IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

DETECTING AND CLASSIFYING MALICIOUS UNIFORM RESOURCE LOCATIONS USING ADVANCED MACHINE LEARNING

[1]G Deepa Vujjini, [2]N Prashanthi, [3]Dr Nazimunnia

[1,2]Assistance Professor, Department of CSE, Sree Dattha Institute of Engineering and Science,

[3]Associate Professor, Department of CSE, Sree Dattha Institute of Engineering and Science

ABSTRACT

In today's digital age, the World Wide Web has become the primary platform for sharing knowledge and conducting economic activities. However, with this undeniable prominence comes the ever-growing concern of cybersecurity. Companies and governments are constantly researching ways to enhance their security measures. Symantec's 2023 Internet Security Report is a comprehensive resource that sheds light on various global threats. These threats range from corporate data breaches and attacks on browsers and websites to spear phishing attempts, ransomware, and other forms of fraudulent cyber activities. The report also exposes the cunning tactics employed by scammers to carry out their malicious deeds. Among the most common cybersecurity vulnerabilities is the existence of malicious websites or Unified Resource Locations (URLs). Every year, billions of rupees are lost due to hosting gratuitous material like spam, malware, unsuitable adverts, and spoofing. These sites often target unsuspecting visitors, luring them into falling for scams through emails, ads, web searches, or connections from other websites. It's crucial to have a reliable system in place that can accurately identify and categorize dangerous URLs as incidents of phishing, spamming, and malware continue to rise. However, implementing such a system is no easy task. The sheer volume of data, ever-changing patterns and technologies, complex relationships between characteristics, lack of sufficient training data, non-linearity, and the presence of outliers make classification challenging. In the proposed work, malicious URLs are detected and classified using advanced machine learning i.e., ensemble modelling. The dataset has been categorized into four types i.e., Phishing, Benign, Defacement and Malware. Here, logistic regression, and ensemble modeling are implemented to detect and classify malicious URLs.

Keywords: Website attacks, Malicious websites, Malware, Phishing scams, Phishing URLs, Email scams, Web search scams, Random Forest

1. INTRODUCTION

Malicious URLs (Uniform Resource Locators) are web addresses that lead to websites or online content designed with harmful intent. These URLs can pose a range of threats, including cybersecurity risks, privacy violations, and financial scams. Here's an overview of malicious URLs:

Types of Malicious URLs:

Phishing URLs: Phishing URLs are designed to impersonate legitimate websites or services, often with the goal of stealing sensitive information such as login credentials, credit card details, or personal data. Users are tricked into entering their information on fake websites.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

Malware Distribution URLs: These URLs lead to websites that host or distribute malicious software (malware), including viruses, trojans, ransomware, and spyware. Visiting such URLs can result in the automatic download and installation of malware onto the user's device.

Scam URLs: Scam URLs often promote fraudulent schemes or offers, including fake online marketplaces, investment opportunities, lottery winnings, and tech support scams. Users be lured into providing personal information or making payments.

Drive-By Download URLs: These URLs exploit vulnerabilities in a user's web browser or operating system to initiate automatic downloads of malicious files without the user's consent or knowledge.

Hijacked URLs: Malicious actors can take control of legitimate websites and change their content to host malicious payloads or redirect visitors to harmful websites. These URLs may appear trustworthy but have been compromised.

Characteristics of Malicious URLs:

Obfuscated URLs: Malicious URLs are often obfuscated using techniques such as URL shortening services, encoding, or URL redirects to hide their true intent.

Misspelled Domains: Attackers may create domains with slight misspellings of legitimate websites to trick users into visiting the malicious site.

IP Address URLs: Malicious URLs may use IP addresses instead of domain names to bypass domain-based blacklists.

Suspicious Domain Extensions: Some malicious URLs use unusual or suspicious domain extensions (e.g., .xyz, .info) to appear less credible.

3. Reporting: If users encounter a suspicious or malicious URL, they should report it to relevant authorities, internet service providers, or the website's hosting provider to help prevent further harm and investigations.

2. LITERATURE SURVEY

Many academics have researched the results of phishing websites. Our approach makes use of critical concepts from prior findings. We examined previous initiatives that used URL attributes to detect phishing, which impacted our present method.

However, in [1], the researchers generate accurate and trustworthy deep features of URLs while using simplistic features to evaluate URL legitimacy. The maximum throughput for this technology is determined by combining the results of all segments. Thorough research on an internet data set demonstrates that our system can keep up with other detection algorithms while spending an appropriate amount of time detecting phishing websites.

