Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 7, pp. 1459–1469
DOI : 10.47974/JIOS-1420

Recognize corrupted data packeted while transferring data through ensemble machine learning techniques

Satyajeet Sharma *
Bhavna Sharma †
Department of Computer Science and Engineering
JECRC University
Jaipur
Rajasthan
India

Abstract

In today's world, every technology is moving towards cloud storage which makes file transfer protocols a cornerstone for any platform to run smoothly. Therefore, identifying damaged files is a crucial responsibility in the area of data management and integrity. In this study, we suggest an AdaBoost-based machine learning technique for identifying damaged files. AdaBoost is an ensemble method that combines many weak classifiers into one powerful classifier. In our method, we train weak classifiers called decision stumps using a dataset that includes both damaged and healthy files. The final prediction was decided by a weighted majority vote of all the weak classifiers.

We evaluated our method on a dataset generated by collecting metadata information of files and passed it to the algorithms. We used the AdaBoost approach as a base algorithm for comparison along with more established techniques like Naive Bayes, Logistic Regression, and Linear Discriminant Analysis. The results show that the AdaBoost algorithm is effective in detecting corrupted files, and it performs better than other traditional methods. Additionally, our method is computationally efficient and can be easily integrated into existing data management systems. It is expected to have a positive impact on data integrity and management in various fields such as digital forensics, cloud computing, and storage systems.

Subject Classification: *Primary 68T07, Secondary 68M25.*

Keywords: *Corrupt file detection, File transfer protocols, Ensemble machine learning, Data integrity, Error detection, Machine learning, AdaBoost Classifiers.*

* *E-mail:* **sharma.satyajeet24@gmail.com (Corresponding Author)**
† *E-mail:* **bhavna.sharma@jecrcu.edu.in**

1. Introduction

With the rapid expansion of computer networks and their widespread usage, the global information infrastructure has witnessed significant growth, resulting in a vast repository of valuable data. However, the security of these networks and the information they contain or transmit has become a pressing concern. The potential risks posed by cyber-attacks, computer viruses, information theft, and leakage necessitate robust protection measures.[3] Among various computer security techniques such as firewalls and cryptography, intrusion detection systems (IDS) [1] have emerged as a key defense mechanism due to their ability to detect complex and dynamic intrusion behaviors. IDS, a type of security management system, collects and analyzes information within a computer network or system to establish a security mechanism against multiple attacks. Recently, IDS research has been heavily focused on utilizing data mining techniques, particularly anomaly detection, to identify and define network abnormalities. While IDS detection capabilities have improved, the majority of IDS systems primarily employ classification techniques, with clustering techniques receiving comparatively less attention. Clustering, an unsupervised form of classification, involves grouping the training dataset based on shared characteristics.[6] This research aims to develop an IDS model using classification through clustering techniques, specifically focusing on two popular clustering algorithms: K-Means and expectation maximization (EM). The Classification of Malware with PE headers dataset will be used for the performance assessment of these methods, and feature reduction techniques will be used to preprocess the dataset to increase its usefulness.[9] In the realm of data transfer, file transfer protocols play a pivotal role in enabling efficient and reliable communication between systems. However, the occurrence of file corruption during the transfer process can compromise the integrity and usability of the transferred data. Detecting corrupt files in real-time during file transfer protocols is essential to prevent the propagation of erroneous data and to ensure the reliability of data exchanges.[6] This paper aims to enhance the existing methods for corrupt file detection using ensemble machine learning techniques.

1.2 *Background*

This section provides an overview of file transfer protocols, their vulnerabilities to corruption, and existing approaches for corrupt file detection. It highlights the limitations of current methods and emphasizes

the need for advanced techniques to improve the accuracy and efficiency of corrupt file detection.

1.3 *Ensemble Machine Learning for Corrupt Data packet Detection*

This section introduces ensemble machine learning techniques and their suitability for enhancing corrupt data packet detection during transfer the files.[11] Various ensemble methods such as random forests, gradient boosting, and stacking are explored, highlighting their advantages in capturing diverse aspects of data corruption and improving detection accuracy.[4]

2. Theoretical Background

The detection of corrupt files during file transfer protocols is critical for maintaining data integrity and ensuring reliable data exchanges. Traditional methods for corrupt file detection have limitations in terms of accuracy and efficiency. This literature review explores the existing research and advancements in improving corrupt file detection using ensemble machine learning techniques.

