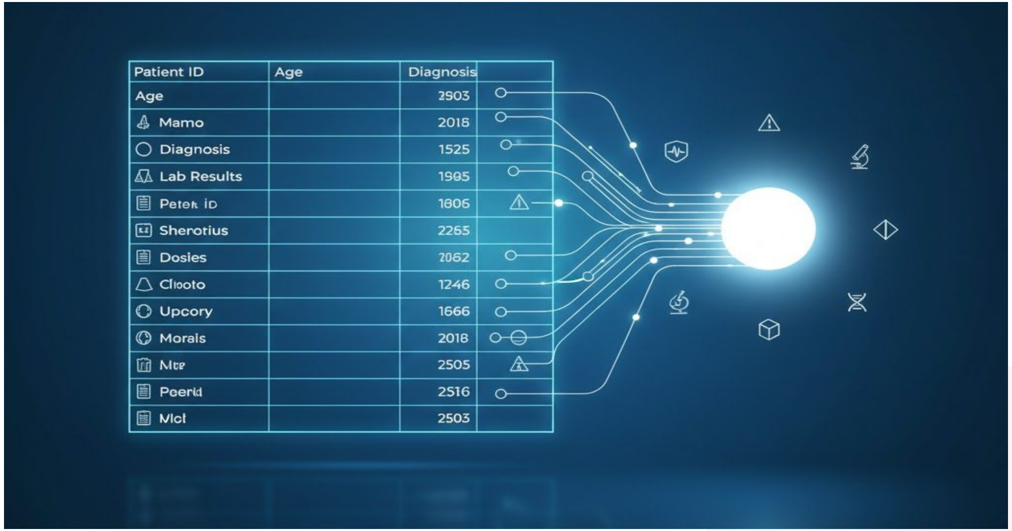
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

Prediction on high-dimensional clinical data



Patient ID	Age	Diagnosis
Age		2503
Mamo		2018
Diagnosis		1525
Lab Results		1985
Peten iD		1806
Sherotius		2265
Dosies		2062
Clioto		1246
Upcory		1666
Morals		2018
Mte		2505
Poerld		2516
Mlol		2503

Supervisor:

Prof. Dr. Myra Spiliopoulou

Team Members:

- Bhavesh Sharad Yeole
- Chethan Chinnabhandara Nagaraj
- Sajith Kumar Santhosh
- Sourav Salkoppalu Lingaraju

Contents

- ❖ Main Dataset Overview
- ❖ Data Preprocessing
- ❖ Pipeline Architecture
- ❖ Evaluation Results
- ❖ Final Steps

Preliminary Dataset Overview: Drug-Induced Autoimmunity

- **Number of Instances:** 477
- **Number of Features:** 195
- Input Features:
    - Continuous Numerical
    - Represents molecular properties
- **Associated Task:** Classification
- **Target Variable:** DIA Positive/Negative

Main Dataset Overview - UNITI Tinnitus Datasets

Dataset 1: Baseline Questionnaire Data

- **Number of Patients:** 376 and **Number of Features:** 622
- **Input Features:**
  - Baseline questionnaire responses
  - Includes baseline THI
  - Clinical measures
  - Psychological and Lifestyle measures
- **Target Variable:** Final THI score
- **Associated Task:** Regression

Main Dataset Overview - UNITI Tinnitus Datasets

Dataset 2: Baseline Questionnaire + Genetic Data

- **Number of Patients:** 250 and **Number of Features:** 624
- **Input Features:**
  - Baseline questionnaire responses
  - Includes baseline THI
  - Clinical measures
  - Psychological and Lifestyle measures
  - Genetic features
- **Target Variable:** Final THI score
- **Associated Task:** Regression

