Library Progress International
Vol.44 No.3, Jul-Dec 2024: P.7211-7224

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

AI-Driven Phishing Email Detection: Leveraging Big Data Analytics for Enhanced Cybersecurity

Sanjay Ramdas Bauskar[1], Chandrakanth Rao Madhavaram[2], Eswar Prasad Galla[3], Janardhana Rao Sunkara[4], Hemanth Kumar Gollangi[5]

[1] Pharmavite LLC Sr. Database Administrator, sanjayramdasbauskar@outlook.com
[2] Microsoft Sr. Technical Support Engineer, Craoma101@outlook.com
[3] Microsoft Sr. Technical Support Engineer, EswarPrasadGalla@outlook.com
[4] Sr. Database Engineer, JanardhanaRaoSunkara@outlook.com
[5] TCS Software Developer, HemanthKumarGollangi12@outlook.com

How to cite this article: Sanjay Ramdas Bauskar, Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Janardhana Rao Sunkara, Hemanth Kumar Gollangi (2024) AI-Driven Phishing Email Detection: Leveraging Big Data Analytics for Enhanced Cybersecurity. *Library Progress International,* 44(3), 7211-7224.

ABSTRACT

Big data analytics and AI are emerging technologies that can help businesses improve their email security. There is a wide range of research that implements big data analytics for email security, and phishing email detection is one dimension of email security. Therefore, the essay emphasizes the use of big data analytics and AI for developing real-time phishing email recognition. Our research demonstrates that a phishing email detection technique utilizing big-data technologies can be used to create a large-scale phishing email dataset, detect phishing emails, visualize additional features, and recognize phishing emails as soon as possible. Chapter 2 outlines present changes in the cybercrime landscape and the current situation of time and defense mechanisms for email security. Then, the concept of harnessing big data analytics and technologies to improve cybersecurity is discussed. In the third segment, an attempt is made to offer a comprehensive list of studies that have been conducted applying big-data analytics to email security and scrutinizing phishing email tactics or technologies for this type of cybercrime. In the final chapter, the implementation of a big-data-based technology utilizing Enron email traffic is highlighted.The essay discusses the application of big data analytics to designing a phishing email detection system in real time. Relevant studies are included in the content as well. Email security has become a chief concern for individuals and organizations. A cybercriminal can victimize anyone after proliferating a phishing email, and millions of phishing emails are distributed to millions of email traffic. With the continual and widespread proliferation of time, cybercriminals are using more sophisticated methods of attacking and have the capability to create more feature-rich phishing emails. This situation necessitates the use of technology to protect us from these methods. The precise recognition of phishing emails reduces their utilization, leading to decreased cybercrimes.

Keywords: AI-driven,Phishing email detection,Big data analytics,Cybersecurity,Machine learning,Email security,Threat detection,Advanced analytics,Cyber threat intelligence,Fraud prevention,Anomaly detection,Data-driven insights,Email filtering,Predictive analytics,Automated response,Neural networks,Behavioral analysis,Data mining,Risk assessment,Security protocols,Malware detection,Natural language processing,Deep learning,Cyber defense strategies,Security analytics.

1. Introduction

This paper aims to highlight the similarities between phishing email detection using machine learning and big data analytics. Thanks to machine learning big-data analytics solutions, a new research avenue is opened to improve cybersecurity based on the latest threats associated with criminal opportunities, terrorist activities, political or personal issues, illicit drugs, actors, or different transactions. The remainder of the paper is organized as follows: Section 2 presents the literature survey. Section 3 provides the early detection of phishing emails using data mining-based machine learning. The big data analytics used in cybersecurity are provided in Section 4. The similarity between the research of big data analytics and the current AI-based phishing email detection process is illustrated in Section 5. Finally, Section 6 concludes the paper.

Detecting and mitigating big data-driven phishing attacks, the first step in implementing a data-driven approach, has been at the forefront of data-driven security research in recent years. This is given that 88% of the phishing emails analyzed in 2019 were equipped with big data processing engines, amounting to 133 billion business and consumer emails per day. Using AI models to prioritize feature selection has been widely hailed as a promising trend in phishing email detection, largely due to researchers' dissatisfaction with the performance of malware detection using engineered features culled solely from a CSV of static file characteristics. Subjective and entirely based on a human's or tool's opinion, these features provide an inaccurate insight into the file's true intentions. The development of increasingly anti-forensic malware has rendered these features critical and impractical in malware detection.For quite some time, the effectiveness of traditional threat mitigation techniques has diminished in contrast to their increasing volume and sophistication, rendering organizations and enterprises equally vulnerable to cyber fraud and thefts. Leveraging big data analytics to detect and mitigate security breaches can contribute to improved cybersecurity. With cybersecurity professionals working around the clock to design techniques for rapidly detecting, classifying, and predicting unseen threats previously believed to be improbable or undiscovered, the field of "data and cybersecurity" has grown broad and promising.Phishing email detection has evolved significantly with the integration of machine learning and big data analytics, marking a pivotal advancement in cybersecurity. As phishing attacks have become more sophisticated and pervasive, analyzing vast amounts of data has become essential for effective threat detection. In 2019, a staggering 88% of phishing emails were processed through big data engines, underscoring the magnitude of the challenge faced. Traditional methods of malware detection, reliant on static features and human-curated data, have proven inadequate against the evolving landscape of cyber threats. The rise of anti-forensic malware has further exposed the limitations of these traditional approaches, emphasizing the need for dynamic and data-driven solutions. Leveraging AI models for feature selection within big data frameworks offers a promising alternative, enabling more accurate and timely identification of phishing attempts. As cybersecurity experts continue to develop innovative techniques to counteract emerging threats, the synergy between machine learning, big data analytics, and cybersecurity is paving the way for more robust defenses against an increasingly complex array of cyber risks.Phishing email detection has undergone significant advancements through the integration of machine learning and big data analytics, crucial for enhancing cybersecurity in an era of increasingly sophisticated cyber threats. In 2019, an alarming 88% of phishing emails were processed using big data engines, highlighting the scale of this challenge. Traditional malware detection methods, which rely on static features and human-curated data, have proven inadequate against evolving tactics, particularly with the rise of anti-forensic malware that circumvents these traditional approaches. In response, leveraging AI models for feature selection within big data frameworks has emerged as a promising strategy, enabling more accurate and timely identification of phishing attempts. This dynamic, data-driven approach empowers cybersecurity professionals to rapidly detect, classify, and predict previously unimagined threats, marking a pivotal shift in the fight against cyber fraud and theft. The synergy between machine learning, big data analytics, and cybersecurity is thus reshaping defenses against a complex and ever-evolving landscape of cyber risks.Phishing email detection has significantly evolved through the integration of machine learning and big data analytics, becoming essential in the fight against increasingly sophisticated cyber threats. By 2019, a striking 88% of phishing emails were processed through big data engines, underscoring the scale of the challenge organizations face. Traditional malware detection methods, which often depend on static features and manually curated datasets, have struggled to keep pace, especially with the emergence of anti-forensic malware that cleverly bypasses these approaches. In this context, employing AI models for feature selection within big data frameworks has proven to be a promising solution, facilitating more accurate and timely detection of phishing attempts. This proactive, data-driven methodology enables cybersecurity professionals to swiftly identify, classify, and predict threats that were once deemed improbable, marking a transformative shift in combating cyber fraud. Ultimately, the collaboration between machine learning, big data analytics, and cybersecurity is redefining how defenses are structured against a constantly evolving array of cyber risks.

