

Automatic phishing detection methods in corporate email systems: a systematic review

Mendoza-Durand Ricardo Hubert¹; Christian Manuel Rojas Ruidias²; Yapo-Cáceres Carlos³
Universidad Tecnológica del Perú SAC, Perú, U19212539@utp.edu.pe; U20306784@utp.edu.pe; C18236@utp.edu.pe

Abstract– *Phishing through email is a growing threat to corporate cybersecurity, where traditional detection methods frequently fail against sophisticated attacks. This review aims to identify and evaluate automatic phishing detection methods (based on Deep Learning and Machine Learning) in contrast with traditional systems, focusing on email analysis and their application in corporate and organizational environments. Method: The review was structured using the PICOC methodology to formulate the research question. A systematic process based on PRISMA was applied, conducting searches in key databases such as Scopus and Web of Science. Articles published between 2021 and 2025 were selected, and after applying inclusion/exclusion criteria, 21 studies were thoroughly reviewed. Results: The findings demonstrate the algorithmic superiority of advanced methods, such as Deep Learning (DL) models and Transformer-based approaches (LLMs), which outperform traditional and shallow ML techniques. These models consistently achieved accuracies above 99% in classification tasks. Conclusions: The study identifies a critical operational gap: this high accuracy is limited to simulated environments and lacks evidence of validation in real corporate settings, such as Office 365. It is concluded that the practical applicability of the most accurate methods remains uncertain, requiring future research to focus on validation in production environments and on mitigating the risk of secondary threats such as ransomware.*

Keywords– *Phishing detection, Machine Learning, Deep Learning, Corporate Sector, Ransomware.*

Una revisión sistemática: métodos de detección automática de phishing en correos electrónicos en el sector corporativo

Mendoza-Durand Ricardo Hubert¹ ; Christian Manuel Rojas Ruidias² ; Yapo-Cáceres Carlos³ 
Universidad Tecnológica del Perú SAC, Perú, U19212539@utp.edu.pe; U20306784@utp.edu.pe; C18236@utp.edu.pe

Resumen— El phishing en correos electrónicos es una amenaza creciente para la seguridad informática en empresas, donde los métodos tradicionales de detección frecuentemente fallan ante ataques sofisticados. Esta revisión tiene como objetivo identificar y evaluar los métodos de detección automática de phishing (basados en Aprendizaje Profundo y Aprendizaje Automático) en contraste con los sistemas tradicionales, enfocados en el análisis de correos electrónicos y su aplicación en el sector corporativo y organizacional. **Método:** La revisión se estructuró a partir de la metodología PICOC para formular la pregunta de investigación. Se aplicó un flujo sistemático basado en PRISMA, realizando búsquedas en bases de datos clave como Scopus y Web of Science. Los artículos fueron seleccionados en un periodo de 2021 a 2025 y, tras aplicar criterios de inclusión/exclusión, se revisaron exhaustivamente 21 estudios. **Resultados:** Los hallazgos demuestran la superioridad algorítmica de los métodos avanzados, tales como los modelos de Deep Learning (DL) y Transformadores (LLM), los cuales superan a los enfoques tradicionales y de ML superficial. Estos modelos alcanzaron Accuracies consistentemente por encima del 99% en la clasificación. **Conclusiones:** El estudio identifica una brecha operacional crítica: esta alta precisión se limita a entornos simulados y carece de evidencia de validación en escenarios corporativos reales, como Office 365. Se concluye que la aplicabilidad práctica de los métodos más precisos sigue siendo incierta, lo que requiere que la investigación futura se enfoque en la validación en producción y la mitigación del riesgo de propagación de amenazas secundarias como el ransomware.

Palabras clave— Detección de Phishing, Aprendizaje Automático, Deep Learning, Sector Corporativo, Ransomware.

I. INTRODUCCIÓN

El phishing por correo electrónico es una de las amenazas más persistentes en el ámbito corporativo, debido a que explota factores humanos (confianza, urgencia y descuido) para obtener credenciales, información financiera o acceso a sistemas críticos. Reportes recientes evidencian su alta incidencia y su impacto transversal en múltiples sectores, lo que refuerza la necesidad de mecanismos de detección más eficaces en entornos organizacionales [1]. En respuesta, la investigación ha transitado desde filtros tradicionales basados en reglas y listas negras hacia enfoques impulsados por inteligencia artificial, particularmente aprendizaje automático, deep learning y técnicas de procesamiento de lenguaje natural, capaces de capturar patrones complejos y adaptarse a variantes cada vez más personalizadas [2].

No obstante, persisten limitaciones relevantes: gran parte de los trabajos prioriza la validación experimental sobre datasets específicos y métricas de clasificación, mientras que aborda de forma insuficiente condiciones críticas de despliegue en

empresas, como escalabilidad, adaptabilidad frente a tácticas cambiantes (p. ej., spear phishing y BEC) e integración con infraestructuras de seguridad existentes [3]. Además, las revisiones previas tienden a cubrir subconjuntos del problema (solo deep learning, amenazas generales o modalidades específicas), sin una síntesis integral centrada en el contexto corporativo [3]. Por ello, este trabajo presenta una Revisión Sistemática de la Literatura sobre detección automática de phishing en correos corporativos, con el objetivo de identificar, clasificar y evaluar enfoques basados en ML/DL frente a métodos tradicionales, considerando desempeño y factores de implementación.

