IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124205

Phishing Attacks Detection Using Hybrid Deep Learning Algorithms

Janani.E[1], Dr.M.S.Anbarasi[2]

Student, Information Technology, Puducherry Technological University, Puducherry, India[1]

Professor, Information Technology, Puducherry Technological University, Puducherry, India[2]

Abstract: The sophistication of phishing assaults is rising, making it challenging to identify them using conventional means. Consequently, there will be a growing need for more advanced techniques to recognise and thwart such attacks. We introduce a hybrid deep learning strategy for phishing attack detection in this research. The proposed method of phishing website detection can be done by combining convolutional neural networks (CNN) and recurrent neural networks (RNN), two alternative deep learning models. The RNN is used to identify temporal dependencies in the data, while features have been determined from the unprocessed information using CNN. The hybrid model is highly accurate at identifying phishing assaults since it is trained on a vast dataset of authentic and phishing websites. We demonstrate that the suggested algorithm outperforms leading-edge methods by comparing its performance to those. Overall, our suggested method offers a reliable defence against phishing assaults and can be applied to increase the security of online systems.

Keywords: Legitimate, Phishing, Cyber Security, Deep learning, Feature Extraction, Websites, CNN, RNN.

I. INTRODUCTION

Phishing assaults are turning into a serious problem for online safety. Phishing attacks employ fraudulent websites or emails to persuade users to divulge private data, including usernames, passwords, and credit card numbers. Attacks like these have the ability to seriously harm people's finances and reputations. Heuristics, blacklists, and manual inspection are the mainstays of conventional phishing assault detection techniques, but they frequently fall short when faced with more complex attacks. Consequently, more advanced techniques that can consistently and automatically identify phishing attacks are required.

Phishing is one of the online criminal activities that intruders employ to access user credentials or sensitive information about victims.The user will browse the website after clicking the link, think it is the real thing, and attempt to enter his information. This method of accessing is accomplished by creating copies of the websites that appear to be exact reproductions of the real websites we frequently visit. To address this issue, we are using a few machine learning algorithms, which will let us to identify phishing websites based on the parameters of the algorithm.

II. RELATED WORK

S.NO	TITLE	OBJECTIVE	ENHANCEMENT	LIMITATION
1.	URL Phishing Detection System Utilizing CatBoost Machine learning Approach (2021)	Then, the analyses and assessments of the CatBoost, Random Forest, and Logistic Regression classifiers' performances were carried out.	Apache Spark is used to generate data and analyse it. Hybridisation of algorithm enhance accuracy.	Catboost uses bigger computational resources because it requires more time to train on and test the dataset.
2.	A hybrid DNN–LSTM model for detecting phishing URLs	LSTM Model was implemented for detection.	The word embedding characteristics were gathered to further enhance the efficiency of the outcomes.	Accurate but takes more time for processing.
3.	Phishing Webpage Classification via	Deep Learning (DL) techniques like DNN,	Experiment with other DL algorithms.	Single algorithm does not provide

	Deep Learning-Based Algorithms: An Empirical Study(2021)	CNN, LSTM, and GRU were employed.		possible accurate result.
4.	An intelligent cyber security phishing detection system using deep Learning techniques (2022)	Detection methods for phishing emails using any deep learning technique.	Develop an automated tool for phishing detection.	To keep up with the phishers' ongoing development of new approaches, feature selection techniques need to be improved substantially.
5.	CCrFS: Combine Correlation Features Selection for Detecting Phishing websites Using Machine Learning (2022)	Leveraging random forest as a technique for selecting characteristics and categorization.	incorporating random forest as a roofing system for feature selection and classification.	Applicable for limited dataset.
6.	PDGAN: Phishing Detection With Generative Adversarial Networks(2022)	Proposed phishing detection using generative adversial network.	Programme for assessing the model's Intricacy.	Does not rely on the content of websites or services provided by outside parties; instead, it basically requires the URL of a website.
7.	Phisher Cop - An Automated Tool Using ML Classifiers for Phishing Detection(2022)	A Stochastic Gradient Descent classifier (SGD) and a Support Vector Classifier (SVC) are the building blocks of Phisher Cop.	Do it in python tool box.	Limited for email phishing detection.
8.	A comparison study of machine learning techniques for phishing detection(2022)	Several diverse machine learning techniques, encompassing Decision Tree, Random Forest, Multilayer Perceptrons, SVM, and XGBoost	For the purpose of making the acquisition system more accurate.	Consumes more time.
9.	Web architecture for URL-based phishing detection based on Random Forest, Classification Trees, and Support Vector Machine (2022)	For the purpose of making predictions, Random Forest, Classification Trees, and Support Vector Machines are trained.	The prototype is going to be enhanced progressively with the goal to create a joint model with nearly flawless precision.	By combining their results, some errors occurs.
10.	Modeling Hybrid Feature-Based Phishing websites Detection Using Machine earning Techniques(2022)	XG Boost technique is used.	In future, we may extended this work by using efficient time.	Consumes more time.

