

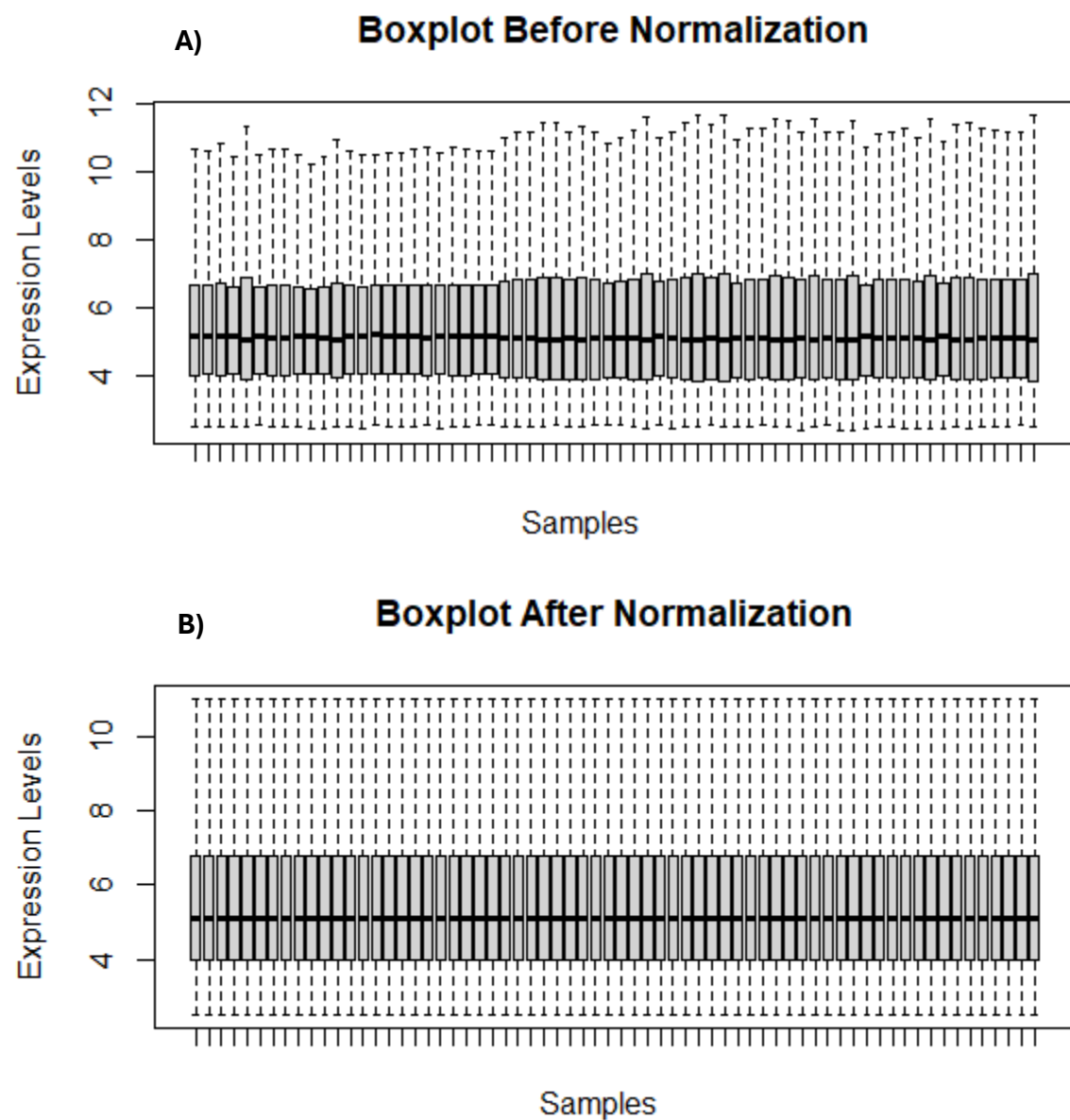


Fig. S 1 Boxplots demonstrating the effect of normalization and batch correction on gene expression data from MDS patients. (A) Raw expression data showing technical variability across samples (uneven medians and interquartile ranges). (B) Processed data after RMA background correction, $\log_2$ transformation, and quantile normalization, revealing aligned distributions essential for downstream analysis.

