Research Paper Evaluation by Implementation

**Paper taken:** Early Detection of Host-based Intrusions in Linux Environment

Xinrun Zhang*∗*, Quamar Niyaz*∗*, Farha Jahan*†*, Weiqing Sun*†*

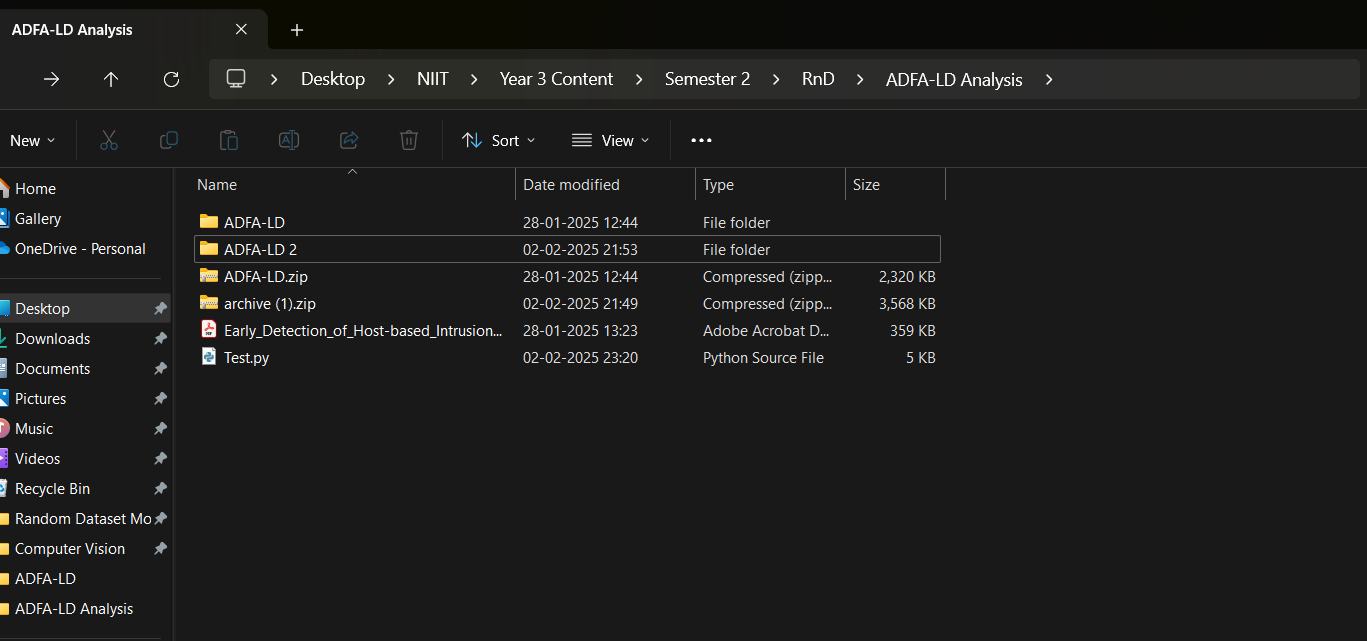
*∗*ECE Department, College of Engineering and Sciences

Purdue University Northwest, Hammond, IN 46323, USA

*†*College of Engineering, The University of Toledo, OH 43607, USA

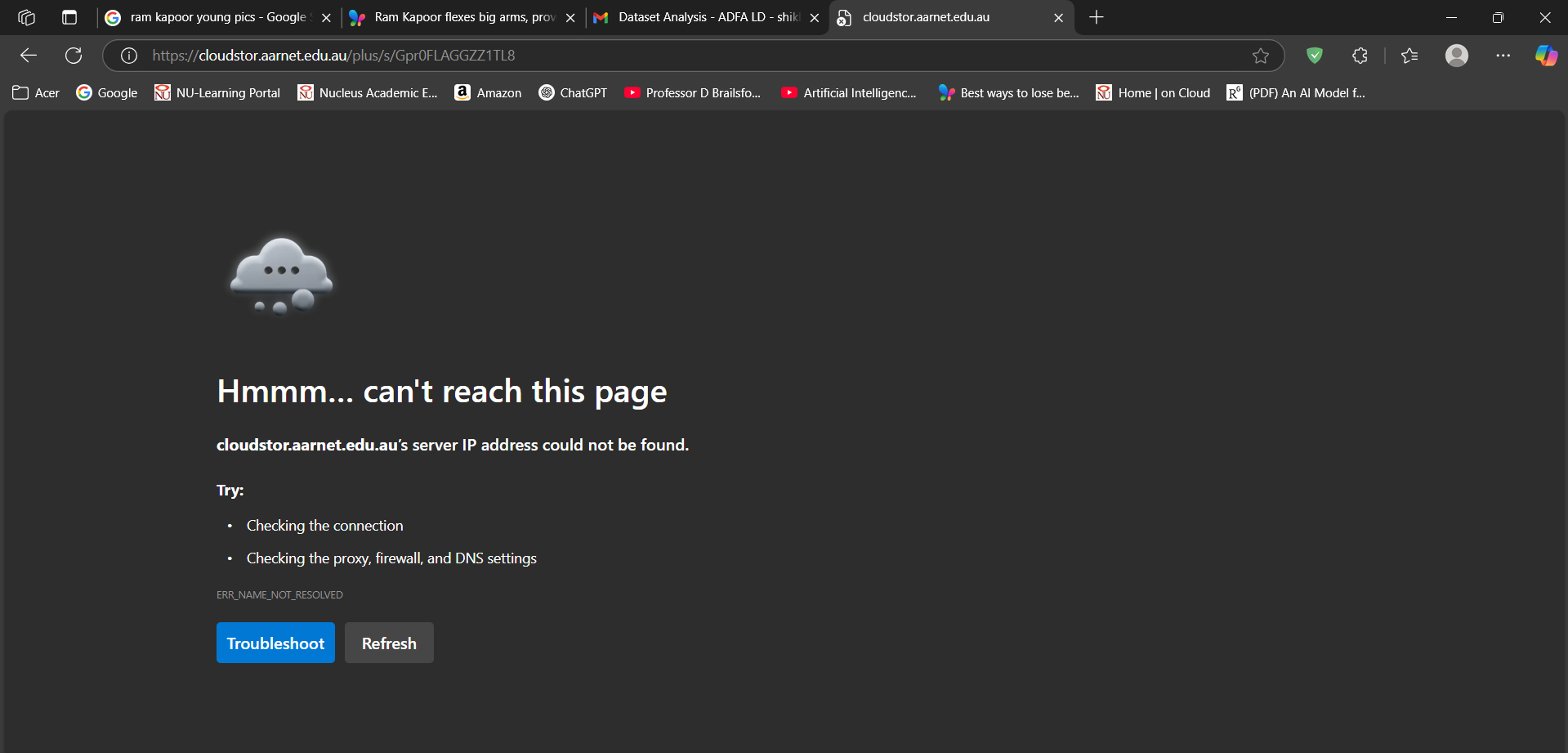
*{*zhan3031, qniyaz*}*@pnw.edu, *{*farha.jahan, [weiqing.sun*}*@utoledo.edu](mailto:weiqing.sun%7d@utoledo.edu)

