International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Harnessing Extra Tree Classifier in Machine Learning for Detection and Classification of Ransomware

Luqman S.M[1], E. R. Ramesh[2]

[1]M.Sc., Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, India
[2]Faculty, Centre of Excellence in Digital Forensics, Dr. MGR Educational and Research Institute, Chennai, India

Abstract:
Malicious programs or attacks, malware and ransomware households for instance, consistently endures to pose critical security issues to cybersecurity and it may cause Catastrophic damages to laptop systems, records centres, web, and cell applications throughout diverse industries and groups Traditional anti-ransomware structures battle to fight against newly created sophisticated attacks. Therefore, state-of-the-art techniques Like conventional and neural network-primarily based totally architectures may be immensely applied in the development of innovative ransomware solutions. In this paper, gifts a feature. Selection-primarily based totally framework with adopting one of a kind gadget mastering algorithms including. Extra Tree Classifier -primarily based totally architectures to categorise the safety stage for ransomware detection and prevention. This proposed method applied multiple machine learning algorithms.

Keywords: Train Test Split , Machine learning Algorithm, Extra tree Classifier, Logistic Regression.

1. INTRODUCTION

Malicious applications or attacks, malware and ransomware households for instance, constantly endures to pose critical protection problems to cybersecurity and it can purpose catastrophic [1], Damages to pc systems, facts centres, web, and mobile applications across various industries and businesses . Most ransomware is designed to save you targeted victims from accessing computer data by applying an Indestructible encrypting method that may be decrypted by the attacker itself solely. Removing the ransomware leads the sufferer to irreversible losses[2], as a result, sufferers are pressured to pay according to the attacker's demands . Failure or denial to follow the attacker's call for will cause dropping data Permanently. With the assist of contemporary-day technology, attackers are transforming conventional ransomware into emerging Ransomware household.

Ransomware is a malevolent program designed to scramble or lock user files until a ransom is paid, which poses a significant cyber security threat[3]. Machine learning algorithms for classification and detection are essential in combating this advancing menace. This intersection of cybersecurity and artificial intelligence enables preventive actions in detecting and mitigating ransomware attacks. The study investigates various machine learning techniques[4], ranging from traditional models to advanced deep

learning approaches, for enhancing the accuracy and efficiency of identifying ransomware. This presentation considers the fundamental principles of discriminating among different types of malware using machine learning techniques. Conventional detection methods are struggling to match the sophistication of current ransomware attacks[6]. With machine learning, it is possible to develop dynamic tactics that can identify complicated patterns and anomalies inherent within ransom ware behaviour[7]. This study goes deeper into key issues, techniques and advancements in utilizing machine learning for robust ransom ware classification and detection as an exciting way forward in improving cyber security.

2. REVIEW OF LITERATURE

Daniel Gibert, Carles Mateu, Jordi Planes [1]. The conflict among protection analysts and malware builders is a unending warfare with the complexity of malware converting as quick as innovation grows. Current state-of-the- artwork studies recognition at the improvement and alertness of system studying strategies for malware detection because of its capacity to maintain tempo with malware evolution. This survey ambitions at supplying a scientific and special evaluate of device mastering strategies for malware detection and in particular, deep mastering strategies. The main contributions of the paper are it provides a complete description of the methods and 16 features in a traditional machine learning workflow for malware detection and classification, (2) it explores the challenges and limitations of traditional machine learning and it analyzes recent trends and developments in the field with special emphasis on deep learning approaches. Furthermore, it provides the studies troubles and unsolved demanding situations of the contemporary strategies and it discusses the brand-new instructions of studies. The survey facilitates researchers to have an expertise of the malware detection subject and of the brand-new traits and instructions of studies explored via way of means of the medical community.

Muhammad Shabbir Abbasi, Harith Al-Sahaf & Ian Welch [2]. Many of the existing ransomware detection and classification models use datasets created through dynamic or behavior analysis of ransomware, hence known as behavior-based detection models. A huge undertaking in computerized behavior-primarily based totally ransomware detection and type is excessive dimensional information with several functions disbursed into diverse groups. Feature choice algorithms commonly assist to cope with excessive dimensionality for enhancing class performance. In connection with ransomware detection and classification, the majority of the feature selection methods used in existing literature ignore the varying importance of various feature groups within ransomware behavior analysis data set. For ransomware detection and classification, we advise a two-level characteristic choice technique that considers the various significance of every of the characteristic corporations withinside the dataset. The proposed technique makes use of particle swarm optimization, a wrapper-primarily based totally function choice algorithm, for choice of the most useful quantity of capabilities from every function institution to supply higher class performance. Although the proposed technique indicates similar overall performance for binary type, it plays appreciably higher for multi-elegance type than current function choice technique used for this purpose

