Two concerns:

1. In three aspects of my approach, I have compared my results with 8 prior works that used different data sets. Those comparisons are not valid due to the use of different datasets.

2. The time difference (all features vs significant features) for execution time is not significant enough to claim credits.

Clarification of those concerns and my viewpoints:

Regarding the first concern:

It is totally unacceptable to compare works using different datasets. In order to compare works, the dataset and the features have to be the same. In Android malware detection, existing works deal with a set of feature types from a range of feature types (*Permissions, API Calls, System Calls, Intents*, etc.), like *Permissions and API Calls*, *Permissions and System Calls*, etc. Also, in one type of feature, there are many variations in the dataset as in each API level update, Android introduces new Permissions and API Calls. So, it is hard to reach a common ground.

I have compared 8 prior works erroneously in my thesis. But none of the 8 prior works compare their performance with other works. Only one of them (Aswini et al.) reported five prior works to justify that their performance is reasonable. Also, I looked through many other relevant papers and noticed none of them compare their performance with other works.

So, I did it really wrong: compared performance with other works that use different data sets.

Also, comparing performance is nothing to do with my research question and goal (see research question at the end).

Therefore, I have planned to omit those comparisons from my thesis as none of the prior works did so.

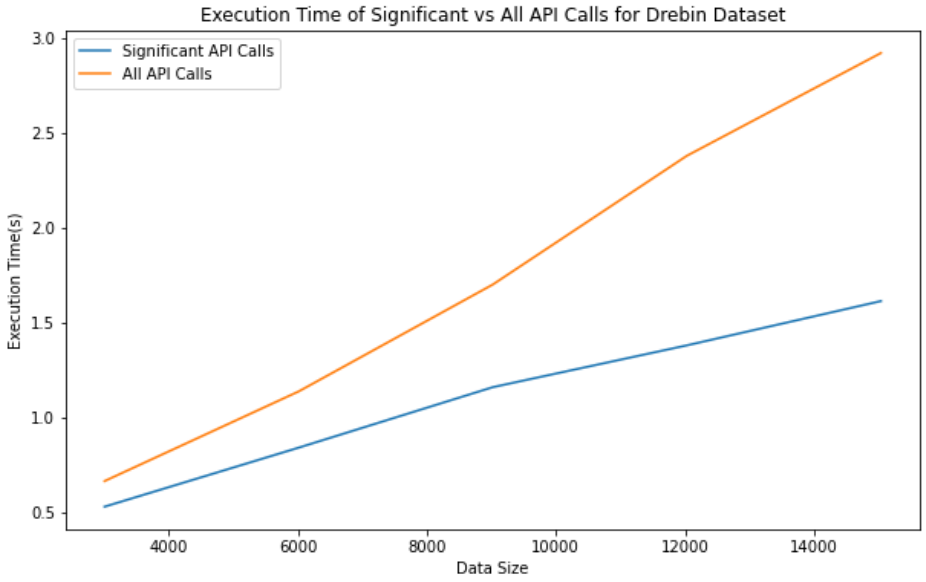
Regarding the second concern:

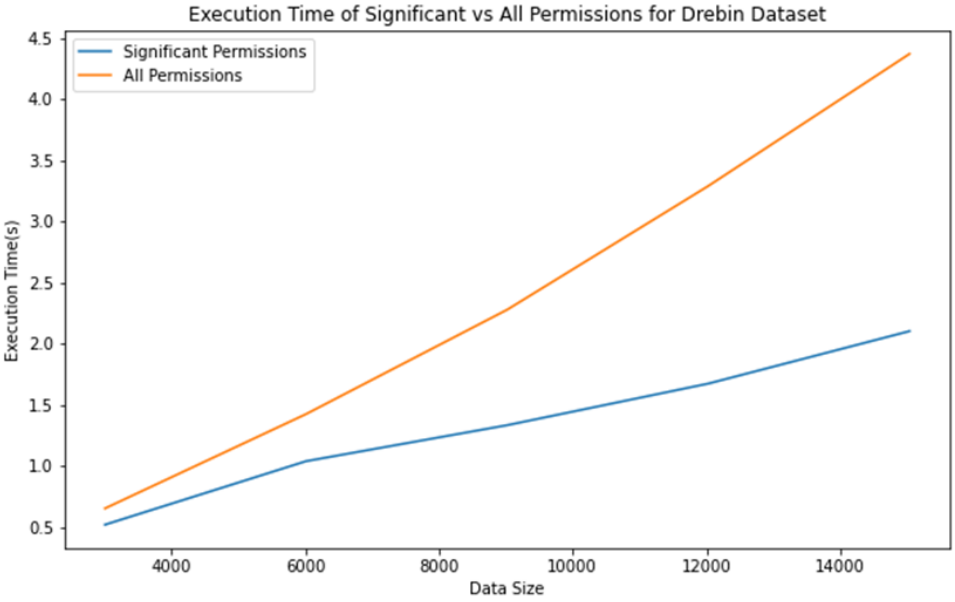
I totally agree with this concern. So, I have to justify my claim.