**Dataset Used**: ADFA-LD Dataset

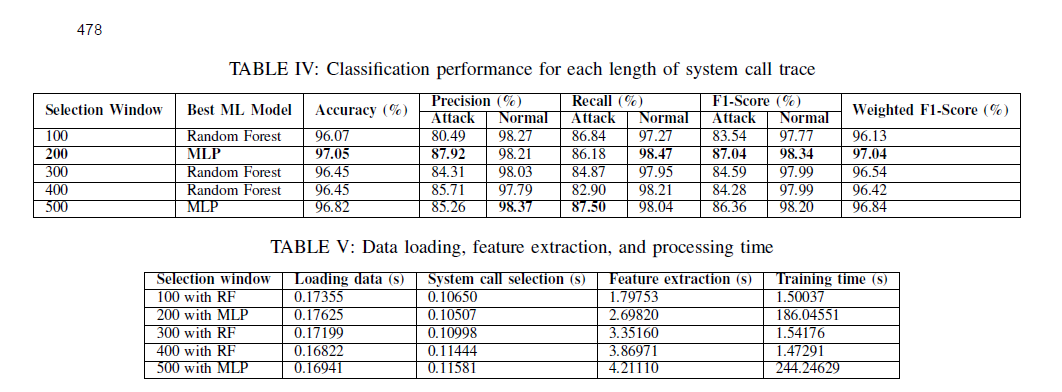


Here You can see, I used 2 ADFA-LD Dataset Types Taken from Different Sources, Kaggle and GitHub. ADFA-LD is from Github and ADFA-LD 2 is from Kaggle