Mohammad Masum, Hossain S [3]. Android, the maximum dominant Operating System (OS), reviews big recognition for clever gadgets for the previous couple of years. Due to its' popularity and open characteristics, Android OS is becoming the tempting aim of malicious apps that might motive excessive protection threat to financial institutions, businesses, and individuals Traditional anti-malware structures do now no longer suffice to fight newly created state-of-the-art malware. Hence, there is an developing need for computerized malware detection solutions to reduce the risks of malicious activities. In 7 recent

years, machine learning algorithms have been showing promising results in classifying malware where most of the methods are shallow learners like Logistic Regression (LR). In this paper, we endorse a deep studying framework, called Droid-NNet, for malware classification. However, our proposed method Droid-NNet is a deep learner that outperforms modern-day modern-day machine studying methods. We finished all of the experiments on datasets (Malgenome-215 & Drebin-215) of Android apps to evaluate Droid-NNet. The experimental quit end result indicates the robustness and effectiveness of Droid-NNet.

Md Jobair Hossain Faruk , Hossain Shahriar, Maria Valero, Alfredo Cuzzocreak, Dan Loss [4]. With the speedy technological advancement, safety has emerge as a chief difficulty because of the growth in malware hobby that poses a severe risk to the safety and protection of both computer systems and stakeholders. To maintain stake holder' s, particularly, end user's security, protecting the data from fraudulent efforts is one of the most pressing concerns. A set of malicious programming code, scripts, lively content, or intrusive software program this is designed to wreck supposed pc structures and packages or cell and net programs is referred to as malware. According to a study, naive customers are not able to differentiate among malicious and benign applications. Thus, pc

structures and cell programs must be designed to discover malicious sports closer to shielding the stakeholders. A wide variety of algorithms are to be had to locate malware sports with the aid of using making use of novel standards which includes Artificial Intelligence, Machine Learning.

Sudeep Tanwar, Ankur Gupta, Rajesh Gupta [5]. Ransomware assaults have emerged as a prime cyber-protection chance in which person records is encrypted upon machine infection. Latest Ransomware strands the usage of superior obfuscation strategies along-side offline C2 Server competencies are hitting Individual customers and massive businesses alike. This hassle has induced enterprise disruption and, of course, economic loss. Since there's no such consolidated framework which can classify, detect and mitigate Ransomware assaults in a single go, we're inspired to give Detection Avoidance Mitigation (DAM), a theoretical framework to check and classify techniques, tools, and techniques to detect, avoid and mitigate Ransomware. We have very well investigated one of kind situations and in-comparison already current country of the artwork assessment studies towards ours. The case examine of the notorious Djvu Ransomware is included to demonstrate the modus-operandi of the cutting-edge Ransomware strands, together with a few hints to comprise its spread.

Aldin Vehabovic; Nasir Ghani; Elias Bou-Harb; Jorge Crichigno; Aysegül [6]. Ransomware makes use of encryption strategies to make statistics inaccessible to valid users. To date a huge variety of ransomware households were advanced and deployed, inflicting huge harm to governments, corporations, and personal users. As the ones cyberthreats multiply, researchers have proposed diverse ransom ware detection and sort schemes. Most of those strategies use superior gadget mastering strategies to technique and examine real-international ransomware binaries and movement sequences. Hence this paper gives a survey of this important area and classifies current answers into numerous categories, i.e., along with network-based, host-based, forensic characterization, and authorship attribution. Key centers and equipment for ransomware evaluation also are supplied along-side open challenges.

Mohammad Masum, Md Jobair Hossain, Faruk Hossain Shahriar [7]. Malicious attacks, malware, and ransomware households pose essential protection troubles to cybersecurity, and it can reason catastrophic damages to pc systems, records centers, web, and cellular packages throughout various industries and businesses. Traditional anti-ransomware structures war to combat towards newly created state-of-the-art attacks. Therefore, latest strategies like conventional and neural network-primarily based totally

architectures may be immensely applied withinside the improvement of progressive ransomware solutions. In this paper, we gift a function selection-primarily based totally framework with adopting exclusive system gaining knowledge of algorithms together with neural network-primarily based totally architectures to categories the safety stage for ransomware detection and prevention.