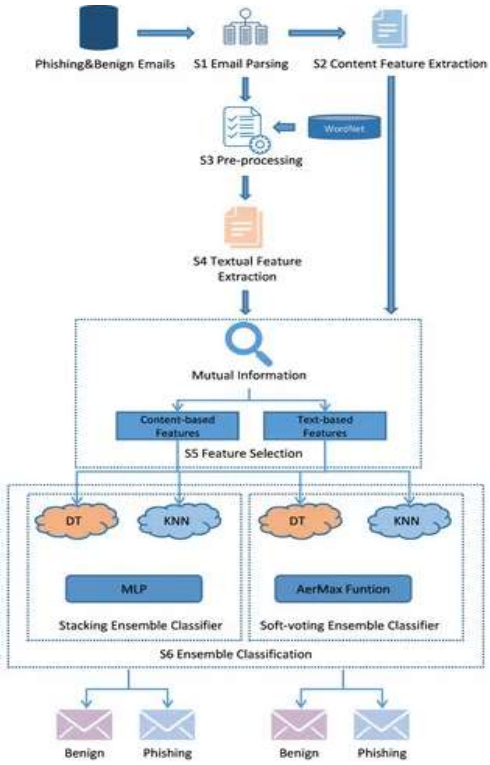

Fig 1 : Enhancing Phishing Email Detection through Ensemble Learning

1.1. Background and Significance

This implantation route motivated us to put forth research to examine and leverage the power of enormous data to provide dual-stage AI-based phishing email detection in the cloud. By building awareness, attitudes, and trust in a new cadre of business practitioners, we aim to provide a secure IAM-friendly, hand-off, SaaS-based proxy. AI-driven, with few false negatives, this solution minimizes response time, dismisses security threats, and resurrects quickly with the most recent intelligence. Phishing refers to a multi-vector attack on enticement to deceive and exploit computer users. Evaluating two main components, wherein the average clicking percentage for phishing attacks was 24.8%, with 18.2% clicking time for the normal campaign. Phishing is the number 1 cyberattack vector leading to cyber exploitation. A comprehensive solution is the need of the hour, not only to mitigate its spread but also to uncover its depths and groves. That is the growth conceptualized here.

Phishing entered the world of computing as early as the 1970s. In 1972, the first worm known as "Creeper Worm" showcased several concepts of subsequent phishing attacks. Since 2018, phishing attacks alone have compromised 70% of smartphone users in the UK, targeting net banking with malware. According to Symantec's Internet Security Threat Report (ISTR) 2021, there was a 67% increase in phishing activity in 2020 compared to 2019. In the first five months of 2021, there were about 1.6 million phishing cases making use of wildcard representations of trusted domains to evade detection. This figure is expected to grow every coming year. Using domain names that refer to a trusted brand, attackers can deceive the victim into giving up their credentials or infecting their devices with malware via email. The catastrophes were caused by simply clicking on a malicious link or attachment or responding to emails. Only about 3% of consumers can distinguish typical scam emails from genuine emails, according to a test involving 2024 consumers.

1.2. Research Objective

Research, therefore, aims to develop a machine learning model that uses big data analytics to assess the overall behavior of the system and demonstrate the proposed approach to using deep learning for AI-driven phishing email detection. A discussion of the existing literature for email phishing detection would further strengthen the rationale for developing our machine learning models. As such, the outcomes of research should include well-documented inferences to a wide range of knowledge users, especially those who are the main stakeholders in developing a cybersecurity system.The objective of this essay is to first evaluate the present state of phishing detection systems and to propose that the integration of big data in deep learning neural networks can improve the performance of an AI model. Specifically, the research identifies this opportunity in the detection of phishing emails to advance towards a better cybersecurity environment. The proposed model will not rely on a human-crafted dataset: it will learn directly from a large amount of raw input data and make decisions to detect suspicious

and malicious content in the emails. While both deep learning techniques and big data methods have previously achieved widespread success across various fields, the aim is to test the effectiveness of combining such techniques for application in cybersecurity. The research objectives have been described below.

Equ 1: Feature Extraction and Representation

1. Feature Extraction

Phishing detection typically starts with extracting features from emails, which may include text content, metadata, and behavioral patterns. Common features include:

- **Textual Features:** Frequency of certain keywords, email structure, etc.
- **Metadata:** Email sender, domain, headers, etc.
- **Behavioral Patterns:** Click-through rates, response patterns, etc.