**Why I did not use the official link from website of ADFA?**

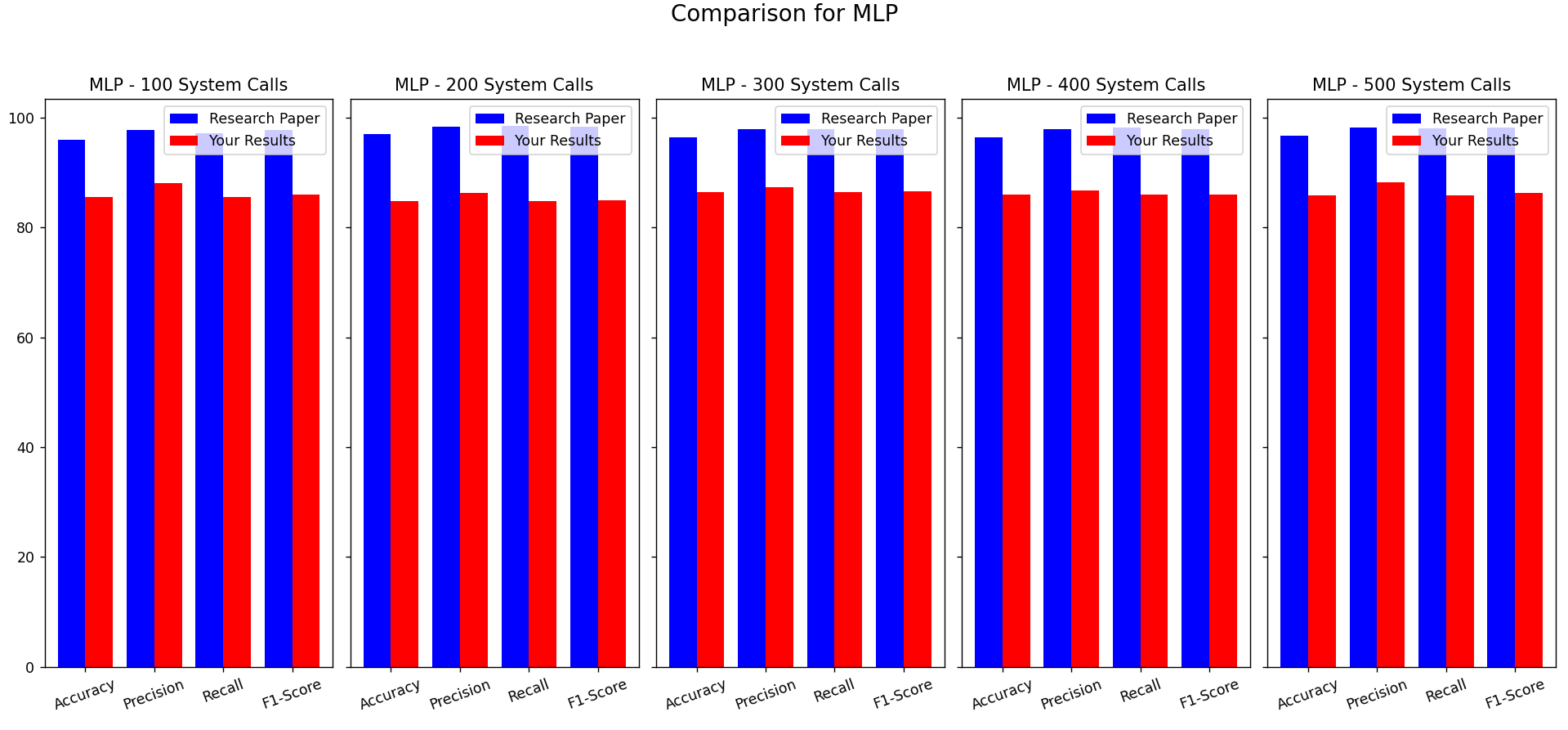
****

**Research Paper Performance Outputs:**

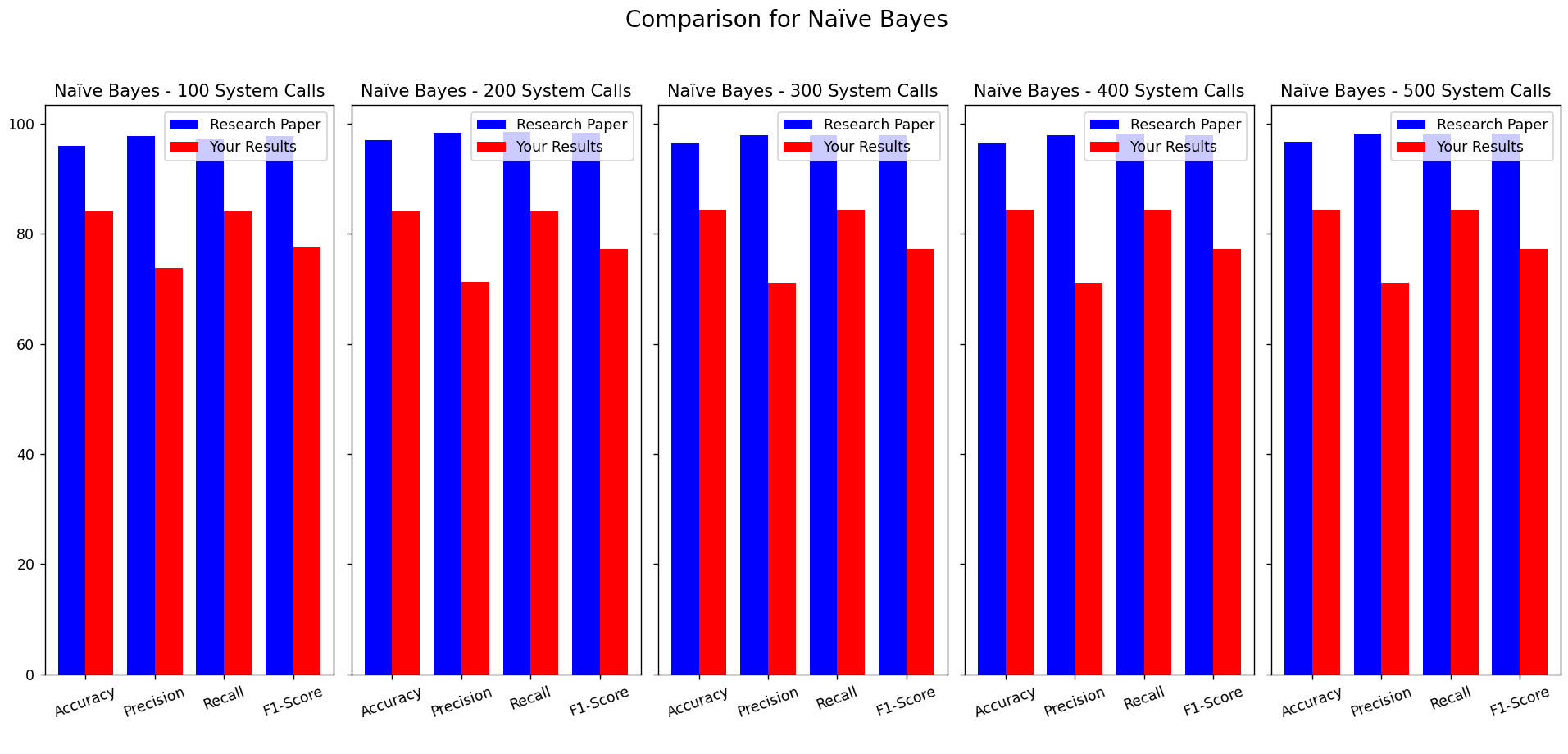


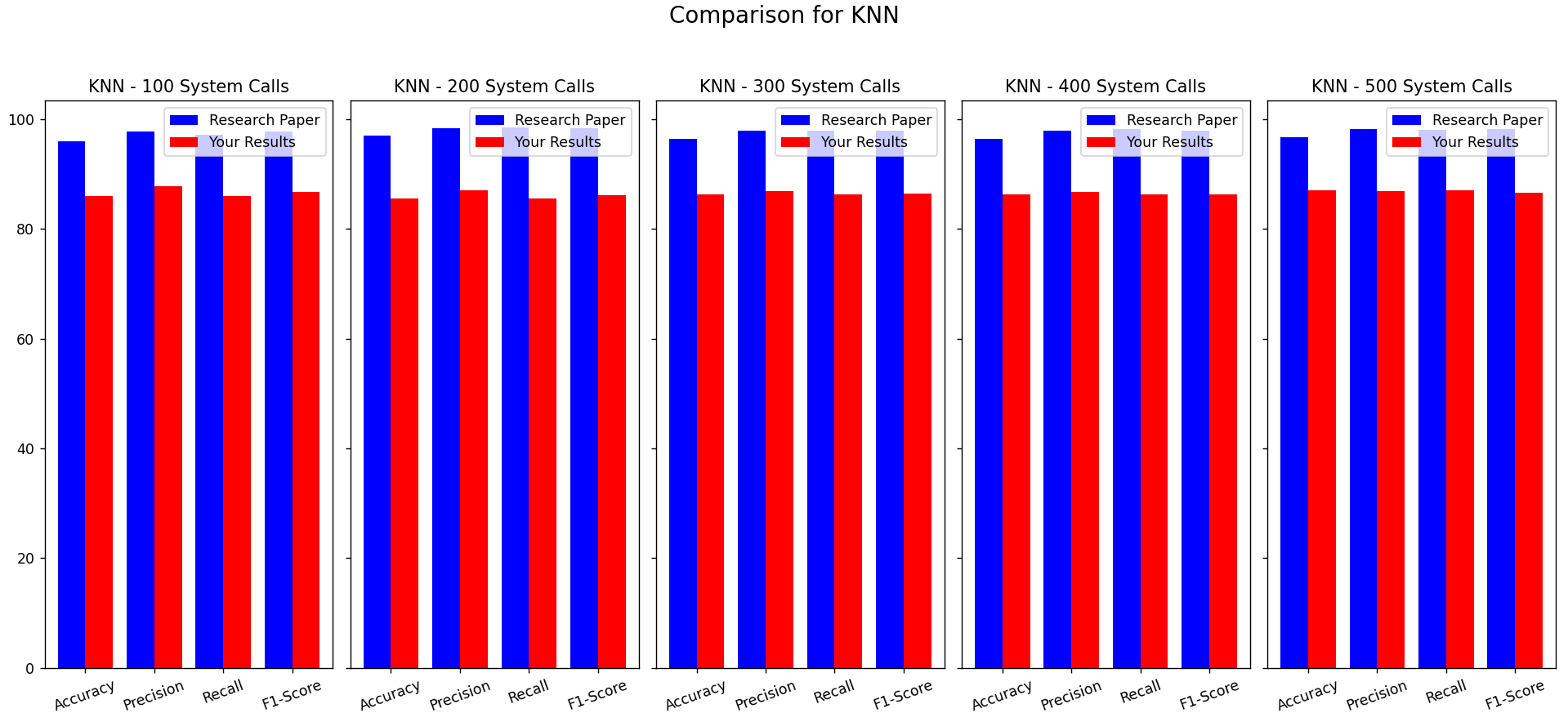
**Outputs Observed on ADFA-LD 2(Kaggle Dataset)**

