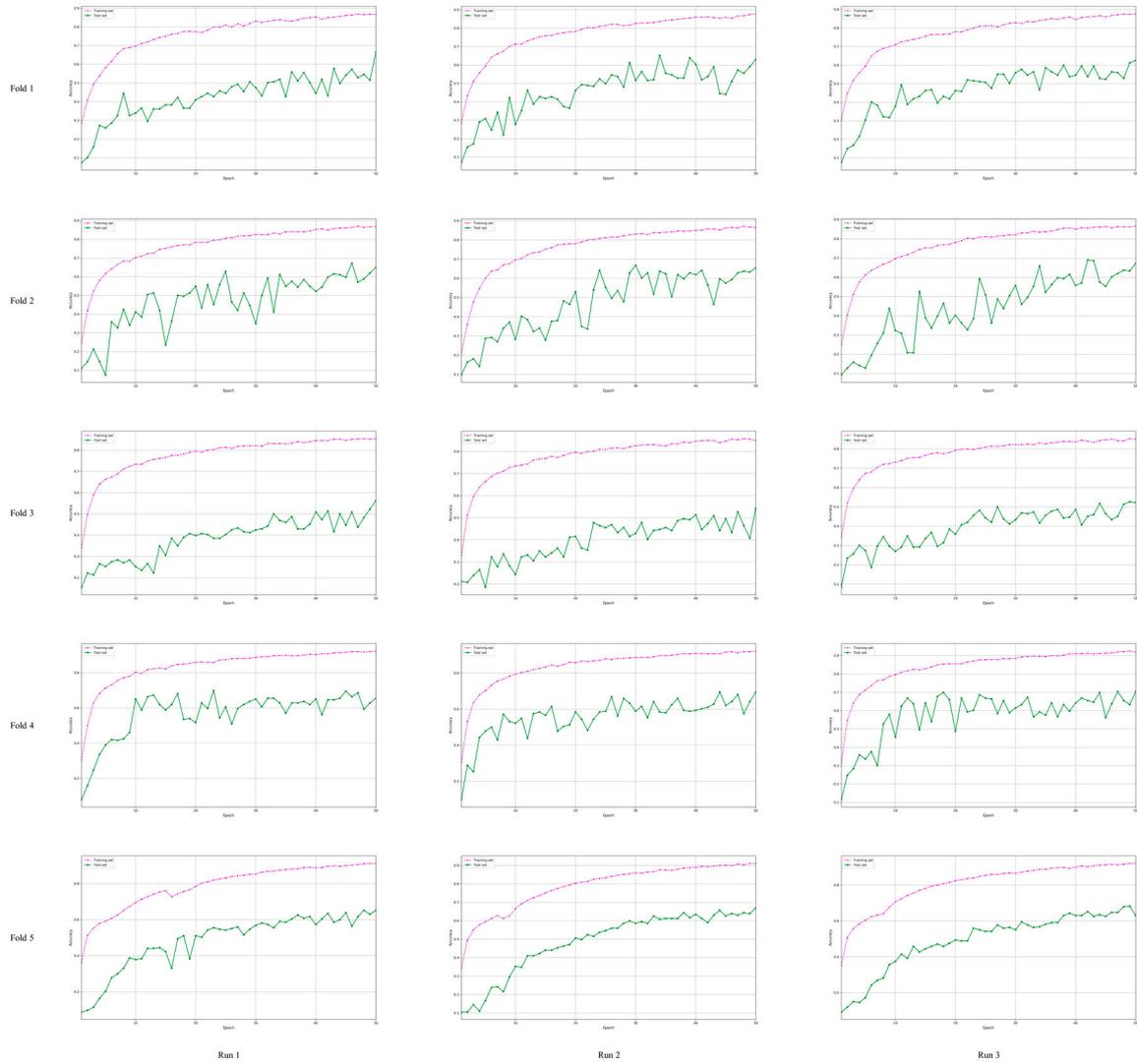
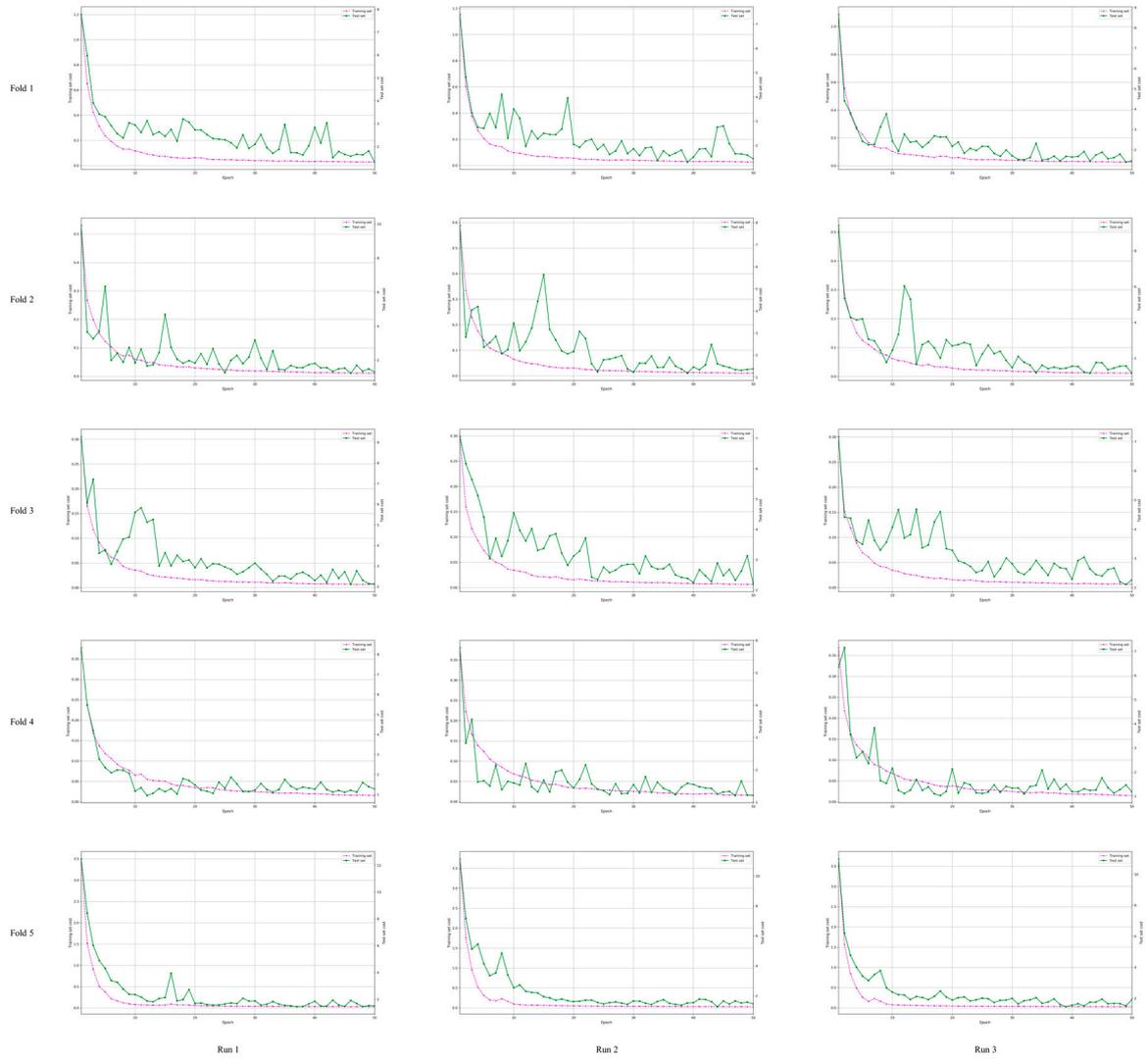


