ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1026

Algorithms Based on Artificial Intelligence for the Detection and Prevention of Social Engineering Attacks: Systematic review

Atuncar Flores, Edgardo[1], Chuan Garcia, Anthony Francisco[1], Pachas Quispe, Gustavo Henry[1], and Raez Martinez, Haymin Teresa[1]
[1]Universidad Tecnológica del Perú, Perú, U18309014@utp.edu.pe, U19202368@utp.edu.pe, C24336@utp.edu.pe, C19240@utp.edu.pe

Abstract*-- In this study, the growing challenge of cybersecurity is addressed by reviewing Artificial Intelligence (AI) based algorithms designed for the detection and prevention of Social Engineering attacks. The research focuses on identifying effective algorithms, with special attention to phishing, a widely prevalent type of attack. Using the PICOC framework, initially, 891 articles from SCOPUS were collected, of which, after applying rigorous criteria through the Prisma methodology, 32 were selected for detailed analysis. The results reveal that among the studied algorithms, XGBoost, Random Forest (RF), and the combination of FastText-CBOW with Random Forest stand out, exhibiting accuracy rates exceeding 99% in the detection of social engineering attacks. This analysis supports the effectiveness of AI-based tools compared to traditional methods, especially in situations of immediate or 'Zero Hour' attacks. In conclusion, AI emerges as a significant alternative to strengthen cybersecurity and protect against increasingly sophisticated threats.

Keywords*-- Social engineering attacks, Artificial intelligence, Detection algorithms, Phishing, Smishing

Algoritmos Basados en Inteligencia Artificial para la Detección y Prevención de Ataques de Ingeniería Social: Revisión Sistemática

Atuncar Flores, Edgardo[1], Chuan Garcia, Anthony Francisco[1], Pachas Quispe, Gustavo Henry[1], and Raez Martinez, Haymin Teresa[1]
[1]Universidad Tecnológica del Perú, Perú, U18309014@utp.edu.pe, U19202368@utp.edu.pe, C24336@utp.edu.pe, C19240@utp.edu.pe

Resumen*-- En este estudio, se aborda el creciente desafío de la inseguridad informática mediante la revisión de algoritmos basados en Inteligencia Artificial (IA) destinados a la detección y prevención de ataques de Ingeniería Social. La investigación se centra en identificar algoritmos efectivos, con especial atención al phishing, un tipo de ataque ampliamente prevalente. Utilizando el esquema PICOC, se recopilaron inicialmente 891 artículos de SCOPUS, de los cuales, tras aplicar criterios rigurosos mediante la metodología Prisma, se seleccionaron 32 para un análisis detallado. Los resultados revelan que, entre los algoritmos estudiados, XGBoost, Random Forest (RF) y la combinación de FastText-CBOW con Random Forest destacan, exhibiendo tasas de aciertos superiores al 99% en la detección de ataques de ingeniería social. Este análisis respalda la eficacia de las herramientas basadas en IA en comparación con métodos tradicionales, especialmente en situaciones de ataques inmediatos o de 'Hora cero'. En conclusión, la IA emerge como una alternativa significativa para fortalecer la ciberseguridad y proteger contra amenazas cada vez más sofisticadas.

***Palabras Clave*-- Ataques de ingeniería social, Inteligencia artificial, Algoritmos de detección, Phishing, Smishing.**

I. INTRODUCCIÓN

En los últimos años se ha visto un crecimiento exponencial en el uso de las Tecnologías de la Información y Comunicación (TIC). Esta evolución ha transformado radicalmente la forma en que vivimos, trabajamos y nos relacionamos, destacando aún más la importancia de la conectividad y la comunicación digital en la sociedad actual [1]. No obstante, como señalan [2], este crecimiento ha dado lugar a que los ataques de Ingeniería Social (SE) dirigidos a la explotación de sistemas de información se conviertan en un desafío significativo en el entorno de la navegación web. Estos ataques han emergido como un vector prominente para el robo de información sensible, el acceso no autorizado a sistemas y la propagación de estafas en línea. Entre los ataques más comúnmente empleados por los ciberatacantes, destaca el phishing. Según la definicióno [3], se trata de una técnica fraudulenta que simula una web normal para adquirir información sensible y confidencial del usuario. Además, [4] añaden que los ciberdelincuentes pueden acercarse a sus objetivos enviando mensajes falsos a través de correo electrónico (por ejemplo, Gmail u Outlook) o redes sociales, con el fin de engañar a las víctimas. Estos ataques que van en aumento cada día se han se han convertido en una grave amenaza y es imperativo detectarlos a tiempo. Sin embargo, aunque la tecnología de anti ataques mejoran constantemente, aún no existe ninguna tecnología eficaz que pueda prevenir completamente los ataques de SE [3].

Por otro lado, el uso de la Inteligencia Artificial (IA) ha demostrado ser altamente eficaz para la detección temprana de ataques de phishing [5]. Gracias a su alta capacidad de análisis de patrones y comportamientos, la IA facilita la identificación de indicios de ataques de ingeniería social con una precisión notablemente mayor. Según la investigación [6], los algoritmos de aprendizaje profundo automático tienen la capacidad de aprender de los datos existentes de manera autónoma y aplicar ese conocimiento a nuevos datos, lo que les otorga un alto potencial en la detección de sitios web de phishing. Una tendencia común ha sido la extracción de características mediante procesamiento del lenguaje natural (PNL), lo que ha demostrado resultados notables en la detección de phishing [7]. Otros métodos han incorporado el aprendizaje por refuerzo y características como URL incrustadas, contenido HTML y encabezado de correo electrónico. Gracias a la disponibilidad de conjuntos de datos sólidos, las tasas de detección de phishing han alcanzado niveles mayores al 98% [2].

En este contexto, la ciudad de Lima, Perú se destaca por la creciente sofisticación de los ataques de Ingeniería Social (SE). Los ciberatacantes han ajustado sus estrategias para ser altamente específicas y persuasivas, uno de los ataques más comunes en lima es el smishing. Según [5], se define como un ataque fraudulento similar al phishing pero que involucra mensajes de texto (SMS) que parecen legítimos, pero redirigen a sitios web maliciosos, donde se solicita información personal, engañando al usuario para que revele datos confidenciales. Adicionalmente, existe un desafío relacionado con la falta de conciencia y educación en ciberseguridad entre la población peruana en general. Las investigaciones [1] indican que muchos usuarios aún no están completamente informados sobre las tácticas de SE y las señales de advertencia a las que deben prestar atención, lo que los vuelve más susceptibles a caer en trampas cibernéticas y revelar información confidencial sin percatarse de ello. Asimismo, la negligencia de las personas juega un papel crítico en el aumento de los incidentes de seguridad, lo que conlleva a un incremento en la efectividad de los ataques [7] .

