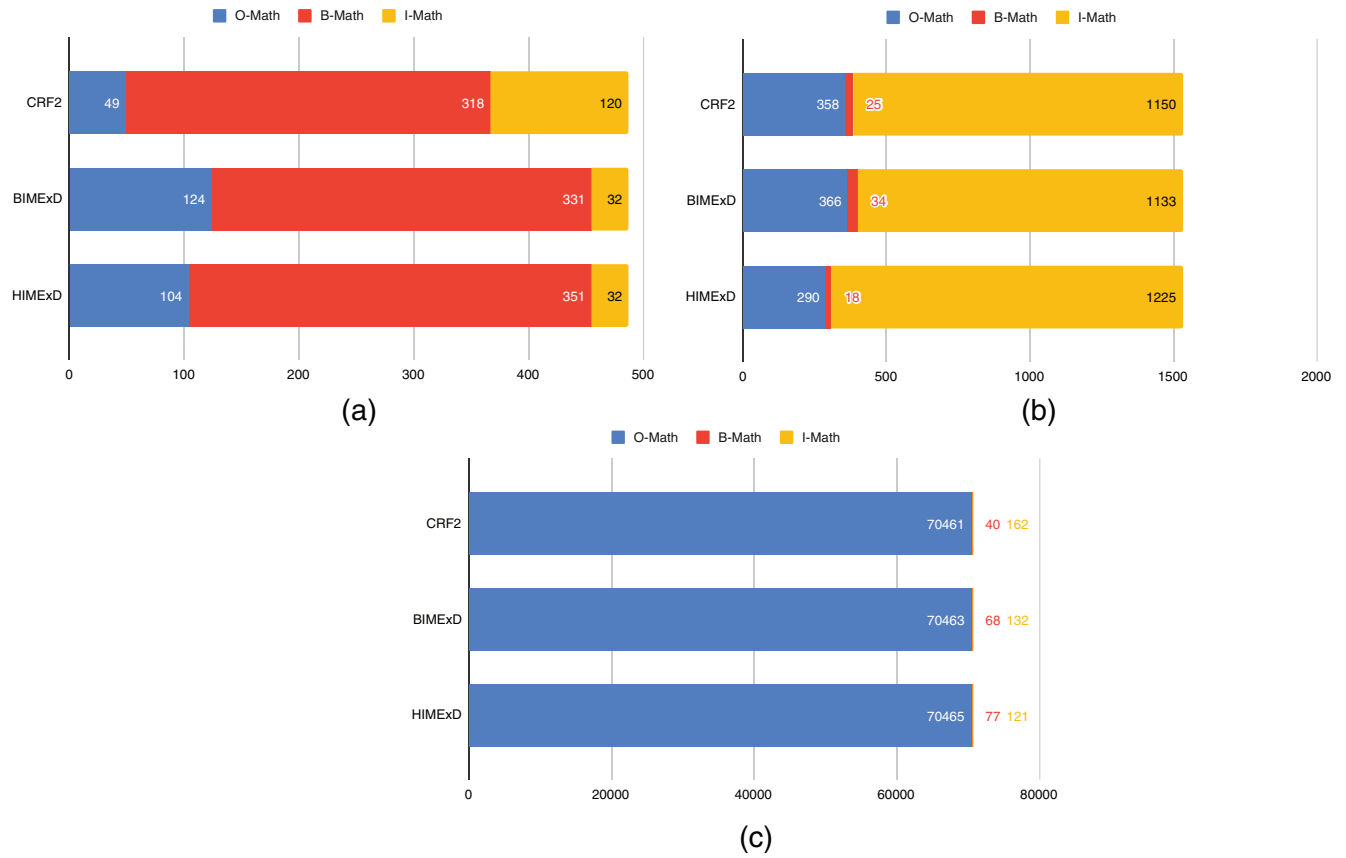

FIGURE 7 Confusion matrix comparison of different classes. (a) B-Math confusion matrix comparison, (b) I-Math confusion matrix comparison, (c) O-Math confusion matrix comparison

The top three best performing methods according to F-Measure (described in Table 3) are selected for confusion matrix comparison. The confusion matrix comparison for all the three classes B-Math, I-Math and O-Math is shown in Figure 7. The confusion matrix comparison for three methods CRF2, BIMExD and HIMExD for B-Math class is shown in Figure 7a. We can observe that a higher number of B-Math instances are predicted as O-Math instances in BIMExD compared to CRF2 and HIMExD. Our proposed approach HIMExD predicted a higher number of B-Math instances correctly compared to CRF2 and BIMExD. For BIMExD and HIMExD methods, the number of B-Math instances predicted as I-Math are lower in number whereas for CRF2 large number (120) of such instances are predicted as I-Math. So, in CRF2 the proposed features are in confusion between B-Math and I-Math. Confusion matrix comparison for I-Math class is shown in Figure 7b. Similar behaviour is observed for I-Math class as B-Math class, that is, more I-Math instances are predicted as O-Math in BIMExD compared to CRF2 and HIMExD. Similarly, more I-Math instances are predicted as I-Math for HIMExD compared to CRF2. The confusion matrix comparison for O-Math class is shown in Figure 7c. All three methods are predicting O-Math instances O-Math most of the times. This is due to more number of O-Math instances are available for training.

In summary, our proposed approach outperforms all the methods in the literature which indicates that the method can be used to successfully detect inline mathematical expressions from scientific documents.

4.5.1 | Result analysis: Different cases

We now perform an in-depth analysis of the failure cases of the algorithms. We see that all the methods except InftyReader perform very well for the O-Math class. The F-Scores for the O-Math class are above 99% for most of the methods indicating both the precision and recall for the O-Math class are high. On the other hand, the performances of the methods vary for the classes that have mathematical symbols. Hence, in this analysis, we look closely at the test cases containing the mathematical symbols. We characterized and categorized the math symbols or expressions in multiple groups, as shown in the Table 5. In the table, we also indicate the performances of the methods for expressions of these types. If a method misclassifies examples from a particular type for more than 70% of the test cases, we consider that the method performs poorly (or fail) for that type. Otherwise, the method is considered to do well (or succeed in identifying) examples from that type.