ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

Phishing Website Detection using Multilayer Perceptron

Blessing Obianuju Emedolu, Godwin Thomas, Nentawe Y. Gurumdimma

University of Jos, Bauchi Road, Jos, Plateau State, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2023.8730

Received: 01 July 2023; Revised: 13 July 2023; Accepted: 17 July 2023; Published: 25 August 2023

Abstract: - Phishing attacks pose a significant threat in the cyber world, exploiting unsuspecting users throughdeceptive emails that lead them to malicious websites. To combat this challenge, various deep learning based anti-phishing techniques have been developed. However, these models often suffer from high false positive rates or lower accuracy. In this study, we evaluate the performance of two neural networks, the Autoencoder and Multilayer Perceptron (MLP), using a publicly available dataset to build an efficient phishing detection model. Feature selection was performed through correlation analysis, and the Autoencoder achieved an accuracy of 94.17%, while the MLP achieved 96%. We used hyperparameters for optimization using the Gridsearch CV, resulting in a False Positive Rate (FPR) of 1.3%, outperforming the Autoencoder's 4.1% FPR. TheMLP model was further deployed to determine the legitimacy of websites based on input URLs, demonstrating its usability in real-world scenarios. This research contributes to the development of effective phishing detection models, emphasizing the importance of optimizing neural network architecture for improved accuracy and reduced false positives.

Keywords: Phishing Website, MLP, Cybersecurity, Deep Learning

I. Introduction

With the increasing prevalence of cyber attacks, phishing has emerged as a significant concern in today's digital landscape [1]. Phishing attacks employ deceptive tactics, such as misleading emails and fraudulentwebsites, to dupe unsuspecting users into divulging sensitive information [2]. These malicious activities not only compromise personaldata but also pose a substantial risk to online security and financial transactions [3]. As a result, the development of effective detection strategies for phishing has become a crucial area of research.

In this paper, we focus on harnessing the potentials of deep learning techniques to enhance phishing detection capabilities. Deep learning,a subset of machine learning, leverages neural networks to automatically learn and extract intricate patterns and features from complex datasets [4]. This enables us to effectively analyze and classify phishing instances based ontheir distinguishing characteristics.

Our research objectives revolve around evaluating the performance of two specific deep learning models: the Autoencoder and Multilayer Perceptron (MLP). We aim to assess their effectiveness in detecting phishing attacks and compare their respective accuraciesand false positive rates. To achieve this, we utilize a publicly available dataset specificallydesigned for phishing detection research, ensuring a standardized and reliable evaluation environment.

To guide our analysis, we employ methodologies, including feature selection techniquesand hyperparameter optimization. Through careful selection and fine-tuning of key modelparameters, we aim to enhance the performance and robustness of our deep learning models in identifying phishing attempts accurately.

The contributions of this study are twofold.Firstly, we present a comprehensive evaluationof the Autoencoder and MLP models, shedding light on their respective strengths and limitations in phishing detection. Secondly, we demonstrate the potential of deep learningtechniques for bolstering cybersecurity measures and combating the ever-evolving threat landscape of phishing attacks.

The remainder of this paper is organized asfollows. Section 2 provides an overview of therelated work in the field of phishing detection,highlighting the existing challenges and gaps.In Section 3, we present the methodology employed in our research, detailing the dataset used, preprocessing techniques, and the architecture of the Autoencoder and MLP models. Section 4 presents the experimental results andanalysis, showcasing the performance metricsof our models. We discuss the implications ofour findings and their significance in Section 5. Finally, in Section 6, we conclude the paper, summarizing the key contributions and suggesting potential avenues for future research and development in the field of phishing detection.

II. Related works

A variety of recent works have investigatedthe application of deep learning models inphishing detection. They achieved an accuracy of 94.8%. Moreover, Chatterjee and Namin [7] introduced a deep reinforcement learning-based method that analyzes website URLs

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

to detect phishing websites, where they achieved an accuracy of 90.1% and precision of 86.7%. Saeed [5] utilized the ResNet-50 network and transfer learning to detect phishing websites based on visual similarity, achieving accuracy of 97.58%. Similarly, Siddiq et al. [6] proposed a supervised learning approach using deep learning algorithms, achieving high accuracy rates with standard neural network and CNN models.