11.	Vote algorithm based probabilistic model for phishing website detection (2022)	In the present investigation, an a two-phase probabilistic methodology is described.	In the future, The accuracy can further be increased	As Compared to single base classifiers, the voting mechanism consumes a greater amount of time.
12.	Towards Web Phishing Detection Limitations and Mitigation (2022)	ML-based detection.	Use some other tool box for better results.	Third party services might not be organised nor easy to extract
13.	Phishing Detection Leveraging Machine Learning and Deep Learning: A Review (2022)	Using both ML and DL classifier.	A URL is the sole source of data that might be used to train a classification model.	The throughput of this pipeline need to be controlled.
14.	A review of spam email detection: analysis of spammer strategies and the dataset shift problem	Adversarial machine learning Spammer Strategies are being used for detection of phishing.	To shore up the classification model's stability.	Spammer adversarial strategies are a challenge for this field.
15.	PhishSim: Aiding Phishing Website Detection With a Feature-Free Tool (2022)	Offer a feature-free solution for identifying phishing websites.	With a processing time of approximately 0.3 seconds, it ought to be possible for this to happen in real systems in the future.	Concentrated primarily with discovering variations of recognised attacks.

III. PROJECT PROPOSAL

As indicated, there are three components to the CNN-RNN algorithm. Feature extraction, categorization, and URL-embedded representation utilising RNN AND CNN. The creation and training of the CNN-RNN algorithm, the phishing detection method, as well as the data pretreatment process.

The URL character sequence is normalised to a fixed-size sequence at the embedded representation stage of the URL by interception or zero-filling, and the normalised sequence is then transformed into a one-hot coding sequence in accordance.

The embedding layer then transforms the sparse one-hot matrix into a dense character embedding matrix. The local deep correlation feature is produced from the embedding matrix during the feature extraction stage using the convolutional layer and maximal CNN pooling. The RNN neural network then receives the pooling result as input to determine the context of the URL sequence.