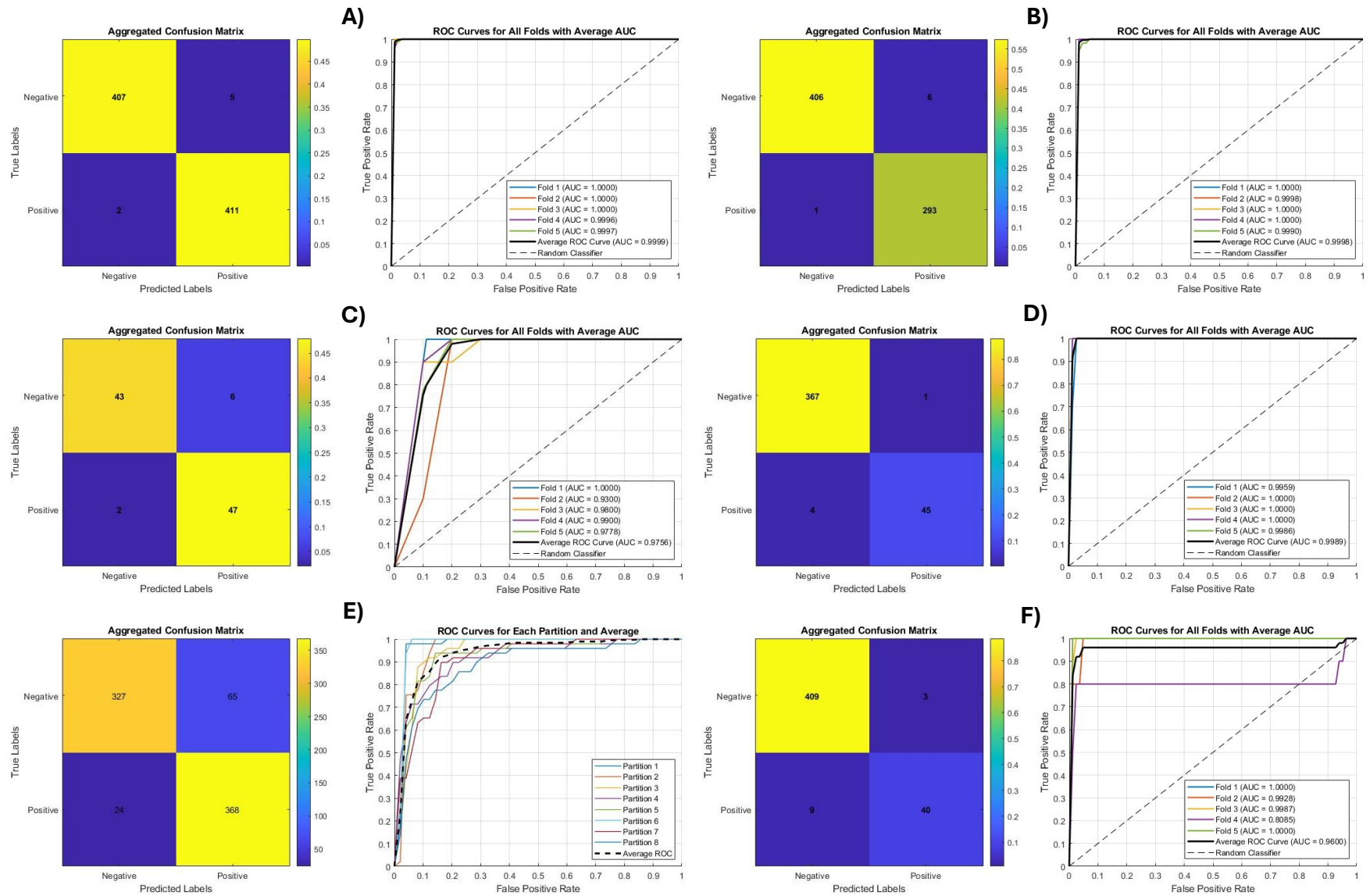


Fig. S 2 Confusion matrices and ROC curves for SVM-RBF classifier using top 10 hub genes under six resampling strategies. a) ADASYN, B) SMOTE, Cluster C) CC, D) ENN, E) Data Partitioning, and F) No Resampling.

