IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Optimized Cybersecurity Solutions: A Multi-Algorithm Ransomware Detection Framework

[1]Tripurapu Bhavya, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions

[2]Dr. T. Ravi Babu, Associate Professor at Miracle Educational Society Group of Institutions

[3]Ravi Nava Ratna, Assistant Professor at Miracle Educational Society Group of Institutions

ABSTRACT

This paper describes a professional system designed to detect cases of ransomware assaults based on data related to processor usage and disk usage. System processes and the activities dealing with files are some of the classical methods used, but at times power effectiveness is compromised and the methods are not very reliable. In order to compensate for these, VMware environment is used in this work in order to obtain HPC as well as I/O events without degrading performance. The machine learning algorithms that were employed in the evaluation of the model included SVM, Random Forest, and XGBoost, where Random Forest and XGBoost achieved 98% accuracy. In addition, other DNN and LSTM deep learning models were applied, and an extension with CNN2D was reported with most accuracy of 98.83%. This self-learning system of detection has changed the way ransomware detection is done without compromising the performance of the system in a big way and is an effective means of dealing with cyber threats.

Keywords:CNN2D, DNN, LSTM

INTRODUCTION:

The era of modern cyber threats have been described by the unprecedented in the evolution of the ransomware attacks known as a ransomware as a service or RaaS where the organizations are able to hire professionals for a price to hack their competitors. Ransomware is a term used to describe a form of malware that is used to encrypt files and operate a system and demand a ransom in exchange for running those files, other instances involve cyber criminals exploiting the ransomware for financial purposes. Ransomware is predicted to be a threat that will occur every 2 seconds in the year 2031, starting from 20 billion of damages in the year 2021 and increasing to 265 billion in 2031. The attacks are modern ransomware attacks which include an exfiltration of the data and involve destroying the critical infrastructure by a stat actor, after which they are persuaded to pay the ransom or their data will be leaked for public access. Poly or metamorphic variants are the hashes that will be stored in the form of documents that will be used to determine the parameters for the detection based on the signature versus the common perception based on the signature, which uses the awareness was a prevalidated approach to the classical machine learning methods. With the help of machine learning models, ransomware

protection and concealment can be performed in real time analyses so that the resistance of the system is improved. The fundamental difference between applications and the ransomware is the higher disk and processing usage that allows the detection of the segmenting of encrypted files as a result of the sophisticated cuffs such that zero insertion costs are added thus doomed to fail completely the cognitive analysis

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

The evidence gap that has been established confirms to pervade research in the following ways that existing processes of detection and tracking the ransomware compromises the overall performance of the system. The remainder, sufficient feature optimization and scale that seem borders the Ransom Flower Drift and SVM have also been supported.

• Although deep learning techniques are encouraging in nature, they do not possess a Lacunae in literature for established Feature Extraction approaches.

Key Gaps:

1. Resources Consumption: Existing conventional detection mechanisms are resource hungry hence low practical applicability in real life scenarios.
2. Low Feature Extraction: The methodologies applied currently do not take advantage of sophisticated feature deduction algorithms or optimization schemes.
3. Low Coverage for Sophisticated Models: Very few studies look at employing say a CNN or any modern deep learning model with an aim of improving the accuracy.
4. Limited Dataset Scope: The majority of the studies are limited to using a few datasets making it difficult for the models to be generalized over a wide range of Ransomware strains.
5. Absence of Benchmark Testing: Severe lack of systematic assessments of advanced and conventional algorithms so as to know which is more effective.

PROBLEM STATEMENT:

Ransomware has attacked lots of organizations in the world resulting to loss of funds by encrypting the data of the victim and asking to pay for it to be decrypted. Many conventional detection strategies such as file system activity monitoring and file activity logging cannot achieve low computational strain and high accuracy performance.