DOI: 10.17148/IJARCCE.2023.124205

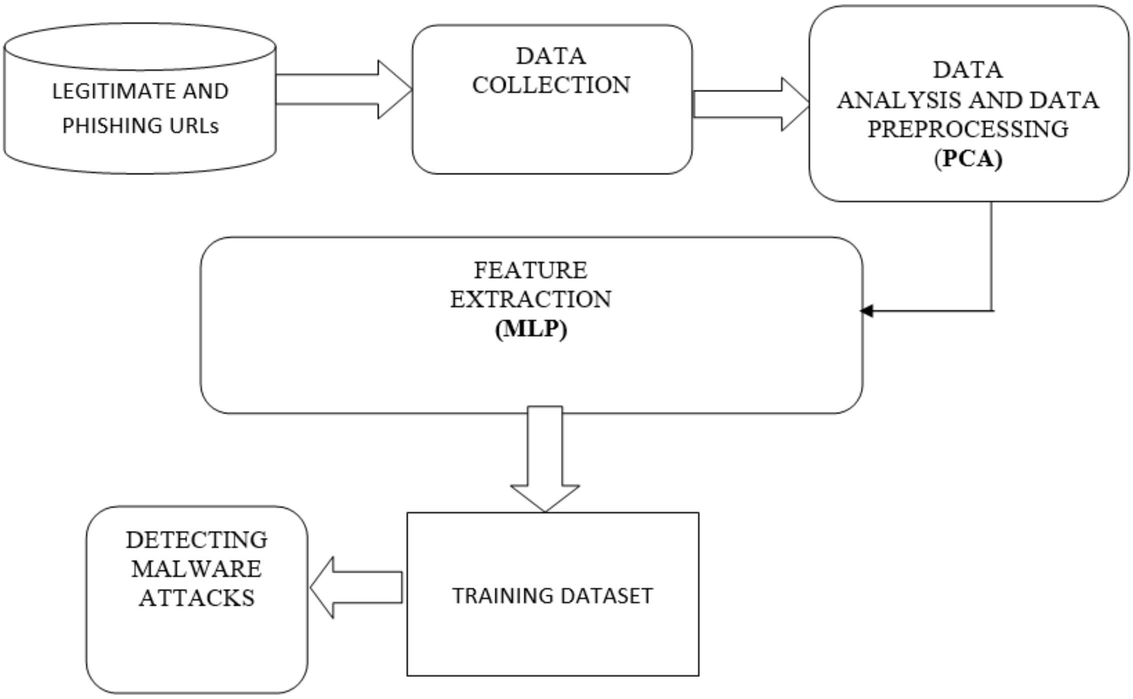

LEGITIMATE AND PHISHING URLs
DATA COLLECTION
DATA ANALYSIS AND DATA PREPROCESSING (PCA)
FEATURE EXTRACTION (MLP)
TRAINING DATASET
DETECTING MALWARE ATTACKS

Fig. 1. Architecture Diagram of Hybrid Phishing Detection System

A.Data Pre-processing

Data Pre-Processing techniques are done by eliminating missing data and omitting some columns, it will change the data to the necessary format. It will first compile a list of the column names it wants to keep or keep. Next, it drops or deletes all columns aside from the ones it wants to keep. Finally, it deletes or removes the rows from the data set that have missing values. Establishing representative network traffic for training and testing includes creating datasets. The labels for these datasets should state whether the connection is typical or abnormal. Network traffic labelling can be a very laborious and difficult task.

The goal of feature creation is to add new features that are more discriminatory than the existing feature set. This has the potential to greatly enhance machine learning algorithms. Features can be created manually or with the use of methods for data mining including frequent-episode mining, association mining, and sequence analysis.

Reduction: is frequently used to eliminate any redundant or unnecessary features from the dataset, hence reducing its dimensionality. It is common practise to ease "the curse of dimensionality" by using an optimisation technique known as feature selection. Feature extraction, which converts the initial feature set into a smaller number of new features, is another method for data reduction. In order to reduce the amount of data, a frequent linear method is principal component analysis (PCA).

B.Feature Extraction

The features that had been removed and preserved as a data set were used to train and test the CNN. This concept attempted to combine CNN, RNN, and a deep learning algorithm and apply them to the data obtained from URLs in order to more precisely detect phishing attempts. Websites that are legitimate and those that are phishing can be distinguished based on an analysis of the features gathered and the knowledge model. Every website can also be assessed to see whether it is legitimate or fraudulent. In order to eliminate duplication in the feature set, the suggested technique compares elements of the webpage's that are similar to the suggested solution.

The deep learning algorithm for the classification process is then taught using the feature set. The idea entails employing two algorithms based on deep learning, CNN and RNN, to analyse a variety of characteristics that could have been collected from websites in order to more accurately estimate phishing operations. The necessary characteristic may be derived from active websites using the extractor algorithm. To determine if the websites are reliable, questionable, or phishing, the knowledge model is used to assess the features that have been gathered. The Hybrid Phishing Detection

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124205

(HPDS) system, which has a three-part, user-friendly warning interface, was developed using the online technique. Websites can be classed as authentic, questionable, or phishing based on the discrepancies between the data collected and the HPDS model.