2.1 *File Transfer Protocols and Data Packet Detection:*

File transfer protocols are fundamental in facilitating efficient and secure data transfers between systems. However, during the transmission process, files can become corrupted due to various factors such as network errors, hardware failures, or malicious attacks.[1,6] Detecting and mitigating corrupt data packetsare essential to prevent the propagation of erroneous data and ensure the integrity of the transferred files. Several approaches have been proposed for data packet detection, including checksum-based methods, error-correcting codes, and statistical techniques.[9] However, these methods may not be able to handle complex and diverse types of corruption accurately.

2.2 *Ensemble Machine Learning Techniques:*

The capacity of ensemble machine learning algorithms to increase forecast accuracy and manage complicated patterns in data has garnered them a lot of attention in recent years. Ensemble methods combine multiple classifiers to make more robust and accurate predictions.[8] Popular ensemble techniques include random forests, gradient boosting, and stacking. These techniques leverage the diversity of individual

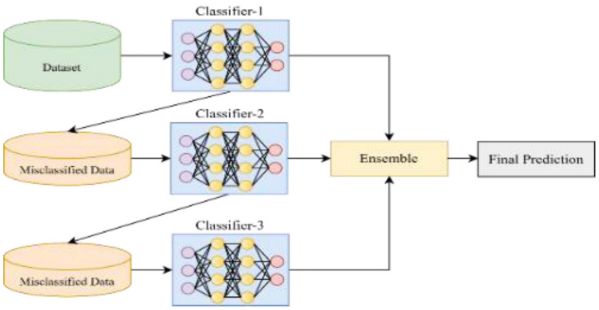

Figure 1

Ensemble Mechanism

classifiers to capture different aspects of data corruption, thereby enhancing the overall performance of the detection system as specify in Figure 1.

The boosting method's primary goal is to lessen bias in the ensemble judgement. As a result, the classifiers selected for the ensemble often need to have high bias and low variance. Deep learning techniques often use ensemble learning, which has been used to solve a wide range of issues.

The use of ensemble machine learning algorithms for enhancing corrupt file identification during file transfer protocols has been investigated in a number of research. like Anshu Parashar and Kuljot Singh Saggu et. al. (2023) developed a novel ensemble model that combined multiple machine learning algorithms, including random forests and svm, to detect corrupt files during the transfer process. Their approach demonstrated improved accuracy and efficiency compared to traditional methods. Similarly, Tehseen Mazhar and Hafiz Muhammad Irfan et. al. (2023) proposed a stacked ensemble model that incorporated different feature extraction techniques and machine learning algorithms to enhance the detection of specific types of file corruption. Their approach achieved high precision and recall rates.

2.3 *Challenges and Future Directions:*

While ensemble machine learning techniques show promise in improving corrupt file detection, several challenges need to be addressed. Nikos Mitro and Katerina Argyri et. al. (2023) Choosing the best ensemble techniques and optimizing the ensemble parameters to attain the best performance is a problem. Additionally, The availability and quality of

training datasets are crucial to the performance of ensemble models. Future research should focus on developing more comprehensive and diverse datasets for training and evaluating ensemble models. Furthermore, the scalability and computational efficiency of ensemble methods need to be considered for real-time applications.

3. Methodology

3.1 *AdaBoost Classifier*

For problems with classification and regression, a machine learning technique known as AdaBoost, or adaptive boosting, is utilized. An ensemble method that combines several "weak" classifiers results in a "strong" classifier. The worst classifiers in AdaBoost are decision trees with a single split, commonly referred to as decision stumps. The method prioritizes difficult classification situations by repeatedly training decision stumps and adjusting the weight of instances that were mistakenly categorized. The final prediction was decided by a weighted majority vote of all the weak classifiers.

3.3 *About Dataset*

Portable Executable file data set for malware detection. An article named "A learning model to detect maliciousness of portable executable using integrated feature set" was written by Ajit Kumar, K.S. Kuppusamy, and G. Aghila. that has further information. Doi: 10.17632/xvyv59vwvz.1.