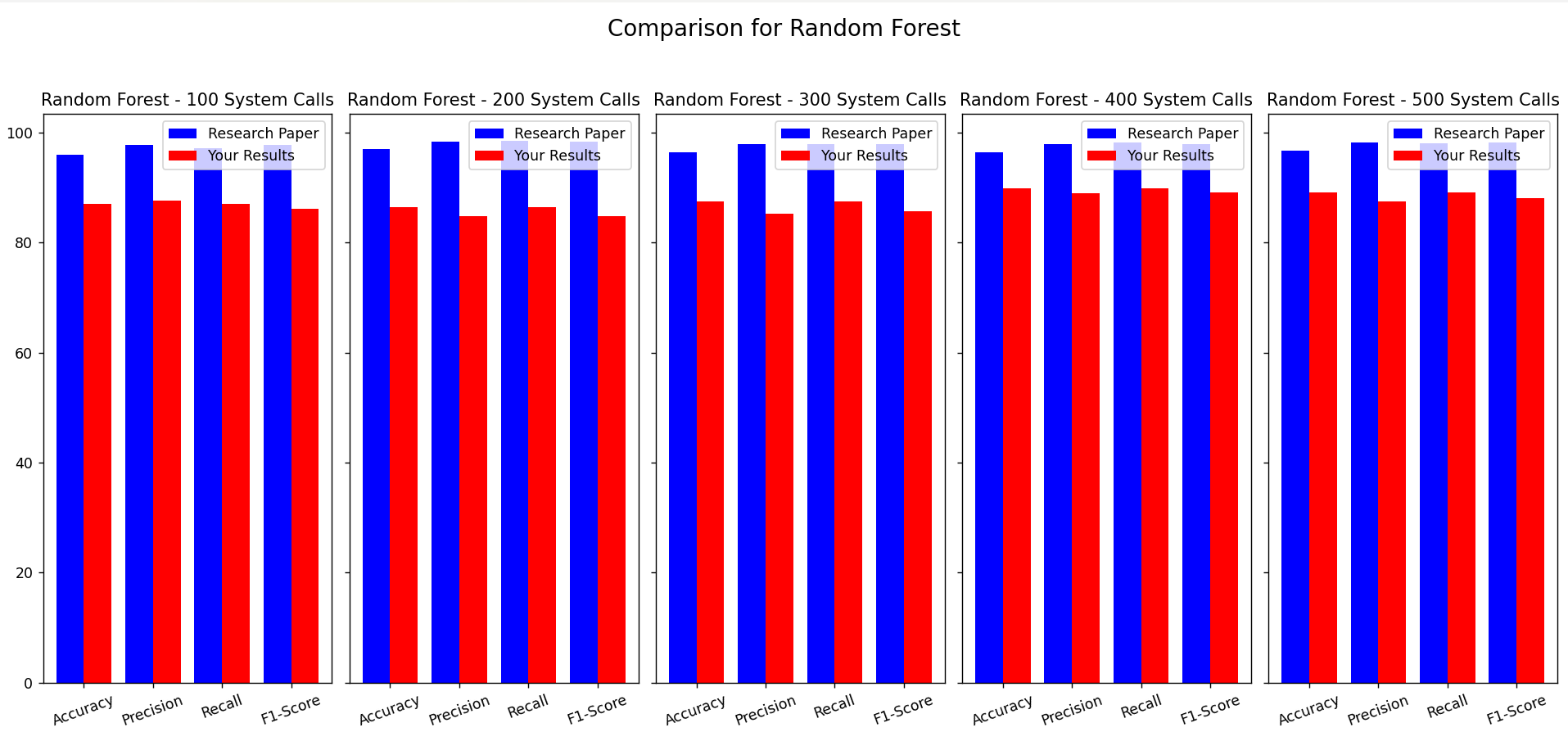
**MLP**

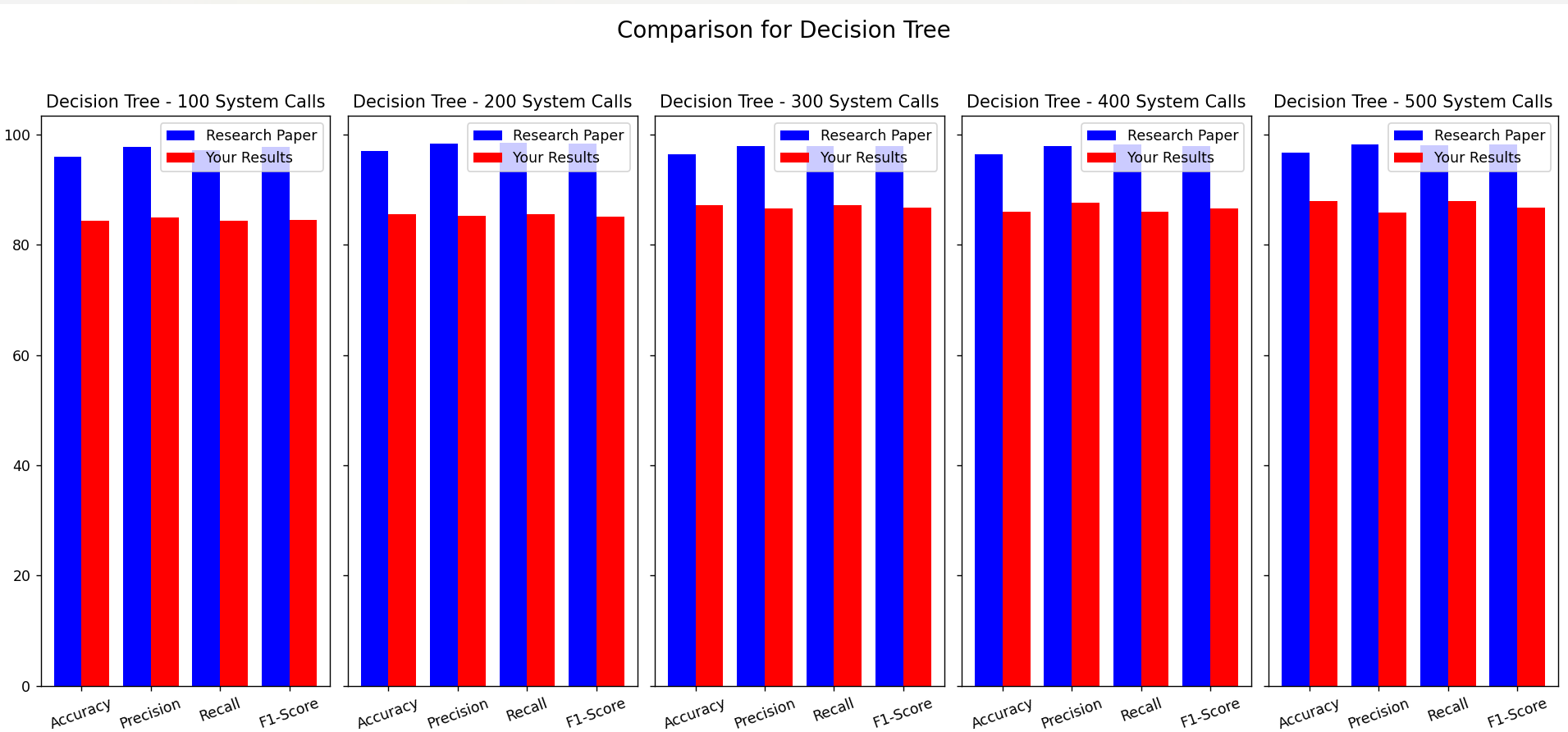


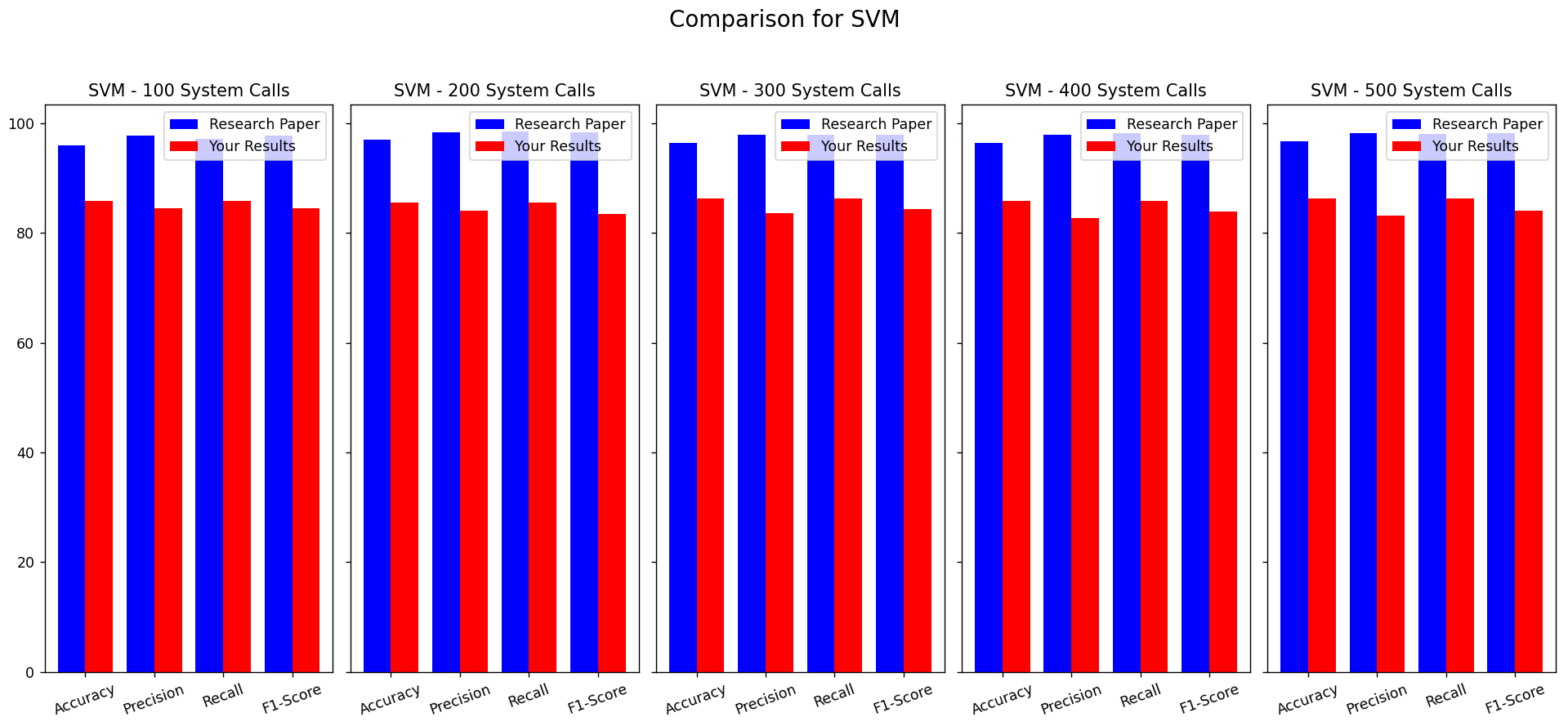
