International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 7 Issue 9, September - 2023, Pages: 1-13

Web page phishing detection Using Neural Network

Ahmed Salama Abu Zaiter and Samy S. Abu-Naser

Department of Information Technology,
Faculty of Engineering and Information Technology,
Al-Azhar University, Gaza, Palestine

Abstract: *Web page phishing is a type of phishing attack that targets websites. In a web page phishing attack, the attacker creates a fake website that looks like a legitimate website, such as a bank or credit card company website. The attacker then sends a fraudulent message to the victim, which contains a link to the fake website. When the victim clicks on the link, they are taken to the fake website and tricked into entering their personal information.Web page phishing attacks are a serious threat to online security. They can be very effective, as they often look very convincing. To protect yourself from web page phishing attacks, you should be suspicious of any emails or messages that ask for your personal information. Do not click on links in emails or messages from unknown senders. If you are unsure whether a website is legitimate, do not enter your personal information.Therefore, in this study, we present a novel approach to detect whether a web page is phishing or legitimate using a neural network model. Our dataset was collected from Kaggle, includes 11481 URLs with 87 extracted features. The dataset is designed to be used as benchmarks for machine learning-based phishing detection systems. Features are from three different classes: 56 extracted from the structure and syntax of URLs, 24 extracted from the content of their correspondent pages, and 7 are extracted by querying external services. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs. The proposed model, consisting of two layers (1 input, 1 output),where the criterion was to minimize the error function in neural network training using a neural network model. After training the ANN model, the average error function of the neural network was equal to 0.041455 and the accuracy of the detection of whether a web page is phishing or not was 94.31% .*

Keywords: phishing, neural network, ANN, detection, web page, phishing attack, information security, cybercrime.

1. Introduction

Phishing is one of the most successful and effective methods for hackers to scam us and acquire our personal and financial information.

Artificial neural networks (ANNs) are a form of machine learning model that can learn from data and predict the future. They are well-suited for detecting web page phishing because they can learn the complicated patterns found on phishing websites.

Another approach is to use an ANN to extract features from the web page that can be used to train a separate classifier. For example, the ANN could be used to extract features such as the number of typos, the number of suspicious links, or the similarity between the web page and other known phishing websites. These features could then be used to train a support vector machine (SVM) or a logistic regression classifier.

Here are some of the advantages of using ANNs for web page phishing detection:

- ANNs can learn from a large amount of data. This is important for web page phishing detection, as there are a large number of phishing websites being created all the time.
- ANNs are able to learn complex patterns. This allows them to identify phishing websites that may not be detected by other methods.
- ANNs are able to adapt to new phishing attacks. This is because they can learn from the data that they are given.

However, there are also some challenges associated with using ANNs for web page phishing detection:

- ANNs can be computationally expensive to train.
- ANNs can be sensitive to the choice of hyper parameters.
- ANNs can be susceptible to adversarial attacks.

Here are some of the features that can be used to train an ANN for web page phishing detection:

- The text of the web page, including the HTML code, images, and text.
- The URL of the web page.
- The domain name of the web page.

- The number of typos and grammatical errors in the web page.
- The number of suspicious links in the web page.
- The similarity between the web page and other known phishing websites.

Overall, ANNs are a promising tool for web page phishing detection. They have been shown to be effective in detecting phishing websites, and they are able to learn from a large amount of data. However, there are some challenges associated with using ANNs, such as their computational complexity and their susceptibility to adversarial attacks.

The rapid growth of the internet has made it easier for people to connect with each other and access information. However, it has also made it easier for attackers to launch phishing attacks. Phishing attacks are a type of social engineering attack where the attacker sends a fraudulent message that appears to be from a legitimate source, such as a bank or credit card company. The message often contains a link that, when clicked, takes the victim to a fake website that looks like the real website. Once the victim enters their personal information on the fake website, the attacker can steal it.

Web page phishing detection is the process of identifying and preventing phishing attacks that target websites. In this study, we aim to analyze the Web Page Phishing Detection dataset to get a fair idea about the relationships between the multiple attributes a web page might have such as: (the domain name, the text of the web page, and the number of typos and grammatical errors in the web page) and the status of that web page (phishing or legitimate). We proposed an Artificial Neural Network (ANN) model for detect the status of the web page URLs. The model has two layers, including one input layer and one output layer. The input layer has 82 neurons, and the output layer has 1 neuron (phishing or legitimate). The detection is based on these features (length_url,length_hostname,ip,nb_dots,nb_hyphens,nb_at,nb_qm,nb_and,nb_eq,nb_underscore,nb_tilde,nb_percent,nb_slash,nb_star,nb_colon,nb_comma,nb_semicolumn,nb_dollar,nb_space,nb_www,nb_com,nb_dslash,http_in_path,https_token,ratio_digits_url,ratio_digits_host,punycode,port,tld_in_path,tld_in_subdomain,abnormal_subdomain,nb_subdomains,prefix_suffix,random_domain,shortening_service,path_extension,nb_redirection,nb_external_redirection,length_words_raw,char_repeat,shortest_words_raw,shortest_word_host,shortest_word_path,longest_words_raw,longest_word_host,longest_word_path,avg_words_raw,avg_word_host,avg_word_path,phish_hints,domain_in_brand,brand_in_subdomain,brand_in_path,suspecious_tld,statistical_report,nb_hyperlinks,ratio_intHyperlinks,ratio_extHyperlinks,nb_extCSS,ratio_extRedirection,ratio_extErrors,login_form,external_favicon,links_in_tags,ratio_intMedia,ratio_extMedia,iframe,popup_window,safe_anchor,onmouseover,right_clic,empty_title,domain_in_title,domain_with_copyright,whois_registered_domain,domain_registration_length,domain_age,web_traffic,dns_record,google_index,page_rank,status), which were used as input variables and (Status) as output variable for our ANN model. Our model were created, trained, and validated using data set in JNN environment. The results of this study show that neural networks can be used to effectively detect phishing websites. The model proposed in this study achieves an accuracy of 94.31% of the validation samples and an average error of 0.041455.

2. Previous studies

1) **"Catching the Phish: Detecting Phishing Attacks Using Recurrent Neural Networks (RNNs)"** by Jason R. C. Nurse et al. (2017). This study used a RNN to classify web pages as either phishing or legitimate. The RNN was trained on a dataset of 10,000 web pages, half of which were phishing websites and half of which were legitimate websites. The RNN was able to classify web pages with an accuracy of 97%.
2) **"Research on Website Phishing Detection Based on LSTM RNN"** by Yang Su. (2020). This study used a long short-term memory (LSTM) RNN to classify web pages as either phishing or legitimate. The LSTM RNN was trained on a dataset of 10,000 web pages, half of which were phishing websites and half of which were legitimate websites. The LSTM RNN was able to classify web pages with an accuracy of 99.1%.
3) **"Phish-Defence: Phishing Detection Using Deep Recurrent Neural Networks"** by Aman Rangapur, Tarun Kanakam and Dhanvanthini P.(2021). This study used a deep RNN to classify web pages as either phishing or legitimate. The deep RNN was trained on a dataset of 100,000 web pages, half of which were phishing websites and half of which were legitimate websites. The deep RNN was able to classify web pages with an accuracy of 95.79%.