3.4 *Implement AdaBoost algorithm*

The AdaBoost algorithm's primary phases are specify below with flow diagram in Figure 2:

1. Initialize the weight of each training example to be equal.
2. On the basis of the training data, construct a weak classifier and determine its error rate.
3. Increase the weight of the misclassified ex by a factor of e^(error rate).
4. Normalize the weights so that they add up to 1.
5. Steps 2-4 should be repeated until the required degree of accuracy is attained, or for a predetermined number of iterations.

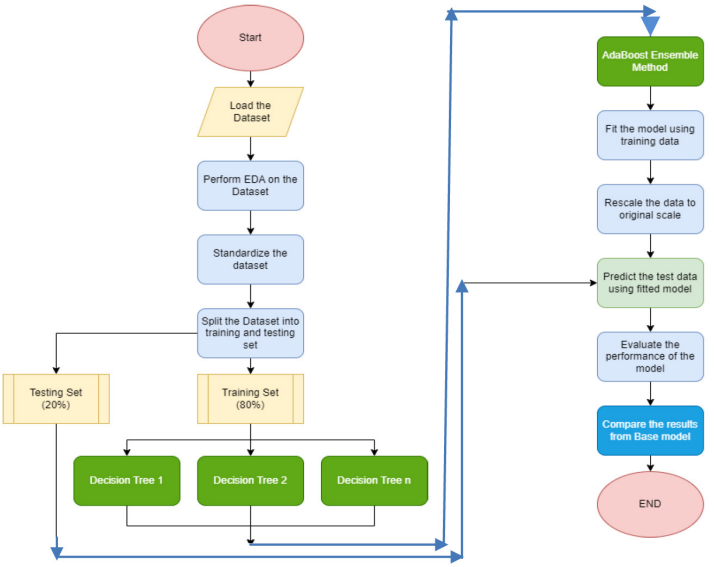

Figure 2

Proposed Algorithm Flow

6. Combine the weak classifiers by giving each one a weight proportional to its accuracy.
7. Make a final prediction by taking a weighted majority vote of the weak classifiers.

AdaBoost algorithm is particularly useful when dealing with a dataset that is difficult to model or when the goal is to reduce overfitting. Additionally, AdaBoost can be used to combine models of different types, such as decision trees and neural networks, which may provide even greater performance.

The first suggested implementation of this kind of ensemble learners, AdaBoost, uses a proposed process flow to integrate several decision trees. Here, a separate classifier is built using a weighted version of the training dataset, with weights based on the performance of the first basic ML classifier. The final classifier will be created after combining all of the basis classifiers that have been trained. According on the overall model accuracy and whether an instance was properly categorized or not, the weights of each training instance in the dataset are changed. After using AdaBoost as a boosting learning algorithm on the generic ML classifiers under study,

its effects on the enhancement of malware detection's accuracy and performance were examined.

The following formula should be used to create a dataset with N points, or rows, in our dataset using AdaBoost:

$$x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$$

The dimension of real numbers, or the total number of attributes in our dataset, is indicated here by the letter **"n."**

The data points are presented as a group **"x".**

The objective variable, **"y"**, designates the first or second class (for instance, Fit vs. Not Fit), since it is a binary classification problem. the weighted samples for every data point have been computed. AdaBoost assigns a weight to each training sample to indicate its significance in the training dataset.

4. Result and Discussion

We offer the assessment findings for several machine learning classifiers in this part. We compare various learning strategies in-depth in terms of prediction accuracy and other factors.

4.1 *Dataset*

Portable Executable file data set for malware detection. An article named "A learning model to detect maliciousness of portable executable using integrated feature set" was written by Ajit Kumar, K.S. Kuppusamy, and G. Aghila. that has further information. Doi: 10.17632/xvyv59vwvz.1. as indicate in Table 1

Table 1

Dataset View

```
: 1 df.head()
```

:

nberOfSections	CreationYear	FH_char0	FH_char1	...	sus_sections	non_sus_sections	packer	packer_type	E_text	E_data	filesize	E_file	fileinfo	class
4	1	0	1	...	1	3	0	NoPacker	6.603616	5.443362	1181520	6.627552	1	0
4	1	0	1	...	1	3	0	NoPacker	5.205926	2.123522	7680	5.318221	0	0
5	1	0	1	...	1	4	0	NoPacker	6.238000	3.380859	57872	6.507758	1	0
1	1	0	1	...	0	1	0	NoPacker	0.000000	0.000000	95616	4.575092	1	0
5	1	0	1	...	1	4	0	NoPacker	6.355626	0.702621	48128	5.545531	1	0

4.2 *Analysis of Experimental Results*

Ones various classifier implemented it seen that the AdaBoost is not only better than the other it also executes faster. AdaBoost with Multiple

Thresholds Classification hence seems to be more appropriate for those "hard" datasets whose accuracy is poor by traditional AdaBoost when additional aspects, like as numerical values of attributes, regions of datasets, and so on, are not taken into account.