II. METODOLOGÍAS

A. Metodología PICOC

Se empleó la metodología PICOC para estructurar de manera precisa y orientada la búsqueda de información dentro de la Revisión Sistemática de la Literatura (RSL). Con esta metodología, se formuló una pregunta de investigación clara y específica que integra los cinco componentes: Población/Problema, Intervención, Comparador, Resultado y Contexto, facilitando la identificación de información relevante y coherente con los objetivos del estudio. Esta aproximación fortalece la estrategia de búsqueda, garantiza la pertinencia de los estudios seleccionados y contribuye a un proceso de revisión más riguroso y fundamentado en evidencia.

1) Pregunta PICOC:

¿Qué métodos automáticos, en comparación con enfoques tradicionales como filtros basados en reglas o listas negras, se han utilizado en el sector corporativo para mejorar la precisión y reducir los falsos positivos en la detección de ataques de phishing en correos electrónicos?

2) Cuadro PICOC:

TABLA I

PALABRAS ELEMENTALES SELECCIONADAS PARA EL MARCO PICOC.

P (Población/Problema)	Ataques de phishing presentes en correos electrónicos
I (Intervención)	Métodos automáticos para la detección de phishing.
C (Comparación)	Métodos tradicionales basados en filtros de reglas o listas negras.
O (Resultado)	Precisión y reducción de falsos positivos.
C (Contexto)	Entornos corporativos o empresariales.

3) Ecuación de Búsqueda

La búsqueda se ejecutó en Scopus y Web of Science. Se utilizó una ecuación estructurada en cinco bloques: (i) phishing/email phishing/spoofing/credential theft, (ii) detección automática e IA/ML/DL/NLP, (iii) métodos tradicionales (rule-based/blacklist/filtering), (iv) métricas de evaluación y (v) contexto corporativo/enterprise. En Scopus se aplicó en TITLE-ABS-KEY; en WoS se aplicó una ecuación equivalente en TS, conservando los mismos términos y operadores lógicos.

Scopus

(TITLE-ABS-KEY):("phishing" OR "phishing attack*" OR "phishing detection" OR "email phishing" OR "phishing email*" OR "phishing message*" OR "email scam*" OR "email fraud*" OR "fraud email*" OR "malicious email*" OR "malicious message*" OR spoof* OR "email spoof*" OR "credential theft*" OR "identity theft" OR "account compromise") AND TITLE-ABS-KEY("automatic detection" OR "intelligent detection" OR "machine learning" OR "deep learning" OR algorithm* OR model* OR "data mining" OR "pattern recognition" OR NLP OR "natural language processing" OR transformer* OR BERT OR RoBERTa OR GPT OR LLM* OR CNN OR LSTM OR RNN OR ANN OR "neural network*" OR GNN OR "graph neural network*" OR "feature extraction" OR "feature selection" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "gradient boosting" OR XGBoost OR LightGBM OR CatBoost OR "ensemble learning" OR meta-learning OR "anomaly detection" OR "intrusion detection" OR "email classification") AND TITLE-ABS-KEY("rule*" OR "rule-based" OR heuristic* OR "signature-based" OR "pattern-based" OR blacklist* OR blocklist* OR filter* OR "spam filter*" OR "email filter*" OR "security policy" OR "security rule*" OR "policy enforcement" OR "traditional method*" OR "conventional method*" OR "manual review" OR "analyst-based") AND TITLE-ABS-KEY("accuracy" OR "precision" OR "recall"

OR "f1-score" OR "false positive*" OR "false negative*" OR "error rate*" OR "detection rate*" OR effectiveness OR efficiency OR performance OR robustness OR reliability OR evaluation OR "benchmark*") AND TITLE-ABS-KEY("enterprise" OR corporat* OR business OR organization* OR company OR firm OR SME OR "small business" OR "mid-size business" OR workplace OR "enterprise environment" OR "business environment" OR "business email" OR "corporate email" OR "enterprise email" OR "email security" OR "Office 365" OR "Google Workspace" OR "Business Email Compromise" OR BEC OR "IT infrastructure" OR "corporate network" OR "enterprise network")

WoS (TS):

Se utilizó la misma ecuación, reemplazando TITLE-ABS-KEY(...) por TS=(...) en cada bloque, manteniendo idénticos términos, comillas, comodines y conectores lógicos.

5) Criterios de Inclusión y Exclusión:

TABLA II
CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN APLICADOS EN LA SELECCIÓN DE ESTUDIOS.