We carried out more than one gadget mastering algorithms: Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR) in addition to Neural Network (NN)-primarily based totally classifiers on a particular range of functions for ransomware classification. We accomplished all of the experiments on one ransomware dataset to assess our proposed framework. The experimental outcomes show that RF classifiers outperform different strategies in phrases of accuracy, F -beta, and precision scores.

3. RESEARCH METHODOLOGY

In the proposed model, the data being processed is in the form of images and videos, and the task is carried The existing method Z-score standardization technique was used to convert each of the variables into a similar scale by centring each of the variables at zero with a standard deviation of 1. In this research, feature selection methods such as variance threshold and variance inflation factor to remove low variant and highly correlated features from the data is been applied, respectively. Removing low variation capabilities from the dataset, a variance threshold rating changed into set 1, because the range of capabilities dramatically dropped from 54 to 13. while threshold. In the second step of feature selection, further checked the multi-collinearity of the high variance features using variance inflation factor (VIF). A VIF rating 10 changed into decided on to become aware of fairly correlated features, that means that a characteristic is diagnosed if the VIF rating is better than 10. Features: Section Mean Raw Size and Section Max Raw Size show multicollinearity by displaying 19.52 and 19.48 VIF scores, respectively.

To enhance the accuracy, the extra tree classifier and logistic regression which scores 0.9952 and 0.9815. By using different data pre-process method and applying various machine learning algorithm, helps in improving the accuracy. The goal is to enhance detection accuracy. This research proposes an enhancement by replacing z-score with Inter quartile range, it's removed a more powerful noise and out of range. Feature selection to optimize the feature with Variance Threshold and correlation, it's improving the classification accuracy.

3.1Abbreviation and Acronyms:

VIF-Variation Inflation Factor, **ETC**-Extra Tree Classifier, **LR**-Logistic Regression**, PCA**- Principal Component Analysis

4. UNITS:

Machine Learning (ML) is a fascinating domain that focuses on training machines to learn from data and make predictions or decisions without explicit programming. Here are the key points: Supervised and Unsupervised Learning: Supervised Learning: In this type of ML, the algorithm learns from labelled data, where input features are associated with known output labels. Examples encompass linear regression, choice trees, and neural networks. Unsupervised Learning: Here, the set of rules learns from unlabeled data. Examples encompass k-way clustering and foremost issue analysis (PCA).

4.1ML Working on this Model:

The Extra Trees (Extremely Randomized Trees) classifier is a powerful machine learning algorithm that can be used to build an accurate predictive model. Here's a detailed explanation of how the Extra Trees

classifier works and how it can be applied to detect the accuracy of a model:

- Ensemble Learning : - Extra Trees is an ensemble method that builds multiple decision trees and merges their results to improve the overall prediction accuracy and control overfitting.
- Training Data: - Unlike traditional decision trees, Extra Trees uses the entire training dataset for each tree without bootstrapping (i.e., it does not create random subsets of the data).
- Randomness in Feature Selection*: - For each split in a tree, a random subset of features is chosen. This is similar to Random.

4.2 Design and Implementation Constraints:

4.2.1 Constraints in Analysis: Constraints as Informal Text, Constraints as Operational Restrictions, Constraints Integrated in Existing Model Concepts Constraints as a Separate Concept, Constraints Implied by the Model Structure Constraints in Design: Determination of the Involved Classes, Determination of the Involved Objects, Determination of the Involved Actions, Determination of the Require Clauses, Global actions and Constraint Realization

4.2.2 Constraints in Implementation: A hierarchical structuring of relations may result in more classes and a more complicated structure to implement. Therefore it's far really useful to convert the hierarchical relation shape to a easier shape consisting of a classical flat one. It is rather straightforward to transform the developed hierarchical model into a bipartite, flat model, consisting of classes on the one hand and flat relations on the other. Flat members of the family are desired on the layout degree for motives of simplicity and implementation.

