Supplemental Material on Additional Datasets for Data organization limits the predictability of binary classification

Fei Jing,[1] Zi-Ke Zhang,[2,*] Yi-Cheng Zhang,[3,†] and Qingpeng Zhang[1,‡]

[1] *Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China*

[2] *Center for Digital Communication Studies, Zhejiang University, Hangzhou 310058, China*

[3] *Department of Physics, University of Fribourg, Chemin du Musée 3, 1700 Fribourg, Switzerland*

CONTENTS

* zkz@zju.edu.cn

† yi-cheng.zhang@unifr.ch

‡ qpzhang@hku.hk

I. EXACT UPPER BOUND OF AUC AND CORRESPONDING OPTIMAL ROC CURVES

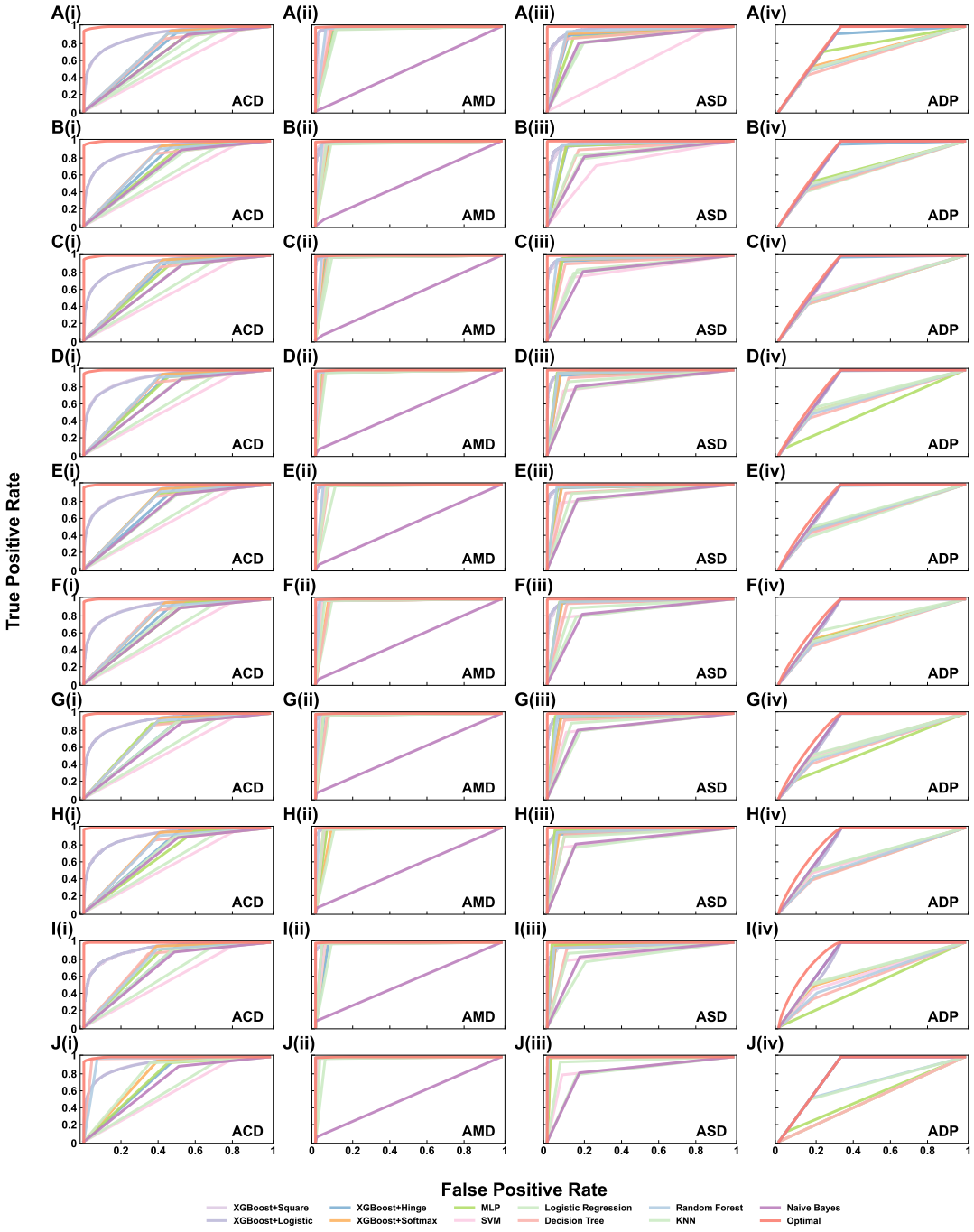

Figure S1. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (ACD, AMD, ASD and ADP) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

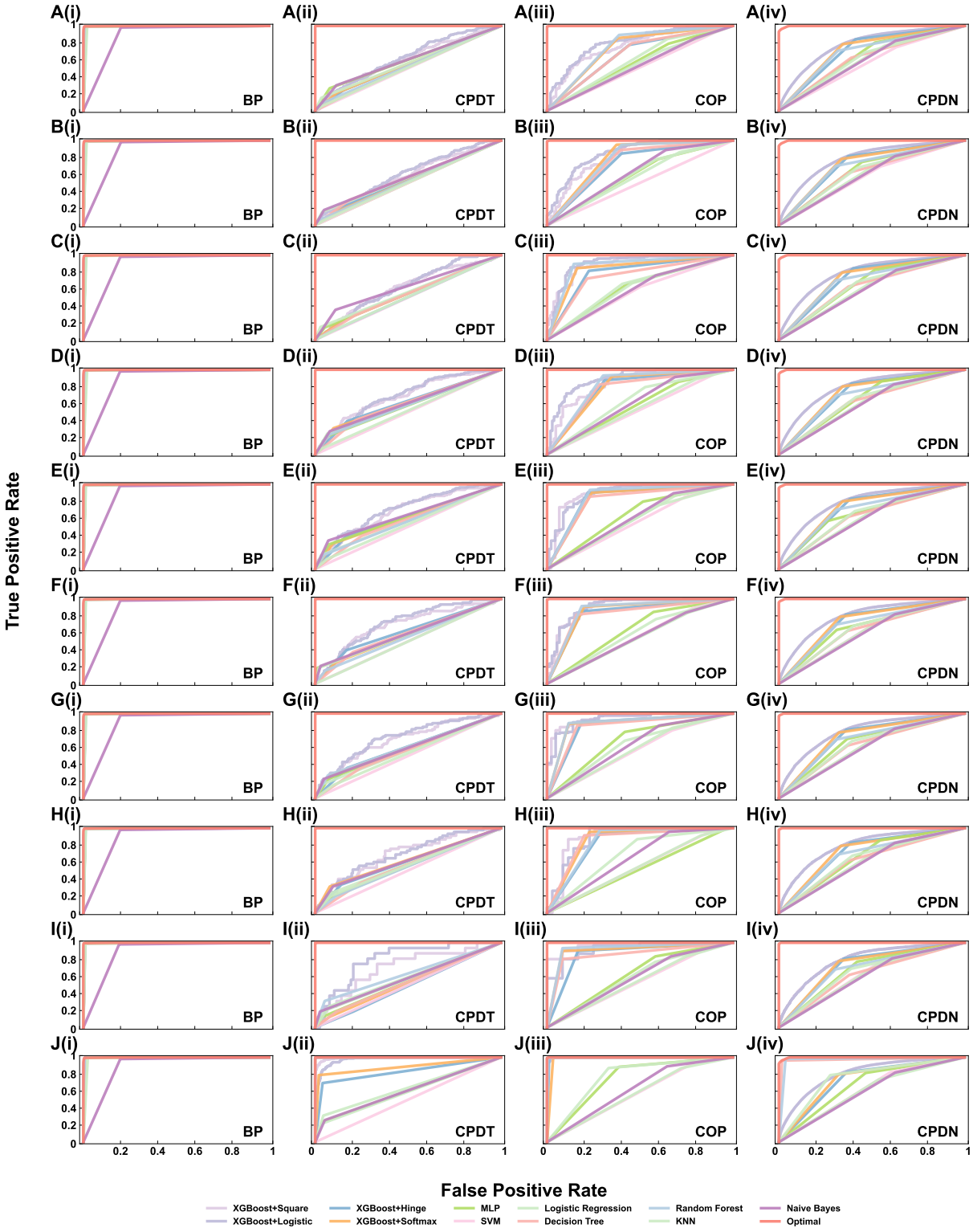

Figure S2. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (BP, CPDT, COP and CPDN) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

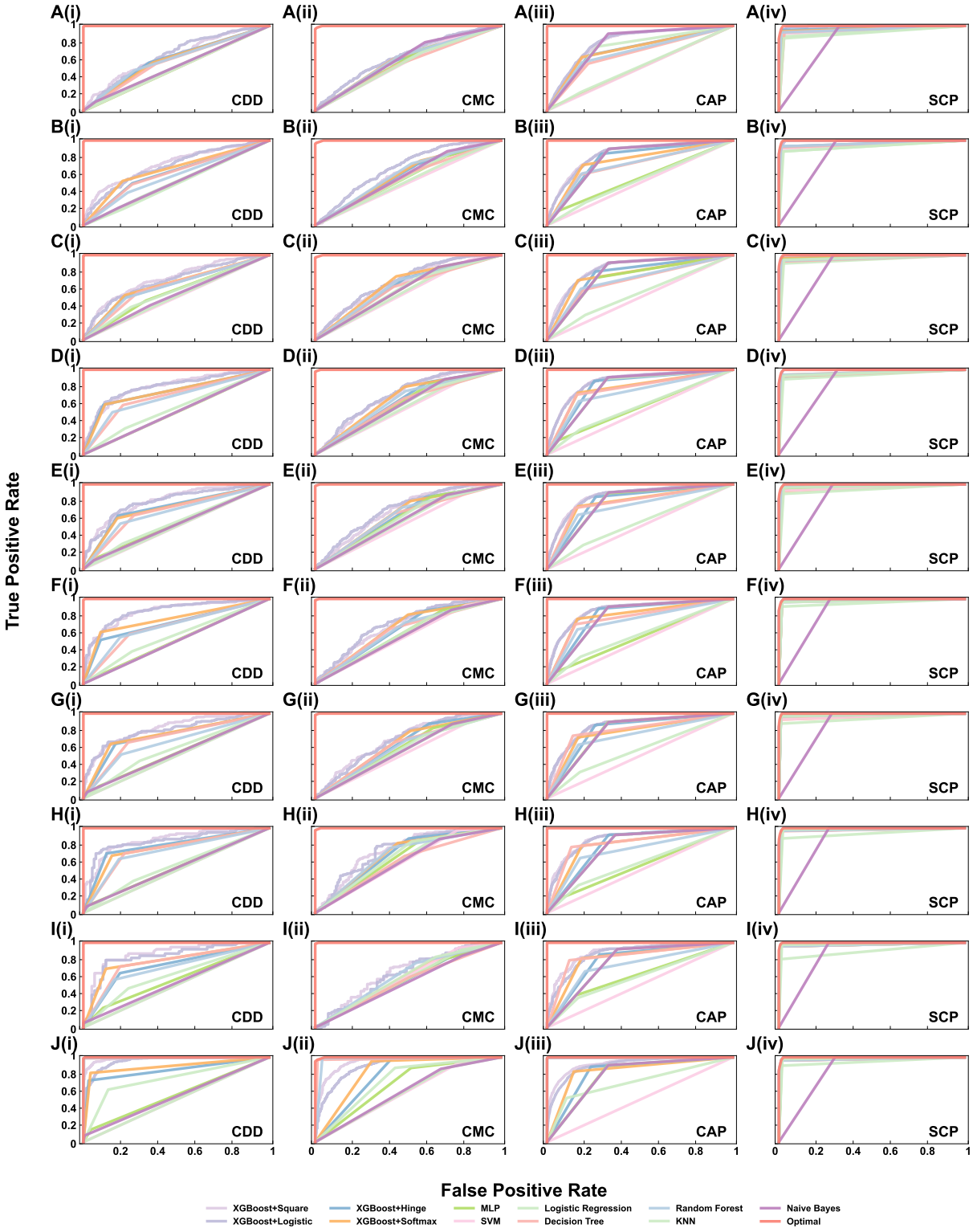

Figure S3. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (CDD, CMC, CAP and SCP) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

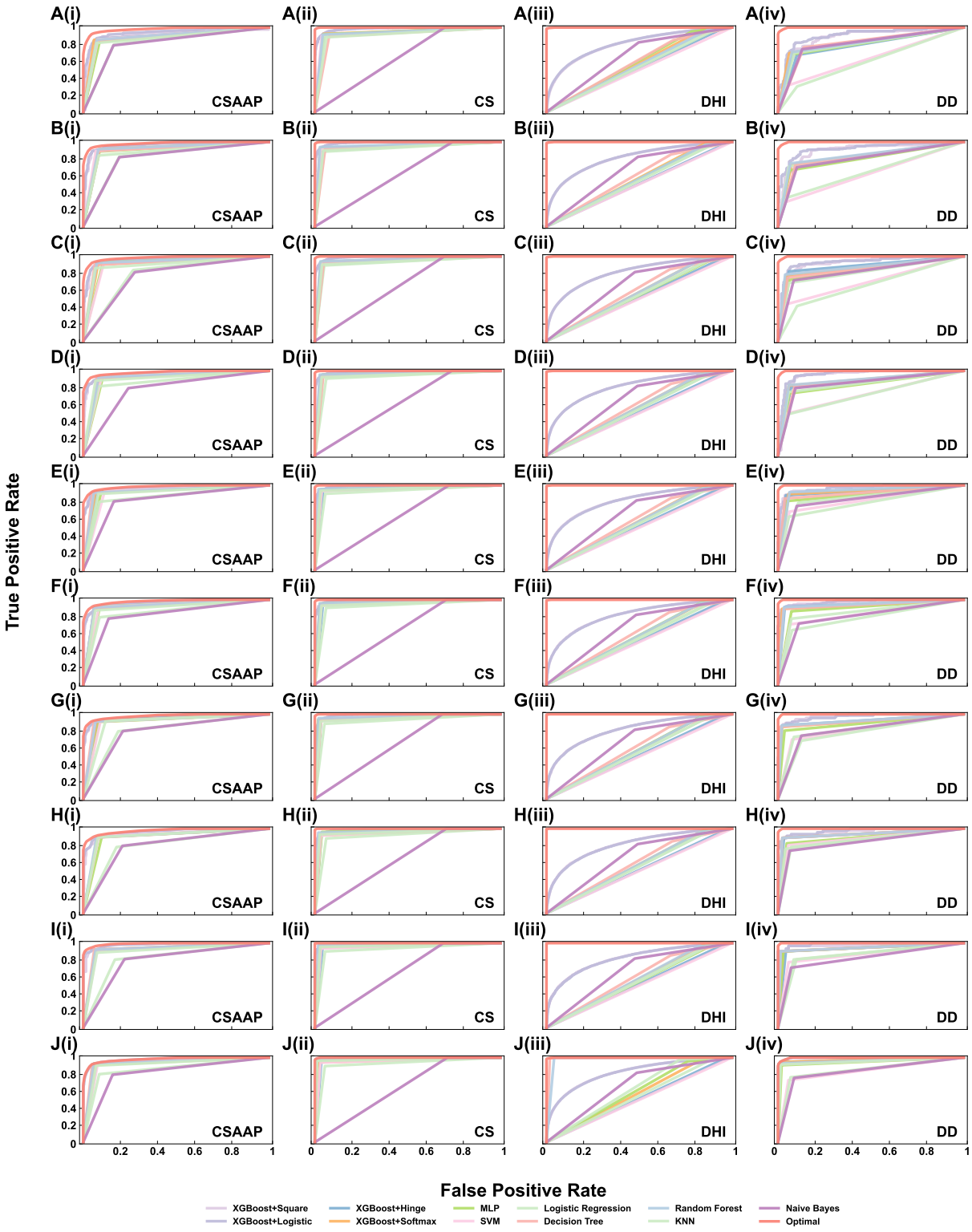

Figure S4. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (CSAAP, CS, DHI and DD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

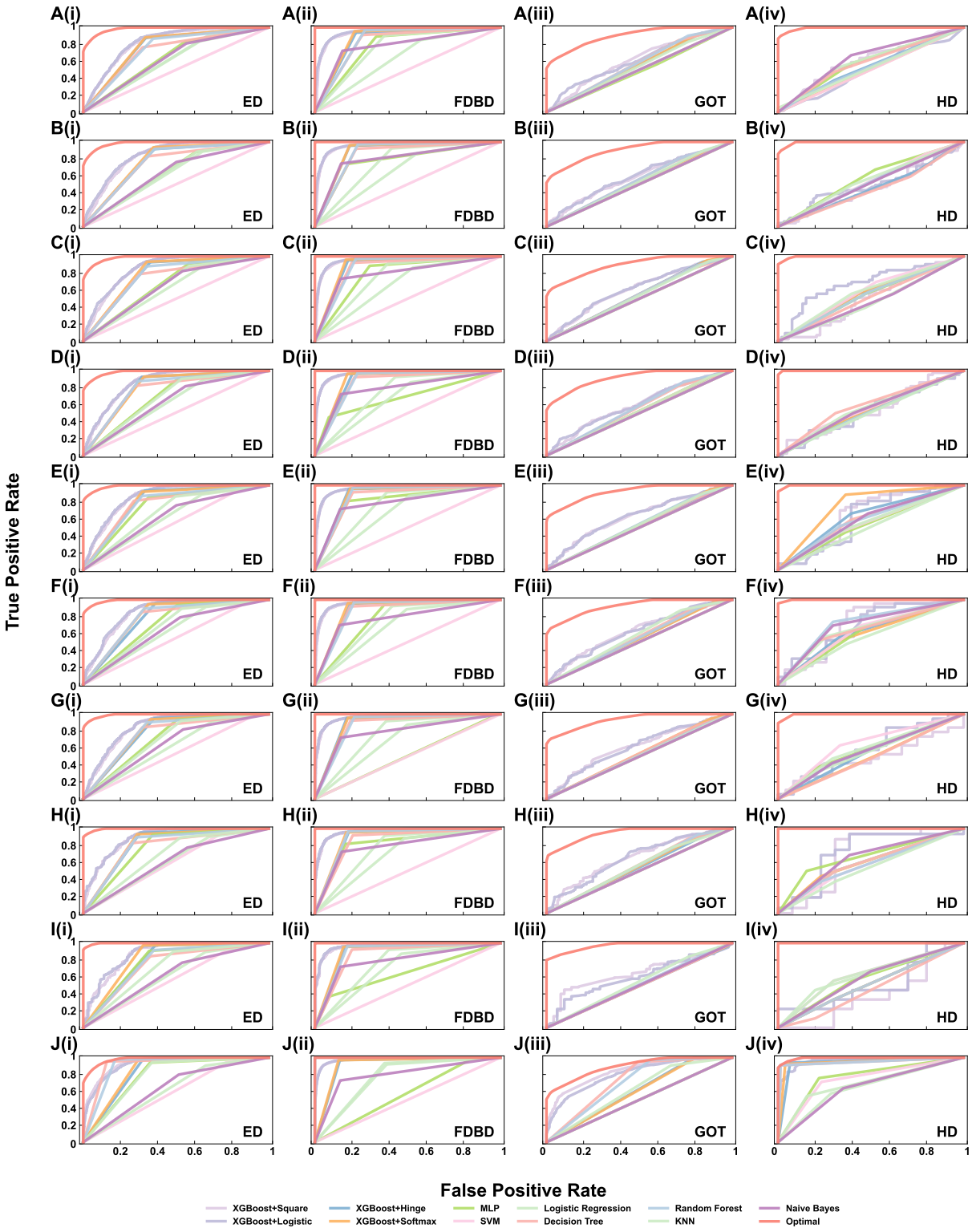

Figure S5. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (ED, FDBD, GOT and HD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

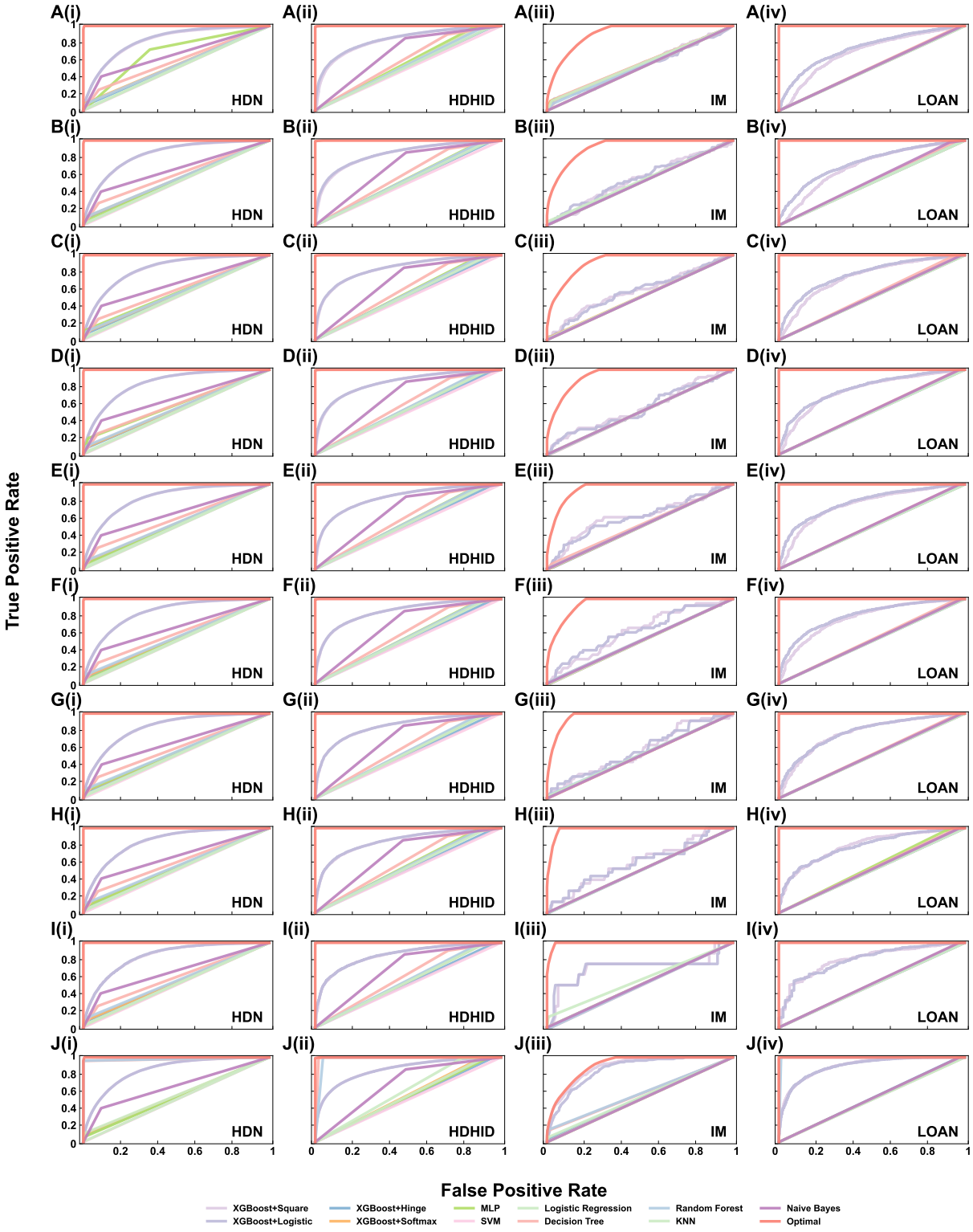

Figure S6. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (HDN, HDHID, IM and LOAN) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

