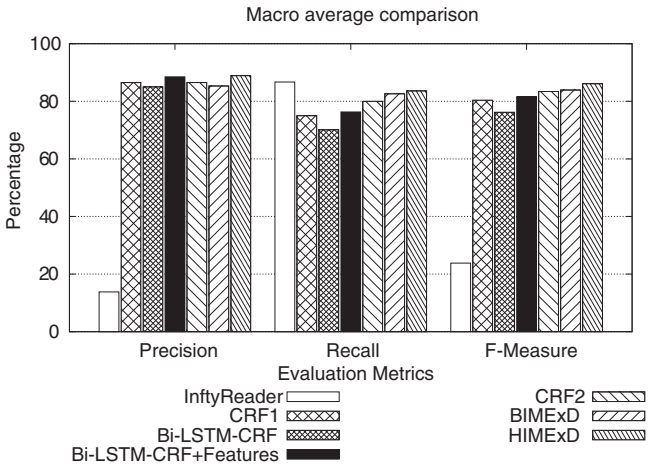
TABLE 3 Results comparison of different models used in our experimentation

(3a) Macro average.			
Method	**Precision**	**Recall**	**F-measure**
InftyReader (Eto and Suzuki (2001))	13.8%	**86.7%**	23.8%
CRF1 (Iwatsuki et al. (2017))	86.5%	75.0%	80.4%
Bi-LSTM-CRF (Huang et al. (2015))	85.0%	70.1%	76.2%
Bi-LSTM-CRF + features	88.5%	76.3%	81.6%
CRF2	87.8%	80.0%	83.4%
BIMExD	85.3%	82.6%	83.9%
HIMExD	**88.9%**	83.6%	**86.1%**
(3b) Micro average.			
Method	**Precision**	**Recall**	**F-measure**
InftyReader (Eto and Suzuki (2001))	13.0%	89.9%	22.8%
CRF1 (Iwatsuki et al. (2017))	94.9%	83.6%	88.9%
Bi-LSTM-CRF (Huang et al. (2015))	98.4%	98.5%	98.5%
Bi-LSTM-CRF + features	98.7%	98.6%	98.7%
CRF2	98.7%	**99.1**%	98.8%
BIMExD	**99.0%**	99.0%	**99.0%**
HIMExD	**99.0%**	**99.1%**	**99.0%**

Note: The first three models are from literature, whereas the remaining ones are proposed in this paper. CRF2 is our feature-based approach, BIMExD & HIMExD are our proposed deep learning based approaches. Bi-LSTM-CRF + Features are built on top of existing Bi-LSTM-CRF architecture. Best results are put in bold.

FIGURE 5 Macro average comparison



successful impact of deep learning models (as they capture a representation of important features efficiently). Bar charts comparing all the baseline and proposed methods for macro average and micro average are shown in Figures 5 and 6 respectively. We can observe that our proposed method F-Measure is performing better than all other baseline approaches.

As we aim to achieve better evaluation results for minor classes (B-Math and I-Math), in this section, we will discuss the scores of each class in detail. The top three methods which are performed best according to the F-Measure (as described in Table 3) are analysed further. The class-specific performance evaluation of different models is shown in Table 4. Results of the Conditional Random Fields (CRF2) model are shown in Table 4. For all the evaluation metrics (*precision*, *recall*, *f-measure*) the score for O-Math class is around 99% which is expected because O-Math class has large number of training instances. For both B-Math and I-math classes, *precision* is better than *recall*. *F-measure* is better for the I-Math class compared to the B-Math class. BIMExD results are shown in Table 5. The performance behaviour for the O-Math class in this method is similar to the behaviour for the O-Math class in the CRF method. *Precision* is better than *recall* for both B-Math and I-Math classes. In BIMExD, *F-Measure* for the I-Math class is higher compared to the B-Math class.