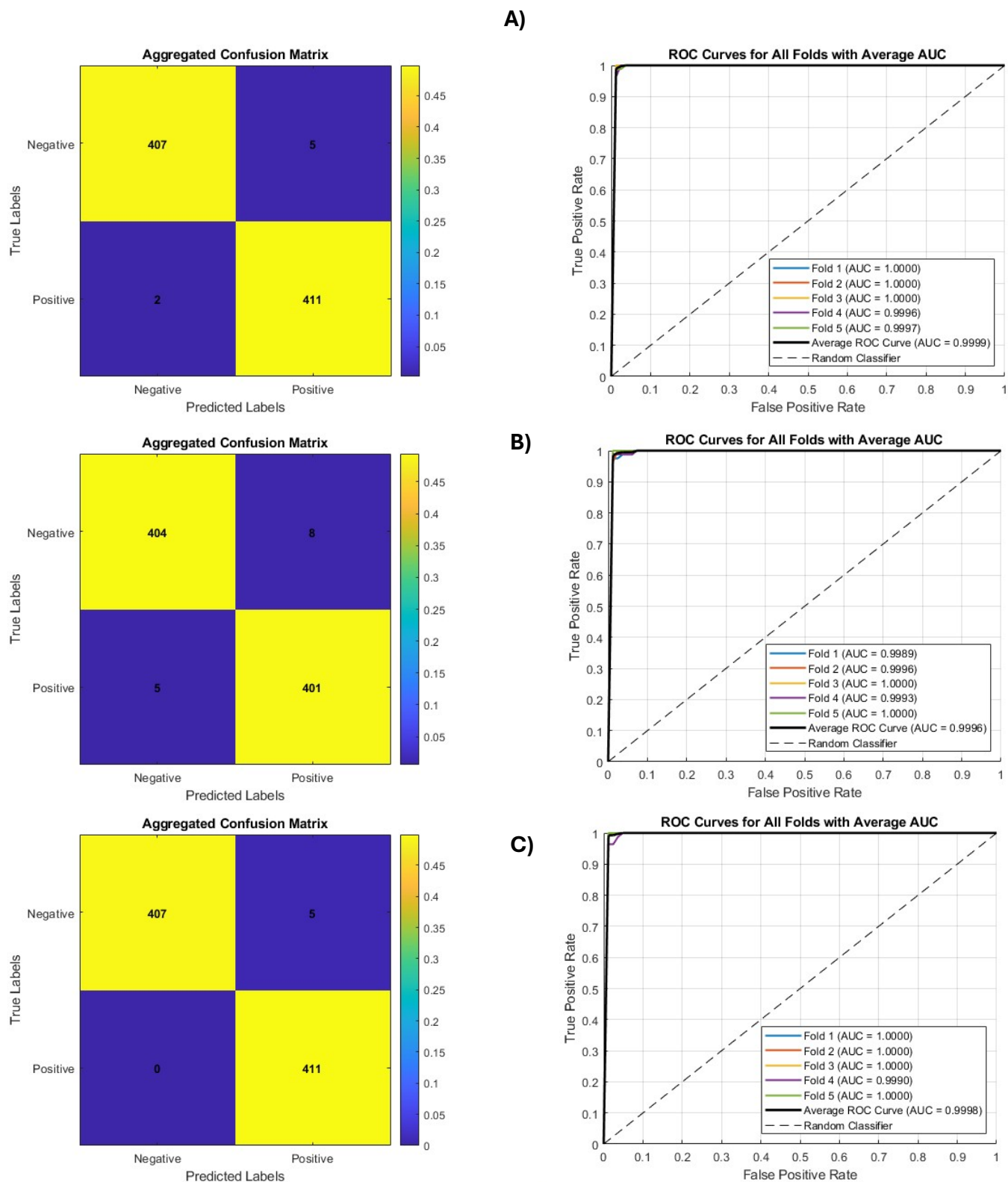
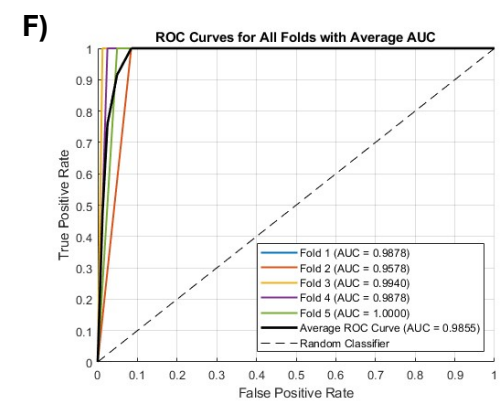
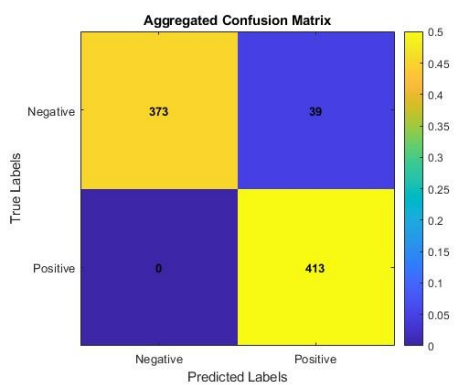
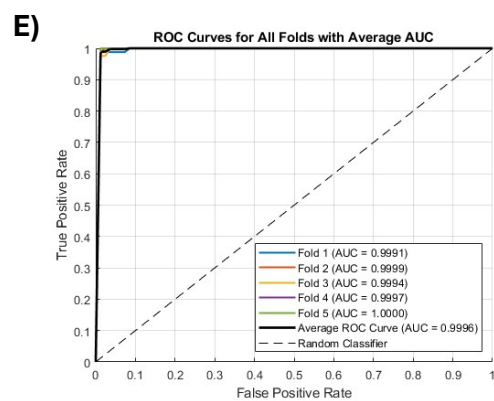
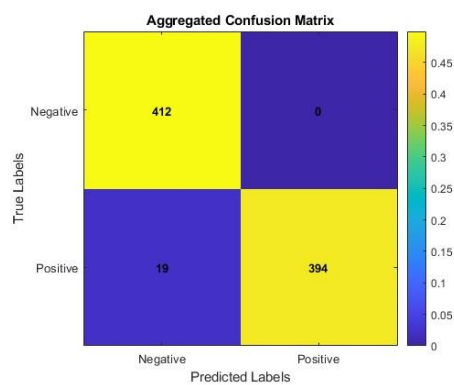
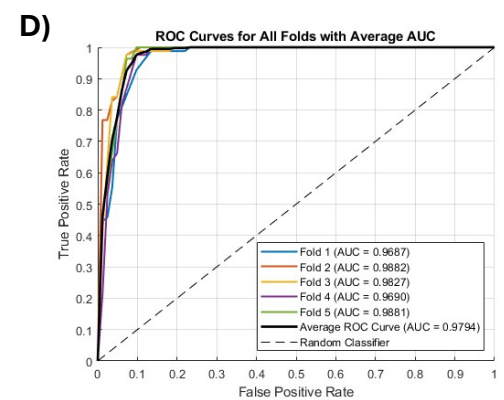