4.3 Other Nonfunctional Requirements

Performance Requirements The software at this aspect controls and communicates with the subsequent 3 principal widespread components. Embedded browser in charge of the navigation and accessing to the web service; Server Tier: The server side contains the main parts of the functionality of the proposed architecture[8]. Safety Requirements . The software may be safety-critical. The software program might not be safety-essential even though it bureaucracy a part of a safety-essential system. [9]If a system must be of a high integrity level and if the software is shown to be of that integrity level, then the hardware must be at least of the same integrity level. There is little point in producing 'perfect' code in some language if hardware and system software (in widest sense) are not reliable[10]. If a computer system is to run software of a high integrity level then that system should not at the same time accommodate software of a lower integrity level. 6. Systems with different requirements for safety levels must be separated[11]. Otherwise, the very best degree of integrity required have to be carried out to all structures withinside the identical environment..

5. ARCHITECTURE DIAGRAM:

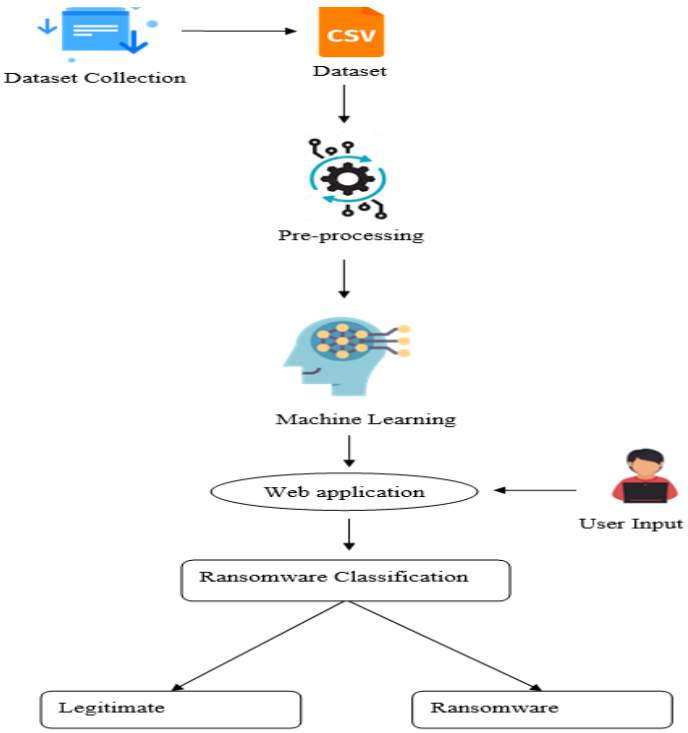

Fig.5.1: Comprehensive overview of system design

5.1 Module Explanation:

5.1.1 Data pre-process: The pre-process method uses inter quartile range and some feature selection method that helps in removing the unwanted noise in the data and further convert categorical variable into numerical variable with a help of label encoder[11]. Using correlation, an independent feature with an highly correlated factor removes the extracted feature.

5.1.2 Model selection: Using machine learning classifier algorithm method, its shows the algorithm and its accuracy score is been chosen as the extra tree classifier like the random forests algorithm, creates many choice trees[12], however the sampling for every tree is random, with out replacement. This creates a dataset for every tree with specific samples. A unique quantity of features, from the entire set of features, also are decided on randomly for every tree[13].

5.1.3 Prediction: In the Prediction module, the user interface is carried out with HTML and Python Flask As a front end and Back End Servers. The framework with all the experiments is applied on the ransomware dataset and evaluated the models performance by a robust comparative analysis among the Train test split data training model[14].