Criterios de inclusión	Criterios de exclusión
CI1. Incluir estudios que aborden la detección de phishing en correos electrónicos maliciosos o fraudes por email, en entornos empresariales u organizacionales (por ejemplo, empresas, corporaciones, pymes o plataformas como Office 365 y Google Workspace).	CE1. Excluir estudios que no traten específicamente la detección de phishing o fraude por correo electrónico, o que aborden los ciberataques de manera general sin énfasis en correos o enlaces.
CI2: Estudios que implementen, describan o comparen métodos de detección automáticos (machine learning, deep learning, NLP, ensemble learning, rule-based, heurísticos, etc.).	CE2: Publicaciones anteriores a 2021.
CI3: Estudios que reporten resultados empíricos con métricas de desempeño (accuracy, precision, recall, f1-score, tasa de falsos positivos, etc.).	CE3: Publicaciones en idiomas distintos al inglés.
CI4: Investigaciones que ofrezcan enfoques, hallazgos o aplicaciones con potencial utilidad para la detección de phishing en el ámbito corporativo u organizacional, independientemente del contexto en que hayan sido desarrolladas.	CE4: Documentos que no sean artículos de revista (excluye papers de congreso, tesis, reportes técnicos, capítulos de libro y material no indexado).
CI5: Estudios que detallen la implementación técnica de los métodos de detección (por ejemplo, arquitectura del modelo, características analizadas o conjunto de datos utilizado), más allá	CE5: Artículos que se centren únicamente en aspectos legales, de concienciación de usuarios o políticas de seguridad, sin proponer ni evaluar un método técnico de

de solo mencionar resultados.	detección.
-------------------------------	------------

B. Metodología PRISMA

Se siguió el flujo PRISMA para la identificación, selección e inclusión de estudios. La búsqueda se realizó en Scopus y Web of Science, aplicando la ecuación definida. Se recuperaron 301 registros en Scopus y 70 en WoS. Posteriormente, se aplicaron filtros (2021–2025, idioma inglés, áreas de Ingeniería/Ciencias de la Computación, artículos científicos y acceso abierto), obteniéndose 41 registros en Scopus y 35 en WoS. Tras la eliminación de duplicados, se consolidaron 48 artículos únicos. El cribado por título y resumen seleccionó 24 estudios para evaluación a texto completo (sin pérdidas por disponibilidad). Finalmente, tras aplicar los criterios de inclusión/exclusión, 3 estudios fueron excluidos y 21 se incluyeron en la síntesis final (Fig. PRISMA).

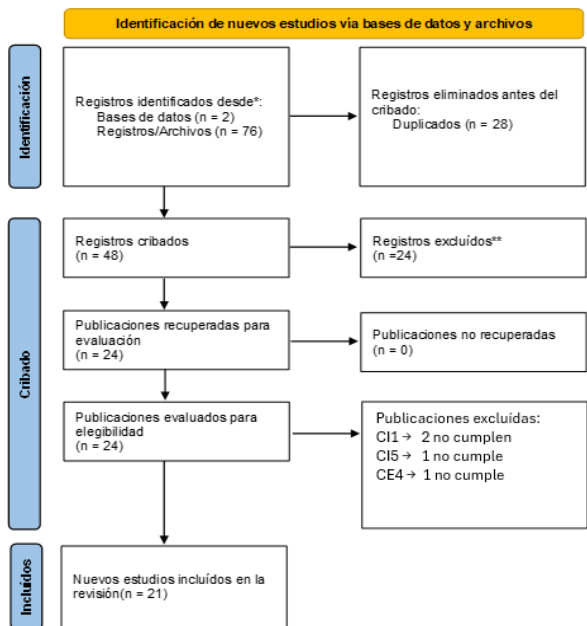


Fig. 1 Diagrama de flujo PRISMA para la selección de estudios. Es importante señalar que, entre los filtros aplicados durante la fase de identificación, se incluyó la restricción a artículos de acceso Abierto (open access). Esta decisión responde a criterios de reproducibilidad y accesibilidad de la revisión, pero constituye una limitación del estudio, dado que potencialmente excluye investigaciones relevantes publicadas en revistas de acceso restringido. En consecuencia, los hallazgos de esta RSL reflejan el estado del conocimiento disponible en la literatura de acceso abierto indexada en Scopus y Web of Science.

III. RESULTADOS

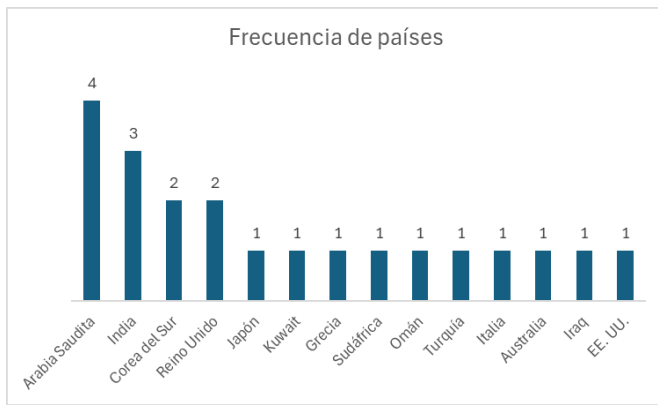
A. Análisis exploratorio de fuentes

Para esta revisión sistemática de la literatura (RSL), se

seleccionaron un total de 21 artículos publicados entre el año 2021 y 2025. Las investigaciones se originaron en los siguientes países: Arabia Saudita, Corea del Sur, India, Japón, Kuwait, Grecia, Sudáfrica, Omán, Turquía, Reino Unido, Italia, Australia, Iraq y EE. UU.