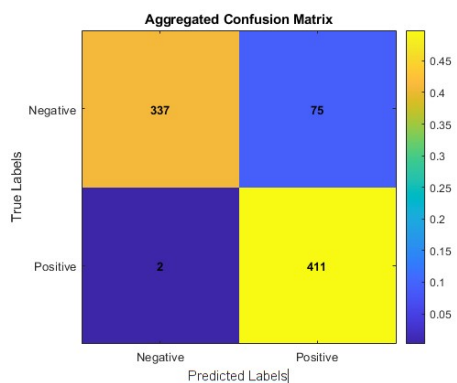
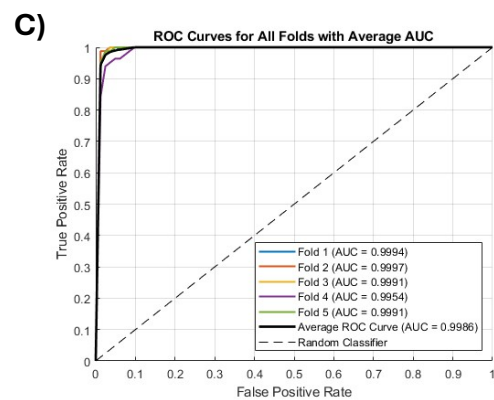
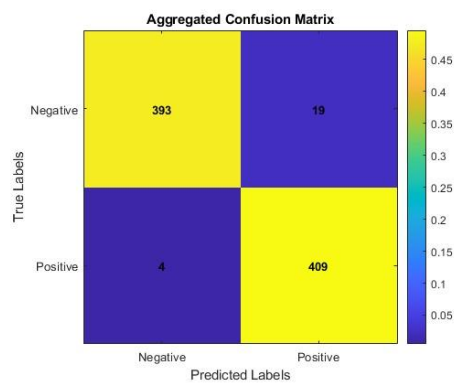
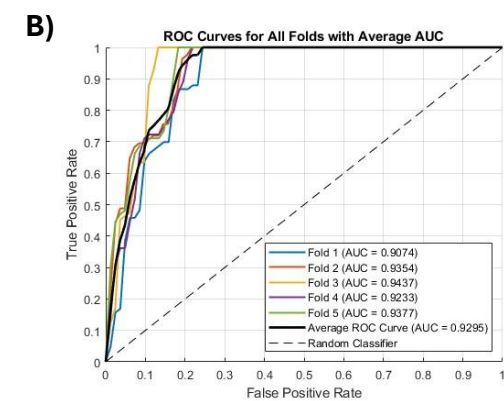
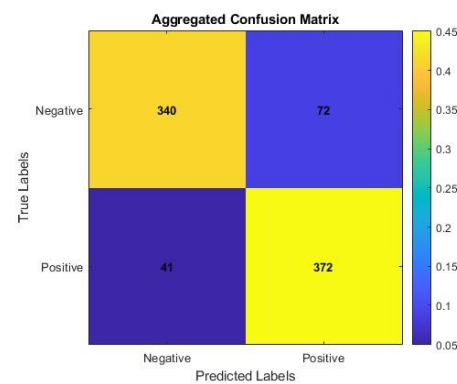
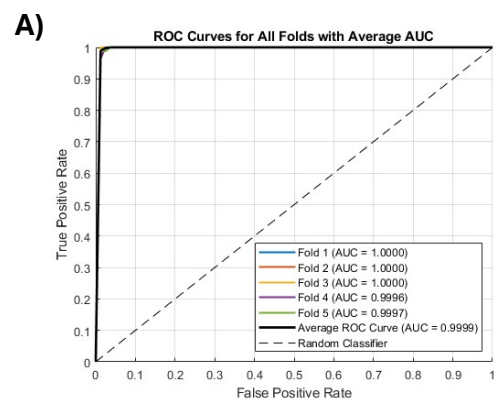
