Name: Asadullah Hill Galib

Exam Roll: 0702

# Response to Reviewer’s Comment
# (Dr. M S Rahman)

Thank you for your constructive review. Mostly, I agree with your comments. So, I have changed my thesis accordingly. In some cases, I would like to mention my reply to your comments. Here it is:

1. Related works mostly overlooked the importance of significant features **– why do you say that this is overlooked?** (Page-3, para 3)

- Because most of the works deal with all the features extracted. They do not employ any feature reduction or analyze feature importance. A few works deal with this issue. I have described and compared those works in my thesis.

2. **Explain how incremental assessment avoids predefined biases.** (Page-37, para 1)

- Several feature selection techniques are assessed incrementally to avoid any predefined biases. Typical feature selection approaches use predefined numbers of features or predefined values, such as threshold, algorithm parameters, etc. to filter out features. All features are evaluated progressively to prevent this bias and to define the minimum number of features. It helps to evaluate performance for a number of features from 1 to n and to choose the minimum number of features that generate

3. Correlation-based Feature Elimination - **I believe this should have been done as the first step. Why haven’t you done it earlier?** (Page-44, heading)

- Yes, I agree. This should have been done as the first step. But my plan was to check the performance of the features set before and after correlation. In doing so, I planned to reduce features with any feature selection technique, then applied correlation-based feature elimination. If I did correlation-based feature elimination in the first place, I had to eliminate features with a predefined correlation threshold which might remove many features without inspecting their importance in malware detection. That is why I did it later, to ensure no important feature is eliminated due to it. As I did it in the later stage, I was able to compare the performance before and after the correlation-based feature elimination (which is computationally simple enough as working with the minimal number features only) which ensured usability of this step.

4. **Why did not you use the ranking in the previous stage?** (Page-45, para 3)

- Yes, I agree. Actually, the ranking is used in the previous stage too. Using it again in the later stage is redundant. So, I changed my writing accordingly.

5. **Where is weka tool used?** (Page-50, para 2)

- Weka tool is used initially for data preprocessing (missing value treatment) and correlation calculation. But later I implemented those parts using python in Jupyter Notebook. So, now I removed the reference of weka from my report.

6. **Since this is RFE would not it be better to horizontally flip these figures?**(Page-52,figure 1)

- I have used RFE in an incremental fashion. Like, I evaluated performance for 1 to n number of features as RFE'sargument (number_of_estimator's). So, it is like RFE with number of estimators to select: 1 feature, 2 features, 3 features, …. n features.

7. **So one of your comments/ hypothesis earlier that: "These growing numbers of features make it complex and would misguide classifiers by over-fitting of data." is not valid in this case.** (Page-58, table)

- Yes, it is not valid in this case. So, extract that comment from my thesis.

8. **How large will be the dataset in a practical case? You need to put your claim in the right context. ~3 secs difference is not a big deal.** (Page-59, para1)

- I have read some papers which use large dataset in their approach. For instance, Marvin (2015) use 150,000 apps, Mobile-Sandbox (2013) use 40,000 apps. I have used 15,000 apps. So, I think in practical case, the dataset can be large enough. And, yes, I agree that, that 3 seconds difference is not a big deal.

9. **Are you saying they found 169 and 1326 as significant?** (Page-59, table 2)

- Yes, according their paper.

10. **Execution times?** (Page-75, table)

- The execution time for those related works is not reported.

11. **What are the execution times of other methods who have all features?** (Page-74, table-2)

- The execution time is not reported in the related works.