TABLA III.
PAÍSES DE LAS FUENTES REVISADAS Y RELEVANCIA SEGÚN H-INDEX

Referencia	H-Index	País
Alkhodhairy, G. et al. [14]	1	Arabia Saudita
Mun, H. et al. [15]	1	Corea del Sur
Alhuzali, A. et al. [16]	5	Arabia Saudita
Hatkar, S. et al. [17]	1	India
Meléndez, R. et al. [18]	1	Japón
Al Shalabi, L. et al. [19]	4	Kuwait
Roumeliotis, K.I. et al. [20]	10	Grecia
Altwaijry, N. et al. [21]	10	Arabia Saudita
Diko, Z. et al. [22]	1	Sudáfrica
Abdul Samad, A.S. et al. [23]	8	Omán
Alshahrani, H.J. et al. [24]	5	Arabia Saudita
al-Tamimi, Y. et al. [25]	1	Turquía
Ji, K. et al. [26]	1	Corea del Sur
Occhipinti, A. et al. [27]	10	Reino Unido
Somesha, M. et al. [28]	3	India
Butnaru, A. et al. [29]	1	Reino Unido
Gallo, L. et al. [30]	5	Italia
Karim, A. et al. [31]	28	Australia
Mohammed, M.A. et al. [32]	57	Iraq
Mohamed, G. et al. [33]	1	India
Jamal, S. et al. [34]	3	EE. UU.



La Fig 2. muestra el número de artículos por país. Arabia Saudita lidera con 4 publicaciones, seguida por India con 3 artículos. Corea del Sur y Reino Unido contribuyen con 2 publicaciones cada uno. Japón, Kuwait, Grecia, Sudáfrica, Omán, Turquía, Italia, Australia, Iraq y EE. UU. presentan una publicación cada uno.

La Fig. 3 presenta los términos más recurrentes identificados en los artículos analizados. Los conceptos con mayor frecuencia fueron machine learning (11 apariciones), phishing (9 apariciones) y deep learning (6 apariciones), evidenciando el predominio de enfoques basados en aprendizaje automático para la detección de amenazas. Asimismo, se destacan términos como electronic mail y machine-learning (6 y 5 apariciones, respectivamente), junto con categorías clave como cybersecurity, decision trees, malware y support vector machines (4 apariciones cada una).

RQ1: “¿Cómo se define y aborda el problema del phishing en entornos corporativos dentro de los estudios recientes?”

Autor (Año)	Enfoque o definición aplicada
Alkhodhairy, G. et al. (2025)	Aborda la detección de emails sospechosos/phishing contra la evolución de amenazas que superan las seguridades tradicionales.
Mun, H. et al. (2025)	Se enfoca en el email de phishing que busca robar información sensible o dirigir a enlaces maliciosos; destaca la ineficacia contra el spear-phishing.
Alhuzali, A. et al. (2025)	Define el phishing como la ciberamenaza más crítica y origen de la mayoría de brechas vía emails de ingeniería social.
Hatkar, S. et al. (2025)	Aborda el phishing como ingeniería social que engaña para revelar información sensible, usando emails, sitios web falsos o SMS.
Meléndez, R. et al. (2024)	El foco es la clasificación binaria de email (phishing vs. no phishing), ante la incapacidad de herramientas existentes para detectar nuevas técnicas.
Al Shalabi, L. et al. (2024)	Define el ataque por emails engañosos y sitios web fraudulentos que explotan la vulnerabilidad humana, abordándose con clasificación binaria.
Roumeliotis, K.I. et al. (2024)	Se aborda la clasificación de spam en correos, donde el phishing es la principal amenaza a mitigar por esta solución.
Altwaijry, N. et al. (2024)	Define el phishing como ataque de ingeniería social por emails de phishing, abordado como tarea de clasificación binaria de documentos.
Diko, Z. et al. (2024)	Aborda las URLs maliciosas como vehículo de ciberataques, incluyendo explícitamente el phishing como una de estas amenazas.
Abdul Samad, A.S. et al. (2023)	Se enfoca en el phishing URL como la táctica principal para ataques masivos, que expone a la organización al uso inadvertido.
Alshahrani, H.J. et al. (2023)	Aborda el aumento de emails maliciosos (phishing y spam) que buscan persuadir para la divulgación de datos personales y credenciales.
al-Tamimi, Y. et al. (2023)	Se centra en la detección de sitios web de phishing para obtener información