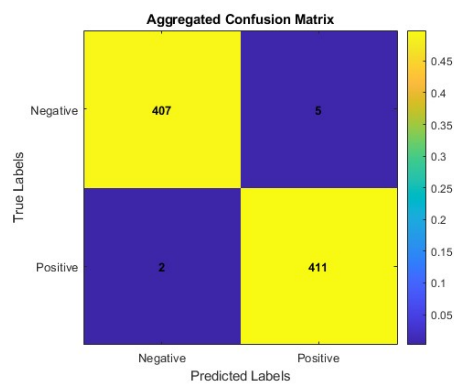
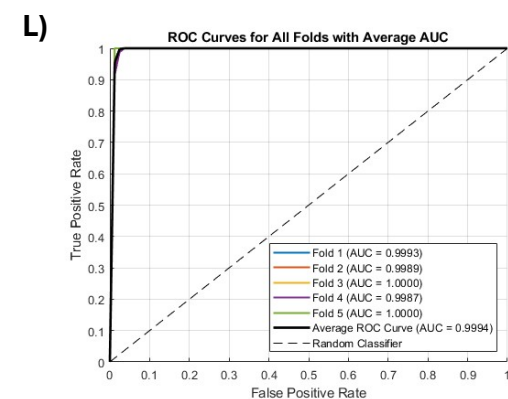
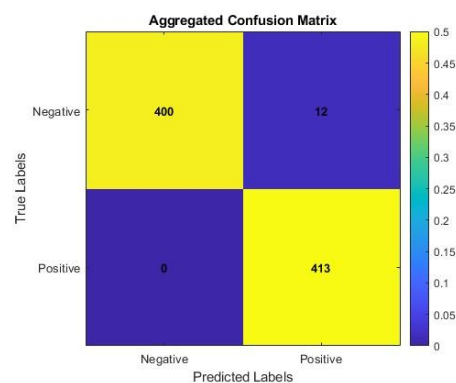
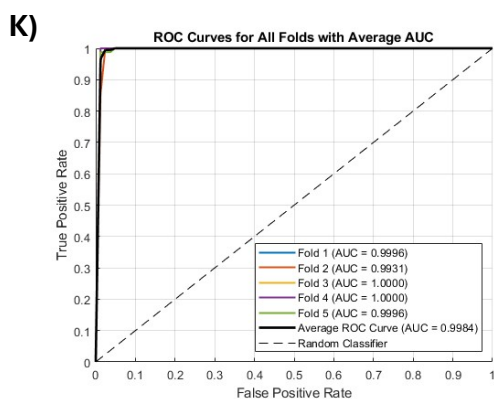
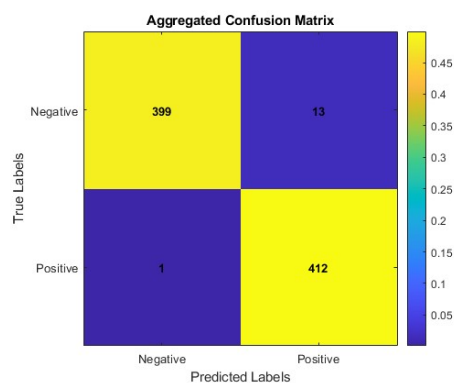
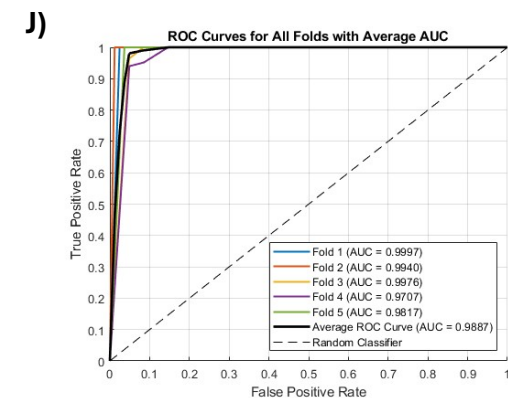