## Deep Learning-Based Morphological Classification of Human Sperm Heads

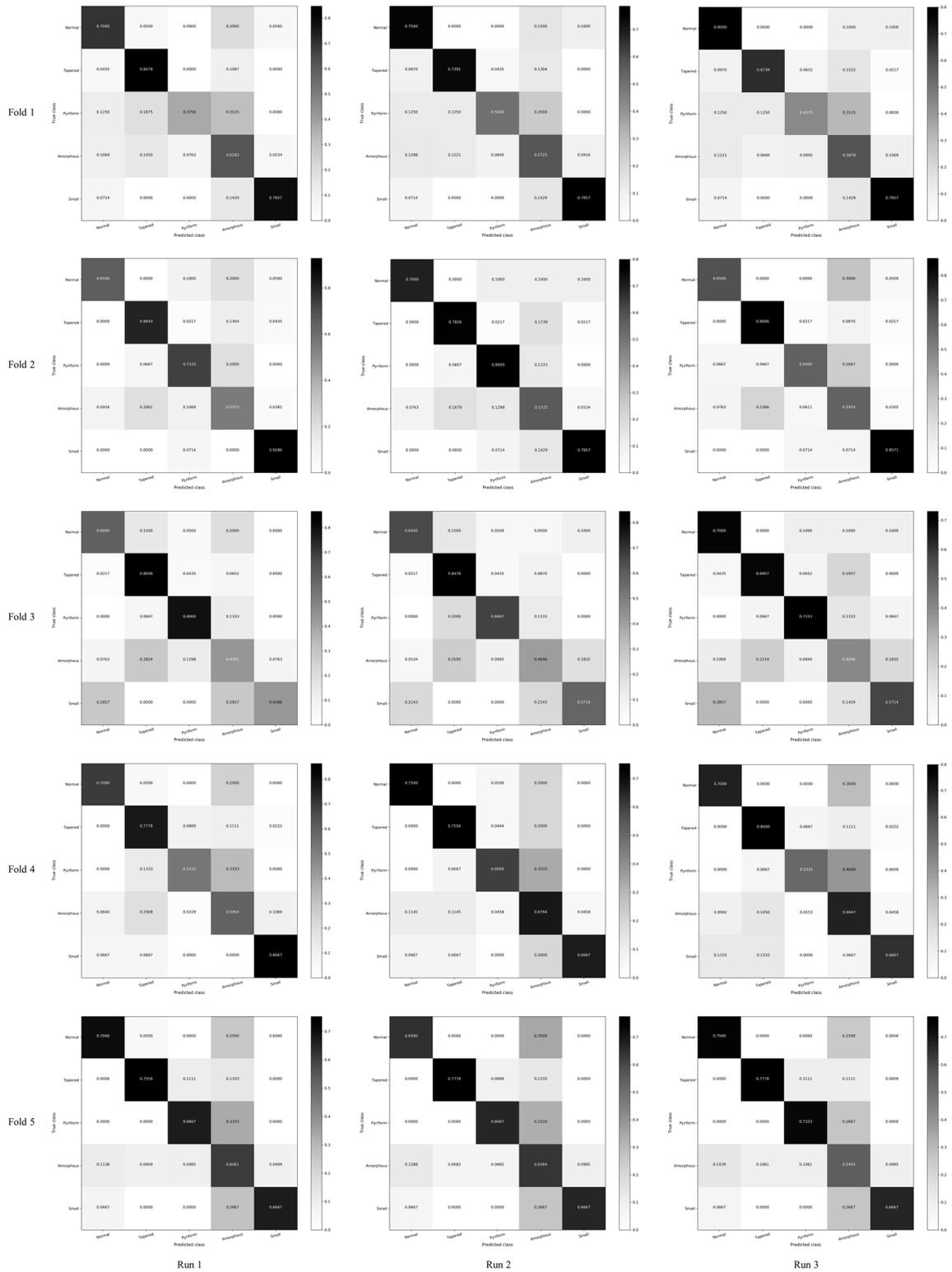
Imran Iqbal <sup>1</sup>, Ghulam Mustafa <sup>2</sup> and Jinwen Ma <sup>1,\*</sup>



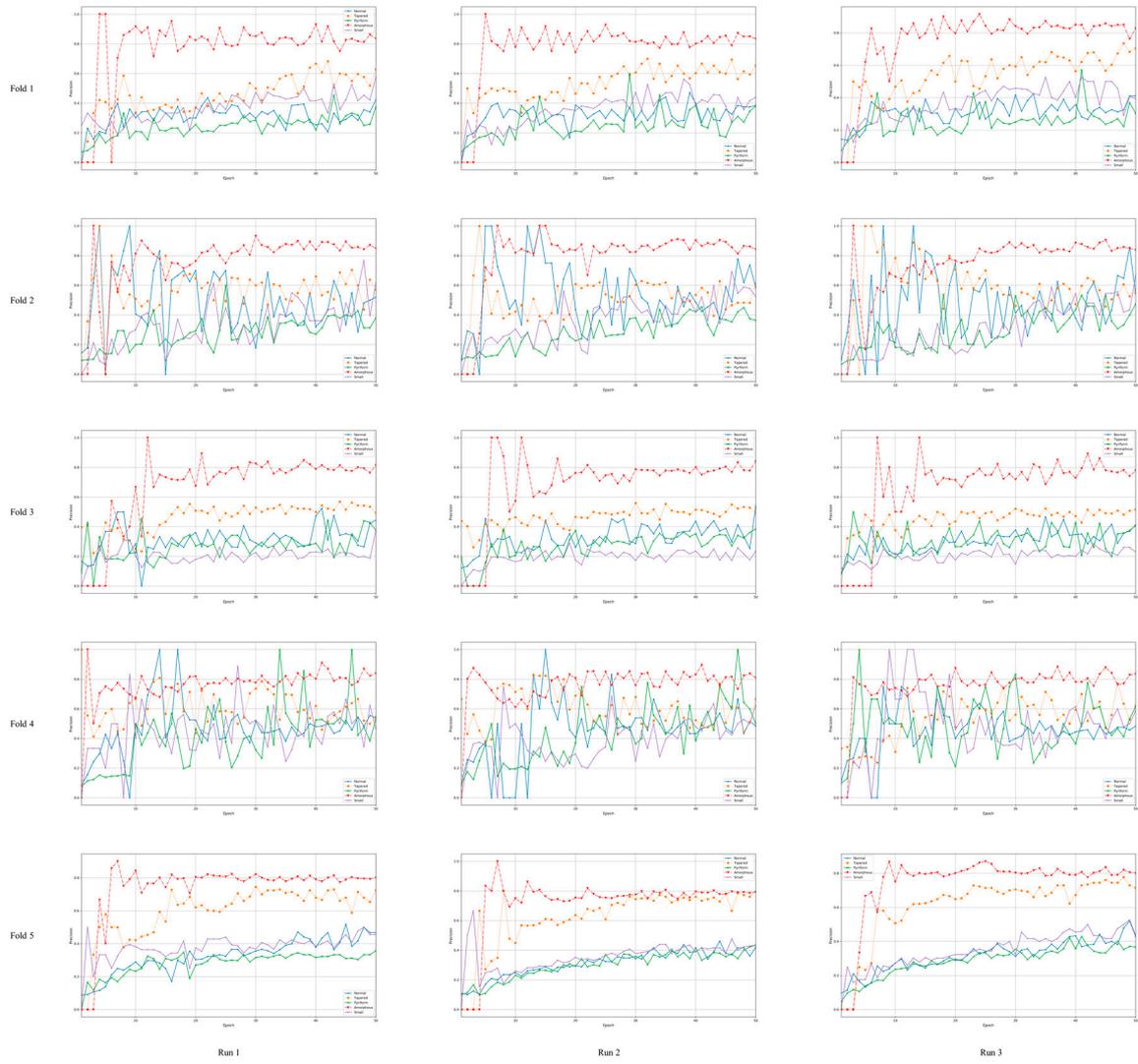
**Figure S1.** Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 classification accuracy curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



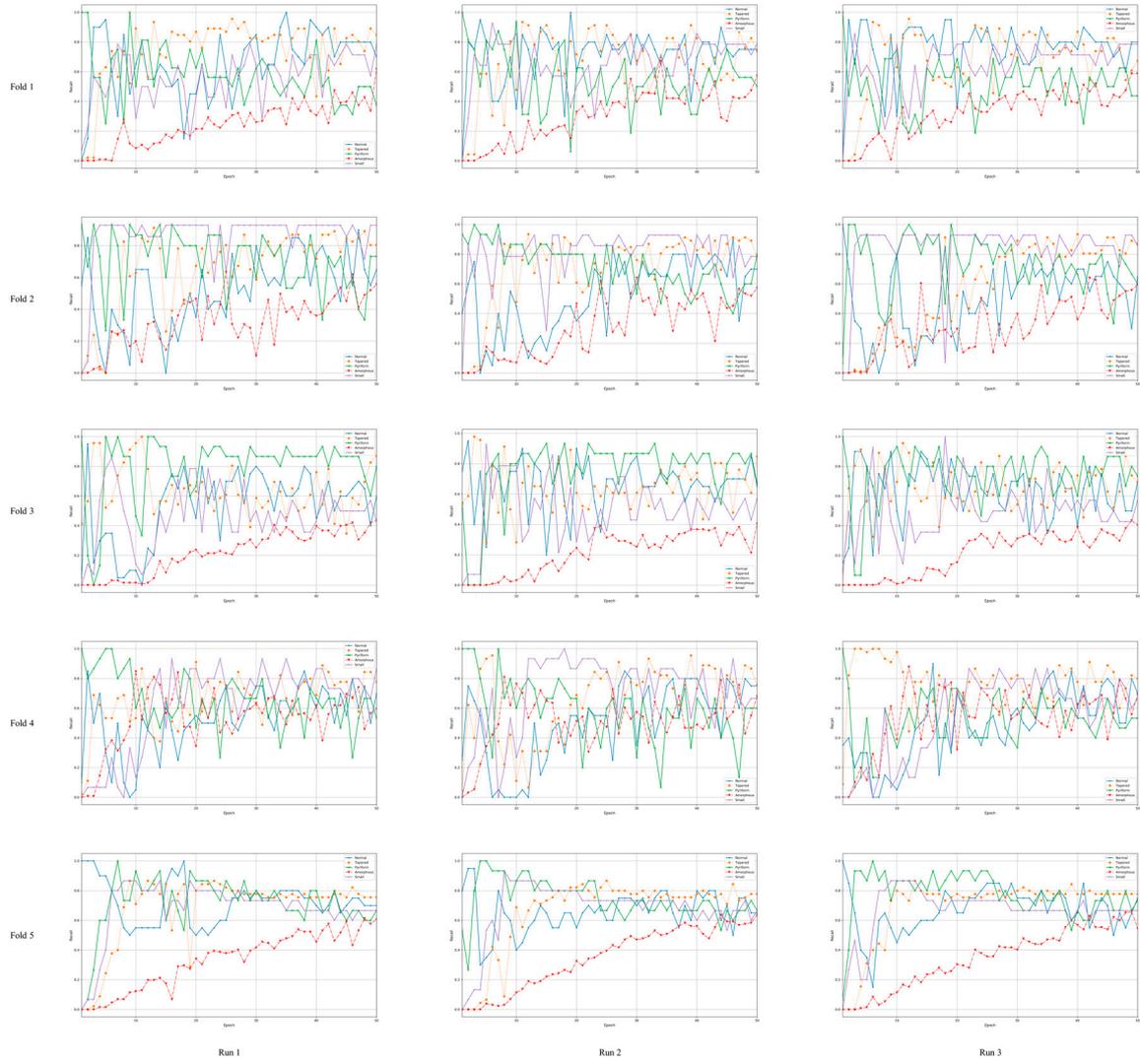
**Figure S2.** Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 cost curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