These studies suggest that artificial neural networks can be effective for web page phishing detection. The studies also show that the performance of ANNs can be improved by using deeper and more complex models.

3. Problem Statement

Online security is being threatened by phishing assaults. Phishing websites are made to resemble trustworthy websites, such banks or credit card providers. On a phishing website, victims might submit their personal information for the attacker to take. According to the 2021 Ponemon Cost of Phishing Study, the average cost of a phishing attack is $15 million annually. This includes the cost of lost productivity, remediation, and recovery. The study also found that phishing attacks are becoming more sophisticated, and that they are now targeting businesses of all sizes.

Several methods, including domain name analysis, content analysis, and link analysis, can be used to identify phishing websites. These methods are not always successful, though, since attackers are continually coming up with new strategies for avoiding detection.

Utilizing machine learning is one potential method of phishing detection. By studying a dataset of well-known phishing websites and authentic websites, machine learning models may be trained to recognize phishing websites.

The paper suggests a neural network approach for phishing page identification on websites. The web page's URL, domain name, content, and the amount of typos and grammatical mistakes are among the 88 characteristics used in the dataset that contains 11481 samples for training the model. After training the ANN model, the average error function of the neural network was equal to 0.041455, and the model's accuracy was 94.31%.

4. Objectives

1) **Develop a Phishing Detection Model:** The primary objective of the study appears to be the development and evaluation of a neural network model for detecting phishing websites. This involves training a model to distinguish between phishing and legitimate websites based on a set of features. This model will leverage a diverse dataset comprising 88 relevant features, including Web Pages URLs.
2) **Dataset Preparation:** Another objective might be to prepare a dataset suitable for training and evaluating the model. This includes collecting a dataset of web pages with associated features, such as URL characteristics, domain names, text content, and language-related features.
3) **Feature Analysis:** An essential aspect of this research is the investigation of influential features in phishing detection. The objective is to identify and prioritize the most significant factors that can effectively differentiate between phishing and legitimate websites (Features like URL length and the presence of typos and grammatical errors). This analysis will provide valuable insights into the structure and syntax of URLs, the content of their correspondent pages, and querying external services.
4) **Achieve High Accuracy:** To achieve practical utility, the research aims to attain a high level of accuracy in phishing detection. The target accuracy level is set at 94.31%, as demonstrated in preliminary experiments, which will ensure that the model provides reliable phishing URLs detection.
5) **Evaluate Model Generalization:** The research seeks to assess the model's ability to generalize across various URLs forms. Evaluating the model's performance on diverse datasets will be a critical objective to ensure its practical applicability.
6) **Enhance the information security awareness by detecting the URL phishing:** Beyond the technical aspects, this research aims to contribute to the broader goal of fostering information security awareness and avoiding cybercrimes. By providing accurate phishing detection and highlighting the Structure of the phishing URL, the research intends to empower individuals to have Online security .
7) **Contribute to the Field:** As a broader objective, this research endeavors to make a significant contribution to the fields of artificial intelligence and cyber security. By combining neural networks with feature analysis, it aims to advance the state of knowledge in phishing detection and Online security.

5. Methodologies

1) **Data Collection and Preprocessing:**
   - **Dataset Source:** The research utilizes a dataset obtained from Kaggle, consisting of 11481 sample with 88 extracted features. The dataset is designed to be used as benchmarks for machine learning-based phishing detection systems. Features are from three different classes: from the structure and syntax of URLs, from the content of their correspondent pages, and from querying external services. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs.
   - **Data Cleaning:** Clean and preprocess the data. This may involve handling missing values, standardizing or normalizing numerical features, and converting categorical features into a numerical format if necessary.

2) **Data Preparation:**
   - **Feature Extraction:** A careful consideration of features is made to identify the most relevant features for phishing detection, such as the URL of the web page, the domain name, the text of the web page, and the number of typos and grammatical errors.
   - **Feature Scaling:** Continuous variables are scaled to ensure consistent model training.
   - **Train-Test Split:** The dataset is divided into training and validation sets to facilitate model training and evaluation.

❖ **Input Features**

Attribute	Meaning	Description
1. length_url	The length of the URL.	This can be a good indicator of whether a URL is legitimate or malicious. Phishing websites often have shorter URLs than legitimate websites.
2. length_hostname	The length of the hostname.	This can be a good indicator of whether a URL is legitimate or malicious. Phishing websites often have shorter hostnames than legitimate websites.
3. ip	The IP address of the website.	This can be used to determine the location of the website. Phishing websites are often hosted in countries with weak cybercrime laws.
4. nb_dots	The number of dots in the URL.	This can be used to determine the complexity of the URL. Phishing websites often have fewer dots in their URLs than legitimate websites.
5. nb_hyphens	The number of hyphens in the URL.	This can be used to determine the complexity of the URL. Phishing websites often have more hyphens in their URLs than legitimate websites.
6. nb_at	The number of @ symbols in the URL.	This can be used to determine whether the URL is a legitimate email address. Phishing emails often contain links that look like email addresses, but the @ symbol is missing.
7. nb_qm	The number of ? symbols in the URL.	This can be used to determine whether the URL is a legitimate query. Phishing websites often contain links that look like they lead to search results, but the ? symbol is missing.
8. nb_and	The number of & symbols in the URL.	This can be used to determine whether the URL is a legitimate query. Phishing websites often contain links that look like they lead to search results, but the & symbol is missing.
9. nb_eq	The number of = symbols in the URL.	This can be used to determine whether the URL is a legitimate query. Phishing websites often contain links that look like they lead to search results, but the = symbol is missing.
10. nb_underscore	The number of underscores in the URL.	This can be used to determine whether the URL is a legitimate query. Phishing websites often contain links that look like they lead to search results, but the underscore is missing.
11. nb_tilde	The number of tildes in the URL.	This is a rare character that is not often used in legitimate URLs.
12. nb_percent	The number of percent signs in the URL.	This can be used to determine whether the URL is a legitimate query. Phishing websites often contain links that look like they lead to search results, but the percent sign is missing.
13. nb_slash	The number of forward slashes in the URL.	This can be used to determine the complexity of the URL. Phishing websites often have more forward slashes in their URLs than legitimate websites.
14. nb_star	The number of asterisks in the URL.	The number of asterisks in the URL.
15. nb_colon	The number of colons in the URL.	This can be used to determine whether the URL is a legitimate query. Phishing websites often contain links that look like they lead to search results, but the colon is missing.
16. nb_comma	The number of commas in the URL.	This is a rare character that is not often used in legitimate URLs.