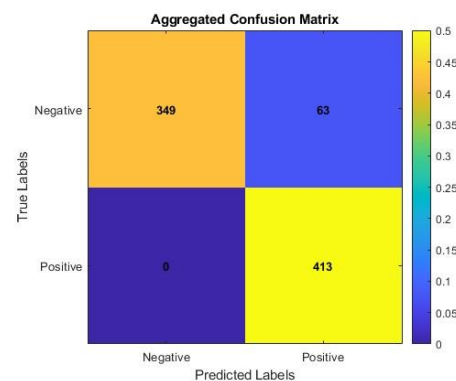
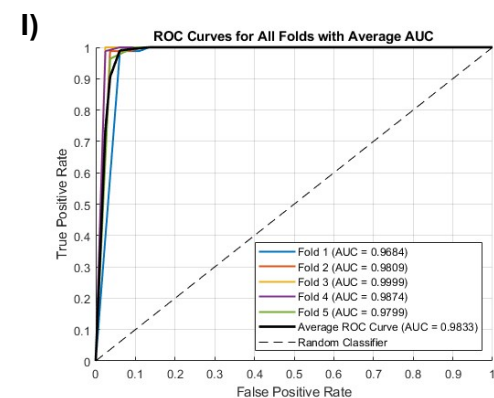
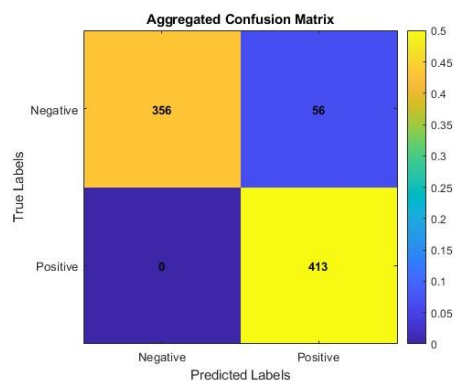
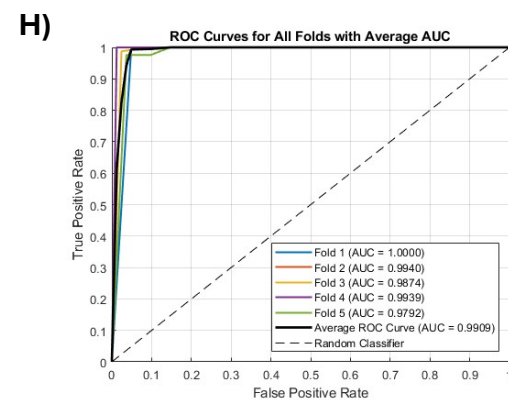
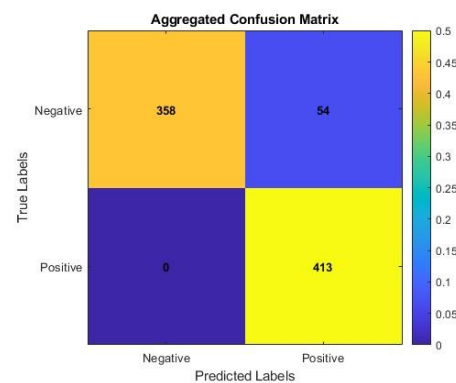
