Machine Learning Final Project Report

Vrishab Prasanth Davey (300438343), Surendar Pala Danasekaran (300401916)

Introduction

The effectiveness of semi-supervised learning (SSL) approaches in comparison to supervised learning with gradient boosting was investigated in this experiment. The dataset was divided into training and testing sets, with the training set being further subdivided into parts that were labeled and those that were not. Various SSL algorithms were implemented, including Self-Training, Co-Training, SemiBoost, and PCA Pretraining. Metrics like accuracy, precision, recall, F1-score, and ROC AUC were assessed for varying percentages of labeled data (10%, 20%, 30%, 40%, and 50%).

Results

Supervised Learning Baseline:

- The supervised Gradient Boosting model consistently achieved the highest accuracy across all experiments, with a peak of **74.32%** accuracy. This indicates its strong performance when sufficient labeled data is available.
- A benchmark for comparing SSL methods is established by metrics such as ROC AUC (**80.71%**), F1-score (**64.29%**), recall (**62.88%**), and precision (**65.75%**).

Self-Training:

- Self-Training demonstrated efficacy across various labeled data partitions, notably with 10% labeled data, attaining an accuracy of **71.43%** and a ROC AUC of **78.55%**.
- Performance improved with an increase in labeled data but plateaued beyond 40%.
- The algorithm showed balanced precision and recall, with F1 scores remaining consistent, highlighting its effectiveness in leveraging unlabeled data.

Co-Training:

- Co-Training underperformed relative to other SSL methods, particularly at lower labeled percentages (e.g., **66.93%** accuracy with 10% labeled data).
- The precision-recall imbalance was notable, with poor recall in some cases (e.g., **23.14%** at 10% labeled data).
- This suggests that Co-Training struggles with highly imbalanced or sparse labeled data.

SemiBoost:

- With an accuracy of 75.12% and an ROC AUC of 79.43%, SemiBoost did better than its competitors, especially with 30% labeled data.
- It consistently outperformed Co-Training, leveraging the ensemble approach to refine pseudo-labeling.
- However, its dependence on similarity measures could limit its generalizability to datasets with less discernible patterns.

PCA Pretraining:

- PCA Pretraining showed moderate effectiveness, performing well with 10% labeled data (**71.59% accuracy, 78.71% ROC AUC**) but plateauing beyond 40% labeled data.
- It found hidden features to help with classification, but the fact that it relies on linear dimensionality reduction could make it less useful for larger, more complicated datasets.

Lessons Learned

Impact of Unlabeled Data:

- SSL methods significantly improve performance when labeled data is scarce (e.g., 10% or 20%). However, their gains diminish as more labeled data becomes available, converging towards supervised learning performance.
- This highlights the importance of choosing SSL for scenarios with limited labeled data.

Algorithm Sensitivity:

- Self-Training and SemiBoost did well with different labeled-unlabeled splits, but Co-Training had trouble, which showed how sensitive it is to feature independence and data quality.
- PCA Pretraining proved useful in datasets with latent structure but requires careful selection of the number of components.

Supervised Learning as a Benchmark:

- The supervised baseline provides a valuable reference point. The fact that SSL methods can't consistently do better than supervised learning when there is enough labeled data shows the trade-off that comes with using fake labels.

Practical Considerations:

- SSL methods require careful tuning of confidence thresholds and similarity metrics to prevent error propagation during pseudo-labeling.
- SemiBoost's reliance on ensemble learning and PCA Pretraining's focus on feature extraction make them suitable for different problem domains.