Para concluir, se ha evidenciado la necesidad de llevar a cabo una Revisión Sistemática de Literatura (RSL) en este campo debido a la creciente amenaza de estafas de phishing, como respaldan las investigaciones [2], estos ataques se han vuelto más prominentes a medida que los conocimientos técnicos disminuyen y los costos de los ataques de phishing son cada vez más accesibles para los atacantes. A pesar de este incremento alarmante, aún persiste la carencia de una

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

caracterización precisa en términos de la identificación de características comunes y la capacidad de anticipar las tendencias futuras en los ataques de phishing [8]. A raíz de esta necesidad práctica, se justifica la elección del tema propuesto para esta Revisión Sistemática de Literatura (RSL). A pesar de la presencia de investigaciones dispersas y estudios individuales sobre el tema, no se ha realizado una revisión sistemática actual y exhaustiva que evalúe de manera integral la literatura científica vigente relacionada con la implementación de IA en un algoritmo para prevenir ataques de Ingeniería Social a través del análisis de aplicaciones de mensajería, correos electrónicos y enlaces web. Este vacío es de gran relevancia, especialmente a la luz de la Cuarta Revolución Industrial, en la que la IA se considera una tecnología crítica para salvaguardar los sistemas de redes informáticas contra ciberataques, malware, phishing y otras amenazas [9].

El objetivo de esta investigación radica en analizar y evaluar los principales algoritmos utilizados para la detección de ataques de ingeniería social, enfocándose en su efectividad y precisión. Se busca comprender el panorama actual de las soluciones basadas en Inteligencia Artificial aplicada a esta problemática, identificando y evaluando los algoritmos más eficaces. El objetivo primordial de este estudio es realizar una revisión sistemática y exhaustiva de los distintos algoritmos basados en IA empleados para la detección de ataques de SE, destacando aquellas que demuestren un alto rendimiento en la identificación y prevención de dichos ataques. Se aspira a proporcionar una visión integral y organizada de las soluciones disponibles en la literatura, mediante una revisión cuidadosa y detallada en este ámbito específico

En este sentido, la estructura de esta investigación se divide de la siguiente manera. En la segunda sección, Metodología, se describe el método utilizado para llevar a cabo la revisión sistemática de la literatura, detallando el proceso desde la selección de nuestra pregunta de investigación hasta llegar a la formulación de la ecuación de investigación que guiará la selección de los artículos científicos para esta RSL. La tercera sección, Resultados, presenta y organiza los principales hallazgos obtenidos en el análisis de los artículos seleccionados en la segunda sección, mostrándolos a través de imágenes y tablas. En la cuarta sección, Discusión, se abordan los resultados obtenidos anteriormente, interpretándolos y discutiendo las limitaciones encontradas. Finalmente, en la quinta sección, Conclusiones, se resumen los hallazgos clave y las limitaciones de la investigación, además de proponer sugerencias para futuras investigaciones que puedan abordar las limitaciones identificadas en esta RSL.

II. METODOLOGÍA

A. *Descripción de la estrategia de búsqueda sistemática*

El proceso metodológico se inició con la formulación de una pregunta de investigación específica. Esta pregunta se estructuró siguiendo el marco PICOC (Population, Intervention, Comparison, Outcome, and Context), que desglosa la pregunta en cinco componentes esenciales: la población de interés (P), la intervención o exposición bajo consideración (I), la comparación o alternativa evaluada (C), los resultados esperados (O) y el contexto relevante. Según Richardson [10], para formular preguntas de manera efectiva, se deben tener en cuenta dos aspectos clave. En primer lugar, la pregunta debe estar estrechamente relacionada con el problema que se está abordando. En segundo lugar, debe estar formulada de manera que facilite la búsqueda de respuestas precisas.

1.1. Pregunta PICOC y sus componentes

En la primera etapa del proceso de la estrategia de búsqueda, se formula la pregunta PICOC. En este caso, la pregunta planteada es:

¿Cuáles son los métodos basados en inteligencia artificial más eficaces para la detección y prevención de ataques de ingeniería social en dispositivos inteligentes?

Una vez formulada la pregunta PICOC, se llevará a cabo un proceso de análisis en el cual dicha pregunta será desglosada en cada uno de los componentes antes mencionados, tal como se muestra en la Tabla I.

TABLA I
COMPONENTES PICOC

P	Población	Ataques de Ingeniería Social
I	Intervención	Métodos basados en inteligencia artificial
C	Comparación	Otros metodos tradicionales de detección
O	Resultados	Detección y Prevención
C	Contexto	Dispositivos Inteligentes

1.2. Palabras claves especializadas pertinentes

Sucesivamente, en la etapa de análisis se procede a reconocer las palabras clave asociadas a cada uno de los elementos presentes en la pregunta PICOC, tal como se muestran en la tabla II. Este proceso es esencial para orientar de manera efectiva la búsqueda de evidencia científica y garantizar que los resultados sean precisos y relevantes para la investigación en curso.

TABLA II
PALABRAS CLAVE

P	Población	Engineering, attacks, phishing, smishing
I	Intervención	detection methods with artificial intelligence, artificial intelligence, detection methodss, Machine Learning
C	Comparación	non artificial intelligence-based methods, conventional methods, traditional methods
O	Resultados	Prevention detection, protection, security, detection accuracy, detection precision, Precision in attack prevention

C	Contexto	smart devices, advanced devices, intelligent devices, smartphones, computer

1.3. Ecuación de búsqueda empleada

Finalmente, como señala la tabla III, en esta tercera fase del proceso, se define la estructura correcta para la ecuación de búsqueda, incorporando el uso del operador OR y la inclusión de comillas (“”) para las expresiones de palabras compuestas. Asimismo, se llevó a cabo la integración de cada uno de los elementos PICOC mediante la aplicación del operador AND.

TABLA III

ECUACIÓN DE BÚSQUEDA

("engineering attacks" OR "phishing" OR "smishing") AND ("detection methods with artificial intelligence" OR "artificial intelligence" OR "detection methodss" OR "Machine Learning") AND ("prevention" OR "detection" OR "protection" OR "security" OR "detection accuracy" OR "detection precision" OR "Precision in attack prevention") AND ("smart devices" OR "advanced devices" OR "intelligent devices" OR "smartphones" OR "computer")

1.4. Definición de Criterios de Inclusión y Exclusión

Posteriormente a la aplicación del esquema PICOC y su análisis, se establecieron criterios de inclusión y exclusión con el objetivo de definir claramente los parámetros que guiarán la selección de documentos para este estudio. La implementación de estos criterios busca asegurar la relevancia y coherencia de los materiales recopilados, optimizando así la calidad y confiabilidad de los resultados obtenidos.