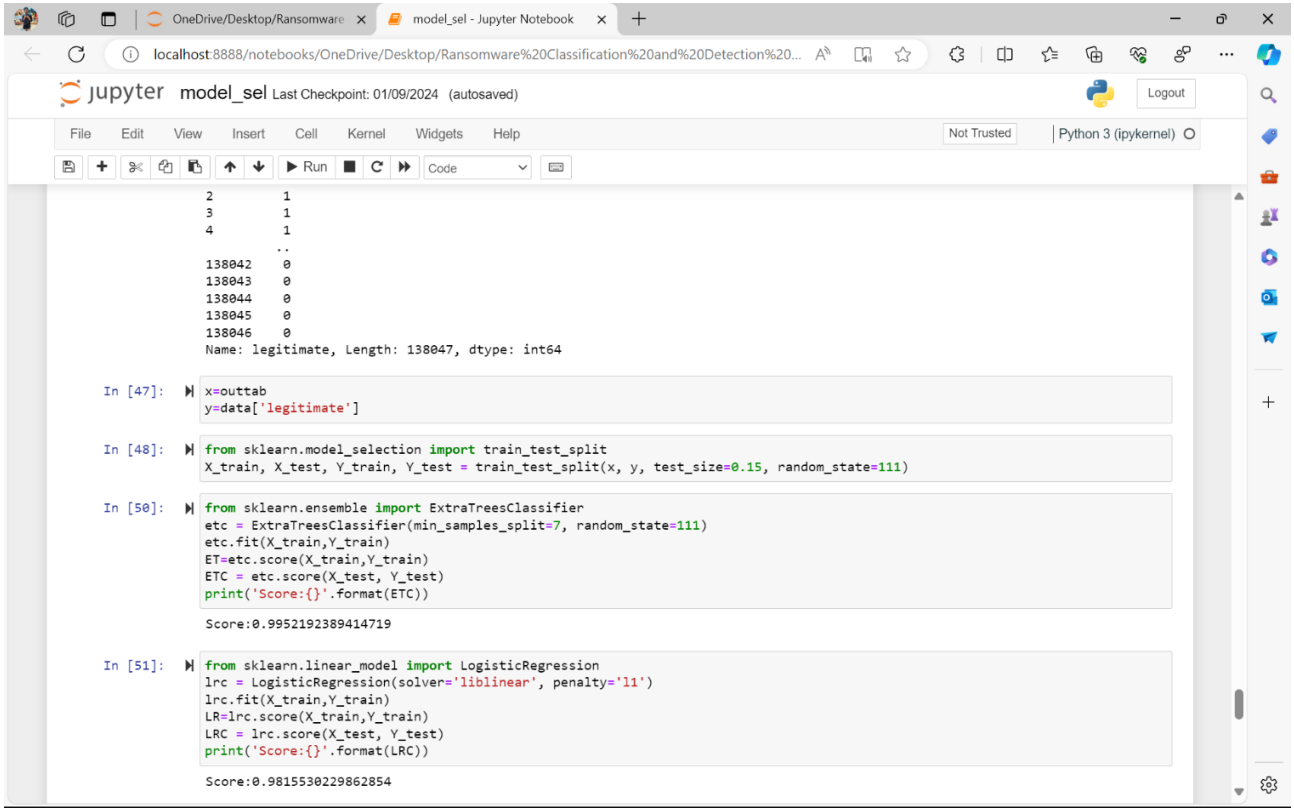

Fig. 5.2: Accuracy score of ET and LR Algorithm

6. FIGURE:

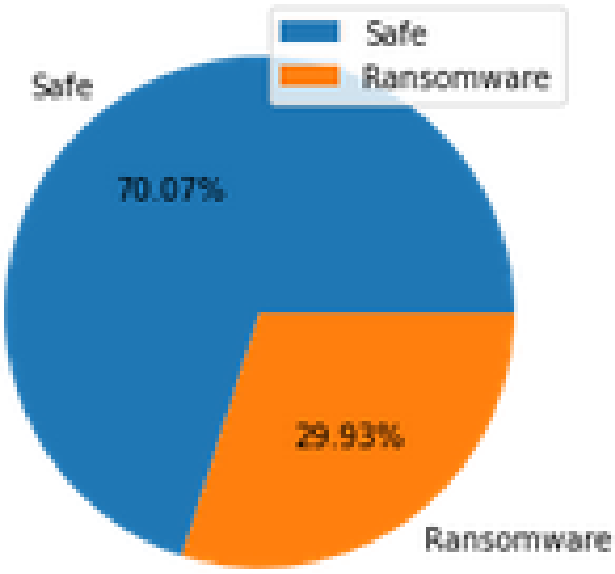

Fig. 3 :Result of the Model

7. APPENDIX:

Fig. 4: Visualization overview of Model

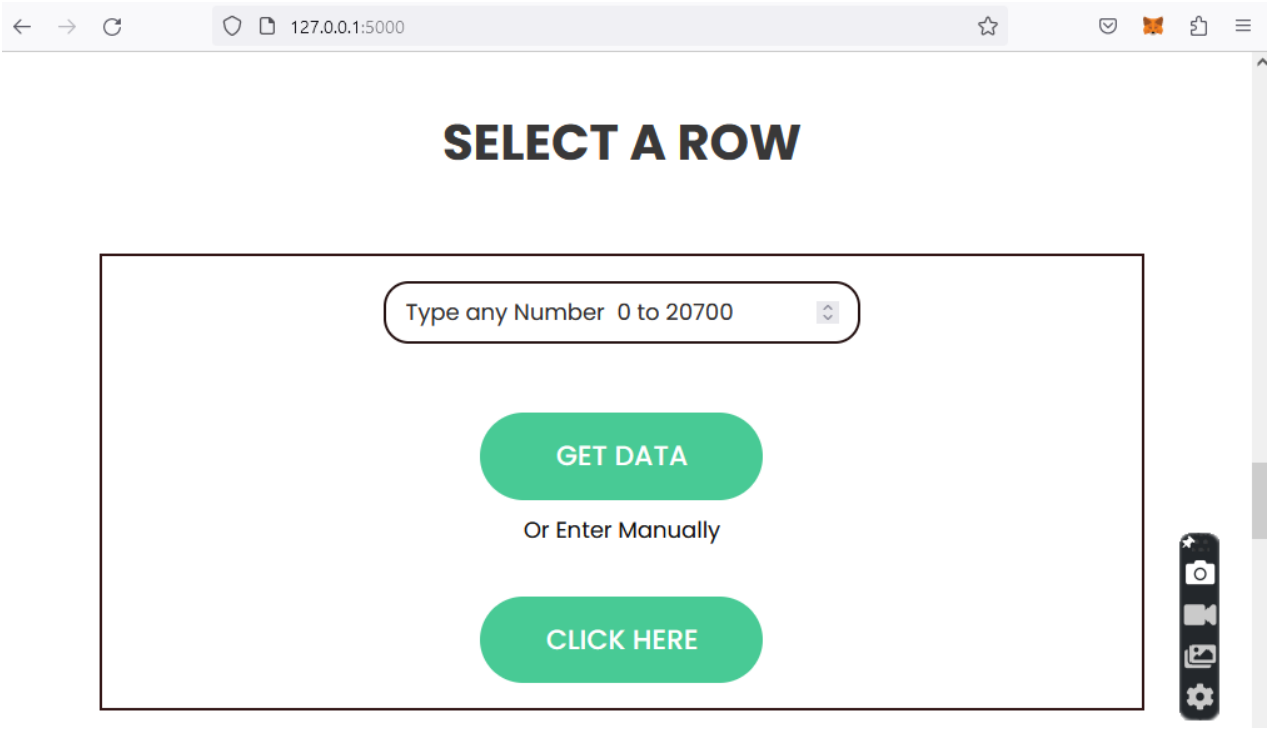

Fig .5: Visualization of User Interface

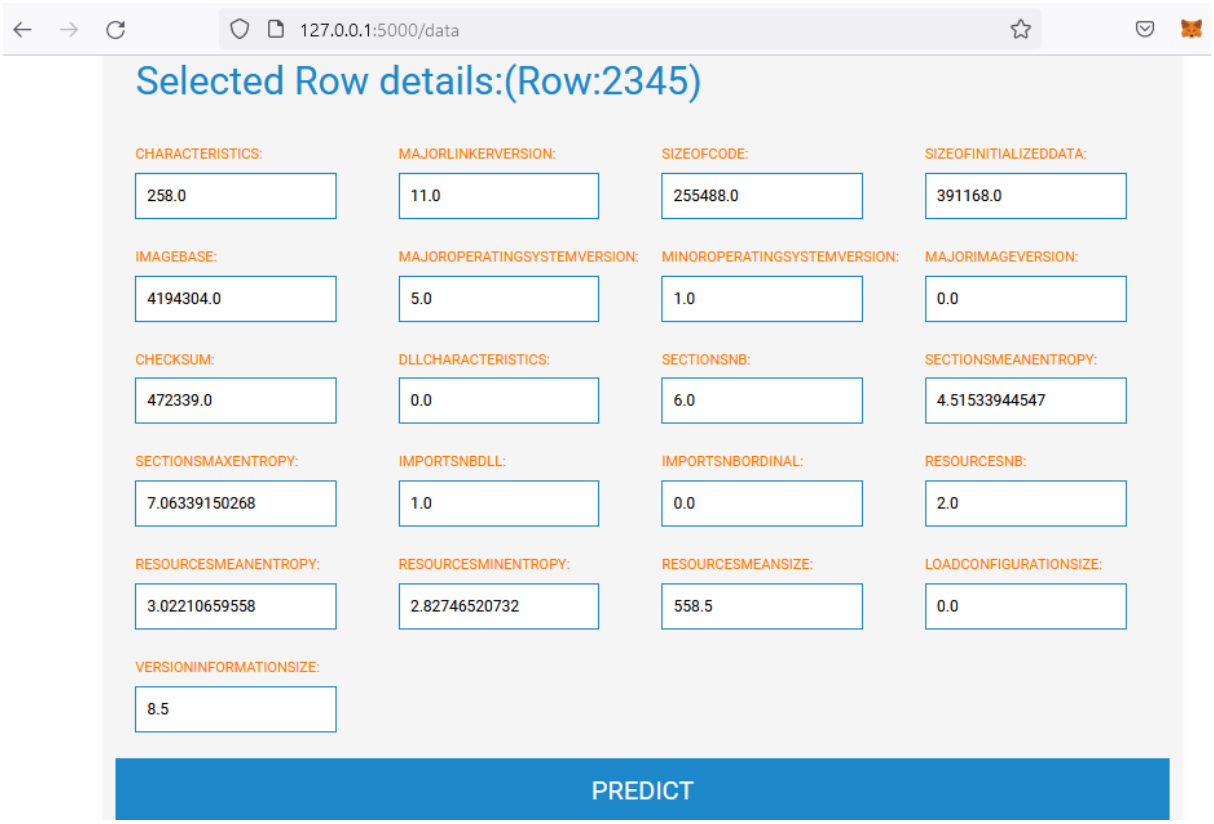

Fig. 6: Overview of the Prediction process

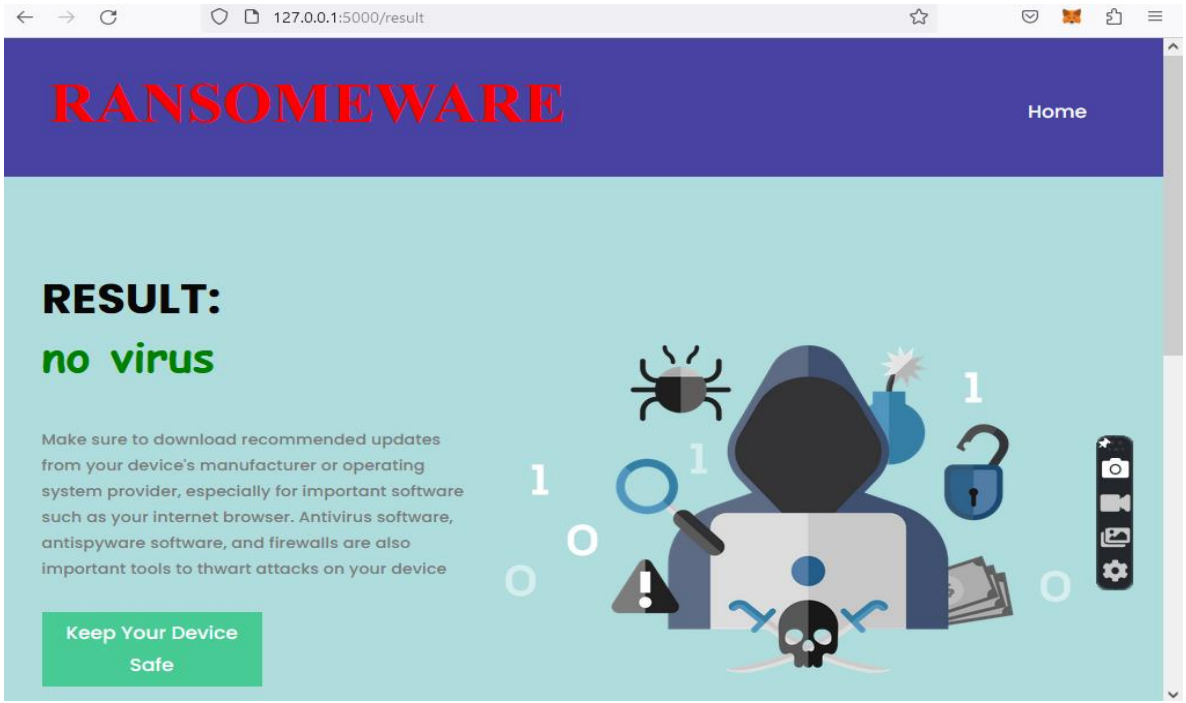

Fig. 7: Displaying the Final Model

8. CONCLUSION

Malware which include ransomware is an increasing number of posing a extreme safety hazard to economic institutions, businesses, and individuals. It is essential to develop an automatic system to effectively classify and detect ransomware and reduce the risk of malicious activities. In this paper, It is presented by a feature selection-based novel framework, adopted different machine learning algorithms including Extra -tree classifiers for effective ransomware classification and detection. In this proposed model applied the framework with all the experiments on a ransomware dataset and evaluated the models' performance by a robust comparative analysis among the Train split Data training model. The experimental results demonstrate that Extra-Tree classifier outperformed other classifiers by achieving the highest accuracy of 0.99521.