	sensible (IDs, contraseñas), señalando la insuficiencia de herramientas existentes.
Ji, K. et al. (2023)	Aborda el filtrado de email malicioso, siendo el phishing la principal causa de daño personal y reducción de eficiencia laboral en negocios.
Occhipinti, A. et al. (2022)	Define el phishing como un método popular derivado del spam que roba información personal, credenciales o detalles bancarios mediante mensajes maliciosos.
Somesha, M. et al. (2022)	El enfoque es el email de phishing (spam especial) con enlaces a sitios suplantados para robar información personal/credenciales.
Butnaru, A. et al. (2021)	Se centra en la detección de phishing basado en URL por interacción social, debido a la ineficacia de los denylists contra la naturaleza dinámica de las URLs.
Gallo, L. et al. (2021)	Aborda la detección de spam malicioso y emails phishing ante la constante evolución de amenazas que dificultan los filtros de operador (carrier-grade).
Karim, A. et al. (2021)	Aborda los ataques de spam (incluyendo phishing spam) que buscan el robo de identidad y el pillaje de datos sensibles.
Mohammed, M.A. et al. (2021)	Se centra en el filtrado de email malicioso que puede contener enlaces a sitios web con phishing, ante la disminución de efectividad de filtros basados en texto.
Mohamed, G. et al. (2022)	Define el phishing como uno de los mayores crímenes dirigido a organizaciones para el robo de datos sensibles mediante sitios web de phishing.
Jamal, S. et al. (2024)	Aborda el phishing como una amenaza persistente de ingeniería social que sirve frecuentemente como punto de entrada para ataques de ransomware.

C. Métodos automáticos aplicados para la detección de phishing

RQ2: ¿Qué métodos automáticos se han aplicado para la detección de phishing en contextos corporativos?

La evidencia muestra predominio de enfoques supervisados: (i) ML clásico con features de URL/cabeceras, (ii) DL (CNN/LSTM/GRU) sobre contenido, y (iii) Transformers (BERT/RoBERTa) cuando el foco es variación lingüística. Los ensembles aparecen como estrategia de robustez, mientras que LLMs/zero-shot se exploran como tendencia emergente. En general, la comparación entre estudios es limitada por heterogeneidad de datasets y protocolos, por lo que las métricas reportadas deben interpretarse con cautela..

TABLA V
MÉTODOS AUTOMÁTICOS UTILIZADOS

Autor (Año)	Tipo de método automático utilizado
Alkhodhairy et al. (2025)	Machine Learning (ML) supervisado + Natural Language Processing (NLP)
Mun et al. (2025)	Deep Learning (DL) e híbrido (DL + reglas/firmas)
Alhuzali et al. (2025)	Machine Learning (ML) y Deep Learning (DL)
Hatkar et al. (2025)	Ensemble Learning y Machine Learning (ML)
Meléndez et al. (2024)	Deep Learning (Transformers) y ML supervisado
Al Shalabi & Jazyah (2024)	Híbrido (Clustering + Clasificación con DL y Decision Tree)
Roumeliotis et al. (2024)	Deep Learning (DL), Large Language Models (LLM) y NLP
Altwaijry et al. (2024)	Deep Learning (DL) basado en 1D-CNN y variantes recurrentes
Diko & Sibanda (2024)	Machine Learning (ML) / Ensemble Learning sobre URLs
Abdul Samad et al. (2023)	Híbrido (NLP + Machine Learning + Deep Learning)
Alshahrani et al. (2023)	Deep Learning (DL) + optimización metaheurística (IFFO-SRRNN)
al-Tamimi & Shkoukani (2023)	Híbrido (Clustering K-Means + ML supervisado con DT/RF)
Ji & Kwon (2023)	Machine Learning (ML) supervisado (Naïve Bayes) sobre Hadoop MapReduce
Occhipinti et al. (2022)	Machine Learning (ML) supervisado + NLP (pipeline con 12 clasificadores)
Somesha & Pais (2022)	ML supervisado + NLP / Word Embedding (FastText, TF-IDF, Count Vectorizer)
Butnaru et al. (2021)	Machine Learning (ML) supervisado ligero basado en URL
Gallo et al. (2021)	Machine Learning (ML) supervisado (Support Vector Machine y Random Forest)
Karim et al. (2021)	Aprendizaje no supervisado (Clustering: OPTICS, DBSCAN, K-means, etc.)

Mohammed et al. (2021)	ML supervisado (Naive Bayes) en sistema multi-agente
Mohamed et al. (2022)	ML supervisado (Neural Network, SVM, Random Forest)
Jamal et al. (2024)	Deep Learning (DL) / LLM basado en Transformers (IPSDM, DistilBERT, RoBERTa)

D. Desempeño comparativo entre métodos automáticos y tradicionales

RQ4: ¿Qué métricas de desempeño reportan los estudios para evaluar la efectividad de los métodos automáticos frente a los tradicionales?