**Figure S3.** Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 confusion matrixes on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



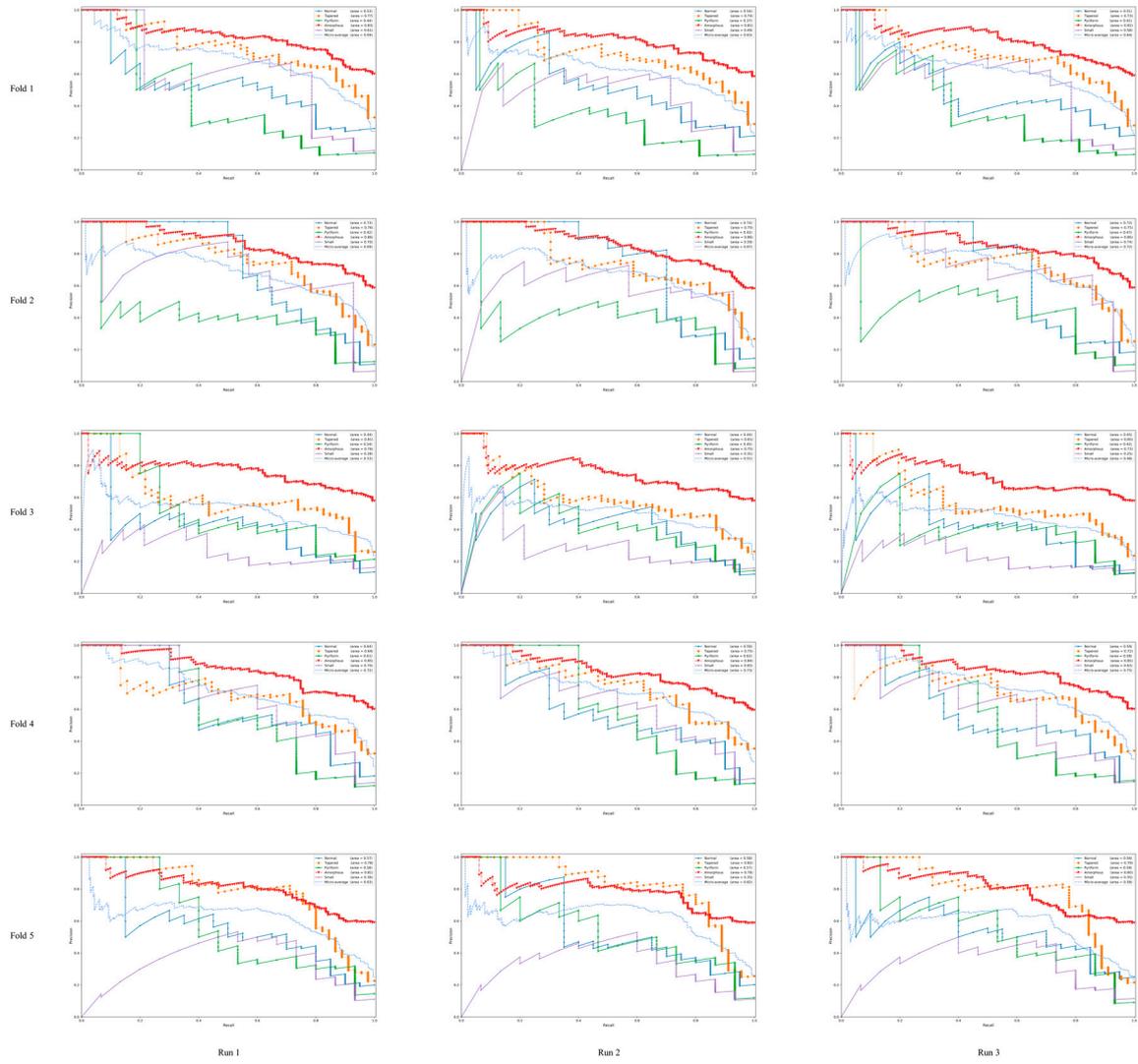
**Figure S4.** Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 precision curves of each class with the number of epochs during the training on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



**Figure S5.** Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 recall curves of each class with the number of epochs during the training on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.

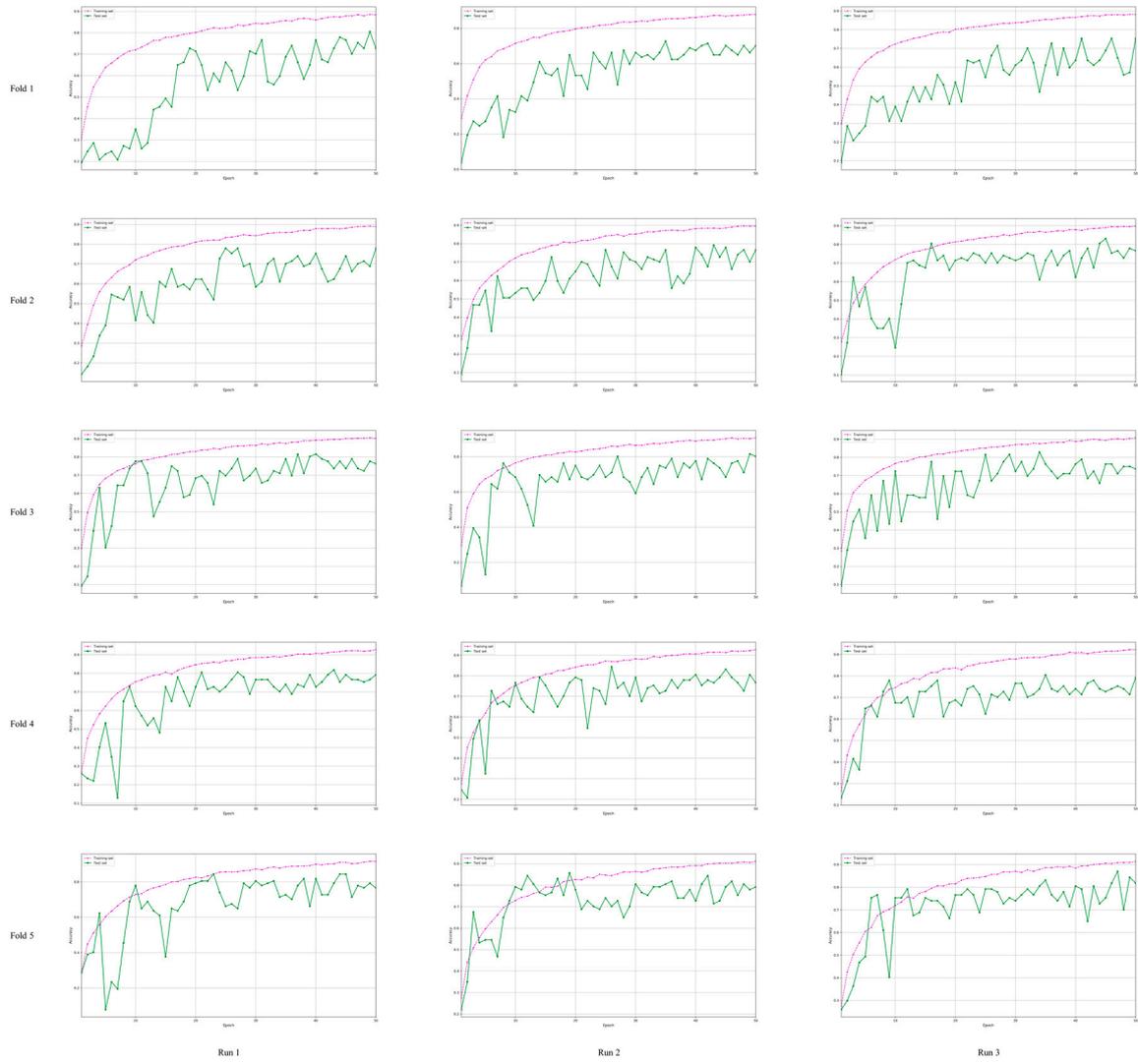


**Figure S6.** Detailed experimental results of the proposed model through 50 epochs in the partial agreement setting of the SCIAN dataset, where 15 F1-score curves of each class with the number of epochs during the training on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.