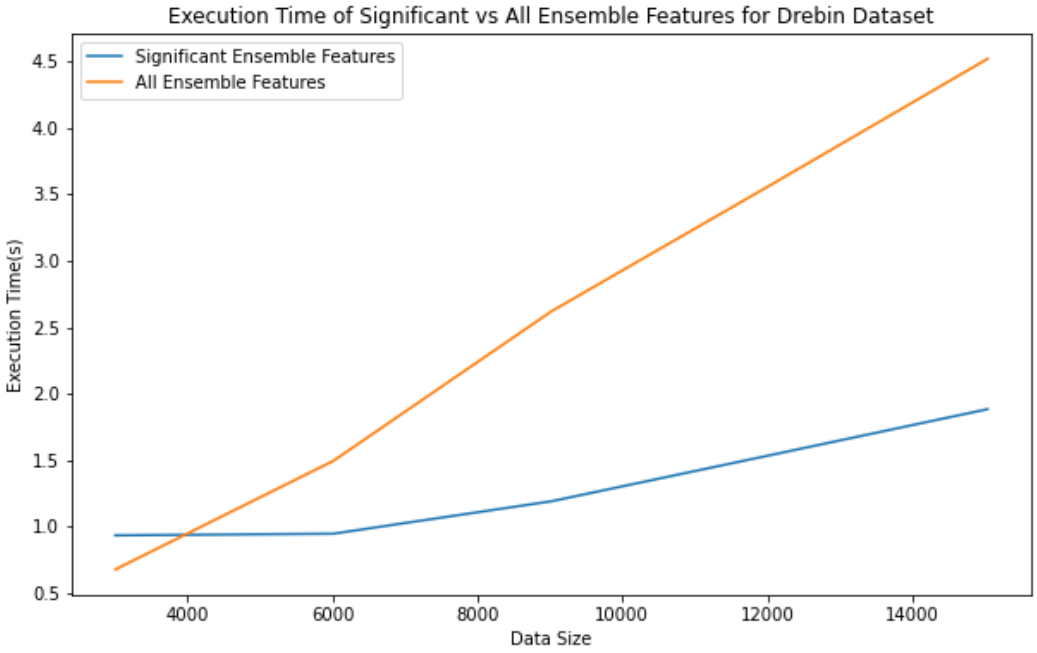
I have planned to analyze the difference in execution time by increasing the size of the datasets. Like, I would like to inspect the trend of the execution time with respect to significant features and all features by taking the portion of the data sets incrementally: 20%, 40%, 60%, 80%, and 100% of the data set. I have already experimented with it and would like to share it here.

For Drebin and Malgenome datasets, I have analyzed the execution time of significant API Calls, Permissions, and ensemble features separately with respect to all API Calls, Permissions, and ensemble features. In all cases, I have taken the average time after running it five times.

For the Drebin dataset, the execution time of significant features and all features is diverging (roughly) with the increase of dataset size in all three aspects (API Calls, Permissions, and Ensemble features).







For the Malgenome dataset, similar pattern occurs.

So, we can assume that for a larger dataset, the difference between significant and all features' execution time would diverge substantially.

Conclusion:

Apart from those two concerns, I would like to share one more thing regarding my thesis. I think I diverted from my actual research question and goal in the thesis defense and report. If you noticed my research question (I mentioned that below), my primary focus should be on significant features identification and their performance evaluation. I have identified and analyzed that. But, apart from that, I compared execution time and prior works that are not supposed to be compared. Because no prior work compared their performance with other works and compared execution time in this domain of work.

Research Question: How can we detect Android malware using significant features based on machine learning techniques?

> **SQ1:** How can we select the significant features for Android malware detection using machine learning techniques?
>
> **SQ2:** How do the significant features perform in Android malware detection?