Key Challenges:

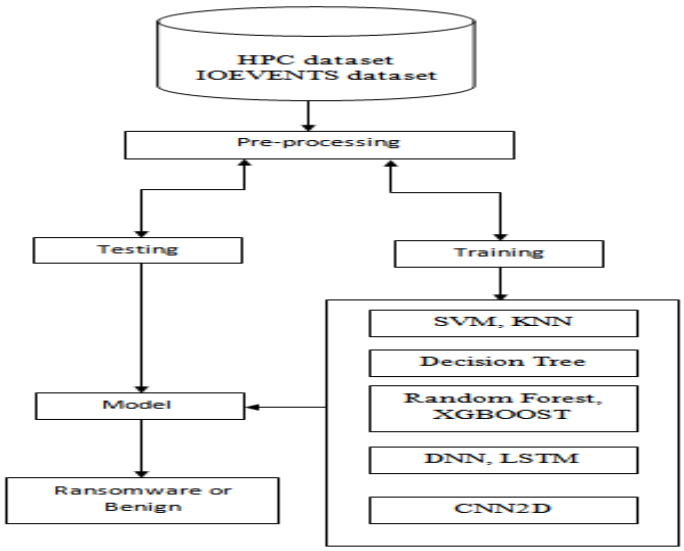
1. Adverse Effect on System Performance: Existing strategies make use of conventional methods which reduce the performance of a system during the detection phase.
2. Low Detection Rate: Current approaches are not effective to classify the different ransomware scripts with sufficient accuracy.
3. Time Sensitive Ransomware Identification: Finding a Consistent method of identifying a ransomware in order to deploy a reaction to it in a timely manner seems to be difficult.
4. Wider Applicability: Testing for effectiveness of solutions in different environments and data sets.

5. Feature Optimization: The absence of advanced techniques to optimize dataset features impacts the efficiency of the models.

PROPOSED METHOD:

This work contributes a novel concept of detecting ransomware using processor and disk usage data during operation which was collected using VMware. Hardware Performance Counters (HPC) and I/O events are gathered in a non-intrusive manner allowing for precise classification of the data. The methodology attempts to use traditional machine learning algorithms such as SVM and Random Forest as well as deep learning approaches such as LSTM and DNN. The subsequent extension with the use of CNN2D allows for better feature optimization which brought the accuracy to 98.83%. Normalization, shuffling and splitting into training and testing data sets is what data pre-processing entails. The performance metrics used were accuracy and confusion matrices to validate the models which have shown CNN2D as the best transfer learning method. Stability of the systems which improves the overall performance in detecting ransomware applications makes the combined framework a viable option for use in the cybersecurity industry.

ARCHITECTURE:



DATASET:

The dataset includes data derived from different applications that are 7ZIP and AES. It integrates information from Hardware Performance Counters (HPC) and I/O events. There are several feature columns that assist in indicating the activity of the system, and the entire dataset is labeled as binary (0 for Benign and 1 for Ransomware). Pre-processing entails procedures such as normalization of features, shuffling of data, and partitioning of the data into 80% for training and 20% for testing. The data set has been designed in such a way that there is no class imbalance between the benign and ransomware samples allowing the models to be properly evaluated. It is freely available as it offers the public unlimited access creating some sort of repository that can be highly utilized for training models especially machine learning algorithms and deep learning algorithms aimed at ensuring ransomware is detected at the least of the systems breathing space.

METHODOLOGY:

1. **Data Collection and Integration:**

   o Retrieved datasets regarding Hardware Performance Counters (HPC) and I/O events from databases that are open to the public.

   o Sought to build a bigger dataset by integrating features from different scripts, such as 7zip and AES.

2. **Data Preprocessing:**

   o Ensured uniformity by adjusting the feature values using normalization.

   o Randomly mixed the dataset to eliminate training bias.

   o Partitioned the dataset into 80%, which is used for training, and 20% that is used for testing.

3. **Feature selection:**

   o Collected relevant information from HPC and I/O events data to make it possible to tell the difference between a benign program and a ransomware program.

   o Leveraged VMware to perform data grabbing tasks while still maintaining acceptable levels of overload on the systems used.

4. **Training the Models using the Conventional Algorithms:**

   o SVM: Was able to get an accuracy of 88%. This model seems to be moderately effective.

   o KNN: Was able to achieve 97% accuracy which proves that this algorithm is strong in classification.

   o Decision Tree: Was able to achieve an accuracy score of 93%, this model can be interpreted.

   o Random Forest: On the other hand, this model achieved a robust classification performance of 98% in accuracy.

   o XGBoost: Matched Random Forest,, which in turn scored an accuracy rate of 98% as well.

5. **Training Deep Learning Models:**

   o DNN: Was trained with HPC and I/O events data, which was able to achieve an accuracy of 88%.

   o LSTM: A sequential processing approach reached an accuracy of 93%.

6. **Extension with CNN2D:**

   o Hooked CNN2D up with multi-convolutional layers to hook up an advanced feature extraction.

   o Achieved a classification accuracy of 98.83% by reducing feature duplication.