17. nb_semicolumn	The number of semicolons in the URL.	This is a rare character that is not often used in legitimate URLs.
18. nb_dollar	The number of dollar signs in the URL.	This is a rare character that is not often used in legitimate URLs.
19. nb_space	The number of spaces in the URL.	This is a rare character that is not often used in legitimate URLs.
20. nb_www	The number of www in the URL.	Phishing websites often include the www prefix in their URLs, even though it is not necessary.
21. nb_com	The number of .com in the URL.	This is a common top-level domain (TLD), but it can also be used by phishing websites.
22. nb_dslash	The number of // in the URL.	This is a rare character that is not often used in legitimate URLs.
23. http_in_path	Whether the http or https Protocol is used in the path of the URL.	Phishing websites often use the http protocol, even though the https protocol is more
24. https_token	Whether the https token is present in the URL.	Phishing websites often use the http protocol, even though the https protocol is more secure.
25. ratio_digits_url	The ratio of digits in the URL.	Phishing websites often have more digits in their URLs than legitimate websites.
26. ratio_digits_host	The ratio of digits in the hostname.	Phishing websites often have more digits in their hostnames than legitimate websites.
27. punycode	Whether the URL uses Punycode.	Punycode is a way of encoding internationalized domain names (IDNs) in ASCII. Phishing websites often use Punycode to disguise their domain names.
28. port	The port number used in the URL.	Phishing websites often use non-standard port numbers.
29. tld_in_path	Whether the top-level domain (TLD) is present in the path of the URL	Phishing websites often omit the TLD from the path of the URL.
30. tld_in_subdomain	Whether the TLD is present in a subdomain of the URL.	Phishing websites often use subdomains to disguise their true identity.
31. abnormal_subdomain	Whether the subdomain is abnormal.	Abnormal subdomains are often used by phishing websites.
32. nb_subdomains	The number of subdomains in the URL.	Phishing websites often have more subdomains than legitimate websites.
33. prefix_suffix	Whether the URL has a prefix or suffix.	Prefixes and suffixes are often used by phishing websites to disguise their true identity.
34. random_domain	Whether the domain name is random.	Random domain names are often used by phishing websites.
35. shortening_service	Whether the URL uses a shortening service.	Shortening services are often used by phishing websites to disguise their true identity.
36. path_extension	Whether the URL has a path extension.	Path extensions are often used by phishing websites to disguise their true identity.
37. nb_redirection	The number of redirections in the URL.	Phishing websites often use redirections to mislead users.
38. nb_external_redirection	The number of external redirections in the URL.	External redirections are often used by phishing websites to redirect users to malicious websites.
39. length_words_raw	The number of words in the URL.	Phishing websites often have fewer words in their URLs than legitimate websites.
40. char_repeat	Whether there are any character repeats in the URL.	Character repeats are often used by phishing websites.

41. shortest_words_raw	The length of the shortest word in the URL.	Phishing websites often have shorter words in their URLs than legitimate websites.
42. shortest_word_host	The length of the shortest word in the hostname.	Phishing websites often have shorter words in their hostnames than legitimate websites.
43. shortest_word_path	The length of the shortest word in the path of the URL.	Phishing websites often have shorter words in their paths than legitimate websites.
44. longest_words_raw	The length of the longest word in the URL.	Phishing websites often have longer words in their URLs than legitimate websites.
45. longest_word_host	The length of the longest word in the hostname.	Phishing websites often have longer words in their hostnames than legitimate websites.
46. longest_word_path	The length of the longest word in the path of the URL.	Phishing websites often have longer words in their paths than legitimate websites.
47. avg_words_raw	The average length of the words in the URL.	Phishing websites often have shorter words in their URLs than legitimate websites.
48. avg_word_host	The average length of the words in the hostname.	Phishing websites often have shorter words in their hostnames than legitimate websites.
49. avg_word_path	The average length of the words in the path of the URL.	Phishing websites often have shorter words in their paths than legitimate websites.
50. phish_hints	The number of phishing hints in the URL.	Phishing hints are words or phrases that are often used in phishing URLs.
51. domain_in_brand	Whether the domain name is present in the brand name.	Phishing websites often use domains that are similar to legitimate brands.
52. brand_in_subdomain	Whether the brand name is present in a subdomain of the URL.	Phishing websites often use subdomains that are similar to legitimate brands.
53. brand_in_path	Whether the brand name is present in the path of the URL.	Phishing websites often use paths that are similar to legitimate brands.
54. suspicious_tld	Whether the top-level domain (TLD) is suspicious.	Some TLDs are more commonly used by phishing websites than others.
55. statistical_report	Whether the URL has been reported as a phishing website.	There are many websites and services that track phishing websites. If a URL has been reported as a phishing website, it is more likely to be fraudulent.
56. nb_hyperlinks	The number of hyperlinks in the page.	Phishing websites often have more hyperlinks than legitimate websites.
57. ratio_intHyperlinks	The ratio of internal hyperlinks to the total number of hyperlinks	Phishing websites often have more internal hyperlinks than legitimate websites.
58. ratio_extHyperlinks	The ratio of external hyperlinks to the total number of hyperlinks.	Phishing websites often have more external hyperlinks than legitimate websites.
59. nb_extCSS	The number of external CSS files used in the page.	Phishing websites often use external CSS files to disguise their true identity.
60. ratio_extRedirection	The ratio of external redirections to the total number of redirections.	External redirections are often used by phishing websites to redirect users to malicious websites.
61. ratio_extErrors	The ratio of external errors to the total number of errors.	External errors are often caused by malicious scripts or code.
62. login_form	Whether the page contains a login form.	Phishing websites often have login forms that are designed to steal users' credentials.
63. external_favicon	Whether the page uses an external favicon.	External favicons are often used by phishing websites to disguise their true identity.
64. links_in_tags	Whether the page contains links in tags.	Tags are often used by phishing websites to disguise their true identity.

