|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| I sincerely thank the reviewers for their valuable feedback. I have addressed all the comments in the revised work. I have incorporated the suggestions given during the Mock presentation, which was made on May 22nd, 2025.  Submitted By,  Scholar: Chaithanya B N,  Roll no:321860304001  Under the guidance of  Dr. Renuka C Herakal,  Assistant Professor,  Department of CSE,  GITAM School of Technology,  Bengaluru. | | | | |
|  | **Suggestions** | | | |
| **1** | **Change the title: Design of an Adaptive Technique for Ransomware Detection and Severity Analysis Using Semantic Similarity and Hybrid Machine Learning Approaches** | | | |
|  | An Adaptive Framework for Ransomware Detection and Severity Analysis Using Semantic Similarity and Hybrid Machine Learning Approaches | | | |
| **2** | **How do you identify priorities among three thresholds? What is the minimum threshold?** | | | |
|  | Three defined thresholds (Low, Medium, and High) are present in each similarity algorithm. Threshold 3, which indicates a high degree of similarity with known dangerous families, is given priority. For Threshold 3, the entry boundary is the lowest threshold (e.g., Levenshtein ≤150). A sample is classified as High if multiple algorithms place it in Threshold 3, Medium if the majority is in Threshold 2, and Low otherwise. Even if one metric performs poorly, the robust severity prediction is guaranteed by this multi-algorithm voting. | | | |
| **3** | **What Happens if token and keyword value increases?** | | | |
|  | The results confirm that increasing token length and keyword dimension beyond optimal points leads to diminished model performance. The best configuration is found at 200 tokens and 200 keywords, where the LSTM effectively captures sequential behavioral patterns of ransomware without overfitting or sparsity. Increasing dimensionality beyond this threshold adds noise and reduces generalization, as reflected in rising loss and falling accuracy at 300-token setups | | | |
| **4** | **Two Techniques are necessary to use which give the same results in the end?**  **Don’t you think it will affect the model?** | | | |
|  | Although TF-IDF and Doc2Vec both serve as feature extraction methods, they are not redundant. TF-IDF captures lexical importance, while Doc2Vec captures semantic relationships. We apply both techniques independently and compare performance across classifiers. This multi-view representation strategy increases robustness and allows the framework to handle both word-based and context-based variations in ransomware/phishing emails. | | | |
| **5** | **What is the purpose of using LSTM in Objective 1, SVM, LR, RF, XGBoost in Obj 2, and hybrid in Objective 3?** | | | |
|  | Objective 1  (ASM-based ransomware detection) | LSTM | We use LSTM because ASM is a sequence, and LSTM can learn opcode transitions and malicious flow logic better than classical ML models. | |
| Objective 2  (Email classification: ham/spam/ransomware) | SVM, LR, RF, XGBoost | * Email content can be represented as high-dimensional, sparse feature vectors (TF-IDF, Doc2Vec). * Classical ML models like SVM and Logistic Regression perform well on linearly or nearly linearly separable data. * RF and XGBoost handle non-linearities, interactions, and feature importance ranking. * We use a diverse set of classifiers to evaluate how different learning biases handle textual and contextual features in emails.” | |
| Objective 3  (Severity estimation using similarity + ML) | Hybrid Model via Soft Voting + Semantic Similarity |  | |
| **6** | **Why is ASM (Assembly) considered a time series?** | | | |
|  | Because .asm (assembly) files represent a sequence of instructions executed in order, they exhibit the sequential dependencies and positional characteristics that are core to time-series data. | | | |
| **7** | **Architecture diagram of 3rd Objective** | | | |
|  |  | | | |
| **8** | **RMSprop** **Optimizer Equation correction** | | | |
|  |  | | | |
| **9** | **Justify how the model is scalable** | | | |
|  | * Accepts .asm, .mbox, .csv, etc. — handles diverse data types. * Classifier, similarity, and decision logic are independent modules — new models or formats can be plugged in easily. * The decision logic can scale to new ransomware families by adding more thresholds. | | | |
| **10** | **Why an Old Dataset (Microsoft BIG 2015) is Used** | | | |
|  | * The dataset provides reliable labeled data, enabling robust training and benchmarking of ransomware detection models. * Most modern datasets do not provide instruction-level disassembly, making BIG 2015 uniquely suitable for behavioral sequence modeling. * Foundational behaviors like API injection, obfuscation, and encryption logic still hold across generations, making BIG 2015 behaviorally relevant. | | | |
| **11** | **Graph update with labels(t\_SNE)** | | | |
|  |  | | |  |
|  | | |  |