Influence of Data Size and Class Balance on Machine Learning Classification Performance and SHAP explanations

Anusha Ihalapathirana,[1] Gunjan Chandra,[1] Piia Lavikainen,[2] Pekka Siirtola,[1] Satu Tamminen,[1] Nirzor Talukder,[1]
Janne Martikainen,[2] and Juha R¨oning [1]
1 Biomimetics and Intelligent Systems Group, University of Oulu, Oulu, FI-90014, Finland
2 School of Pharmacy, University of Eastern Finland, Kuopio, Finland

1 INTRODUCTION

- Some machine learning (ML) models require extensive datasets for optimal training, others give notable results with smaller datasets
- Explainable Artificial Intelligence (XAI) aims to provide insights into the decision-making process of complex algorithms
- SHAP is a post-hoc explanation method that requires a background dataset when interpreting ML models
- The objective is to understand the effect of data size and data imbalance on the performance and the SHAP explanation of machine learning models

Dataset

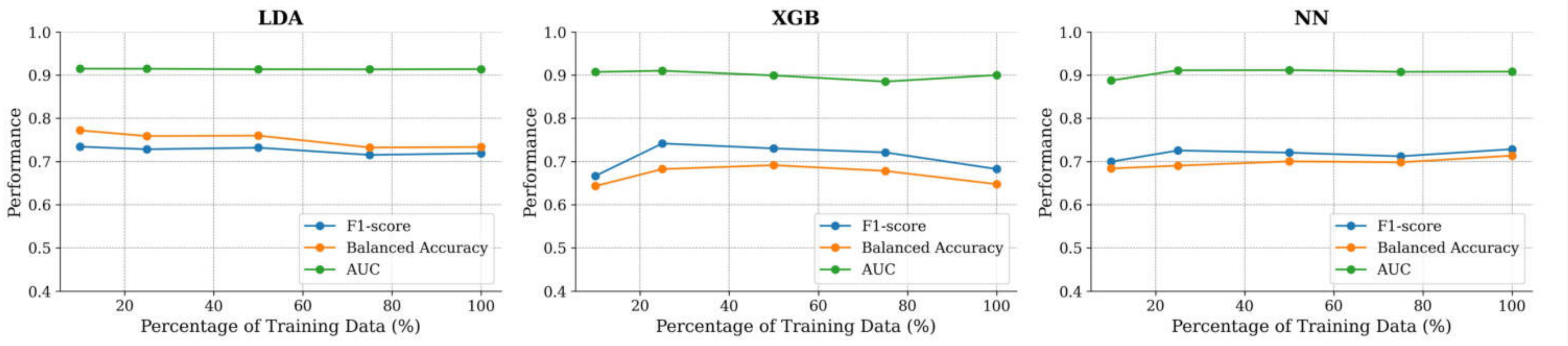
- North Karelia Wellbeing Services County dataset
- Type 2 Diabetes electronic health register (EHR) data collected from Siun sote, Finland

2 METHODOLOGY

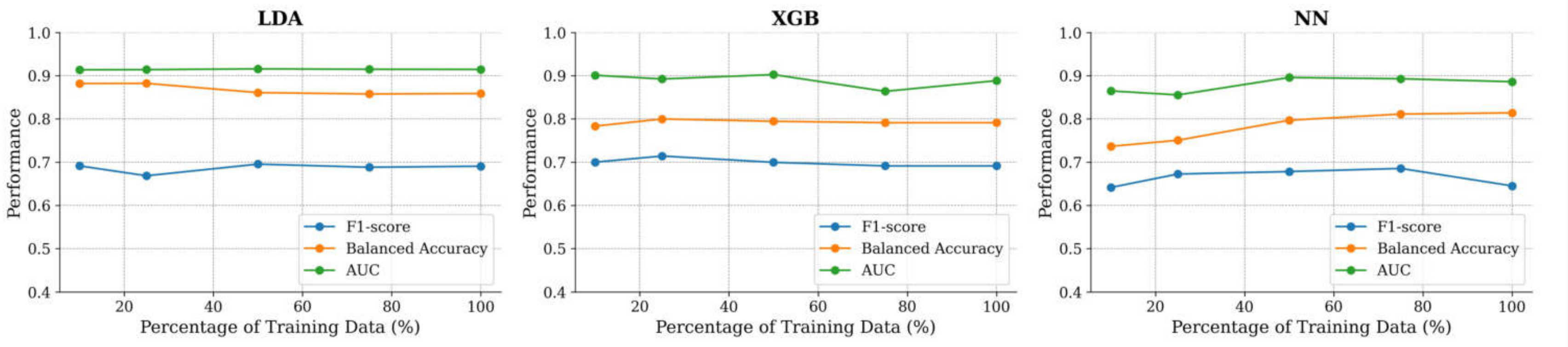
- **Model Performance Assessment**
  - Three models were used
    - Linear Discriminant Analysis (LDA)
    - XGBoost
    - Neural Network (NN)
  - Each model was trained with
    - Five different data sizes
    - Imbalanced Data and Balanced Data (SMOTE applied)
  - Evaluated using balanced accuracy, F1 score, and AUC
- **SHAP Explanations Assessment**
  - Calculated Mean absolute SHAP values across 10-fold cross-validation splits for each feature
  - Assessed under five different background data sizes and class distributions