Conclusion

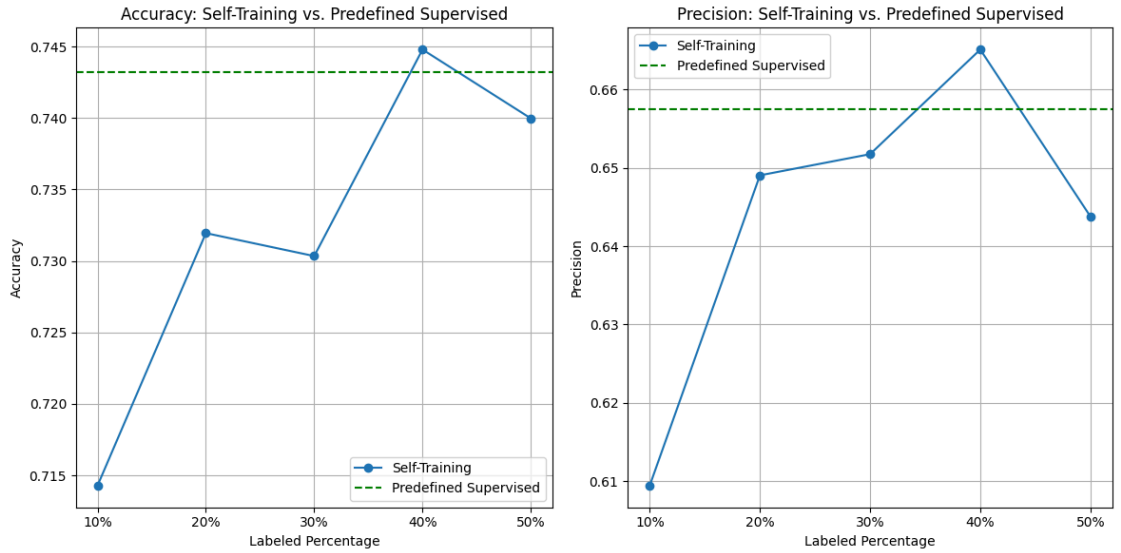
The experiment demonstrates the efficacy of semi-supervised learning in leveraging unlabeled data, especially in contexts where labeled data is limited. SemiBoost and Self-Training consistently surpassed other Semi Supervised algorithms, whereas Co-Training underperformed. PCA Pretraining was constrained by its linear assumptions, but it produced moderate improvements. When there was a large amount of labeled data available, the supervised Gradient Boosting model continued to be the most reliable method. These observations highlight the significance of customizing SSL methods for particular datasets and application scenarios.

