F1 Scores Across Different Methods and Datasets