**MNB**



**KNN**

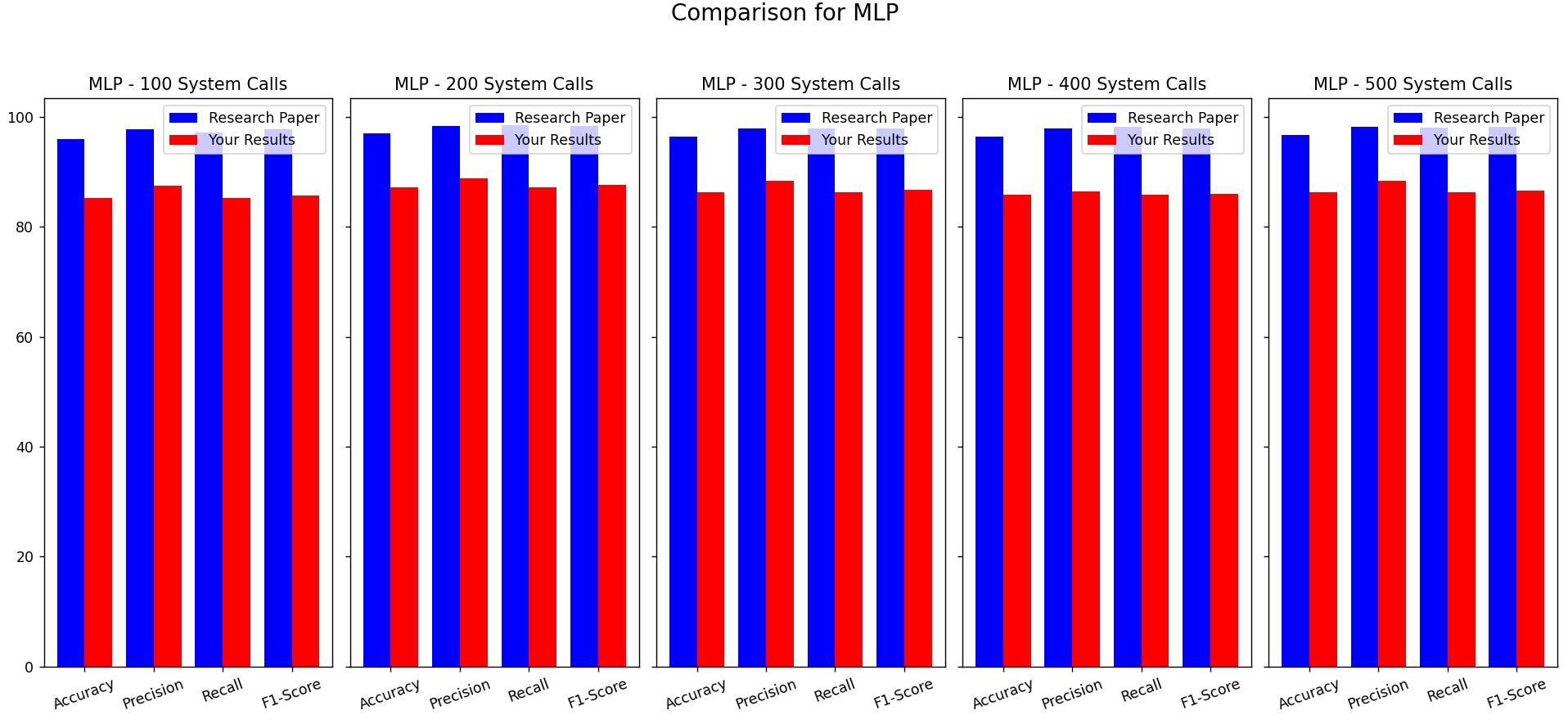
**RANDOM FOREST**

**DECISION TREE**