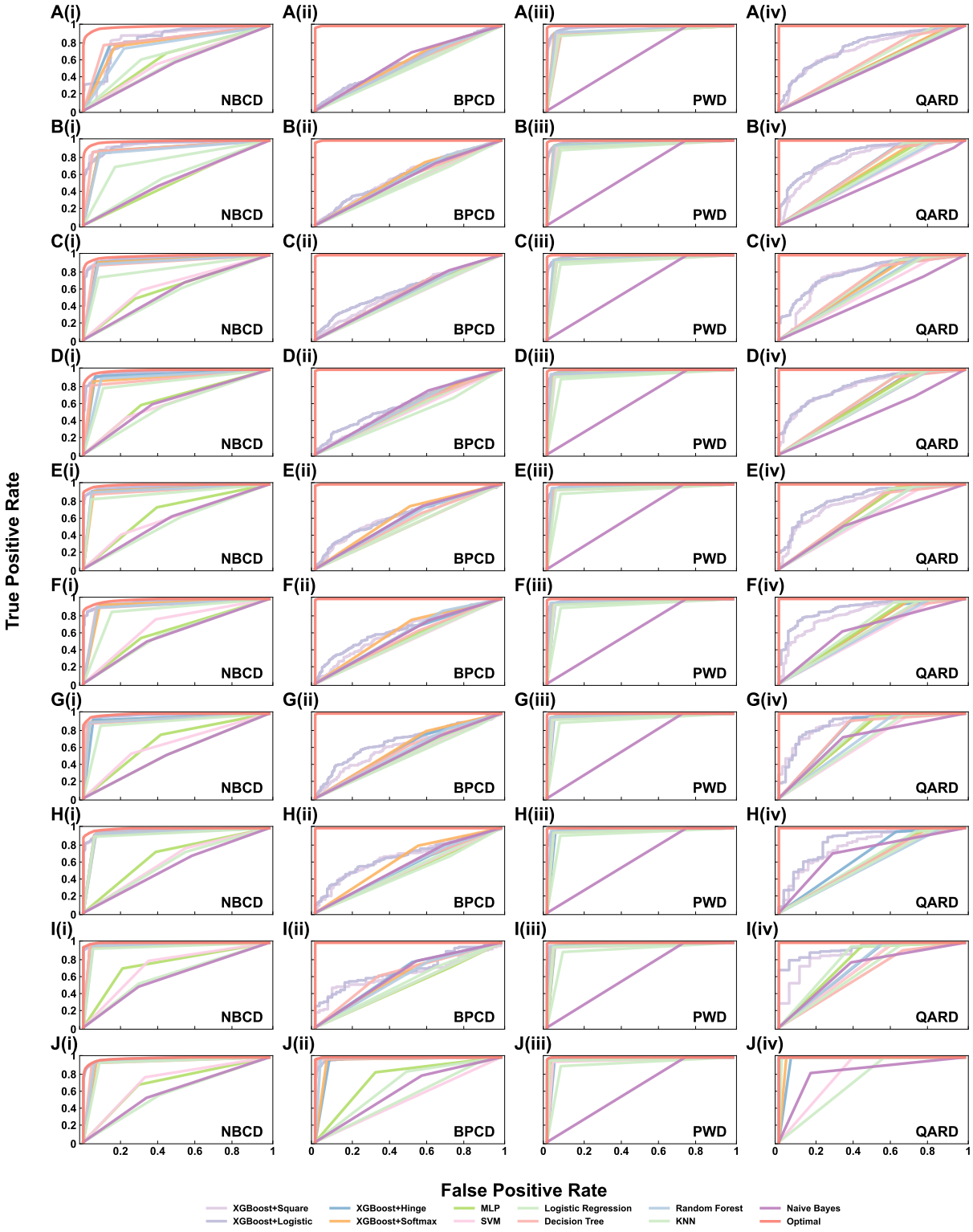

Figure S7. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (NBCD, BPCD, PWD and QARD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

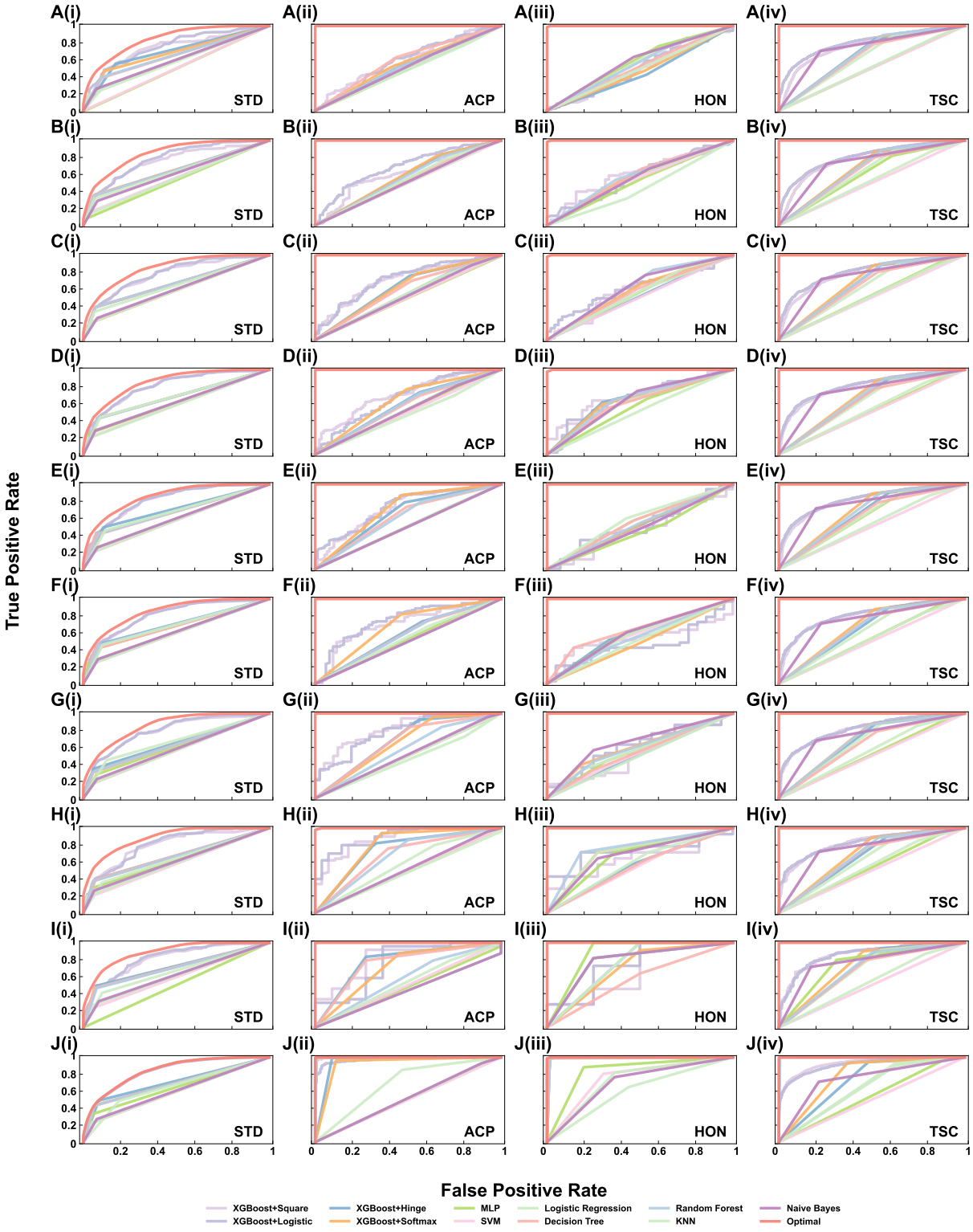

Figure S8. Exact upper bound of AUC and corresponding optimal ROC curves on 4 additional real-world datasets (STD, ACP, HON and TSC) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

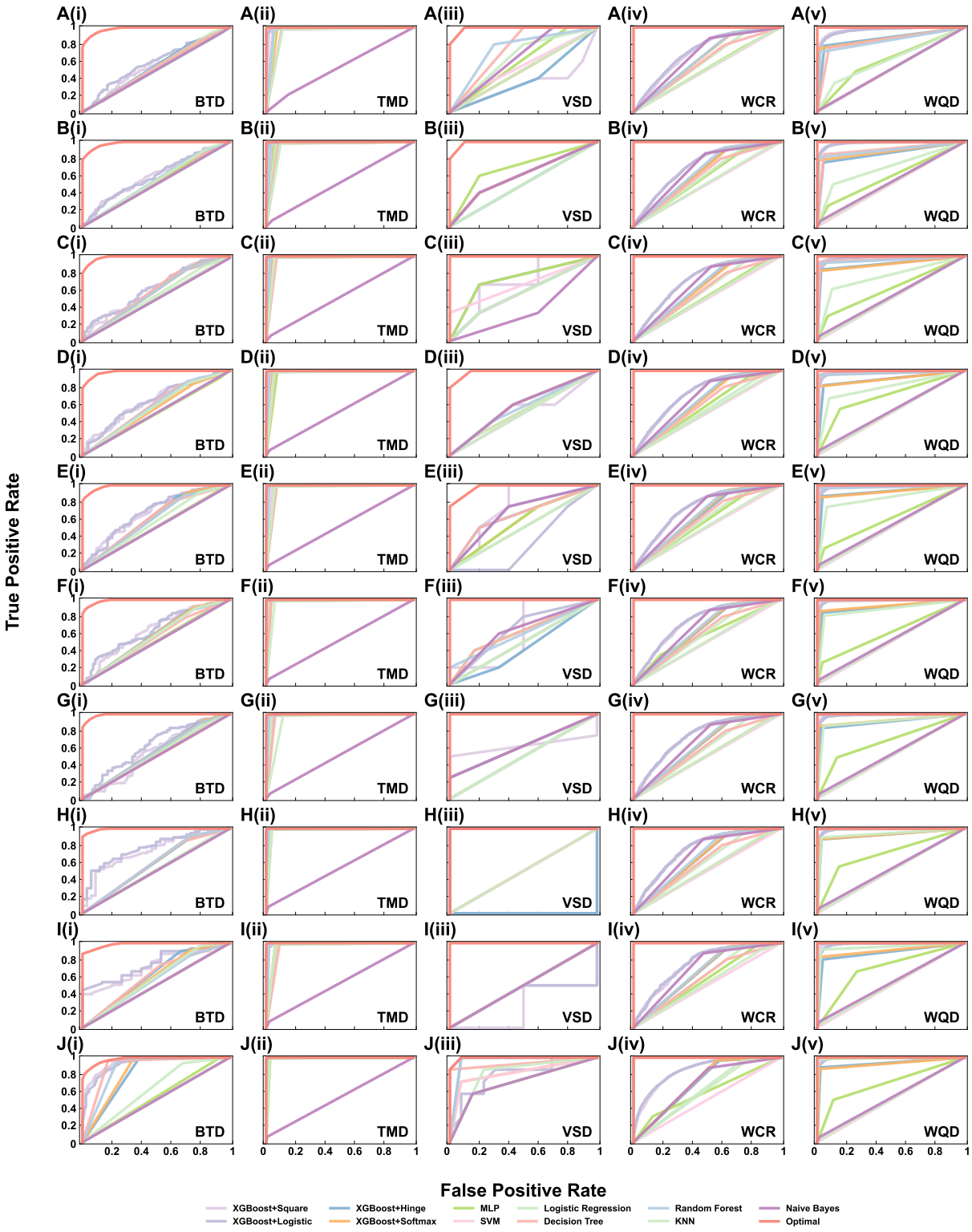

Figure S9. Exact upper bound of AUC and corresponding optimal ROC curves on 5 additional real-world datasets (BTD, TMD, VSD, WCR and WQD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

II. EXACT UPPER BOUND OF AP AND CORRESPONDING OPTIMAL PR CURVES

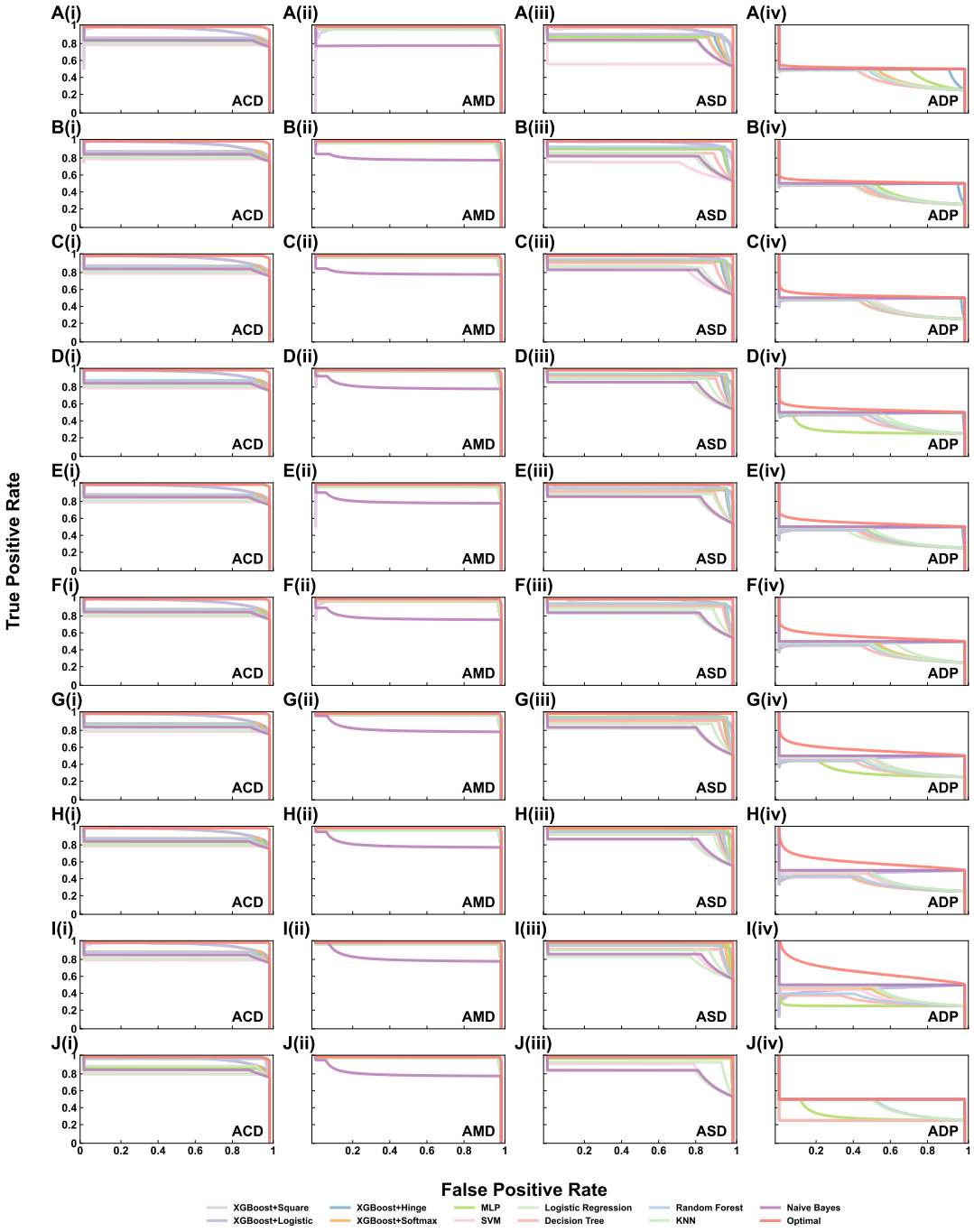

Figure S10. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (ACD, AMD, ASD and ADP) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

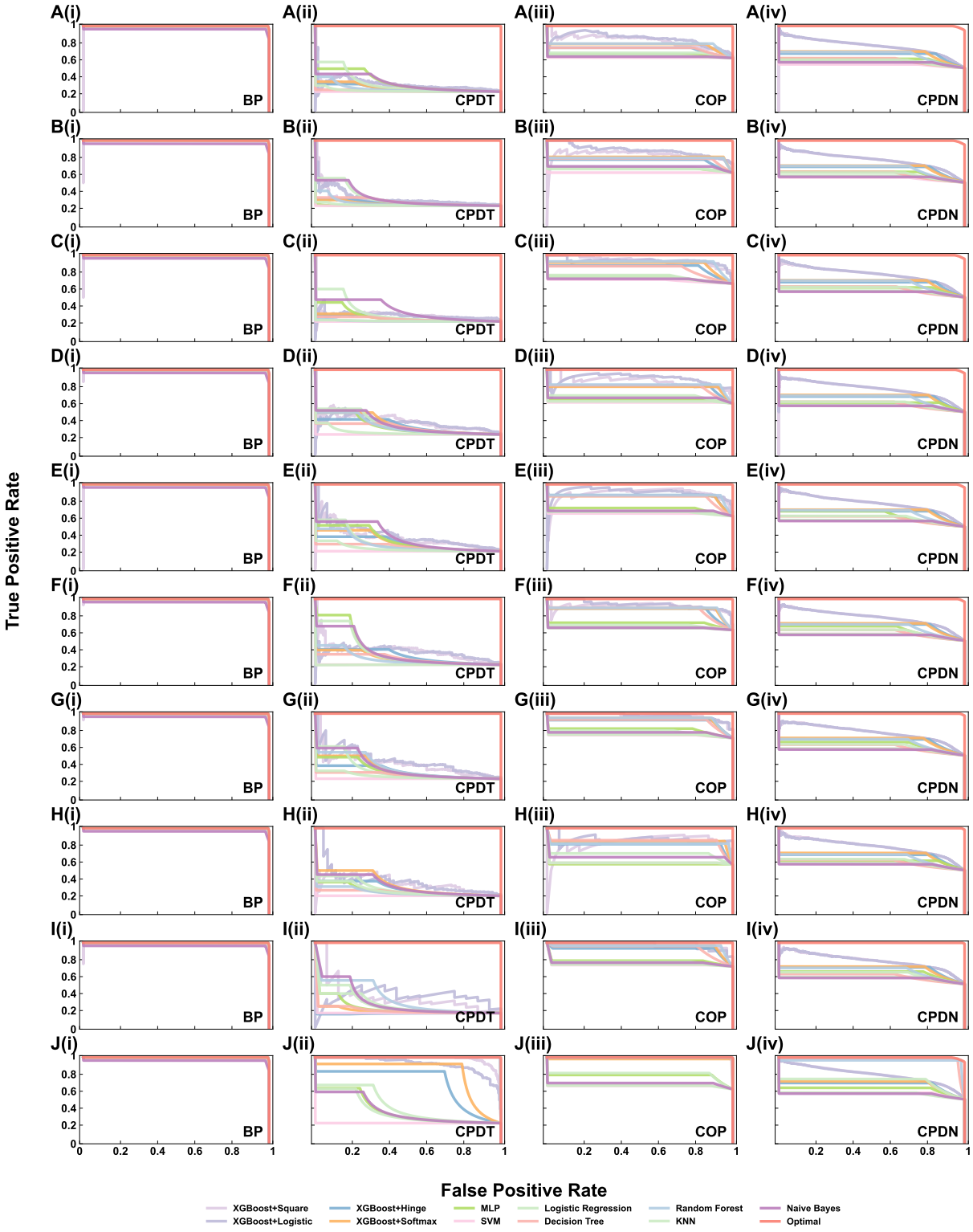

Figure S11. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (BP, CPDT, COP and CPDN) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

