imoncs.org.ng

NIGERIA COMPUTER SOCIETY

IMO STATE CHAPTER

imoncs.org.ng

NIGERIA COMPUTER SOCIETY

IMO STATE CHAPTER

presents

IMO TECHNOLOGY SUMMIT AND WORKSHOP 2023

THEME

Advancing Technology For Sustainable Transformation And Wealth Creation

Heroes Bend Hotel, along Jacob Zuma road, Concorde area, Owerri, Imo State.

25th-27th April 2023

IMO TECHNOLOGY SUMMIT AND WORKSHOP 2023

PROCEEDINGS

imoncs.org.ng

NIGERIA COMPUTER SOCIETY

IMO STATE CHAPTER

Towards advance technology

TABLE OF CONTENTS

A SURVEY ON EFFECTIVENESS OF THE PREDICTING METHODS OF PATIENT RESPONSE TO STEROID IN PRE-TRABECULECTOMY EXAMINATION.

Ajah Ifeyinwa Angela[1], Ukabuiro Ikenna Kelechi[2], Douglas Allswell Kelechi[3]

[1]Ebonyi State University, Abakaliki,
[2]Abia State University, Uturu
[3]Federal University of Technology, Owerri.
Ifyajah@gmail.com[1], ikenna.ukabuiro@abiastateuniversity.edu.ng[2] kelechi.douglas@futo.edu.ng[3],
,

Abstract:
The purpose of this study is to evaluate the accuracy of the various patient response (elevation of intraocular pressure) prediction methods used in the pre-trabeculectomy assessment. The use of a questionnaire for gathering primary data allowed the study to adopt a descriptive and quantitative research technique. Staff from the five federal medical and teaching hospitals in Southeast Nigeria make up the study's target population. However, the sample frame was narrowed to optometrists and ophthalmologists in the chosen hospitals who are specialists in eye diagnosis and care of eye illnesses. With a 95% confidence level as applied, the sample size was calculated using the Taro Yamane formula. Using a descriptive statistical analysis, the skill levels of the respondents' responses were determined. The efficiency of the currently being employed forecasting techniques was evaluated using regression analysis. The study's findings showed that physical examinations, which were previously used to predict patients' responses to steroids during pre-trabeculectomy exams, are ineffective.

Keywords: Trabeculectomy, Steroid, Ophthalmologists, Glaucoma, Artificial Intelligent

1.0 Introduction

A glaucoma procedure known as a trabeculectomy involves operating on the eye to open a new passageway for the drainage of fluid from the eye. This outpatient surgical technique is typically used to reduce eye pressure and stop the progression of glaucoma-related visual loss.

One of the most often prescribed medications; steroids are primarily used to treat a variety of autoimmune and inflammatory disorders [1]. Despite its many advantages, using steroids can have a few negative effects on the eyes, the most significant of which are steroid-induced glaucoma and cataract.

Predictive research, which seeks to forecast future outcomes or events based on patterns within a set of variables, is becoming more and more common in the field of medical research [2]. The future course of a disease or the likelihood of getting one can be predicted accurately using predictive models, which can help patients and doctors make decisions about screening and/or therapy. AI applications have been shown to increase diagnostic accuracy and reliability [3]. Both in the clinical and preclinical contexts, predictors of medication response and resistance have been developed using conventional statistical models and more advanced machine learning techniques. Cancer is a primary cause of death worldwide, according to George et al.'s study in cancer drug response prediction from 2021. The likelihood of a patient's complete recovery can be increased by personalizing drug response prediction using computational models to determine the optimum course of treatment. But regrettably, the computational task of drug response prediction is quite difficult, partly because of the constraints of the available data and partly because of algorithmic flaws.

In clinical ophthalmology, several image-related diagnostic approaches have started to provide previously unheard-of insights into eye illnesses based on morphological datasets with millions of data points, according to [4]. To avoid side effects of the medication that could lead to steroid-induced glaucoma, accurate prediction of patients who react to steroids is essential for a successful trabeculectomy.

There are numerous risk factors for developing steroid-induced glaucoma, according to [5]. Up to 8% of the general population can experience it, although glaucoma patients and their blood relatives experience it far more frequently. In fact, a steroid response occurs in 90% of patients with open-angle glaucoma. The ocular pressure typically returns to pre-steroid levels if responses are correctly predicted, and the steroids are stopped in a timely manner. Regrettably, patients who are repeatedly exposed to steroids run the danger of developing irreversible steroid glaucoma. A lifetime average of one week of steroid use results in a 4% greater risk of developing chronic steroid glaucoma. It's possible that high-risk categories of people should avoid using steroids until essential. Unfortunately, many conditions have non-steroidal alternatives.

1.1 Aim & Objectives

This study seeks to evaluate the accuracy of current patient response to steroid prediction methods in pre-trabeculectomy assessment.

The study's particular goals are as follows:
To find out what experts think about the efficiency of the current technique for anticipating patients' steroid responses and to assess the experts' level of expertise and determine it.

1.2 Research Question

i. How accurate is the current approach to identifying patients who will respond to steroid use during a pre-trabeculectomy operation examination?

2.0 Literature Review

The extent of the examination regimen used will determine the prevalence or incidence of glaucoma in a population study. Decisions regarding the variety of objective tests necessary during a clinical examination may be aided by the results of a thorough history [6].

[7] assert that the measurement of IOP is crucial for treating glaucoma. All clinical IOP measurement techniques are based on calculating the amount of force required to distort the globe. The number of grams required to flatten 3.06 mm2 of central corneal tissue in Goldman tonometry, the accepted method to measure IOP, is multiplied by 10 to produce the IOP in mm Hg. The IOP is therefore 15 mm Hg if 1.5 grams of force are required to flatten the center corneal tissue. Instead of focusing on whether a result is inside or outside of the normal range when examining IOP levels, the IOP should be read in the context of the CCT and health of the optic nerve. While there is uncertainty regarding the cost-effectiveness of the treatments that result, medical artificial intelligence (AI) may improve practitioners' diagnostic accuracy, leading to better treatment decisions at lower costs [3]. There are several research on the accuracy of AI applications used in medicine, but there are very few health economic evaluations of medical AI, and the majority of these have methodological issues [8].

2.1 Pre-consultation AI Based System vs Physical checks Examination.

In dentistry, rapidly developing technology like AI can undoubtedly take the role of physical skill. These technologies must also be used cautiously and under human supervision to decrease errors and oversight. Better patient outcomes result from the earliest and most precise detection of oral illnesses [9].

AI-based systems prior to consultation versus physical examination might be challenging for decision-makers to utilize physical examinations as a tool to forecast a patient's reaction to steroids during a pre-trabeculectomy evaluation. In their study, [10] expressed the opinion that artificial intelligence (AI) has found extensive use in the medical industry and has a wide range of potential applications. However, [11] in research opined that the current systematic review and meta-analysis shown that, when compared to skilled clinicians, artificial intelligence in the form of deep learning is generally an accurate technique for detecting PARL in dental radiographs. The pre-consultation system is an essential addition to the conventional face-to-face consultation, according to their study. The pre-consultation system and AI working together can increase the effectiveness of therapeutic work.

The complex electronic health record (EHR) data analysis and processing are still difficult for the AI to handle. A total of 2,648 pediatric patients participated in their trial from November 2019 to May 2020. Before visiting the doctors in the outpatient department of the Shanghai Children's Medical Center, the patients utilized their model to provide a medical history and obtain the primary diagnosis. The aim was to assess how well doctors and AI could get a primary diagnosis and to examine how the diagnostic performance was affected by the consistency between the medical history provided by the doctors and their model.

The findings demonstrated that the model performed worse than the physicians and had a lower average F1 score (0.825 vs. 0.912) if they did not consider whether the medical history recorded by the AI and doctors was consistent or not.

The model had a higher average F1 score and was closer to the doctors when the major complaint or the history of the present illness provided by the AI and doctors was consistent. Last but not least, when the AI and doctors shared identical diagnostic conditions, the model outperformed them both in terms of average F1 score (0.931 vs. 0.92 for doctors).

This study showed that the automated model could gather a more organized medical history and had a sound diagnostic logic, which would help to increase the outpatient doctors' diagnostic precision and decrease missing and incorrect diagnoses.

IOP variation while a 24-hour day or between visits is one of the factors linked to glaucoma progression [12]; [13]. Glaucoma may be more likely to develop or advance if IOP is exposed because of steroid use.

Topical drugs have been shown to cause elevated intraocular pressure (IOP). In a study by [14], the age and axial length of people who responded to steroids after having cataract surgery were analyzed. This study used a retrospective chart review to examine 1642 eyes that underwent straightforward cataract surgery over the course of a single calendar year at a single ophthalmology clinic. Following surgery, topical 1% prednisolone acetate was given to all patients. Axial length, patient age, and the first postoperative months’ worth of IOP were also noted, in addition to axial length and patient age.

A steroid responder was defined as someone whose IOP increased by at least 25% while taking topical prednisolone (minimum 28 mmHg), and then decreased by at least 25% after stopping the medication. The age and axial length of steroid responder eyes were then compared to non-responder eyes. The findings revealed that 38 eyes had been identified as steroid responders. A higher probability of steroid response was also linked to younger age and longer axial length, especially when IOP rose above 40 mmHg. According to the study's findings, young myopes who undergo straightforward cataract surgery should be more closely watched for a postoperative steroid response.

Artificial intelligence (AI) has been successfully applied in a wide range of digitalization-related fields. [15] conducted research on AI and the ethical implications of digital technology. [16] presented a study on sustainable development with AI for a variety of industries, including building, transportation, healthcare, and more. Regression, reinforcement learning, and AI-based decision support system technologies were shown to be particularly popular for sustainability.

[17] offer a paradigm for the deployment of AI in enterprises. They have determined the criteria of People, Processes, Technology, and Data Incorporation. According to the proposed model, effective adoption also necessitates the preparation of people, processes, and data in addition to technology readiness.

Issues like the practical use of AI and the lack of expertise in its efficient usage have been explored by [18]. The authors provide a conceptual framework that considers the four key factors of decision support, employee and customer interaction, automation, and innovative product and service development. Despite the fact that these issues exist, research has also suggested various ways and answers. With the right strategy, AI can be implemented smoothly and contribute to process automation in any firm.

Big data modeling is always a challenging challenge when deploying artificial and expert systems for problem solving, and there are numerous barriers preventing the actualization of AI automation in a variety of research fields, including the healthcare industry. There are numerous study topics on applying AI to different fields that have covered a variety of difficulties and approaches for effective AI adoption. Some of these studies, according to [19]; [20] and [21], apply AI to human resource management, incorporate ontologies, and apply big data to emerging management disciplines.

3.0 Methodology

The study used a questionnaire to collect primary data during a field survey, which is a descriptive and quantitative research methodology. The Taro Yamane method was used to establish the sample size for the study population, which consisted of 146 optometrists and ophthalmologists in five federal government hospitals in five South Eastern states of Nigeria. A 95% applicable confidence level was used for this. As follows is the formula:

$$n = \frac{N}{1 + N(e)^2}$$

Were

n = Sample size
N = Population (146)
E = tolerable error (100% - 95% = 5%)

Therefore,

$$n = \frac{146}{1 + 146(0.05)^2}$$

$$n = \frac{138}{1 + 146(0.0025)}$$

$$n = \frac{146}{1.365}$$

$n = 107$

The study used the content validity of research instrument to confirm the instrument's validity by ensuring that it was relevant, sufficient, inclusive, and pertinent to the investigation. The goal of research instrument validity is to make sure that the variables that are supposed to be assessed are really measured.

The test-retest reliability test was used in this study to ascertain the reliability of the research instrument. So, a pilot survey of four (4) optometrists and ophthalmologists was done. These copies were given to the respondent by the researcher, who later picked them up. The identical questions were asked to the same responders and then retrieved after a week. On the questions posed by the researcher, the outcome showed an 81% correlation. The data gathered for this study was analyzed using a variety of statistical approaches. Regression analysis was used to evaluate the hypotheses at a 0.05 level of significance. Version 20 of the Statistical Package for Social Sciences (SPSS) was also used.

4.0 Results and Discussion

Table 1: Age of Respondents

Age

		Frequency	%	Valid %	Cumulative %
Valid	20-30	9	11	11	11
	31-40	32	40	40	51
	41-50	18	23	23	74
	51-60	11	14	14	88
	61 and above	10	12	12	100.0
	Total	80	100.0	100.0	

Source: Survey Data, 2022

The age range of 31 to 40 years recorded the highest number of respondents with 40%, while the age range of 61 and above recorded the lowest number of respondents with 12%, according to the data from table 1 above. These findings indicated that most of the specialists were primarily between the ages of 31 and 40, which is indicative of the recent surge in interest among Nigeria's youth in the medical field.

Table 2: Number of Years in the Organization

No. of years in the Organization

		Frequency	%	Valid %	Cumulative %
Valid	Below 5 years	18	22	22	22
	6-10 years	34	43	43	65
	11 and above	28	35	35	100.0
	Total	80	100.0	100.0	

Source: Survey Data, 2022

According to the respondents' years of eye care and trabeculectomy surgery experience, the replies in Table 2 above were compiled. According to the poll results, 43% of respondents have worked for the company for less than six years, while 22% have been there for less than five years. This suggests that the study subjects had the knowledge and skills necessary to appropriately answer the inquiries the researcher aimed to make in light of their prior knowledge.

.

Research Questions

i. How well does the current pre-trabeculectomy surgery examination method identify patients who react or respond to the usage of steroids?

Test of Hypothesis

i. Ho1: Currently, the pre-trabeculectomy evaluation approach for predicting patients' responses to steroids is ineffective.

ii. Ho2: The present technique for anticipating a patient's steroid responses during the pre-trabeculectomy assessment is successful.

Table 3: Descriptive Statistics

	Mean	Std. Deviation	N
Existing predicting methods	4.2030	.96886	80
Effectiveness of predicting patients' response to steroid	4.3218	.93597	80

Table 4: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.213[a]	.113	.588	.32438	. 872

a. Predictors: (Constant), Current predicting methods

b. Dependent Variable: Effectiveness of predicting patients' response to steroid

Table 5: ANOVA

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	167.634	1	167.634	1593.126	.062[b]
	Residual	21.045	78	.105		
	Total	188.678	79			

a.Dependent Variable: Effectiveness of predicting patients' response to steroid

b. Predictors: (Constant), Current predicting methods

Table 6: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.014	.108		-.128	.898
	Current predicting methods	.976	.024	.943	1.428	.062

a. Dependent Variable: Effectiveness of predicting patients' response to steroid

Table 7: Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.9619	4.8647	4.2030	.91323	80
Residual	-.93759	.13529	.00000	.32357	80
Std. Predicted Value	-3.549	.725	.000	1.000	80
Std. Residual	-2.890	.417	.000	.998	80

a. Dependent Variable: Effectiveness of predicting patients' response to steroid

5.0 Summary of Findings

The key findings of the study include:

Most of the specialists were found to be between the ages of 31 and 40, which implies that younger professionals than older ones have been working in the field of eye care recently in the study area. According to survey results, 43% of respondents have a service history of less than 6 to 10 years. This suggests that the study subjects had the knowledge and skills necessary to appropriately answer the inquiries the researcher aimed to make in light of their prior knowledge.

With an R value of 0.213, the patient's response to steroids can only be predicted with very little accuracy. The R2 number indicates the proportion of the dependent variable's overall variation (method effectiveness) that can be accounted for by the independent variable's present forecasting technique (Physical checks). In the selected Federal Medical Hospitals in Southeast Nigeria, Table 4 demonstrates the present predicting method's 11.3% effectiveness in predicting the patient's responses to steroids during the pre-trabeculectomy test.

The study concludes that the current method of predicting patients' responses to steroids in pre-trabeculectomy examination in the study area is ineffective because the p-value (0.062) is greater than Alpha (0.05), that is, 0.062>0.05, and t calculated (1.428) is less than t tabulated (1.960).

6.0 Conclusion

To reduce the rise in steroid-induced glaucoma patients, the study found that the current method (physical checks) of predicting patients' responses to steroids was ineffective. It therefore recommended implementing a modeled artificial intelligent system to assist experts in predicting and identifying patients who may likely react to the use of steroids in pre-trabeculectomy examinations.

References

[1]. M. Razeghinejad & L. Katz. (2012). Steroid-induced iatrogenic glaucoma. . *Ophthalmic Res., 47*(2), 66-80.

[2]. A. Idowu, A. Aladekomo, O. Williams & J. Balogun. (2015). Predictive model for likelihood of Sickle cell aneamia (SCA) among pediatric patients using fuzzy logic. . *Transactions in networks and communications, 31*(1), 31-44.