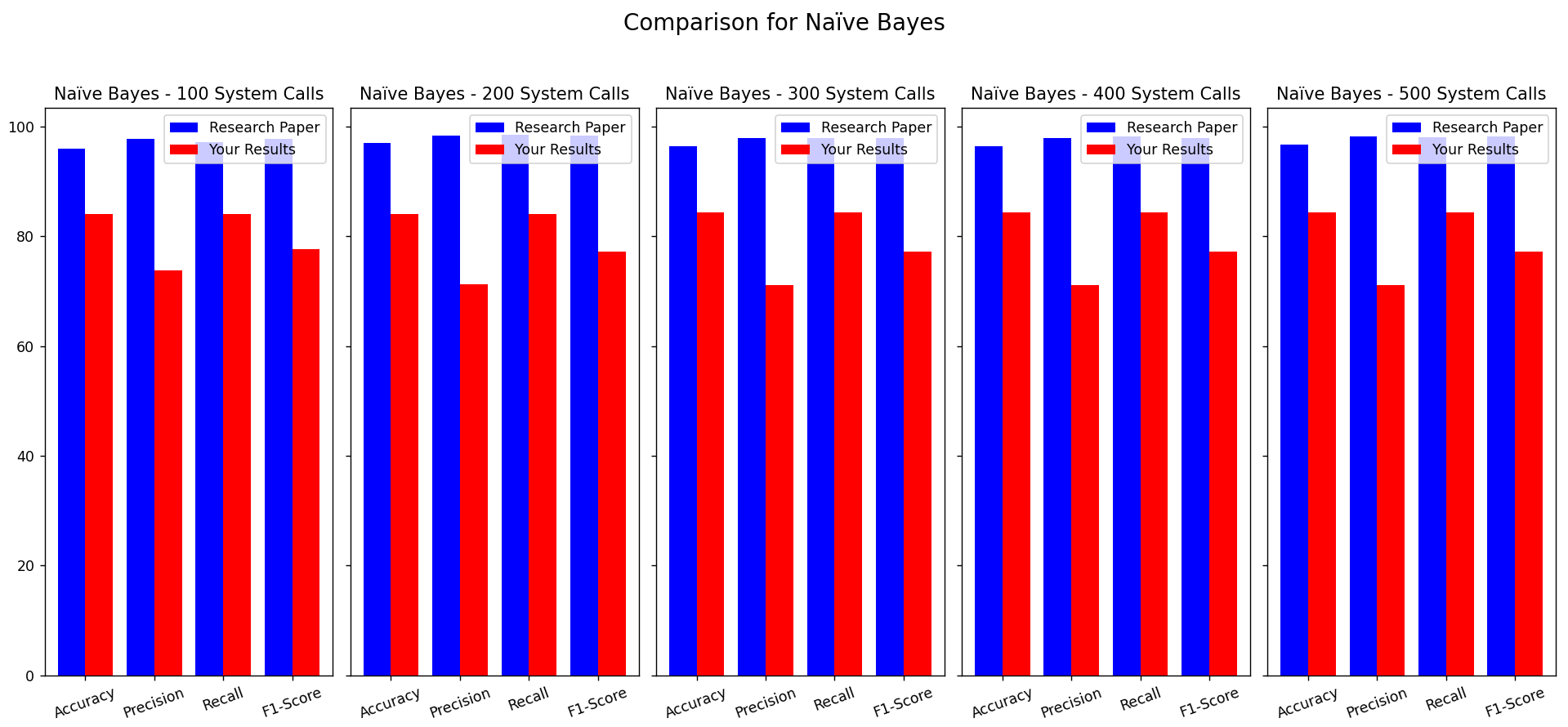
**SVM**

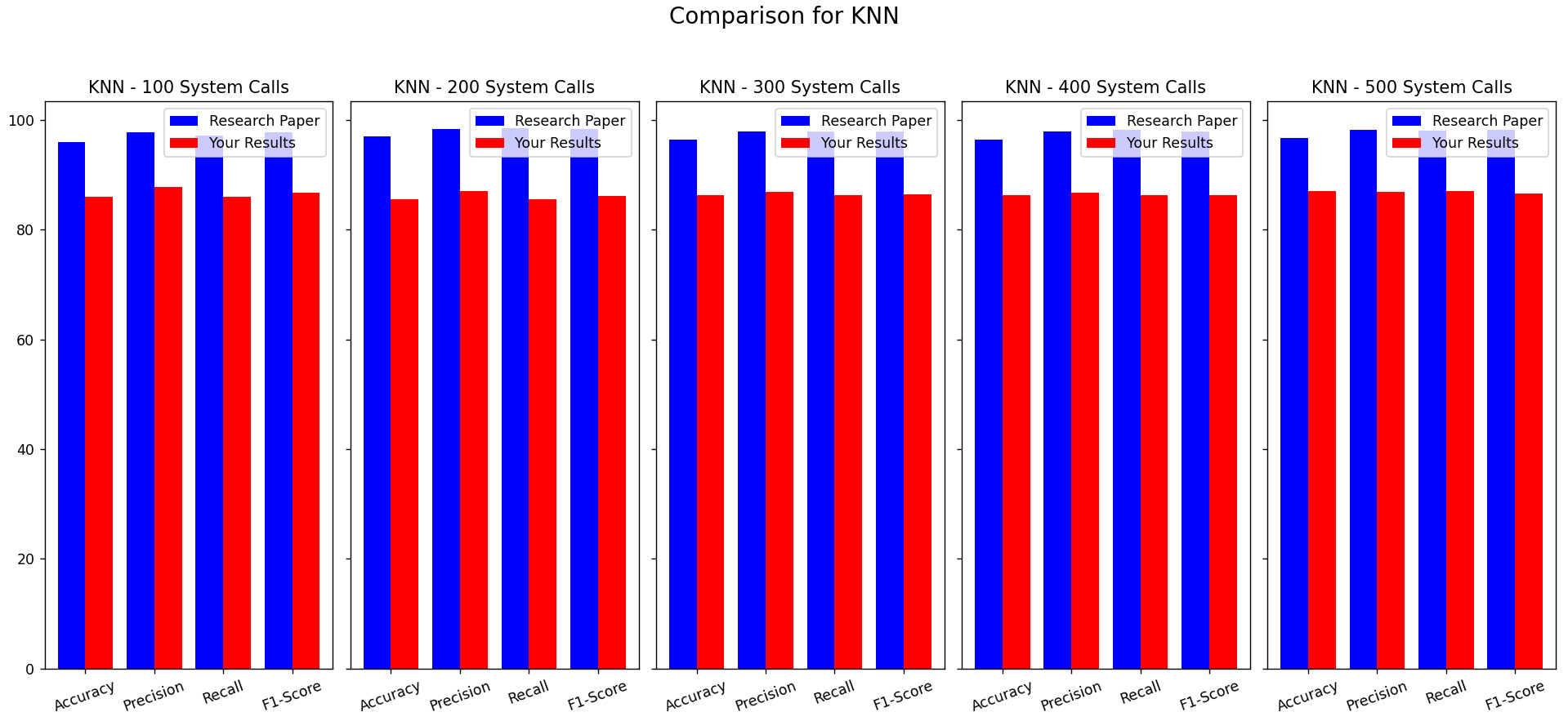
**Outputs Observed on ADFA-LD (Github Dataset)**