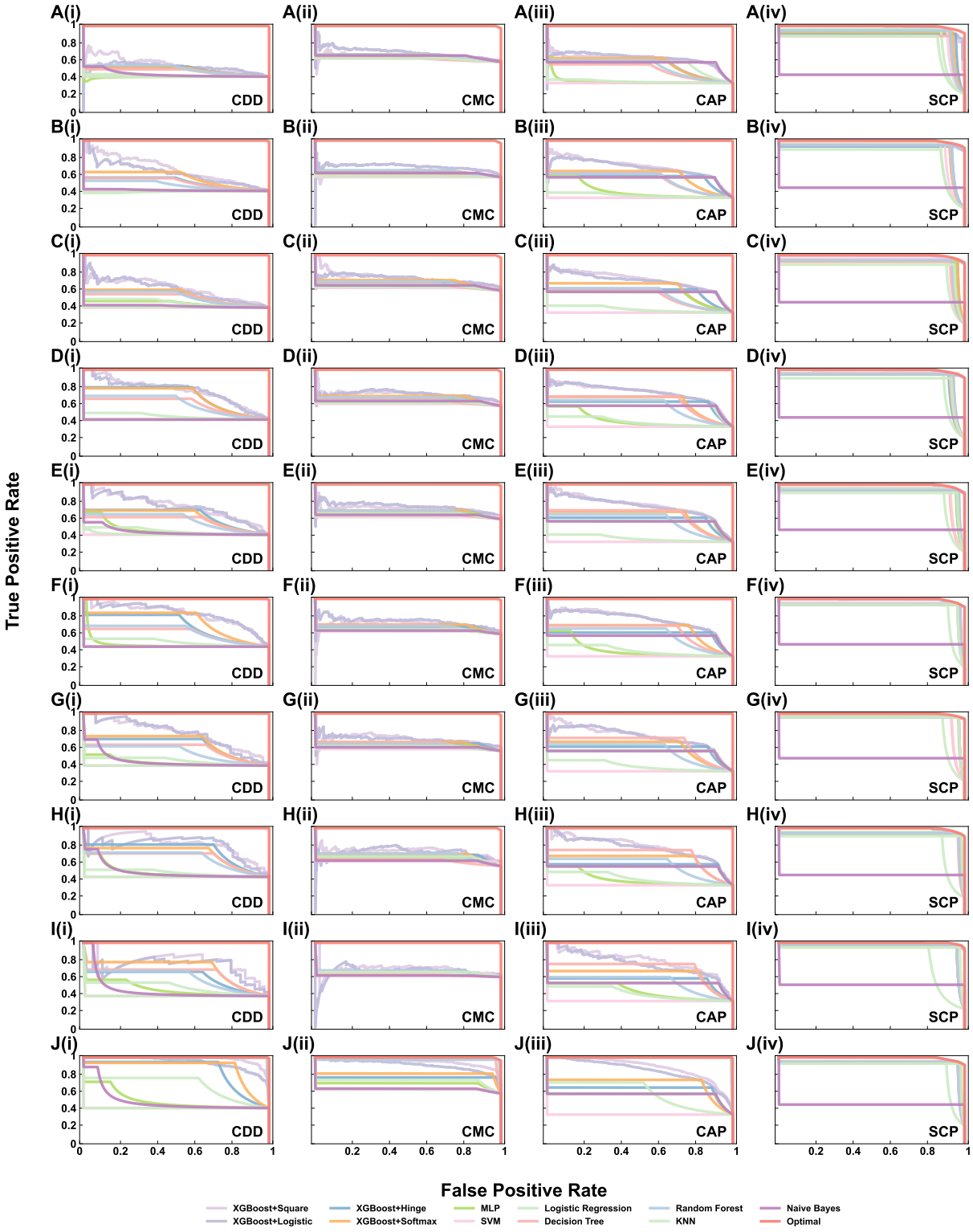

Figure S12. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (CDD, CMC, CAP and SCP) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

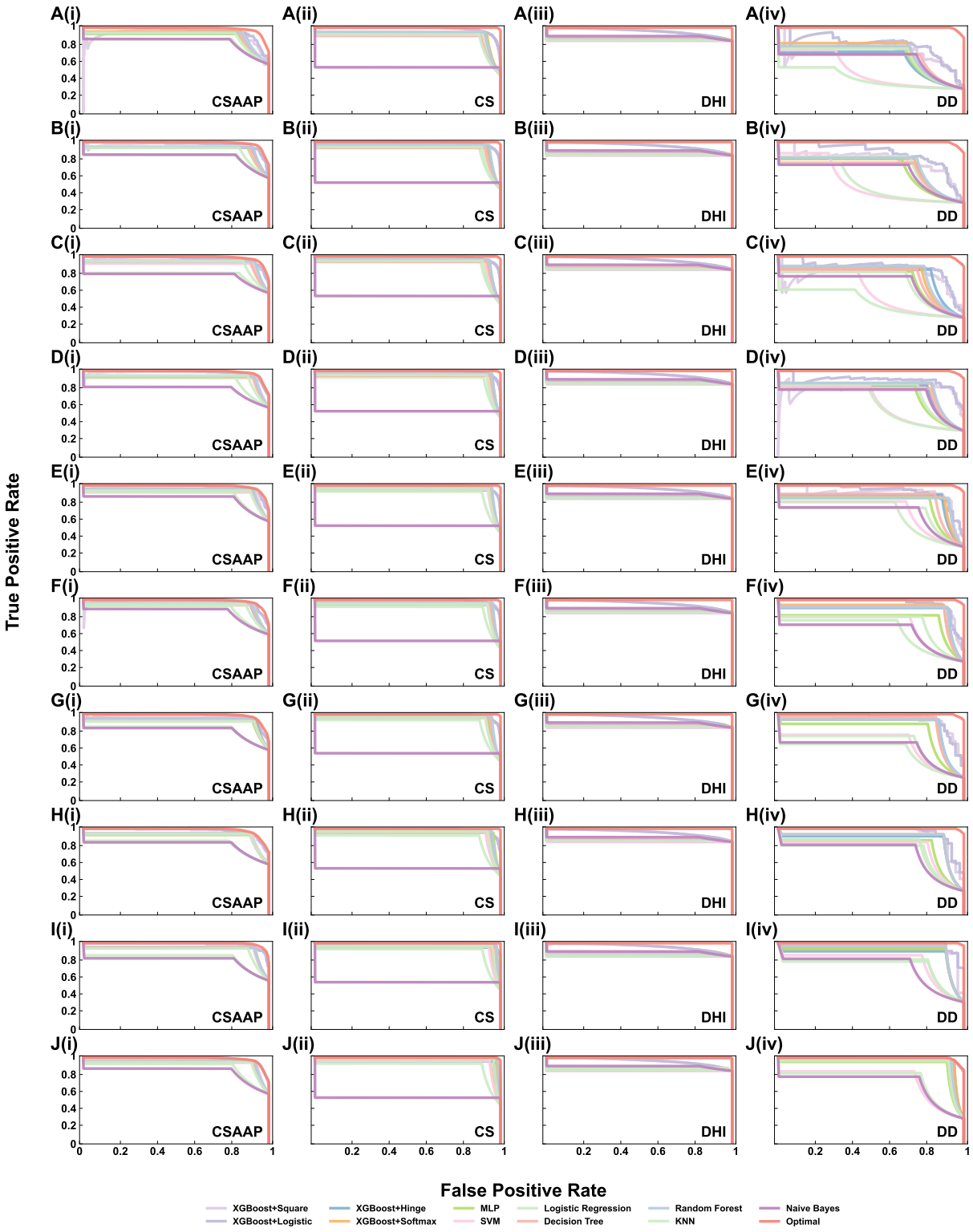

Figure S13. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (CSAAP, CS, DHI and DD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

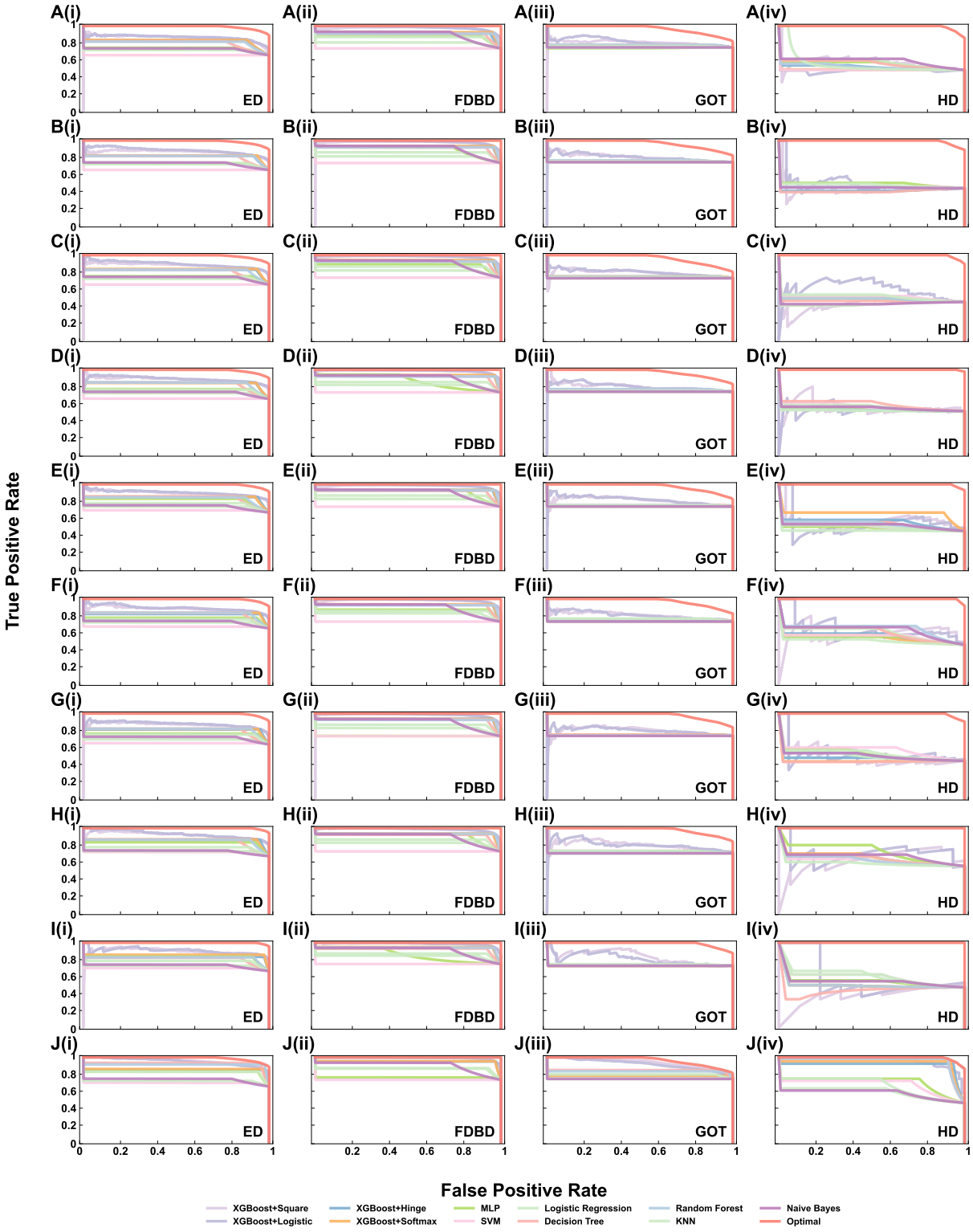

Figure S14. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (ED, FDBD, GOT and HD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

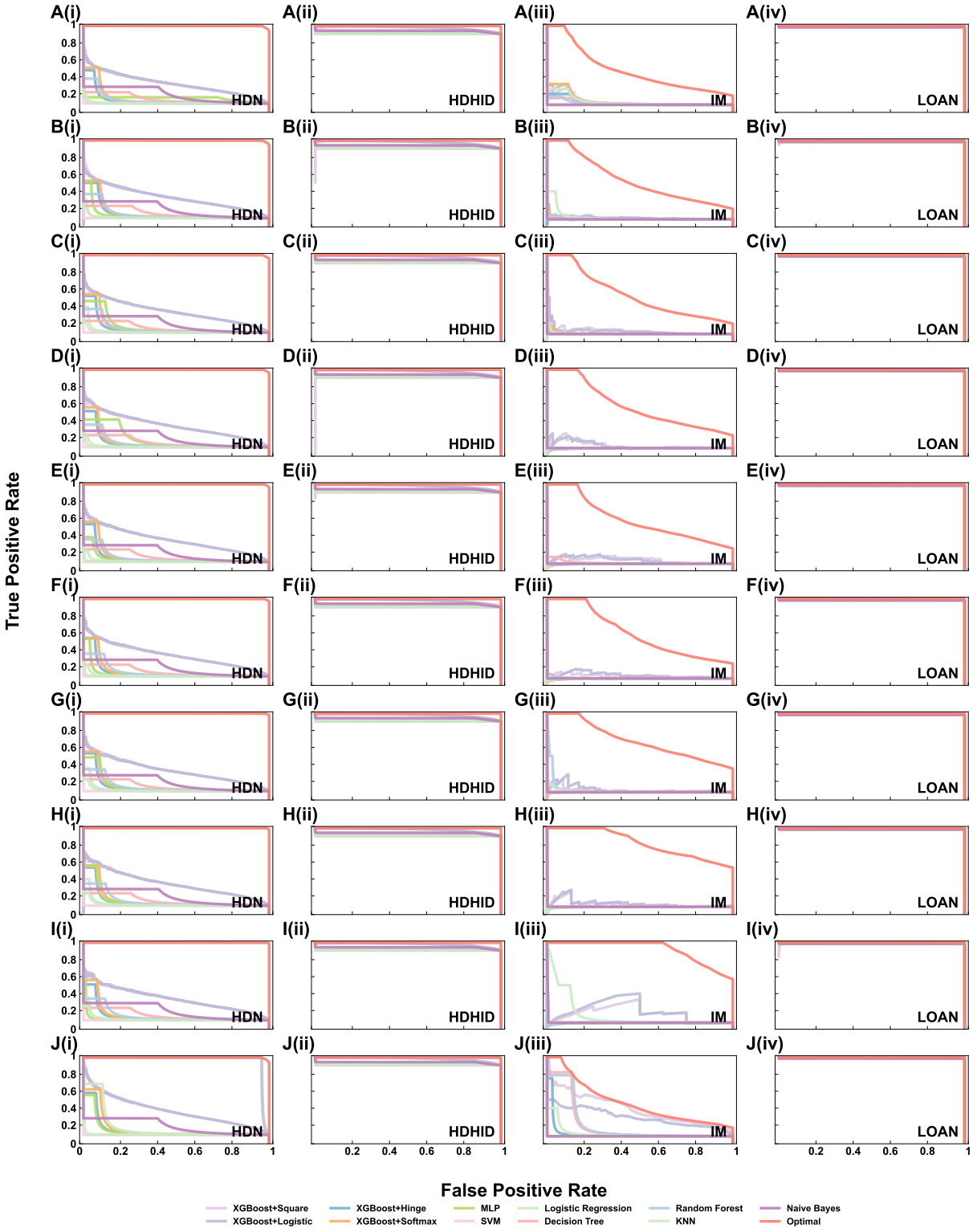

Figure S15. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (HDN, HDHID, IM and LOAN) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

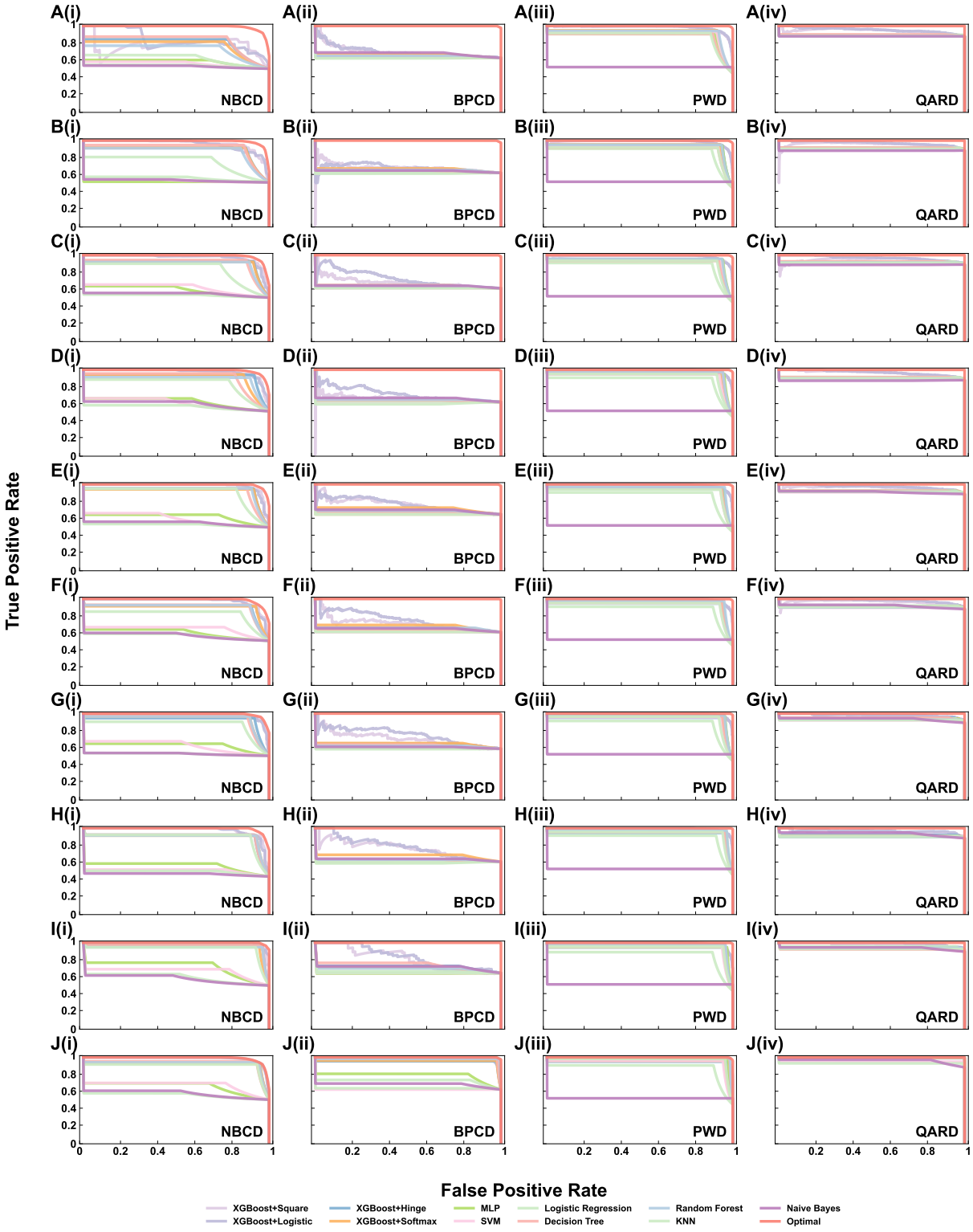

Figure S16. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (NBCD, BPCD, PWD and QARD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

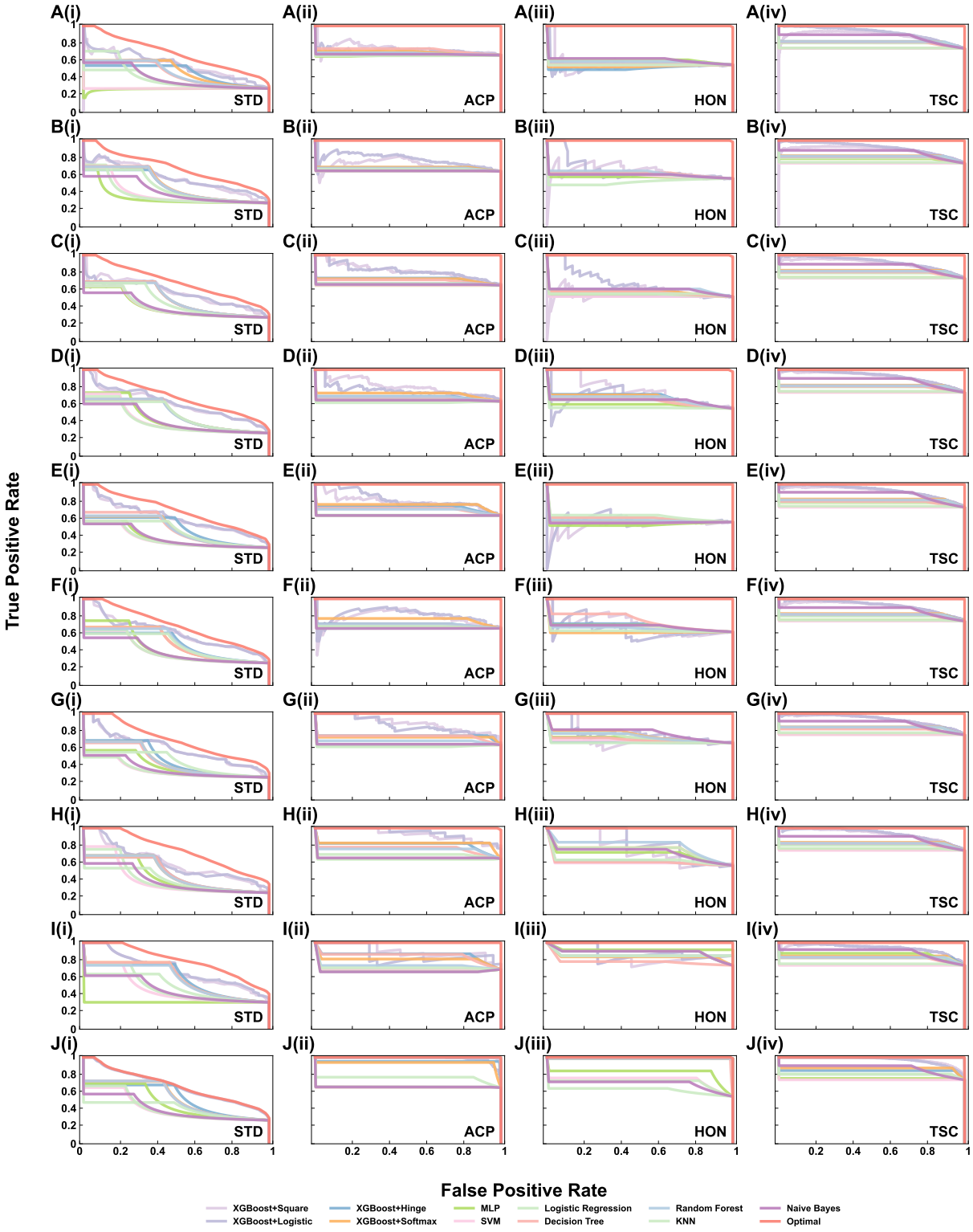

Figure S17. Exact upper bound of AP and corresponding optimal PR curves on 4 additional real-world datasets (STD, ACP, HON and TSC) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