3 RESULTS - MODEL PERFORMACE

Performance of the models with balanced and imbalanced data with different training data sizes.



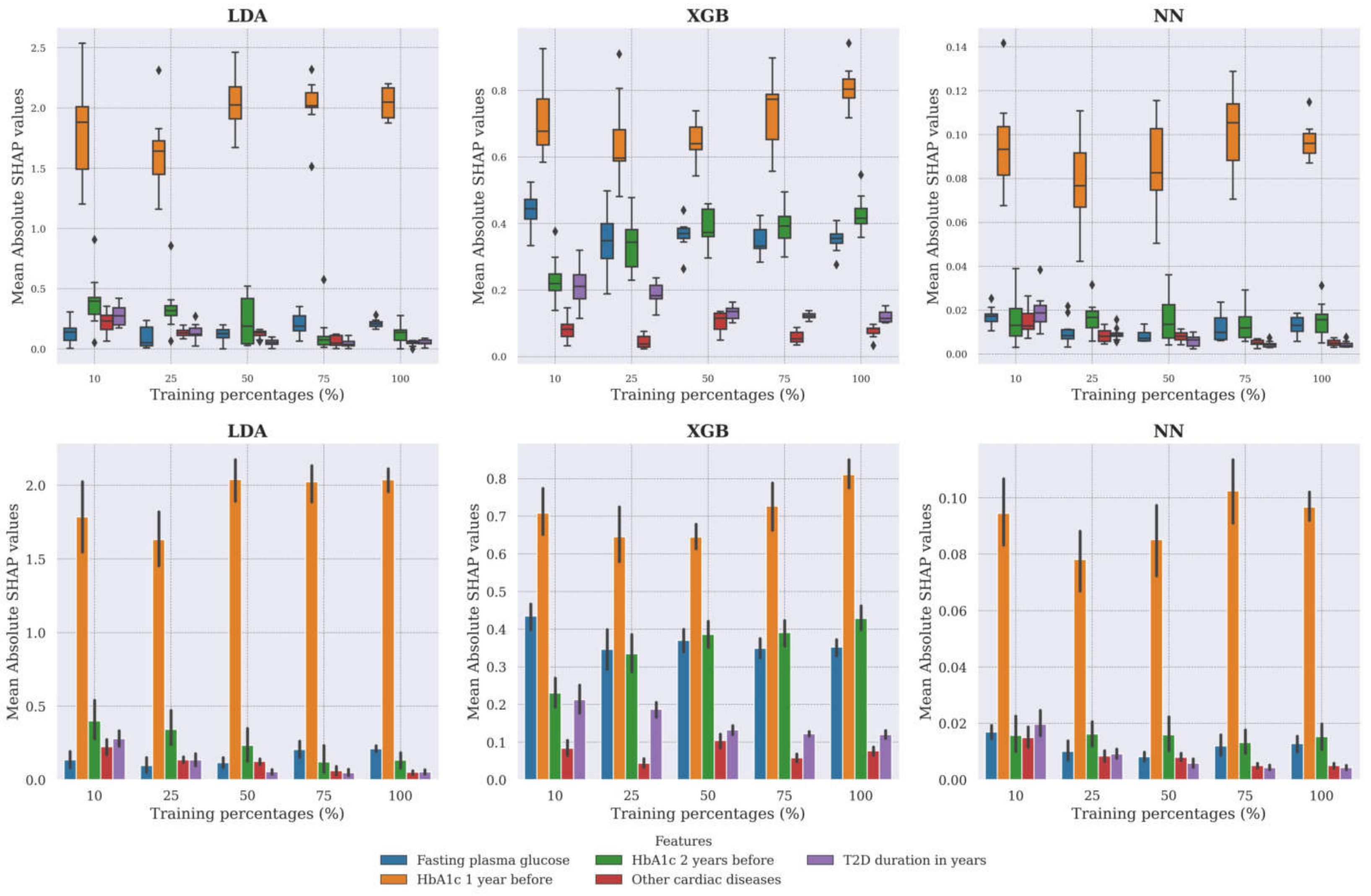
(a) Imbalanced Data



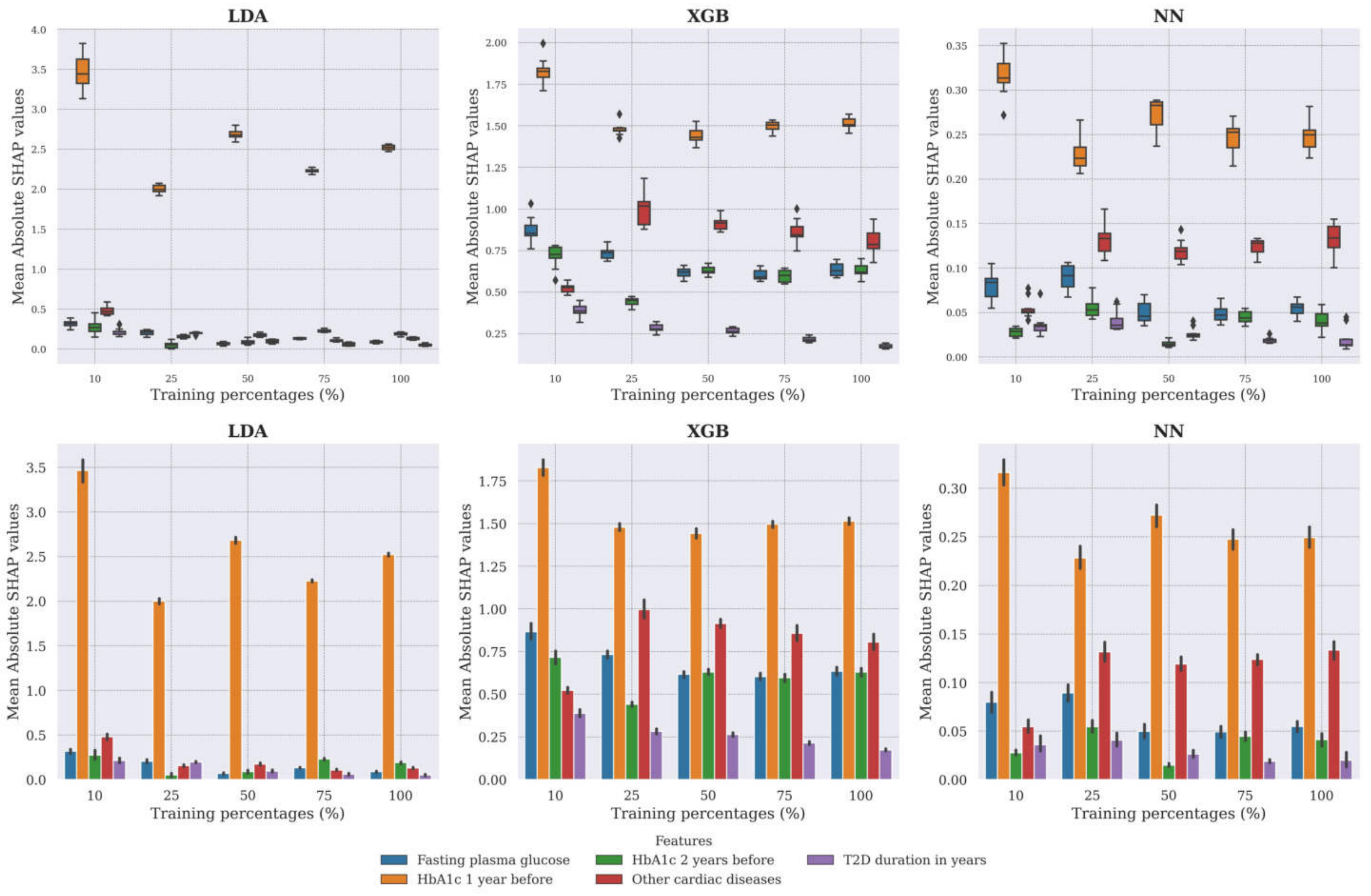
(b) Balanced Data

4 RESULTS - SHAP EXPLANATIONS