La Tabla VIII evidencia que predominan métricas de clasificación (Accuracy, Precision, Recall y F1), con uso ocasional de AUC-ROC y tasas de error (FPR/FNR). La comparabilidad entre estudios es limitada por diferencias de dataset, particiones y definición de phishing, por lo que los valores deben interpretarse como desempeño “dentro del estudio” más que como ranking universal.

TABLA VI
MÉTRICAS DE DESEMPEÑO

Autor (Año)	Datos/métricas del método automático	Datos del método tradicional o baseline	Evidencia de desempeño comparativo
Alkhodhairy & Saleem, 2025	Accuracy = 98.65%	No se reportan métricas numéricas	El modelo ML+NLP supera a los sistemas convencionales de detección de spam
Mun et al. (PhiShield), 2025	LSTM1000 Acc = 0.9855, Prec = 0.9795, Recall = 0.9909, F1 = 0.9852, BERT Acc = 0.9903, Prec = 0.9926, Recall = 0.9875, F1 = 0.9900	Naive Bayes Acc = 0.7049, F1 = 0.7228, FCNN Acc = 0.8681, F1 = 0.8599	LSTM y sobre todo BERT superan claramente a NB y FCNN, BERT es el más preciso aunque más costoso computacionalmente

Alhuzali et al., 2025	Deep Learning Accuracy máx = 99.81% (Zenodo), Random Forest = 99.63%	Naive Bayes Accuracy máx = 98.05%	Los modelos DL son claramente superiores al ML superficial, RF es el mejor algoritmo de ML tradicional
Hatkar et al. (UTM), 2025	Ensemble ML Acc = 97.0%, AUC-ROC = 0.996, FP = 2.3%, FN = 3.1%, latencia ≈ 4.8 ms	Fortinet UTM Acc = 94.5%, latencia ≈ 5–7 ms	El UTM propuesto es más preciso y más rápido que Fortinet UTM en entornos empresariales
Meléndez et al., 2024	RoBERTa Acc = 0.9943, F1 = 0.9939, Prec = 0.9961, Recall = 0.9917	SVM (mejor ML) Acc = 0.9897, F1 = 0.9880	Los modelos Transformer superan a los modelos ML tradicionales, RoBERTa ofrece el mejor rendimiento
Al Shalabi & Jazyah, 2024	ScC-K-mean (DT, DS1) Acc = 100%, ScC-K-mean (DL, DS1) Acc = 99.37%, Prec = 98.90%, Recall = 100%, Fm = 99.45%, AUC = 100%, RS-K-mean (DL) Acc = 99.53%, AUC = 100%	Mejor benchmark literatura 98.95% (PSO+BP NN), RS (DL) sin clustering 94.33%	El clustering K-means tras la selección de características eleva la Accuracy desde ≈92–94% hasta ≈99–100%
Roumeliotis et al., 2024	RoBERTa (LLM fine-tuned) Acc = 99.78%, F1 = 0.9978	Random Forest Acc = 97.66%	Los LLM fine-tuned logran precisión mucho mayor que los modelos NLP/ML tradicionales y CNN

Altwaijry et al., 2024	1D-CNNPD + Bi-GRU Acc = 99.68%, Prec = 100%, Recall = 99.32%, F1 = 99.66%, AUC ≈ 99.85%	Modelos previos ML RF ≈ 97%, Bayes Net ≈ 92%, modelo DL DeepAnti-PhishNet ≈ 97.9%	El modelo propuesto supera a ML tradicional y a otros modelos DL avanzados (DeepAnti-PhishNet, THEMIS)
Diko & Sibanda, 2024	Random Forest Acc = 96%, F1 (phishing) = 0.96, Prec = 0.98, Recall = 0.94, XGB Acc = 95%, LGB Acc = 95%	No se reportan métricas numéricas de listas negras o heurísticas	Los ensembles de ML muestran alta capacidad para detectar URLs maliciosas y se proponen como alternativa superior a listas negras/heurísticas
Abdul Samad et al., 2023	DT+TF-IDF (D2 phishing) Acc = 99.5%, RF+Count = 99.4%, ANN+Embeddings = 99.2%	Sin métricas numéricas para métodos tradicionales, solo discusión de listas negras y heurísticas	Las combinaciones NLP+ML/DL logran accuracies cercanas al 100% en detección de dominios maliciosos
Alshahrani et al., 2023	IFFO-SRRNN Acc = 98.86%, Prec phishing = 94.29%, Recall phishing = 100%, F-score phishing = 97.06%	SVM Acc = 83.46%, Linear Model 83.25%, Naive Bayes 59.67%	El modelo SRRNN optimizado supera ampliamente a todos los modelos ML comparados
al-Tamimi & Shkoukani, 2023	RF con descomposición de clase (4 clases, FS) Acc = 100%, DT con descomposición de clase Acc = 100%	RF baseline con FS Acc = 97.1%, RF baseline sin FS 96.9%	La descomposición de clase mediante clustering aumenta la precisión desde ≈97% hasta 100%
Ji & Kwon, 2023	Hadoop MapReduce Naive Bayes Acc = 99.28%, error de predicción = 0.72%	Python Naive Bayes en servidor bare-metal Acc = 89.83%, error = 10.17%	Naive Bayes sobre Hadoop es 1.11 veces más preciso y reduce la tasa de error ≈14 veces frente a la versión tradicional
Occhipinti et al., 2022	XGBoost F-score = 94%, Prec = 94%, Recall = 93%, AUC = 0.98, MPNN F-score = 92%, tiempo ≈ 0.007 s	Multinomial NB F-score = 85%, AUC = 0.93, Linear SVM F-score = 89%, AUC = 0.95	El pipeline con preprocesamiento NLP y optimización mejora de forma notable el F-score respecto a NB y SVM
Somesha & Pais, 2022	RF + FastText Acc = 99.50%, FPR = 0.053%, RF + TF-IDF Acc = 99.39%, RF + Count Acc = 99.18%	Sin métricas numéricas de listas negras/heurísticas	RF + FastText es la mejor combinación, con accuracy muy alta y tasa de falsos positivos extremadamente baja
Butnaru et al., 2021	Random Forest optimizado Acc = 99.29%, F-measure = 98.22, Prec = 97.40, Recall = 99.06, accuracy temporal media ≈ 94%	Google Safe Browsing accuracy media ≈ 45% sobre URLs activas	El modelo ML más que duplica la accuracy de GSB y mantiene buen rendimiento meses después del entrenamiento
Gallo et al., 2021	SVM Prec = 95.2%, Recall = 91.6%, RF con resultados similares	No se reportan métricas cuantitativas de los filtros carrier-grade	El conjunto de features optimizado con SVM/RF mejora la detección y permitió rediseñar el proceso de gestión de amenazas