7. **Performance Evaluation:**

   o Used model accuracy, confusion matrices and performance graphs to evaluate and validate the models.

   o Analysis of the algorithms made use of bar graphs. CNN2D was the most effective algorithm according to their graphs.

8. **Real Time Detection Validation:**

o The differ- ence was in the manner in which the data were applied – CNN2D was employed on the test data . The model was able to precisely foretell whether the scripts were benign or ransomware.

o The work exhibited practicality by having the same outcomes on data that were not seen before.

9. **System Implementation:**

o All algorithms were coded into the jupyter notebook which facilitated smooth running of the programs.

o Outputs were designed considering the applications for cybersecurity users.

EVALUATION:

Precision:

$$\text{Formula: Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Recall (Sensitivity):

$$\text{Formula: Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

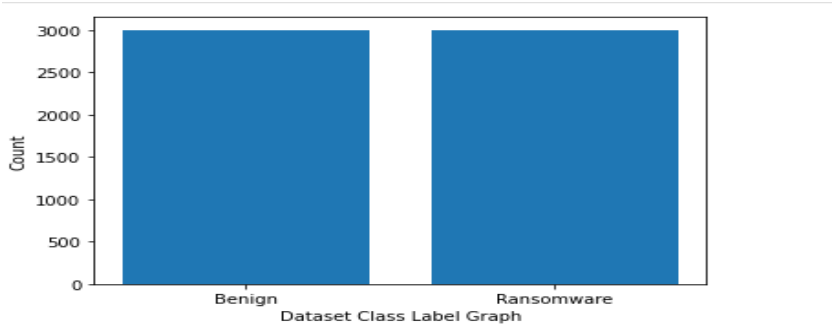
F1 Score:

$$\text{Formula: } F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

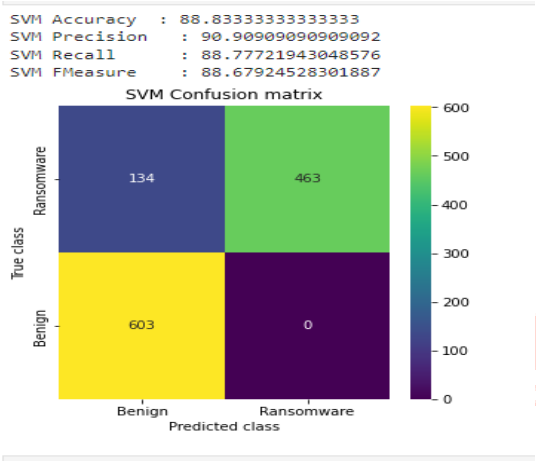
Accuracy:

$$\text{Formula: Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

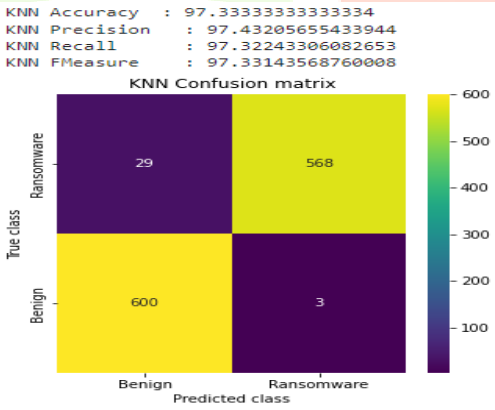
RESULTS:



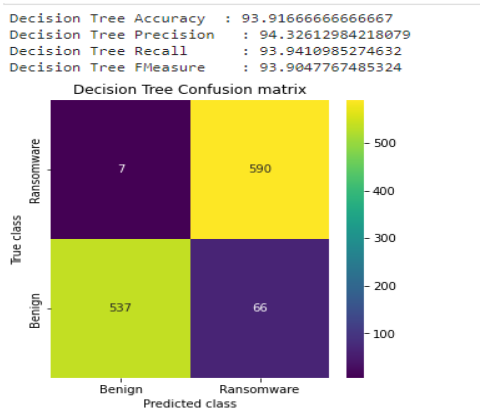
Graph of benign and Ransomware dataset size where x-axis represents class label and y-axis represents count