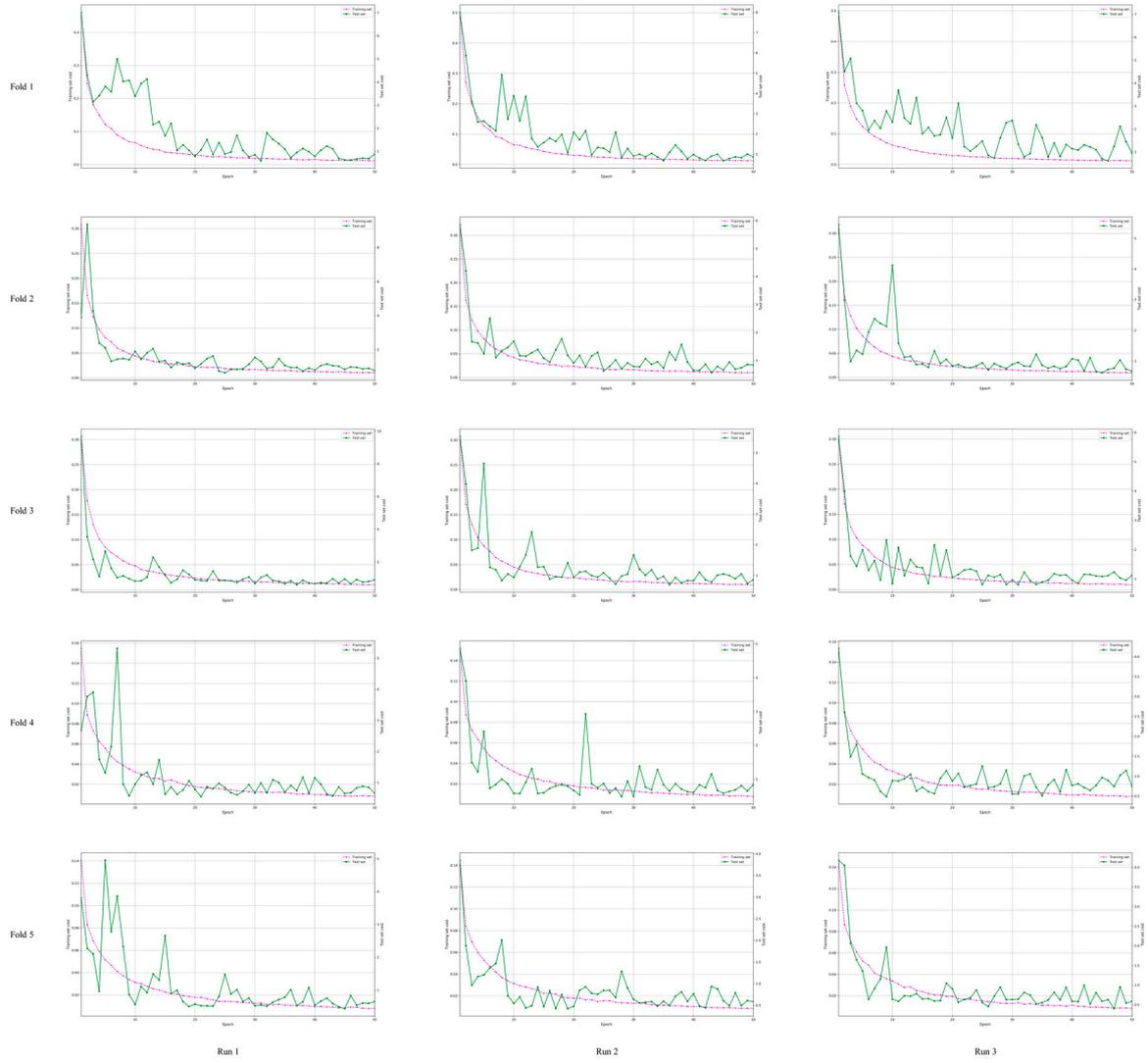


**Figure S7.** Detailed experimental results of proposed model in the partial agreement setting of the SCIAN dataset, where 15 precision-recall curves of each class and their micro-averaging precision-recall curve on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated. A high area under the curve signifies the high precision as well as high recall.