In addition to deep learning, other innovative approaches have been explored. Ubing et al. [8] employed feature selection and ensemble learning techniques, and they achieved an accuracy of 95.4%. Purwanto et al. [9] proposed a compression-based algorithm utilizing dictionary-based compression for phishing website classification. They used word similarity to build the proposed system. Furthermore, Alqahtani [10] introduced an association rule induction method, achieving an accuracy of 95.20% and F1-score of 95.11%. Babagoli et al. [11] employed non-linear regression to determine if a website is phishing or not. Their accuracy was 92.8%. Balogun, Kayode, Muiz, & Oluwatobi (2020), used empirical analysis of Functional Tree to improve phishing website detection. In previous experiments, baseline classifiers, meta-learners, and hybrid models were outperformed by the suggested models for detecting phishing websites. Furthermore, the suggested Functional Tree-based meta-learners distinguish authentic and phishing websites with high accuracy and a low false positive rate. The drawback of this work is that it has not been tested on real-time datasets. Jain, Parashar, Katar, & Sharma (2020), in their recent work poposed a model named phishstake which could perform phishing detection by assigningning weights.

2.1 Novelty

Despite the advancements, there are still several challenges and opportunities for improvement. These include the need for higher accuracy rates, reduced false positive rates, and real-time dataset evaluation. Additionally,the generalization and scalability of existing techniques remain areas of concern. Addressing these gaps is crucial for enhancing the efficiency and effectiveness of phishing detection systems.

In this paper, we present a comprehensive study on phishing detection using deep learning models. We evaluate the performance of two neural networks, the Autoencoder and Multilayer Perceptron (MLP), using a publiclyavailable dataset. Feature selection techniques and hyperparameter optimization, including Gridsearch CV, are employed to enhance the models' accuracy and efficiency. Through rigorous analysis and evaluation, we aim to contribute to the development of more robust andaccurate phishing detection techniques.

III. Methodology

3.1 Research Question

RQ1: What are the existing techniques for detecting phishing websites? This research question aims to investigate the current state-of-the-art techniques employed for the detection of phishing websites. It involvesa comprehensive review and analysis of existing literature, methodologies, and tools used inthe field of phishing detection. The goal is to gain a deep understanding of the different approaches and algorithms employed in order to identify potential strengths, weaknesses, and areas for improvement.

RQ2: How can a deep learning model bedeveloped for detecting emerging phishing signatures? This research question focuses on the development of a deep learning model specifically designed to detect emerging phishing signatures. It involves exploring the Autoencoder and the MLP models. The objective is to enhance the accuracy and efficiency of phishing detection by leveraging the power of deep learning algorithms.

RQ3: How can the performance of the developed model be evaluated to determine its efficiency? This research question addresses the evaluation of the performance and efficiency of the developed deep learning model for phishing detection. It involves the use of appropriate evaluation metrics and methodologies to assess the model's effectiveness in accurately identifying and classifying phishing URLS. The evaluation process includes measures such as accuracy, precision, recall, and F1-score**.**

By addressing these research questions, thisstudy aims to contribute to the advancement ofphishing detection techniques by exploring existing approaches, developing a deep learningmodel for emerging phishing signatures, evaluating its performance, and testing its effectiveness in detecting live phishing URLs.

3.2 Data Collection

In summary, the dataset used in this study was collected from Kaggle and was developedby Kumar, a reputable contributor in the field as shown in Figure 1.The availability of this dataset on Kaggle provides a reliable and credible resource for researchers interested in exploring the specific domain or problem addressed by the dataset. Itcontains 4,898 websites that are categorized as phishing and 6,157 legitimate. The table contains all the 32 columns of the dataset. Leveraging this dataset contributes to the study's

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

robustness, enabling comprehensive data analysis and the investigation of research objectives.