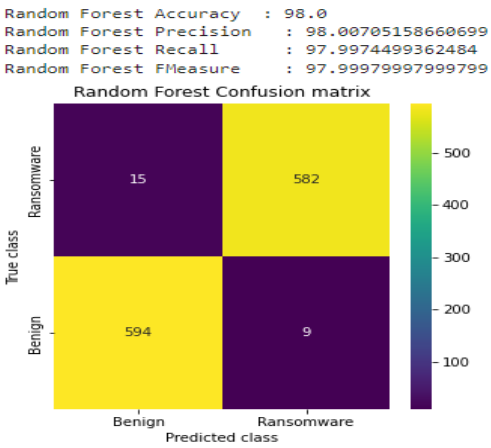
Training SVM algorithm on 80%



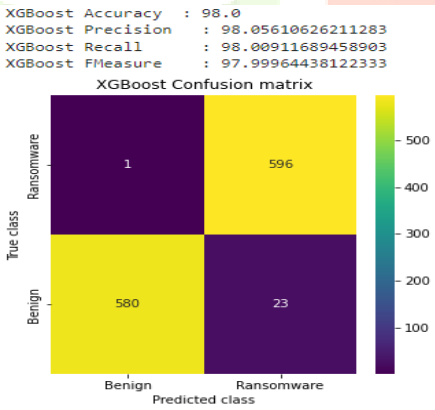
KNN got 97% accuracy



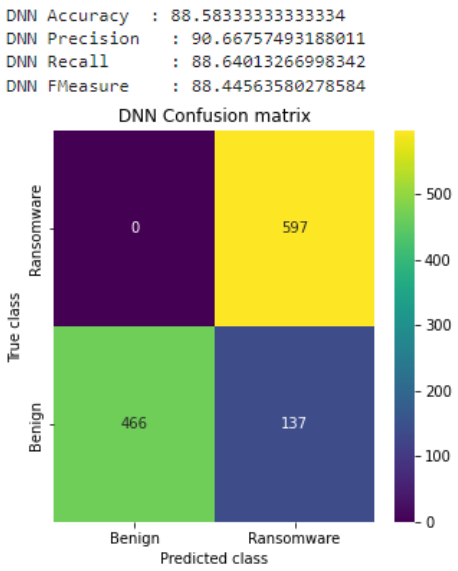
Decision tree got 93% accuracy



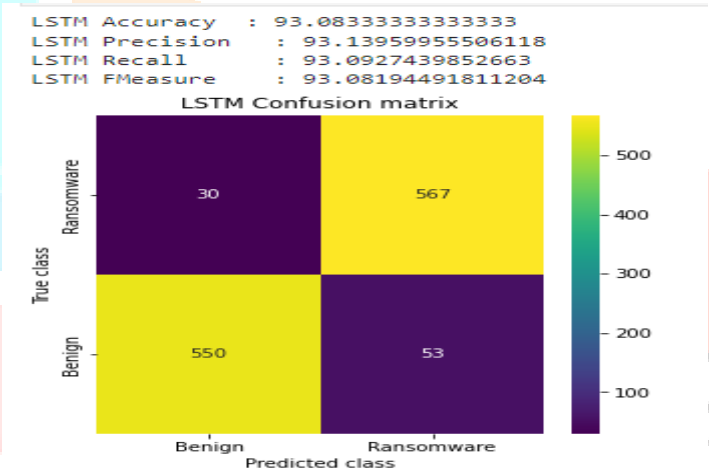
Random Forest got 98% accuracy



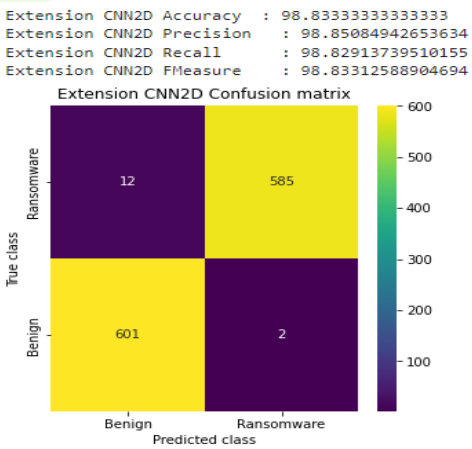
XGBOOST also got 98% accuracy



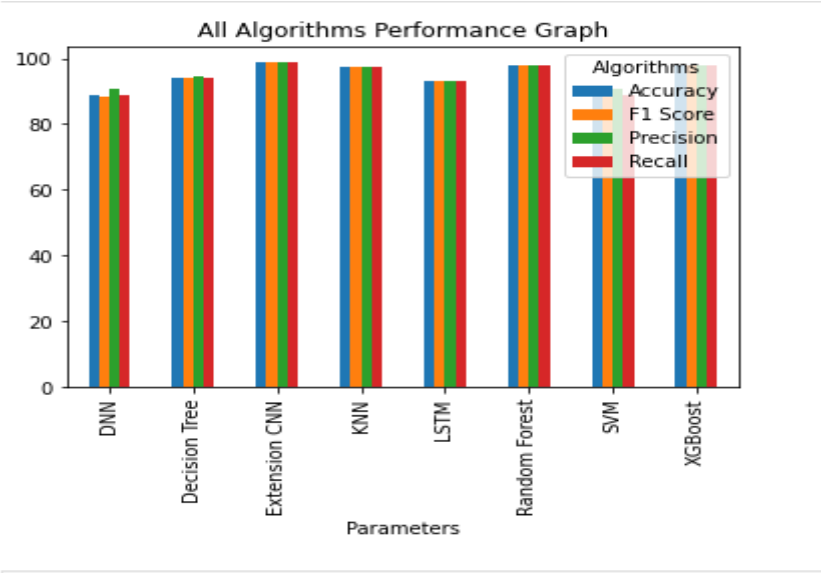
DNN got 88%accuracy



LSTM got 93% accuracy



CNN2d got 98.83% accuracy and in all algorithms extension CNN got high accuracy



In all algorithms Extension CNN got high accuracy

Prediction:

```
Test Data = [2.1506087e+07 8.5300000e+03 8.6983600e+05 5.8991000e+04 3.0000000e+00
 4.0000000e+00 2.4576000e+04 4.4000000e+01 1.3844480e+06 0.0000000e+00
 4.6274600e+05 1.0921356e+07 0.0000000e+00] Predicted AS ====> Benign

Test Data = [2.4040858e+07 6.2620000e+03 1.0111480e+06 8.8860000e+04 2.0000000e+00
 3.5000000e+01 7.9462400e+05 1.2000000e+01 2.4985600e+05 0.0000000e+00
 3.3243050e+06 1.2417450e+06 0.0000000e+00] Predicted AS ====> Ransomware

Test Data = [8.6719703e+07 2.3200000e+02 3.1320000e+03 2.0025900e+05 0.0000000e+00
 0.0000000e+00 0.0000000e+00 0.0000000e+00 0.0000000e+00 0.0000000e+00
 0.0000000e+00 0.0000000e+00 0.0000000e+00] Predicted AS ====> Ransomware

Test Data = [8.6837894e+07 1.4700000e+02 3.7560000e+03 2.1272200e+05 1.0000000e+00
 0.0000000e+00 0.0000000e+00 0.0000000e+00 0.0000000e+00 0.0000000e+00
 0.0000000e+00 0.0000000e+00 0.0000000e+00] Predicted AS ====> Ransomware
```

Predicted values as ‘Ransomware or Benign’.

CONCLUSION