[3]. F. Schwendicke, G. O. Cejudo, J. Krois, G. Cantu, F. Meyer-Luckel & F. Chaurasia. (2022). Artificial Intelligence for Caries Detection: Value of Data and Information. *Jornal of Dental Research SAGE Journals , 101* (11), 122**.**

[4]. Y. Dong, Q. Zhang, Z. Qiao & J. Yang (2017). Classification of cataract fundus image based on deep learning. *IEEE International Conference on Imaging Systems and Techniques (IST).* (pp. 1–5). Beijing: IEEE.

[5]. P. Terry. (2020, October 02). *//www.glaucoma.org/.* Retrieved March 03, 2022, from glaucoma.org:https://www.glaucoma.org/glaucoma/steroids-and-glaucoma-whats-the-connection.php(2020/10)

[6]. M. Tatemichi, T. Nakano & K. Tanaka. (2004). Possible association between heavy computer users and glaucomatous visual field abnormalities: a cross sectional study in Japanese workers. *J Epidemiol Community Health.* , 1021-1027.

[7]. C. Laura, & P. Louis. (2023). *Clinical Characteristics and Current Treatment.* Massachusetts Eye and Ear, Brigham and Women's Hospital, Harvard, Departments of Ophthalmology . Massachusetts: Cold Spring Harbor Laboratory Press.

[8]. J. Wolff, J. Pauling, A. Keck & J. Baumbach. (2020). The economic impact of artificial intelligence in health care: systematic review. *J Med Internet Res. , 22* (2), 22

[9]. S. S. Mahdi, G. Battineni, M. Khawaja, R. Allana, M. Siddigui, & D. Agha, (2023). How does artificial intelligence impact digital healthcare initiatives? A review of AI applications in dental healthcare. *International Journal of Information Management Data Insights*, *3* (1), 100144.

[10]. H. Qian, B. Dong, J. Yuan, F. Yin, Z. Wang, H. Wang, . . . B. Ning. (2021).) Pre-Consultation System Based on the Artificial Intelligence Has a Better Diagnostic Performance Than the Physicians in the Outpatient Department of Pediatrics. Front. Med. . *Front. Med*, 8. doi:doi: 10.3389/fmed.2021.695185

[11]. S. Sadr, M. R. Hossein, S. Zahedrozegar, P. Motie, S. Vinayahalingam, O. Dianat, et al. (2023). Deep Learning for Detection of Periapical Radiolucent Lesions: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy. *Journal of Endodontics , 49* (1), 248-261.

[12]. R. Varma, L. Hwang, J. Grunden, & G. Bean. (2008). Inter-visit IOP range: an alternative parameter for assessing intraocular pressure control in clinical trials. . *Am J Ophthalmol., 145*, 336-342.

[13]. S. Asrani, R. Zeimer, J. Wilensky, D. Gieser, S. Vitale, & K. Lindenmuth. (2010). Rsik Factors Among Cataract Patients for Steroid Reseponse. *Investigative Ophthalmology & Visual Science* (Vol.51, 4563. doi:), 134-142.

[14]. J. Tan & F. Chang. (2010). Risk Factors Among Cataract Patients for Steroid Response. *Investigative Ophthalmology & Visual Science, Vol.51, 4563. doi:*, 1.

[15]. M. Ashok, N. Ashok, R. Madam, A. Joha & U. Sivarajah. (2022). Ethical framework for artificial intelligence and digital technologies. *International journal of Information Management , 62*, 102-433.

[16]. A. Kar, S. Choudhary & V. Singh. (2022). How can artificial intelligence impact sustainability: A systematic literature review. *Journal of Cleaner production* , 134120.

[17]. V. Uren & J. Edwards. (2023). Technology readiness and the organizational journey towards AI adoption: An empirical study. *International Journal of Information Management* (68), 102588.

[18]. A. Borges, F. Laurindo, M. Spinola, R. Goncalves & C. Mattos. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management* (57), 102225.

[19]. A. Votto, R. Valecha, P. Najafirad, & H. Rao. (2021). Artificial intelligence in tactical human resource management: A systematic literature review. *International Journal of Information Management Data Insights , 1* (2), 23.

[20]. A. Antunes, E. Cardoso & J. Barateiro. (2022). Incorporation of ontologies in data warehouse/business intelligence systems—a systematic literature review. *International Journal of Information Management Data Insights , 2* (2), 1.

[21]. A. Kushwaha, A. Kar, & Y. Dwivedi. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights , 1* (2), 5.

URL Based Phishing Website Detection Using Machine Learning

Donatus O. Njoku[1], Callistus T. Ikwuazom[2], Stanley A. Okolie[3], Janefrances E. Jibiri[4], Emmanuel C. Ololo[5], Kelechi Onyemachi[6]

[1]Dept. of Computer Science, Federal University of Technology Owerri, Nigeria
[2]Dept. of Information Technology, Federal University of Technology Minna, Nigeria
[3]Dept. of Computer Science, Federal University of Technology Owerri, Nigeria
[4]Dept. of Information Technology, Federal University of Technology Owerri, Nigeria
[5]Dept. of Computer Science, Imo State Polytechnic, Omuma, Imo state, Nigeria
[6]Dept. of Computer Science, Federal Polytechnic Nekede, Nigeria

Abstract—**Phishing attacks are one of the most common social engineering attacks targeting users' emails to fraudulently steal confidential and sensitive information. They can be used as a part of more massive attacks launched to gain a foothold in corporate or government networks. Over the last decade, a number of antiphishing techniques have been proposed to detect and mitigate these attacks. However, they are still inefficient and inaccurate. Thus, there is a great need for efficient and accurate detection techniques to cope with these attacks. In this paper, we proposed a phishing attack detection technique based on machine learning. We modeled these attacks by selecting 10 relevant features and building a large dataset. This dataset was used to train, validate, and test the machine learning algorithms. For performance evaluation, four metrics have been used, namely probability of detection, probability of miss-detection, probability of false alarm, and accuracy. The experimental results show that better detection can be achieved using an artificial neural network.**

Keywords—URL based, phishing, machine learning, algorithm, detection

1.0 INTRODUCTION

A. Background of the Study

Due to the rapid developments of the global networking and communication technologies, lots of our daily life activities such as social networks, electronic banking, e-commerce, etc. are transferred to the cyberspace. The open, anonymous and uncontrolled infrastructure of the Internet enables an excellent platform for cyberattacks, which presents serious security vulnerabilities not only for networks but also for the standard computer users even for the experienced ones. Although carefulness and experience of the user are important, it is not possible to completely prevent users from falling to the phishing scam [4]. Because, to increase the success of the phishing attacks, attackers also get into consideration about the personality characteristics of the end user especially for deceiving the relatively experienced users [10].

End-user-targeted cyberattacks cause massive loss of sensitive/personal information and even money for individuals whose total amount can reach billions of dollars in a year [15].

Phishing attacks' analogy is derived from "fishing" for victims, this type of attacks has attracted a great deal of attention from researchers in recent years. It is also a promising and attractive technique for attackers (also named as phishers) who open some fraudulent websites, which have exactly similar design of the popular and legal sites on the Internet. Although these pages have similar graphical user interfaces, they must have different Uniform Resource Locators (URLs) from the original page. Mainly, a careful and experienced user can easily detect these malicious web pages by looking at the URLs.

Phishing is defined as a a cybercrime in which a target or targets are contacted by email, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords. The information is then used to access important accounts and can result in identity theft and financial loss.

Machine learning based phishes detection gadget relies upon efficiently on the aspects of accuracy. The most of antiphishers researchers center of attention on optimizing new feature proposals or classification algorithms, where developing proper features analysis and selection techniques is not the important plan. The 12 features of this site are legitimate, phishing-enabled, reaching an effective positive rate of 97% and a false positive rate of 4%. The features are obtained by META tagging, web pages content, URLs, hyperlinks, TF-IDF, and more. Therefore, extraneous aspects might also nonetheless exist, which will increase the price of the technology (i.e. Training time, storage, electricity, etc.), however, it does not affect the average accuracy. Therefore, identifying a truly effective compact feature set requires an efficient Machine Learning based technique for Phishing detection.

The first phishing lawsuit was filed in 2004 against a Californian teenager who created the imitation of the website "America Online". With this fake website, he was able to gain sensitive information from users and access the credit card details to withdraw money from their accounts.

Other than email and website phishing, there's also 'vishing' (voice phishing), 'smishing' (SMS Phishing) and several other phishing techniques cybercriminals are constantly coming up with. The study wants to focus on the various ways phishing can be done and possible solutions to them in form of a machine learning based software.

B. Objectives of the Study

The primary aim of the work is to design a URL based phishing website detector. The specific objectives are:

- To develop a novel approach to detect malicious URL and alert users.

- To apply Machine Learning techniques in the proposed approach in order to analyze the real time URLs and produce effective results.
- Creating a reporting platform for other users of the platform to report fake websites in order to build the knowledge base.
- Studying previous work on the proposed topic and looking for ways to improve them.

2.0 LITERATURE REVIEW

C. Theoretical Framework

A theoretical framework for a URL-based phishing website detector would likely draw on concepts from the fields of computer science, information security, and human-computer interaction. The first component of the theoretical framework would be a technical understanding of how phishing attacks work and the methods that attackers use to spoof legitimate websites. This would include knowledge of techniques such as domain spoofing, URL redirects, and the use of malicious scripts or payloads. The second component would be an understanding of the psychological and social factors that make individuals susceptible to phishing attacks, such as trust in familiar brands or a willingness to provide personal information [15] This would inform the design of user interfaces and interactions that aim to educate and empower users to protect themselves against phishing. The third component would be the use of machine learning and data mining techniques to analyze and detect patterns in website URLs and other features that are indicative of phishing websites. This could include using models such as Random Forest, SVM and Neural Network.

The fourth component would be the use of browser extension or security software that can interact with the user's web browser to warn them of potentially malicious websites or block them entirely, and also providing feedback to machine learning model. The final component would be the evaluation of the effectiveness of the detector, through the use of datasets that contains both phishing and legitimate website URLs, and comparing the performance of the detector with existing state-of-the-art methods, [2]. By combining these various components, the theoretical framework would provide a comprehensive approach to detecting and defending against phishing attacks by using machine learning and user interface strategies, while also taking into account the social and psychological factors that make individuals susceptible to phishing scams.

D. Conceptual Framework

Phishing is the fraudulent attempt to obtain sensitive information or data, such as usernames, passwords, credit card numbers, or other sensitive details by impersonating oneself as a trustworthy entity in a digital communication [7]. Typically carried out by email spoofing, instant messaging, and text messaging, phishing often directs users to enter personal information at a fake website which matches the look and feel of the legitimate site. As of 2020, phishing is by far the most common attack performed by cyber-criminals, with the FBI's Internet Crime Complaint Centre recording over twice as many incidents of phishing than any other type of computer crime.

The first recorded use of the term "phishing" was in the cracking toolkit AOHell created by Koceilah Rekouche in 1995 [7] however it is possible that the term was used before this in a print edition of the hacker magazine *2600*, [1]. The word is a leetspeak variant of *fishing* (*ph* is a common replacement for *f*), probably influenced by phreaking, and alludes to the use of increasingly sophisticated lures to "fish" for users' sensitive information.

Attempts to prevent or mitigate the impact of phishing incidents include legislation, user training, public awareness, and technical security measures [11]. The types of phishing include:

- E-mail phishing: Email phishing is the general term given to any malicious email message meant to trick users into divulging private information. Attackers generally aim to steal account credentials, personally identifiable information (PII) and corporate trade secrets. However, attackers targeting a specific business might have other motives.
- Spear phishing: These email messages are sent to specific people within an organization, usually high-privilege account holders, to trick them into divulging sensitive data, sending the attacker money or downloading malware.
- Whaling and CEO fraud: These messages are typically sent to high-profile employees of a company to trick them into believing the CEO or other executive has requested to transfer money. CEO fraud falls under the umbrella of phishing, but instead of an attacker spoofing a popular website, they spoof the CEO of the targeted corporation.
- Voice phishing: Also known as Vishing, here attackers use voice-changing software to leave a message telling targeted victims that they must call a number where they can be scammed. Voice changers are also used when speaking with targeted victims to disguise an attacker’s accent or gender so that they can pretend to be a fraudulent person
- SMS phishing: This type of phishing attack is also known as smishing. Using SMS messages, attackers trick users into accessing malicious sites from their smartphones. Attackers send a text message to a targeted victim with a malicious link that promises discounts, rewards or free prizes.
- Watering hole: A compromised site provides endless opportunities, so an attacker identifies a site used by numerous targeted users, exploits a vulnerability on the site, and uses it to trick users into downloading malware. With malware installed on targeted user machines, an attacker can redirect users to spoofed websites or deliver a payload to the local network to steal data [8]

E. Anti-Phishing Systems

Anti-phishing software consists of computer programs that attempt to identify phishing content contained in websites, e-mail, or other forms used to accessing data (usually from the internet) and block the content, usually with a warning to the user (and often a choice to view the content regardless). It is often integrated with web browsers and email clients as a toolbar that displays the important name for the web site the viewer is visiting, in an effort to prevent fraudulent websites from masquerading as other legitimate websites [16].

Most popular web browsers come with built-in anti-phishing and anti-malware protection services, but almost none of the alternate web browsers have such protections. [13] Password managers also can be wont to help defend against phishing, as can some mutual authentication techniques.

An independent study conducted by Carnegie Mellon University CyLab titled "Phinding Phish: An Evaluation of Anti-Phishing Toolbars" and released November 13, 2018 tested the power of ten anti-phishing solutions to block or warn about known phishing sites and not block or warn about legitimate sites (not exhibit false-positives), also because the usability of every solution. Of the solutions tested, Netcraft Toolbar, EarthLink ScamBlocker and SpoofGuard were able to correctly identify over 75% of the sites tested, with Netcraft Toolbar receiving the highest score without incorrectly identifying legitimate sites as phishing. Severe problems were however discovered using SpoofGuard, and it incorrectly identified 38% of the tested legitimate sites as phishing, resulting in the conclusion that "such inaccuracies might nullify the benefits SpoofGuard offers in identifying phishing sites." [12]. Google Safe Browsing (which has since been built into Firefox) and Internet Explorer both performed well, but when testing ability to detect fresh phishes Netcraft Toolbar scored as high as 96%, while Google Safe Browsing scored as low as 0%, possibly thanks to technical problems with Google Safe Browsing. The testing was performed using phishing data obtained from Anti-Phishing working party, PhishTank, and an unnamed email filtering vendor.

Another study, conducted by SmartWare for Mozilla and released November 14, 2018, concluded that the anti-phishing filter in Firefox was more effective than Internet Explorer by over 10%. The results of this study are questioned by critics, noting that the testing data was sourced exclusively from PhishTank, which itself is an anti-phishing provider. The study only compared Internet Explorer and Firefox, leaving out (among others) Netcraft Toolbar and therefore the Opera browser, both of which use data from PhishTank in their anti-phishing solutions. This has led to speculations that, with the limited testing data, both Opera and Netcraft Toolbar would have gotten an ideal score had they been a part of the study.

While the two directly aforementioned reports were released just one day apart, Asa Dotzler, Director of Community Development at Mozilla, has skilled the criticism of the Mozilla commissioned report by saying, "so you're agreeing that the most recent legitimate data puts Firefox ahead. Good enough for me."

Since these studies were conducted, both Microsoft and Opera Software have started licensing Netcraft's anti-phishing data, bringing the effectiveness of their browser's built-in anti-phishing on par with Netcraft Toolbar and beyond [6].

F. Machine Learning

With machine learning algorithms, AI was able to develop beyond just performing the tasks it was programmed to do. Before Machine Learning entered the mainstream, AI programs were only used to automate low-level tasks in business and enterprise settings. This included tasks like intelligent automation or simple rule-based classification. This meant that AI algorithms were restricted to only the domain of what they were processed for. However, with machine learning, computers were able to move past doing what they were programmed and began evolving with each iteration.

Machine learning is fundamentally set apart from artificial intelligence, as it has the capability to evolve. Using various programming techniques, machine learning algorithms are able to process large amounts of data and extract useful information. In this way, they can improve upon their previous iterations by learning from the data they are provided, [9]

We cannot talk about machine learning without speaking about big data, one of the most important aspects of machine learning algorithms. Any type of AI is usually dependent on the quality of its dataset for good results, as the field makes use of statistical methods heavily. Machine learning is no exception, and a good flow of organized, varied data is required for a robust Machine learning solution. In today's online-first world, companies have access to a large amount of data about their customers, usually in the millions. This data, which is both large in the number of data points and the number of fields, is known as big data due to the sheer amount of information it holds.