Karim et al., 2021	OPTICS balanced accuracy media $\approx 75.76\%$, purity media $\approx 82.12\%$, F-score spam ≈ 0.86 , accuracy de clase $\approx 85.81\%$	K-means balanced accuracy $\approx 66.74\%$	OPTICS y DBSCAN generan clusters de mejor calidad que K-means y Spectral, OPTICS es el mejor algoritmo evaluado
Mohammed et al., 2021	Mejor run (6) Acc = 98.4% , FP = 0.08% , FN = 2.90% , Prec = 99% , F-measure = 97.65% , Recall = 96.9%	Fette et al. 2007 Prec = 98.92% , F-measure = 97.64% , Recall = 96.38%	El modelo multi-agente propuesto mejora ligeramente precisión y recall frente al benchmark clásico de Fette
Mohamed et al., 2022	Neural Network Acc = 95.18% , SVM = 85.45% , RF = 78.89%	No hay método tradicional adicional, SVM y RF son baselines	La NN es el modelo con mejor rendimiento y adecuada para detectar ataques de phishing tipo zero-day
Jamal et al. (IPSDM), 2024	IPSDM/ROBERTA Acc = 99.00% , F1 = 0.985 , IPSDM/DistilBERT Acc = 97.10% , F1 = 0.9710	ROBERTA base Acc = 93.29% , DistilBERT base Acc = 88.95%	El fine-tuning de IPSDM mejora consistentemente a los modelos base, con hasta 5.71 puntos extra de Accuracy

F. Entornos empresariales y plataformas corporativas analizadas

RQ5: “¿En qué tipos de entornos corporativos, plataformas empresariales o infraestructuras se evalúan los métodos de detección de phishing y cuáles son sus características?”

La literatura evaluada analiza la detección de phishing principalmente en entornos de simulación (datasets/benchmarks y configuraciones controladas), y solo de forma excepcional en escenarios corporativos reales. Para mayor claridad, se clasificó cada estudio en tres categorías: (a) simulación pura, (b) entorno híbrido con mención de aplicabilidad corporativa pero evaluación sobre datos controlados, y (c) validación real en entornos operativos. En lo referente a la pregunta: como muestra la Tabla VII, 16 de 21 estudios corresponden a simulación pura, 3 a entornos híbridos y únicamente 2 —Gallo et al. [30] y Hatkar et al. [17]— reportan validación empresarial real. La Tabla VII resume los entornos/plataformas reportados por los estudios incluidos.

TABLA VII
ENTORNO O PLATAFORMA ESTUDIADA

Autor (Año)	Entorno o plataforma estudiada
Alkhodhairi, G. et al. (2025)	Simulación (benchmark dataset).
Mun, H. et al. (2025)	Red de usuario local (extensión de navegador) / Compatible con Gmail y Outlook.
Alhuzali, A. et al. (2025)	Simulación (datasets públicos).
Hatkar, S. et al. (2025)	Redes empresariales / Sistemas UTM (PyMES).
Meléndez, R. et al. (2024)	Simulación / Comparación algorítmica.
Al Shalabi, L. et al. (2024)	Simulación / Plataforma ML open-source H2O.
Roumeliotis, K.I. et al. (2024)	Simulación (SpamAssassin y Enron).
Altwaijry, N. et al. (2024)	Simulación / Instalación en Edge Network Operation (ENO) de correo.
Diko, Z. et al. (2024)	Simulación / Detección de URLs a nivel de la red de la organización.
Abdul Samad, A.S. et al. (2023)	Simulación / Detección de URLs en el ciberespacio.