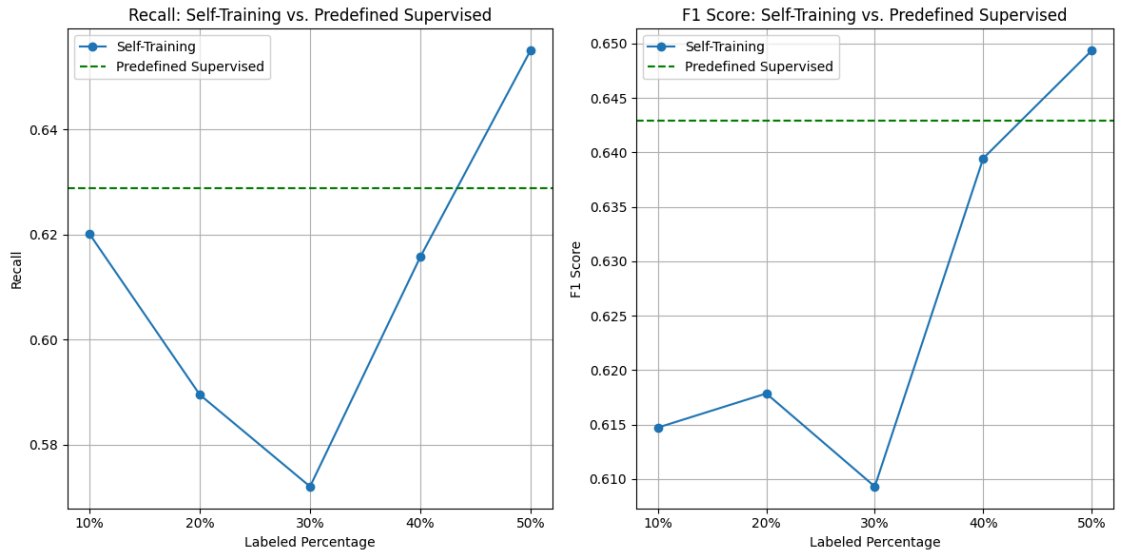
Results

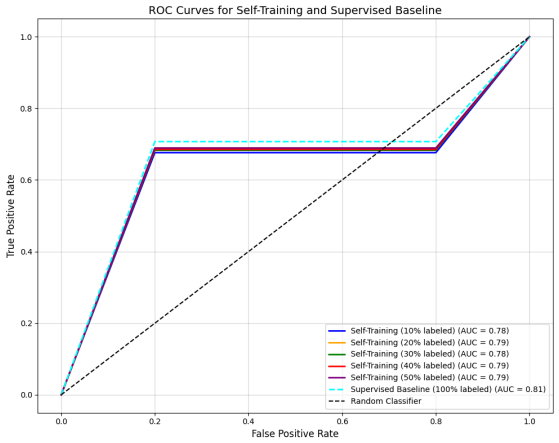
Self-Training vs. Supervised Baseline

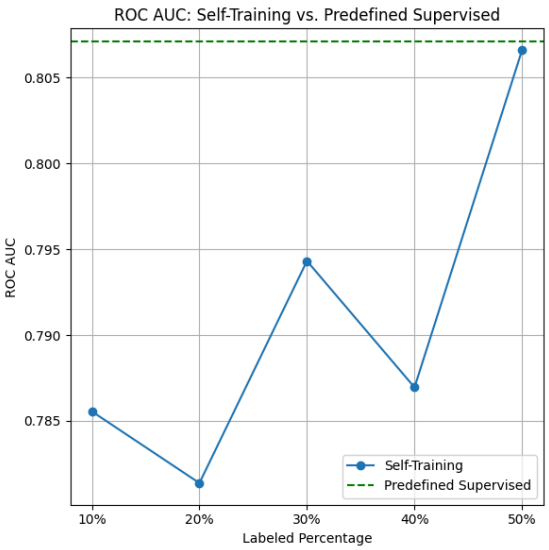
Self-Training vs. Predefined Supervised Model Results:

	Labeled Percentage	Accuracy	Precision	Recall	F1 Score
0	10%	0.714286	0.609442	0.620087	0.614719
1	20%	0.731942	0.649038	0.589520	0.617849
2	30%	0.730337	0.651741	0.572052	0.609302
3	40%	0.744783	0.665094	0.615721	0.639456
4	50%	0.739968	0.643777	0.655022	0.649351
5	100% (Predefined Supervised)	0.743200	0.657500	0.628800	0.642900