Big data is time-consuming and difficult to process by human standards, but good quality data is the best fodder to train a machine learning algorithm. The more clean, usable, and machine-readable data there is in a big dataset, the more effective the training of the machine learning algorithm will be.

As explained, machine learning algorithms have the ability to improve themselves through training [14].

3. 0 METHODOLOGY AND SYSTEM ANALYSIS

A. Facts Finding

Fact finding is an approach taken to acquire data about a specific or subject with the aim of analyzing and synthesizing the analyzed data to come up with a better system. Fact finding for this study was done by examining related publications, research work, journals and books.

The phishing detection systems are generally divided into two groups: List Based Detection Systems and Machine Learning Based Detection Systems.

1. *List based detection systems:* List-based phishing detection systems use two list, whitelists and blacklists, for classifying the legitimate and phishing web pages. Whitelist-based phishing detection systems make secure and legitimate websites to provide the necessary information. Blacklists are created by URL records, which are known as phishing websites. These list entries are derived from a number of sources, such as spam detection systems, user notifications, third party organizations, etc. The use of blacklists makes it impossible for attackers to attack again via same URL or IP address, which are previously used for attack.

2. *Machine learning based detection systems:* One of the popular methods of malicious websites' detection is the use of machine learning methods. Mainly, detection of phishing attack is a simple classification problem. In order to develop a learning-based detection system, training data must contain lots of features, which are related to phishing and legitimate website classes. By the use of a learning algorithm, it can be easy to detect the unseen or not classified URLs with a dynamic mechanism.

B. Proposed System Design

The system as extensively described in previous chapters seeks to use the standard software development models which in this case is the Waterfall model, to create a standardized anti-phishing system. To achieve this goal above, we:

- Ensure that user details are kept secure.
- Ensure proper maintenance in terms of update of the knowledge base.
- Ensure only admins are granted admin a privilege access to affect the database tables.

C. Architectural Design of the Proposed System

This is where the programs that will run the modules identified in the control centre are specified. This will enable the researcher to capture the complete working picture of the application and how each component is related to another. The system architecture is shown below:

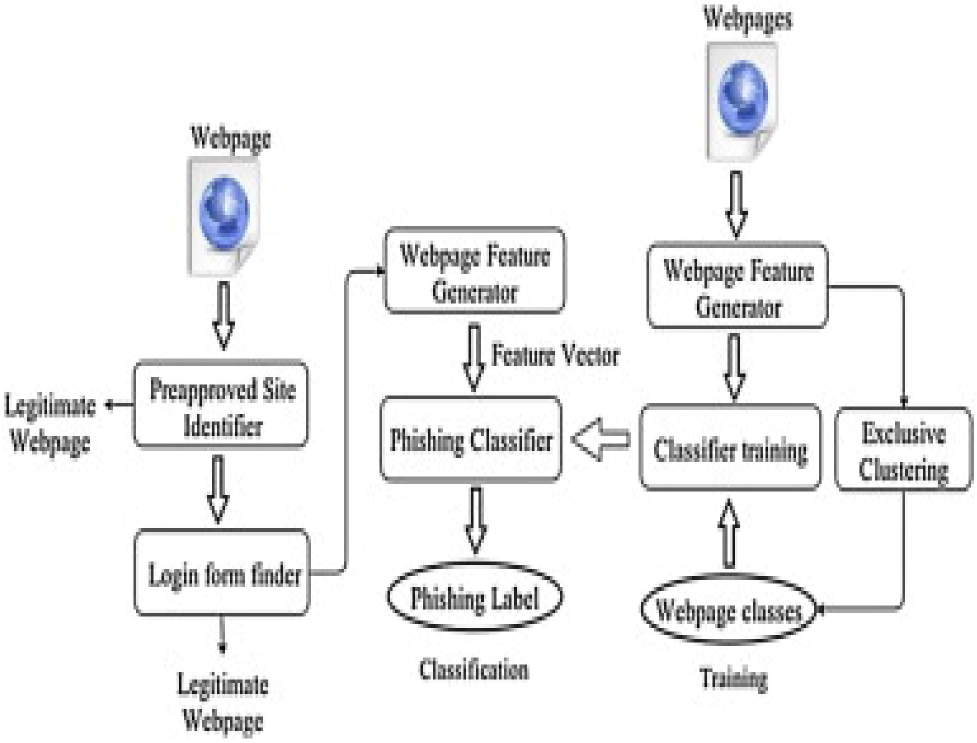


Fig. 1. Architectural design of the anti-phishing system

4.0 SYSTEM DESIGN AND IMPLEMENTATION

This chapter discusses the deployment and testing of the phishing detection system after the design and development. The Hardware and Software Requirements as well as Development tools are identified in this chapter.

A. *Objectives of design / overall system description*

- To accurately identify and classify phishing websites: The detector should use machine learning algorithms and data mining techniques to analyze website URLs and other features to accurately distinguish phishing websites from legitimate ones.
- To provide real-time protection: The detector should be integrated with a user's browser as an extension and operate in real-time, providing users with immediate warnings or blocks when they attempt to access a potentially malicious website.
- To be user-friendly: The detector's user interface should be intuitive and easy to understand, providing clear and concise warnings to users when a potentially malicious website is detected.
- To be efficient: The detector should be designed to be computationally efficient, using minimal system resources and memory to avoid slowing down the user's browsing experience.

- To improve over time: The detector should be designed to continuously improve its performance over time through the use of machine learning and feedback mechanisms'

The overall system description for such a detector would likely include the following components:

- URL Analysis: A module that analyzes website URLs using machine learning algorithms and data mining techniques to identify patterns and features that are indicative of phishing websites.
- Real-time protection: A module that integrates with a user's browser as an extension, providing real-time warnings or blocks when a potentially malicious website is detected.
- User interface: A module that provides a user-friendly interface for users to interact with the detector, receive warnings, and access educational resources.
- Feedback mechanism: A module that allows users to provide feedback on the detector's performance and provide feedback to the machine learning algorithm.
- Update mechanism: A module that allows detector to update itself with latest phishing websites and improve the performance over time.

B. *Program / system design*

1. *Data collection:* The first step would be to collect a dataset of both phishing and legitimate website URLs. This data would be used to train and test machine learning models.
2. *Feature extraction:* Next, the URLs would be preprocessed and features would be extracted, such as domain name, path, number of subdomains, presence of special characters, etc, which would be used as input to the machine learning model

3. Model development: Machine learning models such as Random Forest, SVM, Neural Network, etc. would be developed using the extracted features. The models would be trained and tested using the collected dataset.

4. *Model evaluation:* The performance of the models would be evaluated using metrics such as accuracy, precision, recall, F1 score, etc. to determine which model performs best.

5. *Integration with browser*: The detector would be integrated with a user's browser as an extension, providing real-time warnings or blocks when a potentially malicious website is detected.

The objectives of the design and overall system description for a URL-based phishing website detector would include the following:

- To accurately identify and classify phishing websites: The detector should use machine learning algorithms and data mining techniques to analyze website URLs and other features to accurately distinguish phishing websites from legitimate ones.

- To provide real-time protection: The detector should be integrated with a user's browser as an extension and operate in real-time, providing users with immediate warnings or blocks when they attempt to access a potentially malicious website.
- To be user-friendly: The detector's user interface should be intuitive and easy to understand, providing clear and concise warnings to users when a potentially malicious website is detected.
- To be efficient: The detector should be designed to be computationally efficient, using minimal system resources and memory to avoid slowing down the user's browsing experience.
- To improve over time: The detector should be designed to continuously improve its performance over time through the use of machine learning and feedback mechanisms.

This approach would provide a systematic way to develop a URL-based phishing website detector that can effectively detect and protect against phishing attacks while also being user-friendly and efficient.

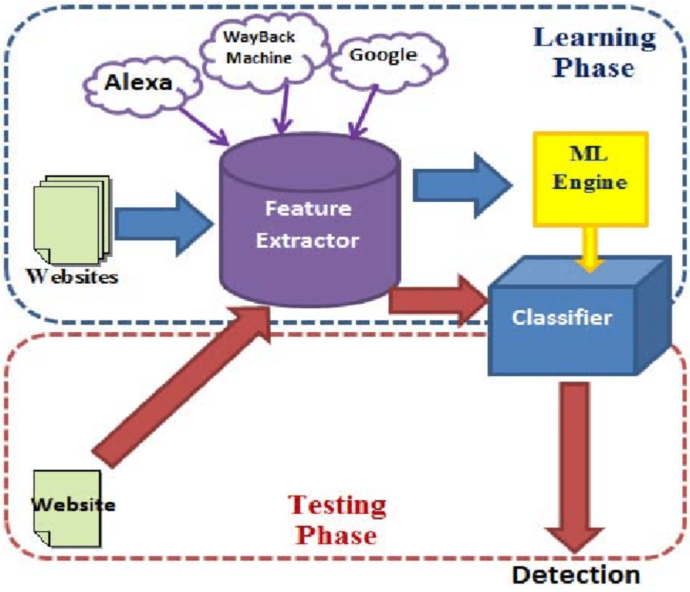


Fig. 2. System architecture

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

C. Algorithm

Natural language algorithm:

1. Take the URL of the website in question as input.

2. Check if the URL is on a list of known phishing websites.
    a. If the URL is found on the list, flag it as a phishing website and display a warning to the user.
    b. If the URL is not found on the list, proceed to the next step.
3. Check if the URL's domain name matches that of a known legitimate website, but with slight variations such as the addition or replacement of certain characters (e.g. "g00gle.com" instead of "google.com").
    a. If a match is found, flag it as a phishing website and display a warning to the user.
    b. If no match is found, proceed to the next step.
4. Check if the website has a valid SSL/TLS certificate.
    a. If the certificate is valid, proceed to the next step.
    b. If the certificate is invalid or missing, flag it as a phishing website and display a warning to the user.
    c. Perform a Google Safe Browsing check on the website.
5. If the website is found to be unsafe, flag it as a phishing website and display a warning to the user.
6. If the website is found to be safe, proceed to the next step.
    a. Check if the website has been reported as a phishing website by a reputable source.
7. If the website has been reported, flag it as a phishing website and display a warning to the user.
8. If the website has not been reported, proceed to the next step.
    a. Perform a machine learning classification on the website using a pre-trained model to detect phishing websites.
9. If the model detects the website as a phishing website, flag it as a phishing website and display a warning to the user.
10. If the model does not detect the website as a phishing website, flag it as safe.

D. Specification

The review of the existing system brought about identification of key areas that need to be improved on and they were also considered in the development of this project.

They include:

1. Data Validation
2. Speed and Reliability
3. Correctness
4. Understandability

E. Hardware and software requirements

Tables 4.1 and 4.2 identify the requirements both hardware and software required to successfully implement the system.

TABLE I. MINIMUM HARDWARE REQUIREMENTS

Minimum Hardware Requirements		
S/N	Server-Side Specification	Client-Side Specification
1	2GHz and above of CPU speed	2GHz and above of CPU speed
2	2GB and above of RAM	512MB and above of RAM
3	10 GB and above of hard disk space	512MB and above of hard disk space
4	Webserver (Apache)	Internet Connectivity
5	Database server (Sql)	

TABLE II. MINIMUM SOFTWARE REQUIREMENTS

Minimum Software Requirements		
	Server-Side Specification	Client-Side Specification
1	Windows OS	Windows OS
2	Apache	JavaScript enabled web browser
3	MySQL	

F. Communication Interfaces

Communication interfaces in this project include (and not limited to) TCP/IP (Transmission Control Protocol/Internet Protocol), HTTPS (Secured Hyper Text Transfer Protocol), FTP (File Transfer Protocol)

G. System Maintenance

Maintaining a machine learning-based phishing detector system that operates on URLs would involve several key steps to ensure its continued effectiveness. Some of these steps might include:

- Regularly updating the system's training data: As new phishing techniques are developed, the system's training data should be updated to reflect these changes so that it can continue to accurately detect new types of phishing attempts.
- Monitoring the system's performance: Regularly monitoring the system's performance metrics such as accuracy, false positive rate and false negative rate, allows to detect if there is any drift and retrain or fine-tune the model.
- Refining the system's parameters: As the system is used, its parameters may need to be adjusted to optimize its performance. This might involve adjusting the weights of different features used by the system, or tuning its threshold for classifying URLs as phishing or legitimate.
- Managing the system's infrastructure: Regularly updating and maintaining the underlying infrastructure of the system is important to ensure its continued reliability. This might involve patching security vulnerabilities, scaling the system to handle increasing traffic, and monitoring for potential system failures.

H. User Competence

The end user of this system just like any other web based application that employs a payment system, should be at least literate in English Language to be able to understand the options and the requests made from and to the server. The users should also have a basic understanding of internet security.

I. Experimental results

This section gives the experimental details of the proposed model's classification algorithms and used feature extraction types (NLP based features, Word Vectors, and Hybrid) are detailed.

1. *Used classification algorithms:* We have used seven different classification algorithms (Naive Bayes, Random Forest, kNN (n = 3), Adaboost, K-star, SMO and Decision Tree) as machine learning mechanism of the proposed system and then compared their performances. The Naïve Bayes classification is a probabilistic machine learning method, which is not only straightforward but also powerful. Due to its simplicity, efficiency and good performance, it is preferred in lots of application areas such as classification of texts, detection of spam emails/intrusions, etc. It is based on the Bayes theorem, which describes the relationship of conditional probabilities.

By the use of data preprocessing as detailed in previous sections, it can be easy to extract some distinctive features. These features are extracted by using the Natural Language Processing (NLP) operations. Therefore, these features depend on the used language. For the efficiency of the system, features are extracted according to the English language; however, according to aim it can be easily adapted to any language. Selection and design of these features are very trivial issues to accomplish, and most of the works focus on phishing detection used different feature list according to their algorithms. The selected features mainly need to parameterize the URL of the web page. Therefore, the text form of web address must be decomposed to the words that it contains. However, this is not an easy task. Because a web address can contain some combined texts in, which finding each word is a trivial task. In this decomposition operation, firstly the URL was parsed by taking into account some special characters such as ("?", "/", ".", "= ", "&").