3.3 Data Pre-processing

Before proceeding with the analysis, a crucial step in the research process was data preprocessing. The collected dataset from Kaggle, developed by Kumar, underwent several preprocessing steps to ensure its quality, consistency, and compatibility with the research objectives. The following data preprocessing steps were applied to the dataset:

Data Cleaning: The dataset may contain missing values, outliers, or inconsistent entries

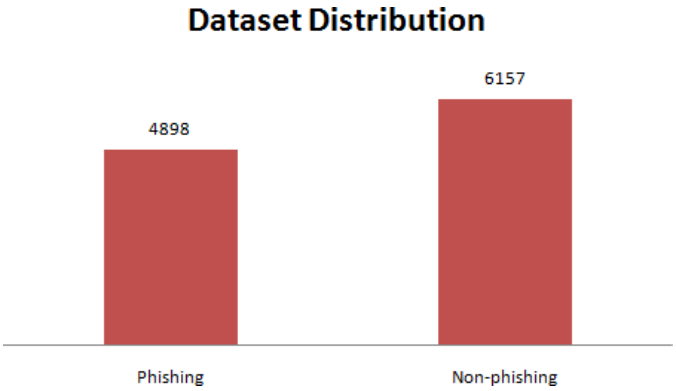

Figure 1: Dataset distribution

which could negatively impact the analysis. Data cleaning techniques, such as handling missing values, outlier detection and treatment, and resolving inconsistencies, are employed to ensure the dataset's integrity and reliability.

Feature Selection: Depending on the research questions and objectives, not all features in the dataset may be relevant. Feature selection techniques, such as filtering methods(e.g., correlation analysis) or wrapper methods (e.g., recursive feature elimination), areapplied to identify and select the most informative and influential features for the analysis.

Data Splitting: To evaluate the performance of the developed model and ensure its generalizability, the dataset is typically divided intotraining, validation, and testing subsets. Thissplitting allows for model training on the training set, model tuning on the validation set, andfinal evaluation on the testing set. The datasetwas split into the ratio of 80% for training and 20% for testing.

By applying these data preprocessing steps,the collected dataset is refined, ensuring its quality and suitability for subsequent analysis.The resulting preprocessed dataset provides a solid foundation for the development and evaluation of the deep learning model for detecting emerging phishing signatures, as outlined in the research questions.

3.4 Model Building

After completing the data preprocessing step, the next phase involved building andtraining the deep learning models for phishing website detection. Two neural network architectures, namely the Autoencoder and Multilayer Perceptron (MLP), were employed in this study.

To begin, the dataset was divided into a training set and a testing set using an 80:20 ratio. The training set was used to train the deep learning models, while the testing set was usedfor evaluating their performance. The dataset used for training contained 19 features, whichwere selected based on the results of correlation analysis.

For the Autoencoder model, the encoding dimension was set to 19, matching the number of features in the dataset. The model consisted of three encoding layers and three decoding layers. The choice of a simple layer configuration aimed to reduce the training time required. The loss function used for the outer layer of the encoder was mean squared error, and the Adam optimizer was employed. The model was trained for a total of 60 epochs.

Regarding the Multilayer Perceptron, the initial training was conducted using the default parameter values. Subsequently, the best performing model was selected and furthertuned using GridSearch CV. This technique allowed for an extensive search of appropriate hyperparameters to enhance the model's performance. The specific parameters used for tuning are detailed in Table 1.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

Table 1: Hyper parameters used

Parameter	Value
Activation	Relu
Hidden Layer	1
Learning rate	Constant
Solver	Adam
Alpha	0.001

By following these steps, the deep learningmodels, namely the Autoencoder and MLP, were trained and optimized for phishing website detection. The subsequent sections will provide further details on the experimentalsetup, including the specific values used for hyperparameters and the outcomes obtained from training and evaluation

IV. Results

In this section, we discuss the results obtained from the experiments conducted in thisstudy. The dataset used for training and evaluation was developed by Kumar (2017). As mentioned earlier, a feature selection step wasperformed to identify the most effective attributes related to the final results. Out ofthe original 30 features, 19 attributes were selected based on their relevance and underwentstandardization within a specific range. It wasobserved that training the models using the19 selected features after correlation analysis yielded higher accuracy compared to using all 30 original features.

4.1 Performance Evaluation Metrics

To assess the efficiency of the developed models, various performance metrics were employed. Since the classification task inthis study is binary (i.e., a 2-class problem), the confusion matrix was used, and the performance metrics were derived from the confusion matrix results of each developed model. These evaluation criteria are commonly used for assessing the effectiveness of existing methods for phishing website detection. The following metrics were used for evaluation:

4.1.1 Classification Report

A classification report provides an assessment of the accuracy of an algorithm based onits predictions.