2. Phishing Attacks and Email Security

Email has increasingly become the primary communication channel at work, but it is plagued by numerous security issues. Email security has become a major concern that is linked to cybercrime in small and medium enterprises, as email is identified as a key threat vector. The email security threat can eventually compromise the network as a result of phishing attacks. Technologies have been developed to avoid human-related problems, but these mitigations can never completely eliminate harmful messages. Furthermore, sending an email with proper attachments and with a document for the users to click on is not treated as a bad practice in an enterprise organization where developers carry out software updates via emails for users. Cybersecurity incidents are evolving and they are more sophisticated. The more the aspect of deception is built into these attacks, the more the email filtering technologies will fail, making it easier for an adversary to achieve their goal. The challenge of unwanted emails has been a long-standing problem for the security community because unwanted emails that disseminate quickly are an efficient way to carry out spam attacks.Several cyber-attacks on individuals and organizations begin with a phishing email. Phishing involves tricking people into revealing sensitive information such as credentials, personal information, and company information. As the target of a phishing attack, you might not realize you were attacked at all. This type of cybercrime often involves collecting the personal information of targets hoping to perform a highly targeted attack. These attacks, known as spear phishing attacks, are sent only to specific individuals who have valuable information that the attacker wants. Spear phishing has multiple subtypes such as whaling, vishing, and quid pro quo. There is no clear definition for any of these subtypes, but they are similar to standard spear phishing except that they use a different form of communication for the social engineering stage. Researchers have also noted differences in intent: while spear phishing targets individuals, whaling goes after giant institutions.

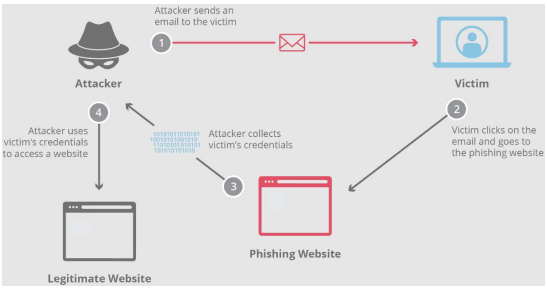

Fig 2 : Phishing attack

2.1. Definition and Types of Phishing Attacks

Phishing is a social engineering attack in which a cybercriminal poses as a trustworthy entity or leverages impersonation tools to deliver fraudulent communications (typically emails) in an attempt to trick individuals into providing sensitive information such as credit card details, bank account numbers, social security numbers, login credentials, and the like. The primary objective of a phishing attack is for the attacker to unrightfully gain access to restricted digital assets that can be monetized in various ways or used to launch more sinister attacks like ransomware or business email compromise. Historically, a phishing email informed the recipient that they had won a lottery and needed to claim their reward, hence prompting the user to share their personal details like postal address, date of birth, or to pay a fee. However, with the evolution of technology and the improvement in cyber threat-awareness training, phishing strategies have become more sophisticated and evolved. Current phishing strategies take advantage of social engineering tactics to impersonate companies and services that the recipient of the email is likely to be familiar with and open communications from. Some examples of phishing tactics and

types are detailed in the literature.In its standalone state, pure phishing is an email attack in which cybercriminals leverage psychological manipulation to trick email recipients into sharing sensitive information. However, the following is a list of major phishing tactics that cybercriminals typically execute:
- Pharming attacks: Redirect users to fake sites by corrupting the DNS servers of popular websites or default DNS servers that redirect requests to hostbypositive.com. - Spear Phishing: A more targeted and specific phishing attack in which an email recipient is selected and convinced that they are receiving information from a trusted source. Green Card offers, often purporting to be from the United States Citizenship and Immigration Services (USCIS). Links to such emails often redirect users to malicious sites from where spyware, trojans, rootkits, and other kinds of malware can be downloaded onto their systems. The goal of spear phishing emails is to induce victims to click on a link, open an attachment, or share sensitive data.

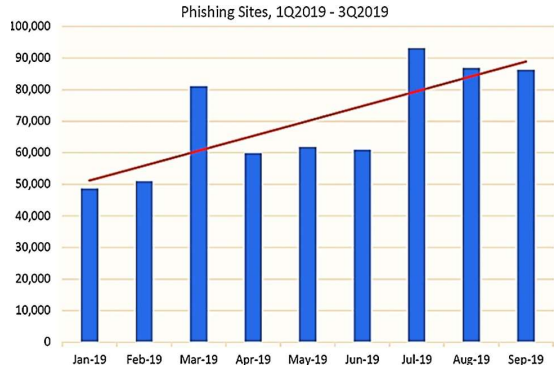

Fig : A comprehensive survey of AI-enabled phishing attacks detection techniques

2.2. Challenges in Email Security

Although some phishing emails exhibit anomalies in their text such as spelling mistakes, visual discrepancies, mismatched URLs, or use of salient brand names, it is becoming increasingly difficult to rely on text-based systems to detect phishing since attackers are finding more subtle methods by which to trigger these systems.
While the new technological advancements and current sophisticated systems have greatly impacted the automation of the cyber threats detection process, there are still many challenges in classifying phishing from legitimate email. Along with the increasingly sophisticated systems, there are email threats from phishing, spoofing, and spear phishing. Computers are slightly better than Chance at distinguishing threatening social-engineering email text between real and synthetic.To effectively classify phishing attacks, both natural language processing and domain-specific features should be employed while designing the system. Phishing by email is ages old, but it still works in the current internet and corporate environments. Phishing can either be used to exfiltrate sensitive data, spread malware, or cause denial of service by abusing server access. Hundreds of millions of user accounts are compromised per year due to phishing, causing billions of dollars in financial damage.
Blockchain offers a new possible approach to phishing email detection. Whereas the previously mentioned research focuses on the actual content of the email, blockchain email signatures take into account the source and domain of the sender by validating that these originate within the correct chain that is distributed and not fraudulent. This anti-phishing method is particularly powerful in regard to email.