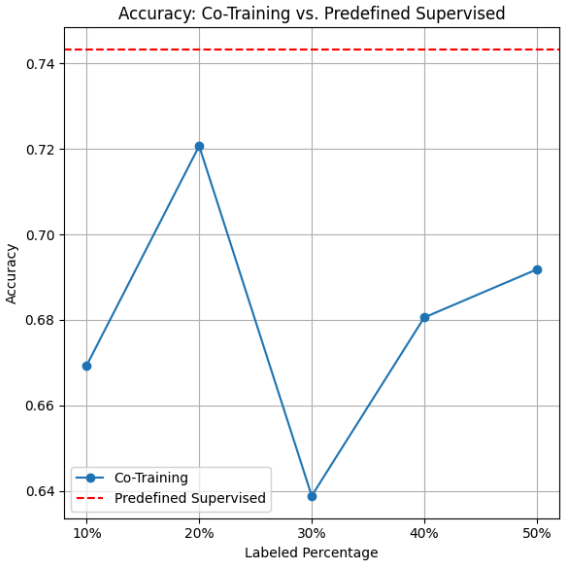


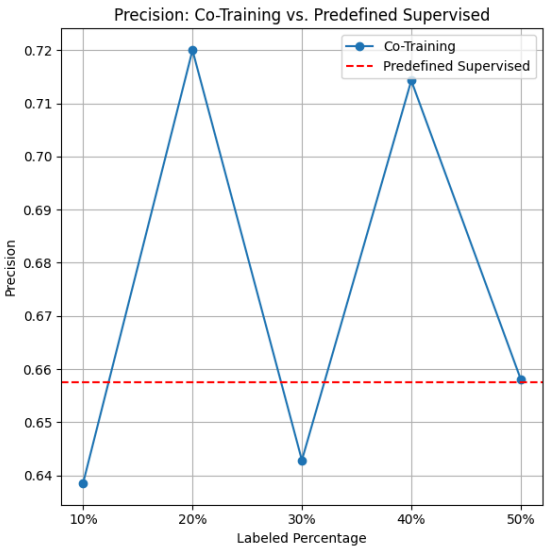


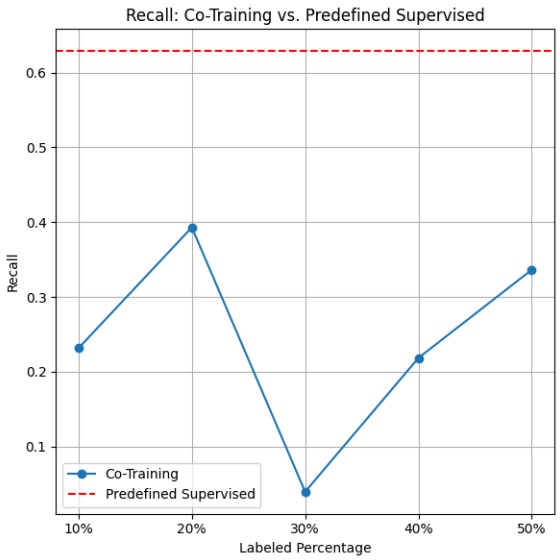
Co-Training vs. Supervised Baseline

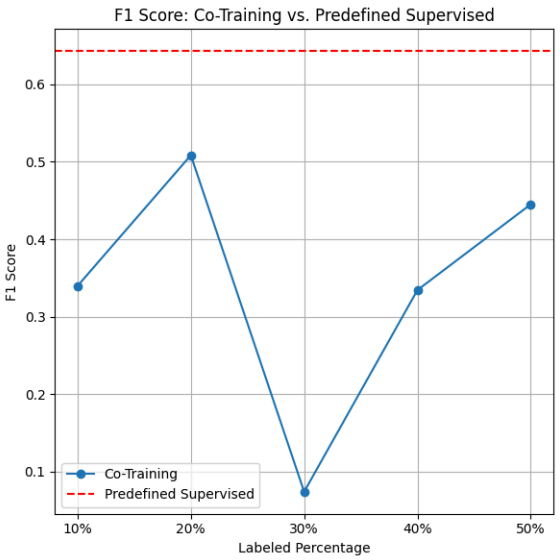
Co-Training vs. Predefined Supervised Model Results:

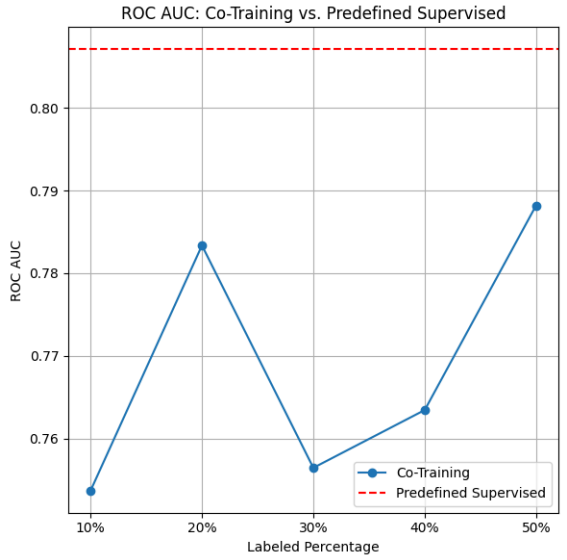
	Labeled Percentage	Accuracy	Precision	Recall	F1 Score
0	10%	0.669342	0.638554	0.231441	0.339744
1	20%	0.720706	0.720000	0.393013	0.508475
2	30%	0.638844	0.642857	0.039301	0.074074
3	40%	0.680578	0.714286	0.218341	0.334448
4	50%	0.691814	0.658120	0.336245	0.445087
5	100% (Predefined Supervised)	0.743200	0.657500	0.628800	0.642900