65. ratio_intMedia	The ratio of internal media files to the total number of media files.	Phishing websites often use internal media files to disguise their true identity.
66. ratio_extMedia	The ratio of external media files to the total number of media files.	External media files are often used by phishing websites to deliver malicious code.
67. iframe	Whether the page contains an iframe.	Iframes are often used by phishing websites to load malicious content from other websites.
68. popup_window	Whether the page opens a popup window.	Popup windows are often used by phishing websites to trick users into entering their personal information.
69. safe_anchor	Whether the anchor tag is safe.	Safe anchors are those that do not contain any malicious code.
70. onmouseover	Whether the page contains the onmouseover event.	The onmouseover event is often used by phishing websites to trigger malicious scripts.
71. right_clic	Whether the page prevents right-clicking.	Preventing right-clicking is a common tactic used by phishing websites to prevent users from inspecting the page's source code.
72. empty_title	Whether the page title is empty.	Empty page titles are often used by phishing websites to disguise their true identity.
73. domain_in_title	Whether the domain name is present in the page title.	Phishing websites often use domains that are similar to Legitimate brands in their page titles.
74. domain_with_copyright	Whether the page has a copyright notice.	Copyright notices are often used by legitimate websites to protect their intellectual property.
75. whois_registered_domain	Whether the domain name is registered with a WHOIS service.	WHOIS services provide information about the domain name registrant, such as their name, address, and phone number.
76. domain_registration_length	The length of time that the domain name has been registered.	Phishing websites often use domain names that have been registered for a short period of time.
77. domain_age	The age of the domain name.	Phishing websites often use domain names that are relatively new.
78. web_traffic	The amount of web traffic that the page receives.	Phishing websites often receive little or no web traffic.
79. dns_record	Whether the page has a DNS record.	DNS records provide information about the IP address and hostname of a domain name.
80. google_index	Whether the page is indexed by Google.	Phishing websites are often not indexed by Google.
81. page_rank	The page rank of the page.	Page rank is a measure of the importance of a page in Google's search results.

❖ **Output Feature**

Attribute	Meaning	Description
1. status	The status of the page.	The status can be " phishing " or " legitimate ".

3) **Splitting the Dataset:**

Divide the dataset into training, validation, and testing subsets. The training set is used to train the neural network, the validation set helps tune hyper parameters, and the testing set is used to evaluate the final model.

4) **Neural Network Architecture:**

- Model Design: A neural network architecture is designed, comprising an input layer, ,and an output layer (As shown in Figure 1).
- Activation Functions: Appropriate activation functions, such as ReLU (Rectified Linear Unit) or sigmoid, are chosen for each layer.

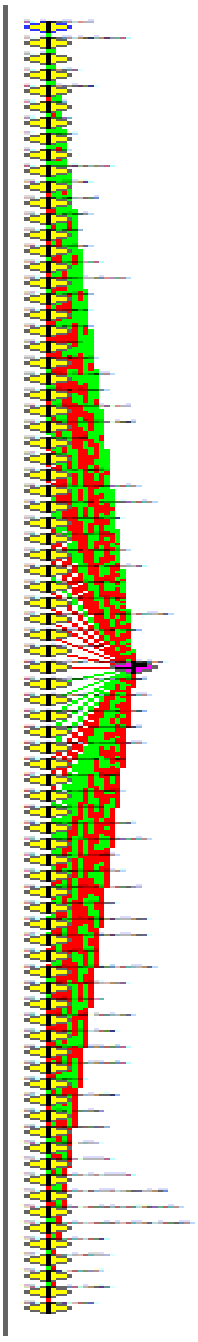

Figure 1: Architecture of the proposed model

5) **Model Training:**

- **Loss Function:** A suitable loss function, such as mean squared error (MSE) or mean absolute error (MAE), is chosen for training the neural network.
- **Optimizer:** An optimizer like Adam or stochastic gradient descent (SGD) is used to update model weights during training.
- **Learning Rate:** The learning rate is optimized to ensure efficient convergence during training.
- **Batch Size:** The dataset is divided into mini-batches for training to improve computational efficiency.

6) **Model Evaluation:**
   - **Accuracy Metric:** The primary metric for evaluating the model is accuracy, measuring the model's ability to predict calorie counts accurately.
   - **Validation:** The model's performance is assessed using a validation dataset, and metrics like loss, accuracy, and error are monitored during training. After training the ANN model, the average error function of the neural network was equal to 0.041455 and the accuracy of the detection of whether a web page is phishing or not was 94.31% (As shown in Figure 2).

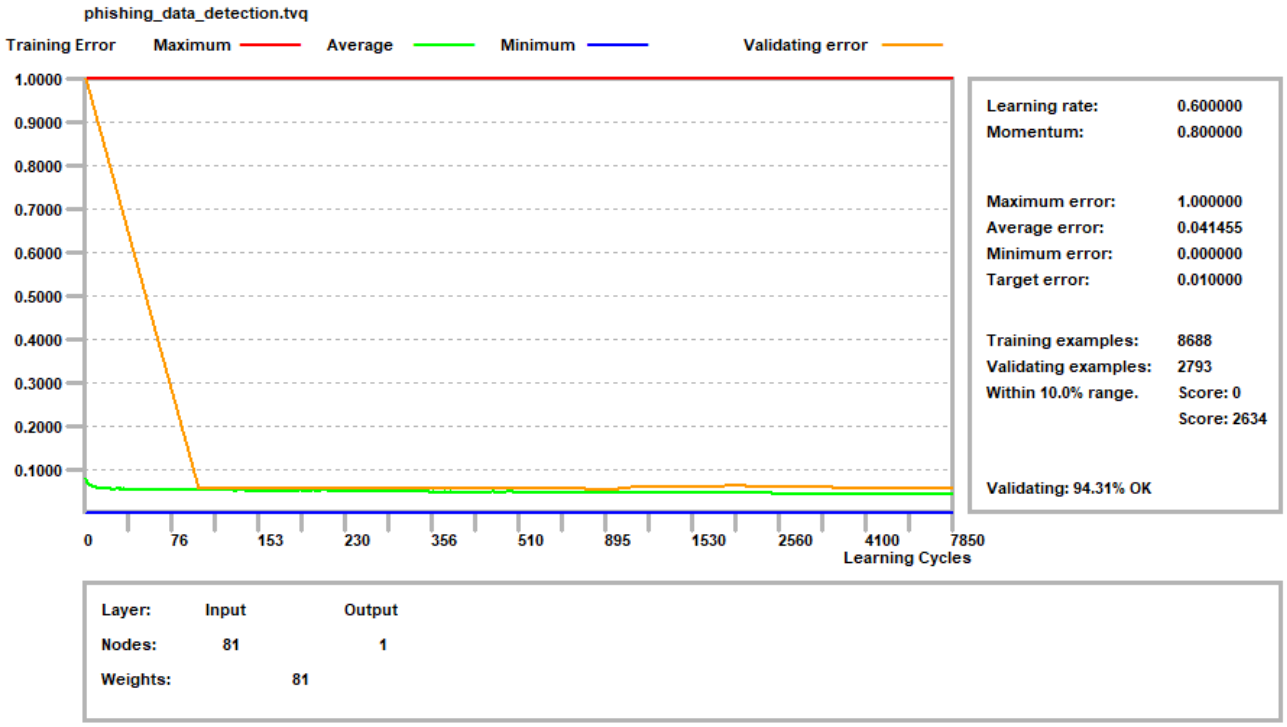

Figure 2: History of training and validation

7) **Final Model Evaluation:** Evaluate the final model using the testing dataset to provide an unbiased assessment of its performance.
8) **Feature Importance Analysis:**
   - **Feature Ranking:** A feature importance analysis is conducted to identify and rank the most influential features in web page phishing detection. After training the ANN model was able to identify and prioritize the most significant factors that can effectively differentiate between phishing and legitimate websites.
   - **Visualization:** Visual representations, such as feature importance plots or heat maps, are created to illustrate the significance of each feature (As shown in Figure 3).

phishing_data_detection.tvq 7942 cycles. Target error 0.0100 Average training error 0.041449

The first 81 of 81 Inputs in descending order.