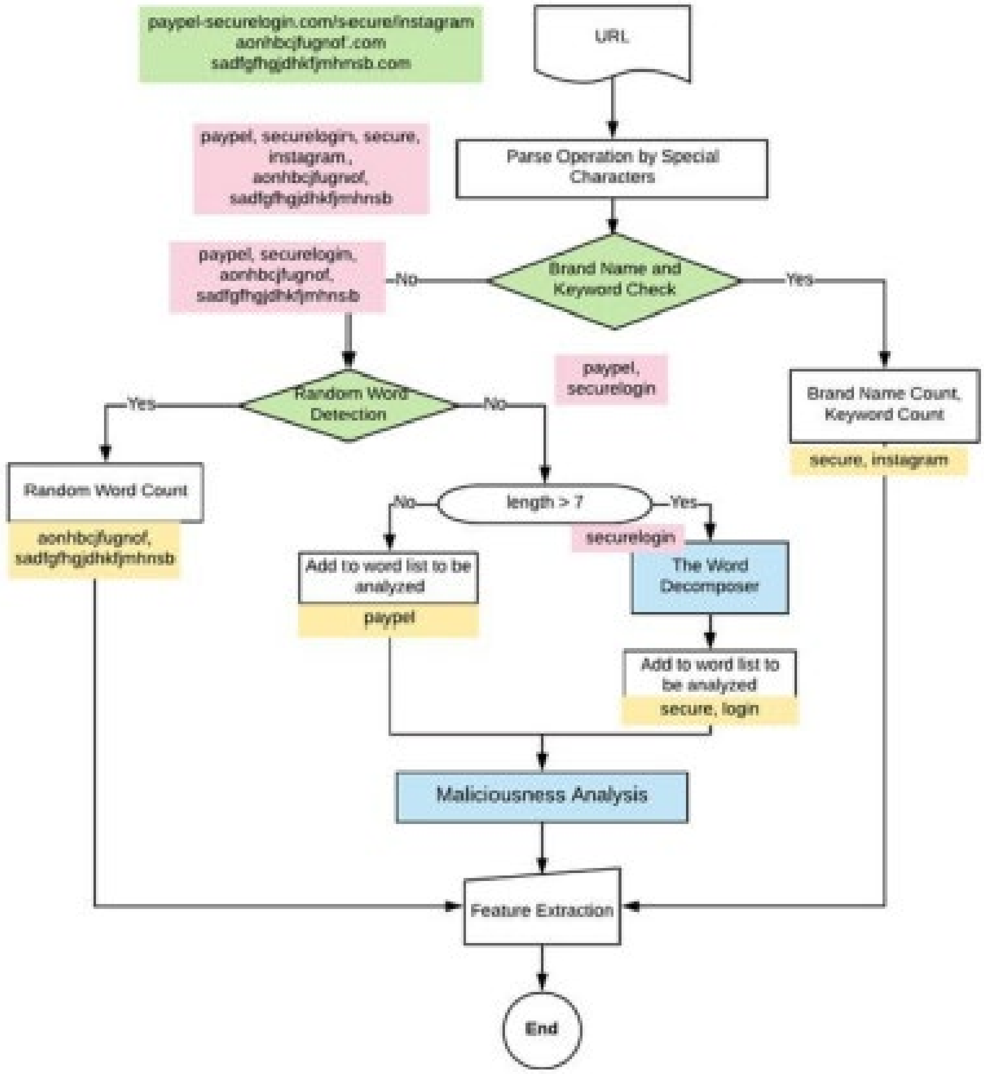

paypel-securelogin.com/secure/instagram
aonhbcjfugnof.com
sadfgfhgjdhkfjmhnsb.com
URL
paypel, securelogin, secure, instagram, aonhbcjfugnof, sadfgfhgjdhkfjmhnsb
Parse Operation by Special Characters
paypel, securelogin, aonhbcjfugnof, sadfgfhgjdhkfjmhnsb
No
Brand Name and Keyword Check
Yes
paypel, securelogin
Brand Name Count, Keyword Count
secure, instagram
Yes
Random Word Detection
No
Random Word Count
aonhbcjfugnof, sadfgfhgjdhkfjmhnsb
No
length > 7
Yes
securelogin
Add to word list to be analyzed
paypel
The Word Decomposer
Add to word list to be analyzed
secure, login
Maliciousness Analysis
Feature Extraction
End

Fig. 3. Execution of malicious analysis module

Then, a raw word list is reached in, which each word can have meaning alone or can be combined with the use of two or more distinct words in a meaningful order. The latter one is especially preferred for the attackers to convince the victim as if it is a legitimate web page. To deceive the users, attackers can use different techniques.

2. *Word Vectors:* In the text processing or text mining approaches, converting words into vectors is mostly preferred for reaching some crucial features. In our system, we are related to the URL of the web page, which is mainly constructed as a text that contains lots of words in it. Instead

of converting these words manually, an automatic vectorization process is preferred. In this module, each URL is converted into word vectors with the help of a specific function of Weka named as “StringtoWordVector”. After obtaining the related vectors, they can be easily used in the selected machine learning algorithm. In the proposed system, 73,575 URLs are used for the testing. In the vectorization process, 1701 word-features are extracted. Then a feature reduction mechanism is applied to decrease the number of features in the list by using a feature selection algorithm named as “CfsSubsetEval”algorithm, which runs with the best first search method. With this reduction mechanism, the sufficient number of features has dropped from 1701 to 102.

3. *Hybrid features:* To increase the efficiency of the proposed system we wanted to combine both features (NLP features and word vectors) in a hybrid model. After the implementation of the word vectorization step, we have totally 1701 word-features, and then we joined them with the 40 NLP features and there were 1741 total features before making a hybrid test. Then a similar feature reduction mechanism is executed, and the total number is decreased to 104 features.
4. *Test results:* One of the important problems for testing the proposed system is the use of a worldwide accepted dataset. We cannot reach this dataset, therefore, produced our own dataset as detailed. The dataset is also published in (Ebbu2017 Phishing Dataset, 2017). Due to its huge size and lack of test de- vice capacity, we have performed our test on this dataset, which contains 73,575 URLs. This dataset contains 36,400 legitimate URL and 37,175 phishing URLs. Experiments are executed on a MacBook Pro device with 2.7 GHz Intel Core i5 processor and 8 GB of 1867 MHz DDR3 RAM. For testing the proposed system Weka was used with some pre- developed libraries. 10-fold Cross Validation and the default parameter values of all algorithms were used during the tests. Each test set is executed with seven different machine learning algorithms. Firstly, the confusion matrix for the tested learning algorithms is constructed.

5.0 SUMMARY, CONCLUSION, AND RECOMMENDATIONS

A. Summary

A URL-based phishing website detector using machine learning is a system that uses machine learning algorithms to analyze the features of a website's URL and determine whether it is a legitimate website or a phishing website. These features can include the structure of the URL, the presence of certain keywords, and other characteristics that are commonly associated with phishing websites. The system is trained on a large dataset of both legitimate and phishing URLs, allowing it to learn the patterns and characteristics that differentiate the two. Once it is trained, the system can be used to automatically classify new URLs as legitimate or phishing. It is important to regularly update the system's training data and fine-tune the system to maintain its effectiveness in detecting new phishing techniques. Additionally, regular monitoring of the system's performance and compliance with laws and regulations is also essential

B. Conclusion

In conclusion, a URL-based phishing website detector using machine learning is a powerful tool for detecting and protecting against phishing attacks. By analyzing the features of a website's URL, the system can accurately identify and flag potential phishing websites, helping to protect users and organizations from falling victim to these types of attacks. However, it's important to keep in mind that machine learning-based phishing detectors are not foolproof, they require maintenance and monitoring to keep up to date with the latest phishing techniques. Additionally, regular monitoring of the system's performance and compliance with laws and regulations is also essential.

C. Recommendation

For future works, improvements can be made in terms of user identification and verification. Data security, data retrieval and fraud detection and reporting should be a vital consideration in development of any further web based machine learning systems.

Based on the analysis of existing and the proposed anti-phishing website detection systems, some recommendations for developing an anti-phishing website detection system will include:

- Incorporate multiple detection techniques: To increase the system's accuracy in identifying phishing websites, it should incorporate multiple detection techniques such as website structure analysis, content analysis, URL analysis, and reputation analysis.
- Use of Machine Learning techniques: To increase the system's ability to adapt to new and evolving phishing tactics, machine learning techniques such as supervised, unsupervised, and deep learning should be used.
- Incorporate real-time processing: To increase the system's effectiveness in blocking phishing attempts, it should incorporate real-time processing capabilities, which allow it to quickly identify and block phishing websites.
- Incorporate user feedback: To improve the system's accuracy, it should incorporate user feedback by allowing users to report potential phishing websites and receive alerts when a potential phishing website is detected.
- Incorporate browser extension: The system could also be integrated with browser extensions, to enable real-time identification and blocking of phishing websites while a user is browsing the internet.

REFERENCES

[1] Ramzan, Zulfikar (2019). "Phishing attacks and countermeasures". In Stamp, Mark; Stavroulakis, Peter (eds.). *Handbook of Information and Communication Security*. Springer. ISBN 978-3-642-04117-4.

[2] Van der Merwe, A J, Loock, M, Dabrowski, M. (2020), Characteristics and Responsibilities involved in a Phishing Attack, Winter International Symposium on Information and Communication Technologies, Cape Town, January 2020.

[3] "Landing another blow against email phishing (Google Online Security Blog)". (2021).

[4] Dudley, Tonia.(2019) "Stop That Phish". Archived from the original on 21 March 2021.

[5] "What is Phishing?". (2016). Archived from the original on 16 October 2016.

[6] "Internet Crime Report (2020)" (PDF). *FBI Internet Crime Complaint Centre*. U.S. Federal Bureau of Investigation. Retrieved 21 March 2021.

[7] Wright, A; Aaron, S; Bates, DW (2016). "The Big Phish: Cyberattacks Against U.S. Healthcare Systems". *Journal of General Internal Medicine*. **31** (10): 1115–8. doi:10.1007/s11606-016-3741-z. PMC 5023604. PMID 27177913.

[8] Ollmann, Gunter(2016). "The Phishing Guide: Understanding and Preventing Phishing Attacks". *Technical Info*. Archived from the original on 2011-01-31.

[9] Mitchell, Anthony (2020). "A Leet Primer". TechNewsWorld. Archived from the original on April 17, 2019

[10] "Phishing". *Language Log, September 22, 2019*. Archived from the original on 2016-08-30.

[11] Jøsang, Audun; et al. (2007). "Security Usability Principles for Vulnerability Analysis and Risk Assessment". *Proceedings of the Annual Computer Security Applications Conference 2017 (ACSAC'07)*. Archived from the original on 2021-03-21. Retrieved 2020-11-11.

[12] Aleksandersen, Daniel (16 August 2016). "Most of the alternate web browsers don't have fraud and malware protection". *Slight Future*. Retrieved 25 August 2016.

[13] Carnegie Mellon University(2016)"Phinding Phish: An Evaluation of Anti-Phishing Toolbars" (PDF). Archived from the original (PDF) on 2017-06-10. Retrieved 2018-05-25.

[14] Barraclough, P.A., Hossain, M.A., Tahir, M.A., Sexton, G., &Aslam, N. (2018) Intelligent Phishing Detection and Protection Scheme for Online Transactions. Expert Systems with Applications, 40, pp 4697-4706.

[15] Purkait, S. (2019) Phishing Counter Measures and Their Effectiveness – Literature Review.Information Management and Computer Security, 20 (5), pp 382-420.

[16] Ma, L., Ofoghi, B., Watters, P. & Brown, S. (2019) Detecting Phishing Emails Using Hybrid Features. Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, p. 493–497

Enhancing Industrial Chemical Process based on Delay Cancellation and Improved Transient Response Performance

Callistus T. Ikwuazom[1], Donatus O. Njoku[2], Janefrances E. Jibiri[3], Perpetual N. Ibe[4]

[1]Department of Information Technology, Federal University of Technology,
Minna-Niger State-Nigeria
[2] Department of Computer Science Federal University of Technology, Owerri,
Imo State-Nigeria
[3]Department of Information Technology, Federal University of Technology, Owerri
Imo State-Nigeria
[4]Department of Computer Science, Imo State Polytechnic,Omuma
Imo State-Nigeria

Abstract: This paper presents enhancing industrial chemical process based on delay cancellation and improved transient response performance. It is desired to enhance the performance of an industrial chemical process. In order to achieve this, the dynamic characteristics of a chemical reactor that involves pH concentration in water treatment carried out using continuous stirred tank reactor (CSTR). A compensator was designed using the Control and Estimation Tools Manager (CETM). The results from the simulation conducted in MATLAB environment indicated that the compensator was able to cancel the time delay effect suffered by process and improved the transient response performance.

Keywords: CETM, Compensator, CSTR, Delay cancellation, Transient response

1. Introduction

As industrial chemical process and their control systems become more complex, achieving higher efficiency and improved control performance becomes more challenging. Also, the process become harder to control and maintain a steady state. For this reason, there is need to develop a system which will assist plant operators and enhance the general production process while ensuring effective control. Example of such industrial chemical process is the chemical reactor.

Chemical reactors are indispensable and influential factors in industrial chemical process. Chemical reactions in reactors can be in the form of continuous processing or batch processing. Continuous processing offers advantages over batch processing such as reduce labour, increased throughput, reduced production cost and so on. One area of industrial processes where chemical reactors have been largely deployed is the wastewater treatment.

Wastewater from industrial process comes out as pollutants. Treating some of these pollutants is costly and difficult. Also, the characteristics of wastewater may change significantly with respect to industrial activities. Nevertheless, the most essential possibility of wastewater treatment process is to regulate the effectiveness of such harmful wastewater. Since the pH is the most important characteristic of wastewater [11], pH-control to maintain desired level of wastewater concentration that is not harmful becomes worthwhile. This is known as neutralization.

Many different control algorithms and techniques have been developed and applied to wastewater treatment in chemical reactor. Also, application of computational control technique in chemical processes has been examined in [8]. This paper presents a Proportional-Integral-Derivative (PID) Tuned Compensator (PID-TC) for delay prone pH process in chemical reactor.

2. Literature Review

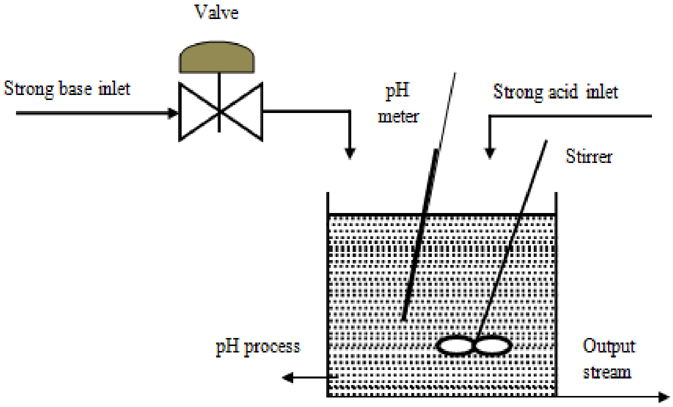
In this section the review of related works are presented based on the control technique or method used to achieved desired performance result. Conventional and particle Swam optimization based PID Control technique. According to [5] presented turning of proportional integral and derivative (PID) controllers for unstable continuous stirred tank reactors (CSTR). It proposed two PIDs controller design techniques for unstable Second Order plus Time Delay System with a Zero (SOPTDZ) based on internal model control (IMC) and Stability Analysis (SA) principle. The controllers were used to control unstable CSTRs whose reaction of order irreversible reaction performance comparison was carried out with the proposed method and the synthesis method. The simulation, results obtained showed that the controller designed by proposed showed more robust performance than the one designed using synthesis method on nonlinear unstable CSTRs. [9] presented Proportional Integral Derivative (PID) of a continuous stirred tank reactor (CSTR). It used non isothermal CSTR and different control modes. Simulation parameters were chosen at equilibrium state and dynamic point. It maintained that simulation results obtained indicated that stability of the non-isothermal CSTR at different turning point and disturbances. In [2] presented modelling and control of CSTR with PID Controller. A model of dynamics control for CSTR in methanol synthesis in a three-phase system was developed. Simulation of the reactor was performed for steady and transient states. Efficiency ratio for achieving maximum performance of the output for a unit reactor volume was calculated. Simulation in closed loop was conducted which allow the control process to receive data for optimum production capacity, with the elimination of local hot spot or temperature runaway. [6] studied design and analysis of PID controller for CSTR process. The objective of the study was to control temperature and load disturbance rejection of CSTR. Simulation result showed that the PID controller provided a less percent overshoot with efficient load disturbance rejection with a minimum setting time. [10] studied design of fractional order PID controller for a CSTR process. It proposed the application of fractional order PID (FOPID) controller in CSTR process controls. It employed soft computing techniques which comprises genetic algorithm (GA) and particle swarm optimization (PSO) to model the CSTR and for obtaining model parameters. It maintained that the developed model was able to compensate for the nonlinearity present in the CSTR. Simulation results were presented in terms PID and FOPID. The performance was analyzed with respect to Integral Square Error (ISE). It was observed that FOPID provided better performance than PID. [1] presented design and implementation of a PID controller for a CSTR system using particle swarm algorithms. It applied proportional integral (PI) and PID controller tuned with PSO, Adaptive Weighted PSO (AWPSO) algorithms to CSTR process to take care of the temperature and concentration control. Three error criteria were used to achieve the optimization process. These include integral of square error (ISE), the Integral of Absolute Error (IAE) and integral of Time Absolute Error (ITAE). In order to test the robustness of temperature and concentration performance of the process, some of the parameters of CSTR were altered. It maintained that better performance algorithm was observed. [4] carried out tuning of PID Control using optimization techniques for a multi-input-multi-output (MIMO) process. It considered two processes which consists of Quadruple Tank process and CSTR process. Dynamic model of the two processes was performed by linearization of the system due to MIMO process. In order to tune the controller parameters, two optimization techniques consisting of PSO, and GA were use. Performance comparison was for the two different optimization techniques used for tuning of PID controller gain parameters for the two process considered. It stated that the simulation results showed that PSO based turned provided better response than that of GA whereas, for the Quadruple Tank process, both optimization techniques provided almost the same response with slight difference in their peak overshot values.

3. System Design

This section presents the mathematical description of the chemical reactor process in a continuous stirring tank reactor (CSTR) and the subsequent design of a compensator that will be implemented as part of the networked system for ensuring that a predetermined pH concentration is maintained.

3.1 Mathematical Description of pH Process

A pH neutralization process for wastewater treatment is shown in Fig. 1. In this process, strong acid (HCl) and strong base (NaOH) of 1 molarity [11] are used.