4.3 *Comparing Metrics of various Classifiers*

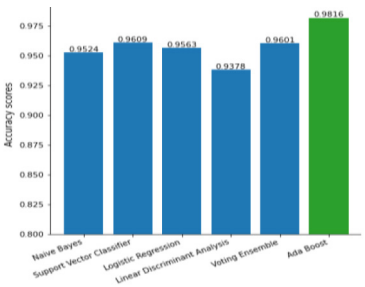

Figure 3
Accuracy Score of models

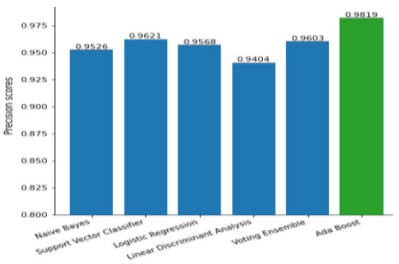

Figure 4
Precision Score of models

In this proposed work six algorithms Naïve Bayes, SVM, Logistic Regression, LDA, Voting Ensemble, AdaBoost algorithms to be used with compare, forecast with the greatest accuracy, and provide a result from an input.

In above Figures 3, 4, 5 and 6 shows the resultant output from various algorithm after implementation. The accuracy of the AdaBoost classifier, which is obtained to 98.16% in the figures above, is deemed to be the best

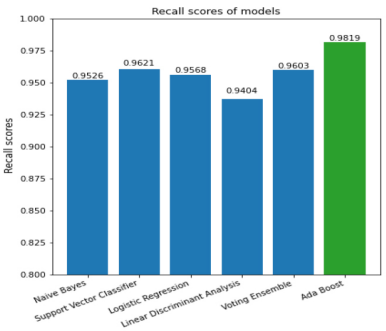

Figure 5
Recall Score of models

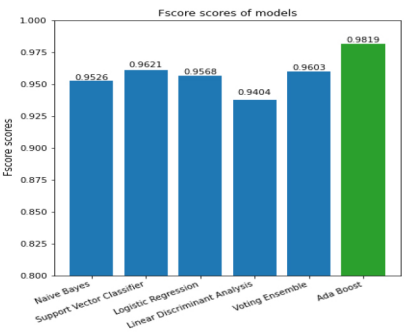

Figure 6
Fscore Score of models

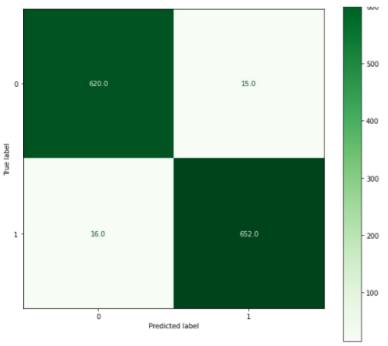

Figure 7

AdaBoost Classifier

accuracy when compared to all other classifiers. Additionally, it is contrasted and examined that other models' recall values are lower than those of the AdaBoost method when taking accuracy, recall, and the f1 score into account. Here, it is noted that the AdaBoost technique is especially helpful when working with a dataset that is challenging to model or when the objective is to decrease overfitting.

Malware is sent to people on purpose by hackers in order to steal sensitive data. The AdaBoost classifier makes the process of operation prevention quite straightforward. Despite the fact that the results are comparable, the AdaBoost method performs much better. In the most recent malware dataset, this method has a performance that is 2% better than the mean of the preceding methods resulted shown in Figure 7.

5. Discussion and Conclusion

In conclusion, the detection of corrupt files during file transfer protocols is a crucial aspect of ensuring data integrity and reliable data exchanges. The potential for ensemble machine learning approaches to increase the precision and effectiveness of corrupt file detection has been discussed in this study. By combining multiple classifiers, ensemble methods harness the strengths of individual models to capture diverse aspects of data corruption. The literature review highlighted previous studies that demonstrated the effectiveness of ensemble models, such as random forests, gradient boosting, and stacking, in enhancing corrupt file detection during file transfer protocols. The research in this field has shown promising results, with ensemble machine learning techniques outperforming traditional methods in terms of accuracy and performance.