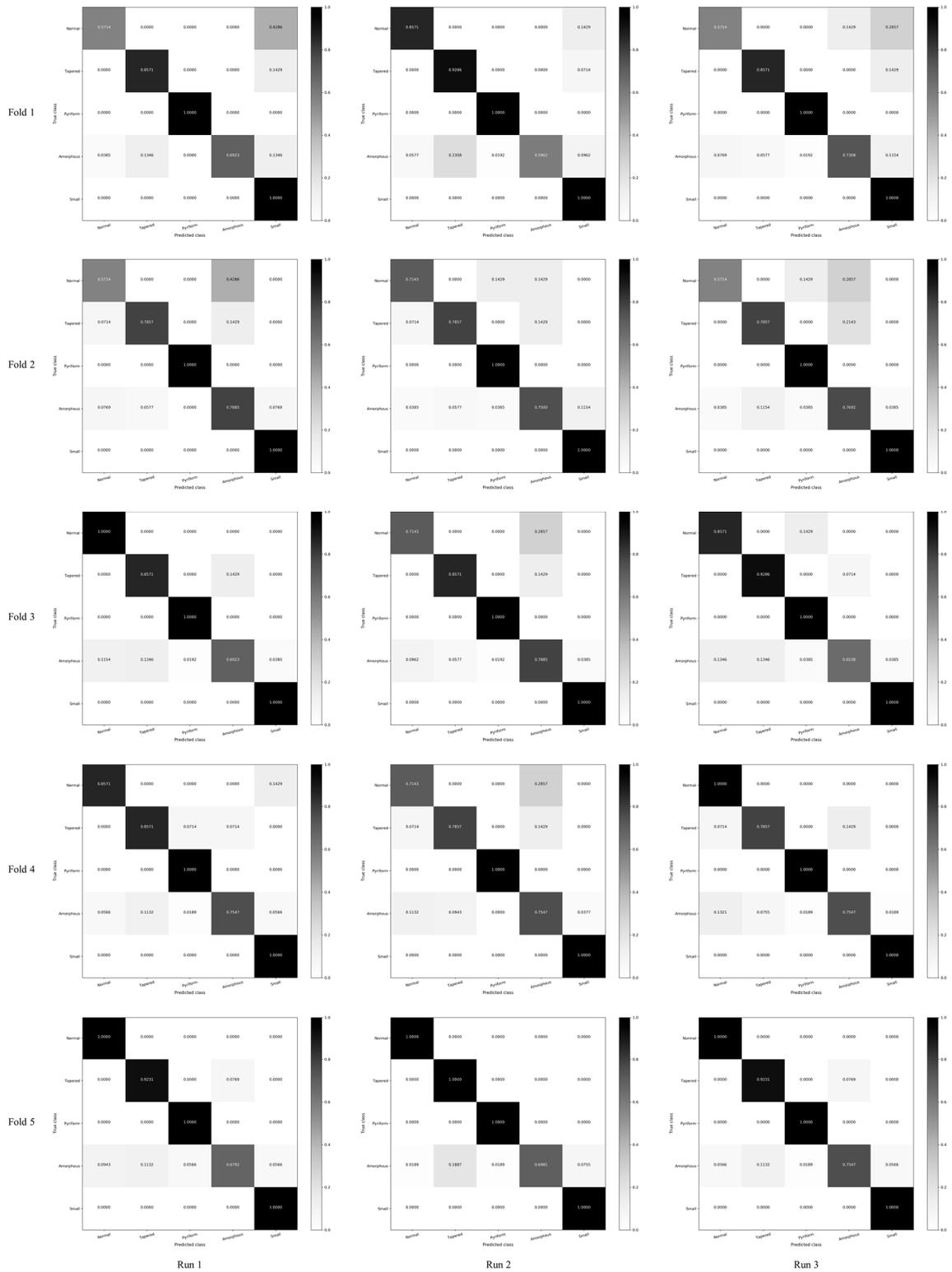




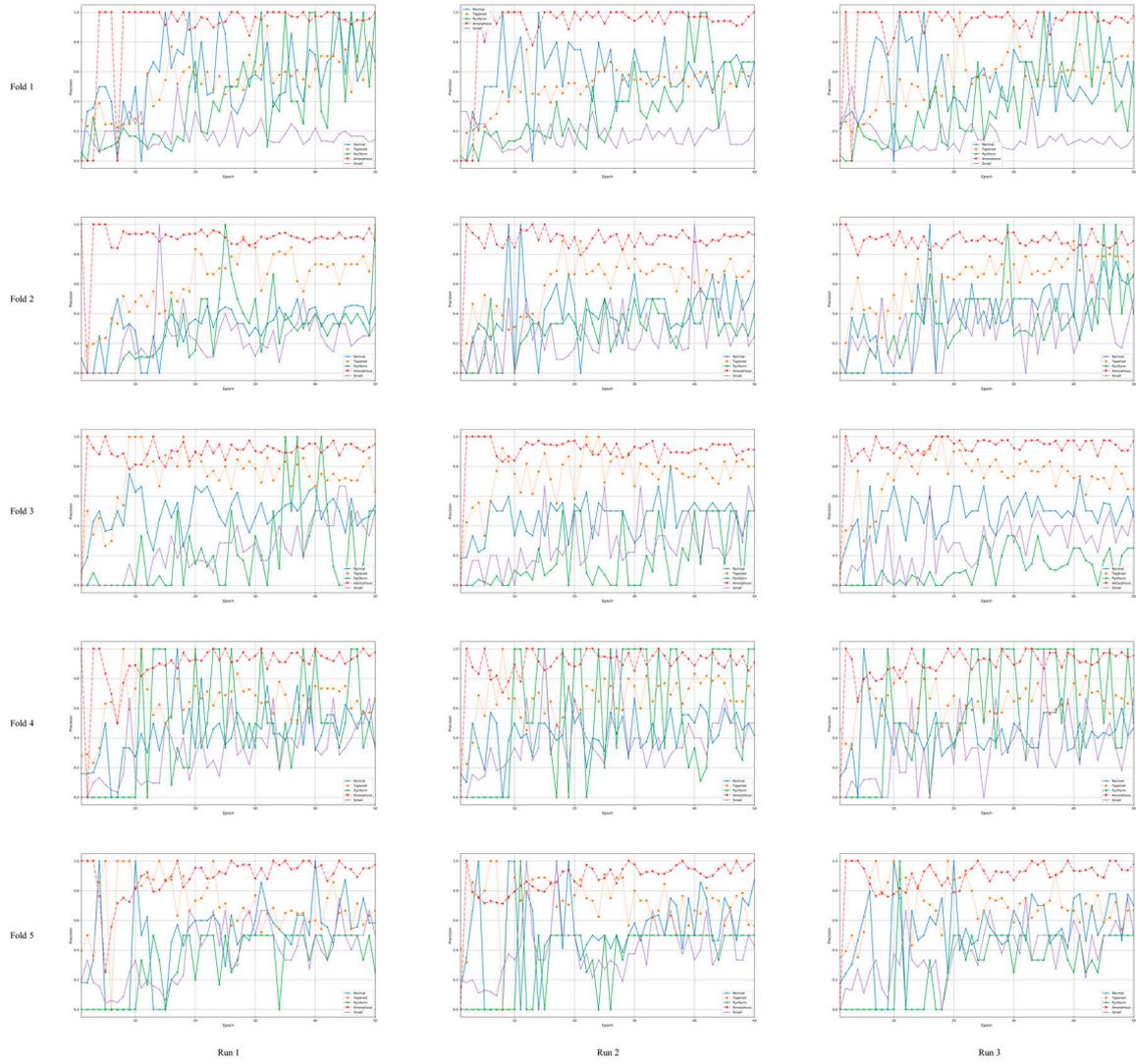
**Figure S9.** Detailed experimental results of the proposed model through 50 epochs in the total agreement setting of the SCIAN dataset, where 15 classification accuracy curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



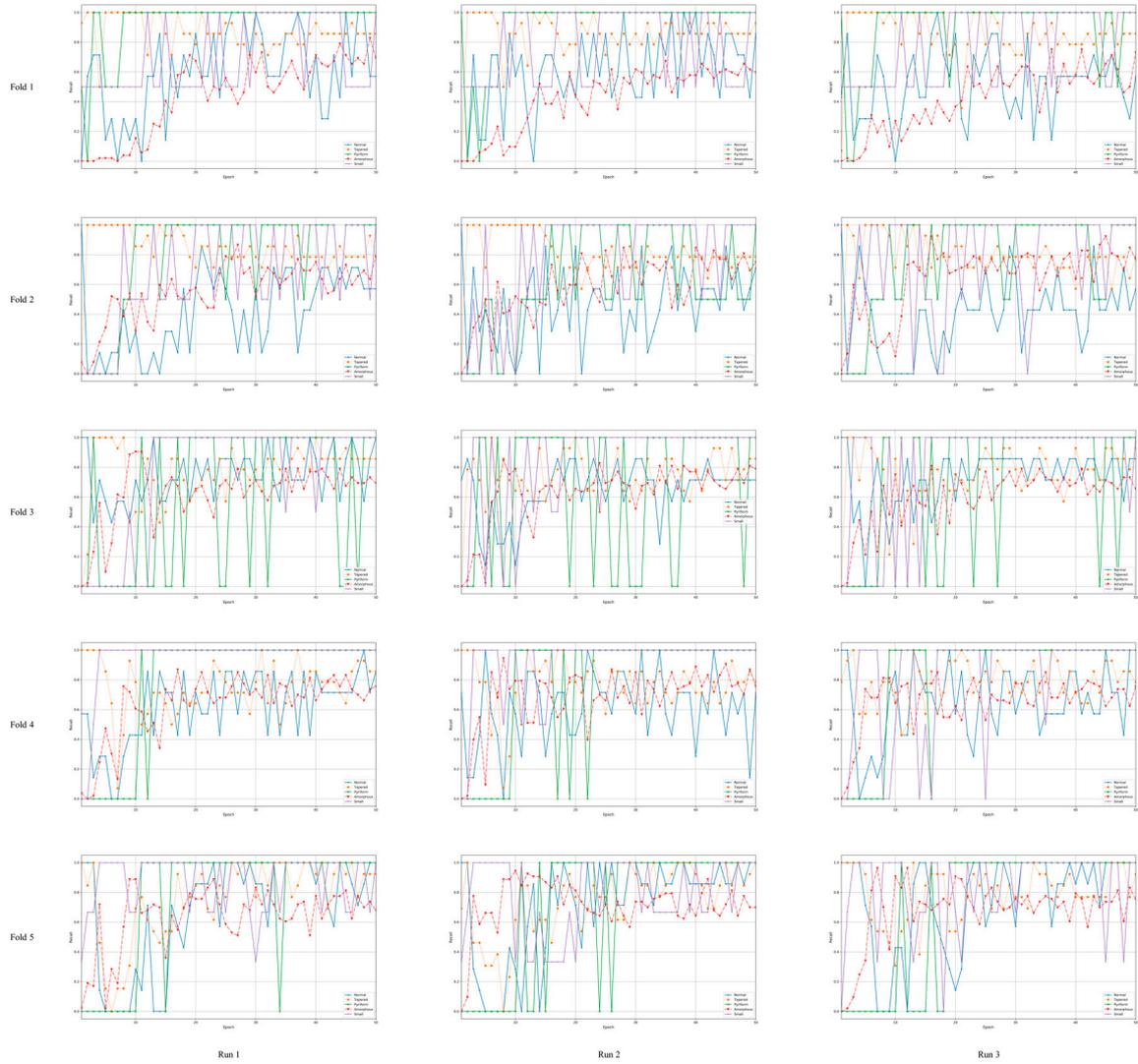
**Figure S10.** Detailed experimental results of the proposed model through 50 epochs in the total agreement setting of the SCIAN dataset, where 15 cost curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



**Figure S11.** Detailed experimental results of the proposed model through 50 epochs in the total agreement setting of the SCIAN dataset, where 15 confusion matrixes on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