Criterios de Inclusión

CI1: Documentos que sean artículos científicos

CI2: Documentos publicados a partir del año 2020

CI3: Documentos que estén en idioma Ingles

CI4: Documentos que presenten resultados favorables con una tasa de éxito mayor al 90%

Criterios de Exclusión

CI1: Documentos diferentes de artículos científicos

CI2: Documentos anteriores al 2020

CI3: Documentos que no tenga acceso libre

CI4: Documentos que no estén en idioma ingles

B. Descripción del Proceso de Selección

En base a mejorar la claridad, transparencia, calidad y valor de los informes sistemáticos, se adoptó un enfoque basado en la metodología PRISMA. De acuerdo con la investigación [11], esta metodología ha sido desarrollada con el objetivo específico de validar las fuentes y llevar a cabo un análisis objetivo y sistemático de los resultados de estudios empíricos relacionados con un problema de investigación particular. Este método responde a la creciente necesidad de sintetizar de manera efectiva la información más relevante en un contexto de expansión continua de la investigación científica. El proceso de selección se rige por criterios de inclusión/exclusión y sigue una metodología documental establecida, respaldada por investigaciones sistemáticas recientes en diversas áreas de investigación.

1.1. Resultados obtenidos del proceso de búsqueda de literatura científica

En primera instancia, se realizó la búsqueda de información en la base de datos Scopus utilizando una ecuación de búsqueda previamente definida. En esta búsqueda se obtuvieron 891 resultados, como se observa en la Fig 1. Estos resultados se someterán a un proceso de selección para identificar y definir las fuentes que se incluirán en la revisión sistemática de literatura (RSL).

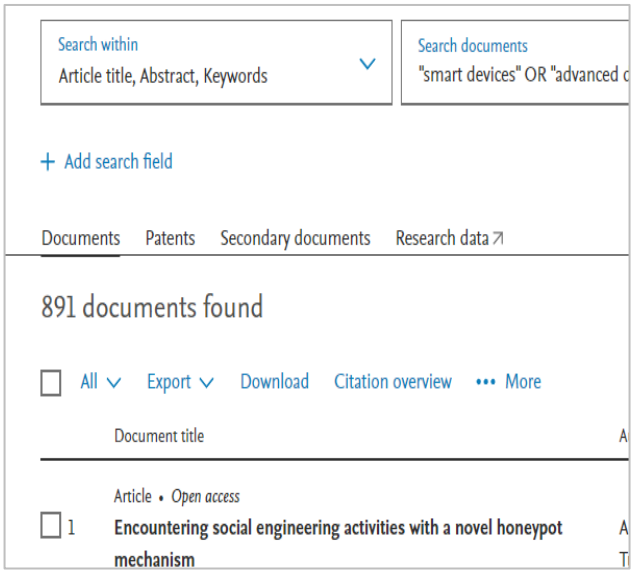

Fig. I. Resultados de búsqueda en Scopus aplicando la ecuación de búsqueda.

1.2. Descripción de la lógica de selección considerada (PRISMA)

Seguidamente, después de definir los resultados de la búsqueda, se llevaron a cabo diversas etapas. En primer lugar, se identificaron los estudios a través de Scopus como fuente principal de búsqueda, y se registró el número total de estudios identificados en estas etapas iniciales. Luego, se procedió a eliminar duplicados utilizando la herramienta Mendeley. Seguidamente, se realizó el análisis de los resúmenes de los estudios únicos para evaluar su relevancia de acuerdo con los criterios de inclusión. Posteriormente, se llevó a cabo el análisis de los criterios de exclusión e inclusión. Sucesivamente, se identificaron y excluyeron aquellos estudios que no se ajustaban al esquema PICO previamente definido. Finalmente, se resumió el número total de estudios que cumplían con todos los criterios y que serán considerados en la revisión sistemática de literatura. Este proceso garantiza la inclusión de estudios relevantes y adecuados para responder a la pregunta de investigación de manera rigurosa y objetiva en la RSL.

1.3. Descripción detallada de los pasos del proceso de selección y sus resultados (PRISMA)

Primeramente, durante la etapa de identificación, se emplearon palabras clave específicas para recuperar registros pertinentes. Este proceso arrojó un total de 891 registros.

Seguidamente, en la etapa de detección y gestión de duplicados, se aprovechó la funcionalidad de Mendeley para identificar y eliminar eficazmente los artículos duplicados presentes en la base de datos Scopus. Durante este proceso, se llevaron a cabo exhaustivas comparaciones entre los registros obtenidos en Scopus, utilizando diversos criterios adicionales. Además de la comparación de autores, los identificadores de objetos digitales (DOI), y las fechas de publicación, esta metodología meticulosa garantizó la eliminación precisa de duplicados. En este caso se identificó solamente un registro duplicado, el cual fue eliminado con éxito, resultando en un total de 890 registros finales provenientes de Scopus.

Posteriormente, en la fase de elegibilidad de estudios aptos, se llevó a cabo la exclusión de registros que no se alineaban con la temática de nuestra investigación, basándonos en la evaluación de los títulos y/o resúmenes. Como resultado de este proceso, se obtuvieron un total de 186 registros que se consideraron pertinentes y relevantes para el estudio.

Asimismo, en el proceso de análisis de los criterios de inclusión y exclusión, se llevaron a cabo tres etapas distintas para asegurar la inclusión de los registros más relevantes. Primero, se descartaron 114 documentos que no cumplían con la categoría de "Artículo", lo que redujo la cantidad a 72. Seguidamente, se excluyeron 17 documentos anteriores al año 2020, con lo que se llegó a un total de 55 registros. En la tercera fase, se procedió a eliminar 17 documentos que no estaban disponibles en acceso abierto, culminando así con 38 registros que cumplen con todos los criterios establecidos para nuestro estudio. Este logro fue posible gracias a la utilización de las herramientas de filtrado proporcionadas por la plataforma Scopus, tal como se ilustra en la Fig III. Este proceso de selección rigurosa garantiza que los documentos incluidos sean pertinentes y adecuados para la investigación.