Korkmaz et al. [2] proposed a method for identifying phishing webpages based on Hypertext Markup Language (HTML) Tag Distribution Language (HTDL) and URL attributes. They also constructed concise HTDL and URL properties that allowed them to create HTDL string-embedding functions without relying on third-party infrastructure, allowing one approach to work in a legitimate recognition app. Researchers tested their technique on a

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

legitimate database of over 30,000 HTDL and URL characteristics. According to the authors, their arrangement obtained 98.24% accuracy, a 3.99% True Positive (TP) rate, and a 1.74% False Negative (FN) rate [3].

Aljofey et al. [3] described a strategy for identifying phishing relying on the URL of this publication. The researchers used multiple algorithms to evaluate the URLs of a couple of distinct data points using different types of DL approaches and hierarchical structures to compare the findings. The first method checks several URL attributes; the second investigates the website's legitimacy by reviewing where it is featured and who administers it and the next investigates the website's illustration appearance. DL algorithms and techniques are applied to examine the many aspects of web pages and URLs. They created [5] a unique method for identifying phishing websites that concentrate on analyzing a URL, which was described as a precise and effective detection technique. We have divided our new NN structure into multiple concurrent parts to give you a better idea. One approach is to start by removing shallow URL attributes.

To detect zero-day phishing schemes, Stokes et al. [6] suggested a unique approach for intelligent PD based on website text properties. The researchers used the principles of consistent resource identification and sequence strategies that usually match their framework. According to the researchers, the suggested technique successfully detects phishing and zero-day attacks, with a TP rate of 95.38%. Previous studies used website text structures to create PD frameworks. Meanwhile, phishers avoided detection by using content from other websites.

In their research on a method for spoofing site forecast that uses hyperlinks as a data feed for DL models, Yerima et al. [7] investigated the efficacy of the long short-term memory (LSTM) classifier. The researcher compares an RF classifier-based method against a new RNN-based method. They performed their scientific study of web addresses using fourteen criteria. First, researchers built a model using LSTM that accepts a hyperlink as a sequence of text as input and identifies whether the URL is legitimate or fraudulent. Despite the lack of specialist expertise required to create the features, users discovered that the LSTM algorithm outperforms the RF classifier in terms of average accuracy rate. Though without preprocessing and feature formation, their approach achieves 97.4% accuracy [8]. Their research, moreover, was restricted to the text-feature perspective of webpages. Integrating other elements such as frame features and website images might enhance the efficiency of their model.

3. PROPOSED SYSTEM

The process of malicious URL detection involves data preprocessing to prepare the dataset, random under-sampling to address class imbalance, and the use of a RFEC to build a robust and accurate model for identifying malicious URLs. This combination of techniques helps enhance cybersecurity efforts by identifying and blocking potentially harmful URLs. Figure 4.1 shows the block diagram of proposed system.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

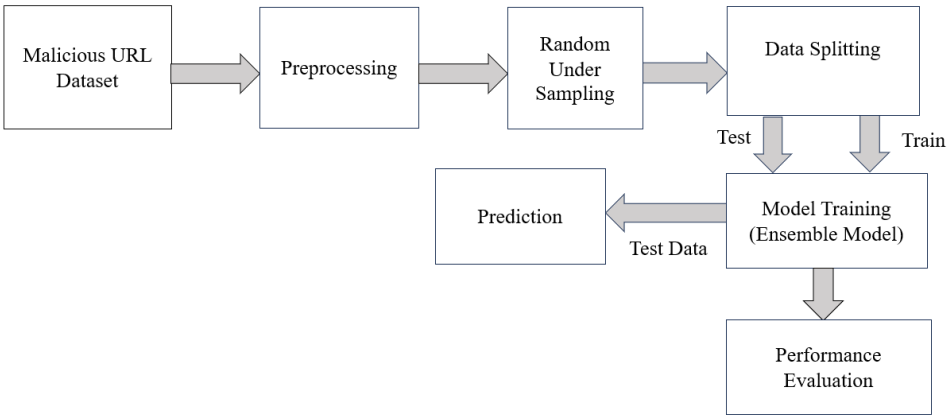

Fig. 1: Block diagram of proposed system.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