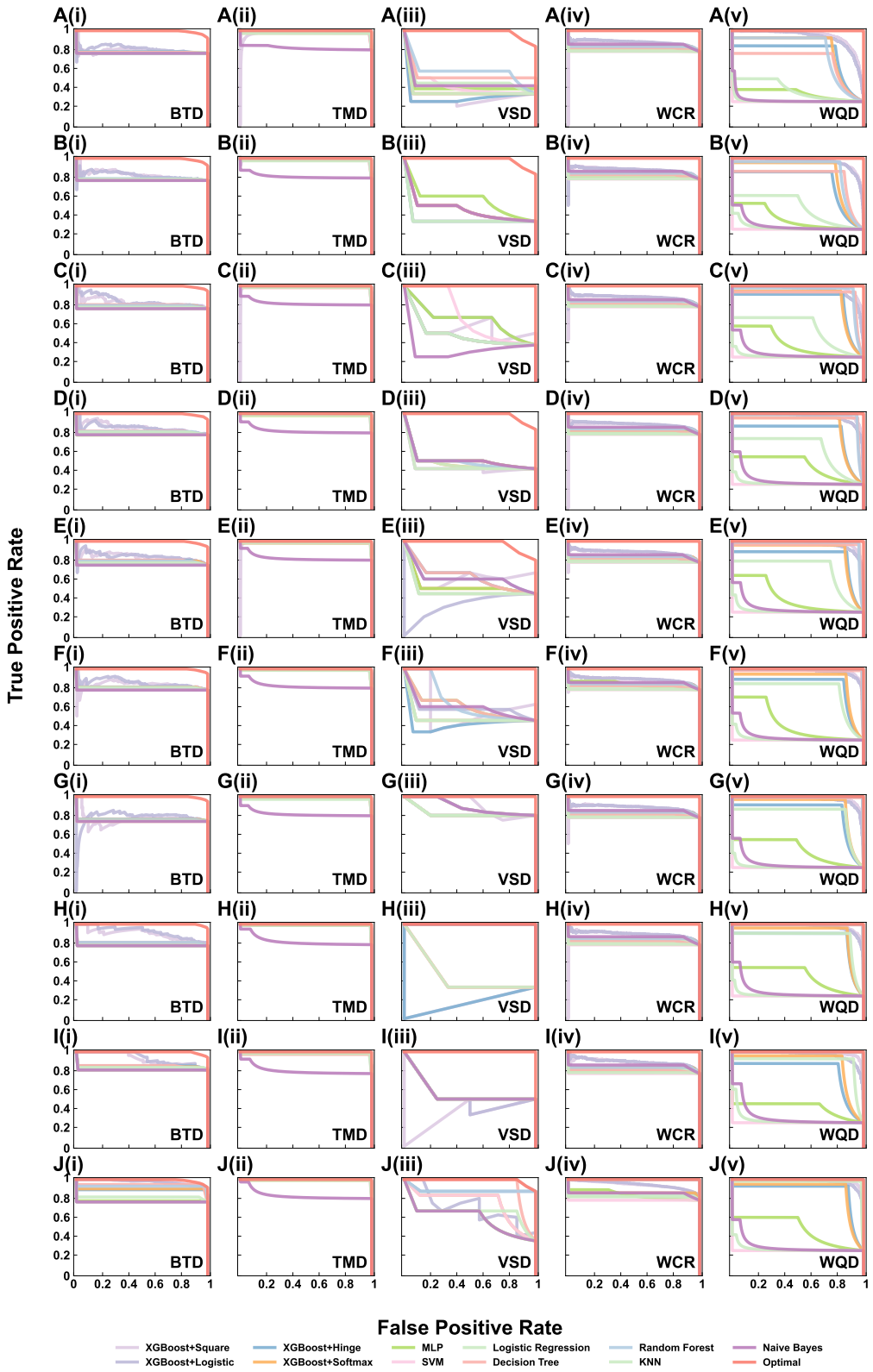

Figure S18. Exact upper bound of AP and corresponding optimal PR curves on 5 additional real-world datasets (BTD, TMD, VSD, WCR and WQD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I) and $|\mathcal{S}_{train}|/|\mathcal{S}| = 1$ (J). The binary classifiers we used in this experiment include XGBoost, MLP, SVM, Logistic Regresion, Decision Tree, Random Forest, KNN and Naive Bayes. Red curves represent the theoretical optimal ROC curves.

21

III. THE ERROR DYNAMICS DURING TRAINING AND TESTING PHASES

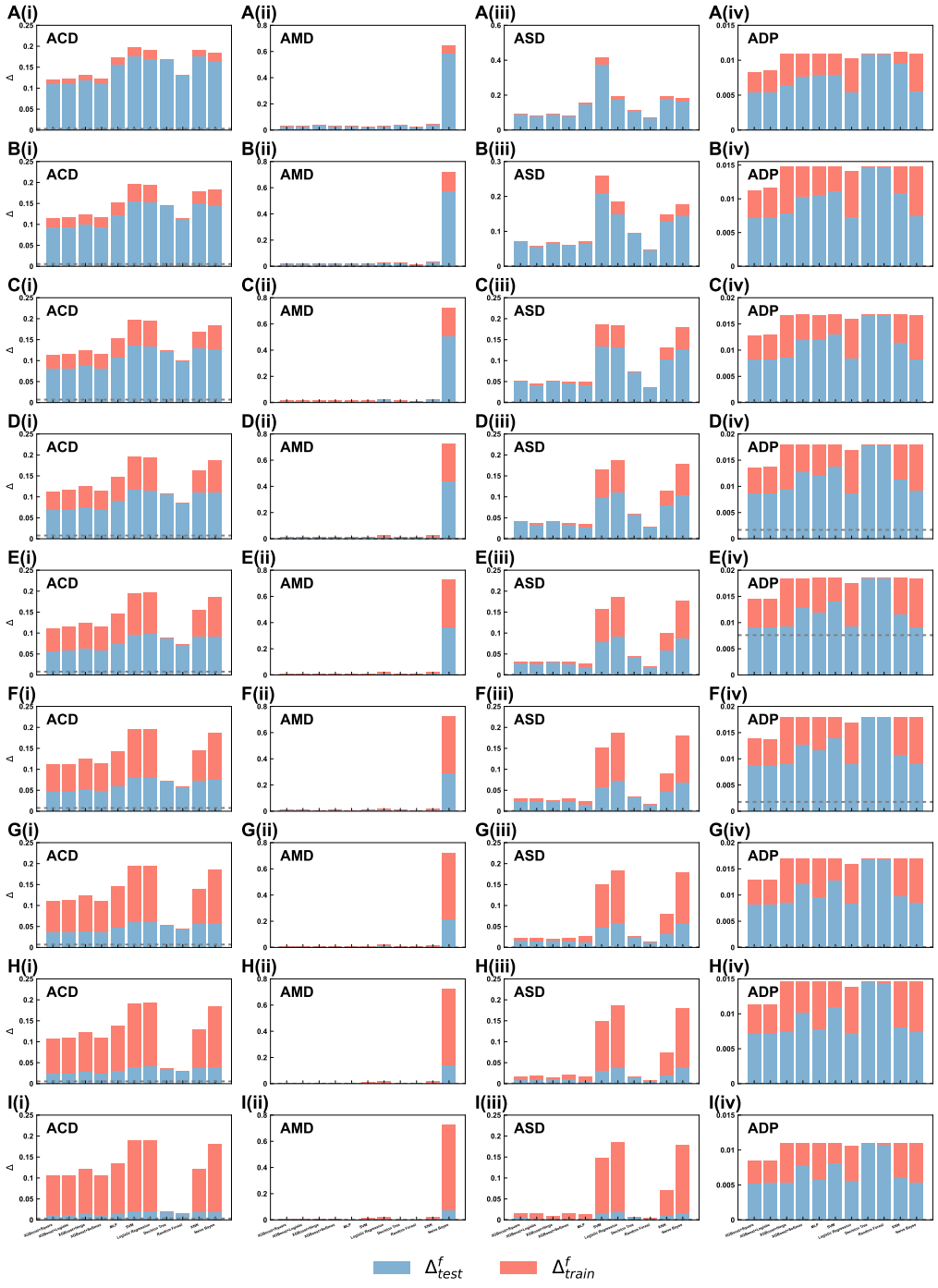

Figure S19. The error dynamics on 4 additional datasets (ACD, AMD, ASD and ADP) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

Figure S20. The error dynamics on 4 additional datasets BP, CPDT, COP and CPDN) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

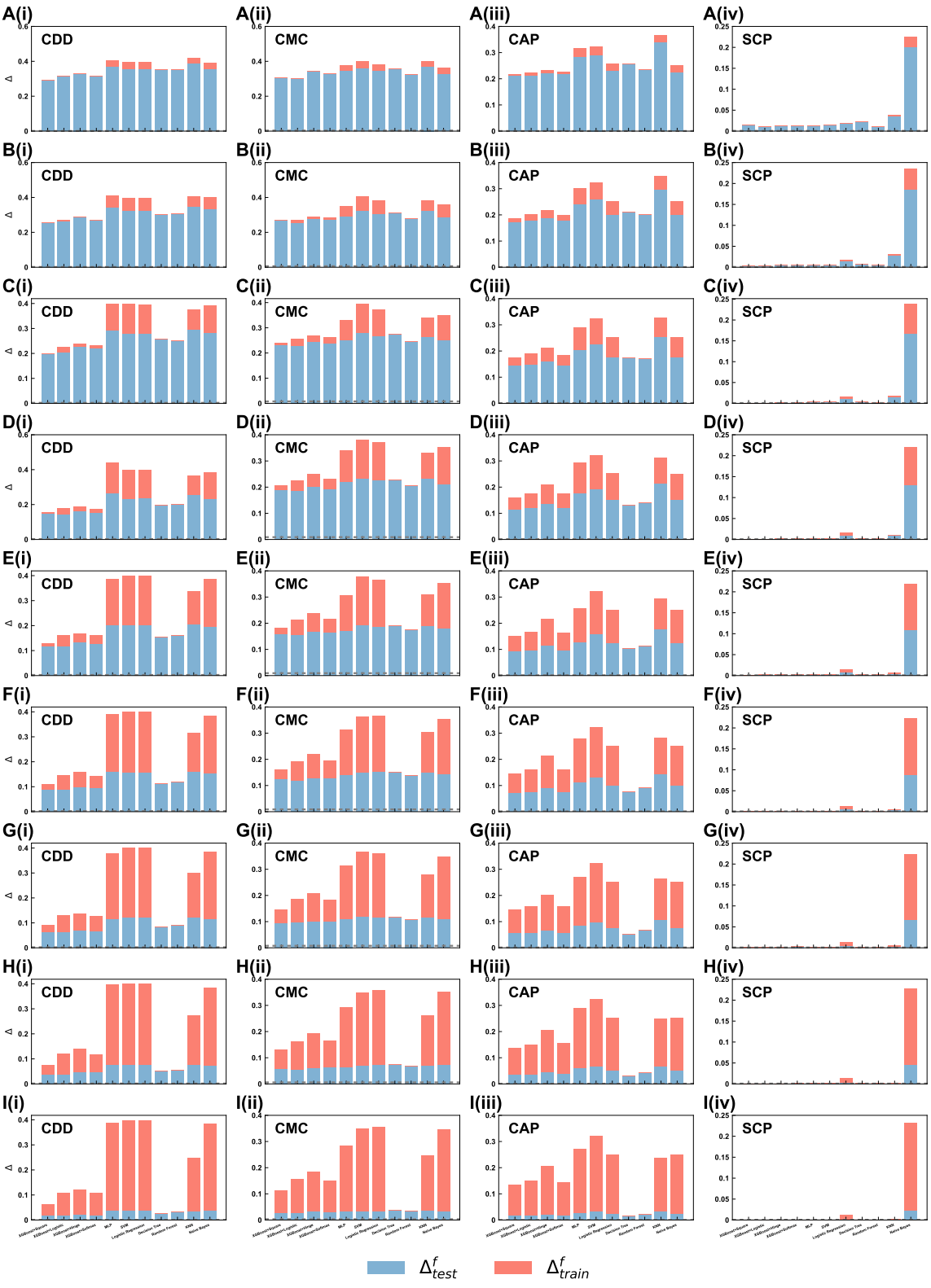

Figure S21. The error dynamics on 4 additional datasets (CDD, CMC, CAP and SCP) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

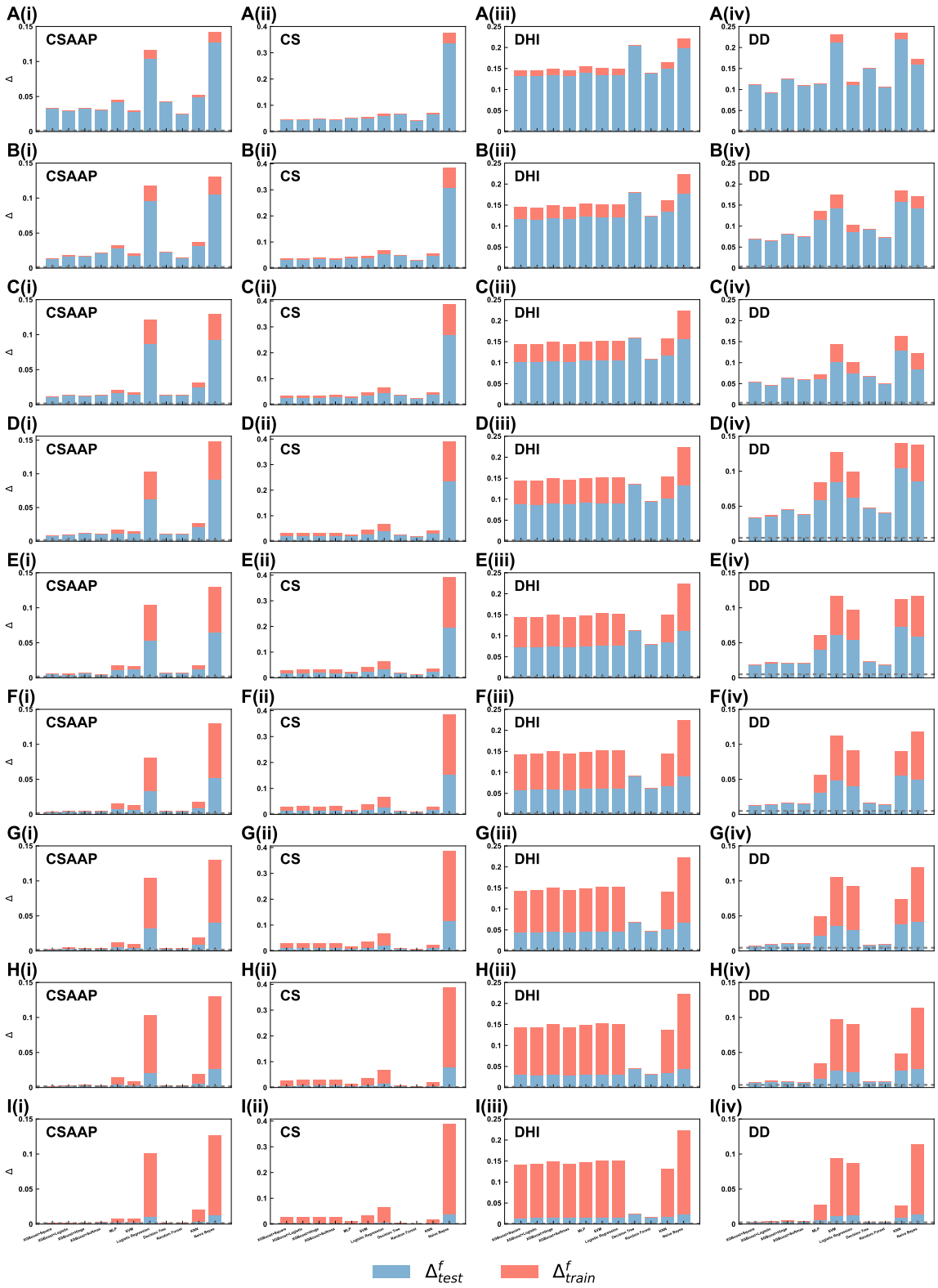

Figure S22. The error dynamics on 4 additional datasets (CSAAP, CS, DHI and DD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

Figure S23. The error dynamics on 4 additional datasets (ED, FDBD, GOT and HD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

Figure S24. The error dynamics on 4 additional datasets (HDN, HDHID, IM and LOAN) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

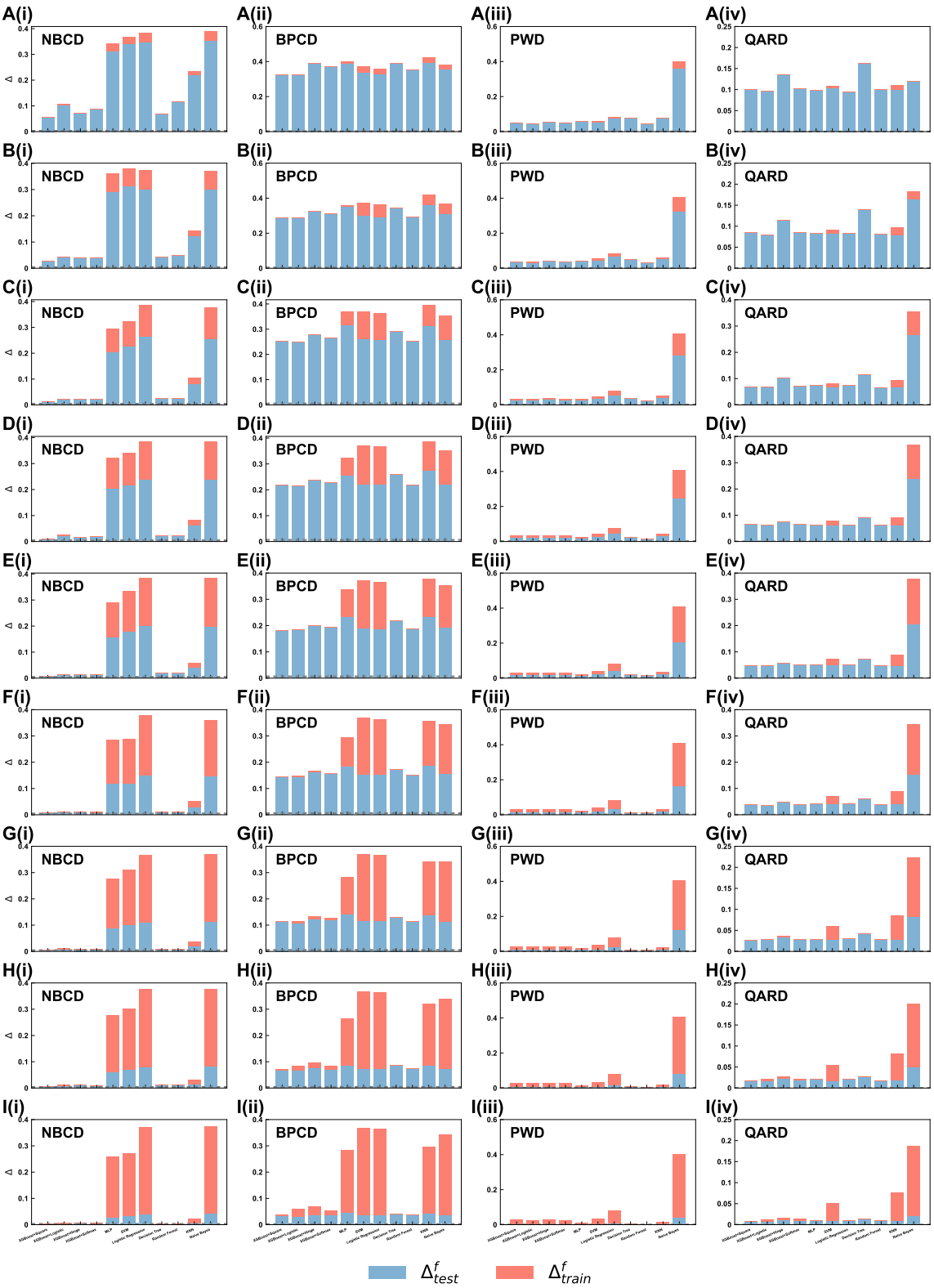

Figure S25. The error dynamics on 4 additional datasets (NBCD, BPCD, PWD and QARD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

Figure S26. The error dynamics on 4 additional datasets (STD, ACP, HON and TSC) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