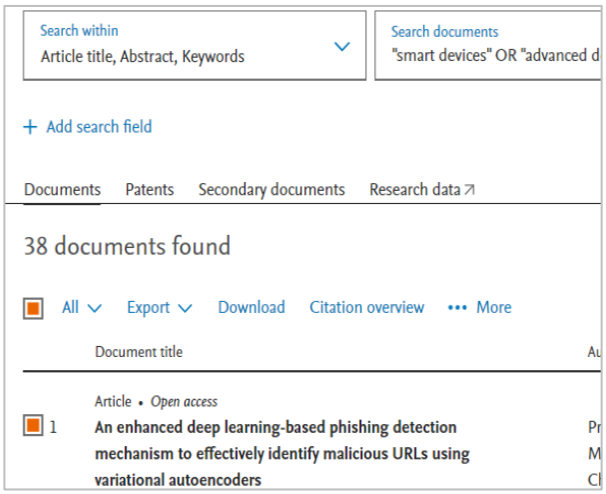

Fig. II. Resultados posteriores a los criterios de inclusión y exclusión

Finalmente, en la fase de exclusión basada en el esquema PICOC, se llevó a cabo un análisis detallado de cada artículo para garantizar que solo se incluyeran aquellos que cumplían rigurosamente con los criterios preestablecidos y que contribuían significativamente al objetivo de la investigación. Se excluyeron los artículos que no presentaban resultados favorables en la detección de ataques de Ingeniería Social (SE), y aquellos que carecían de una profundización sustancial en la investigación.

Además, se descartaron los artículos que abordaban la temática sin aportar información sustancial a la investigación en curso. Este proceso de exclusión se realizó para garantizar que los artículos seleccionados no solo estuvieran alineados con el esquema PICOC, sino valiosos y relevantes para el análisis y la revisión sistemática, fortaleciendo así la calidad y la coherencia de la investigación final. El resultado de esta fase fue la identificación y retención de un total de 29 artículos que cumplían con los criterios más estrictos de inclusión y contribuían significativamente al corpus de conocimiento sobre la detección de ataques de Ingeniería Social.

1.4. Diagrama de flujo PRISMA que refleje gráficamente el proceso

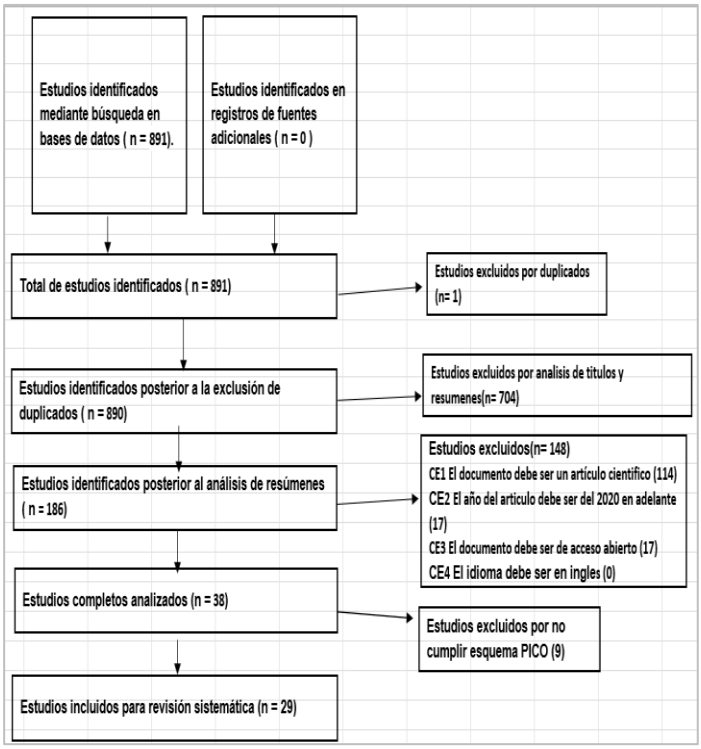

Fig III. Diagrama de Flujo PRISMA

III. RESULTADOS

Se realizó un análisis de los 29 artículos científicos seleccionados en la etapa previamente descrita. Estos artículos se publicaron en diversas revistas y conferencias académicas entre los años 2020 y 2023 y provienen de países de todo el mundo, lo que destaca la diversidad geográfica en la investigación de este campo. Los estudios abordaron una variedad de contextos, con un enfoque principal en la investigación cuantitativa y la investigación experimental. El año más antiguo de publicación fue 2020, y el más reciente, 2023, lo que muestra la actualidad de la investigación en esta área. La editorial más asociada con estos estudios fue IEEE, y

la revista más utilizada fue IEEE Access. Este análisis refleja la diversidad de fuentes y enfoques utilizados en el estudio de la detección de ataques de ingeniería social en ciberseguridad como se puede observar en la tabla IV.

Tabla IV
Datos Generales de Referencias

Referencia	Año de publicación	Título de la revista	Editorial	País de origen del estudio	Contexto de aplicación	Tipo de estudio
[12]	2020	IEEE Access	IEEE	Nigeria	ciberseguridad	Investigación cuantitativa
[13]	2022	Journal of Cyber Security and Mobility	River Publishers	India	Ciberseguridad	Investigación cuantitativa
[14]	2022	IEEE Access	IEEE	Suiza	Ciberseguridad	Investigación cuantitativa
[15]	2023	Neural Computing and Applications	Springer	India	Ciberseguridad	Investigación cuantitativa
[16]	2020	IEEE Access	IEEE	Australia	Ciberseguridad	Investigación cuantitativa
[7]	2022	IEEE Access	IEEE	Grecia	Ciberseguridad	Investigación cuantitativa
[5]	2022	IEEE Access	IEEE	Japón	Ciberseguridad	Investigacion experimental
[17]	2020	IEEE Access	IEEE	Brasil	Ciberseguridad	Investigación cuantitativa
[8]	2021	IET Information Security	IET	Baréin	Ciberseguridad	Investigación cuantitativa
[18]	2020	IEEE Access	IEEE	China	Ciberseguridad	Investigación cuantitativa
[19]	2022	IEEE Access	IEEE	España	Ciberseguridad	Investigación cuantitativa
[20]	2023	IEEE Access	IEEE	Italy	Ciberseguridad	análisis del estado del arte
[6]	2023	IEEE Access	IEEE	India	Ciberseguridad	Investigación cuantitativa
[21]	2020	IEEE Access	IEEE	Corea del sur	Ciberseguridad	Investigación cuantitativa
[22]	2022	Journal of Cyber Security and Mobility	River Publishers	India	Ciberseguridad	Investigación experimental
[23]	2023	IEEE Access	IEEE	Pakistan	Ciberseguridad	Investigación experimental
[24]	2020	IEEE Access	IEEE	Arabia Saudita	Ciberseguridad	Investigación experimental
[25]	2022	IEEE Access	IEEE	Arabia Saudita	Ciberseguridad	Investigación experimental
[26]	2023	IEEE Access	IEEE	Grecia	Ciberseguridad	Investigación experimental
[27]	2023	IEEE Access	IEEE	Brasil	Ciberseguridad	Investigación experimental
[28]	2022	IEEE Access	IEEE	Malasia	Ciberseguridad	Investigación experimental
[29]	2023	IEEE Access	IEEE	Bangladesh	Ciberseguridad	Investigación experimental
[30]	2022	IEEE Access	IEEE	Sri Lanka	Ciberseguridad	Investigación experimental
[31]	2022	IEEE Access	IEEE	Turquía	Ciberseguridad	Investigación experimental
[32]	2020	IEEE Access	IEEE	Japón	Ciberseguridad	Investigación experimental
[3]	2022	IEEE Access	IEEE	Canadá	Ciberseguridad	Investigación experimental
[2]	2023	IEEE Access	IEEE	Arabia Saudita	Ciberseguridad	Investigación experimental
[1]	2023	IEEE Access	IEEE	Iraq	Ciberseguridad	Investigación experimental
[4]	2023	IEEE Access	IEEE	China	Ciberseguridad	Investigación experimental