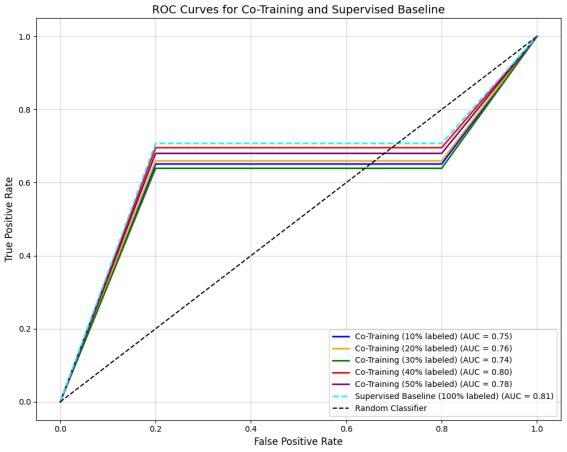








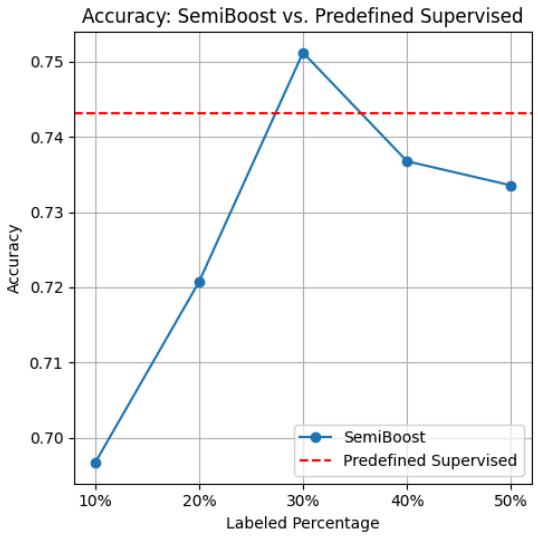


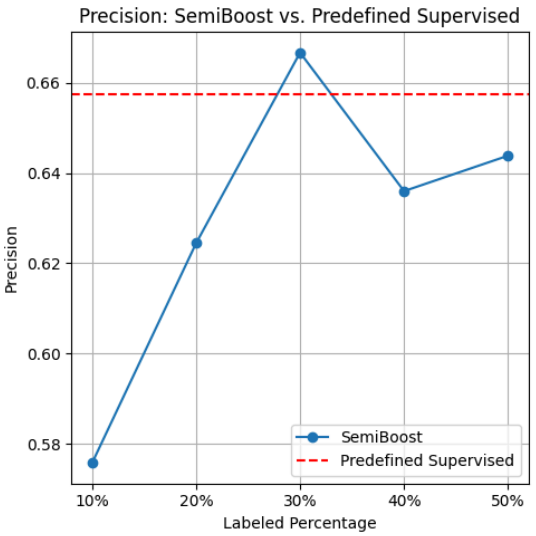


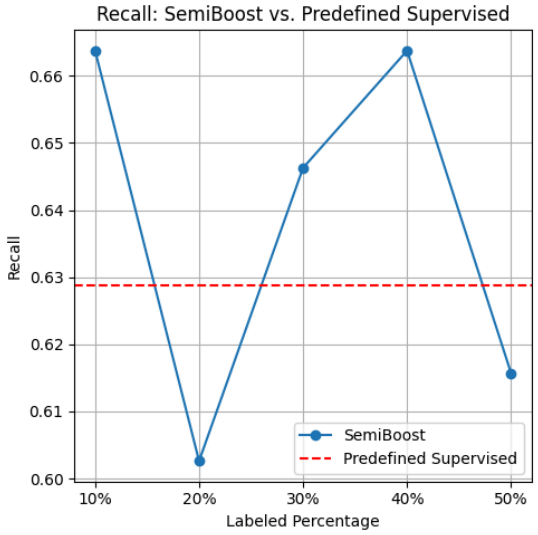
SemiBoost vs. Supervised Baseline

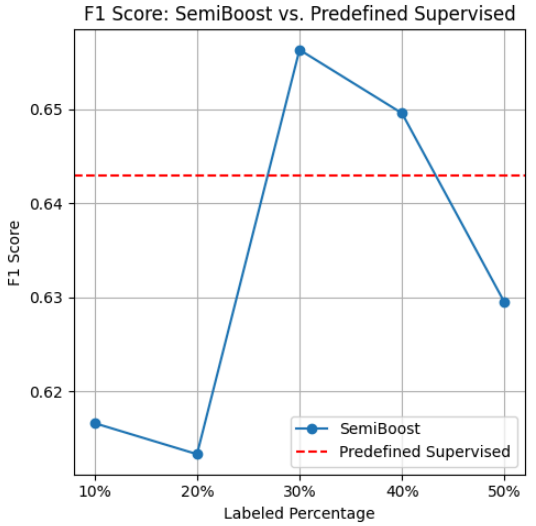
Comparison of SemiBoost and Predefined Supervised Metrics:

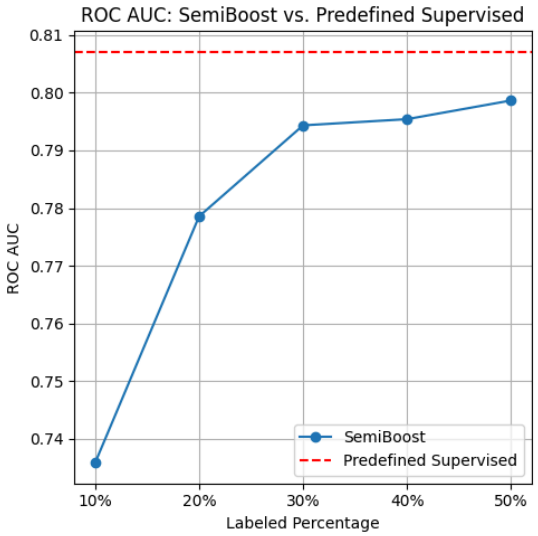
	Labeled Percentage	Accuracy	Precision	Recall	F1 Score
0	10%	0.696629	0.575758	0.663755	0.616633
1	20%	0.720706	0.624434	0.602620	0.613333
2	30%	0.751204	0.666667	0.646288	0.656319
3	40%	0.736758	0.635983	0.663755	0.649573
4	50%	0.733547	0.643836	0.615721	0.629464
5	100% (Predefined Supervised)	0.743200	0.657500	0.628800	0.642900

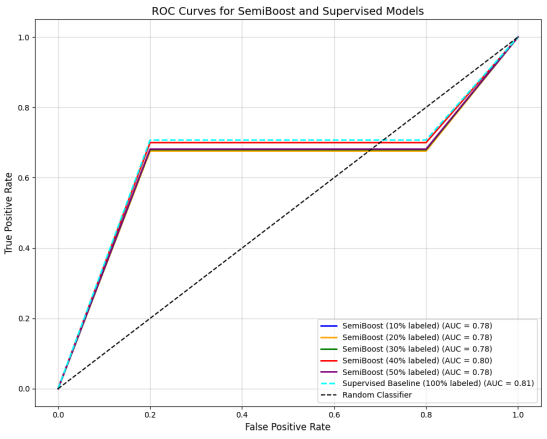

Accuracy: SemiBoost vs. Predefined Supervised
Accuracy
Labeled Percentage
SemiBoost
Predefined Supervised
10%
20%
30%
40%
50%
0.70
0.71
0.72
0.73
0.74
0.75


Precision: SemiBoost vs. Predefined Supervised
Precision
Labeled Percentage
SemiBoost
Predefined Supervised
10%
20%
30%
40%
50%
0.58
0.60
0.62
0.64
0.66


Recall: SemiBoost vs. Predefined Supervised
Recall
Labeled Percentage
SemiBoost
Predefined Supervised
10%
20%
30%
40%
50%
0.60
0.61
0.62
0.63
0.64
0.65
0.66

