F1 score distribution for each model and dataset