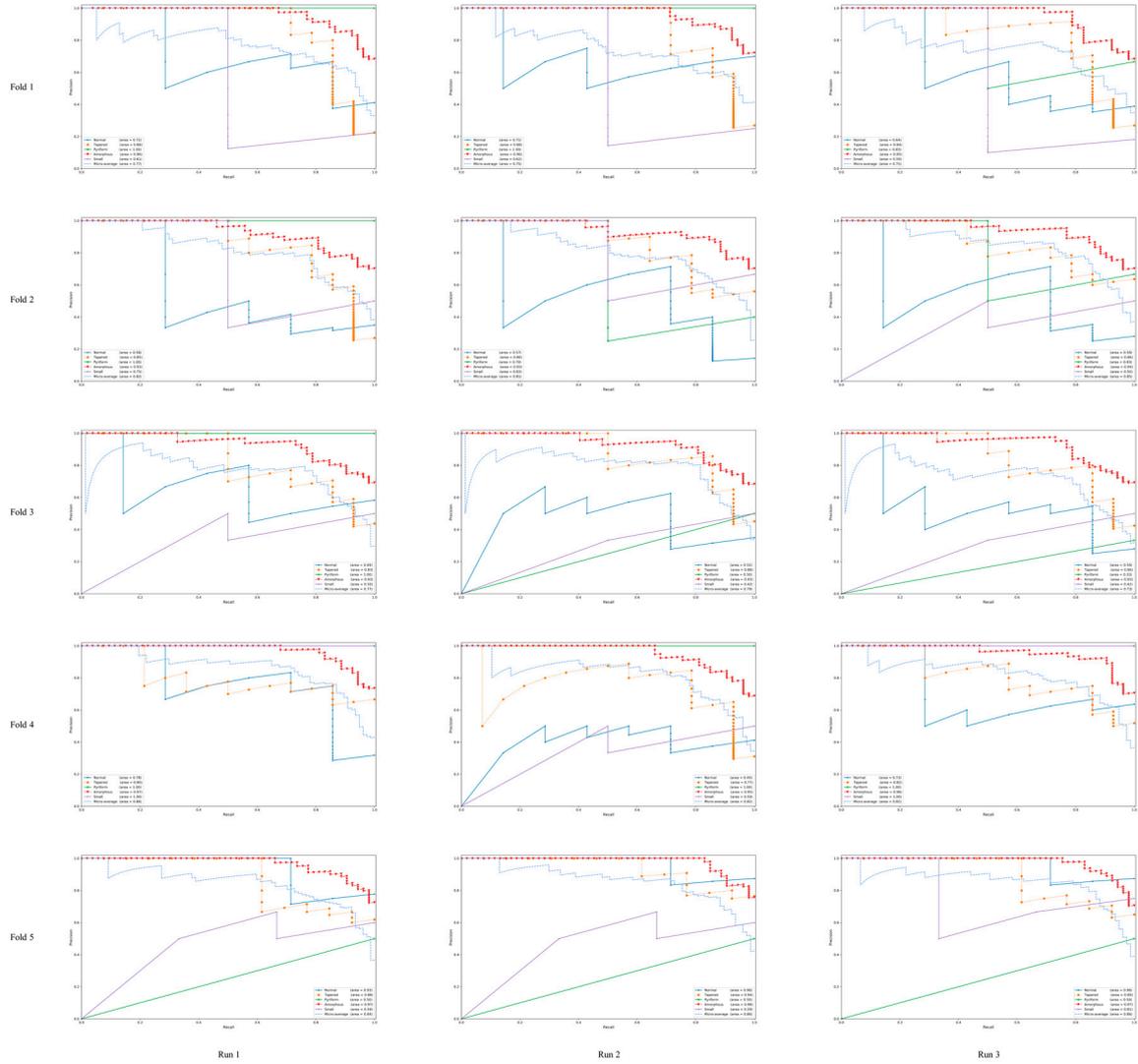
**Figure S12.** Detailed experimental results of the proposed model through 50 epochs in the total agreement setting of the SCIAN dataset, where 15 precision curves of each class with the number of epochs during the training on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



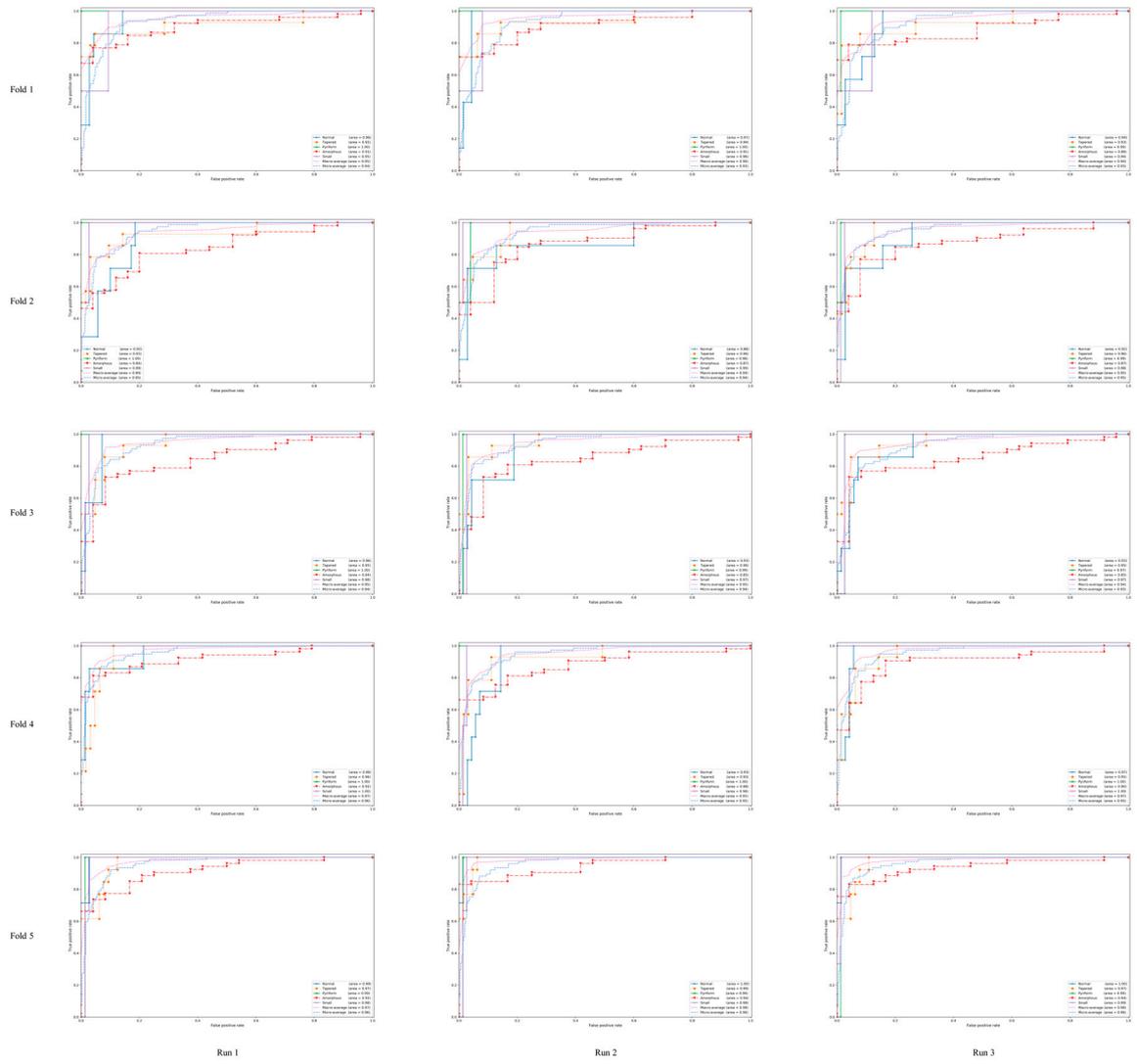
**Figure S13.** Detailed experimental results of the proposed model through 50 epochs in the total agreement setting of the SCIAN dataset, where 15 recall curves of each class with the number of epochs during the training on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



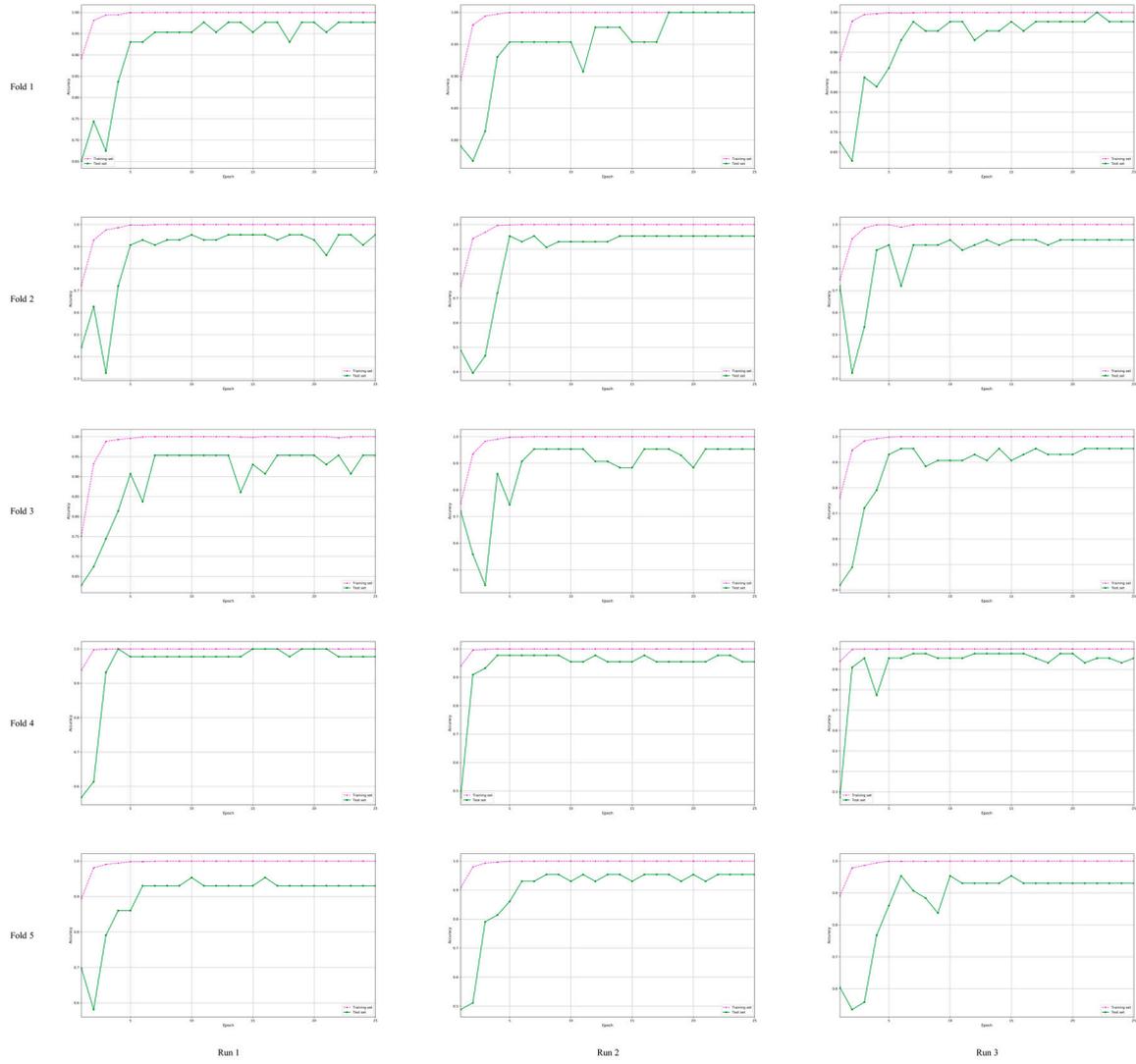
**Figure S14.** Detailed experimental results of the proposed model through 50 epochs in the total agreement setting of the SCIAN dataset, where 15 F1-score curves of each class with the number of epochs during the training on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