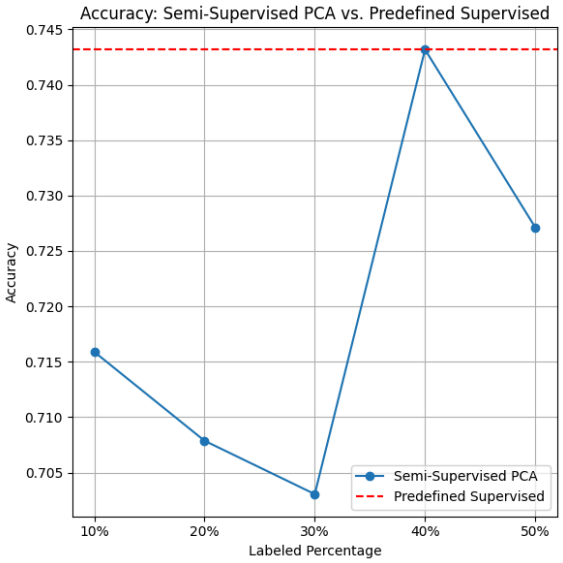

F1 Score: SemiBoost vs. Predefined Supervised
F1 Score
Labeled Percentage
SemiBoost
Predefined Supervised
10%
20%
30%
40%
50%
0.62
0.63
0.64
0.65

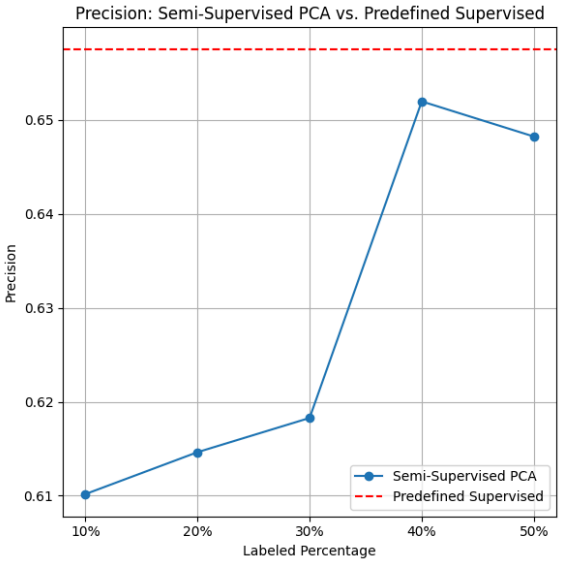

ROC AUC: SemiBoost vs. Predefined Supervised
ROC AUC
Labeled Percentage
SemiBoost
Predefined Supervised
10%
20%
30%
40%
50%
0.74
0.75
0.76
0.77
0.78
0.79
0.80
0.81

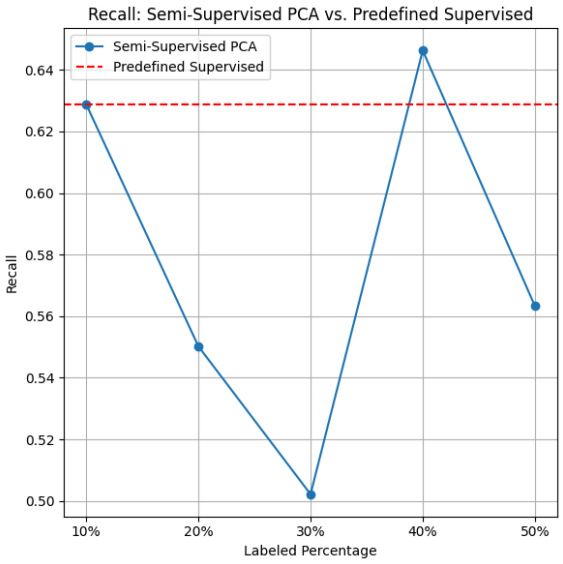

ROC Curves for SemiBoost and Supervised Models
True Positive Rate
False Positive Rate
SemiBoost (10% labeled) (AUC = 0.78)
SemiBoost (20% labeled) (AUC = 0.78)
SemiBoost (30% labeled) (AUC = 0.78)
SemiBoost (40% labeled) (AUC = 0.80)
SemiBoost (50% labeled) (AUC = 0.78)
Supervised Baseline (100% labeled) (AUC = 0.81)
Random Classifier