Fig, 1 Model of a pH process [7]

Assuming no chemical reaction at given level, the equation for material balance can be expressed by:

[Rate of accumulation within vessel volume] = [Inflow rate to the pH process] -[Outflow rate from the pH process] (1)

$$V\frac{dX}{dt} = U - FX \qquad (2)$$

where V is the volume of the mixture in the pH process, X is the state variable of the nonlinear pH process, U is the pH process input flow rate, F is the pH output flow rate. The back titration curve for process flow is given by:

$$T_{(pH)} = X \qquad (3)$$

In a chemical reaction, the components of the system for the back titration curve (TC) produce the nonlinearity of the process flow and can be represented by:

$$C_{TC} = \frac{A[pH] + \sum_{i=1}^{n} a_i[pH]C_i}{\sum_{i=1}^{n} a_i[pH]X_i} \qquad (4)$$

wherea_i[pH] is the acid-base weighting factor (-1 for Strong acid and + 1 for strong base), n is the number of ions present in the reactor, C_i ion concentration of the i^{th} kind of process flow, X_i is the concentration of the i^{th} kind of neutralization liquid.

The value of pH is defined as the negative logarithmic value of the concentration of Hydrogen $[H^+]$ ions.

$$pH = -\log\left(H^+\right) \qquad (5)$$

For the neutralization process

$$A[pH] = 10[-pH_{sv}] - 10[pH_{sv} - 14] \qquad (6)$$

where pH_{sv} is the setpoint (or desired) value

In the neutralization process, the difference value between the actually measured pH and the set point value in line with the nonlinear conversion is given by:

$$Y = T(pH_{sv}) - T(pH) \qquad (7)$$

The First order differential equation is given by Eq.(2) and since X is equal to T(pH), substituting into Equation (2) gives:

$$V\frac{d[pH]}{dt} = U - F[pH] \qquad (8)$$

Taking the Laplace transform of Eq. (8) and rearranging gives:

$$VspH(s) + FpH(s) = U(s) \qquad (9)$$

Further rearrangement of Eq. (8) gives:

$$\frac{pH(s)}{U(s)} = \frac{\frac{1}{F}}{\frac{V}{F}s + 1} \qquad (10)$$

where V/F is equal to the process time constant τ and 1/F is the process gain, K.

The pH process is a first order system with time delay due to pipe line and detection process for the measuring instrument (sensor). Hence it takes the general transfer function model of first order plus time delay (FOPTD) given by:

$$G(s) = \frac{Ke^{-T_D s}}{\tau s + 1} \qquad (11)$$

where $e^{-T_D s}$ is the time delay.

Substituting parameters obtained from real time experiments conducted in Dinesh and Deepika (2014) which was applied in Ram et al (2016), with K = 0.276, $\tau = 3.2$, and T_D = 5.005, Equation (11) becomes:

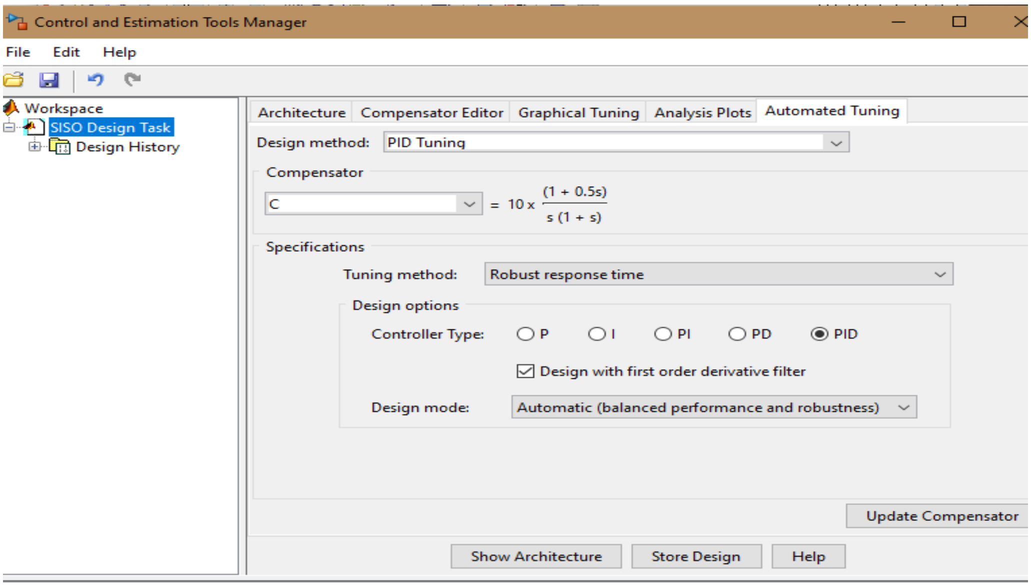
$$G(s) = \frac{0.276e^{-5.005s}}{3.2s + 1} \qquad (12)$$

Therefore, Eq. (12) is the established transfer function model for a pH neutralization process for wastewater treatment considered in this paper. It can be observed that the process is prone to delay.

3.2 Design of Compensator

Compensators are used as part of industrial process control loop or network whenever the response or output of the process is unstable and required to be stabilized to meet specific performance [4] or it is desired to eliminate certain limitation to proper or efficient process performance such as overcome the effect of time delay. A PID-TC was designed using the Control and Estimation Tools Manager (CETM) of the MATLAB to eliminate the effect process delay in a chemical reactor.

The MATLAB CETM tool can be used to design single input single out (SISO) closed loop network to regulate the transient characteristics and output (or response) of an industrial process. A PID-TC compensator has been implemented in [3] and [4], where it has shown to provide robust performance and ability to handle disturbance effect. Thus, similar approach is used in this paper to design a compensator for pH neutralization process in a CSTR. The tuning method employed is robust response time [3],[4]. However, the design mode employed was automatic (balanced and performance and robustness). Figure 2 shows the CETM too graphical user interface used to design the compensator. The designed compensator C



Design method: PID Tuning

C = 10 x (1 + 0.5s) / s (1 + s)

Tuning method: Robust response time

Controller Type: ○ P ○ I ○ PI ○ PD ● PID

☑ Design with first order derivative filter

Design mode: Automatic (balanced performance and robustness)

Fig. 2 Graphical user interface tuning of the MATLAB

The designed compensator C, with gain K = 10, real pole and real zero locations of -1 and -2 is given by:

$$C = 10 \times \frac{(1+0.5s)}{s(1+s)} \qquad (13)$$

(13)

The designed system, which is proposed to compensate the closed loop network for regulating a chemical reactor process involving pH neutralization in water treatment, is shown in Fig. 3.

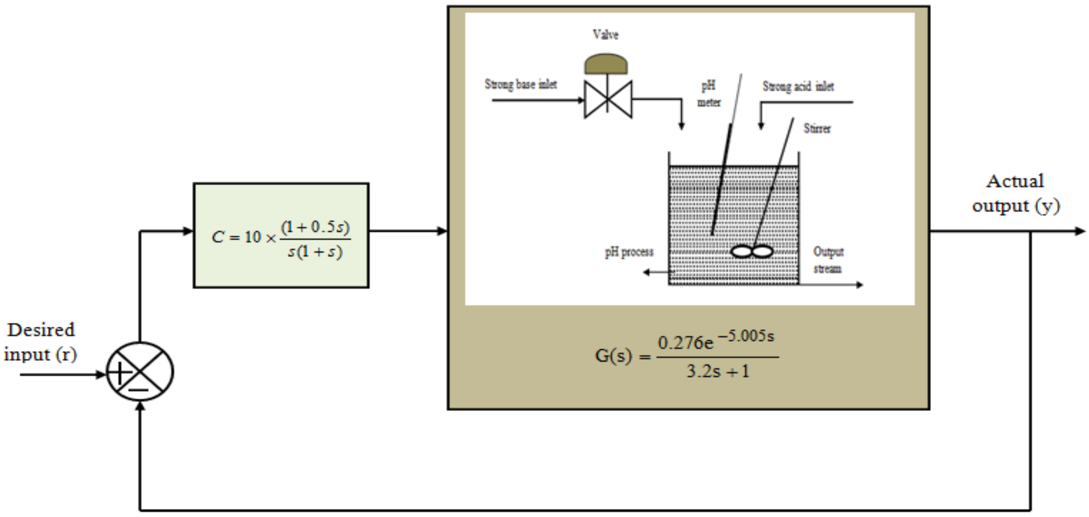


Fig. 3 System configuration

The closed loop network shown in Fig. 3 described the configuration of compensated chemical reactor process examined in this paper. As shown, the actual output, which is the response (or current pH concentration of the water treatment process) is fed back to the summing point where it is compared with the actual value or pH level expected (desired input). The difference or deviation between the desired level of pH concentration and the actual level of pH concentration is fed into compensator, which in turns manipulate the information via mathematical computation in order to send a correctional signal or command that ensures that chemical process is adjusted to meet the desired level at the input. This process continues until little or no deviation exists between the desired input and actual output such that the system settles.

4. Results

This section presents the computer simulation analysis carried to evaluate the response performance of the water treatment reactor using MATLAB. The simulations were conducted for three scenarios which include: the open loop network

of the water treatment process, the closed loop network of the water treatment process without compensator, and the closed loop network of the water treatment process with compensator. In the open loop network, the system was evaluated assuming the actual output (or system response) and the desired input (that is the expected level of concentration) were not relatively compared. That is assuming no fraction of the actual output (the current pH concentrate level) is compared against the desired level. In the closed loop network without compensator, the system was analysed assuming a fraction of the actual output is compared relatively to the desired level but with no technique to ensure that the result of the comparison was used to ensure the chemical process is forced or made to meet a specified level of concentration. For the closed loop network with compensator, the system was evaluated to ensure that the result of the comparison is utilized by a technique that offers command to make sure the process response meets the desired level of concentration while eliminating the delay in the process. The simulation curves for the three scenarios are shown in Figures 4, 5, and 6. Table 1 is the numerical analysis of the system performance in terms of the transient characteristics of the response in time domain for each simulation curve obtained from the three scenarios.

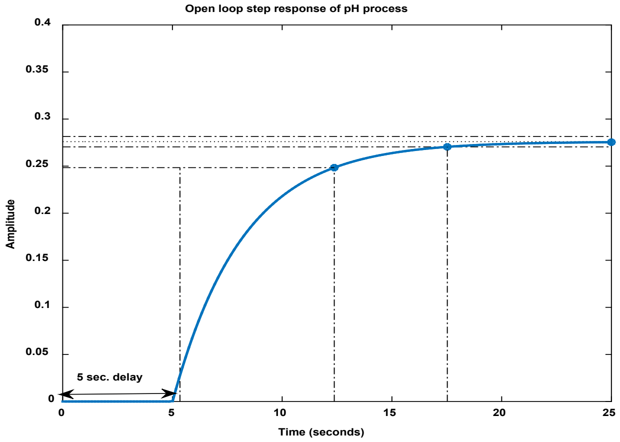


Figure 4 Step response of pH process in open loop

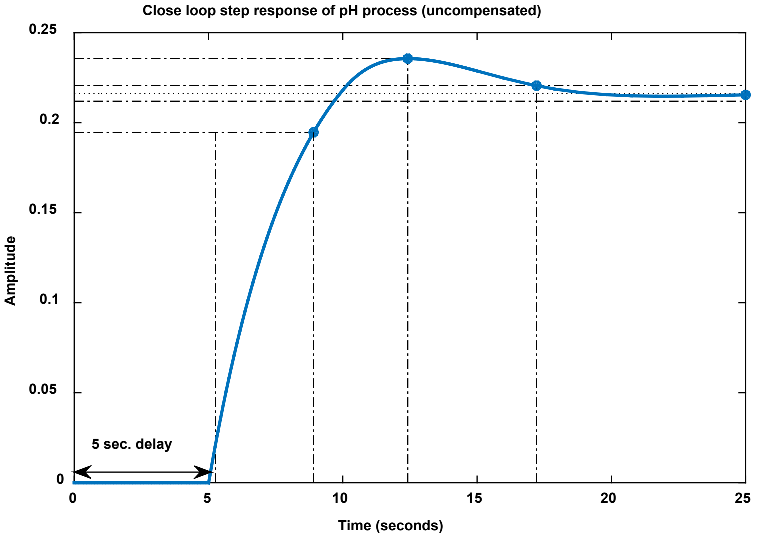


Figure 5 Step response of pH process in uncompensated closed loop

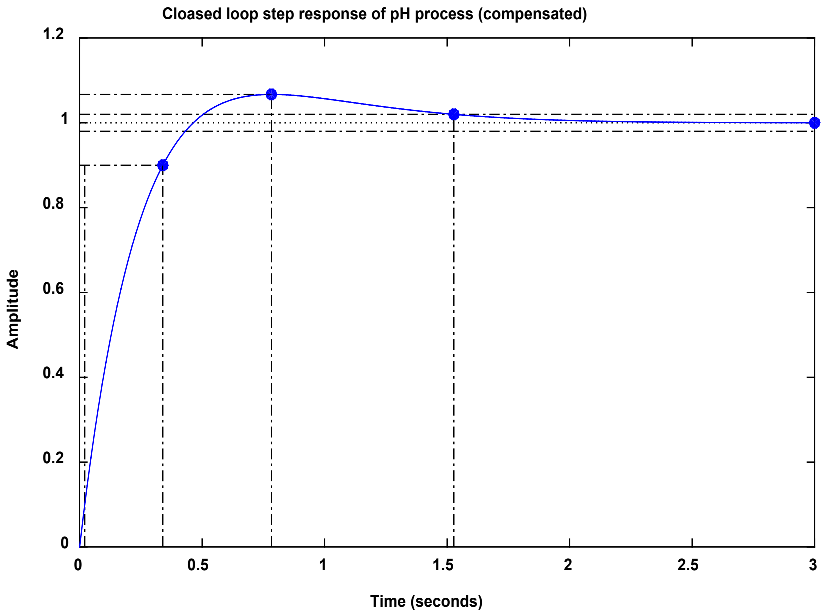


Figure 6 Step response of pH process in compensated closed loop

Table 1 Numerical analysis of transient response Characteristics

Case	Rise time (s)	Peak time (s)	Overshoot (%)	settling time (s)	Final value
Open loop network	7.03	25	0	17.5	0.275
Closed loop network without compensator	3.65	12.4	8.94	17.2	0.217
Closed loop network with compensator	0.319	0.783	6.7	1.53	1

Figure 4 is the simulation result of the evaluation of the step response performance of the system in open loop network arrangement. It can be seen that the still suffers a delay in response to input forcing signal for 5 seconds, and this largely affects the rise time and settling time as the values of these transient parameters can be seen to be very high 7.03 seconds and 17.5 seconds respectively. Thus, if the system is allowed to run in this mode, it will take 7.03 seconds before it will come up for chemical reaction process to start and even when it begins to run, it will take such a long time (17.5 seconds) for it to settle. Though, the system show promising performance in terms of overshoot (0%), this benefit is crushed due to the associated delay and the fact that the output (0.275 or 27.5%) fall short (by 72.5%) of the desire level of concentration, which is assumed as unit step input (that is 1 or 100%). In Figure 5, the uncompensated closed loop network is evaluated to observe the behaviour of transient parameters. As shown in Table 1, the system in this condition still suffers the effect of 5 seconds time delay, but outperformed the open loop system in terms of rise time, peak time, and settling time which are 3.65 seconds, 12.4 seconds and 17.2 seconds respectively. Nevertheless, the uncompensated closed loop system falls short in terms of overshoot and final value compared to the open loop system. The level of the actual output concentration (which is 0.217 or 21.7%) is 78.3% less than the desired level of concentration. This is unsatisfactory. With the compensator added to the closed loop network, simulation analysis revealed that the time delay suffered by the system in the previous modes was completely eliminated as shown in Figure 6. In addition, the system offers rise time of 0.319 second, peak time of 0.783 second, overshoot 6.7%, settling time of 1.53 seconds, and final value of 1. Thus, with the proposed compensator, the chemical process can achieve expected pH concentration at a very fast time (in terms of rise time) and with very much reduced overshoot including rapidly reaching and settling at the desired level of concentration. Thus 100% concentration level can be achieved using the developed compensator.

5. Conclusion

Simulations have been carried out in MATLAB environment to evaluate the performance of chemical reactor process involving maintaining a given pH concentration in water treatment facility. It was observed that the process suffered from delay which has duration of 5 seconds. The simulation analysis of the system using the proposed compensator revealed that the transient response of the system was improved and the delay was eliminated.