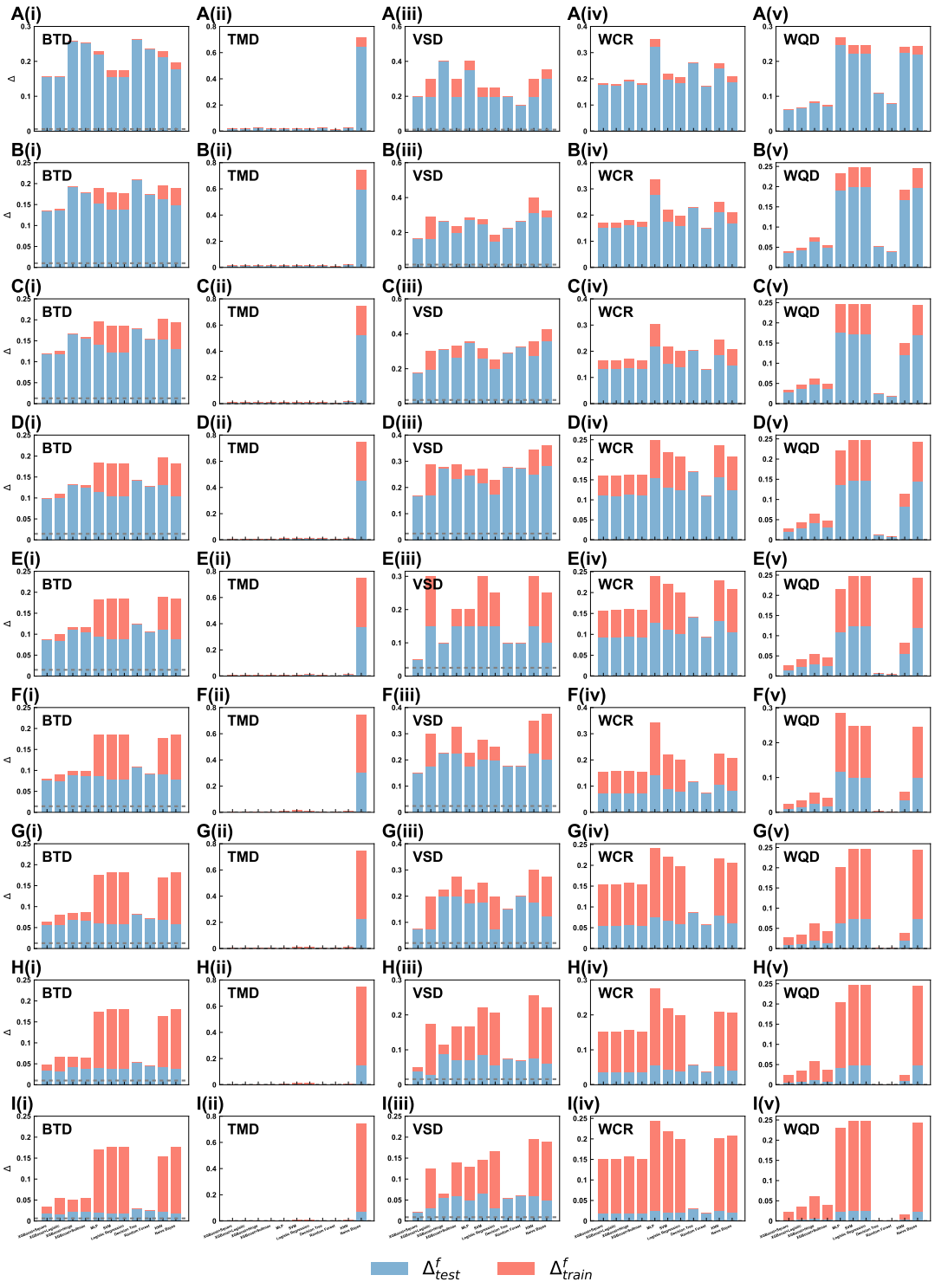

Figure S27. The error dynamics on 5 additional datasets (BTD, TMD, VSD, WCR and WQD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Dash line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

22

IV. THE CORRELATION BETWEEN Δ^f_{train} AND Δ^f_{test}

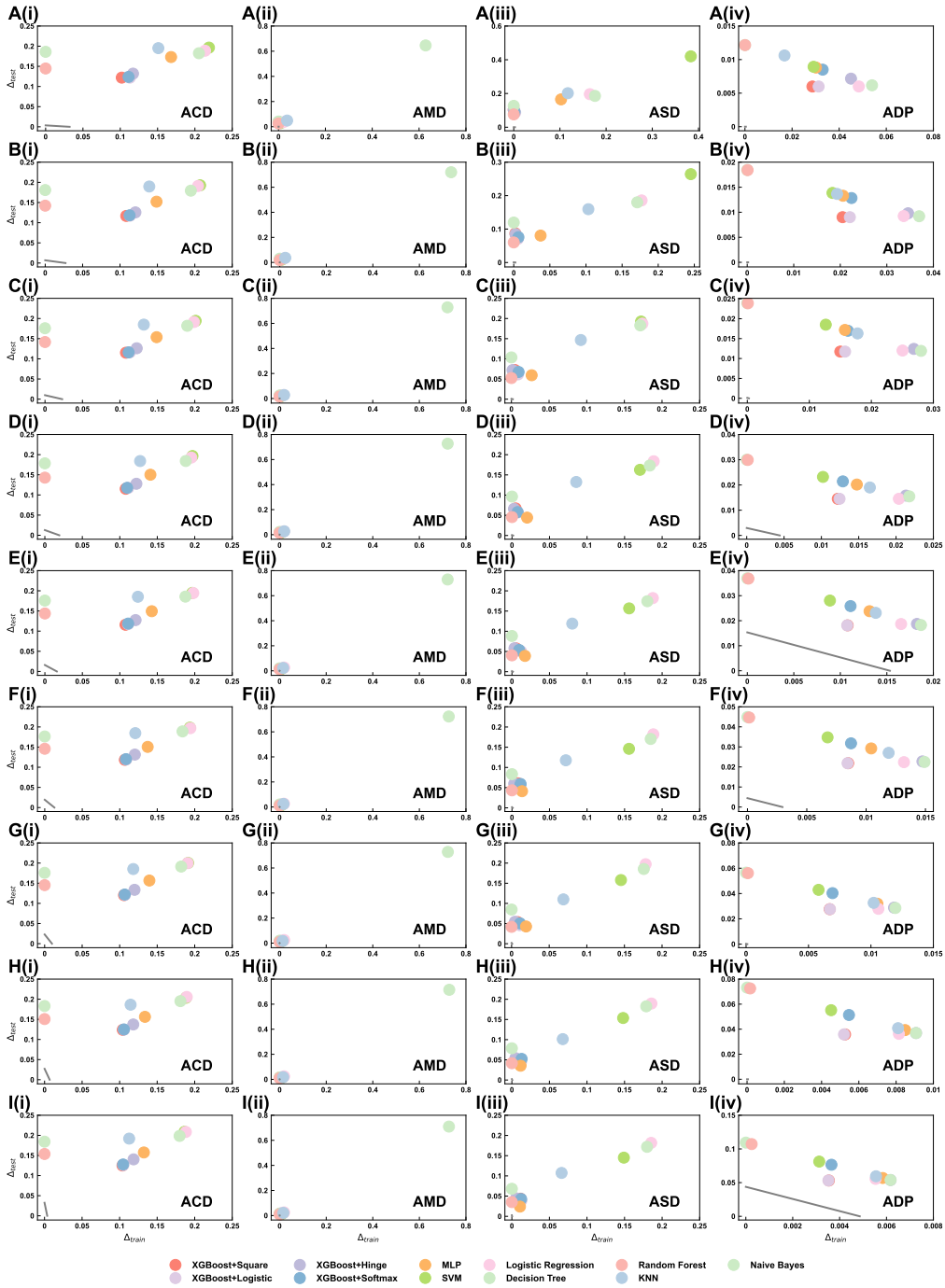

Figure S28. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (ACD, AMD, ASD and ADP) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

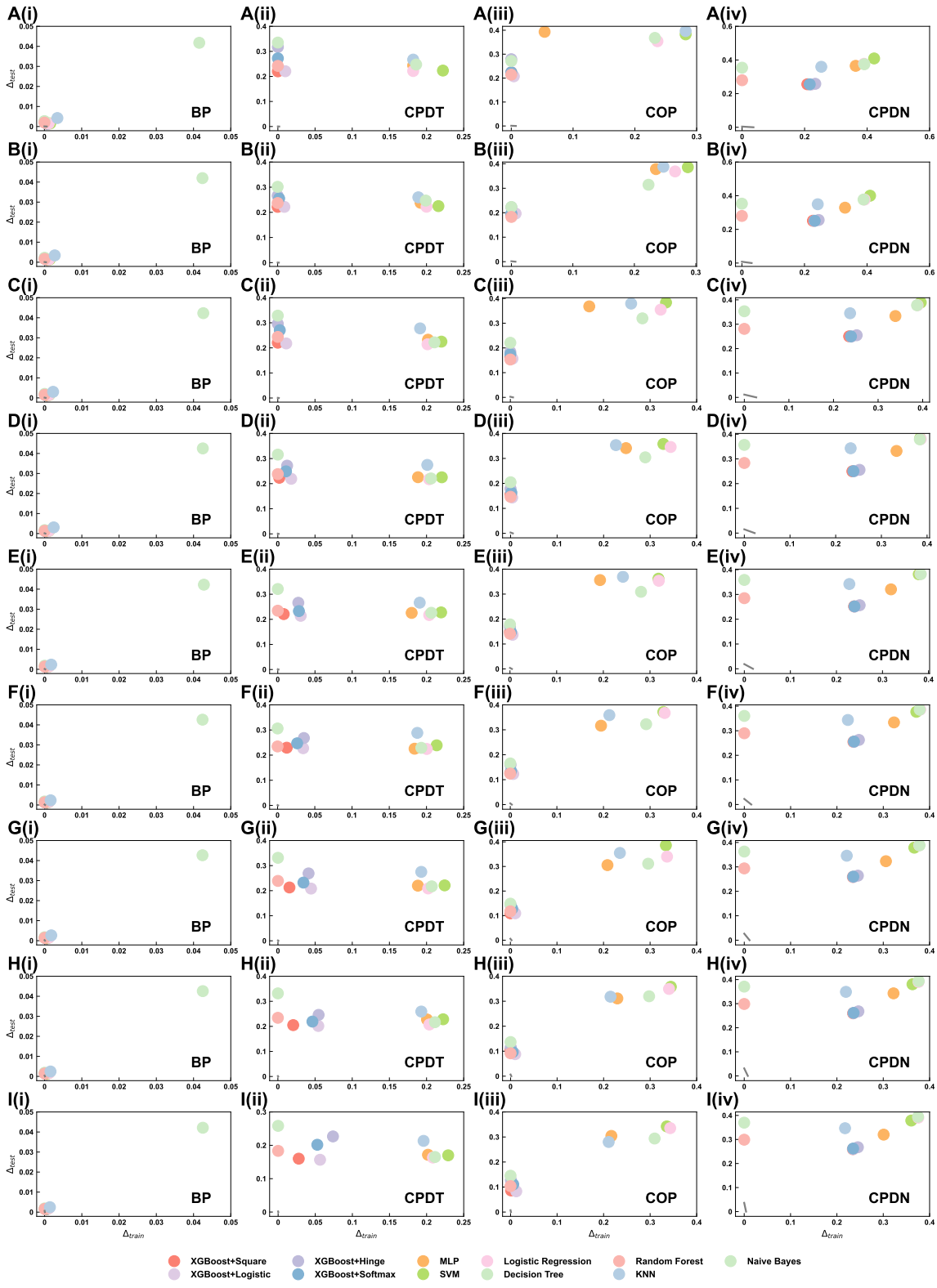

Figure S29. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (BP, CPDT, COP and CPDN) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

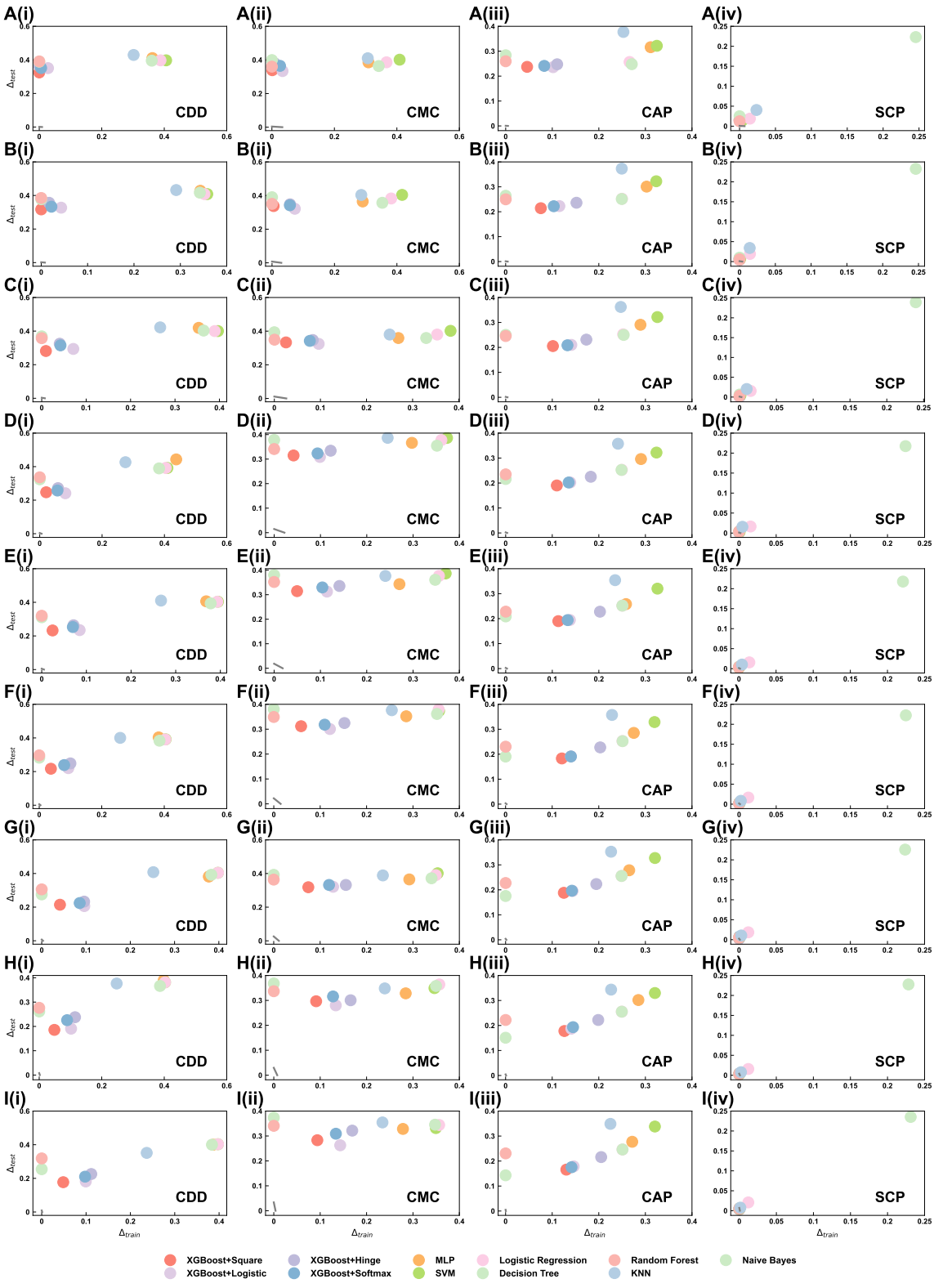

Figure S30. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (CDD, CMC, CAP and SCP) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

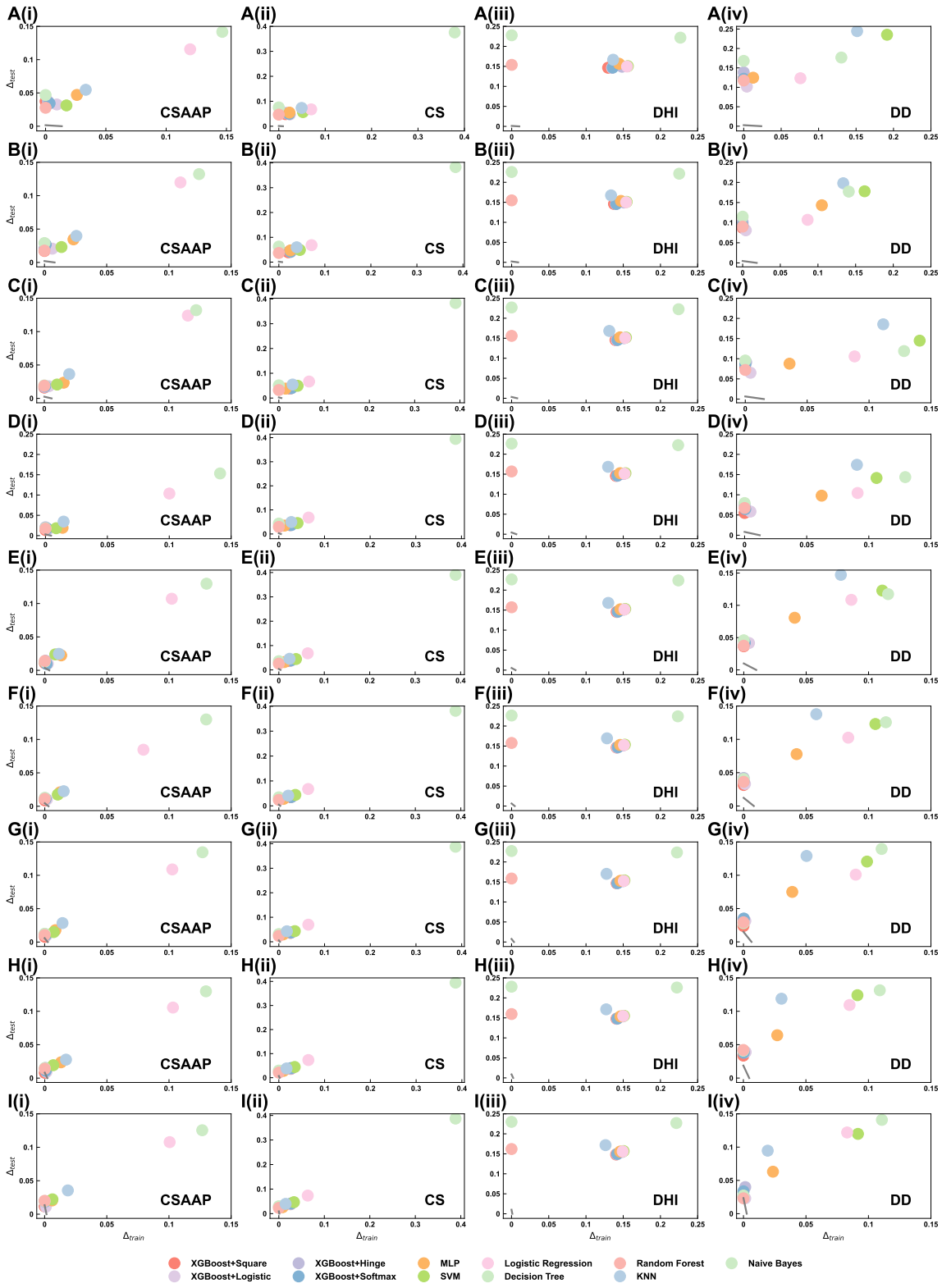

Figure S31. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (CSAAP, CS, DHI and DD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

Figure S32. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (ED, FDBD, GOT and HD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

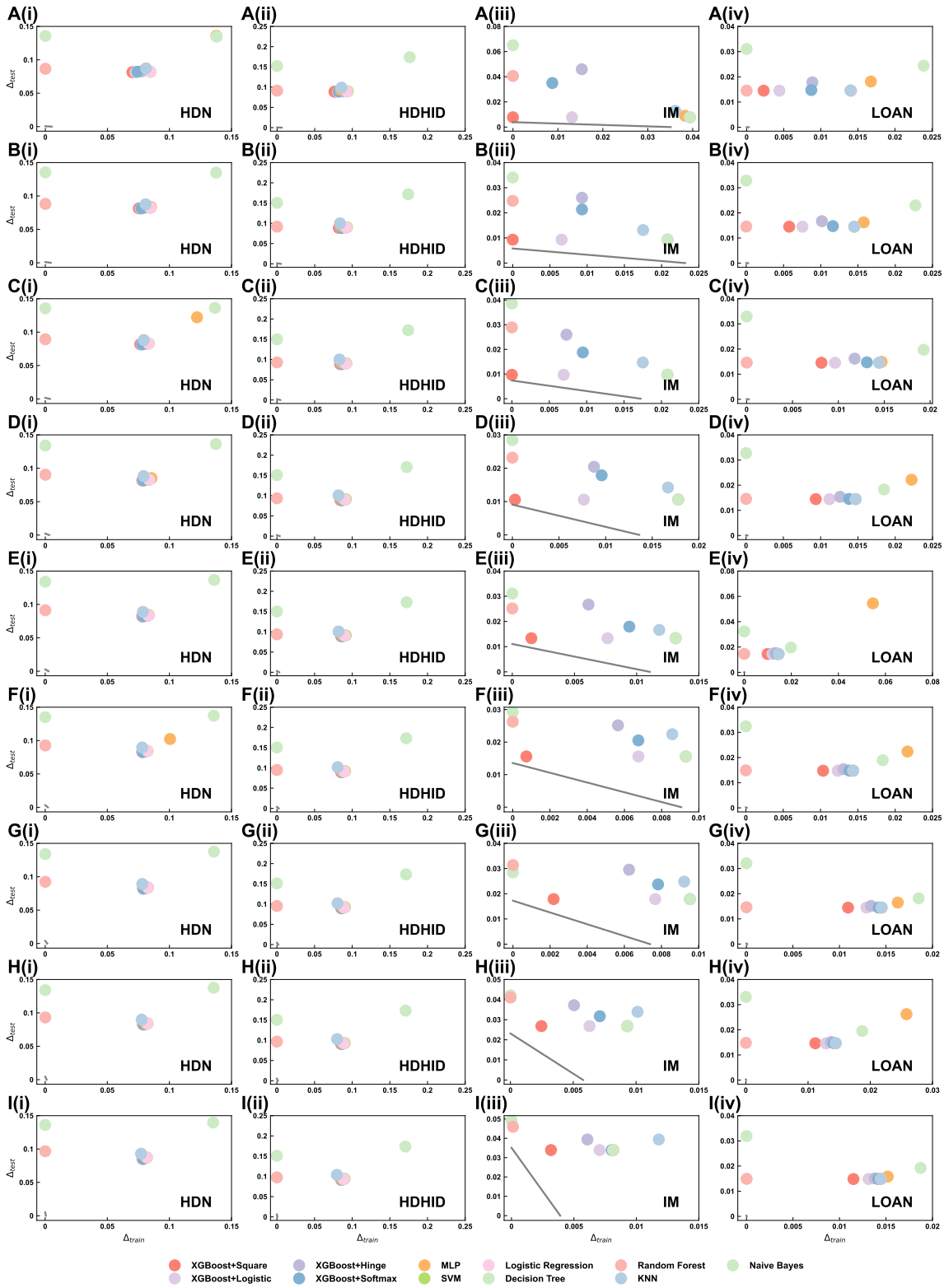

Figure S33. The correlation between Δ^f_{train} and Δ^f_{test} onr 4 additional datasets (HDN, HDHID, IM and LOAN) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