En la Fig. IV se presentan los recuentos de frecuencias de algoritmos utilizados en los estudios de detección de ataques de ingeniería social (SE), donde se destacan las máquinas de vectores soporte (SVM) con un total de 12 menciones en los

estudios [6], [7], [13], [16], [17], [22], [24], [25], [27], [29], [31], [32]; así como Random Forest (RF), que también cuenta con un total de 12 menciones en los mismos [2], [3], [8], [13], [15], [17], [22], [23], [24], [25], [31], [32]. En segundo lugar, se encuentra el algoritmo de árbol de decisión (DT) con 10 menciones [2], [6], [8], [13], [15], [16], [17], [23], [25], [31].

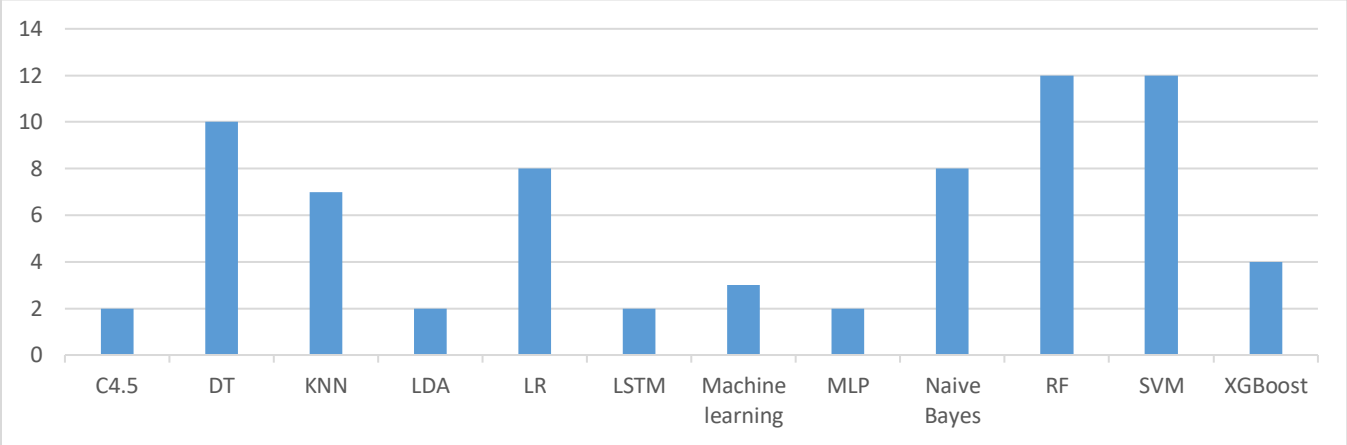

Fig. IV Frecuencia de algoritmos basados en IA

Seguidamente, los algoritmos de Naive Bayes y regresión logística (LR) son ampliamente utilizados, con 8 menciones cada uno en los estudios [6], [15], [16], [17], [22], [24], [25], [27] y [6], [13], [17], [19], [22], [23], [27], [31], respectivamente. Por otro lado, K-Nearest Neighbors (KNN) se destaca con 7 menciones en los estudios [16], [17], [23], [24], [25], [28], [31], lo que subraya su relevancia en la detección de ataques de ingeniería social. Asimismo, XGBoost ha sido ampliamente utilizado, con 4 menciones en los estudios [13], [17], [23], [27], consolidando su importancia en este ámbito. En total, se aplicaron 59 algoritmos diferentes en los estudios de detección de ataques de ingeniería social (SE), lo que demuestra varios enfoques para abordar el problema de seguridad cibernética.

Continuando con el análisis de los datos posteriores a la recopilación de resultados de los artículos, se destaca que el fenómeno de ingeniería social más abordado en los estudios es el phishing, contabilizando 26 menciones. También, se han identificado distintas modalidades de ataques, tales como el smishing mencionado en 4 ocasiones, y otros tipos de ataques de SE mencionados en al menos un estudio. Por otro lado, se observa que los ataques de fuerza bruta, la suplantación de identidad, las estafas de citas en línea, las estafas de entrega o envío, las estafas de soporte técnico, el vishing, etc., fueron referidos una vez en cada caso. Dichos resultados ponen de manifiesto que el phishing se configura como el ataque de ingeniería social predominante en los estudios objeto de análisis, seguido de cerca por otros variados tipos de ataques (véase la Fig. V).