3. AI in Cybersecurity

Artificial intelligence (AI) promises to aid a variety of industries, and one of the most notable areas for AI implementation in security systems is cybersecurity. Machine learning and other AI-based tools are critical in both identifying potential threats and preventing future security breaches. AI algorithms are fine-tuned to leverage a multitude of historical (structured and unstructured) data to develop patterns and understandings of how intruders and spammers/phishers function across a wide-ranging number of attack types and contexts.
These capabilities offer businesses a way to uniquely root out security threats. Additionally, machine learning tools are applicable in creating more advanced understandings of language, and how to successfully identify various types of attacks. Natural language processing tools that humanize the data can aid in improving cybersecurity and catching multiple threats from a linguistic perspective. From authenticating user identity to preventing potential fraud, AI/ML offers a multitude of dynamic cyber protection mechanisms that give corporations flexibility and intelligence when it comes to cybersecurity, stirring up a lot of demand for jobs in robotics, cybersecurity, and associated sectors.Training an AI engine to spot known phishing emails is crucial. These AI models need to encapsulate the subtle linguistic differences in order to establish the difference between junk emails and phishing attempts. The further sophistication of AI models helps strengthen the defenses of

corporations against a myriad of network attacks.Artificial intelligence (AI) is revolutionizing cybersecurity by offering advanced tools and methodologies to identify and mitigate potential threats. Leveraging machine learning algorithms, AI systems analyze vast amounts of historical data—both structured and unstructured—to recognize patterns associated with various types of cyberattacks, including phishing, spamming, and intrusions. These algorithms are fine-tuned to detect subtle linguistic nuances in phishing emails, distinguishing them from legitimate communications and enhancing the accuracy of threat detection. Furthermore, natural language processing (NLP) tools contribute to this effort by humanizing and contextualizing data, which improves the identification of threats from a linguistic perspective. As AI continues to evolve, its applications in authenticating user identities and preventing fraud are becoming increasingly sophisticated, making it a critical component in modern cybersecurity strategies. This technological advancement is driving significant demand for expertise in robotics, cybersecurity, and related fields, highlighting the growing importance of AI in safeguarding digital environments.

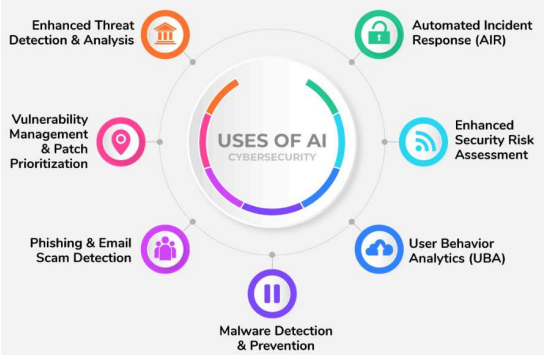

Fig 3 : AI in Cybersecurity

3.1. Applications of AI in Cybersecurity

AI is released in the various dimensions of the information technology sector that positively affect society by introducing basic methods, enhancing and adjusting suitable methods in response to interference or disruption resulting from unavoidable security breaches. Five such ways, to begin with, are raising the degree of biodiversity in the system, introducing diversity and redundancy, data sharing, data integrity, and encouraging encrypted communications. However, two common security breaches that severely affect the QoE in cyber-physical systems are largely eradicated using artificial intelligence (AI) deployed on the substantial volume of big data that bombards and deluges the system networks. To begin with, a substantial increase in risks attached to phishing emails necessitates the use of AI to enhance and decipher sophisticated email files and databases in order to increase overall security.

The bulk of online assaults, security breaches, and hacking have become increasingly direct, focused, and concentrated on individuals and businesses in recent years. A sizable collection of incidents on a limited budget, few personnel, and sponsored university students in the dark web have been capable of creating ransomware. Cybersecurity uses a big set of digital conditions and policies to safeguard data and systems from dynamic cyber threats, consisting of three perspectives such as confidentiality, the integrity of the data, and the availability of it. To begin with, having access to key resources (for example, networks and websites) and forums after they have been compromised. UserRepository of usernames and passwords is the public directive of judgment and permission management in cybersecurity.

Equ 2: Machine Learning Models

2. **Machine Learning Models**

a. **Logistic Regression**

Logistic regression is used to classify emails as phishing or non-phishing based on features.

- Logistic Function:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

where $P(Y=1|X)$ is the probability of an email being phishing given features X, and β are the model parameters.

b. **Naive Bayes**

Naive Bayes classifiers use Bayes' theorem with an assumption of feature independence:

- Bayes' Theorem:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

where C represents the class (phishing or non-phishing), X represents the feature vector, and $P(X|C)$ is typically computed using a multinomial distribution.

c. **Support Vector Machines (SVM)**

SVM finds the optimal hyperplane to separate classes in high-dimensional space.

- Decision Function:

$$f(x) = \text{sign}(w^T x + b)$$

where w is the weight vector, x is the feature vector, and b is the bias term.