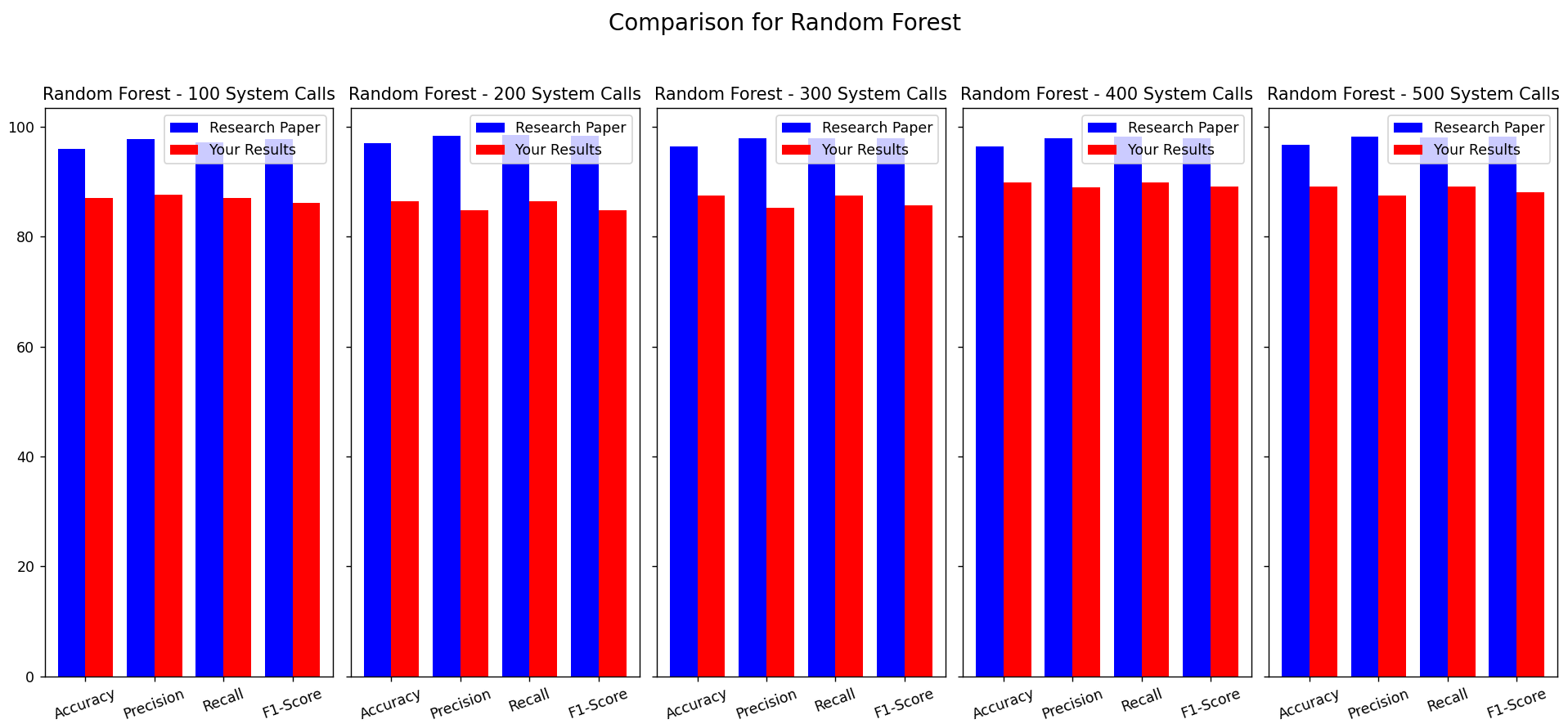
**MLP**



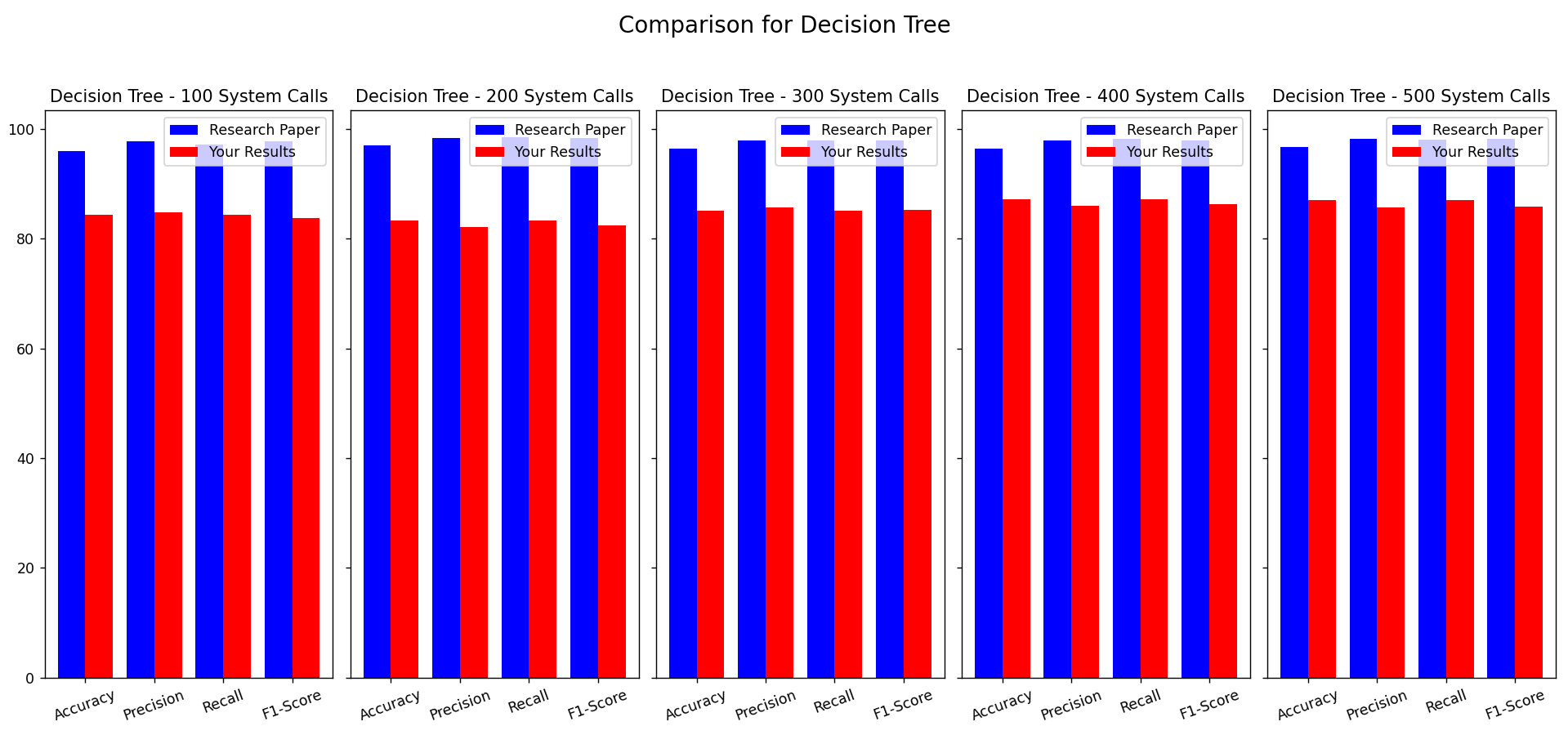
**MNB**



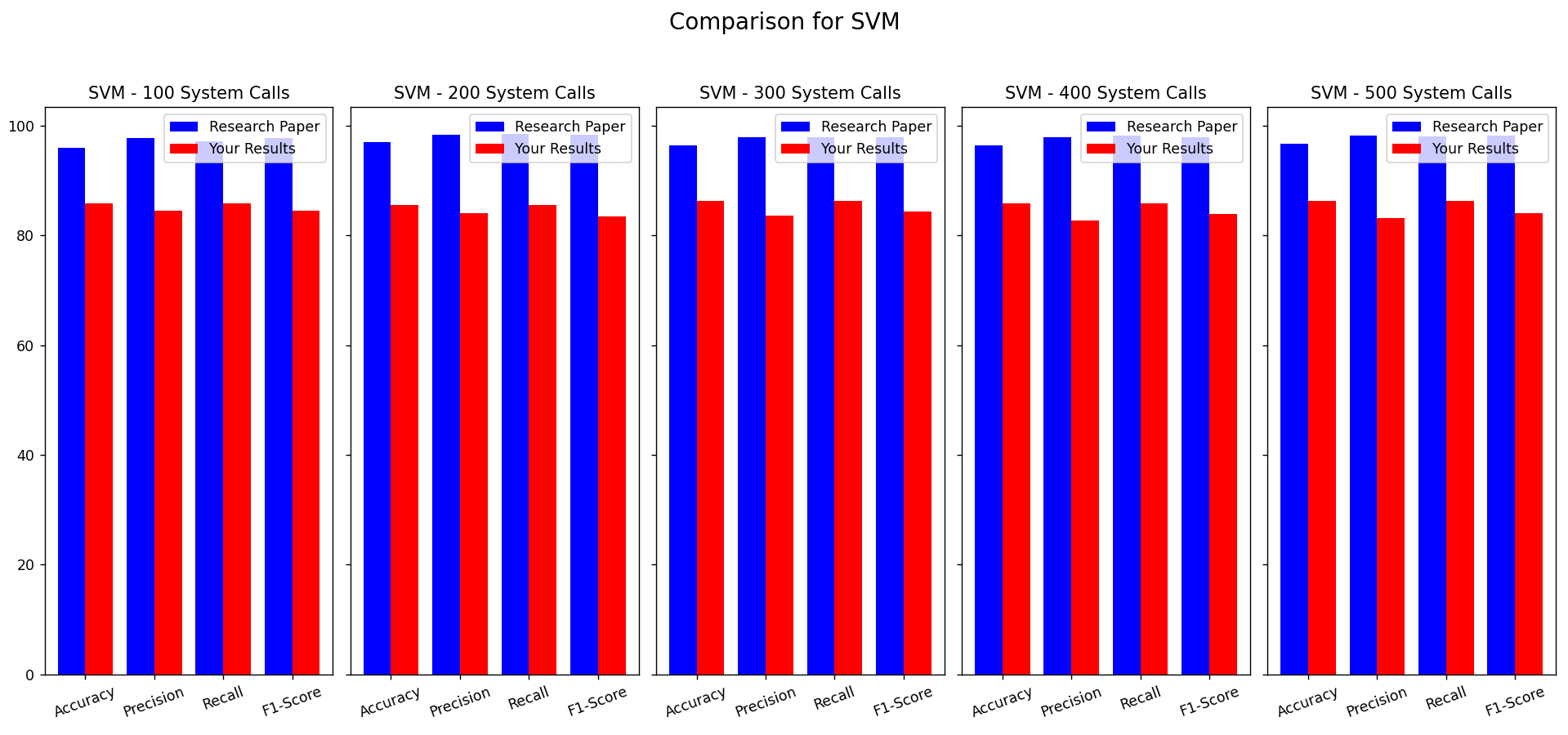
**KNN**

**RANDOM FOREST**

**DECISION TREE**



**SVM**



**1. Why Are Results Different?**

Your accuracy is much lower than what the paper claims (~87-89% vs. their 96-97%).

Possibility 1: You Missed a Step in Preprocessing

The research paper uses N-gram TF-IDF representation, but are you sure your feature extraction matches exactly?

Maybe the system call selection method differs (e.g., how traces are sampled).

Possibility 2: Parameter Differences

Check if your ML models' hyperparameters exactly match Table III.

Even a small difference in parameters (e.g., wrong C in SVM or different n\_estimators in RF) could cause this.

Possibility 3: Dataset Differences

Are you 100% sure your ADFA-LD dataset version matches theirs?

You got yours from Kaggle/GitHub—was their dataset modified for better results?

Check if file counts match their description.

3. What to Do Next?

Step 1: Compare Feature Extraction

Print your final dataset (first few rows) and see if it looks like N-gram TF-IDF vectors.

If not, fix it.

Step 2: Check Hyperparameters

Double-check every model’s settings match Table III.

Run grid search if needed.

Step 3: Compare Class Distribution

Print class-wise number of samples in your dataset vs. the paper’s dataset.

If there's a big difference, their dataset might have been balanced or modified.

Step 4: Try the Same Models Again

Fix any mismatches found in the above steps.

If results still differ, you might have just debunked a research paper!

Final Thought

If you fix any errors and still don’t reach 96-97%, you have legitimate proof that their methodology doesn’t generalize well. This is still a valuable research outcome!