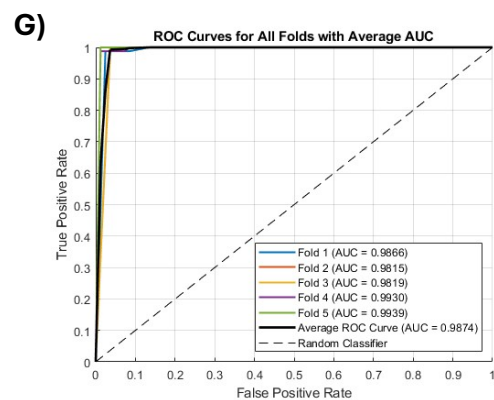
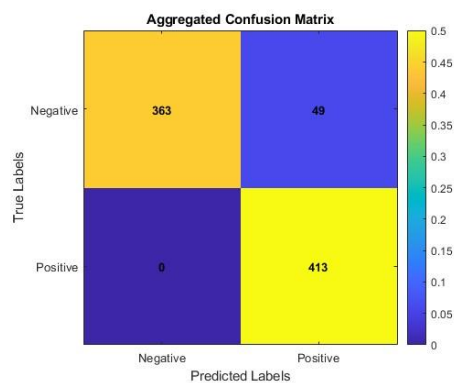


Fig. S 3 Comparative confusion matrices and ROC curves for performance analysis of SVM-RBF classifiers using different hub gene subsets (10, 15, and 20 genes, respectively).