Column	Input Name	Importance
56	nb_hyperlinks	338.7526
44	longest_words_raw	317.7023
1	length_hostname	307.3601
50	phish_hints	244.9051
39	length_words_raw	239.3560
4	nb_hyphens	196.1292
13	nb_slash	182.1841
23	http_in_path	154.0304
48	avg_word_host	140.4609
15	nb_colon	124.4607
5	nb_at	118.4532
46	longest_word_path	109.9961
17	nb_semicolumn	82.8408
42	shortest_word_host	75.6211
86	page_rank	74.2902
16	nb_comma	65.3465
10	nb_underscore	63.7918
9	nb_eq	63.7877
19	nb_space	61.1873
7	nb_and	50.7378
28	port	48.1166
41	shortest_words_raw	44.4531
81	domain_registration_le+	41.8140
26	ratio_digits_host	41.3735
21	nb_com	40.5207
85	google_index	40.3432
20	nb_www	38.6725
6	nb_qm	37.9678
82	domain_age	37.0930
53	brand_in_path	30.5965
0	length_url	28.3424
43	shortest_word_path	27.6437
60	nb_extCSS	26.6473
47	avg_words_raw	24.1463
3	nb_dots	23.4486
31	abnormal_subdomain	21.0865
62	ratio_extRedirection	18.5385
58	ratio_extHyperlinks	16.5408
52	brand_in_subdomain	16.0536
11	nb_tilde	15.8739
55	statistical_report	15.7066
73	popup_window	14.9978
75	onmouseover	14.5517
72	iframe	12.7594
80	whois_registered_domain	12.7087
35	shortening_service	12.5831
84	dns_record	12.5678
22	nb_dslash	12.3811
29	tld_in_path	10.8938
57	ratio_intHyperlinks	10.0271
25	ratio_digits_url	9.6836
74	safe_anchor	9.6168
65	login_form	9.1430
32	nb_subdomains	8.6469
24	https_token	8.0006
64	ratio_extErrors	7.8595
12	nb_percent	7.8045
37	nb_redirection	7.3860
78	domain_in_title	7.2267
54	suspecious_tld	6.9325
70	ratio_extMedia	6.5900
51	domain_in_brand	5.7235
2	ip	5.4229
33	prefix_suffix	5.4085
69	ratio_intMedia	4.6122
79	domain_with_copyright	3.8487
66	external_favicon	3.1015
49	avg_word_path	3.0466
67	links_in_tags	2.5735
76	right_clic	2.3055
34	random_domain	2.2407
77	empty_title	1.8192
36	path_extension	1.6363
83	web_traffic	1.5936
40	char_repeat	0.9554
45	longest_word_host	0.9199
18	nb_dollar	0.2270
38	nb_external_redirection	0.2269
14	nb_star	0.2262
27	punycode	0.2255
30	tld_in_subdomain	0.0017

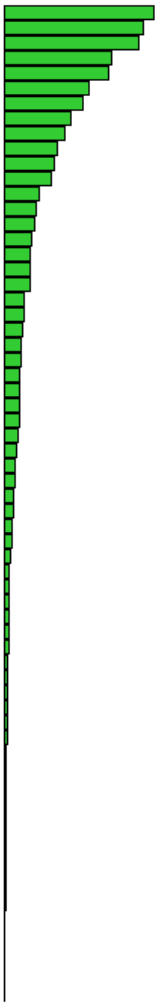

Figure 3: The most influential Features

9) **Practical Implications:**

- **Application Scenarios:** The practical implications of this study include improved online security, reduced risks for internet users, and enhanced cybersecurity measures through the adoption of an effective and relatively straightforward neural network model for detecting phishing websites.

10) **Results and Discussion:**

The purpose of this experiment was to detect the phishing web pages to provide online security. We used Backpropagation algorithm, which provides the ability to perform neural network learning and testing. Our neural network is the front feed network, with one input layer (82 inputs), and one output layer (1 output) [as shown in Figure 1]. The proposed model is implemented in Just Neural Network (JNN) environment. The dataset for Web page Phishing Detection was gathered from Kaggle which contains 11481 samples with 88 attributes [as shown in Figure 4].

This model was used to determine the value of each of the variables using JNN which they are the most influential factor on phishing detection [as shown in Figure 3]. After training and validating, the network, it was tested using the test data and the following results were obtained. The accuracy number of phishing web pages detection was (94.31%). The average error was 0.041455. The training cycles (number of epochs) were 7942. The training examples were 8688. The number of validating examples was 2793 [as shown in Figure 2]. The control parameter values of the model is shown in[Figure 5] and the detail summary of the proposed model is shown in[Figure 6].