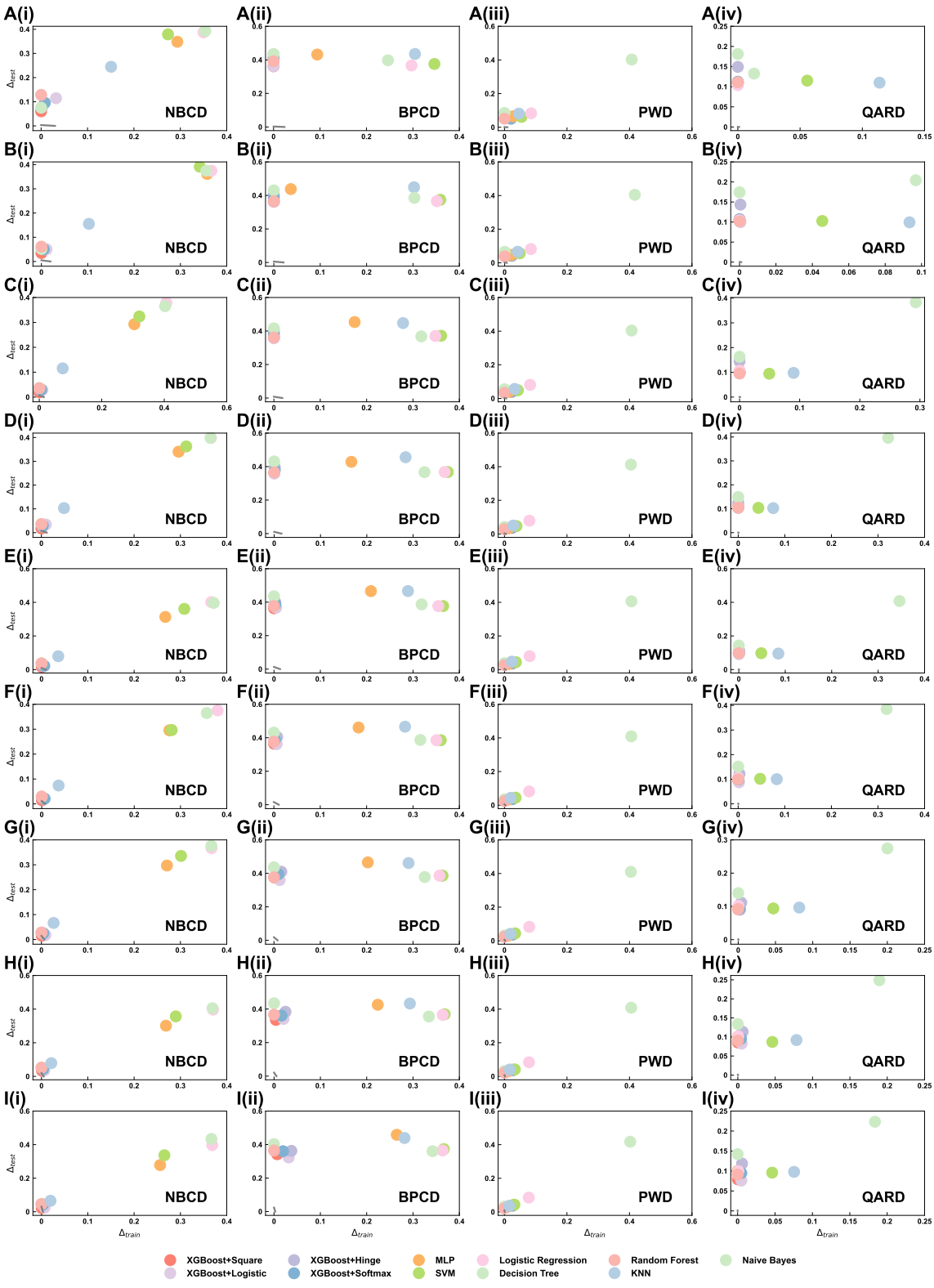

Figure S34. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (NBCD, BPCD, PWD and QARD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

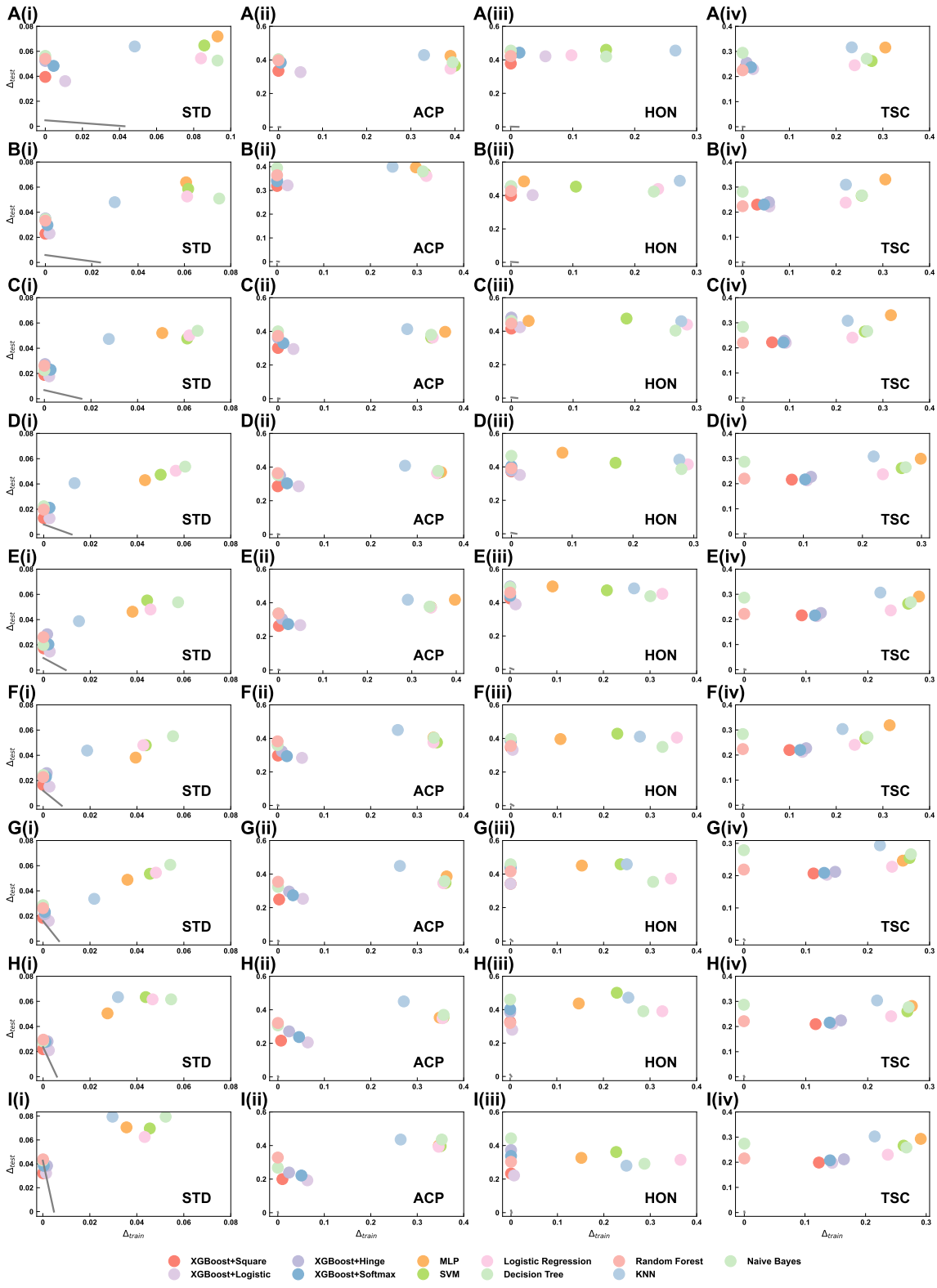

Figure S35. The correlation between Δ^f_{train} and Δ^f_{test} on 4 additional datasets (STD, ACP, HON and TSC) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

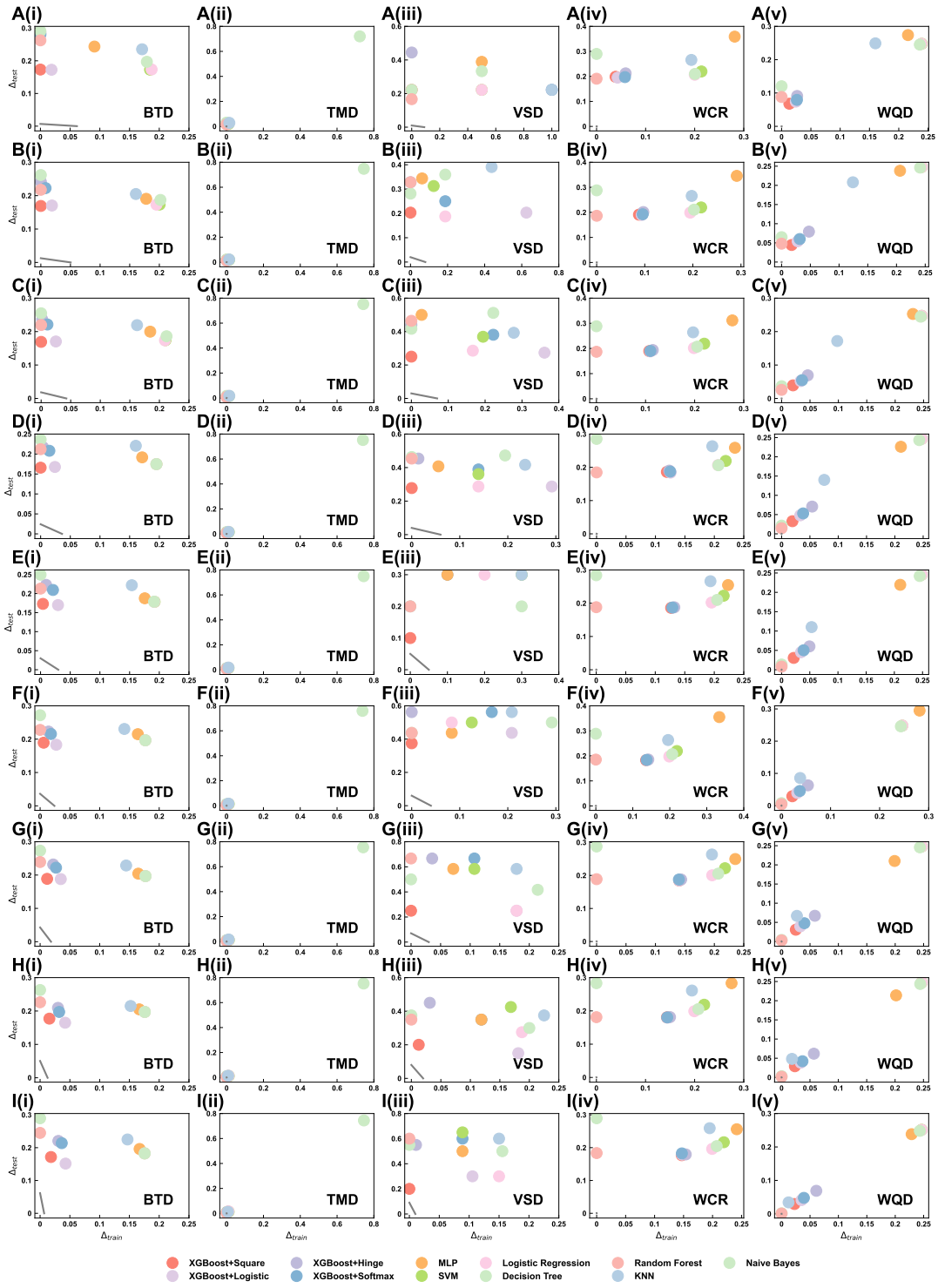

Figure S36. The correlation between Δ^f_{train} and Δ^f_{test} on 5 additional datasets (BTD, TMD, VSD, WCR and WQD) when $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.1$ (A), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.2$ (B), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.3$ (C), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.4$ (D), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.5$ (E), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.6$ (F), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.7$ (G), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.8$ (H), $|\mathcal{S}_{train}|/|\mathcal{S}| = 0.9$ (I). Gray line represents the expected error of optimal classier based on Supplemental Material, Eq. 94.

V. THE LOSS ERRORS OF DIFFERENT BINARY CLASSIFIERS

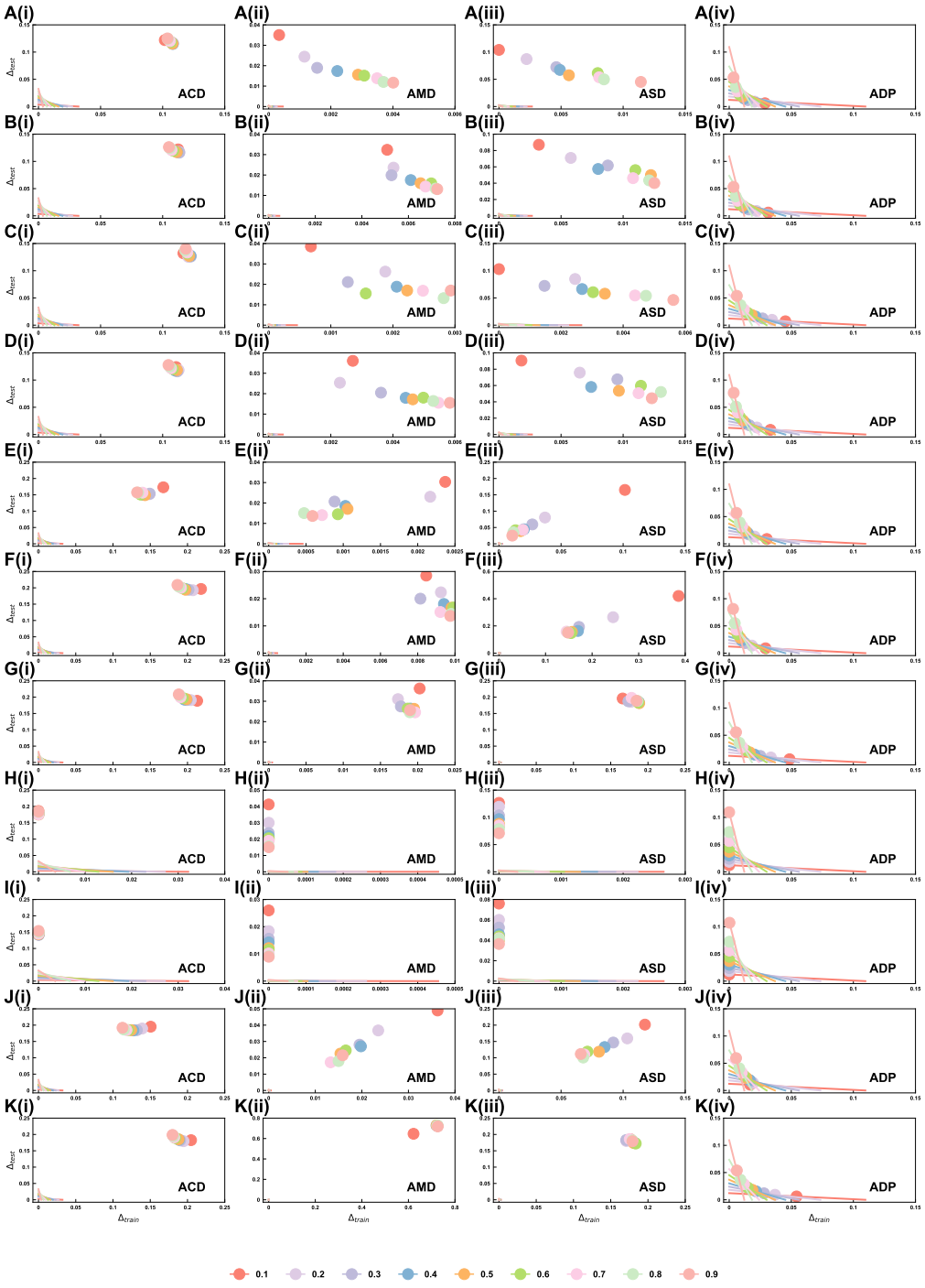

Figure S37. The loss errors on four additional datasets (ACD, AMD, ASD and ADP) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

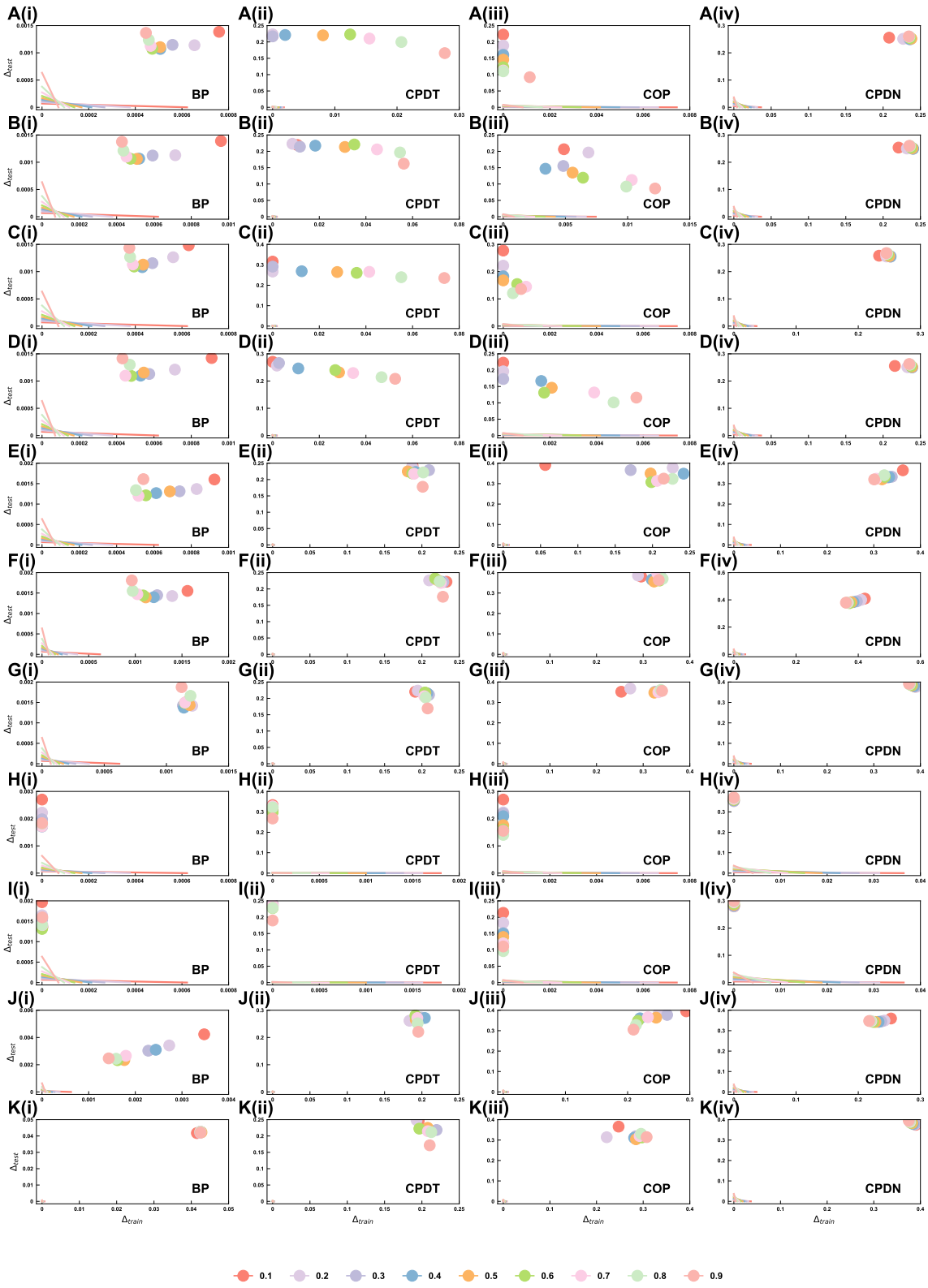

Figure S38. The loss errors on four additional datasets (BP, CPDT, COP and CPDN) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