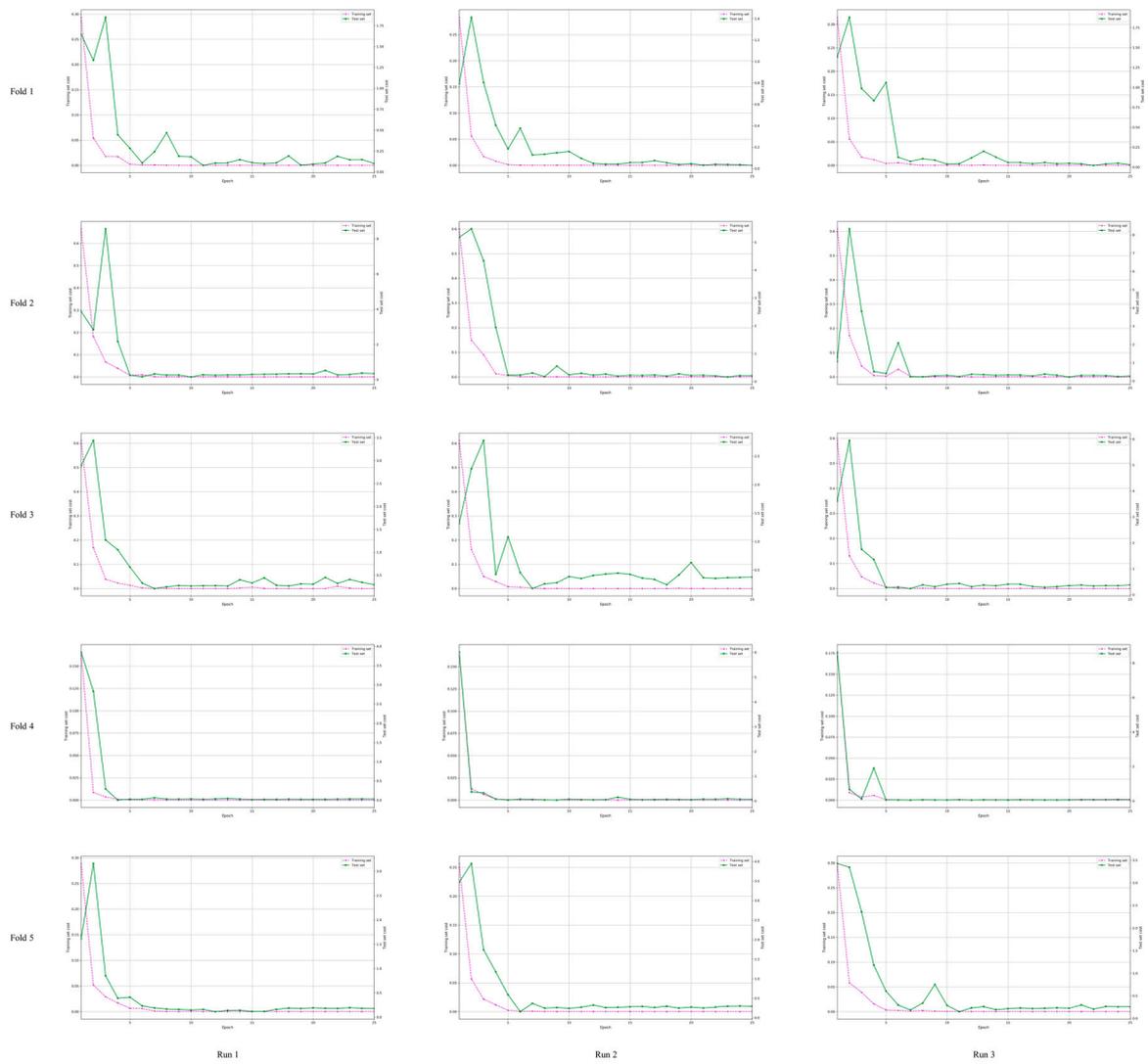
**Figure S15.** Detailed experimental results of proposed model in the total agreement setting of the SCIAN dataset, where 15 precision-recall curves of each class and their micro-averaging precision-recall curve on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated. A high area under the curve signifies the high precision as well as high recall.



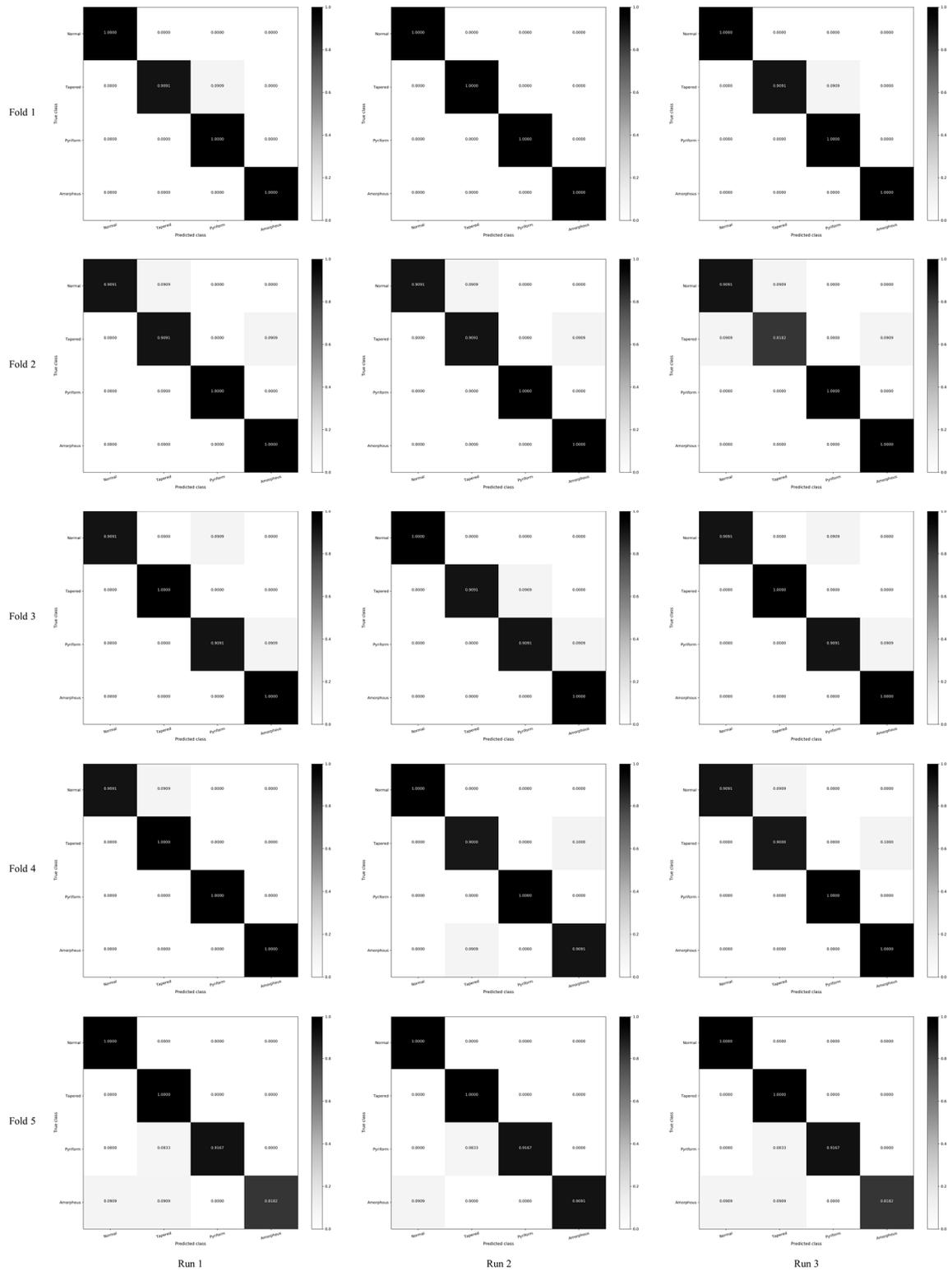
**Figure S16.** Detailed experimental results of proposed model in the total agreement setting of the SCIAN dataset, where 15 receiver operating characteristic (ROC) curves of each class and their macro and micro-averaging ROC curves on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated. These plots show the tradeoff between the true positive rate and false positive rate.