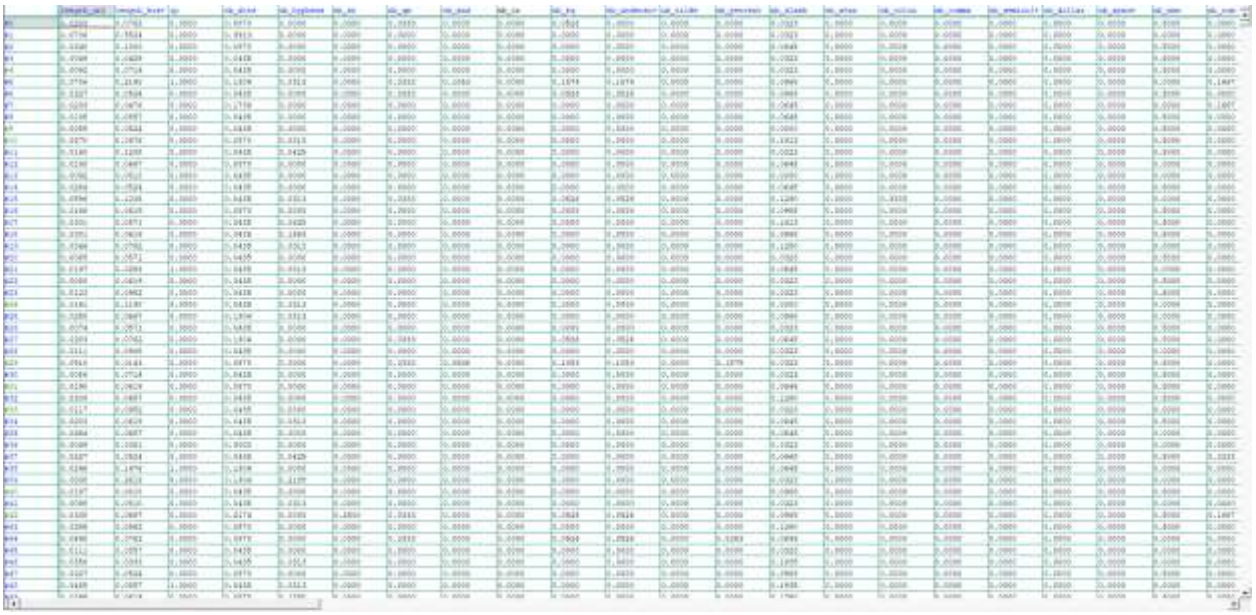

Figure 4: Dataset after cleaning

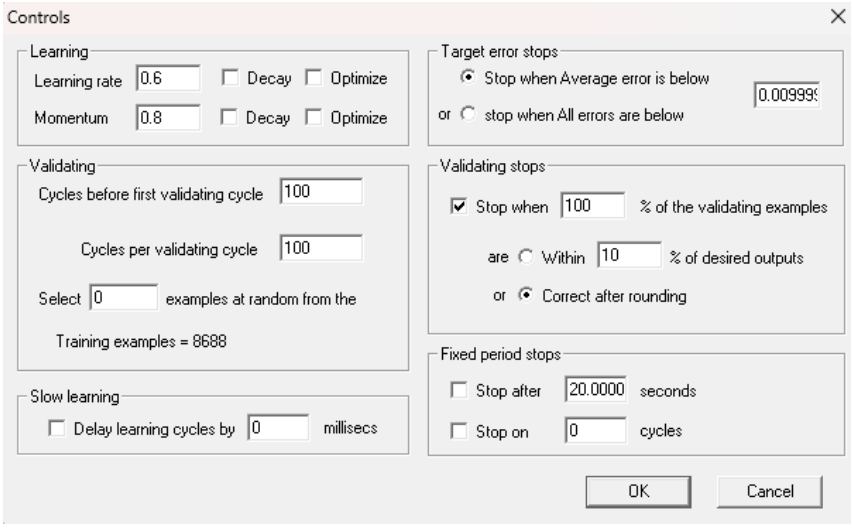

Figure 5: Controls of the proposed models

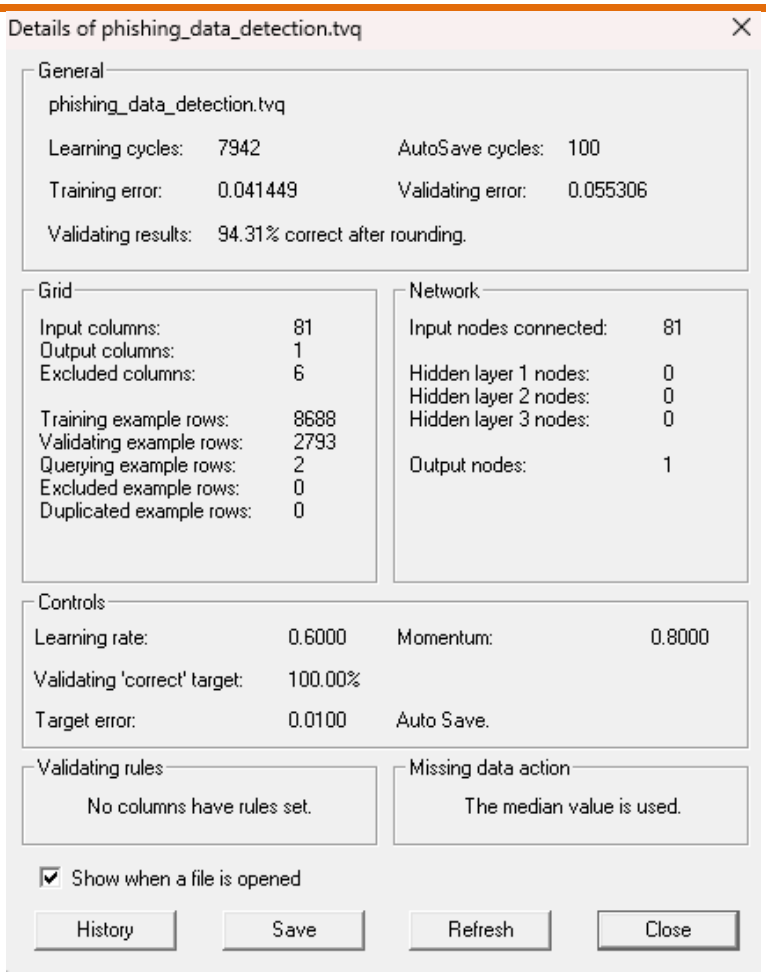

Figure 6: Details of the proposed model

6. Conclusion

Using a neural network model, this research paper's comprehensive method for spotting phishing web pages is presented. The study uses a dataset with 11,481 samples and 88 extracted characteristics, which includes properties relating to URL structure, content, and external services. Creating a neural network model that can discriminate between legal websites and phishing websites is the main goal of this project.

By illustrating how well neural networks perform at detecting web page phishing, the study makes a contribution to the field of information security and the fight against cybercrime. Given its great accuracy, the model has the potential to be used in real-world scenarios to protect internet consumers from phishing scams.

It's important to note, though, that while neural networks have potential in this field, they also have drawbacks, including computational complexity and susceptibility to hostile attacks. The accuracy and robustness of phishing detection systems might be substantially improved by future study on strategies to address these issues. The actual application of such models in real-time online security systems is another area that needs more research.