4. Big Data Analytics in Cybersecurity

Big data analytics customarily used different tools to manage a massive collection of data as per Gartner's agenda publication. These tools are popular for data storage of security operations. The adoption of Hadoop, NoSQL, and MapReduce by private and government organizations is due to the characteristics of varieties, veracity, and volume with ample velocity adaptation of structured or unstructured data. The paper consists of important pioneer use cases of big data analytics in cybersecurity, for instance: monitoring the network, risk management, fraud detection, and development of vulnerability that can be detected by public security operations (SOCs). A combination of one or more detection methods can be used to improve the accuracy of the detection process, there is no use of a single method. The work focuses on several types of active defense cyber-threats, including using dynamic malware analysis, Internet blacklisting and whitelisting, phishing websites detection, APT detection (short report), and malware-c2 traffic detection. Big data and AI-based frameworks are used to tackle the upcoming challenges.Cybersecurity is the guarantor of privacy and confidentiality of the data from unauthorized access, fast attacks, and intelligence congregate disparate sources, like massive IoT devices, social platforms, cloud databases, and end-to-end networks. Big data analytics is a vital cog in cybersecurity, which addresses the challenges of handling massive volumes peculiarly due to the characteristic of phishing emails to elude traditional methods of detection. ML with its self-sufficient AI—machine learning behavior requires the invention of a new mode of phishing email detection. Big Data analytics enables cybersecurity to handle the daunting challenge of detecting and protecting garbled data. These data are either structured or unstructured and encompass computer network visual information (packet payloads), social media networks, and external threats.Big data analytics plays a crucial role in cybersecurity by leveraging tools like Hadoop, NoSQL, and MapReduce to manage and analyze vast amounts of data generated from various sources such as IoT devices, social platforms, and cloud databases. As organizations face challenges related to the variety, veracity, volume, and velocity of data, these tools are essential for addressing complex security issues. Key use cases include monitoring network activity, risk management, fraud detection, and vulnerability assessment. By combining multiple detection methods—such as dynamic malware analysis, Internet blacklisting and whitelisting, phishing website detection, APT detection, and malware-command and control traffic analysis—organizations can enhance the accuracy of threat detection. The integration of big data and AI frameworks is pivotal in tackling emerging threats, particularly in the realm of phishing emails that evade traditional detection methods. Machine learning, with its advanced AI capabilities, is continually evolving to create new techniques for identifying and mitigating phishing attempts, thereby bolstering the protection of both structured and unstructured data against unauthorized access and sophisticated attacks.

Fig 4 : Data Analytics in Combating Cybercrime

4.1. Role of Big Data Analytics in Cybersecurity

Derived from the domains of big data and information technology, cybersecurity is a major field of knowledge and application that represents the practice of protecting information, networks, systems, and data from security breaches, data loss, and other potential forms of theft that can be used to damage an institution or a country. Big data analytics in the current era offers the most efficient performance for structuring, handling, and mining big data, which are considered object-oriented data generated and compiled on a large scale. Big data analytics can predict, sense, and respond to the huge amount of network and security data, identify potential threats to the company, and help make risk decisions.The incorporation of big data in cybersecurity has made it possible to identify anomalies, such as missing data, sluggish reporting, timely statistics, new data utilized statically, and ad-hoc data interpreted which further enhances the detection and prevention process built on viruses and spyware. The acquisition of big data in cybersecurity resolves issues concerning detection, redundancy in response, verification, and sharing of threats. The adoption maximizes the value of data, applies new security solutions in order to quickly detect sophisticated threats, and secures context-aware defense strategies that can logically sense the setting of a threat. Big data analytics assists organizations and businesses to depend on the combination of sophisticated tools, technologies, and human technological competence to sustain their security posture and reduce threats. Current trends of big data analytics on cybersecurity and conducting existing surveys, software tools, methods, and algorithms are discussed.Cybersecurity has evolved significantly with the integration of big data analytics, revolutionizing how organizations protect their information, networks, and systems. data to identify potential threats and vulnerabilities. This integration enables the detection of anomalies such as Big data analytics helps in resolving issues related to threat detection, redundancy in responses, and verification, by providing a more comprehensive view of security incidents. It facilitates context-aware defense strategies and the rapid identification of threats through advanced tools and algorithms. As a result, organizations can enhance their security posture, make informed risk decisions, and apply innovative solutions to stay ahead of emerging threats, making big data analytics a crucial component in modern cybersecurity strategies.

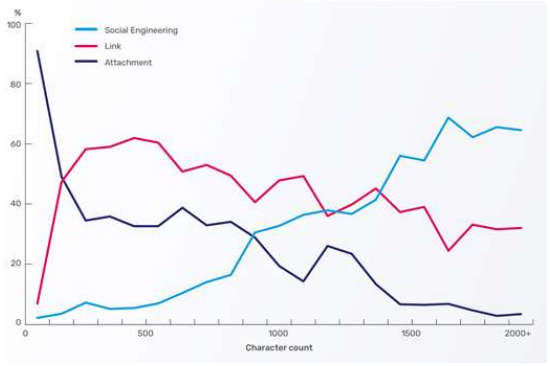

Fig : AI-Generated Phishing Emails Almost Impossible to Detect, Report Finds

5. Integration of AI and Big Data Analytics for Phishing Email Detection

The integration of AI and big data analytics can lead to a dramatic increase in detection performance, thereby saving the organization from various attacks and loss of revenue. The extraction and integration of multiple types and structuring of data then feeding them to the AI system may change the dimension of threat identification and mitigation. For instance, AI systems can work with data from different sources, such as firewall logs, IP addresses, features derived from captured traffic, action logs, etc., to identify any abnormal behavior taking place by the inside or outside threat agents.This results in more effective and accurate threat detection. For email attacks, more