References

[1] Aboelela M.A.S., Hennas, R.H.M., Dorrah, H.T. (2015). Design and Implementation of a PID Controller for a Continuous Stirred Tank Reactor (CSTR) System using particle Swarm Algorithms Conference Paper, 1-12: https://www.researchgate.net/publication/281068326

[2] Arthur, W. (2016). Modelling and Control of Continuous Stirred Tank Reactor with PID Controller. Ecological Engineering, 49, 195-201.

[3] Ekengwu, B. O., Eze, P. C., Nwawelu, U. N., and Udechukwu, F. C. (2021). Effect of PID Tuned Digital Compensator on Servo-based Ground Station Satellite Antenna Positioning Control System. 2nd International Conference on Electrical Power Engineering (ICEPENG 2021), 18 – 22 May, 97-101.

[4] Eze, P. C., Ugoh, C. A., and Inaibo, D. S. (2021). Positioning Control of DC Servomotor-Based Antenna Using PID Tuned Compensator. Journal of Engineering Sciences, 8(1), E9–E16. doi:10.21272/jes.2021.8(1).e2

[5] Krishna, D., Suryanarayana, K., Apara, G., and Padmasrez R. (2012). Tuning of PID Controllers for Continuous Stirred Tank Reactors. Indian Chemical Engineer, 54(3), 157-179

[6] Maheshwari, N. Jain, N., Jingar, A., and Suthehar M. (2016). Design and Analysis of PID Controller for CSTR Process. International Journal of Science Engineering and Technology Research 5(2)

[7] Njoku, D. O., Juliet, O., Nwokorie, E. C., and Nwokonkwo, O. C. (2022). A Hybrid Intelligent Control Model for Regulating pH In Industrial Chemical Process. Journal of Electrical Engineering, Electronics, Control and Computer Science, 8(29), 1-8.

[8] Njoku, D. O., Nwokorie, E. C., Asagba, P. O., and Nwokonkwo, O. C. (2020). Application of Computation Control Techniques in Industrial Chemical Process: A Review.

International Conference on Emerging Applications and Technologies for Industry 4.0, EATI 2020, 124-131.

[9] Ojiabo, T.K., and Igbokwe, P.K. (2015). Proportional Integral Derivative Control (PID) of a Continious Stirred Tank Reactor (CSTR). The International Journal of Engineering and Science, 4(10), 66-73.

[10] Poovarasan, J., Kayalvizhi, R. And Pongianan, R.K. (2014). Design of Fractional Order PID Controller for a CSTR Process. International Referenced Journal of Engineering and Science, 3(1), 08-14.

[11] Ram, S.S., Kumar, D.D. and Meenakshipriya, B. (2016). Designing of PID Controllers for pH Neutralization Process. Indian Journal of Science and Technology, 9(12), 1-5.

WEBSITE PHISHING DETECTION USING MACHINE LEARNING ALGORITH

UFUOMA CYRIL OGUDE[1], RAHAMAT OYAMINE NASIRU[2], ONWUACHU UZOCHUKWU C[3].

[1]*Department of Computer Sciences, University of Lagos, Akoka-Yaba, Lagos, Nigeria ,*
uogude@unilag.edu.ng
[2]*National Open University of Nigerian, Victoria Island, Lagos, Nigeria*
nasirurahamat@gmail.com
[3]*Department of Computer Science, Imo State University, Owerri, Imo State, Nigeria.*
onwuachu.uzochuku@imsu.edu.ng

ABSTRACT

Phishing attacks cost internet users and organization billions of dollars every year and has become a rapidly growing threat in the cyberspace. It is illegal to gather sensitive information from consumers through a number of social engineering techniques such as Email, instant messaging, pop-up messages, web pages, and other forms of communication can all be used to identify phishing tactics. This work offers a model that can determine whether a URL link is legitimate or phishing. The data set used for the classification was sourced from the University of New Brunswick dataset bank, which has a collection of benign, spam, phishing, malware, and defacement URLs, as well as from an open-source service called "Phish Tank," which contains phishing URLs in multiple formats such as CSV, JSON, etc. Phishing URLs are identified using deep neural network models. This paper create a web application software that can easily identify phishing URLs from a database of more than 10,000 URLs that have been randomly selected, divided into 50% training samples and 50% testing samples, and have up to 24,442 phishing and 5000 legitimate URLs. To distinguish between legal and phishing URLs, the URL dataset is trained and tested using feature selections like address bar-based features, domain-based features, HTTPS& JavaScript-based features. The result offered a strategy for categorizing URLs into real and phishing URLs by authenticating every link that is sent to them.

Keywords: Phishing, Deep Neural Network, Cyberspace, Features extraction and communication

1.0 INTRODUCTION

The Internet, particularly social media, has become a significant component of our lives for gathering and spreading information. Pamela (2021) claims that the Internet is a network of computers that houses important data. Security measures strive significantly more to keep users' data and devices secure when they easily give away their data or access to their computers. As a result, Imperva (2021) describes social engineering (a sort of attack designed to acquire user data, such as login passwords and credit card details) as one of the most common types of social engineering assaults. When an attacker deceives a victim into opening an email, instant message, or text message that appears to be from a trusted source, the attack occurs. When the recipient clicks the link, they wrongly believe they've gotten a present and unknowingly click a harmful link, which leads to the installation of malware, the freezing of the machine during a ransom ware assault, or the release of private data.

Due to the rapid adoption of technological advancements, there has been a significant growth in computer security threats in recent years, which has also increased the vulnerability of human exploitation. Users should be informed of the methods used by phishers as well as ways to help against falling victim to phishing. As technology develops, cybercriminals' tactics get more sophisticated. There are other ways to get consumers' personal information aside from phishing. According to KnowBe4 (2021), the following methods applies:

a) Vishing (also known as voice phishing) involves the phisher calling the victim to obtain personal information regarding the bank account. The use of a fake caller ID is the most typical method of phone phishing.
b) Smishing (SMS Phishing): Smishing is the practice of sending phony messages using the Short Message Service (SMS). By delivering a link to a phishing website, it is a technique for seducing a target using the SMS text message service.
c) Ransomware: A ransomware attack is a kind of attack that denies users access to a device or data unless they pay a ransom.
d) Malvertising: Malvertising is malicious advertising that use live scripts to push unwanted material or download malware onto your machine. Exploits for Adobe PDF and Flash are the most often utilized methods in malicious advertisements.

Consequently, this poses a growing threat to both large and small businesses as well as to people. Now that criminals have access to industrial-strength services on the dark web, there are more phishing URLs and emails being sent out and, more worrisomely, they are getting better and harder to spot.

2.0 LITERATURE REVIEW

Anjum et al, (2016) published a thorough study with the title A Literature Review on Phishing Crime, Prevention Review, and Investigation of Gaps. Various reviews of prior works of literature are offered. In order to combat phishing, the report suggests using CRI, which stands for Crime, Prevention Review and Investigation of Research Gap.

Ashritha et al, (2019) reviewed Detection of Phishing Websites Using Machine Learning suggested many algorithms (models), as well as various elements of phishing assaults and strategies to detect phishing websites. The paper's discussion of works and various phishing detection techniques is one of its strong points. Additionally, it presents a suggested mechanism for precisely predicting phishing websites. The inquiry hole gives researchers additional room to explore phishing detection.

Kiruthiga. & Akila, (2019) outlined an innovative technique for employing machine learning algorithms to identify phishing websites. Additionally, they evaluated the performance of five machine learning algorithms: Generalized Linear Model (GLM), Generalized Additive Model (GAM), Gradient Boosting (GBM), Random Forest (RF), and Decision Tree (DT) (Shad & Sharma, 2018). Each algorithm's accuracy, precision, and recall assessment metrics were computed and compared. The performance of the top three algorithms, Decision Tree, Random Forest, and GBM, was compared in the table. The Random Forest algorithm yielded the maximum 98.4% accuracy, 98.59% recall, and 97.70% precision, according to the tables of accuracy, recall, and performance.

Sönmez et al. (2018), propose a categorization approach to classify phishing attacks. Website classification and feature extraction from web pages are included in this approach. The ideas for phishing feature extraction have been explained, and thirty features have been extracted from the UCI Irvine machine learning repository data set. The data was classified using these features using Support Vector Machine (SVM), Naive Bayes (NB), and Extreme Learning Machine (ELM) (Sönmez et al., 2018). The Extreme Learning Machine (ELM), which exceeded SVM and NB in accuracy with 95.34%, used six activation functions. The results were assisted by the usage of MATLAB.

Peng et al.(2018) offer a method for identifying phishing email attacks using machine learning and natural language processing. In order to find malicious intent, the text is subjected to a semantic analysis. Each sentence is parsed using a natural language processing (NLP) technique to determine the semantic roles of the words in relation to the predicate. The Nazario phishing email set dataset is utilized in conjunction with Python programs to create this technique. Comparison of Net-craft with SEAHound results (Peng et al., 2018) reveals precision of 98% and 95%, respectively.

The Table 2.1 shows related algorithms proposed by several researchers in Machine Learning to detect phishing websites. On reviewing their papers, they concluded that most of the work done is by using familiar machine learning algorithms like Naïve Bayesian, SVM, Decision Tree, and Random Forest. Some authors proposed a new system like Phish Score and Phish Checker for detection. The combinations of features with regards to accuracy, precision, recall, etc. were used. Experimentally successful techniques in detecting phishing website URLs were summarized in Table 2.1.

Table2.1 :Outline of related algorithms used to detect phishing website

Algorithm used	**Reference paper**	**No. of Features**	**Dataset**	**Language /Tools**	**Conclusion**
Decision Tree (DT), Random Forest(RF), Gradient Boosting (GBM), Generalized Linear Model (GLM), Generalized Additive Model(GAM)	Shad, and Sharma, (2018)	30	Not Mentioned	Python,R Language	Random Forest highest accuracy 98.4%
Support vector Machine (SVM),Naïve Bayes (NB) and Extreme Learning Machine(ELM)	Sönmez, et al (2018)	30	UCI-Machine Learning Repository	MATLAB	ELM achieved 95.34% accuracy.
Natural Language Processing	Peng, et al. (2018)	-	Nazario phishing Emailset	Python	Proposed SEA Hound provides 95% accuracy
Random Forest	Saimadhu. (2017)	8	Phish tank,	R Studio	95% accuracy
Neural network model Adam AdaDelta and SGD	Shreya, (2020)	URL length	Phishtank	Chainer	Accuracy of Adam 94.18%
Convolution neural network(CNN) and SNN long short-term memory(CNN-LSTM)	Kondeti et al. (2021)	-	Phishtank, Open Phish, Malware Domain list, Malware Domain	Tensor Flow in conjunction with Keras	CNN-LSTM obtained98% accuracy
Logistic Regression and Support Vector Machine(SVM)	Noel, (2016).	19	UCI machine learning repository	BigData	SVM accuracy95.62%

daBoost, Bagging, Random Forest, and SMO	Kartik, (2021)	11	Direct Industry Anti-Phishing Alliance of China	BigData	Only Semantic Features of word embedding obtained high accuracy.
C4.5 decision tree	Almomani et al. (2015)	9 Features and heuristic values	Phishtank Google	-	89.40%
KNN,SVM and Random Forest	Gupta et al (2016)	22	UCI-Machine Learning Repository	HTML, JavaScript, CSS, Python	Random Forest high accuracy
Naïve Bayesand Sequential Minimal Optimization(SMO)	Rishikesh & Irfan, (2018).	133	Phishtank Google	C# programming and R programming WEKA	SMO Beat accuracy Than NB
Heuristic feature root mean square Error(RMSE)	Rami et al. (2015),	6	PhishTank	MYSQL.PHP	97%
Phish Score	Shaikh et al. (2016)	12	PhishTank	-	94.91%
Phish Checker	Abdelhamid et al. (2017)	5	PhishTank and Yahoo directory set	Microsoft Visual Studio Express2013 and C# language	96%

3.0 MATERIALS AND METHODS

The new phishing detection system makes use of Random Forest, Multilayer Perceptions, Auto Encoder Neural Network, Support Vector Machine, Decision Tree, and XGBooster. These models were chosen based on several comparisons between the results of various machine learning methods. These models are each tested and trained using a website content-based feature that is taken from phishing and authentic datasets. Therefore, the most accurate model is chosen and implemented into a web application that will allow a user to determine whether a URL link is authentic or phishing

Data Collection

Different open-source platforms provide the data that is utilized to create the datasets used to train the models. The dataset collection includes both legal and phishing URL datasets. The collection of phishing URLs comes from Cisco Talos Intelligence Group's open-source Phish Tank service. This site offers a collection of phishing URLs that are updated every hour in a variety of forms, including CSV, JSON, and others. The dataset was collected from the phishtank.com website. Over 24,442 random phishing URLs are gathered from this dataset to train the ML models.

The University of New Brunswick's open datasets provide the set of legitimate URLs accessible on the university website. This dataset contains a collection of URLs that aren't malicious, spamming, phishing, or defacement. The legitimate URL dataset is taken into consideration for this study out of all of these types. Over 5000 randomly selected valid URLs from this dataset are gathered to train the ML models.

A. Preprocessing

The first and most important step after data collection is data preprocessing. By eliminating redundant and erroneous data and encoding the raw dataset for phishing detection using the One-Hot Encoding approach, the raw dataset was made ready for the machine learning model.

B. Exploratory Data Analysis

Following a number of data cleaning steps, the dataset was subjected to exploratory data analysis (EDA). The dataset was examined, explored, and summarized using the data visualization technique. To find patterns and insights in data, these visualizations use heat maps, histograms, box plots, scatter plots, and pair plots.

C. Feature Extraction

y extracting new features from the current ones in a dataset, feature extraction seeks to lower the overall number of features in the dataset. As a result, phishing and legitimate datasets were used to extract website content-based features, such as the address bar-based feature, which has 8 features, the domain-based feature, which has 3 features, and the HTML & JavaScript-based feature, which has 4 features. 15 features were thus extracted in total for phishing detection.

Architectural design focuses on understanding how a system should be set up and developing the overall structure of that system. It demonstrates how the system's various parts interact to accomplish its primary goals. It is the procedure for determining the various components that make up a system as well as the framework for sub-system coordination and communication.

The architectural design of the suggested system is shown graphically in the diagram below. When a user enters a URL link, the link passes through several trained machine learning and deep neural network models before the best model with the highest accuracy is chosen. The chosen model is implemented as an API (Application Programming Interface) and then incorporated into a web application. As a result, a user engages with the online application, which is available on many display devices including PCs, tablets, and mobile devices. The use case scenarios for the phishing detection system are shown in Figure 1

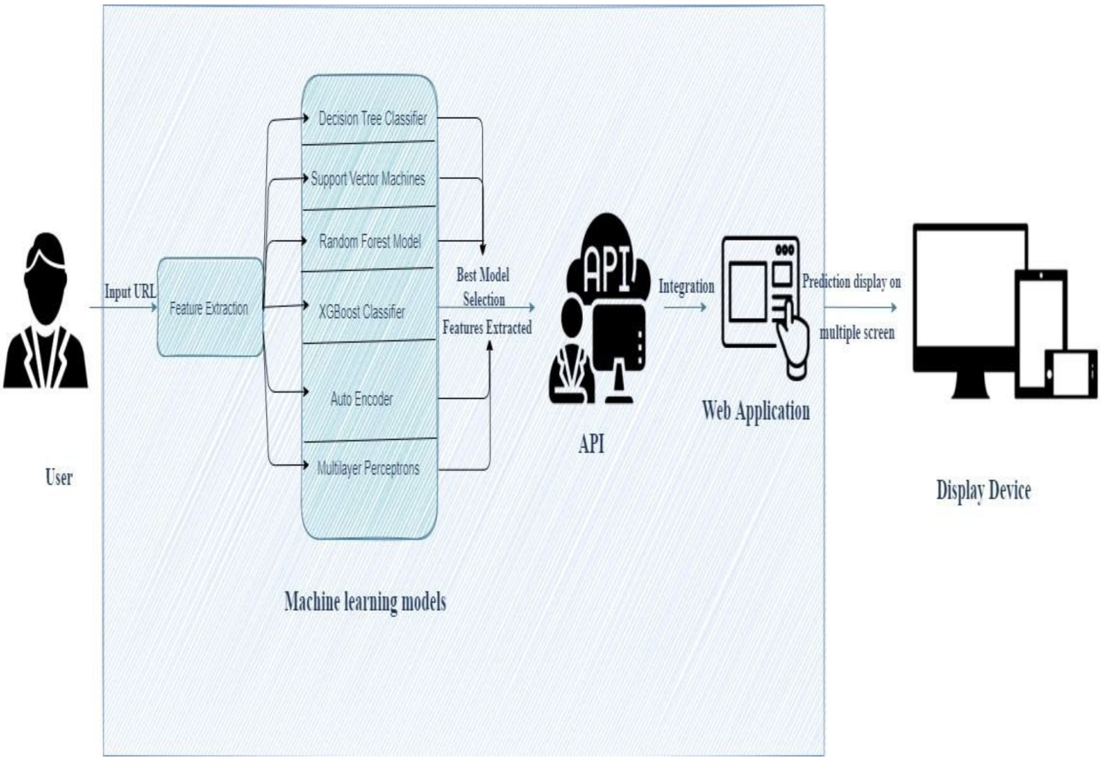


Figure 1: Architectural Design of the Proposed System

Figure 2 shows the functionality of the system as designed from the requirements is described in the use case diagram, which also provides an overview of the system's users. It represents the observable interactions between actors and the developing system as a behavior diagram. Actors, the system, associated use cases, and relationships between them are all included in the use case diagram.