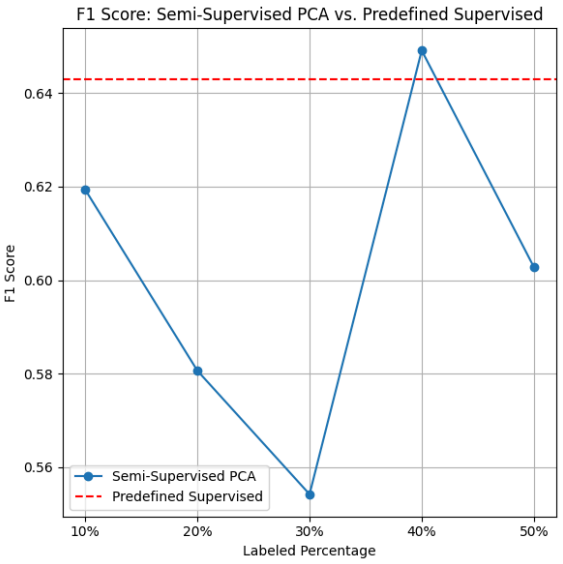
Semi-Supervised PCA Pretraining vs. Supervised Baseline

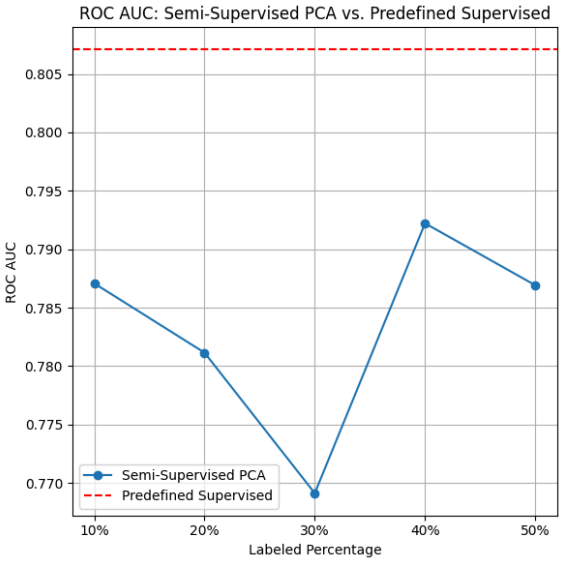
```
Semi-Supervised PCA Pretraining vs. Predefined Supervised Results:
             Labeled Percentage  Accuracy  Precision    Recall  F1 Score
0                           10%  0.715891   0.610169  0.628821  0.619355
1                           20%  0.707865   0.614634  0.550218  0.580645
2                           30%  0.703050   0.618280  0.502183  0.554217
3                           40%  0.743178   0.651982  0.646288  0.649123
4                           50%  0.727127   0.648241  0.563319  0.602804
5  100% (Predefined Supervised)  0.743200   0.657500  0.628800  0.642900
```

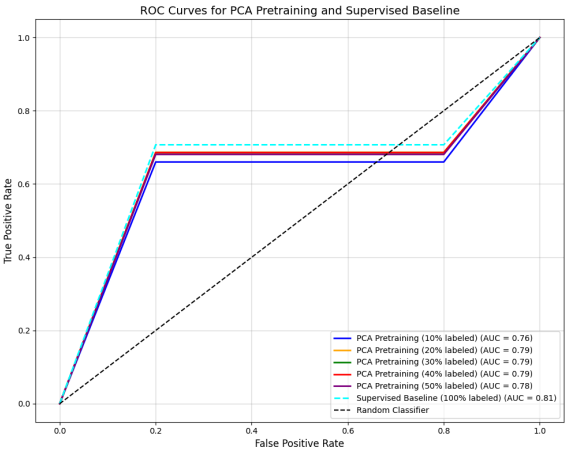









ROC AUC: Semi-Supervised PCA vs. Predefined Supervised
ROC AUC
Labeled Percentage
10%
20%
30%
40%
50%
0.805
0.800
0.795
0.790
0.785
0.780
0.775
0.770
Semi-Supervised PCA
Predefined Supervised


ROC Curves for PCA Pretraining and Supervised Baseline
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
PCA Pretraining (10% labeled) (AUC = 0.76)
PCA Pretraining (20% labeled) (AUC = 0.79)
PCA Pretraining (30% labeled) (AUC = 0.79)
PCA Pretraining (40% labeled) (AUC = 0.79)
PCA Pretraining (50% labeled) (AUC = 0.78)
Supervised Baseline (100% labeled) (AUC = 0.81)
Random Classifier