References

1. Zaid, A. A., et al. (2020). "The Impact of Total Quality Management and Perceived Service Quality on Patient Satisfaction and Behavior Intention in Palestinian Healthcare Organizations." Technology Reports of Kansai University 62(03): 221-232.
2. Sultan, Y. S. A., et al. (2018). "The Style of Leadership and Its Role in Determining the Pattern of Administrative Communication in Universities-Islamic University of Gaza as a Model." International Journal of Academic Management Science Research (IJAMSR) 2(6): 26-42.
3. Salman, F. M. and S. S. Abu-Naser (2019). "Expert System for Castor Diseases and Diagnosis." International Journal of Engineering and Information Systems (IJEAIS) 3(3): 1-10.
4. Saleh, A., et al. (2020). Brain tumor classification using deep learning. 2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech), IEEE.
5. Salama, A. A., et al. (2018). "The Role of Administrative Procedures and Regulations in Enhancing the Performance of The Educational Institutions-The Islamic University in Gaza is A Model." International Journal of Academic Multidisciplinary Research (IJAMR) 2(2): 14-27.
6. Nassr, M. S. and S. S. Abu Naser (2018). "Knowledge Based System for Diagnosing Pineapple Diseases." International Journal of Academic Pedagogical Research (IJAPR) 2(7): 12-19.
7. Nasser, I. M., et al. (2019). "Artificial Neural Network for Diagnose Autism Spectrum Disorder." International Journal of Academic Information Systems Research (IJAISR) 3(2): 27-32.
8. Nasser, I. M. and S. S. Abu-Naser (2019). "Predicting Tumor Category Using Artificial Neural Networks." International Journal of Academic Health and Medical Research (IJAHMR) 3(2): 1-7.
9. Musleh, M. M., et al. (2019). "Predicting Liver Patients using Artificial Neural Network." International Journal of Academic Information Systems Research (IJAISR) 3(10): 1-11.
10. Musleh, M. M. and S. S. Abu-Naser (2018). "Rule Based System for Diagnosing and Treating Potatoes Problems." International Journal of Academic Engineering Research (IJAER) 2(8): 1-9.
11. Mettleq, A. S. A., et al. (2020). "Mango Classification Using Deep Learning." International Journal of Academic Engineering Research (IJAER) 3(12): 22-29.
12. Mettleq, A. S. A. and S. S. Abu-Naser (2019). "A Rule Based System for the Diagnosis of Coffee Diseases." International Journal of Academic Information Systems Research (IJAISR) 3(3): 1-8.
13. Masri, N., et al. (2019). "Survey of Rule-Based Systems." International Journal of Academic Information Systems Research (IJAISR) 3(7): 1-23.
14. Madi, S. A., et al. (2018). "The Organizational Structure and its Impact on the Pattern of Leadership in Palestinian Universities." International Journal of Academic Management Science Research (IJAMSR) 2(6): 1-26.
15. Madi, S. A., et al. (2018). "The dominant pattern of leadership and Its Relation to the Extent of Participation of Administrative Staff in Decision-Making in Palestinian Universities." International Journal of Academic Management Science Research (IJAMSR) 2(7): 20-43.
16. Kashkash, K., et al. (2005). "Expert system methodologies and applications-a decade review from 1995 to 2004." Journal of Artificial Intelligence 1(2): 9-26.
17. Hilles, M. M. and S. S. Abu Naser (2017). "Knowledge-based Intelligent Tutoring System for Teaching Mongo Database." EUROPEAN ACADEMIC RESEARCH 6(10): 8783-8794.
18. Elzamly, A., et al. (2015). "Classification of Software Risks with Discriminant Analysis Techniques in Software planning Development Process." International Journal of Advanced Science and Technology 81: 35-48.
19. Elsharif, A. A. and S. S. Abu-Naser (2019). "An Expert System for Diagnosing Sugarcane Diseases." International Journal of Academic Engineering Research (IJAER) 3(3): 19-27.
20. Elqassas, R. and S. S. Abu-Naser (2018). "Expert System for the Diagnosis of Mango Diseases." International Journal of Academic Engineering Research (IJAER) 2(8): 10-18.
21. El-Mashharawi, H. Q., et al. (2020). "Grape Type Classification Using Deep Learning." International Journal of Academic Engineering Research (IJAER) 3(12): 41-45.
22. El Talla, S. A., et al. (2018). "The Nature of the Organizational Structure in the Palestinian Governmental Universities-Al-Aqsa University as A Model." International Journal of Academic Multidisciplinary Research (IJAMR) 2(5): 15-31.
23. El Talla, S. A., et al. (2018). "Organizational Structure and its Relation to the Prevailing Pattern of Communication in Palestinian Universities." International Journal of Engineering and Information Systems (IJEAIS) 2(5): 22-43.
24. Dheir, I. and S. S. Abu-Naser (2019). "Knowledge Based System for Diagnosing Guava Problems." International Journal of Academic Information Systems Research (IJAISR) 3(3): 9-15.
25. Dahouk, A. W. and S. S. Abu-Naser (2018). "A Proposed Knowledge Based System for Desktop PC Troubleshooting." International Journal of Academic Pedagogical Research (IJAPR) 2(6): 1-8.
26. Barhoom, A. M. and S. S. Abu-Naser (2018). "Black Pepper Expert System." International Journal of Academic Information Systems Research (IJAISR) 2(8): 9-16.
27. Ashqar, B. A. M. and S. S. Abu-Naser (2019). "Identifying Images of Invasive Hydrangea Using Pre-Trained Deep Convolutional Neural Networks." International Journal of Academic Engineering Research (IJAER) 3(3): 28-36.
28. Anderson, J., et al. (2005). "Adaptation of Problem Presentation and Feedback in an Intelligent Mathematics Tutor." Information Technology Journal 5(5): 167-207.
29. AlZamily, J. Y. and S. S. Abu-Naser (2018). "A Cognitive System for Diagnosing Musa Acuminata Disorders." International Journal of Academic Information Systems Research (IJAISR) 2(8): 1-8.
30. Al-Shawwa, M. and S. S. Abu-Naser (2019). "Knowledge Based System for Apple Problems Using CLIPS." International Journal of Academic Engineering Research (IJAER) 3(3): 1-11.
31. Alshawwa, I. A., et al. (2020). "Analyzing Types of Cherry Using Deep Learning." International Journal of Academic Engineering Research (IJAER) 4(1): 1-5.
32. Al-Nakhal, M. A. and S. S. Abu Naser (2017). "Adaptive Intelligent Tutoring System for learning Computer Theory." EUROPEAN ACADEMIC RESEARCH 6(10): 8770-8782.
33. Almurshidi, S. H. and S. S. Abu Naser (2017). "Design and Development of Diabetes Intelligent Tutoring System." EUROPEAN ACADEMIC RESEARCH 6(9): 8117-8128.
34. Almasri, A., et al. (2019). "Intelligent Tutoring Systems Survey for the Period 2000-2018." International Journal of Academic Engineering Research (IJAER) 3(5): 21-37.
35. Almasri, A., et al. (2018). "The Organizational Structure and its Role in Applying the Information Technology Used In the Palestinian Universities-Comparative Study between Al-Azhar and the Islamic Universities." International Journal of Academic and Applied Research (IJAAR) 2(6): 1-22.
36. Al-Habil, W. I., et al. (2017). "The Impact of the Quality of Banking Services on Improving the Marketing Performance of Banks in Gaza Governorates from the Point of View of Their Employees." International Journal of Engineering and Information Systems (IJEAIS) 1(7): 197-217.