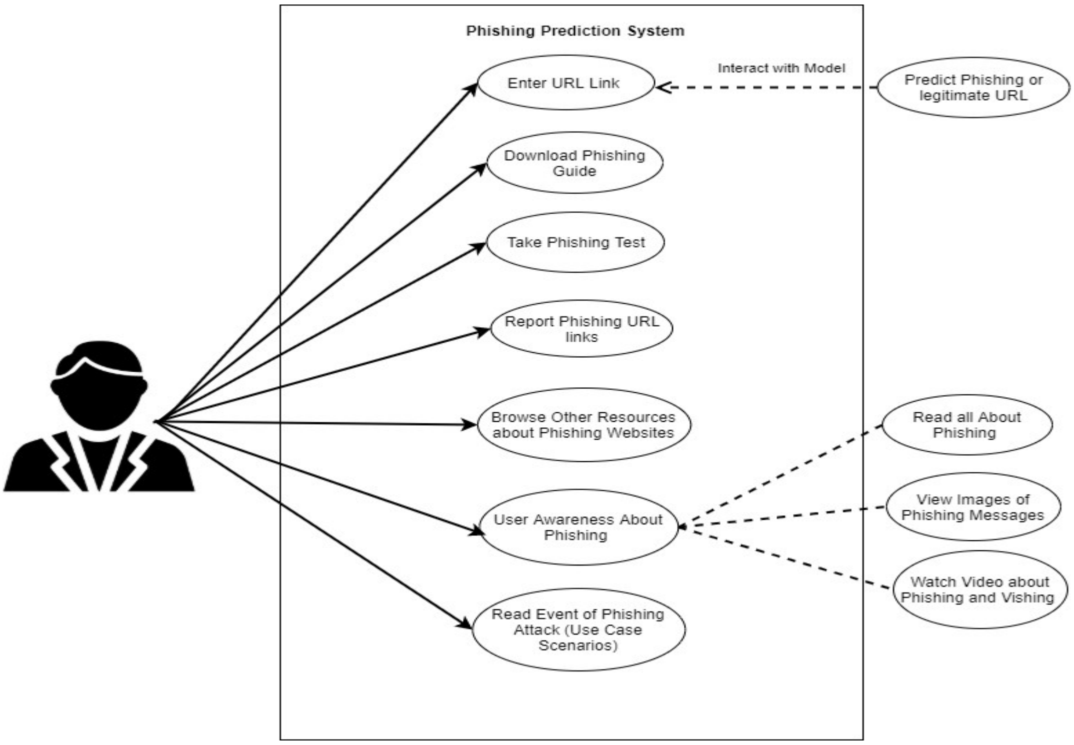

Phishing Prediction System
Enter URL Link
Interact with Model
Predict Phishing or legitimate URL
Download Phishing Guide
Take Phishing Test
Report Phishing URL links
Browse Other Resources about Phishing Websites
User Awareness About Phishing
Read all About Phishing
View Images of Phishing Messages
Watch Video about Phishing and Vishing
Read Event of Phishing Attack (Use Case Scenarios)

Figure 2: Use Case diagram for Proposed System

A flowchart is a diagram that shows how a system, computer algorithm, or process works. It is a graphical depiction of the system's stages to be carried out, listing them in chronological sequence. It is intended to convey complex processes in simple, understandable representations and to show how algorithms run. The machine learning technique used by phishing detection systems is depicted in Figure 3.

The phishing detection web interface system is displayed in Figure 4. When a user enters a URL link, the website analyzes the URL's format and then determines whether the link is legitimate or phishing.

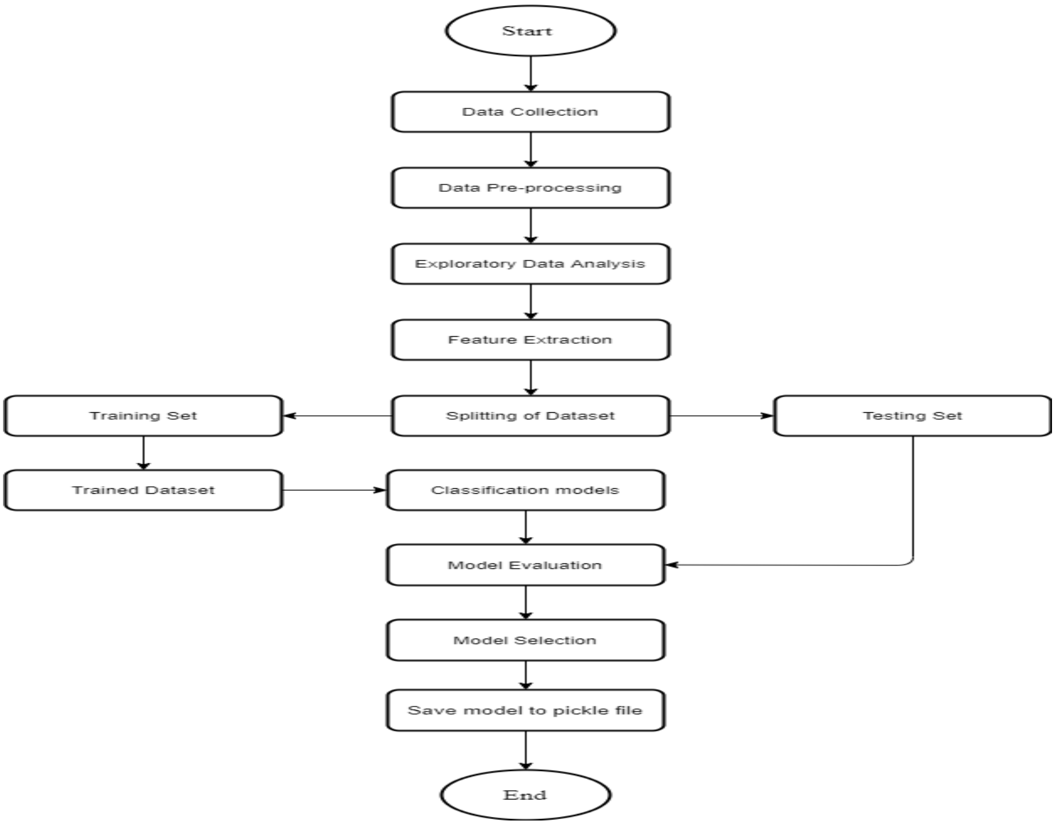


Figure 3: Flowchart of the proposed System

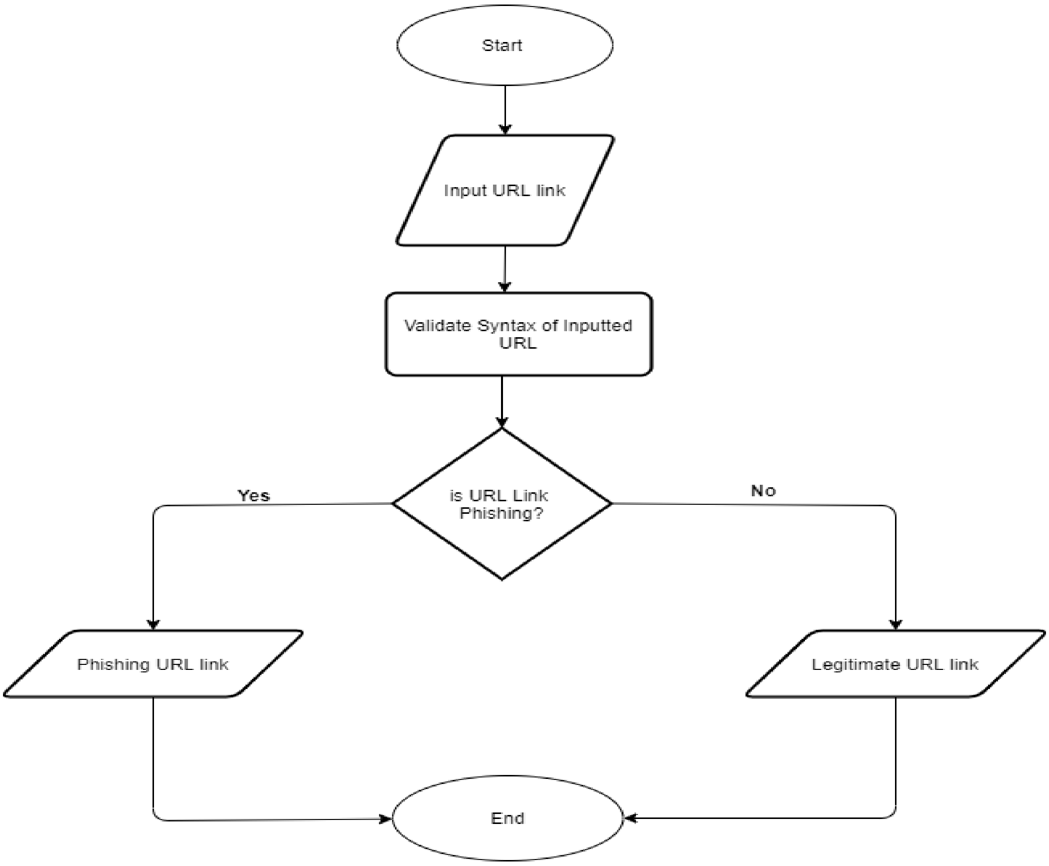


Figure 4: Flowchart of the web interface

4.0 EXPERIMENT AND RESULTS

"PHISH-BOT" is a one-page phishing detection web application can be used with any browser. Python was the only programming language used to create the application. The following pages in figure 5 are part of the phishing detecting website application:

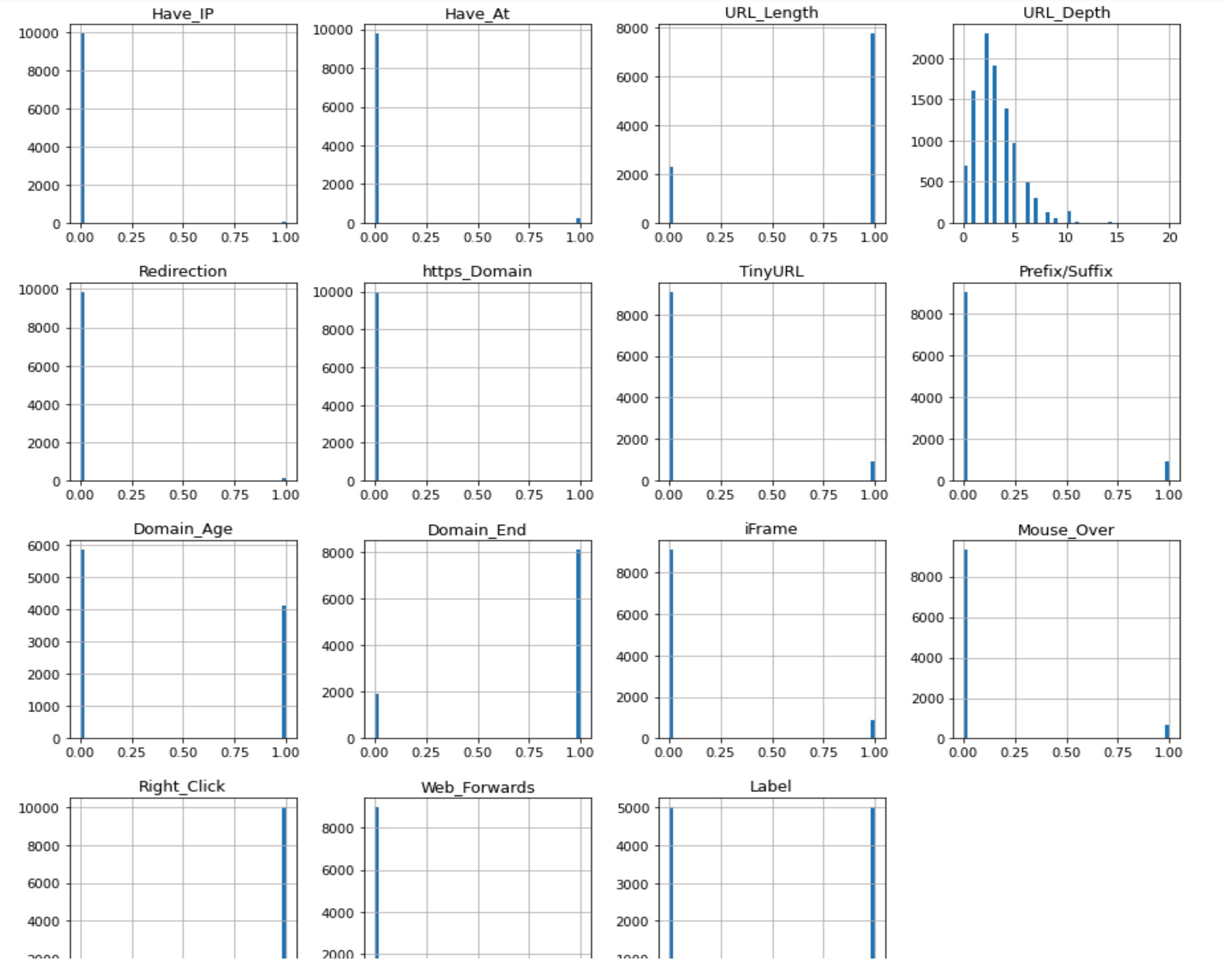


Figure 5: Dataset distribution plot based on the chosen features

On the home page, there is a session where a user can enter a URL and determine whether it is phishing or not. It forecasts the URL's current state. This page's users can use it to verify a URL link and to access a variety of phishing attack materials. To learn how to recognize phishing messages and URLs, the User can explore the resource tab.

The Predict URL page

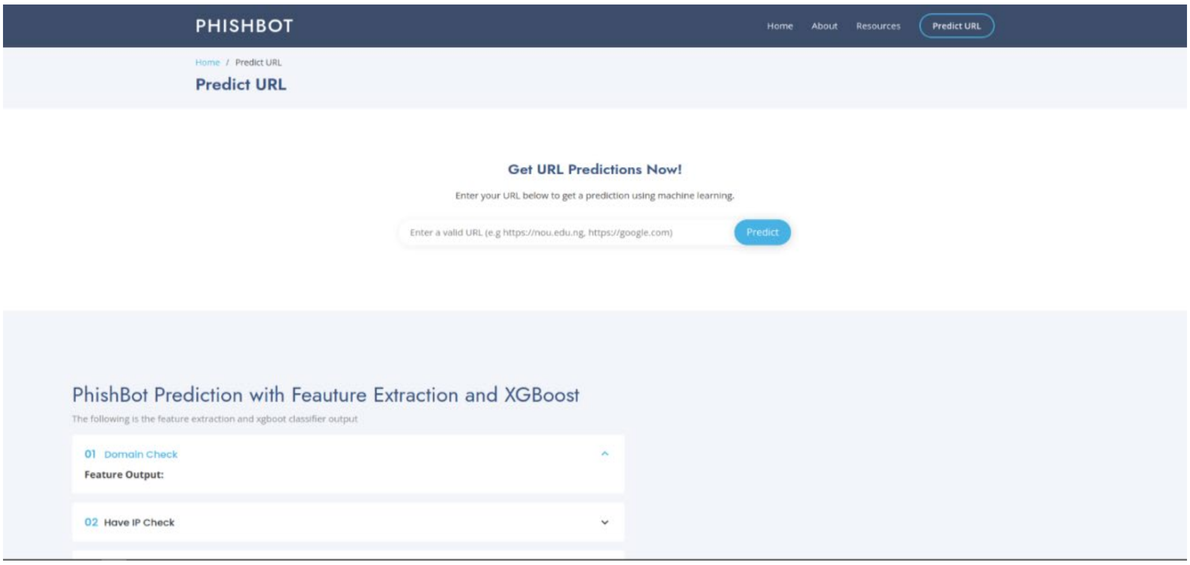
The predict URL page as shown in figure 6(a) and figure 6(b). This is the page users will input the suspicious URL to get the prediction. The output will determine if the URL is legitimate or phishing.