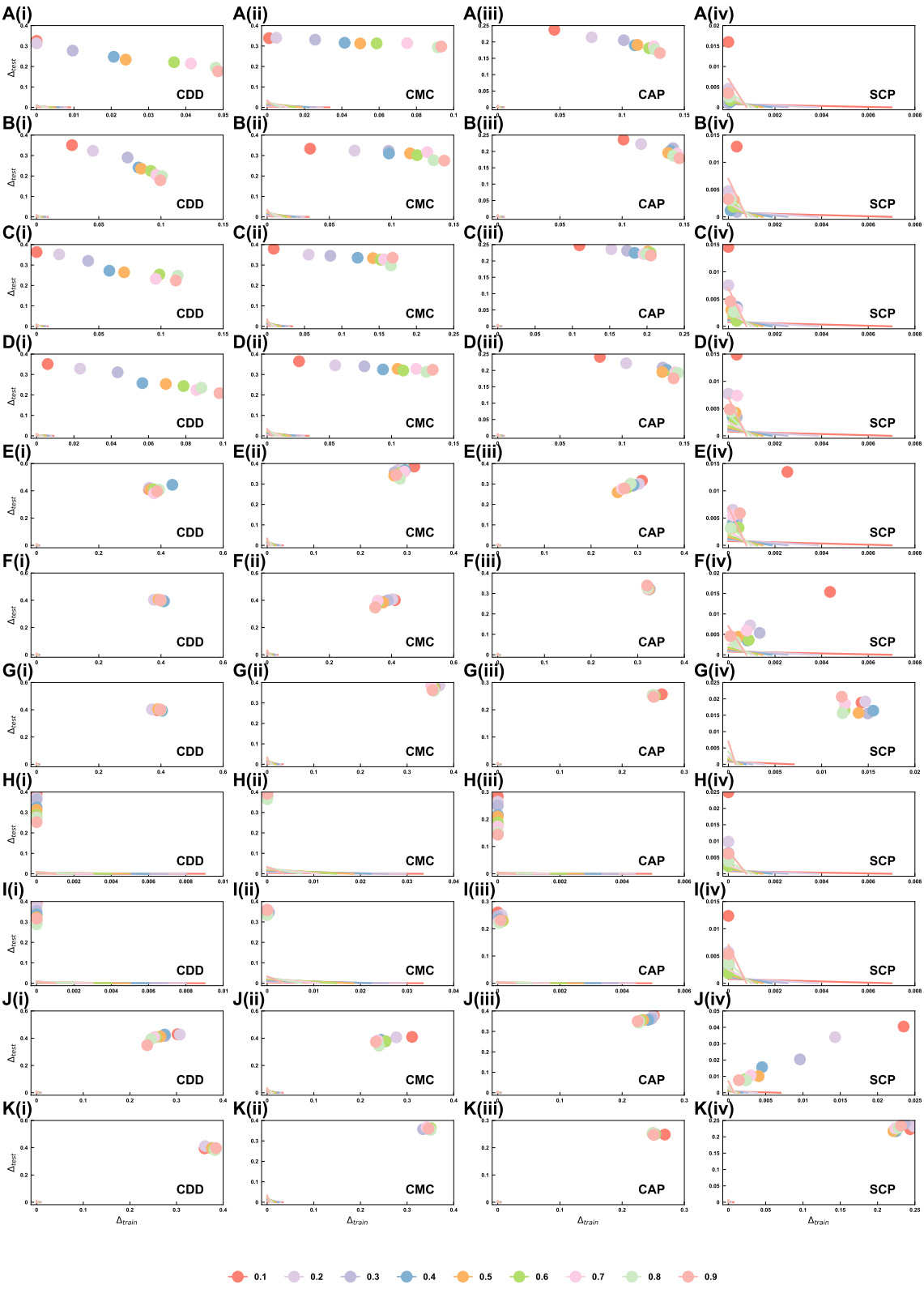

Figure S39. The loss errors on four additional datasets (CDD, CMC, CAP and SCP) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

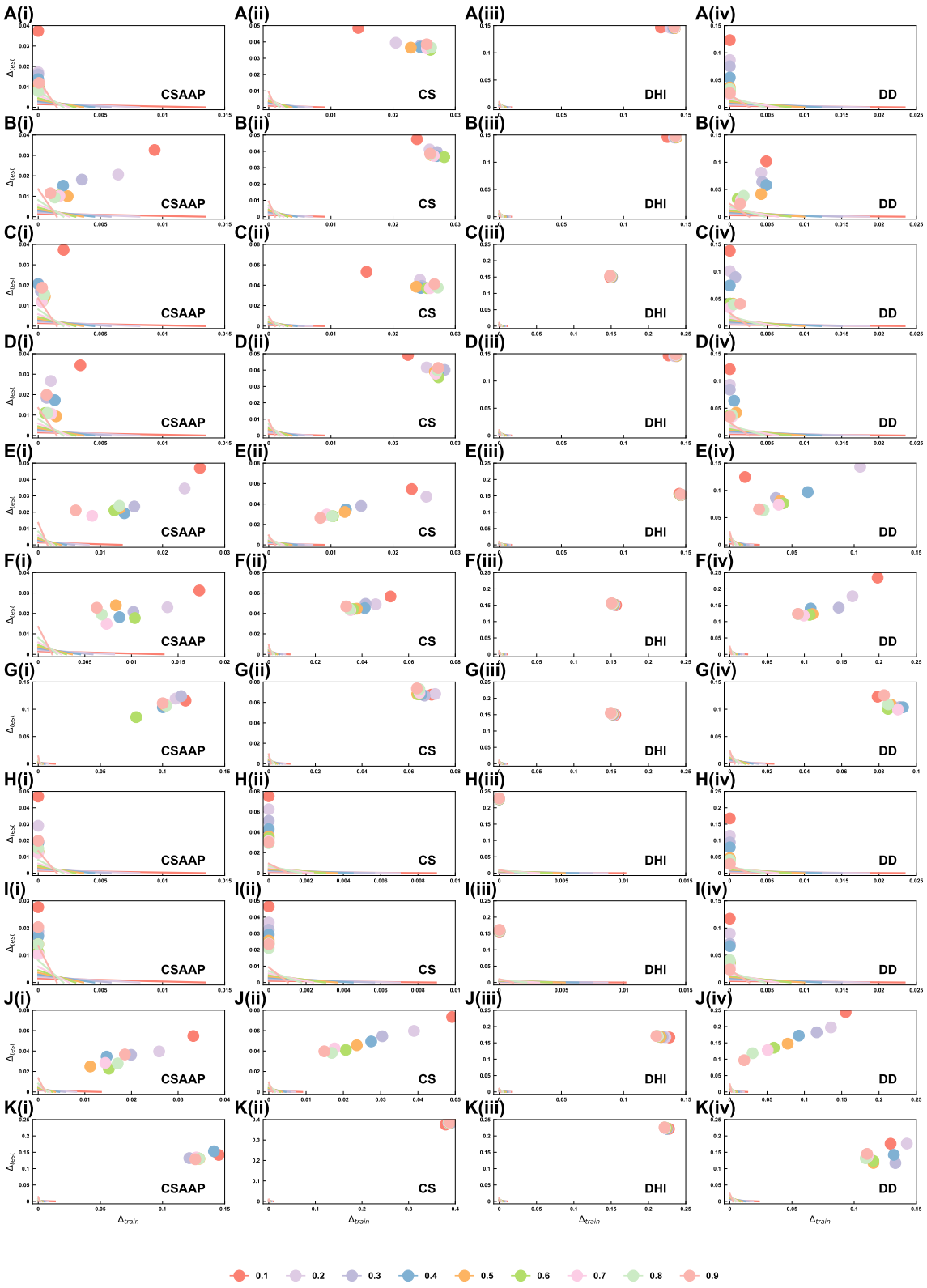

Figure S40. The loss errors on four additional datasets (CSAAP, CS, DHI and DD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

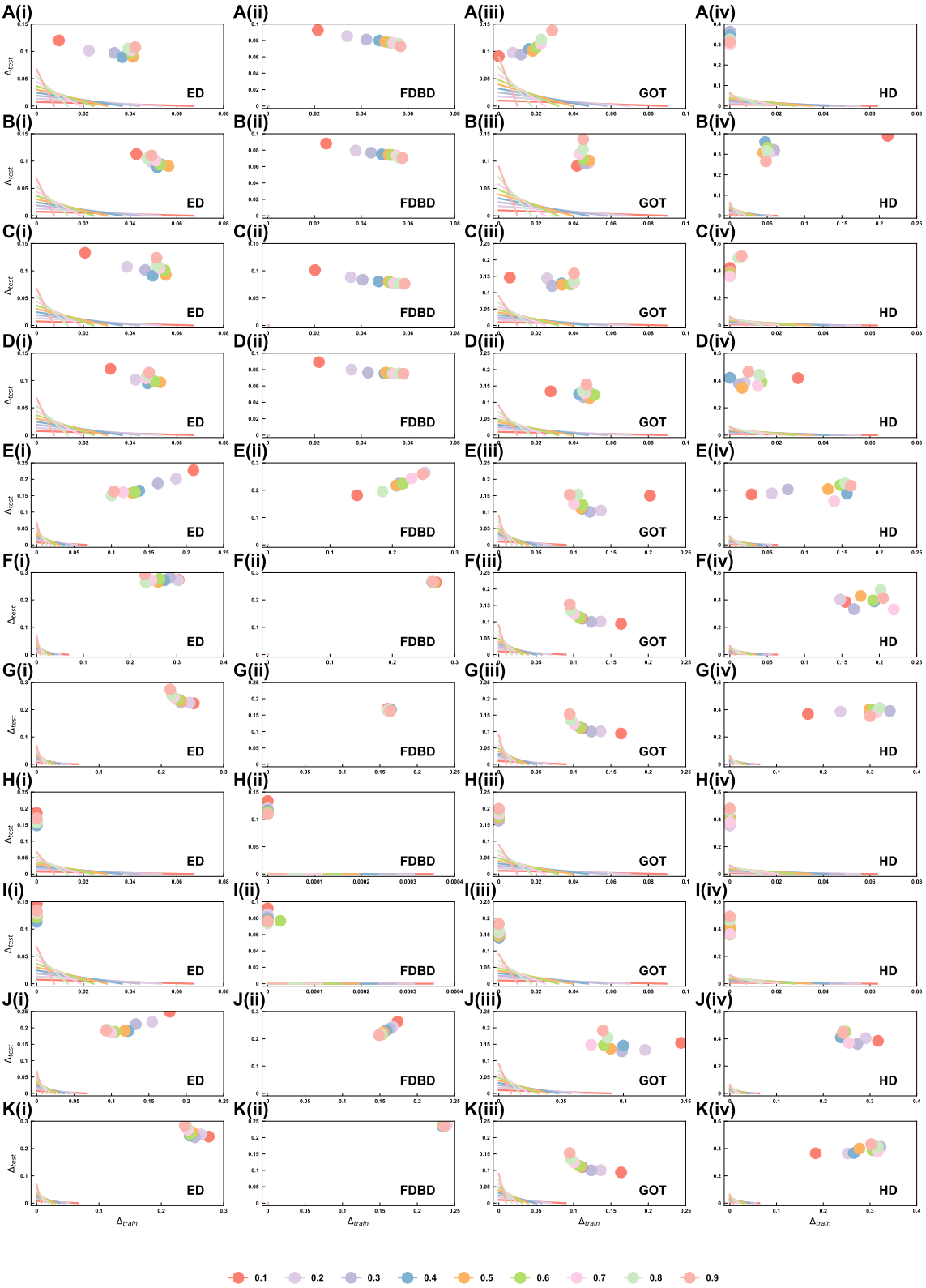

Figure S41. The loss errors on four additional datasets (ED, FDBD, GOT and HD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

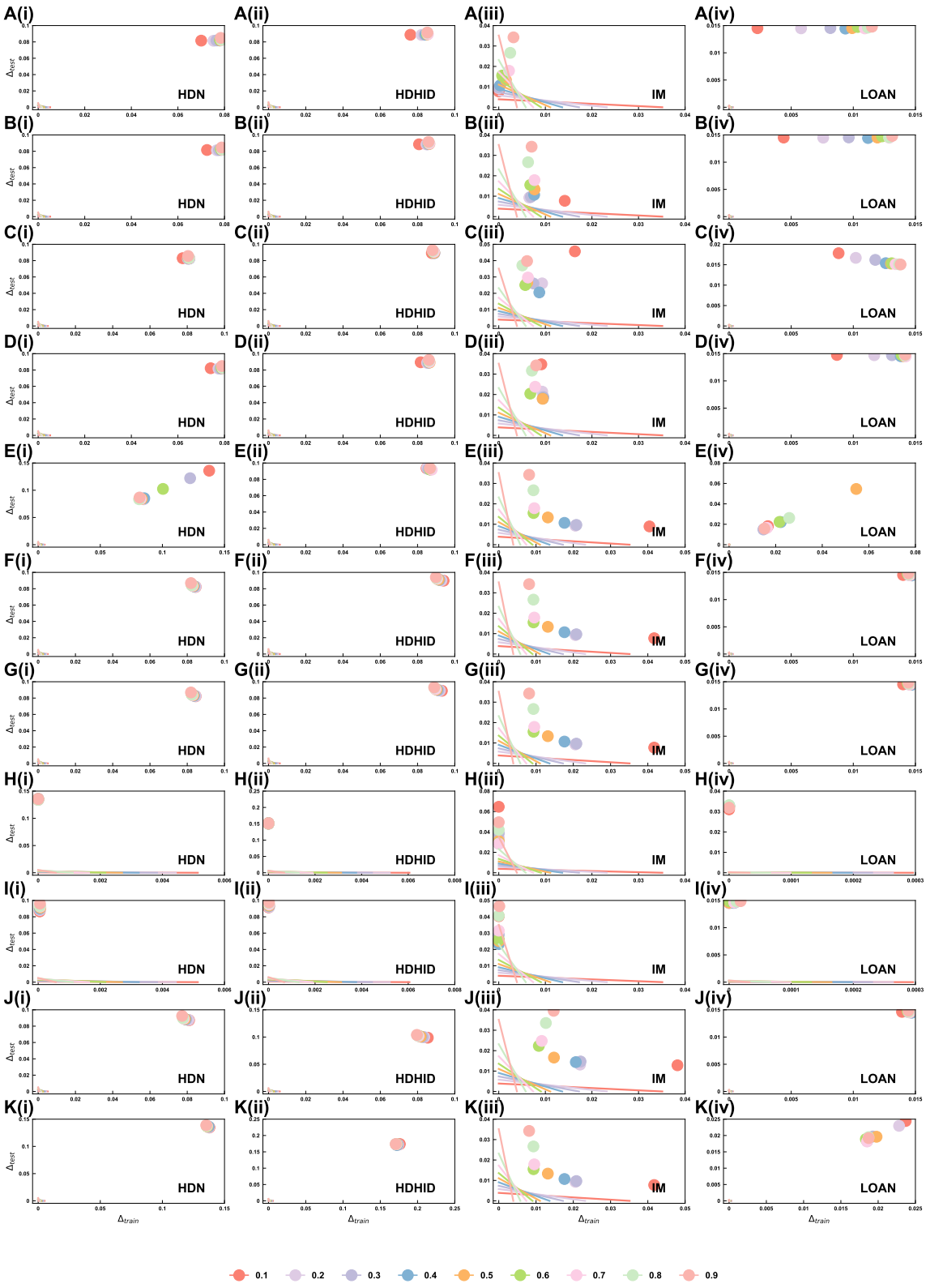

Figure S42. The loss errors on four additional datasets (HDN, HDHID, IM and LOAN) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

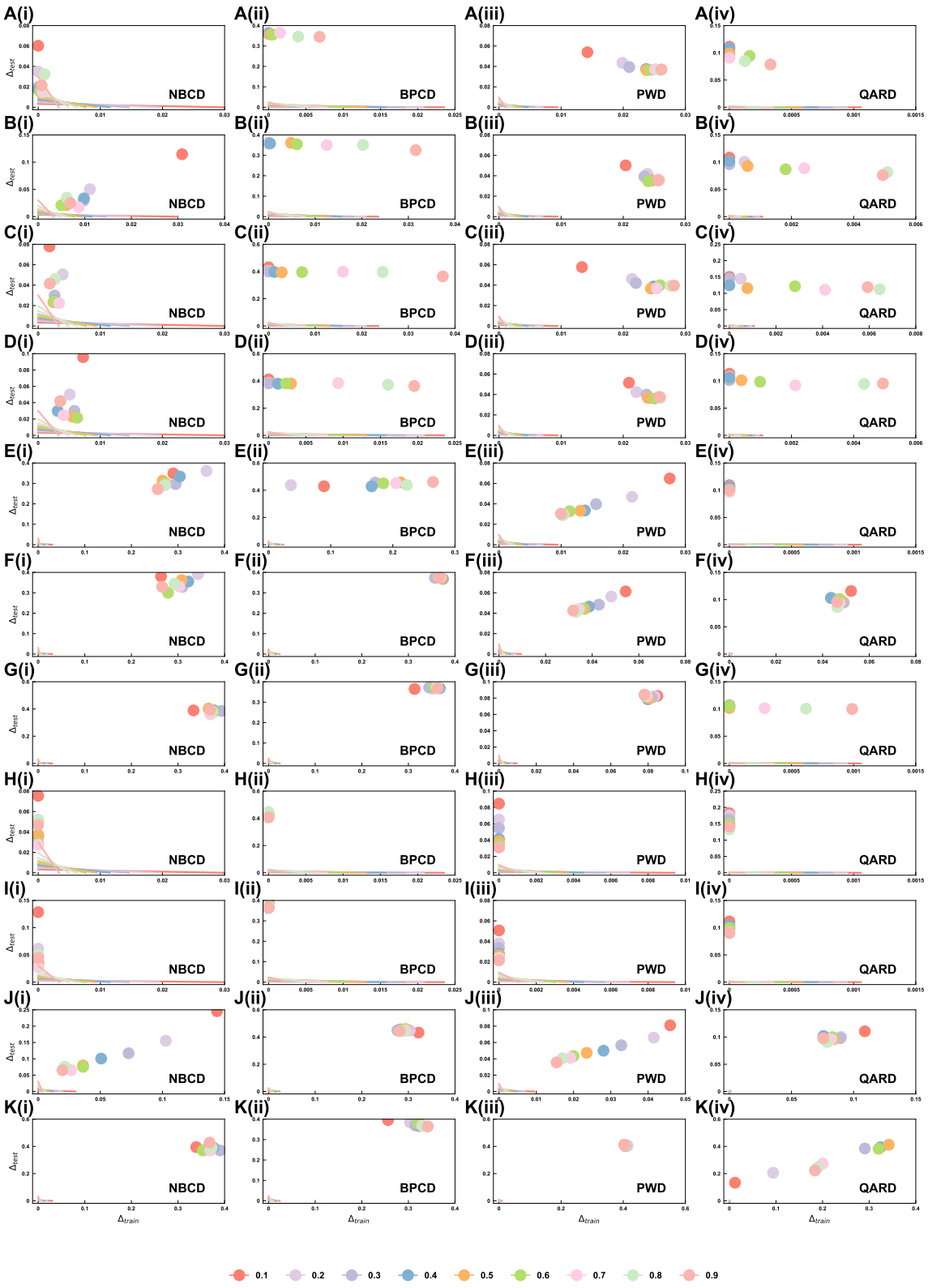

Figure S43. The loss errors on four additional datasets (NBCD, BPCD, PWD and QARD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

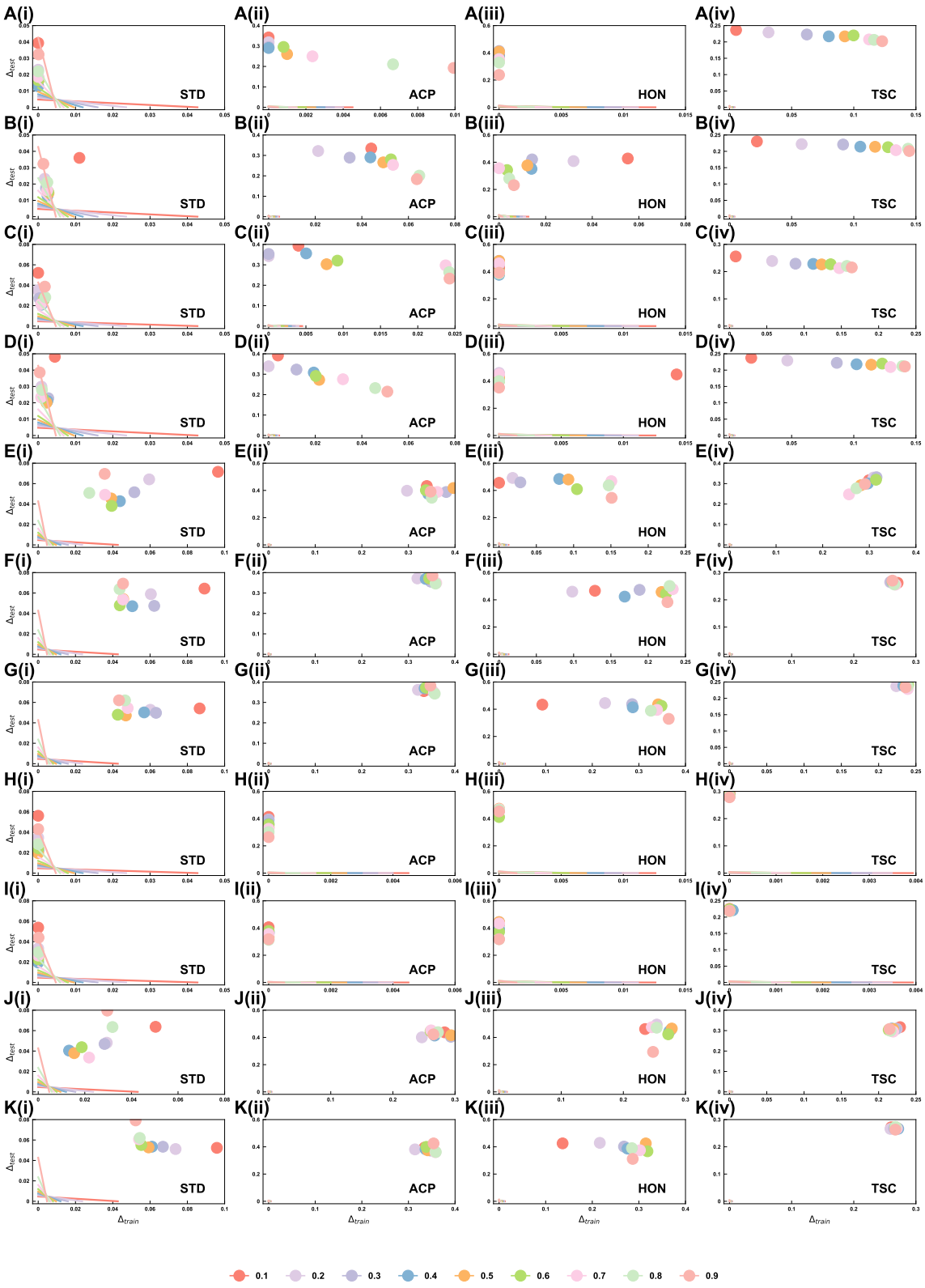

Figure S44. The loss errors on four additional datasets (STD, ACP, HON and TSC) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