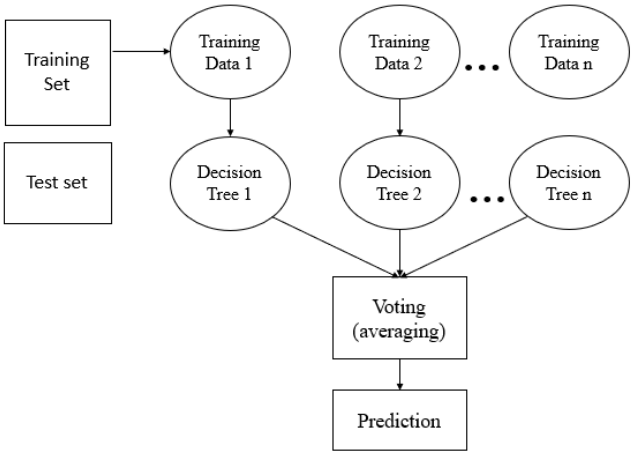

Fig. 2: Random Forest algorithm.

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

4. RESULTS

Figure 3 displays a portion of the dataset that is used for detecting malicious URLs. It include columns like 'URL', 'Label', and other features relevant to the task.

Figure 4 displays a pie chart representing the distribution of different categories or classes in the target column (e.g., benign, defacement, malware, phishing). Each slice of the pie represents a different category, and the size of the slice corresponds to the proportion of samples belonging to that category.

Figure 5 is a heatmap that visually represents the correlation between different variables in the dataset. It helps to identify relationships and dependencies between variables. Darker colors indicate stronger positive correlations, while lighter colors represent weaker or negative correlations.

	url	type
0	br-icloud.com.br	phishing
1	mp3raid.com/music/krizz_kaliko.html	benign
2	bopsecrets.org/rexroth/cr/1.htm	benign
3	http://www.garage-pirenne.be/index.php?option=...	defacement
4	http://adventure-nicaragua.net/index.php?optio...	defacement
...	...	...
651186	xbox360.ign.com/objects/850/850402.html	phishing
651187	games.teamxbox.com/xbox-360/1860/Dead-Space/	phishing
651188	www.gamespot.com/xbox360/action/deadspace/	phishing
651189	en.wikipedia.org/wiki/Dead_Space_(video_game)	phishing
651190	www.angelfire.com/goth/devilmaycrytonite/	phishing

651191 rows × 2 columns

Figure 3: Sample dataset used for malicious URL detection

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

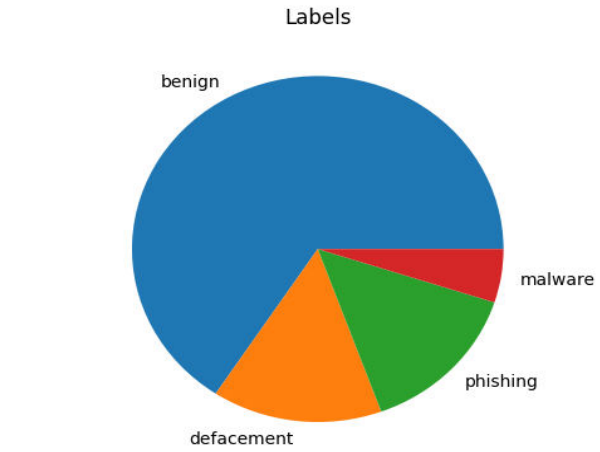

Figure 4: Pie chart used for count of target column

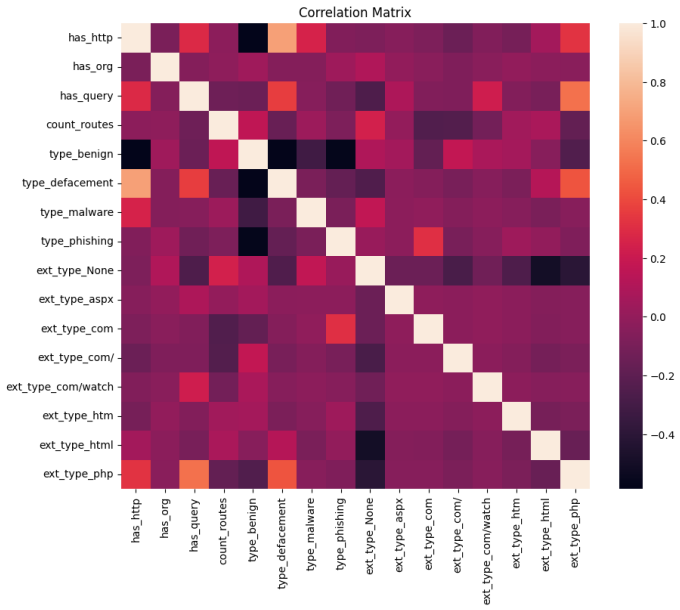

Figure 5: The heatmap represents the correlation between variables in the dataset

Figure 6 displays a portion of the dataset after various preprocessing steps have been applied. Preprocessing include tasks like one-hot encoding, handling missing values, feature extraction, etc.