Every website is individually assessed to see whether it is authentic or fraudulent (phishing). When website feature extractions are input into the HPDS, a knowledge model developed using deep learning techniques (CNN and RNN) verifies the feature information before classification can take place. If the requested URL is a phishing webpage, the text directive with the status of phishing is the first element in the web-based plugin warning system. If the requested site is questionable, written instructions with an amber status are the next option. In the third way, a text command and written instructions are combined, and if the requested website has legitimate pages, a legitimate signal is provided.

C.Detection of Phishing Attacks

After the pre-trained learning is transferred, the output size of the fully integrated layer was changed to reflect the number of categories that would need to be divided into three categories: valid, suspicious, and phishing attacks. The learning rate factor and each bias learning rate factor were both set to 10. An additional layer was connected after the first categorization layer was removed. The recently linked categorization layer was checked, but no mistakes were found. The deep network structure was then added along with the new network. The extracted image data set was then transferred into the image data storage and processed to produce a number of features by gathering the accelerated robust characteristics from each image using a grid approach. The data was then divided into 30% for training and 70% for testing using holdout cross-validation.

D. Creating Web Application

Through the creation of a browser add-on. The extension receives a URL from the user using the method known as GET and passes it along to the Python code using the extension's Python script. The URL's features have been retrieved by the python code and then transferred to an array.

IV. PSEUDOCODE

```
Step 1: Collect the dataset from both phishing and legitimate dataset .
Step 2:  And then Stores the collected data in the database.
Step 3:  Collected dataset are then sended for data analysis and preprocssing.
Step 4:  The Collected dataset of processed dataset are passed into the feature extraction.
Step 5:  The extracted features are given to the proposed hybrid model.
Step 6:  This hybrid model detects the phishing attacks
Step 7:  End
```

V. RESULTS AND DISCUSSION

The Study involves the small discussion about the detection of Phishing Website using the Hybrid deep learning algorithm by creating the web application through that it is possible to detect the phishing website detection based on the features of URL and also can be able to detect the Website as phishing by entering the URL which we want to detect in for attacks detection.

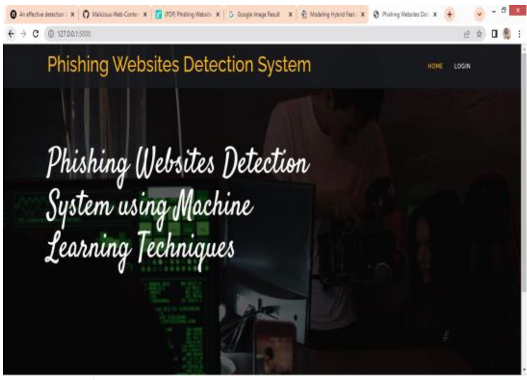
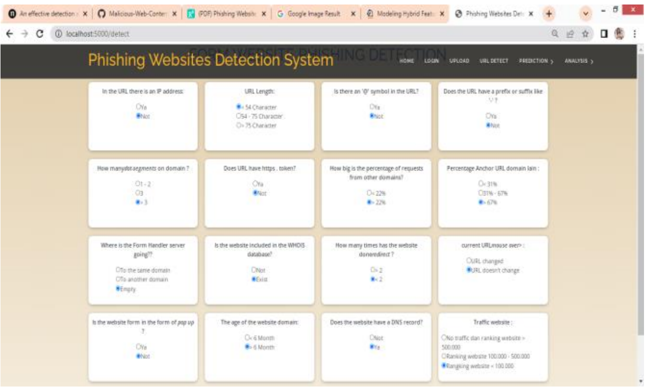


Fig. 1. Admin page



In the URL there is an IP address:
○Ya
●Not

URL Length:
●< 54 Character
○54 - 75 Character
○> 75 Character

Is there an '@' symbol in the URL?
○Ya
●Not

Does the URL have a prefix or suffix like '-' ?
○Ya
●Not

How manydot-segments on domain ? :
○1 - 2
○3
●> 3

Does URL have https . token?
○Ya
●Not

How big is the percentage of requests from other domains?
○< 22%
●> 22%

Percentage Anchor URL domain lain :
○< 31%
○31% - 67%
●> 67%

Where is the Form Handler server going??
○To the same domain
○To another domain
●Empty