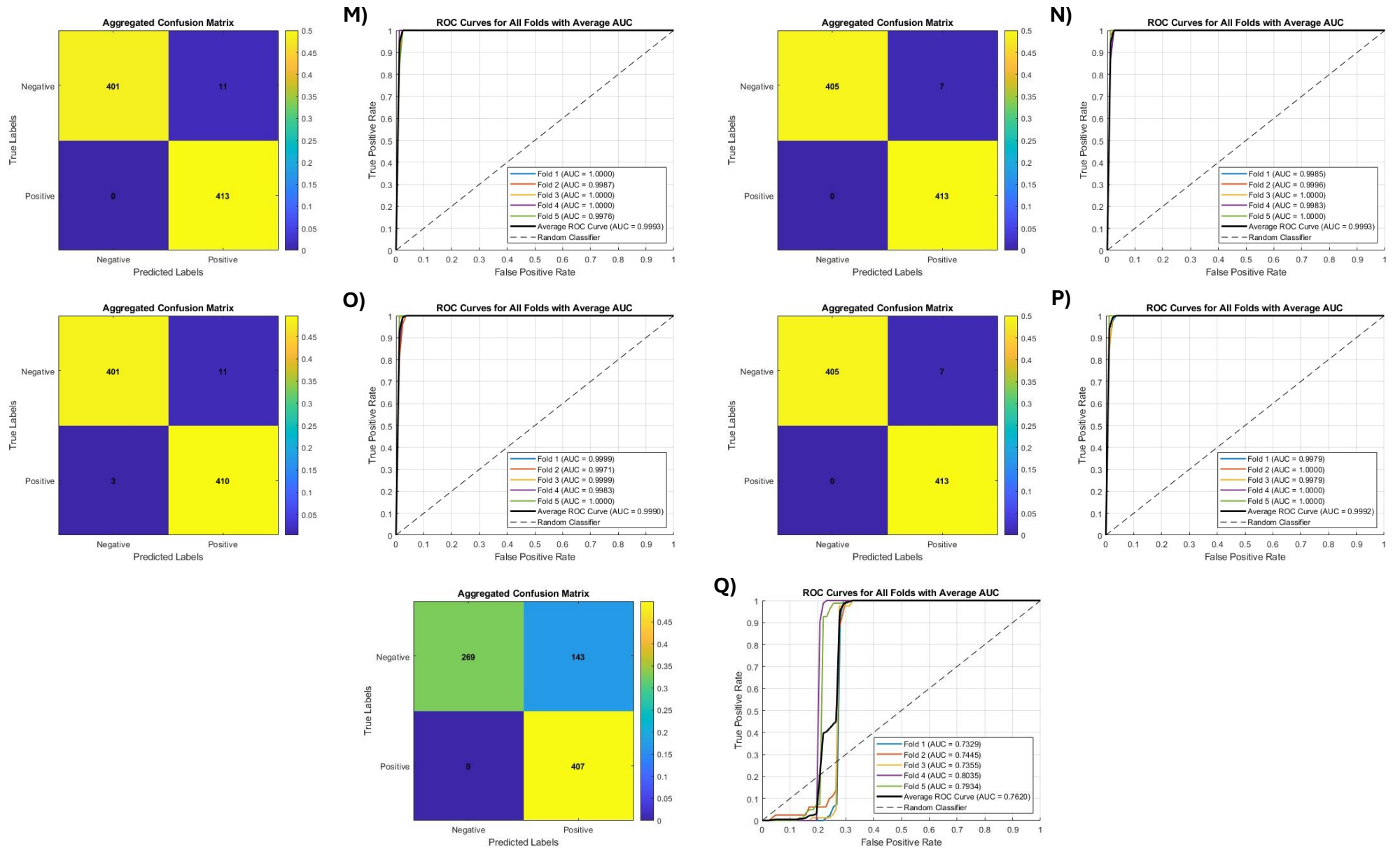


Fig. S4 Confusion matrices (A–Q) illustrating classification performance of machine learning models for MDS classification using the top 10 hub genes, corresponding to the results in Table 5. (A) SVM, (B) NB, (C) DT, (D) AdaBoost, (E) QDA, (F) KNN = 3, (G) KNN = 5, (H) KNN = 7, (I) KNN = 9, (J) KNN = 11, (K) GAM, (L) RF = 50, (M) RF = 100, (N) RF = 150, (O) RF = 200, (P) RF = 250, (Q) ELM.