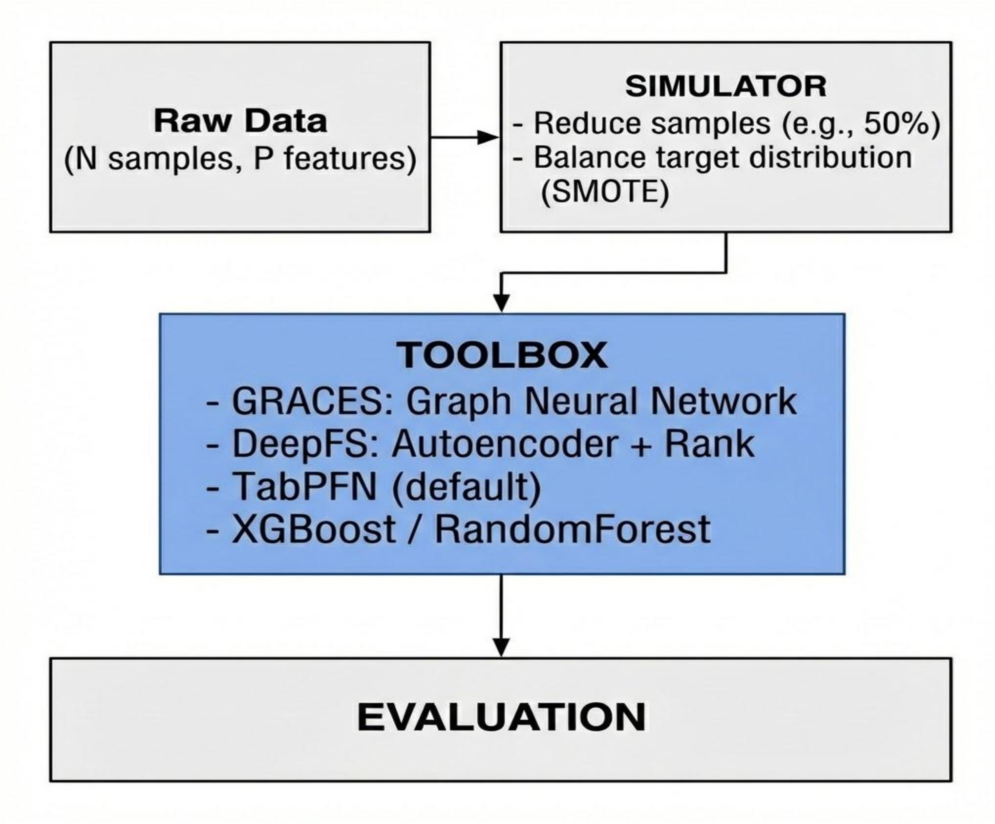
Data Preprocessing

- Used baseline (± genetic) features; **final THI** as **target**.
- Dropped columns with **>95% missingness;** remaining gaps handled per model.
- Removed constant, duplicate and ID-like columns.
- Converted features to **numeric**.
- Built scaled data for similarity models
- Built raw data for foundation models.

Regression Task Implementation

- **Objective:** Predict final visit **THI score** from baseline patient data
- **Modeling Approaches:**
  - **TabPFN:**
  - **GRACES + XGBoost**
  - **DeepFS + XGBoost**
  - **Validation Strategy:**
  - **K-fold cross-validation** to ensure robust performance estimates and reduce overfitting
- **Evaluation Metrics:**
  - **RMSE** for prediction error magnitude
  - **R²** for explained variance

PIPELINE :



Raw Data
(N samples, P features)

SIMULATOR
- Reduce samples (e.g., 50%)
- Balance target distribution (SMOTE)

TOOLBOX
- GRACES: Graph Neural Network
- DeepFS: Autoencoder + Rank
- TabPFN (default)
- XGBoost / RandomForest

EVALUATION

Modular Pipeline

- **Simulator**: sample reduction + SMOTE
- **Toolbox**: TABPFN, GRACES, DeepFS feature selection
- Evaluation for Regression: **RMSE, MAE, R²**
- Evaluation for Classification: **Accuracy, Weighted F1 Scores**
- **Pipeline**: orchestration & execution

STEPS: Import necessary libraries → Load and prepare data → Run Pipeline

```
results = run_pipeline(X_train, y_train, X_test, y_test, red_perc,
no_features, model_type)
```

Simulator (Reduction + SMOTE/SMOGN)

1. **SMOTE (Synthetic Minority Over-sampling Technique)**
2. **SMOGN (Synthetic Minority Over-sampling with Gaussian Noise)**to better model continuous distributions

Why SMOTE or SMOGN?

- Our dataset: 75% Negative, 25% Positive (3:1 ratio)
- Models biased toward majority class
- **Generates** NEW synthetic samples for minority class
- **Creates** balanced dataset without information loss