Is the website included in the WHOIS database?
○Not
●Exist

How many times has the website doneredirect ?
○> 2
●< 2

current URLmouse over> :
○URL changed
●URL doesn't change

Is the website form in the form of *pop up* ?
○Ya
●Not

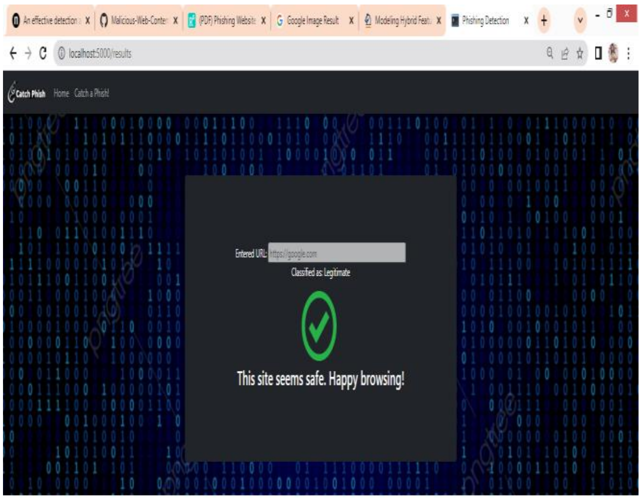
The age of the website domain:
○< 6 Month
●> 6 Month

Does the website have a DNS record?
○Not
●Ya

Traffic website :
○No traffic dan ranking website > 500.000
○Ranking website 100.000 - 500.000
●Rangking website < 100.000

Fig. 2. Feature Based Detection

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

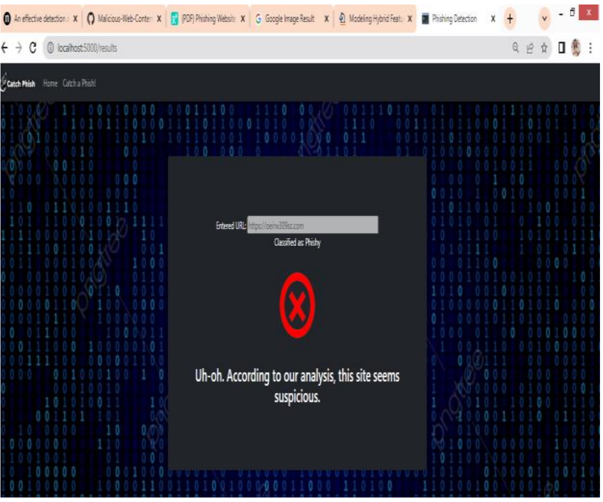


Entered URL: https://google.com

Classified as: Legitimate

This site seems safe. Happy browsing!

Fig. 3. Safe Website Detection



Entered URL: https://oeirw309sc.com

Classified as: Phishy

Uh-oh. According to our analysis, this site seems suspicious.

Fig. 4. Phishing Website Detection

VI. CONCLUSION AND FUTURE WORK

The capacity to tell the difference between distinctive legitimate URLs and phishing URLs utilising two methods they are CNN and RNN and are as a combined classifier in a cutting-edge method known as the HPDS. The HPDS classification for phishing website verification was created using a deep learning algorithm because of its ability to conduct thorough examination of text. Excellent classification accuracy of 93.28% was provided by the proposed HPDS. Based on the behavioural characteristics discovered from earlier data sets, the methodology can screen harmful websites. The HPDS was able to react immediately in real time and could verify a URL before loading it on the user's PC in an average of 25 s.

REFERENCES

[1] Almutairi, W., Saeed.F, Al-Sarem.M, Al-Shamani.M," Hybrid Filter and Wrapper Feature Selection Method for Enhancing Detection Process of Phishing Websites" https://doi.org/10.1007/978-981-16-5559-3_34 (Springer 2022).

[2] Cagatay Catal, Görkem Giray, Bedir Tekinerdogan, Sandeep Kumar & Suyash Shukla "Applications of deep learning for phishing detection: a systematic literature review" https://doi.org/10.1007/s10115-022-01672 (Springer 2022).