9. REFERENCES

1. K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "A Survey on Machine Learning Techniques for Cyber Security in the Last Decade," *IEEE Access*, vol. 8, pp. 222310–222354, 2020, doi: 10.1109/ACCESS.2020.3041951.
2. N. Shah and M. Farik, "Ransomware-Threats, Vulnerabilities And Recommendations," *Int. J. Sci. Technol. Res.*, 2017, [Online]. Available: https://www.ijstr.org/final-print/june2017/Ransomware-Threats-Vulnerabilities-And-Recommendations.pdf.
3. M. J. Hossain Faruk *et al.*, "Malware Detection and Prevention using Artificial Intelligence Techniques," *Proc. - 2021 IEEE Int. Conf. Big Data, Big Data 2021*, 2021, [Online]. Available: https://www.researchgate.net/publication/357163392_Malware_Detection_and_Prevention_using_Artificial_Intelligenc e_Techniques.
4. F. Noorbehbahani, F. Rasouli, and M. Saberi, "Analysis of machine learning techniques for ransomware detection," *Proc. 16th Int. ISC Conf. Inf. Secur. Cryptology, Isc. 2019*, pp. 128–133, 2019, doi: 10.1109/ISCISC48546.2019.8985139.
5. U. Adamu and I. Awan, "Ransomware prediction using supervised learning algorithms," *Proc. - 2019 Int. Conf. Futur. Internet Things Cloud, FiCloud 2019*, pp. 57–63, 2019, doi: 10.1109/FiCloud.2019.00016.
6. K. Savage, P. Coogan, and H. Lau, "The Evolution of Ransomware," *Res. Manag.*, vol. 54, no. 5, pp. 59–63, 2015, [Online].Available:http://openurl.ingenta.com/content/xref?genre=article&issn=08956308&volume=54&issue=5&spage=59.
7. 7D. W. Fernando, N. Komninos, and T. Chen, "A Study on the Evolution of Ransomware Detection Using Machine Learning and Deep Learning Techniques," *IoT*, vol. 1, no. 2, pp. 551–604, 2020, doi: 10.3390/iot1020030.
8. F. Noorbehbahani and M. Saberi, "Ransomware Detection with Semi-Supervised Learning," *2020 10h Int. Conf. Comput. Knowl. Eng. ICCKE 2020*, pp. 24–29, 2020, doi: 10.1109/ICCKE50421.2020.9303689.
9. L. Chen, C.-Y. Yang, A. Paul, and R. Sahita, "Towards resilient machine learning for ransomware detection," 2018, [Online]. Available: http://arxiv.org/abs/1812.09400.
10. A. M. Abiola and M. F. Marhusin, "Signature-based malware detection using sequences of N-grams," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 120–125, 2018, doi: 10.14419/ijet.v7i4.15.21432.
11. D. Nieuwenhuizen, "A behavioural-based approach to ransomware detection," *MWR Labs*, 2017,
12. Y. L. Wan, J. C. Chang, R. J. Chen, and S. J. Wang, "Feature-Selection-Based Ransomware Detection

with Machine Learning of Data Analysis," *2018 3rd Int. Conf. Comput. Commun. Syst. ICCCS 2018*, pp. 392–396, 2018, doi: 10.1109/CCOMS.2018.8463300.

13. M. Masum and H. Shahriar, "Droid-NNet: Deep Learning Neural Network for Android Malware Detection," Proc. -2019 IEEE Int. Conf. Big Data, Big Data 2019, pp. 5789–5793, 2019, doi: 10.1109/BigData47090.2019.9006053.
14. H. Ghanei, F. Manavi, and A. Hamzeh, "A novel method for malware detection based on hardware events using deep neural networks," J. Comput. Virol. Hacking Tech., vol. 17, no. 4,pp. 319–331, 2021, doi: 10.1007/s11416-021-00386-y