Figure 7 displays the target column (e.g., 'Label') after preprocessing. It show the distribution of different classes or categories.

Figure 8 includes metrics such as precision, recall, F1-score, and support for each class in the classification task performed using a Naive Bayes classifier. It provides a detailed evaluation of the classifier's performance.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

	has_http	has_org	has_query	ext_type_None	ext_type_aspx	ext_type_com	ext_type_com/	ext_type_com/watch	ext_type_htm
0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	1
3	1	0	1	0	0	0	0	0	0
4	1	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
651186	0	0	0	0	0	0	0	0	0
651187	0	0	0	1	0	0	0	0	0
651188	0	0	0	1	0	0	0	0	0
651189	0	1	0	1	0	0	0	0	0
651190	0	0	0	1	0	0	0	0	0

651191 rows × 16 columns

Figure 6: data frame of variables in dataset after preprocessing

```
0          3
1          0
2          0
3          1
4          1
          ..
651186     3
651187     3
651188     3
651189     3
651190     3
Name: type, Length: 651191, dtype: int64
```

Figure 7: Target column of a dataset after preprocessing

Naive Bayes classifier classifiaction report

	precision	recall	f1-score	support
0	0.85	0.92	0.88	189539
1	0.64	0.99	0.78	43333
2	0.75	0.63	0.69	17389
3	0.80	0.22	0.35	49461
accuracy			0.80	299722
macro avg	0.76	0.69	0.68	299722
weighted avg	0.81	0.80	0.77	299722

Figure 8: classification report of Naive Bayes classifier

Figure 9 displays a confusion matrix specific to the Naive Bayes classifier. A confusion matrix provides a more detailed breakdown of the classifier's performance by showing the number of true positives, true negatives, false positives, and false negatives.

Figure 10 is a classification report, but it pertains to the evaluation of a Random Forest classifier.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

Figure 11 displays a confusion matrix, but it's specific to the Random Forest classifier.

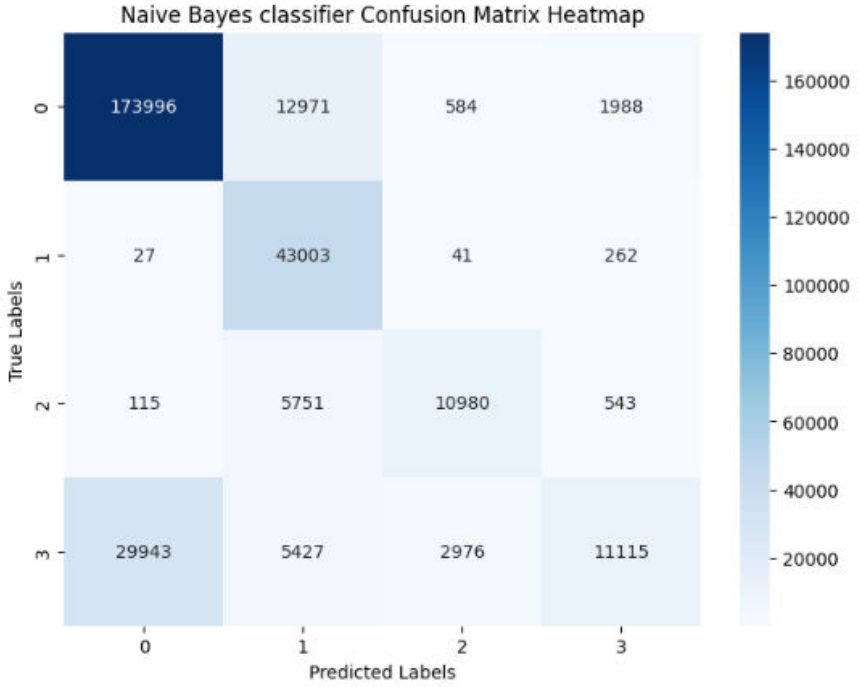

Figure 9: confusion matrix of Naïve Bayes Classifier

RandomForestClassifier Classification Report:

	precision	recall	f1-score	support
0	0.96	0.97	0.96	189539
1	0.99	0.99	0.99	43333
2	0.99	0.97	0.98	17389
3	0.86	0.83	0.84	49461
accuracy			0.95	299722
macro avg	0.95	0.94	0.94	299722
weighted avg	0.95	0.95	0.95	299722