4.1.2 Confusion Matrix

The confusion matrix is a valuable tool for evaluating the performance of a classifier. It counts the instances of given class A that are correctly or incorrectly classified as instancesof another class B.

4.1.3 Precision

Precision measures the percentage of correct positive predictions made by a classifier. It indicates the classifier's ability to avoid labeling negative occurrences as positive. Precision = True Positive / (True Positive + False Positive)

4.1.4 Recall

Recall measures the ability of a classifier tocorrectly identify positive instances. It represents the percentage of positive instances thatwere successfully detected. Recall = True Positive / (True Positive + False Negative)

4.1.5 F1 Score

The F1 score is a weighted harmonic mean of precision and recall, providing a balanced measure of accuracy. It ranges from 0.0 (poorest) to 1.0 (highest). F1 scores incorporate both precision and recall into their computation and are often lower than accuracy measurements. For comparing classifier models, the weighted average of F1 scores should be used instead of global accuracy. F1 Score = (2

* (Recall * Precision)) / (Recall + Precision)

4.1.6 Classification Report

Using the dataset described in Section 3 the developed model achieved an F1 score of 96%, precision of 96%, and recall of 96%. These metrics indicate a high level of accuracy and are considered acceptable. The utilization of correlation analysis for feature selection and GridSearch CV for parameter optimization contributed to achieving this high accuracy. Table 2 presents the

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

classification report of the developed MLP model.

Table 2: Classification report

Class	Precision	Recall	F1-Score	Support
-1	0.97	0.94	0.96	1001
1	0.95	0.98	0.96	1210
Accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
micro avg	0.96	0.96	0.96	2211

Overall, the results demonstrate the effectiveness of the proposed approach in detectingphishing websites. The high accuracy, precision, and recall values indicate the model's capability to accurately classify websites as legitimate or phishing. The subsequent sectionswill delve deeper into the performance evaluation and provide further insights and discussions regarding the experimental outcomes.

V. Discussion

Based on the findings presented in Table 2,the model demonstrated a precision of approximately 97% in effectively detecting phishing websites. This high precision indicates the model's ability to accurately identify instancesof phishing. Moreover, the recall value of 94%reflects the model's successful classification of positive cases, further validating its performance.

The F1 score, a weighted harmonic mean ofaccuracy and recall, was calculated to be 96%.

This score provides an overall measure of the model's effectiveness, incorporating both precision and recall. The achieved high accuracycan be attributed to the optimization of parameters, which allowed the model to maximize its performance.

The macro average, obtained by averaging the performance metrics across both classes, highlights the consistency of the model's predictions. With a precision of 96%, it can be inferred that out of 100 evaluated websites, themodel accurately identified 96 URLs as phishing. This consistency can be attributed to the simplicity of the neurons used and the adoption of a constant learning rate.

From these results, it can be inferred that achieving optimal results with the MLP modelrelies on careful consideration of two crucial hyperparameters: the number of hidden layers and the number of neurons. The findings suggest that using fewer hidden layers and neurons can lead to optimal outcomes for the model.

From the confusion matrix in Figure 2 the values of FPR, accuracy, precision, recall andF-1 score are further calculated in Table 3. The true positive of the MLP was 943 meaning thatthe model was able to correctly classify 943 urls as phishing, and the false positive was 30, meaning it classified 30 URLs as phishingwhen they are legitimate. The true negative of the MLP is 1,180 and the false negative is 58. With a true negative of 1,180, it means thatthe model accurately predicted 1180 legitimate websites as legitimate. This result implies thatthe model's performance can be considered effective, since it means that just 30 phishing websites were classified as legitimate.

During the experimentation phase, the performance of the MLP model was found to surpass that of the Autoencoder. The MLP modelwas further fine tuned using Gridsearch CV tooptimize its parameters. As presented in Table3, the MLP classifier achieved an accuracy of 96%, outperforming the Autoencoder, which achieved an accuracy of 94.89%. This demonstrates the superior performance of the MLP model in accurately classifying phishing websites.