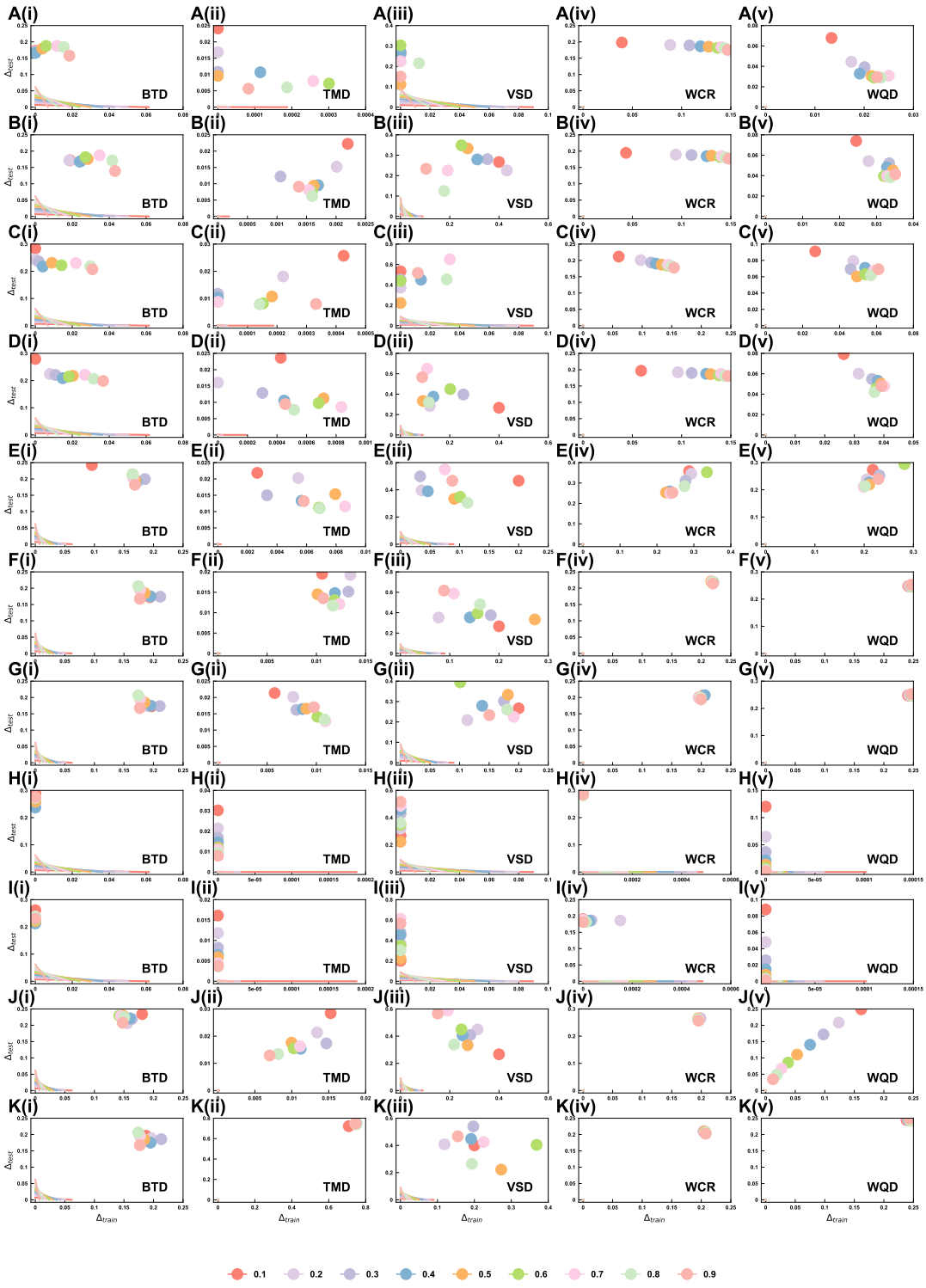

Figure S45. The loss errors on five additional datasets (BTD, TMD, VSD, WCR and WQD) in training (Δ^f_{train}) and test sets (Δ^f_{test}) of different binary classifiers, including XGBoost with four classical objectives (A-D), MLP (E), SVM (F), Logistic Regression (G), Decision Tree (H), Random Forest (I), KNN (J). Colorful dots and lines represent different $|\mathcal{S}_{train}|/|\mathcal{S}|$ ranging from 0.1 to 0.9.

VI. MINIMUM HINGE LOSS

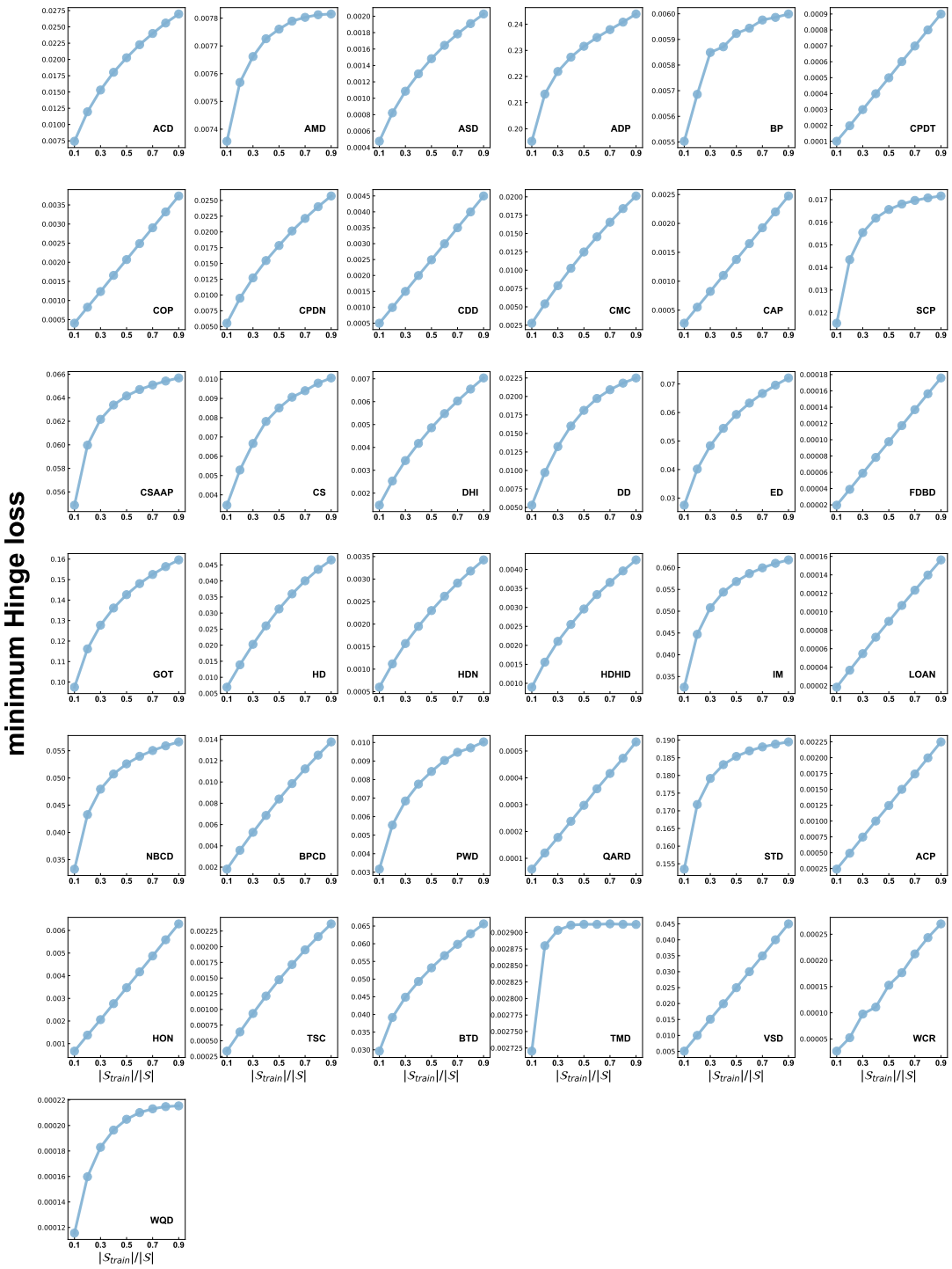

Figure S46. Minimum Hinge loss for 37 additional datasets under different random divisions In these panels, dots correspond to numerical results derived from the data divisions, while lines represent the theoretical predictions adjusted for the respective division ratios (see Eqs. 54 and 56).

25

VII. UPPER BOUND OF ACCURACY

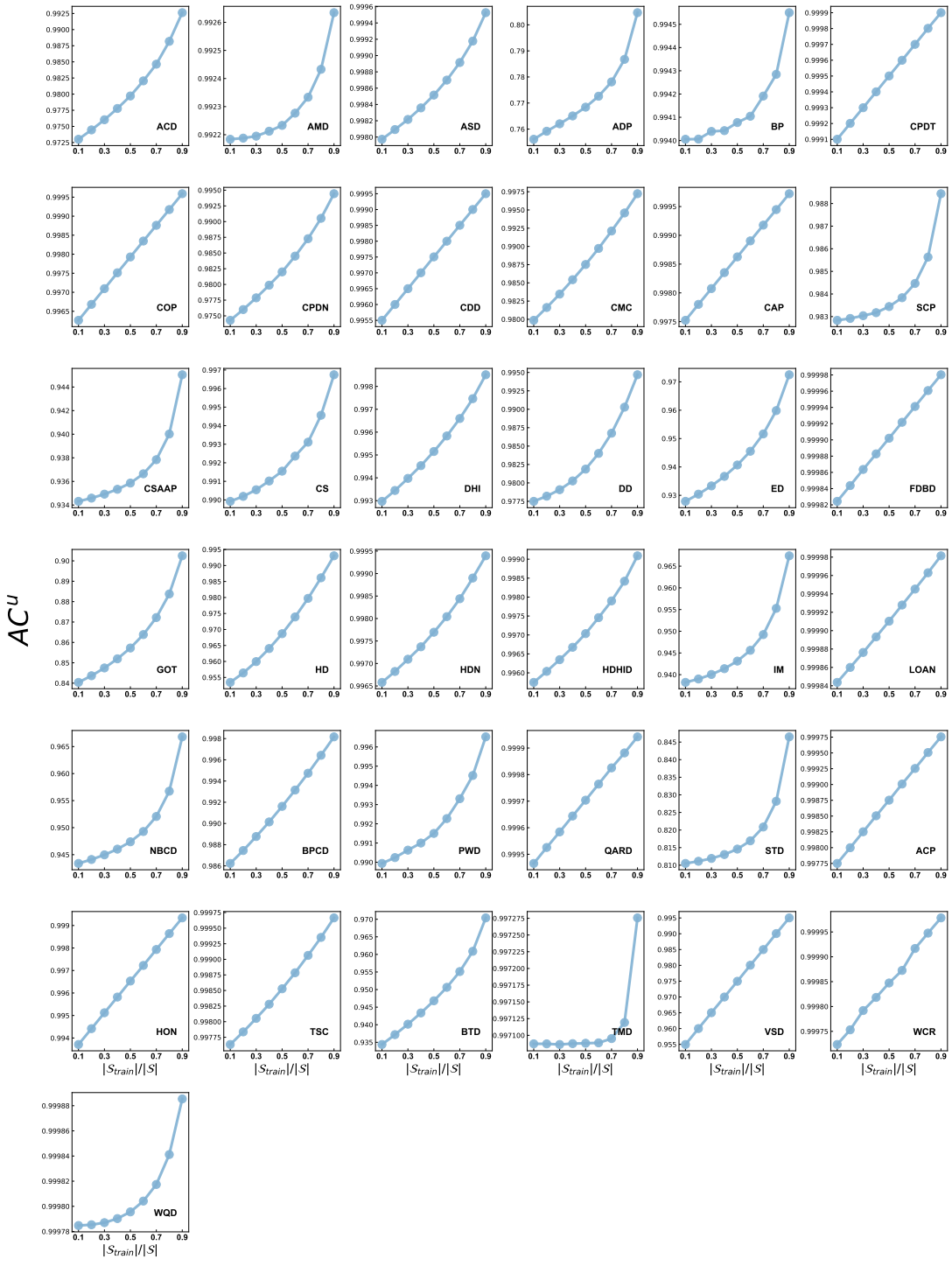

Figure S47. Upper bound of accuracy (AC^u) for 37 additional datasets under different random divisions In these panels, dots correspond to numerical results derived from the data divisions, while lines represent the theoretical predictions adjusted for the respective division ratios (see Eqs. 54 and 56).

26

VIII. ANTICIPATED OPTIMAL ERRORS

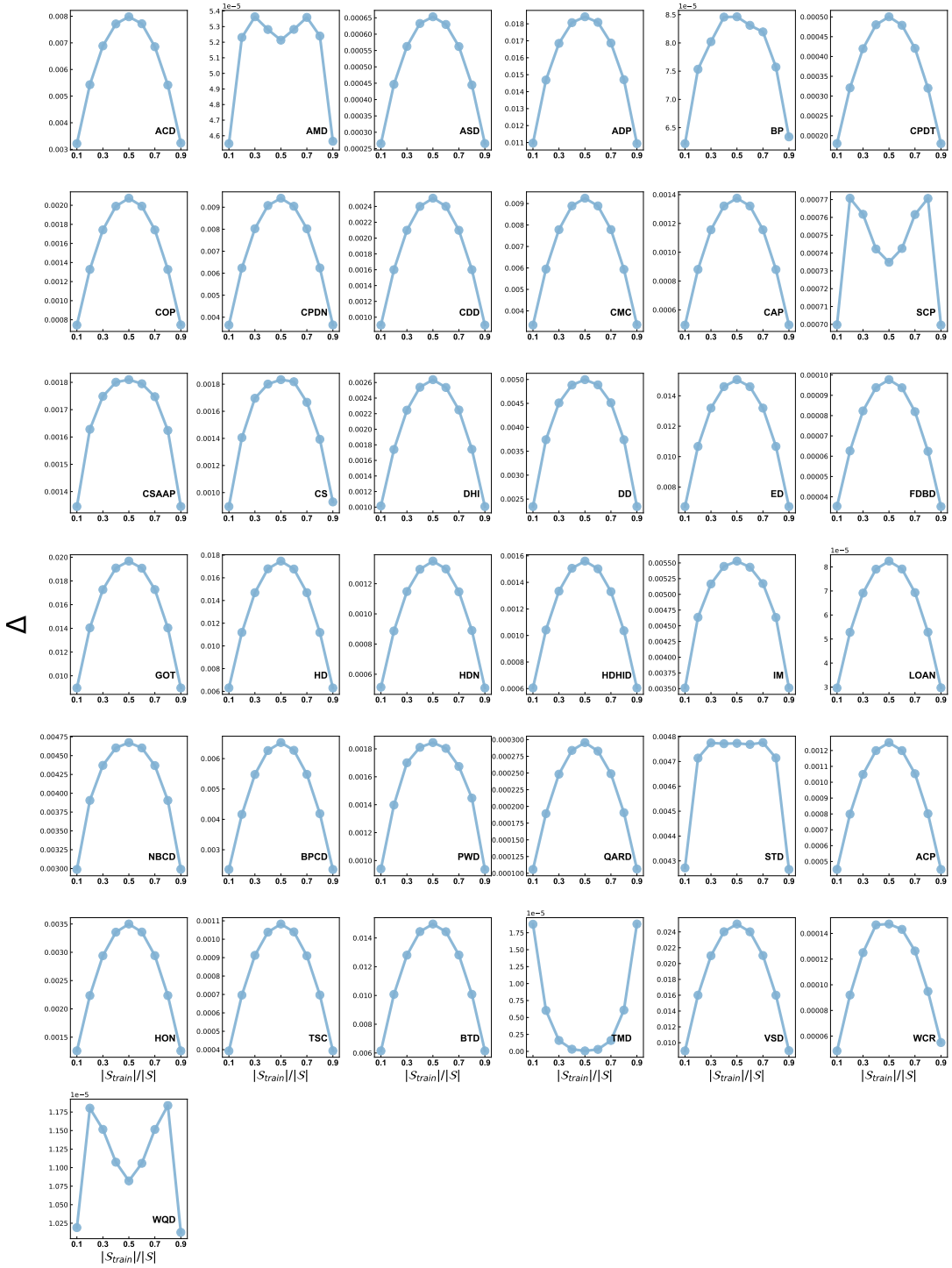

Figure S48. Anticipated optimal errors (Δ) for 37 additional datasets under different random divisions In these panels, dots correspond to numerical results derived from the data divisions, while lines represent the theoretical predictions adjusted for the respective division ratios (see Eqs. 54 and 56).

IX. THE $AR^u_{k_0}$ IN FEATURE SELECTION

Figure S49. The $AR^u_{k_0}$ versus the optimal k_0 feature subset in feature selection (blue lines and dots) for 37 additional datasets. After we selected the optimal k_0 feature subset, we would use the feature extraction skill (LDA) to create new extracted features and add them into the original k_0 feature one by one (see red lines and dots).

28

X. THE $D_S^{k_0}$ IN FEATURE SELECTION

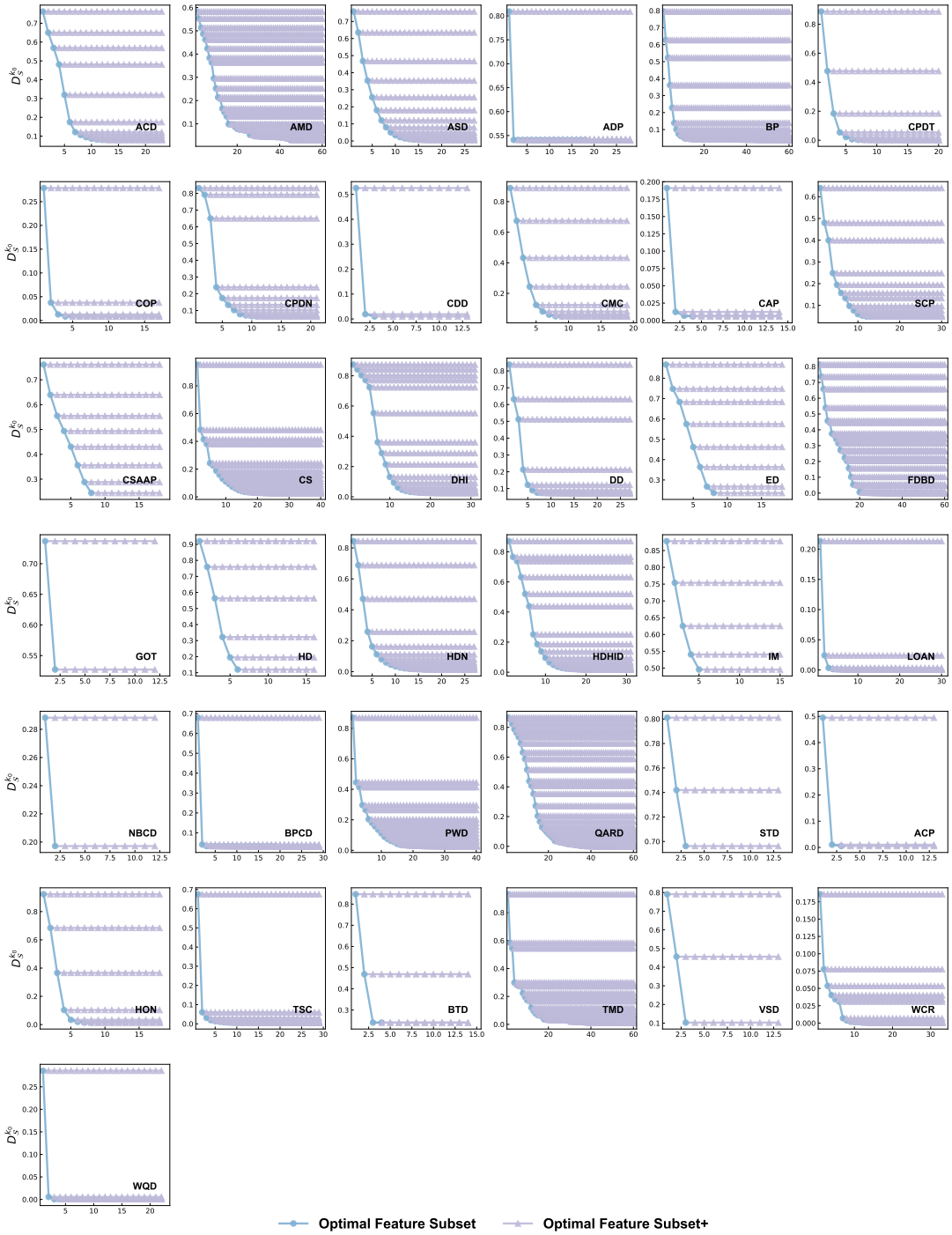

Figure S50. The $D_S^{k_0}$ versus the optimal k_0 feature subset in feature selection (blue lines and dots) for 37 additional datasets. After we selected the optimal k_0 feature subset, we would use the feature extraction skill (LDA) to create new extracted features and add them into the original k_0 feature one by one (see red lines and dots)