Figure 10: Classification report of Random Forest classifier

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

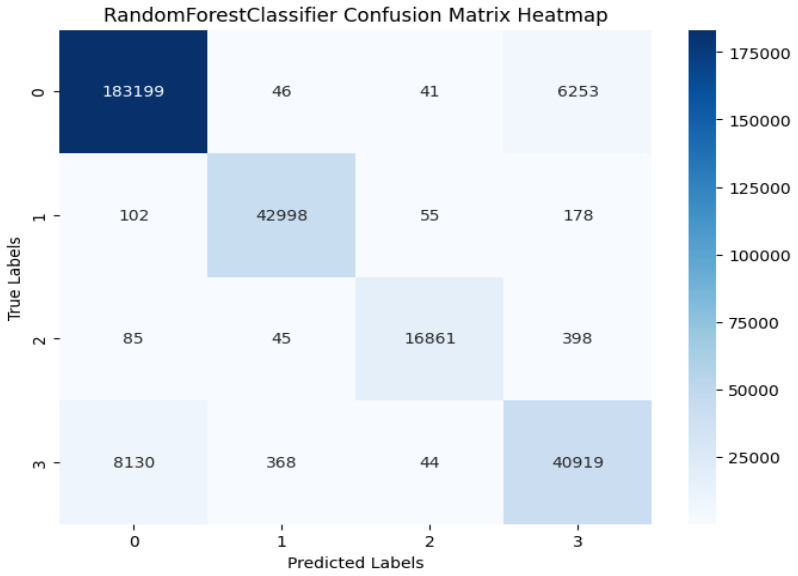

Figure 11: Confusion matrix of Random Forest classifier.

5. CONCLUSION

In conclusion, malicious URL detection is a critical cybersecurity task aimed at identifying harmful URLs to protect users and systems from online threats. This step involves data collection, cleaning, feature extraction, and splitting data into training and testing subsets to prepare the dataset for machine learning. To address imbalanced datasets where benign URLs significantly outnumber malicious ones, random under-sampling is used. It involves selecting a subset of benign URLs from the training data to balance the dataset, training the model on this balanced data, and evaluating its performance on unbalanced testing data. The RFEC, an ensemble machine learning algorithm, is commonly used for malicious URL detection. It consists of multiple decision trees trained on different subsets of data, with a majority voting mechanism to make predictions. Random Forest is known for its robustness and ability to handle complex and noisy data. Together, these techniques create a comprehensive approach to identifying and classifying malicious URLs, contributing to the overall cybersecurity efforts by helping to detect and mitigate online threats effectively.

REFERENCES

[1] Korkmaz, M.; Kocyigit, E.; Sahingoz, O.K.; Diri, B. Phishing Webpage Detection Using N-Gram Features Extracted from URLs. In Proceedings of the 3rd International Congress on Human–Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 11–13 June 2021; pp. 1–6.

[2] Do, N.Q.; Selamat, A.; Krejcar, O.; Yokoi, T.; Fujita, H. Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical Study. Appl. Sci. 2021, 11, 9210.

[3] Aljofey, A.; Jiang, Q.; Qu, Q.; Huang, M.; Niyigena, J.-P. An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. Electronics 2020, 9, 1514.

[4] Tajaddodianfar, F.; Stokes, J.W.; Gururajan, A. Texception: A Character/Word-Level Deep Learning Model for Phishing URL Detection. In Proceedings of the IEEE

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2857–2861.

[5] Yerima, S.Y.; Alzaylaee, M.K. High Accuracy Phishing Detection Based on Convolutional Neural Networks. In Proceedings of the 3rd International Conference on Computer Applications and Information Security (ICCAIS), Riyadh, Saudi Arabia, 19–21 March 2020; pp. 1–6.

[6] Prabakaran, M.K.; Chandrasekar, A.D.; Sundaram, P.M. An enhanced deep learning-based phishing detection mechanism to effectively identify malicious URLs using variational autoencoders. IET Inf. Secur. 2023.

[7] Gopinath, M.; Sethuraman, S.C. A Comprehensive Survey on Deep Learning Based Malware Detection Techniques. Comput. Sci. Rev. 2023, 47, 100529.