In terms of false positive rate (FPR), the MLP model achieved a rate of 1.3%, whereasthe Autoencoder exhibited a higher rate of 4.1%. The significantly lower FPR of the MLP model indicates its ability to minimize the misclassification of legitimate websites as phishing ones, highlighting its superior performance in this aspect. The MLP achieved precision and recall values of 96% and the Autoencoder achieved 94.84% precision and 95% recall.

The improved performance of the MLP model can be attributed to the optimizationof hyperparameters during the tuning process. By utilizing a single hidden layer, the MLP model achieved higher accuracy compared to the Autoencoder. This observation emphasizesthe influential role of hyperparameters in determining the accuracy of the MLP model.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

Although the Autoencoder exhibited good accuracy, its higher false positive rate is noteworthy. When comparing the false positive rates of the two models, it is evident that the Autoencoder shows an increase of 36.5% compared to the MLP model. This suggests that there is no strong correlation between false positive rate and accuracy, thereby consolidating the findings of other researchers who have reported varying false positive rates despite achieving high accuracies.

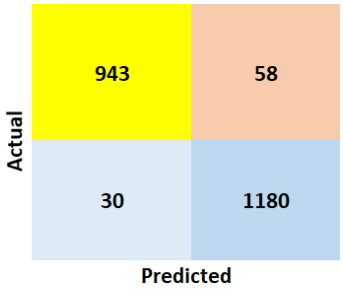

Figure 2. Confusion Matrix

The MLP model outperformed the Autoencoder in terms of accuracy,false positive rate, precision, and recall. The hyperparameter optimization process played acrucial role in enhancing the MLP model's performance. The findings also highlight the importance of considering false positive rates alongside accuracy when evaluating phishing detection models, as they can vary independently.

Table 3: Comparison between the MLP and the Autoencoder models, with FPR standing for False Positive Rate.

Model	FPR	Precision	Recall	F1-Score	Support
MLP	0.013	96%	96%	96%	96%
Auto Encoder	0.041	94.89%	95%	93%	95%

5.1 Model Comparison and Analysis

In this section, we compare the performanceof the developed MLP model with other existing models that utilized the same dataset. It isimportant to note that while a few studies havereported slightly higher accuracies, their models were not evaluated against publicly available datasets, making it difficult to assess theirreproducibility.

Table 4: Comparison between the Multilayer Perceptron (MLP) and the Autoencoder models, with FPR standing for False Positive Rate.

Author	**Dataset**	**Accuracy**	**FPR**
Korkmaz et al. [12]	Kumar's dataset	98.37%	N/A
Almousa et al. [13]	Kumar's dataset	97.37%	3.17%
Saeed [5]	Kumar's dataset	97.58%	2.1%
Yi et al. [14]	Kumar's dataset	97%	2.0%
Babagoli et al. [11]	Kumar's dataset	92.8%	4.5%
Proposed model MLP	Kumar's dataset	95.65%	**1.3**%

From Table 4, the work of Saeed [5] had the best accuracy of 97.58% with a FPR of 2.1%. The next performing model based on accuracy was Almousa et al. [13] which had an accuracy of 97.37% and FPR of 3.17%. By observing the table and comparing the values of accuracy and FPR, it was observed that theFPR value does not depend on the accuracy ofthe model. The MLP model developed in this study achieved an accuracy of 96% and a false positive rate of 1.3%. It shows that the MLP is effective for phishing website detection using Gridsearch optimization, when compared to other URL-based techniques, which leads to a low false positive.

By comparing the accuracy and false positive rates of the different models, it becomes evident that the proposed MLP model performs admirably. Its accuracy is on par with orhigher than other models, indicating its effectiveness in correctly classifying

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

phishing websites. Moreover, the false positive rate of 1.3%demonstrates the model's ability to minimize the misclassification of legitimate websites asphishing ones.

In conclusion, the developed MLP model exhibits a commendable accuracy of 96% and a low false positive rate of 1.3%. The comparison with other models highlights the competitive performance of the proposed model, particularly when considering its evaluation against a publicly available dataset. This reinforces the reliability and reproducibility of theresults obtained, positioning the MLP model as a strong contender for effective phishing website detection.