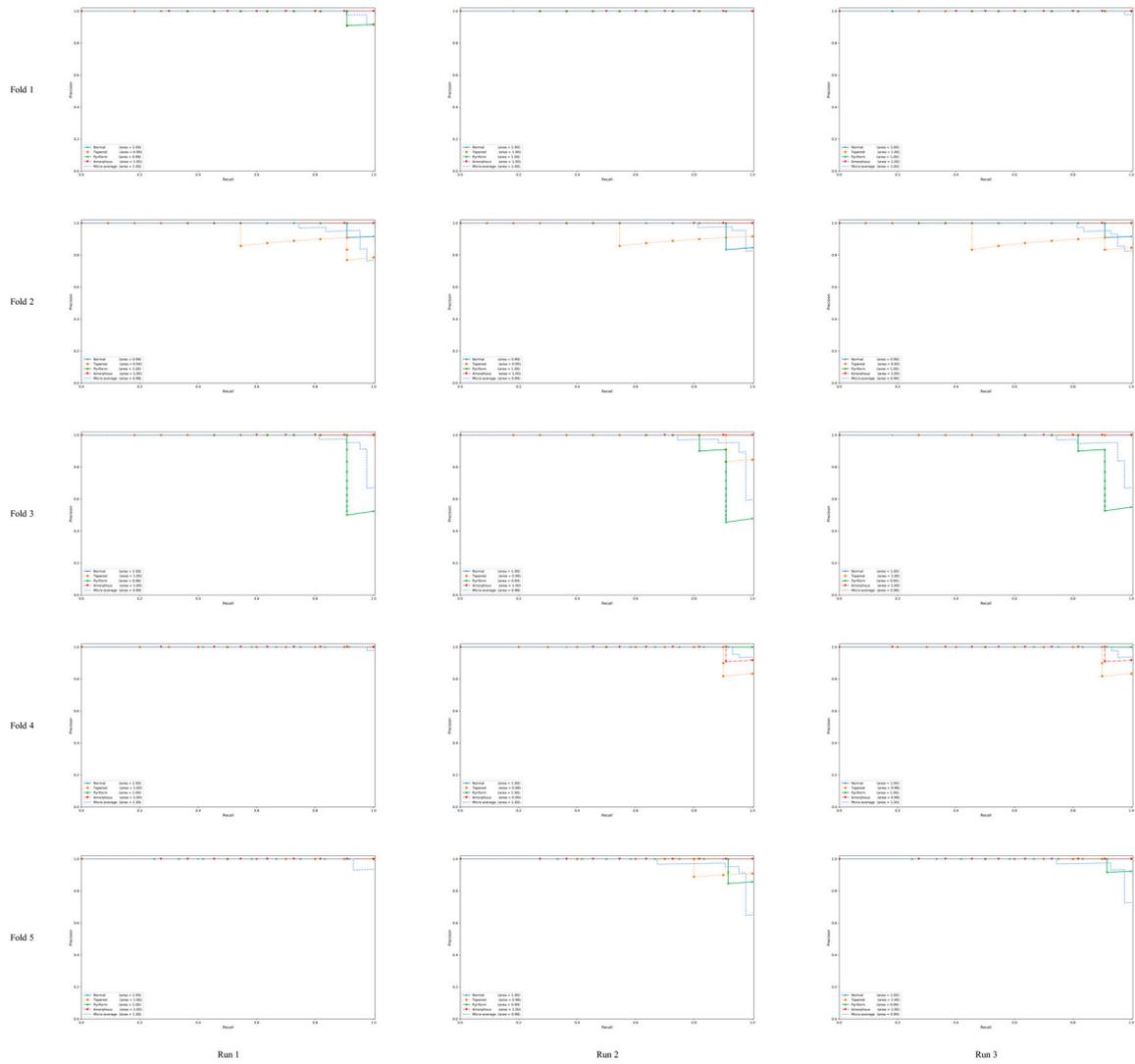
**Figure S17.** Detailed experimental results of the proposed model through 25 epochs of the HuSHeM dataset, where 15 classification accuracy curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



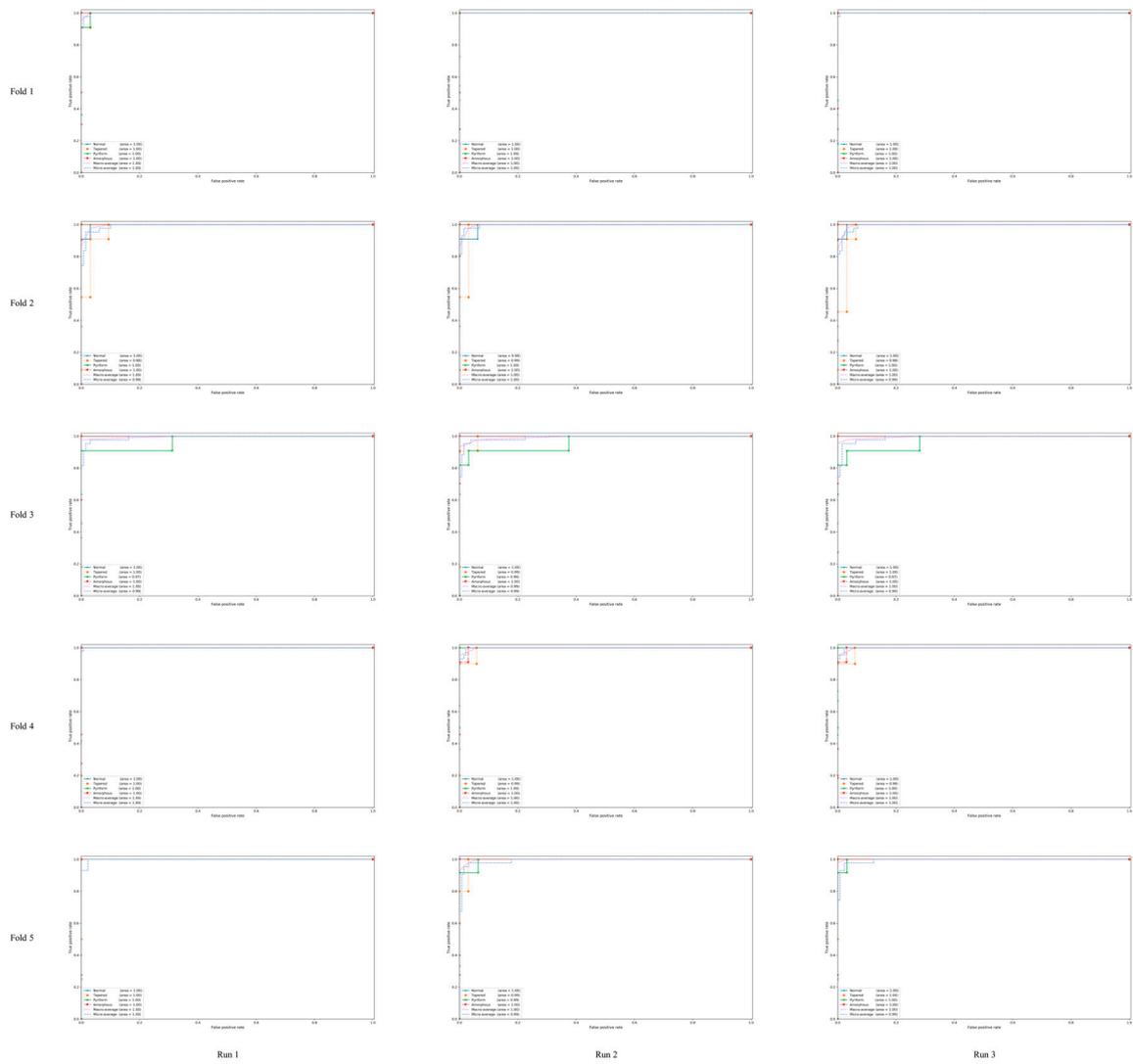
**Figure S18.** Detailed experimental results of the proposed model through 25 epochs of the HuSHeM dataset, where 15 cost curves with the number of epochs during the training on each of five possible choices of the training and test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



**Figure S19.** Detailed experimental results of the proposed model through 25 epochs of the HuSHem dataset, where 15 confusion matrixes on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated.



**Figure S20.** Detailed experimental results of proposed model of the HuSHeM dataset, where 15 precision-recall curves of each class and their micro-averaging precision-recall curve on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated. A high area under the curve signifies the high precision as well as high recall.



**Figure S21.** Detailed experimental results of proposed model of the HuSHeM dataset, where 15 receiver operating characteristic (ROC) curves of each class and their macro and micro-averaging ROC curves on each of five possible choices of the test sets for 3 runs (5 folds  $\times$  3 runs) are illustrated. These plots show the tradeoff between the true positive rate and false positive rate.