This project presents an effective ransomware detection framework that employs VMware for advanced data extraction and machine learning models. Using HPC and I/O events data, the method ensures high accuracy with minimal compromise to system performance. Traditional models such as Random Forest and XGBoost performed well with an accuracy level of 98%, while CNN2D was found to do even better with a staggering 98.83%. The method deals with the important issue of feature optimization and the cost of computations ensuring a robust and flexible approach in cyber defence. By making sure that the system resources are not overtaxed while getting accurate detection of ransomware, this framework provides a standard for measuring how effective future studies are against ransomware attacks.

REFERENCES:

[1] SR Department. (2022). Ransomware victimization rate 2022. Accessed: Apr. 6, 2022. [Online]. Available: https://www.statista. com/statistics/204457/businesses-ransomware-attack-rate/ While our model limits its applicability to VMs, we plan to adapt it to stand-alone machines in our future work. We have not evaluated whether the models developed for a machine configuration work well for another machine configuration, such as increased memory or more CPU cores. We plan to investigate this in the future.

[2] D .Braue. (2022). Ransomware Damage Costs. Accessed: Sep. 16, 2022. [Online]. Available: https://cybersecurityventures.com/globalransomware-damage-costs-predicted-to-reach-250-billion-usd-by-2031/

[3] Logix Consulting. (2020). What is Signature Based Malware Detection. Accessed: Apr. 3, 2023. [Online]. Available: https://www.logixconsulting. com/2020/12/15/what-is-signature-based-malware-detection/

[4] W. Liu, P. Ren, K. Liu, and H.-X. Duan, ‘‘Behavior-based malware analysis and detection,’’ in Proc. 1st Int. Workshop Complex. Data Mining, Sep. 2011, pp. 39–42.