37. Alhabbash, M. I., et al. (2016). "An Intelligent Tutoring System for Teaching Grammar English Tenses." EUROPEAN ACADEMIC RESEARCH 6(9): 7743-7757.
38. AlFerjany, A. A. M., et al. (2018). "The Relationship between Correcting Deviations in Measuring Performance and Achieving the Objectives of Control-The Islamic University as a Model." International Journal of Engineering and Information Systems (IJEAIS) 2(1): 74-89.
39. Al-Bastami, B. G. and S. S. Abu Naser (2017). "Design and Development of an Intelligent Tutoring System for C# Language." EUROPEAN ACADEMIC RESEARCH 6(10): 8795.
40. Alajrami, M. A. and S. S. Abu-Naser (2018). "Onion Rule Based System for Disorders Diagnosis and Treatment." International Journal of Academic Pedagogical Research (IJAPR) 2(8): 1-9.
41. Al Shobaki, M., et al. (2018). "Performance Reality of Administrative Staff in Palestinian Universities." International Journal of Academic Information Systems Research (IJAISR) 2(4): 1-17.
42. Al Shobaki, M. J., et al. (2018). "The Level of Organizational Climate Prevailing In Palestinian Universities from the Perspective of Administrative Staff." International Journal of Academic Management Science Research (IJAMSR) 2(5): 33-58.
43. Al Shobaki, M. J., et al. (2017). "Learning Organizations and Their Role in Achieving Organizational Excellence in the Palestinian Universities." International Journal of Digital Publication Technology 1(2): 40-85.
44. Al Shobaki, M. J., et al. (2017). "Impact of Electronic Human Resources Management on the Development of Electronic Educational Services in the Universities." International Journal of Engineering and Information Systems 1(1): 1-19.
45. Al Shobaki, M. J., et al. (2016). "The impact of top management support for strategic planning on crisis management: Case study on UNRWA-Gaza Strip." International Journal of Academic Research and Development 1(10): 20-25.
46. Al Shobaki, M. J. and S. S. Abu Naser (2016). "The reality of modern methods applied in process of performance assessments of employees in the municipalities in Gaza Strip." International Journal of Advanced Scientific Research 1(7): 14-23.
47. Al Shobaki, M. J. and S. S. Abu Naser (2016). "Performance development and its relationship to demographic variables among users of computerized management information systems in Gaza electricity Distribution Company." International Journal of Humanities and Social Science Research 2(10): 21-30.
48. Al Shobaki, M. J. and S. S. Abu Naser (2016). "Decision support systems and its role in developing the universities strategic management: Islamic university in Gaza as a case study." International Journal of Advanced Research and Development 1(10): 33-47.
49. Ahmed, A. A., et al. (2018). "The Impact of Information Technology Used on the Nature of Administrators Work at Al-Azhar University in Gaza." International Journal of Academic Information Systems Research (IJAISR) 2(6): 1-20.
50. Abu-Saqer, M. M., et al. (2020). "Type of Grapefruit Classification Using Deep Learning." International Journal of Academic Information Systems Research (IJAISR) 4(1): 1-5.
51. Abu-Saqer, M. M. and S. S. Abu-Naser (2019). "Developing an Expert System for Papaya Plant Disease Diagnosis." International Journal of Academic Engineering Research (IJAER) 3(4): 14-21.
52. Abu-Nasser, B. S. and S. S. Abu Naser (2018). "Rule-Based System for Watermelon Diseases and Treatment." International Journal of Academic Information Systems Research (IJAISR) 2(7): 1-7.
53. Abu-Naser, S. S., et al. (2011). "An intelligent tutoring system for learning java objects." International Journal of Artificial Intelligence & Applications (IJAIA) 2(2): 86-77.
54. Abu-Naser, S. S. and M. J. Al Shobaki (2016). "Computerized Management Information Systems Resources and their Relationship to the Development of Performance in the Electricity Distribution Company in Gaza." EUROPEAN ACADEMIC RESEARCH 6(8): 6969-7002.
55. Abu-Naser, S. S. and M. A. Al-Nakhal (2016). "A Ruled Based System for Ear Problem Diagnosis and Treatment." World Wide Journal of Multidisciplinary Research and Development 2(4): 25-31.
56. Abu-Naser, S. S. (2016). "ITSB: An Intelligent Tutoring System Authoring Tool." Journal of Scientific and Engineering Research 3(5): 63-71.
57. Abu-Naser, S. S. (2009). "Evaluating the effectiveness of the CPP-Tutor, an Intelligent Tutoring System for students learning to program in C++." Journal of Applied Sciences Research 5(1): 109-114.
58. Abu-Naser, S. S. (2008). "JEE-Tutor: An Intelligent Tutoring System for Java Expression Evaluation." Information Technology Journal 7(3): 528-532.
59. AbuEloun, N. N. and S. S. Abu Naser (2017). "Mathematics intelligent tutoring system." International Journal of Advanced Scientific Research 2(1): 11-16.
60. Abu Naser, S. S., et al. (2017). "Trends of Palestinian Higher Educational Institutions in Gaza Strip as Learning Organizations." International Journal of Digital Publication Technology 1(1): 1-42.
61. Abu Naser, S. S., et al. (2016). "Measuring knowledge management maturity at HEI to enhance performance-an empirical study at Al-Azhar University in Palestine." International Journal of Commerce and Management Research 2(5): 55-62.
62. Abu Naser, S. S. and M. J. Al Shobaki (2016). The Impact of Management Requirements and Operations of Computerized Management Information Systems to Improve Performance (Practical Study on the employees of the company of Gaza Electricity Distribution). First Scientific Conference for Community Development.
63. Abu Naser, S. S. (2008). "Developing an intelligent tutoring system for students learning to program in C++." Information Technology Journal 7(7): 1055-1060.
64. Abu Naser, S. S. (2006). "Intelligent tutoring system for teaching database to sophomore students in Gaza and its effect on their performance." Information Technology Journal 5(5): 916-922.
65. Abu Naser, S. S. (1999). "Big O Notation for Measuring Expert Systems complexity." Islamic University Journal Gaza 7(1): 57-70.
66. Abu Naser, S. S. (1993). A methodology for expert systems testing and debugging, North Dakota State University, USA.
67. Abu Nada, A. M., et al. (2020). "Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach." International Journal of Academic Information Systems Research (IJAISR) 4(8): 6-9.
68. Abu Nada, A. M., et al. (2020). "Age and Gender Prediction and Validation Through Single User Images Using CNN." International Journal of Academic Engineering Research (IJAER) 4(8): 21-24.
69. Abu Amuna, Y. M., et al. (2017). "Understanding Critical Variables for Customer Relationship Management in Higher Education Institution from Employees Perspective." International Journal of Information Technology and Electrical Engineering 6(1): 10-16.
70. Abu Amuna, Y. M., et al. (2017). "Strategic Environmental Scanning: an Approach for Crises Management." International Journal of Information Technology and Electrical Engineering 6(3): 28-34