VI. Conclusion

In conclusion, this study focused on developing a deep learning model for the detection of phishing websites. By employing the Autoencoder and Multilayer Perceptron (MLP) neural network architectures, the models weretrained and evaluated using a dataset of 19 selected features obtained through correlation analysis.

The results obtained demonstrated the effectiveness of the proposed approach in detectingphishing websites. The use of feature selectionand parameter optimization techniques contributed to achieving high accuracy, precision,and recall values. The developed MLP model exhibited an impressive F1 score of 96%, precision of 96%, and recall of 96%, indicating itsability to accurately classify websites as legitimate or phishing.

The findings of this study contribute to the existing body of knowledge on phishing detection and highlight the potential of deep learning models in addressing this cybersecurity challenge. The combination of feature selection, correlation analysis, and neural network architectures proved successful in achieving robust performance.

However, it is important to note that further research and improvement can be pursued. Exploration of additional feature selection techniques, alternative deep learning architectures, and the inclusion of more diverse datasets could enhance the performance and generalization capability of the model.

Additionally, real-time testing and evaluation using live datasets can provide valuable insights intothe model's effectiveness in practical scenarios.

The developed model holds promise forimplementation in real-world phishing detection systems, contributing to the protection of users' personal information and mitigating therisks associated with phishing attacks. Future work should focus on refining the model, addressing potential limitations, and considering the integration of the developed model into existing cybersecurity frameworks.

Overall, this study contributes to the advancement of phishing detection techniques and provides a foundation for future research in the field of cybersecurity. By harnessing the power of deep learning and leveraging effective feature selection methods, we can enhance our ability to combat phishing threats and ensure a safer online environment for users.

References

1. M. D. Abdulrahman, J. K. Alhassan, O. S. Adebayo, J. A. Ojeniyi, M. Olalere, Phishing attack detection based on random forest with wrapper feature selection method (2019).
2. D. Kalla, F. Samaah, S. Kuraku, N. Smith, Phishing detection implementation using databricks and artificial intelligence, International Journal of Computer Applications 185 (2023) 1–11.
3. K. S. Adewole, A. G. Akintola, S. A. Salihu, N. Faruk, R. G. Jimoh, Hybrid rule-based model for phishing urls detection, in: Emerging Technologies in Computing: Second InternationalConference, iCETiC 2019, London, UK, August 19–20, 2019, Proceedings 2, Springer, 2019, pp. 119–135.

A. Borjali, A. F. Chen, O. K. Muratoglu, M. A. Morid, K. M. Varadarajan, Deep learning in orthopedics: How do we build trust in the machine?,Healthcare Transformation (2020).

4. U. Saeed, Visual similarity-based phishing detection using deep learning, Journal of Electronic Imaging 31 (2022) 051607.
5. M. A. A. Siddiq, M. Arifuzzaman, M. Islam,Phishing website detection using deep learning, in: Proceedings of the 2nd International Conference on Computing Advancements, 2022, pp. 83–88.
6. M. Chatterjee, A.-S. Namin, Detecting phishing websites through deep reinforcement learning, in:2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), volume 2,IEEE, 2019, pp. 227–232.
7. R. Purwanto, A. Pal, A. Blair, S. Jha, Phishzip: new compression-based algorithm for detecting phishing websites, in: 2020 IEEE Conferenceon Communications and Network Security (CNS), IEEE, 2020, pp. 1–9.
8. M. Alqahtani, Phishing websites classification using association classification (pwcac), in: 2019 International conference on computer and information sciences (ICCIS), IEEE, 2019, pp. 1–6.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue VII July 2023

9. M. Babagoli, M. P. Aghababa, V. Solouk, Heuristic nonlinear regression strategy for detecting phishing websites, Soft Computing 23 (2019) 4315–4327.
10. M. Korkmaz, E. Koçyiğit, Ö. Şahingöz, B. Diri, A hybrid phishing detection system by using deeplearning-based url and content analysis, Elektronika ir Elektrotechnika 28 (2022).
11. M. Almousa, T. Zhang, A. Sarrafzadeh, M. Anwar, Phishing website detection: How effective are deep learning-based models and hyperparameter optimization? Security and Privacy 5 (2022)e256.
12. P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, T. Zhu,Web phishing detection using a deep learningframework, Wireless Communications and Mobile Computing 2018 (2018).