Resource page

It includes many materials on phishing, including explanations of the term, types of phishing attacks, and strategies, as well as references to the sources from which the content was gathered. Additionally, it has two (4) sub-session links: the google safe browsing, google search help, intradyn and jigsaw phish quiz as seen in Figure 7.

Web application Source Code

As seen in figure 8, the web application's source code is divided into pages and is written in python.



Predict URL

Get URL Predictions Now!

Enter your URL below to get a prediction using machine learning.

Enter a valid URL (e.g https://nou.edu.ng, https://google.com) Predict

PhishBot Prediction with Feauture Extraction and XGBoost

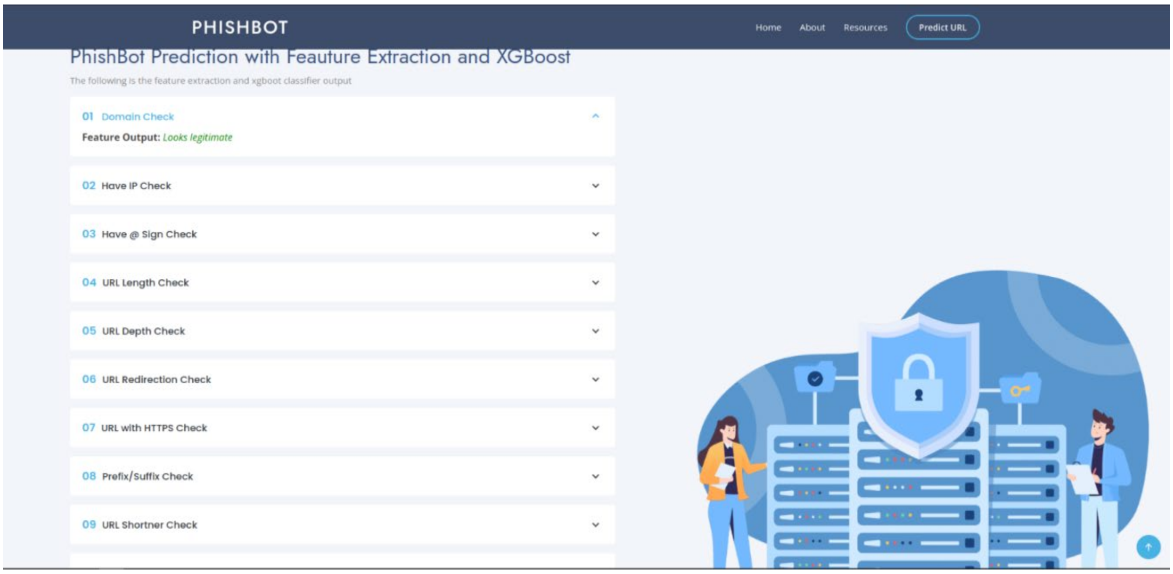
The following is the feature extraction and xgboot classifier output

01 Domain Check

Feature Output:

02 Have IP Check

Figure 6. (a):The Predict URLpage



PhishBot Prediction with Feauture Extraction and XGBoost

The following is the feature extraction and xgboot classifier output

01 Domain Check

Feature Output: *Looks legitimate*

02 Have IP Check

03 Have @ Sign Check

04 URL Length Check

05 URL Depth Check

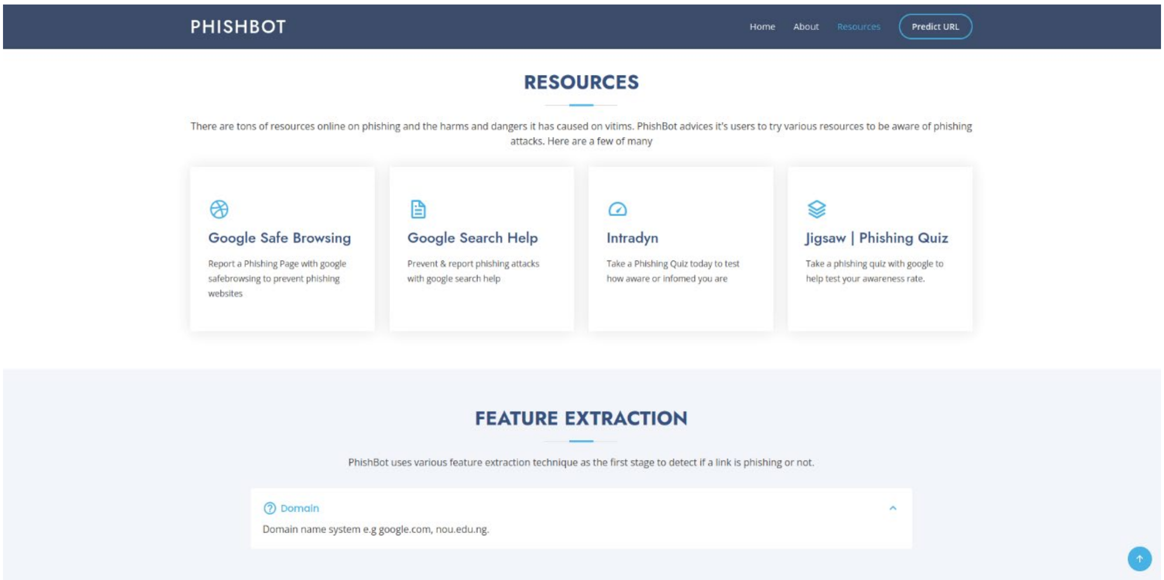
06 URL Redirection Check

07 URL with HTTPS Check

08 Prefix/Suffix Check

09 URL Shortner Check

Figure 6 (b): The Predict URL page



RESOURCES

There are tons of resources online on phishing and the harms and dangers it has caused on vitims. PhishBot advices it's users to try various resources to be aware of phishing attacks. Here are a few of many

Google Safe Browsing

Report a Phishing Page with google safebrowsing to prevent phishing websites

Google Search Help

Prevent & report phishing attacks with google search help

Intradyn

Take a Phishing Quiz today to test how aware or infomed you are

Jigsaw | Phishing Quiz

Take a phishing quiz with google to help test your awareness rate.

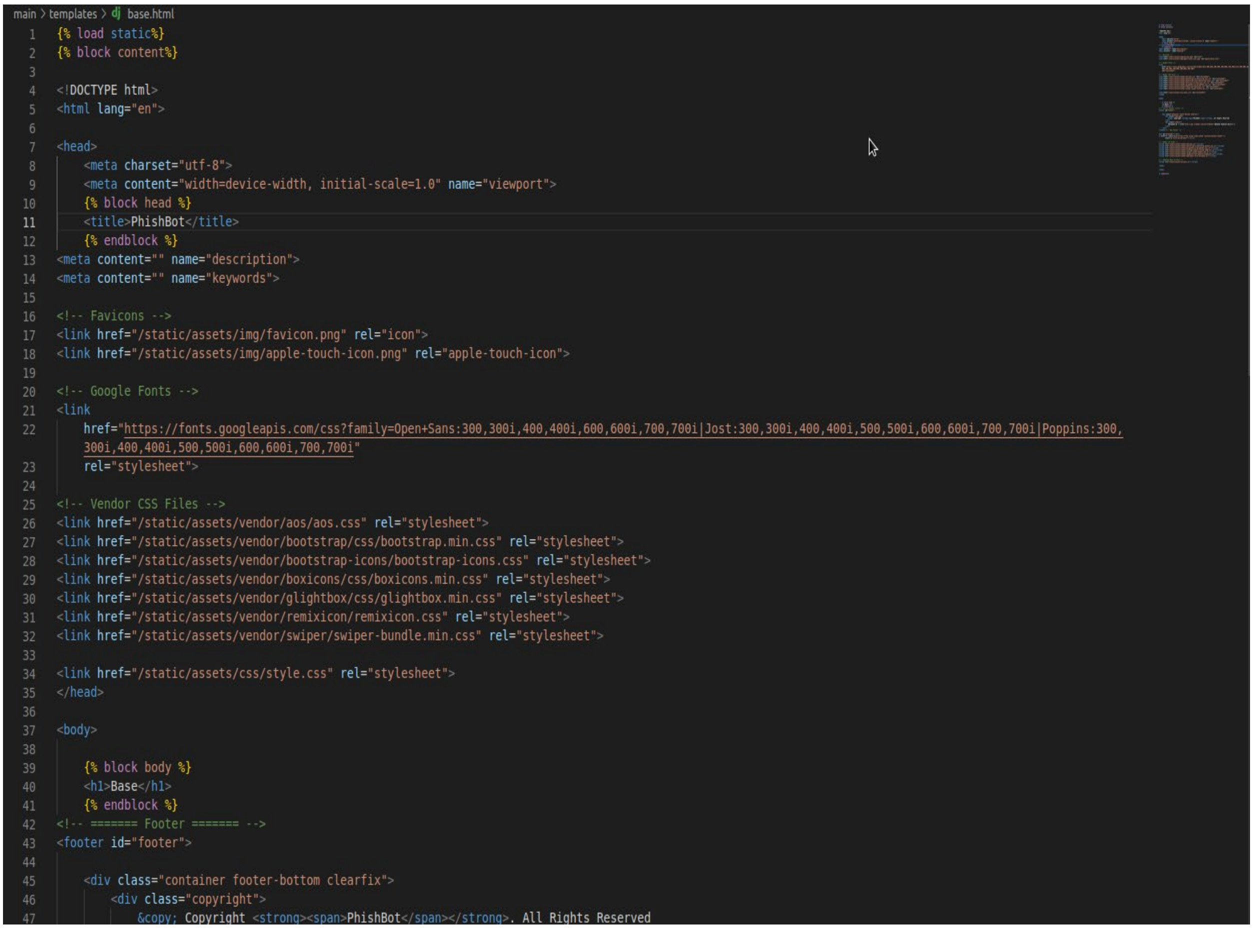
FEATURE EXTRACTION

PhishBot uses various feature extraction technique as the first stage to detect if a link is phishing or not.

Domain

Domain name system e.g google.com, nou.edu.ng.

Figure 7: The Resource page



```
{% load static%}
{% block content%}

<!DOCTYPE html>
<html lang="en">

<head>
    <meta charset="utf-8">
    <meta content="width=device-width, initial-scale=1.0" name="viewport">
    {% block head %}
    <title>PhishBot</title>
    {% endblock %}
<meta content="" name="description">
<meta content="" name="keywords">

<!-- Favicons -->
<link href="/static/assets/img/favicon.png" rel="icon">
<link href="/static/assets/img/apple-touch-icon.png" rel="apple-touch-icon">

<!-- Google Fonts -->
<link
    href="https://fonts.googleapis.com/css?family=Open+Sans:300,300i,400,400i,600,600i,700,700i|Jost:300,300i,400,400i,500,500i,600,600i,700,700i|Poppins:300,
    300i,400,400i,500,500i,600,600i,700,700i"
    rel="stylesheet">

<!-- Vendor CSS Files -->
<link href="/static/assets/vendor/aos/aos.css" rel="stylesheet">
<link href="/static/assets/vendor/bootstrap/css/bootstrap.min.css" rel="stylesheet">
<link href="/static/assets/vendor/bootstrap-icons/bootstrap-icons.css" rel="stylesheet">
<link href="/static/assets/vendor/boxicons/css/boxicons.min.css" rel="stylesheet">
<link href="/static/assets/vendor/glightbox/css/glightbox.min.css" rel="stylesheet">
<link href="/static/assets/vendor/remixicon/remixicon.css" rel="stylesheet">
<link href="/static/assets/vendor/swiper/swiper-bundle.min.css" rel="stylesheet">

<link href="/static/assets/css/style.css" rel="stylesheet">
</head>

<body>

    {% block body %}
    <h1>Base</h1>
    {% endblock %}
<!-- ======= Footer ======= -->
<footer id="footer">

    <div class="container footer-bottom clearfix">
        <div class="copyright">
            &copy; Copyright <strong><span>PhishBot</span></strong>. All Rights Reserved
```

Figure8 :Code for the web application

5.0 Result Discussion

PHISH-BOT" is a one-page phishing detection web application can be used with any browser. Python was the only programming language used to create the application. The pages displayed in figure 5 are part of the phishing detecting website application: On the home page, there is a session where a user can enter a URL and determine whether it is phishing or not. The developed system can forecast the URL's current state. The page's users can use it to verify a URL link and to access a variety of phishing attack materials. To learn how to recognize phishing messages and URLs, the User can explore the resource tab. The predict URL page as shown in figure 6(a) and figure 6(b). This is the page users will input the suspicious URL to get the prediction. The output will determine if the URL is legitimate or phishing.

The new system includes many materials on phishing with the explanations of the term, types of phishing attacks, and strategies, as well as references to the sources from which the content was gathered. additionally, it has two (4) sub-session links: the google safe browsing, google search help, intradyn and jigsaw phish quiz as seen in figure 7. web application source code as seen in figure8, the web application source code is divided into pages.

CONCLUSION

The developed system provides users with access to new and quicker technique to determine if a URL link is real or phishing as well as an instructional material regarding phishing attacks. It uses deep neural network methods and machine learning models to determine whether a URL link is real or phishing. Phishing URLs were specifically identified using feature extraction and models applied to the dataset, which also improved the performance accuracy of the models. It is also remarkably effective at determining whether a URL link is legitimate.

REFERENCES

Abdelhamid, N., Thabtah F., & Abdel-Jaber, H. Phishing detection: A recent intelligent machine learning comparison based on models' content and features," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2017, pp.

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2015). A survey of phishing email filtering techniques, Proceedings of IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2070–2090.

Anjum N. S., Antesar M. S., & Hossain M.A. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA).

Ashritha, J. R., Chaithra, K., Mangala, K., & Deekshitha, S. (2019). A Review Paper on Detection of Phishing Websites using Machine Learning.Proceedings of International

Journal of Engineering Research & Technology (IJERT), 7, 2. Retrieved from www.ijert.

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2016). Fighting against phishing attacks: state of the art and future challenges, Neural Computing and Applications.

Imperva. (2021). Phishing attacks. Retrieved from https://www.imperva.com/learn/application-security/phishing-attack-scam/.

Kartik, M. (2021). Everything You Need to Know About Feature Selection in Machine Learning. Retrieved from https://www.simplilearn.com/tutorials/machine-.

Kiruthiga, R., Akila, D. (2019). Phishing Websites Detection Using Machine Learning. Retrieved from https://www.researchgate.net/publication/337049054 Phishing Websites Detection.

KnowBe4 (2021). Phishing Techniques. Retrieved from https://www.phishing.org/phishing-techniques.

Kondeti, P. S., Konka, R. C., & Kavishree, S. (2021). Phishing Websites Detection using Machine Learning Techniques. International Research Journal of Engineering and Technology, 08(4), Page 1471-1473. Retrieved from https://www.irjet.net/archives/V8/i4/I.

Noel, B. (2016). Support Vector Machines: A Simple Explanation. Retrieved from https://www.kdnuggets.com/2016/07/support-vector-machines-simple-.

Pamela (2021). Phishing attacks. Retrieved fromhttps://www.khanacademy.org/computing/computersandinternet/xcae6f4a7ff01e7d:online-data-security/xcae6f4a7ff015e7d:cyber-attacks/a/phishing-attacks

Peng, T., Harris, I., & Sawa, I. (2018). Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301.

Rami, M. M., Fadi, T., & Lee, M. (2015). Phishing Websites Features. Retrieved from https://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf.

Rishikesh, M., & Irfan, S. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 45. doi:10.5120/ijca2018918026.

Saimadhu, P. (2017). How the random forest algorithm works in machine learning. *Retrieved from https://dataaspirant.com/random-forest-algorithm-machine-*.

Shad, J., & Sharma, S. (2018). A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology, pp. 425–430.

Shaikh, A.N., Shabut, A.M., Hossain, M.A. (2016, December 15-17). A literature review on phishing crime, prevention review, and investigation of gaps. Paper presented at the Tenth International Conference on Software, Knowledge.

Shreya, G. (2020). Phishing website detection by machine learning techniques. Retrieved from https://github.com/shreyagopal/Phishing-Website-Detection-by-.

Sönmez, Y., Tuncer, T., Gökal, H., & Avci, E. (2018). Phishing web sites features classification based on extreme learning machine. 6th Int. Symp. Digit. Forensics Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5.