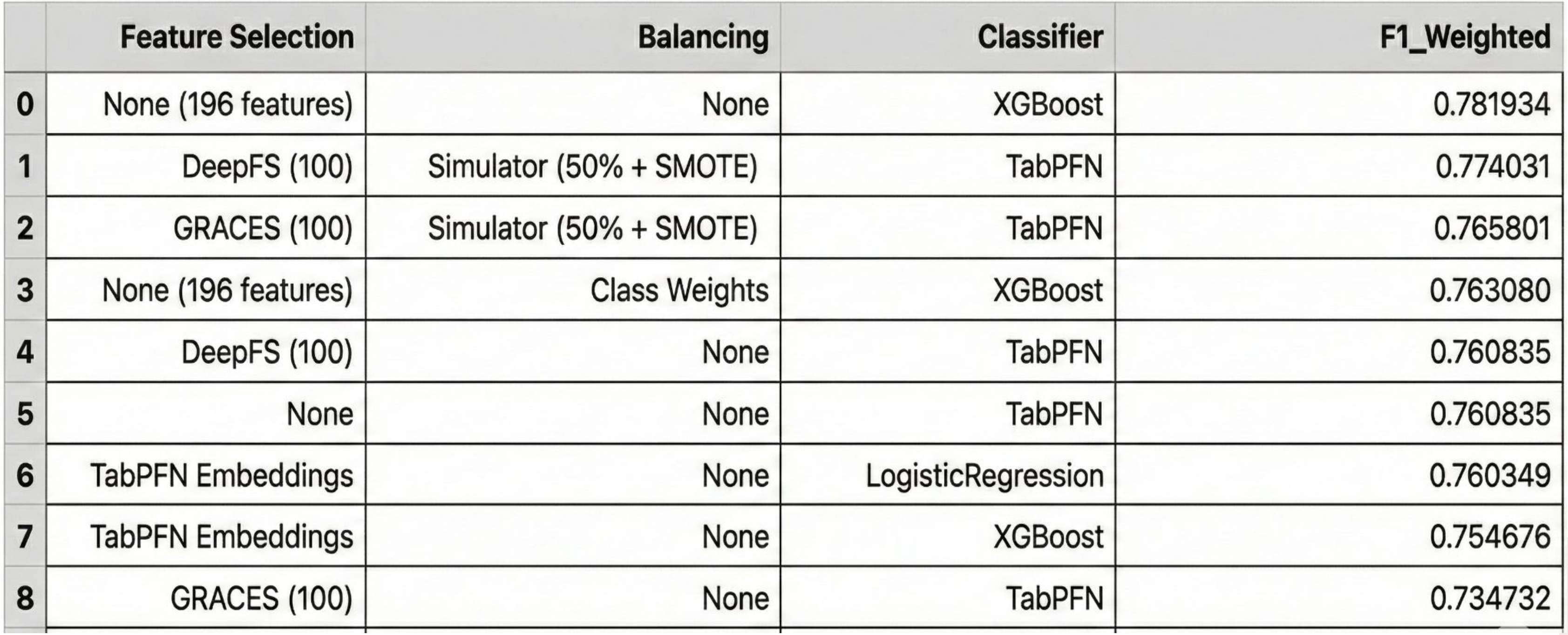
Evaluation Results (Regression)

Strategy	Input Features	Avg RMSE (±SD)	Avg R²
GRACES + XGBoost v1	GRACES(50 features from baseline data + genetic data)	15.59 ± 1.37	0.405
TabPFN v1	Raw baseline data+ genetic features (all 487 features)	15.19 ± 2.15	0.427
GRACES + XGBoost v2	Selected features (50 features from baseline data)	16.41 ± 1.33	0.354
TabPFN v2	Raw baseline data (NaNs preserved, no scaling)	14.67 ± 1.45	0.479

Key Insights

- **Regression strategies predicted final THI** well from baseline features.
- **TabPFN slightly outperformed GRACES + XGBoost** in the genetic dataset (R^2 0.427 vs 0.405).
- **Genetic features improved predictive power** over questionnaire-only data.
- **GRACES + XGBoost offers interpretable top features**, highlighting key clinical and psychological measures.
- **Models are robust**, confirmed by 5-fold cross-validation.

Evaluation Results (Classification)



	Feature Selection	Balancing	Classifier	F1_Weighted
0	None (196 features)	None	XGBoost	0.781934
1	DeepFS (100)	Simulator (50% + SMOTE)	TabPFN	0.774031
2	GRACES (100)	Simulator (50% + SMOTE)	TabPFN	0.765801
3	None (196 features)	Class Weights	XGBoost	0.763080
4	DeepFS (100)	None	TabPFN	0.760835
5	None	None	TabPFN	0.760835
6	TabPFN Embeddings	None	LogisticRegression	0.760349
7	TabPFN Embeddings	None	XGBoost	0.754676
8	GRACES (100)	None	TabPFN	0.734732

Final Steps

- Implement **DeepFS + XGBoost** on Main Dataset.
- Identify most predictive features:
  - **Compare features** selected by GRACES and DeepFS.
  - Determine common features as the **most predictive subset**
- **Test filtered features** with TabPFN to evaluate if performance improves
- Explore regression variation:
  - **Predict baseline THI** using baseline features
  - **Compare results** with final THI prediction models

References

- K. Li, F. Wang, L. Yang, and R. Liu, *"Deep Feature Screening: Feature Selection for Ultra High-Dimensional Data via Deep Neural Networks,"* 2023. [Online].
- C. Chen, S. T. Weiss, and Y.-Y. Liu, *"Graph Convolutional Network-based Feature Selection for High-dimensional and Low-sample Size Data,"* Channing Division of Network Medicine, Harvard Medical School, 2025.
- N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, *"Accurate predictions on small data with a tabular foundation model,"* Nature, 2025.

Thank you

Any Questions?