credible and legitimate email-based attacks from crowd-based junk are possible and can be mitigated. To that extent, utilizing AI in conjunction with big data for phishing email detection could be a step in the right direction that has not been given attention. There have been separate studies looking at the use of AI in phishing email detection. However, the study that combines AI with big data for phishing email detection is very limited. Thus, in line with the research gap and background given, this paper introduces the investigation of the landscape of phishing email detection. It aims to show that integrating big data analytics with AI stands to have a significant impact on the identification and mitigation of phishing emails.The integration of AI and big data analytics holds the potential to significantly enhance threat detection and response, safeguarding organizations from various forms of cyberattacks and preventing financial losses. By extracting and integrating diverse types of data—such as firewall logs, IP addresses, traffic features, and action logs—AI systems can more effectively identify abnormal behaviors and potential threats from both internal and external sources. This approach not only improves the accuracy of threat detection but also addresses challenges such as distinguishing between legitimate email-based attacks and spam. Despite existing research on AI's role in phishing email detection, there is a notable lack of studies combining AI with big data analytics for this purpose. This paper seeks to bridge this gap by exploring how integrating these technologies can enhance the identification and mitigation of phishing emails, demonstrating that such an approach could significantly improve cybersecurity measures and protect against sophisticated email threats.Integrating AI with big data analytics presents a transformative opportunity for enhancing cybersecurity, particularly in the realm of phishing email detection. By harnessing diverse data sources—such as firewall logs, IP addresses, captured traffic features, and action logs—AI systems can achieve a more nuanced and accurate detection of abnormal behaviors indicative of phishing attempts. This integration allows for a more sophisticated analysis, distinguishing between genuine threats and benign activities with greater precision. Although AI's effectiveness in phishing detection has been studied, the combination of AI with big data analytics remains underexplored. This paper aims to address this gap by examining how the synergy of these technologies can improve the identification and mitigation of phishing emails, ultimately bolstering an organization's defenses against increasingly sophisticated cyber threats and preventing potential financial losses.

Equ 3: Natural Language Processing (NLP)

a. **Term Frequency-Inverse Document Frequency (TF-IDF)**

TF-IDF is used to evaluate the importance of words in an email.

- TF-IDF Calculation:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

where:

- Term Frequency (TF):

$$\text{TF}(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

- Inverse Document Frequency (IDF):

$$\text{IDF}(t) = \log\left(\frac{N}{\text{Number of documents containing term } t}\right)$$

where N is the total number of documents.

b. **Word Embeddings**

Word embeddings (e.g., Word2Vec, GloVe) convert words into dense vectors:

- Word2Vec Embedding:

$$\mathbf{v}_w = \mathbf{W}_w \mathbf{h}$$

where $\mathbf{v}_w$ is the word vector, $\mathbf{W}_w$ is the weight matrix for the word, and $\mathbf{h}$ is the hidden layer representation.

5.1. Techniques and Algorithms

The algorithms introduced are artificial intelligence (AI) and data analytics-based platform algorithms capable of integrating multiple and heterogeneous algorithms. This section includes three main algorithms along with the solution approaches and a description of each technique. Firstly, the two-step artificial intelligence (AI)-based technique for online phishing email optics (OPE) called "ABC-PEM-HPSO-RFO algorithm" with the help of high-performance parallel swarm operators (HPSO) including ridge filter operator (RFO) as well as tracking of human phishing susceptibility score and cerebral biometric mechanisms such as mismatch response and neural response, to identify and filter out the highest priority OPE in the received email box (i.e. of email box participating in email exchange protocols) for automatic and proactive phishing attack prevention. In the second algorithmic

layer, detection of the remainder OPE (RMOOP) for incidents of UPA in emails are either executed by exact and fuzzy searching for exploiting the s-Symbols, i-Symbols, and comparison Symbols (i.e. Phrase-Symbols) of potty taxonomies.Phishing email is one of the persistent attacks on the IT infrastructures of different domains. Attackers persistently lure users by employing various kinds of social engineering concepts such as urgency, curiosity, intimidation, and fear to persuade the user to open the mail and click on the embedded malicious URL. Phishing emails are usually designed using obfuscated embedded links in emails, embedded links in fake documents, and criminals using instant messaging apps. The anatomical piece of phishing emails can be the embedded link, which can be used for launching further coordinated security attacks.

With two-step AI-based detection and prevention algorithms, we have developed an AI- and big data-based architecture for detecting and preventing ubiquitous phishing email abnormalities (UPA). Techniques and algorithms utilized and integrated to develop the proposed AI-driven big data architecture include techniques for anomaly detection, k-means clustering, string matching (exact and fuzzy), big velocity, big volume, big variety, online and offline learning, and automated attack feedback as adaptive machine learning features.

Equation 4: Anomaly Detection

Anomaly detection methods identify unusual patterns that may indicate phishing.

a. **Z-Score**

The Z-score measures the number of standard deviations a data point is from the mean.

- Z-Score Calculation:

$$Z = \frac{X - \mu}{\sigma}$$

where X is the feature value, μ is the mean of the feature, and σ is the standard deviation.

b. **Isolation Forest**

Isolation Forest isolates observations by randomly selecting features and splitting values:

- Anomaly Score:

$$\text{Score}(x) = 2^{-\frac{E(x)}{c(n)}}$$

where $E(x)$ is the average path length for point x, and $c(n)$ is a constant based on the number of observations.

5.2. Benefits and Limitations

Despite the summarized challenges, there are also a number of significant advantages to integrating AI and big data analytics in the context of phishing email detection. First of all, the very limited amount of humans required in conjunction with AI for a quick and effective response to phishing email campaigns makes this approach an optimal one in case of increasingly growing numbers of new detection-worthy email instances. Furthermore, if big data becomes available in the form required, the combination of AI and big data analytics can lead to higher detection rates compared to a system using only AI, as experiments in this work show in the example of hyperparameter optimized XGBoost.Using AI cannot be taken as a panacea, however. In many cases, it is unfortunately all too easy to overcome the majority of trained systems. One example here is adversarial attacks where the attacker crafts targeted changes to inputs that are still similar to human observers, but previously trained classifiers produce completely different output. Additionally, the question of which and how much data is big data needs to be addressed as well. Furthermore, accessing raw large volumes of mail server contagions and talking to clients of companies is neither accessible nor ethically unproblematic in many cases. Moreover, building on that, the creation of a large trustworthy labeled dataset and ensuring data quality is expensive, time-consuming, and may prove challenging in some companies. Therefore, not all mentioned benefits might be realistically achievable.Integrating AI and big data analytics into phishing email detection offers notable advantages, including the reduction in human intervention required for swift and effective responses, especially given the increasing volume of phishing attempts. This integration can enhance detection rates, as evidenced by experiments using hyperparameter optimized XGBoost, which demonstrate improved performance over systems relying solely on AI. However, AI alone is not a cure-all; challenges such as adversarial attacks, where attackers subtly manipulate inputs to deceive classifiers, highlight the limitations of current AI models. Additionally, defining and accessing 'big data' presents its own set of problems, including ethical concerns, the difficulty of obtaining and managing large volumes of raw email data, and the substantial resources required to create and maintain a high-quality

labeled dataset. These factors underscore that while the integration of AI and big data has the potential to significantly advance phishing email detection, realizing these benefits in practice may be constrained by technical, ethical, and logistical challenges.