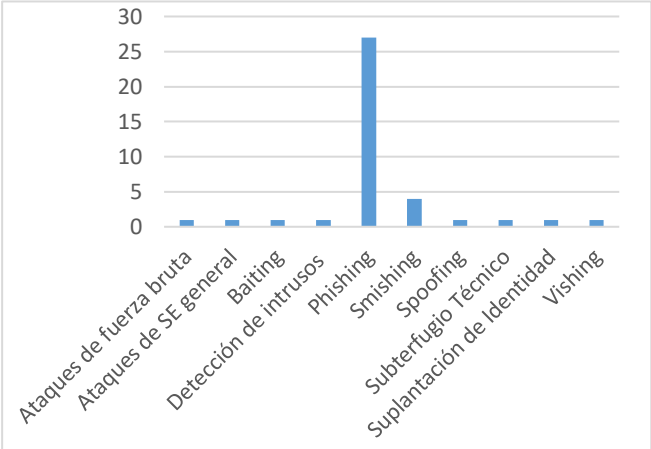

Fig. V. Ataques de SE más comunes tratados en los estudios

Siguiendo con la revisión de los algoritmos utilizados en la detección de ataques de ingeniería social, se destacan varios enfoques altamente eficaces. En este punto específico, es relevante señalar que se recopilaron datos de 22 de los 29 estudios disponibles, ya que algunos de ellos no proporcionaron datos cuantitativos. En este contexto, se determinó que el algoritmo "XGBoost" sobresale como el más destacado, exhibiendo una precisión del 99.84% y una puntuación F1 de 0.9984 [17]. Inmediatamente después, se encuentra "RF" en el mismo eswtudio con una precisión del 99.74%. Además, es importante resaltar que es uno de los algoritmos más recurrentes en numerosos estudios, aunque su precisión no siempre se especifica con claridad, generalmente sobresale entre los resultados más destacados. En la misma línea, "FastText-CBOW Combinado con Random Forest (RF)" alcanza una precisión del 99.5% [13]. Por su parte, "RNN GRU" también muestra un rendimiento sólido, logrando una precisión del 99.18% [3]. Seguidamente, se encuentra "SVM" con una precisión solida del 99% [16].

Finalmente, "RF + PNL" presenta una precisión del 98.38% y una puntuación F1 de 0.9837, respaldando su desempeño efectivo [23]. En el caso de "KNN", demuestra una precisión del 98%, y a pesar de una baja tasa de falsos positivos del 0.01%, no registra falsos negativos [28]. Es crucial tener en cuenta que algunos algoritmos, como "CSE-PUC" y "LR+SVC+DT," presentan tasas de eficacia inferiores

y requieren una revisión más detallada para su aplicación efectiva [6], [7] (véase la tabla V).

TABLA V

ALGORITMOS MÁS EFICACES DESTACADOS EN LOS ESTUDIOS

Referencia	Algoritmos	Precisión	Falso Positivo	Falso Negativo	F1
[12]	LBET	97.576	0.018	0.033	-
[13]	FastText-CBOW con Random Forest (RF)	99.5	-	-	-
[16]	SVM	99	-	-	-
[15]	Backpropagation	97.93	-	-	-
[7]	CSE-PUC	71.6	-	-	-
[5]	Machine learning	95	-	-	-
[17]	XGBoost	99.84	-	-	0.9984
	RF	99.74	-	-	0.9896
[8]	RF	-	-	-	-
[18]	SPWalk	95	-	-	-
[19]	TF-IDF combinado con N-gram y LR	96.5	-	-	-
[6]	LR+SVC+DT	96.77	-	-	-
[21]	PhishHaven	98	-	-	-
[22]	RF + PNL	97.98	-	-	
[23]	RF	98.38	-	-	0.9837
[24]	BPNN	95.88	-	-	-
[25]	MLSELM	98.44	-	-	-
[27]	XGBoost	-	-	-	0.9995
[28]	KNN	98	0.01	-	-
[31]	RF	96.0733	-	-	-
[32]	SVM con Doc2vec	99	-	-	-
[12]	LBET	-	-	-	-
	ABET	-	-	-	-
[3]	RNN GRU	99.18	-	-	-
	RF	-	0.0047	-	-

En la tabla VI, se evidencia los algoritmos de detección de SE basados en inteligencia artificial, representados por FastText-CBOW combinado con Random Forest (RF) [13] , XGBoost y RF [17], exhiben una alta precisión en la detección de ataques de ingeniería social, con valores de precisión del 99.5%, 99.84% y 99.74%, respectivamente. Aunque estos métodos moderadamente rápidos, son inferiores en velocidad en comparación con los métodos tradicionales; no obstante, demuestran ser efectivos contra los ataques en hora cero. Por otro lado, los métodos tradicionales, como las Listas Negras y Listas Blancas[20], [30], si bien pueden ser rápidos en términos de velocidad, no son efectivos contra dichos ataques y su eficacia en general es variable. De igual manera, el método basado en Similitudes [20], aunque lento y complejo de implementar, muestra efectividad contra ataques de hora cero.

TABLA VI

COMPARACIÓN DE LOS MÉTODOS TRADICIONALES CON LOS ALGORITMOS MÁS EFICACES EN LA DETECCIÓN DE ATAQUES SE

Referencia	Algoritmo	Precisión	Velocidad	Efectividad contra ataques en hora cero
[13]	FastText-CBOW con Random Forest (RF)	99.5	Moderada	Sí
[17]	XGBoost	99.84	Moderada	Sí
[20], [30]	Listas Negras	Variable	Rápida	No
[20], [30]	Listas Blancas	Variable	Rápida	No
[20]	Similitudes	Variable	Lenta	Sí

En la Fig. VI se exhiben los recuentos de aplicaciones de los algoritmos utilizados en los estudios abordados Destacando en primer lugar, se observa un énfasis en la detección de sitios web de phishing, con un total de 12 aplicaciones en estudios [3], [5], [6], [12], [16], [18], [20], [22], [23], [24], [25], [31]. En segundo lugar, se aplica la detección phishing en un contexto general, abordada en 6 estudios [1], [2], [9], [15], [22], [27]. Asimismo, se aprecia la aplicación en la detección de correos electrónicos de phishing con 3 menciones [14], [17], [22], así como la detección de smishing con aplicación en 2 estudios [5], [15]. Además de estos contextos, se identifican diversas aplicaciones adicionales, las cuales varían según los objetivos específicos de cada estudio y su enfoque particular.