[5] (2021). Polymorphic Malware. Accessed: Apr. 3, 2023. [Online]. Available: https://www.thesslstore.com/blog/polymorphic-malware-andmetamorphic-malware-what-you-need-to-know/

[6] M. Loman. (2021). LockfileRansomware’s Box of Tricks: Intermittent Encryption and Evasion. Accessed: Nov. 16, 2021. [Online]. Available: https://news.sophos.com/en-us/2021/08/27/lockfile-ransomwares-box-oftricks-intermittent-encryption-and-evasion/

[7] N. Pundir, M. Tehranipoor, and F. Rahman, ‘‘RanStop: A hardwareassisted runtime crypto-ransomware detection technique,’’ 2020, arXiv:2011.12248.

[8] S. Mehnaz, A. Mudgerikar, and E. Bertino, ‘‘RWGuard: A real-time detection system against cryptographic ransomware,’’ in Proc. Int. Symp. Res. Attacks, Intrusions, Defenses. Cham, Switzerland: Springer, 2018, pp. 114–136.

[9] J. Demme, M. Maycock, J. Schmitz, A. Tang, A. Waksman, S. Sethumadhavan, and S. Stolfo, ‘‘On the feasibility of online malware detection with performance counters,’’ ACM SIGARCH Comput. Archit. News, vol. 41, no. 3, pp. 559–570, Jun. 2013.

[10] A. Tang, S. Sethumadhavan, and S. J. Stolfo, ‘‘Unsupervised anomalybased malware detection using hardware features,’’ in Proc. Int. Workshop Recent Adv. Intrusion Detection. Cham, Switzerland: Springer, 2014, pp. 109–129.

[11] S. Das, J. Werner, M. Antonakakis, M. Polychronakis, and F. Monrose, ‘‘SoK: The challenges, pitfalls, and perils of using hardware performance counters for security,’’ in Proc. IEEE Symp. Secur. Privacy (SP), May 2019, pp. 20–38.

[12] S. P. Kadiyala, P. Jadhav, S.-K. Lam, and T. Srikanthan, ‘‘Hardware performance counter-based fine-grained malware detection,’’ ACM Trans. Embedded Comput. Syst., vol. 19, no. 5, pp. 1–17, Sep. 2020.

[13] B. Zhou, A. Gupta, R. Jahanshahi, M. Egele, and A. Joshi, ‘‘Hardware performance counters can detect malware: Myth or fact?’’ in Proc. Asia Conf. Comput. Commun. Secur., May 2018, pp. 457–468.

[14] S. Aurangzeb, R. N. B. Rais, M. Aleem, M. A. Islam, and M. A. Iqbal, ‘‘On the classification of microsoft-windows ransomware using hardware profile,’’ PeerJComput. Sci., vol. 7, p. e361, Feb. 2021.

[15] M. Alam, S. Bhattacharya, S. Dutta, S. Sinha, D. Mukhopadhyay, and A. Chattopadhyay, ‘‘RATAFIA: Ransomware analysis using time and frequency informed autoencoders,’’ in Proc. IEEE Int. Symp. Hardw. Oriented Secur. Trust (HOST), May 2019, pp. 218–227.

[16] K. Thummapudi, R. Boppana, and P. Lama, ‘‘HPC 41 events 5 rounds,’’ Harvard Dataverse, 2022, doi: 10.7910/DVN/MA5UPP.

[17] K. Thummapudi, R. Boppana, and P. Lama, ‘‘IO 41 events 5 rounds,’’ Harvard Dataverse, 2022, doi: 10.7910/DVN/GHJFUT.

[18] K. Thummapudi, R. Boppana, and P. Lama, ‘‘HPC 5 events 7 rounds,’’ Harvard Dataverse, 2022, doi: 10.7910/DVN/YAYW0J.

[19] K. Thummapudi, R. Boppana, and P. Lama, ‘‘Io 5 events 7 rounds,’’ Harvard Dataverse, 2022, doi: 10.7910/DVN/R9FYPL.

[20] K. Thummapudi, R. Boppana, and P. Lama, ‘‘Scripts to reproduce results,’’ Harvard Dataverse, 2023, doi: 10.7910/DVN/HSX6CS.