[3] Frank Cremer, Barry Sheehan, Michael Fortmann, Arash N. Kia, Martin Mullins, Finbarr Murphy & Stefan Materne "Cyber risk and cybersecurity: a systematic review of data availability" https://doi.org/10.1057/s41288-022-00266-6 (Springer 2022).

[4] Grega Vrbancic, Iztok Fister Jr., Vili Podgorelec " Datasets for phishing websites detection" Volume 33, December 2020, 106438 https://doi.org/10.1016/j.dib.2020.106438 (Elsevier 2020).

[5] Huseyin Ahmetoglu , Resul Das "A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions" https://doi.org/10.1016/j.iot.2022.100615 (Elsevier 2022).

[6] I. Kara, M. Ok and A. Ozaday, "Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites With Machine Learning Methods," vol. 10, pp. 124420-124428 doi: 10.1109/ACCESS.2022.3223111 (IEEE Access 2022)

[7] Imen Souiden , Mohamed Nazih Omri , Zaki Brahmi "A survey of outlier detection in high dimensional data streams" https://doi.org/10.1016/j.cosrev.2022.100463 (Elsevier 2022).

[8] Jain, A.K., Debnath, N. & Jain, A.K. APuML: "An Efficient Approach to Detect Mobile Phishing Webpages using Machine Learning" https://doi.org/10.1007/s11277-022-09707 (Springer 2022).

[9] K. Sushma, M. Jayalakshmi and T. Guha "Deep Learning for Phishing Website Detection" doi:10.1109/MysuruCon55714.2022.9972621 (IEEE Explore 2022).

[10] Kibreab Adane & Berhanu Beyene "Phishing Website Detection with and Without Proper Feature Selection Techniques: Machine Learning Approach" https://doi.org/10.1007/978-3-031-24475-9_61 (Springer 2023).

[11] M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models" Riyadh, Saudi Arabia, 2022, pp. 175-180, doi: 10.1109/CDMA54072.2022.00034 (IEEE Explore 2022).

[12] Manuel Sánchez-Paniagua, Eduardo Fidalgo, Enrique Alegre, Rocío Alaiz-Rodríguez" Phishing websites detection using a novel multipurpose dataset and web technologies features" https://doi.org/10.1016/j.eswa.2022.118010 (Elsevier 2022).

DOI: 10.17148/IJARCCE.2023.124205

[13] Mayra Macas , Chunming Wu , Walter Fuertes "A survey on deep learning for cybersecurity: Progress, challenges, and opportunities" https://doi.org/10.1016/j.comnet.2022.109032 (Elsevier 2022).

[14] Mohammed Alshehri , Ahed Abugabah , Abdullah Algarni , Sultan Almotairi "Character-level word encoding deep learning model for combating cyber threats in phishing URL detection" https://doi.org/10.1016/j.compeleceng.2022.107868 (Elsevier 2022).

[15] Mughaid, A., AlZu'bi, S., Hnaif, A. et al. "An intelligent cyber security phishing detection system using deep learning techniques, https://doi.org/10.1007/s10586-022-03604-4 (Springer 2022).

[16] N. Q. D. A. Selamat, O. Krejcar, E. Herrera-Viedma and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions" vol. 10, pp. 36429-3646, doi: 10.1109/ACCESS.2022.3151903 (IEEE Access 2022).

[17] T.A and A.John "Phishing Website Detection Using LGBM Classifier With URL-Based Lexical Features" doi: 10.1109/SILCON55242.2022.10028793 (IEEE Explore 2022).

[18] Vigneshwaran P, Roy, A.S. Sathvik, B.S. Nasirulla, D.M. Chowdary, M.L, "Multidimensional Features Driven Phishing Detection Based on Deep Learning. In: García Márquez, F.P, https://doi.org/10.1007/978-3-030-92905-3_45 (Springer 2022).

[19] Vinden Wylde, Nisha Rawindaran, John Lawrence, Rushil Balasubramanian, Edmond Prakash, Ambikesh Jayal, Imtiaz Khan, Chaminda Hewage & Jon Platts "Cybersecurity, Data Privacy and Blockchain: A Review"https://doi.org/10.1007/s42979-022-01020-4 (Springer 2022).