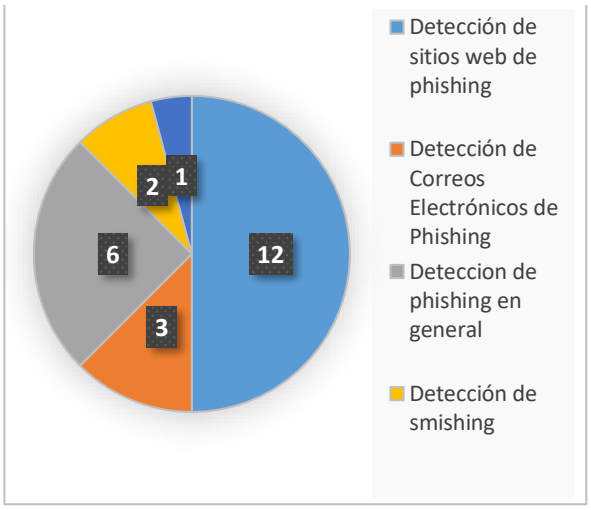

Fig VI. Contextos donde se aplicaron los métodos de detección de ataques de SE

IV. DISCUSIÓN

En este estudio exhaustivo, se ha evidenciado que el phishing es el tipo de ataque de ingeniería social más ampliamente estudiado, respaldado por 25 estudios [1], [2], [3], [4], [5], [6], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [27], [28], [29], [30], [31], [32], seguido del smishing [1], [5], [15], [16]. Así mismo, se considera el cumplimiento del objetivo inicial tras identificar los algoritmos para combatir estos ataques, los estudios destacan los métodos de detección basados en inteligencia artificial (IA), como XGBoost, SVM y Random Forest, resaltando por su eficacia en la identificación de estos ataques. A pesar de la excepcional precisión del 99.84% alcanzada por XGBoost, es crucial destacar que RF [17] alcanzo una precisión de 99.74% y FastText-CBOW Combinado con Random Forest (RF) [13] también se posicionó como uno de los mejores métodos, logrando una precisión del 99.5%. No obstante, RF cuenta con un mayor respaldo de investigaciones [3], [8], [13], [17], [22], [23], [31], a diferencia de XGBoost, que está respaldado únicamente por un estudio [17].

Estos resultados enfatizan la relevancia de los enfoques basados en IA en la ciberseguridad, superando ampliamente a los métodos tradicionales, como las Listas Negras y Listas Blancas y enfoques basados en Similitudes [20], [30], que, aunque son rápidos, mostraron variabilidad en su efectividad, especialmente ante ataques inmediatos o de "hora cero". Por otro lado, los métodos basados en IA demostraron una notable capacidad para identificar estos ataques con alta precisión, como evidencia la excepcional precisión lograda por XGBoost [17]. Estos resultados tienen implicaciones significativas, subrayando la importancia de adoptar enfoques basados en IA en la detección de ataques de ingeniería social en entornos de ciberseguridad.

Una limitante significativa a esta investigación es la diversificación en la aplicación de estos algoritmos, estos fueron aplicados en su mayoría en la detección de sitios web de phishing [1], [2], [3], [4], [5], [6], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [27], [28], [29], [30], [31], [32]. Además, se reconoce la necesidad de una revisión adicional para fortalecer el respaldo de XGBoost, en contraste con el algoritmo RF, que destaca como uno de los más precisos en la mayoría de los estudios que lo mencionan, XGBoost solo cuenta con el respaldo favorable de uno de los cuatro estudios en los que se le menciona. Esta diferencia en el respaldo subraya la importancia de una evaluación más detallada y exhaustiva para garantizar la validez y robustez de los resultados obtenidos mediante el uso de XGBoost en la detección de ataques de ingeniería social.

V. CONCLUSIONES

En síntesis, esta investigación se enfocó en resaltar los principales algoritmos basados en inteligencia artificial para la detección y prevención de ataques de ingeniería social. Dentro de la gran variedad de ataques de ingeniería social, se observó que el phishing fue el tipo de ataque más abordado en los estudios. Asimismo, tras analizar los estudios abordados, se llegó a la conclusión, con una precisión del 99.84 %, de que el algoritmo más eficaz para detectar el phishing fue XGBoost. No obstante, otros algoritmos también demostraron una alta capacidad en esta detección, como Random Forest (RF) con un 99.74% de precisión, FastText-CBOW combinado con RF obteniendo un 99.5% de precisión, Support Vector Machine (SVM) con un 99%, y SVM con Doc2vec, también con una precisión del 99%.

Estas herramientas demostraron consistentemente su capacidad para identificar y mitigar ataques de phishing. Sin embargo, es crucial resaltar que los estudios analizados durante esta RSL destacaron por la aplicación de dichos algoritmos en la detección de páginas web de phishing, si bien estos algoritmos pueden considerarse seguros para la detección de ataques de ingeniería social, aún se requiere una aplicación más amplia en diversos contextos y tipos de ataques, como smishing, vishing, entre otros, los cuales todavía no cuentan con un respaldo suficiente por los estudios para considerarse completamente contrarrestados por los algoritmos anteriormente mencionados.

Por consiguiente, es importante señalar que para futuras investigaciones se recomienda enfocar la efectividad de estos algoritmos en la detección de otros tipos de ataques de ingeniería social u otros contextos en su aplicación, como la identificación de correos electrónicos fraudulentos o mensajes de texto maliciosos, entre otros contextos. Ampliar el análisis hacia estos frentes permitirá comprender mejor la verdadera capacidad y alcance de estos algoritmos en la defensa contra una gama más extensa de ataques de ingeniería social

AGRADECIMIENTO

A la Universidad Tecnológica del Perú, por incentivar la práctica investigativa para un mejor país.

REFERENCIAS

[1] R. M. Abdulla, H. A. Faraj, C. O. Abdullah, A. H. Amin, and T. A. Rashid, “Analysis of Social Engineering Awareness among Students and Lecturers,” *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3311708.

[2] L. R. Kalabarige, R. S. Rao, A. R. Pais, and L. A. Gabralla, “A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites,” *IEEE Access*, vol. 11, pp. 71180–71193, 2023, doi: 10.1109/ACCESS.2023.3293649.

[3] L. Tang and Q. H. Mahmoud, “A Deep Learning-Based Framework for Phishing Website Detection,” *IEEE Access*, vol. 10, pp. 1509–1521, 2022, doi: 10.1109/ACCESS.2021.3137636.

[4] S. Asiri, Y. Xiao, S. Alzahrani, S. Li, and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," *IEEE Access*, vol. 11, pp. 6421–6443, 2023, doi: 10.1109/ACCESS.2023.3237798.

[5] F. Charmet, T. Morikawa, and T. Takahashi, "Toward a Better Understanding of Mobile Users' Behavior: A Web Session Repair Scheme," *IEEE Access*, vol. 10, pp. 99931–99943, 2022, doi: 10.1109/ACCESS.2022.3206402.

[6] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing Detection System Through Hybrid Machine Learning Based on URL," *IEEE Access*, vol. 11, pp. 36805–36822, 2023, doi: 10.1109/ACCESS.2023.3252366.