Mean Absolute SHAP Values for Features Across 10-Fold Cross-Validation Splits

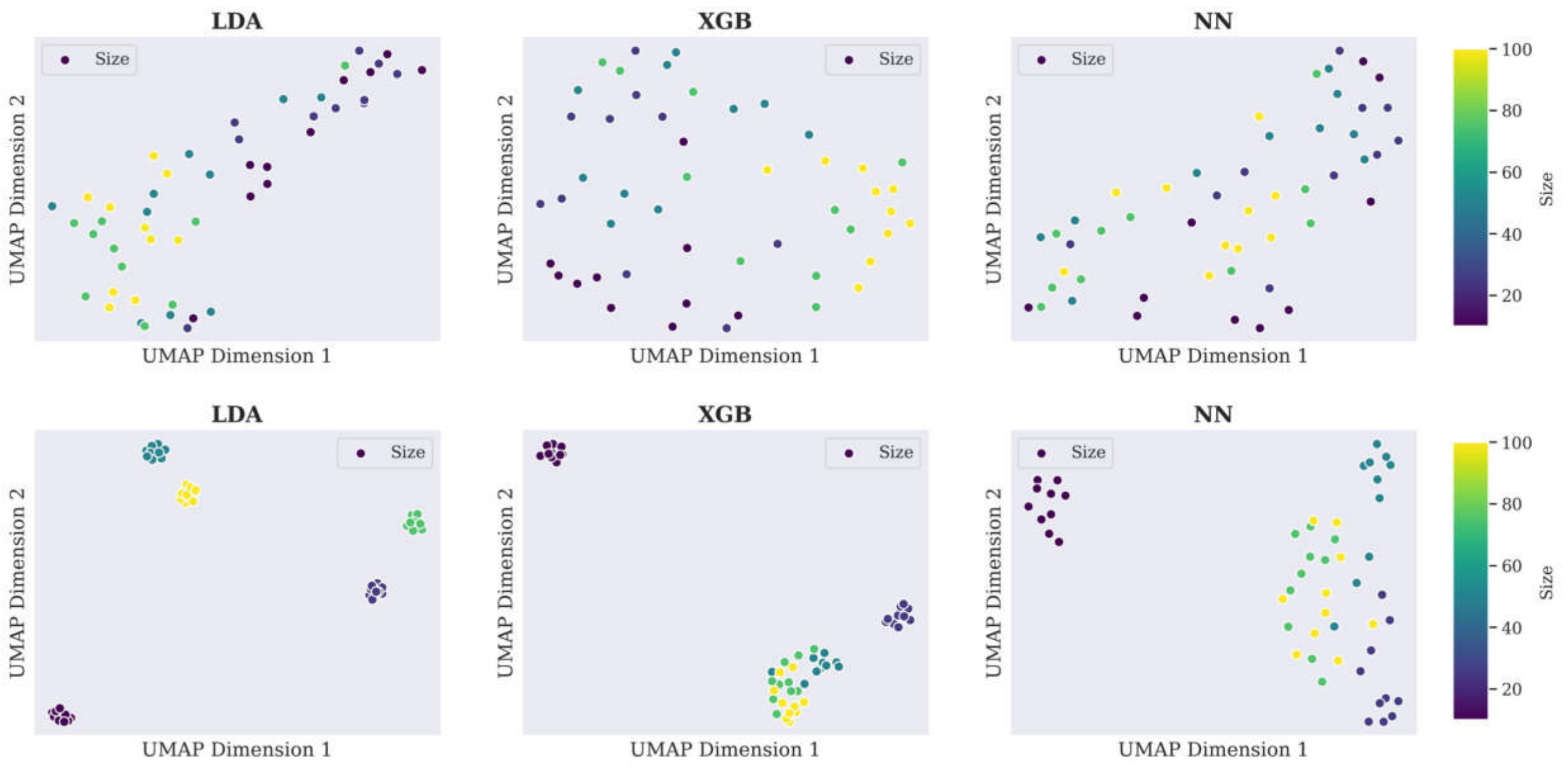


(a) Imbalanced Data



(a) Balanced Data

UMAP Visualization of Mean Absolute SHAP Values Across 10-fold Cross-Validation Splits



Imbalanced Data

Balanced Data

5 CONCLUSION

- Different ML models perform optimally with different sizes of training data, and the impact of data imbalance on performance varies by the metrics
- SHAP explanations benefit from balanced background data and become more stable with larger background datasets
- To ensure reliable SHAP explanations, avoid excessively small background data sizes