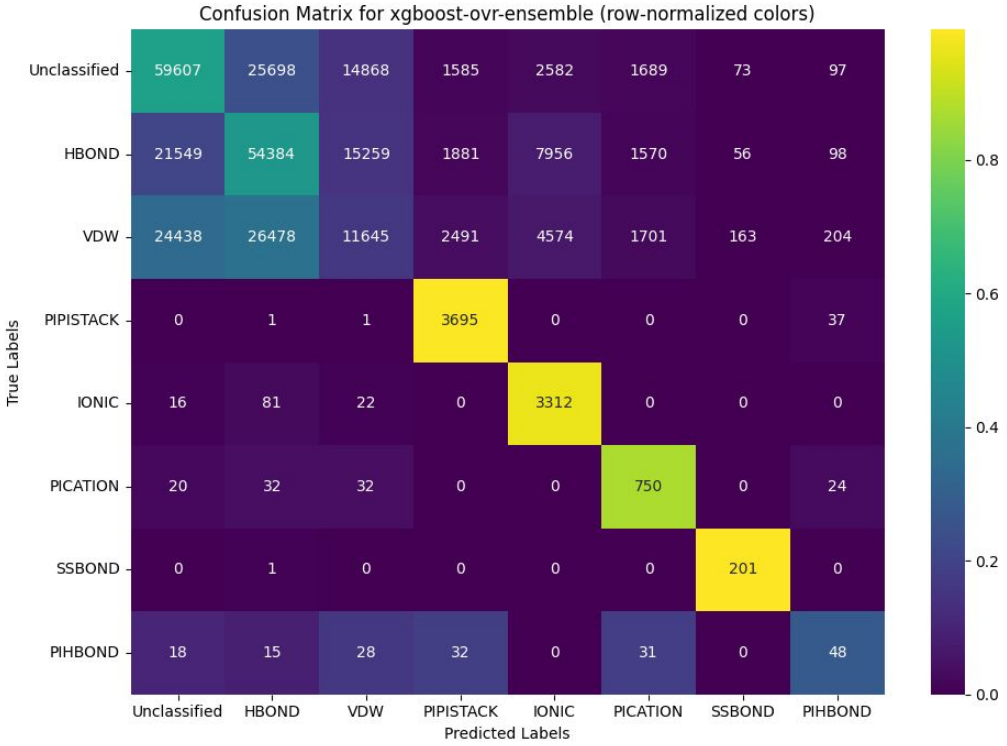
The label corresponding to the highest predicted probability is then assigned as the interaction type.

Results

Metric	Value
Balanced Accuracy	0.622
AUC-ROC	0.923
Matthews Correlation Coefficient	0.231
Average Precision	0.355

Table 3. Performance metrics for the XGBoost OvR ensemble

- **AUC-ROC is strong:** The ensemble ranks the correct class well in each OvR task.
- **Average Precision is moderate:** Meaningful signals are captured, but false positives remain, especially among majority classes.
- **MCC is modest:** Frequent errors persist, mainly misclassifying majority classes.
- **Balanced Accuracy is decent:** Reasonable performance across classes, balancing true positives despite imbalance.



Discussion

- The ensemble met performance goals but **struggled to distinguish between majority classes** (HBOND, VDW, Unclassified).
  - The difficulty in distinguishing HBOND from VDW may reflect their physical and chemical similarities.
- **The resampling strategy significantly impacted performance:**
  - Increasing VDW samples improved VDW classification but negatively affected other majority classes.
  - Optimal results were achieved with a balanced minority-to-majority class ratio of approximately 1:3.
  - Oversampling SSBOND and PIHBOND distorted data distributions and degraded overall performance.
- While the larger validation set stabilized estimates, it may have led to a slight reduction in overall accuracy.
- The main challenges encountered were **non-linear feature** separability and **extreme class imbalance**.