Fig 5 : Benefits of AI in Cybersecurity

6. Conclusion

AI-driven phishing email detection is only logical, given the security challenges that the playback and reproduce approach to testing legacy systems cannot accommodate. It is indicative of a new wave of cyber hacks that require countermeasures that are equally as sophisticated. It ensures that the threat surface is contained, ensuring that the weakest links in an organization are safer by identifying suspicious emails before they are deleted or tagged as spam. With steadily advancing AI, it stands to reason that using big data AI could be one of several new solutions. Going further, designers could feasibly integrate more than just email datasets, cross-referencing networks and network latency with big data could be a means to find compromised and controlled devices far sooner as well as generate more intelligence on how they operate and spread.
In parallel, future work suggests a robust data and pattern-checking approach that leverages hashtags to pass verb+noun strings back and forth; if there's anything other than a VPN, it would detect a potential phish, for example. By leveraging shared concerns on cybersecurity via projects like CIDER, the storage of big data of email and email meta-tagging is likely to make it more useful than ever. Indeed, projects such as these are seeing a resurgence in funding and participation, in part due to their timely remit and focus on cybersecurity concerns. Given that well-behaved institutions enable convergence, it is not unreasonable to assume that they are able to attract attention. A future volume in this series, covering highlights from the CIDER 2016 workshop, is currently being prepared.

6.1. Future Trend AI-driven phishing detection is very promising; however, numerous obstacles need to be conquered to make it successful. In the future, attention should be focused on the development of a robust and effective AI-based phishing email detection system that is optimized for massive-scale security operations centers (SOCs). The application of big data analytics for cybersecurity lends itself to exponential improvements in the strength of AI tools, with the potential to harness much larger datasets. As a result of this development, this enhancement also stands to improve the accuracy and strength of AI-driven phishing detection systems.Analysts were prompted to transfer from rule-based systems to machine-learning classification systems as the quantity of phishing spam continuously increased. Over the past four years, the emphasis has progressively evolved to distinguish AI and non-AI. To attain maximum detection rate and precision, deep learning (DL), frequently linked through word and character embeddings, is the most recent tendency of non-AI approaches. Despite its remarkable achievements in a range of machine learning (ML) domains, state-of-the-art (SOTA) "out of the box" deep learning (DL) methods for the classification of temperate-sequence information have historically been outperformed by feature-engineered systems. However, the ACER-NSIT-BMIL and ACER-NSIT-IBM labels are merged by the deep learning model (created by combining word and character embeddings) as ACER-BMIL, but they cannot be separated appropriately and indicate a potential concern. The utilization of few-shot learning, which facilitates the transfer of knowledge from one area to another, holds notable possibilities for the improvement of AI-driven phishing email detection. With advances in deep learning solutions for few-shot learning, research becomes restricted to the expansion of dataset size and variety.

7. References

[1] Kommisetty, P. D. N. K., & Nishanth, A. (2024). AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI. In IARJSET (Vol. 9, Issue 10). Tejass Publishers. https://doi.org/10.17148/iarjset.2022.91020

[2] Vaka, D. K. (2023). Achieving Digital Excellence In Supply Chain Through Advanced Technologies. Educational Administration: Theory and Practice, 29(4), 680-688.

[3] Purshotam S Yadav. (2024). Optimizing Serverless Architectures for Ultra-Low Latency in Financial Applications. European Journal of Advances in Engineering and Technology. https://doi.org/10.5281/ZENODO.13627245

[4] Mahida, A. Secure Data Outsourcing Techniques for Cloud Storage.

[5] Zanke, P., Deep, S., Pamulaparti Venkata, S., & Sontakke, D. Optimizing Worker's Compensation Outcomes Through Technology: A Review and Framework for Implementations.

[6] Chintale, P., Khanna, A., Korada, L., Desaboyina, G., & Nerella, H. AI-Enhanced Cybersecurity Measures for Protecting Financial Assets.

[7] Pillai, S. E. V. S., Avacharmal, R., Reddy, R. A., Pareek, P. K., & Zanke, P. (2024, April). Transductive–Long Short-Term Memory Network for the Fake News Detection. In 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-4). IEEE.

[8] Vaka, D. K. Empowering Food and Beverage Businesses with S/4HANA: Addressing Challenges Effectively. J Artif Intell Mach Learn & Data Sci 2023, 1(2), 376-381.

[9] Kommisetty, P. D. N. K., & Abhireddy, N. (2024). Cloud Migration Strategies: Ensuring Seamless Integration and Scalability in Dynamic Business Environments. In the International Journal of Engineering and Computer Science (Vol. 13, Issue 04, pp. 26146–26156). Valley International. https://doi.org/10.18535/ijecs/v13i04.4812

[10] Yadav, P. S. (2024). Fast and Efficient UserID Lookup in Distributed Authentication: A Probabilistic Approach Using Bloom Filters. In the International Journal of Computing and Engineering (Vol. 6, Issue 2, pp. 1–16). CARI Journals Limited. https://doi.org/10.47941/ijce.2124

[11] Mahida, A., Chintale, P., & Deshmukh, H. (2024). Enhancing Fraud Detection in Real Time using DataOps on Elastic Platforms.