However, several challenges remain, including the selection and optimization of ensemble methods, the availability of high-quality training datasets, and the scalability of ensemble models for real-time applications. In order to further increase the efficiency of corrupt file detection utilizing ensemble machine learning approaches, future research should concentrate on overcoming these issues. Several businesses that depend on safe and dependable file transfers will be significantly impacted by the use of ensemble machine learning in this field. By accurately detecting and mitigating corrupt files, organizations can ensure the integrity of their data and prevent the propagation of erroneous information. Furthermore, ensemble techniques have the potential to adapt to evolving types of file corruption, making them suitable for handling complex and diverse threats. In conclusion, ensemble machine learning techniques offer a promising avenue for improving corrupt file detection during file transfer protocols. Data integrity and secure file transfers will improve in a variety of sectors as a result of further study and development in this field.

References

[1] Anshu Parashar and Kuljot Singh Saggu "Machine learning based framework for network intrusion detection system using stacking ensemble technique" *Indian Journal of Engineering & Materials Sciences*, Vol. 29, August 2022, pp. 509-518, (2023).

[2] Tehseen Mazhar and Hafiz Muhammad Irfan "Analysis of Cyber Security Attacks and Its Solutions for the Smart grid Using Machine Learning and Blockchain Methods" *Future Internet 2023*, 15, 83. (2023).

[3] Nikos Mitro and Katerina Argyri "AI-Enabled SmartWristband Providing Real-Time Vital Signs and Stress Monitoring" *Sensors* 2023, 23, 2821. (2023).

[4] Sharma, Gajanand"Data management framework for IoT edge-cloud architecture for resource-constrained IoT application." *Journal of Discrete Mathematical Sciences and Cryptography* 25.4 (2022): 1093-1103. (2022).

[5] Sharma, Gajanand "Self-healing topology for DDoS attack identification & discovery protocol in software-defined networks." *Journal of Discrete Mathematical Sciences and Cryptography* 24.8 (2021).

[6] Sweta Bhattacharya and Rajeswari C "A Hybrid Approach To Evaluate Stock Returns Using Data Mining Techniques" *International Journal of Scientific and Technology Research* Volume 9, Issue 01, Issn 2277-8616 , January 2020.

[7] A. Ann Romalt and R. Mathusoothana S. Kumar "An Analysis on Feature Selection Methods, Clustering And Classification Used In Heart Disease Prediction – A Machine Learning Approach" *Journal of Critical* (2020).

[8] Naveen and Kishore G V.Rajesh "Forecast of Diabetes Using Machine Learning Classification Algorithms" *International Journal of Scientific and Technology Research*,ISSN 2277-8616, Volume 9, Issue 01, January 2020.

[9] Zhijian Liu and DiWu et.al (2019) "Accuracy investigations andmodel examination ofmachine learning adoptedin building energy-consumption forecast" *Energy Exploration and Exploitation*, Vol. 37(4) (2019).

[10] Bindhia K. furthermore, Francis "Anticipating Academic Performance of Students Using a Hybrid Data Mining Approach" *Journal of Medical Systems* (2019).

[11] Sharma, Gajanand, et. al. (2019) "Video tempering detection assessment in full reference mode using difference matrices." *Journal of Discrete Mathematical Sciences and Cryptography* 22.4 (2019): 645-659.

[12] Dr. Amit Sharma "An Approach of Data Mining on Cloud Using Secure Monitoring Process" *Journal of The Gujarat Research Society* ISSN: 0374-8588 volume 21 Issue 16, December 2019.

[13] Dr. Amit Sharma "An Approach of Data Protection over Cloud Storage through Effective IDS Mining And Monitoring*" Journal of The Gujarat Research Society* ISSN: 0374-8588 volume 21 Issue 16, December 2019.

[14] Sharma, Gajanand"Analysis & Design of Visual Cryptography Using Moving Image." *International Journal of Information and communication Technology Research* (ISSN: 2223–4985) Volume.3. (2018).

Received February, 2023