[7] N. Tsinganos, I. Mavridis, and D. Gritzalis, "Utilizing Convolutional Neural Networks and Word Embeddings for Early-Stage Recognition of Persuasion in Chat-Based Social Engineering Attacks," *IEEE Access*, vol. 10, pp. 108517–108529, 2022, doi: 10.1109/ACCESS.2022.3213681.

[8] M. Hammad, N. Hewahi, and W. Elmedany, "T-SNERF: A novel high accuracy machine learning approach for Intrusion Detection Systems," *IET Inf Secur*, vol. 15, no. 2, pp. 178–190, 2021, doi: 10.1049/ise2.12020.

[9] S. A. A. Bokhari and S. Myeong, "The Influence of Artificial Intelligence on E-Governance and Cybersecurity in Smart Cities: A Stakeholder's Perspective," *IEEE Access*, vol. 11, pp. 69783–69797, 2023, doi: 10.1109/ACCESS.2023.3293480.

[10] W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward, "The well-built clinical question: a key to evidence-based decisions.," *ACP J Club*, vol. 123, no. 3, 1995.

[11] H. Robles *et al.*, "Language learning apps for visually impaired users: a systematic review," *Res Pract Technol Enhanc Learn*, vol. 19, 2024, doi: 10.58459/rptel.2024.19012.

[12] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites," *IEEE Access*, vol. 8, pp. 142532–142542, 2020, doi: 10.1109/ACCESS.2020.3013699.

[13] M. Somesha and A. R. Pais, "Classification of Phishing Email Using Word Embedding and Machine Learning Techniques," *Journal of Cyber Security and Mobility*, vol. 11, no. 3, pp. 279–320, 2022, doi: 10.13052/jcsm2245-1439.1131.

[14] T. Sutter, A. S. Bozkir, B. Gehring, and P. Berlich, "Avoiding the Hook: Influential Factors of Phishing Awareness Training on Click-Rates and a Data-Driven Approach to Predict Email Difficulty Perception," *IEEE Access*, vol. 10, pp. 100540–100565, 2022, doi: 10.1109/ACCESS.2022.3207272.

[15] S. Mishra and D. Soni, "DSmishSMS-A System to Detect Smishing SMS," *Neural Comput Appl*, vol. 35, no. 7, pp. 4975–4992, 2023, doi: 10.1007/s00521-021-06305-y.

[16] W. A. Al-Khater, S. Al-Maadeed, A. A. Ahmed, A. S. Sadiq, and M. K. Khan, "Comprehensive review of cybercrime detection techniques," *IEEE Access*, vol. 8, pp. 137293–137311, 2020, doi: 10.1109/ACCESS.2020.3011259.

[17] E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa, and C. G. Duque, "The Answer is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering," *IEEE Access*, vol. 8, pp. 223529–223547, 2020, doi: 10.1109/ACCESS.2020.3043396.

[18] X. Liu and J. Fu, "SPWalk: Similar Property Oriented Feature Learning for Phishing Detection," *IEEE Access*, vol. 8, pp. 87031–87045, 2020, doi: 10.1109/ACCESS.2020.2992381.

[19] M. Sanchez-Paniagua, E. F. Fernandez, E. Alegre, W. Al-Nabki, and V. Gonzalez-Castro, "Phishing URL Detection: A Real-Case Scenario Through Login URLs," *IEEE Access*, vol. 10, pp. 42949–42960, 2022, doi: 10.1109/ACCESS.2022.3168681.

[20] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," *IEEE Access*, vol. 11, pp. 18499–18519, 2023, doi: 10.1109/ACCESS.2023.3247135.

[21] M. Sameen, K. Han, and S. O. Hwang, "PhishHaven - An Efficient Real-Time AI Phishing URLs Detection System," *IEEE Access*, vol. 8, pp. 83425–83443, 2020, doi: 10.1109/ACCESS.2020.2991403.

[22] E. S. Vishva and D. Aju, "Phisher Fighter: Website Phishing Detection System Based on URL and Term Frequency-Inverse Document Frequency Values," *Journal of Cyber Security and Mobility*, vol. 11, no. 1, pp. 83–104, 2022, doi: 10.13052/jcsm2245-1439.1114.

[23] A. Raza, K. Munir, M. S. Almutairi, and R. Sehar, "Novel Class Probability Features for Optimizing Network Attack Detection with Machine Learning," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3313596.

[24] W. Ali and S. Malebary, "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection," *IEEE Access*, vol. 8, pp. 116766–116780, 2020, doi: 10.1109/ACCESS.2020.3003569.

[25] L. R. Kalabarige, R. S. Rao, A. Abraham, and L. A. Gabralla, "Multilayer Stacked Ensemble Learning Model to Detect Phishing Websites," *IEEE Access*, vol. 10, pp. 79543–79552, 2022, doi: 10.1109/ACCESS.2022.3194672.

[26] A. Dimitriadis, E. Lontzetidis, B. Kulvatunyou, N. Ivezic, D. Gritzalis, and I. Mavridis, "Fronesis: Digital Forensics-Based Early Detection of Ongoing Cyber-Attacks," *IEEE Access*, vol. 11, pp. 728–743, 2023, doi: 10.1109/ACCESS.2022.3233404.

[27] E. S. Gualberto, R. T. De Sousa, T. P. B. De Vieira, J. P. C. L. Da Costa, and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," *IEEE Access*, vol. 8, pp. 76368–76385, 2020, doi: 10.1109/ACCESS.2020.2989126.

[28] A. H. H. Kabla, M. Anbar, S. Manickam, and S. Karupayah, "Eth-PSD: A Machine Learning-Based Phishing Scam Detection Approach in Ethereum," *IEEE Access*, vol. 10, pp. 118043–118057, 2022, doi: 10.1109/ACCESS.2022.3220780.

[29] Z. Azam, M. M. Islam, and M. N. Huda, "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," *IEEE Access*, vol. 11, pp. 80348–80391, 2023, doi: 10.1109/ACCESS.2023.3296444.

[30] S. Ariyadasa, S. Fernando, and S. Fernando, "Combining Long-Term Recurrent Convolutional and Graph Convolutional Networks to Detect Phishing Sites Using URL and HTML," *IEEE Access*, vol. 10, pp. 82355–82375, 2022, doi: 10.1109/ACCESS.2022.3196018.

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

[32] M. Mimura, "An Improved Method of Detecting Macro Malware on an Imbalanced Dataset," *IEEE Access*, vol. 8, pp. 204709–204717, 2020, doi: 10.1109/ACCESS.2020.3037330.