[12] Pamulaparti Venkata, S., & Avacharmal, R. (2023). Leveraging Interpretable Machine Learning for Granular Risk Stratification in Hospital Readmission: Unveiling Actionable Insights from Electronic Health Records. Hong Kong Journal of AI and Medicine, 3(1), 58-84.

[13] Chintale, P., Deshmukh, H., & Desaboyina, G. Ensuring regulatory compliance for remote financial operations in the COVID-19 ERA.

[14] Vaka, D. K. "Artificial intelligence enabled Demand Sensing: Enhancing Supply Chain Responsiveness.

[15] Avacharmal, R. (2024). Explainable AI: Bridging the Gap between Machine Learning Models and Human Understanding. Journal of Informatics Education and Research, 4(2).

[16] Kommisetty, P. D. N. K., & dileep, V. (2024). Robust Cybersecurity Measures: Strategies for Safeguarding Organizational Assets and Sensitive Information. In IJARCCE (Vol. 13, Issue 8). Tejass Publishers. https://doi.org/10.17148/ijarcce.2024.13832

[17] Yadav, P. S. (2024). Advanced Authentication and Authorization Mechanisms in Apache Kafka: Enhancing Security for High-Volume Data Processing Environments. In Journal of Engineering and Applied Sciences Technology (pp. 1–6). Scientific Research and Community Ltd. https://doi.org/10.47363/jeast/2024(6)e110

[18] Mahida, A. (2024). Integrating Observability with DevOps Practices in Financial Services Technologies: A Study on Enhancing Software Development and Operational Resilience. International Journal of Advanced Computer Science & Applications, 15(7).

[19] Vaka, D. K. " Integrated Excellence: PM-EWM Integration Solution for S/4HANA 2020/2021.

[20] Pamulaparti Venkata, S. (2023). Optimizing Resource Allocation For Value-Based Care (VBC) Implementation: A Multifaceted Approach To Mitigate Staffing And Technological Impediments Towards Delivering High-Quality, Cost-Effective Healthcare. Australian Journal of Machine Learning Research & Applications, 3(2), 304-330.

[21] Chintale, P., Korada, L., WA, L., Mahida, A., Ranjan, P., & Desaboyina, G. RISK MANAGEMENT STRATEGIES FOR CLOUD-NATIVE FINTECH APPLICATIONS DURING THE PANDEMIC.

[22] Avacharmal, R., Pamulaparti Venkata, S., & Gudala, L. (2023). Unveiling the Pandora's Box: A Multifaceted Exploration of Ethical Considerations in Generative AI for Financial Services and Healthcare. Hong Kong Journal of AI and Medicine, 3(1), 84-99.

[23] Kommisetty, P. D. N. K., vijay, A., & bhasker rao, M. (2024). From Big Data to Actionable Insights: The Role of AI in Data Interpretation. In IARJSET (Vol. 11, Issue 8). Tejass Publishers. https://doi.org/10.17148/iarjset.2024.11831

[24] Yadav, P. S. (2023). Enhancing Software Testing with AI: Integrating JUnit and Machine Learning Techniques. North American Journal of Engineering Research, 4(1).

[25] Mahida, A. Explainable Generative Models in FinCrime. J Artif Intell Mach Learn & Data Sci 2023, 1(2), 205-208.

[26] Pamulaparti Venkata, S., Reddy, S. G., & Singh, S. (2023). Leveraging Technological Advancements to Optimize Healthcare Delivery: A Comprehensive Analysis of Value-Based Care, Patient-Centered Engagement, and Personalized Medicine Strategies. Journal of AI-Assisted Scientific Discovery, 3(2), 371-378.

[27] Chintale, P., & Desaboyina, G. (2018). FLUX: AUTOMATING CLUSTER STATE MANAGEMENT AND UPDATES THROUGH GITOPS IN KUBERNETES. International Journal of Innovation Studies, 2(2).

[28] Avacharmal, R., Sadhu, A. K. R., & Bojja, S. G. R. (2023). Forging Interdisciplinary Pathways: A Comprehensive Exploration of Cross-Disciplinary Approaches to Bolstering Artificial Intelligence Robustness and Reliability. Journal of AI-Assisted Scientific Discovery, 3(2), 364-370.

[29] Vaka, D. K. (2020). Navigating Uncertainty: The Power of 'Just in Time SAP for Supply Chain Dynamics. Journal of Technological Innovations, 1(2).

[30] Kommisetty, P. D. N. K., & Nishanth, A. (2024). AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI. In IARJSET (Vol. 9, Issue 10). Tejass Publishers. https://doi.org/10.17148/iarjset.2022.91020

[31] Yadav, P. S. REAL-TIME INSIGHTS IN DISTRIBUTED SYSTEMS: ADVANCED OBSERVABILITY TECHNIQUES FOR CLOUD-NATIVE ENTERPRISE ARCHITECTURES.

[32] Mahida, A. (2023). Machine Learning for Predictive Observability-A Study Paper. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-252. DOI: doi. org/10.47363/JAICC/2023 (2), 235, 2-3.

[33] Tilala, M., Pamulaparti Venkata, S., Chawda, A. D., & Benke, A. P. Explore the Technologies and Architectures Enabling Real-Time Data Processing within Healthcare Data Lakes, and How They

Facilitate Immediate Clinical Decision-Making and Patient Care Interventions. European Chemical Bulletin, 11, 4537-4542.

[34] Perumal, A. P., & Chintale, P. Improving operational efficiency and productivity through the fusion of DevOps and SRE practices in multi-cloud operations.

[35] Avacharmal, R., Gudala, L., & Venkataramanan, S. (2023). Navigating The Labyrinth: A Comprehensive Review Of Emerging Artificial Intelligence Technologies, Ethical Considerations, And Global Governance Models In The Pursuit Of Trustworthy AI. Australian Journal of Machine Learning Research & Applications, 3(2), 331-347.