Alshahrani, H.J. et al. (2023)	Simulación / Aplicabilidad en uso comercial.
al-Tamimi, Y. et al. (2023)	Simulación / Soluciones anti-phishing.
Ji, K. et al. (2023)	Entorno Distribuido (Hadoop) / Filtrado de correo electrónico masivo.
Occhipinti, A. et al. (2022)	Simulación / Corpus de spam empresarial (Enron).
Somesha, M. et al. (2022)	Corpus interno en tiempo real / Solución anti-phishing para organizaciones.
Butnaru, A. et al. (2021)	Simulación / Aplicabilidad en un entorno de producción real.
Gallo, L. et al. (2021)	Validación empírica en empresa grande (large company) / Flujo de correo corporativo.
Karim, A. et al. (2021)	Simulación / Sistemas de IA en entornos de negocios.
Mohammed, M.A. et al. (2021)	Simulación / Sistema Multi-Agente (JADE) / Comunicación corporativa.
Mohamed, G. et al. (2022)	Simulación / Desarrollo algorítmico.
Jamal, S. et al. (2024)	Simulación / Integración en chatbot, aplicaciones web y ciberdefensa.

IV. DISCUSIÓN

Esta sección interpreta los hallazgos y resalta una tensión central en la literatura: la brecha entre eficacia algorítmica y adopción operativa en entornos corporativos. Aunque se reportan desempeños muy altos (p. ej., *Accuracy* >99%) en la clasificación de correos de phishing [16], [18], [21], [34], estos resultados provienen mayoritariamente de evaluaciones en entornos controlados y datasets estandarizados [14], [16], [18], [20], [22], [23], [29], con evidencia limitada de

validación en escenarios empresariales reales [30]. En consecuencia, rara vez se evalúan variables críticas para despliegue corporativo, como escalabilidad, latencia e interoperabilidad con plataformas específicas (p. ej., Office 365), lo que deja la aplicabilidad práctica insuficientemente demostrada.

En términos metodológicos, se observa un avance consistente hacia enfoques de Deep Learning, especialmente Transformers (BERT/RoBERTa) y arquitecturas recurrentes, que tienden a mostrar mejoras frente a enfoques más convencionales en múltiples estudios [15], [17], [19], [30]. De forma complementaria, algunos trabajos sugieren que la infraestructura (p. ej., procesamiento distribuido) puede mejorar el rendimiento de ciertos modelos [26] y que, en escenarios puntuales, enfoques avanzados superan mecanismos tradicionales de filtrado [29]. Sin embargo, persiste un desajuste entre el foco predominante (clasificación de correos/URLs) y el rol del phishing como vector de entrada a incidentes de mayor impacto, como el ransomware [34].

El análisis cruzado de entornos (Tabla VII) revela que aunque 19 de 21 estudios declaran relevancia coportativa, solo Gallo et al. [30] y Hatkar et al. [17] reportan validación en escenarios empresariales reales; los demás operan sobre datasets públicos, por lo que las accuracies 99% [16], [18], [21], [23] deben interpretarse como rendimiento en condiciones controladas. Asimismo, se identifica una taxonomía de tres generaciones: *MI Clásico* sobre features de URL/cabeceras [22, 25, 29]; *Deep learning* (CNN, LSTM, GRU) [15], [21], [24] y *Transformes* [18], [20], [34] cuya superioridad no es unívoca; Butnaru et al. [29] alcanzan 99.29% con Random Forest ligero frente al [98].86% dl SRRNN de Alsharani et al. [24], lo que indica que la elección óptima depende del context operative.

La principal limitación de esta RSL es la escasez de estudios que detallen la implementación y el desempeño operativo en producción, agravada por la restricción a literatura de acceso Abierto. En consecuencia la investigación futura debería priorizar: (i) validación empírica en condiciones reales con métricas de operación (p. ej., latencia y costos), (ii) integración y evaluación específica con suites corporativas ampliamente usadas (p. ej., Office 365) y (iii) diseño de estudios que midan el impacto preventivo de la detección de phishing sobre la reducción efectiva de incidentes encadenados, particularmente ransomware, que solo un estudio ha abordado[34] de forma explícita

V. CONCLUSIONES

Esta Revisión Sistemática identificó, clasificó y evaluó métodos de detección automática de phishing en correos electrónicos en contexto corporativo, sintetizando su desempeño reportado y sus condiciones de evaluación. En conjunto, la evidencia indica que los enfoques basados en Deep Learning, particularmente Transformers, han alcanzado un alto grado de madurez técnica y suelen reportar desempeños elevados en tareas de clasificación frente a alternativas más convencionales. No obstante, los resultados

también evidencian una brecha operacional: la mayoría de estudios valida en entornos simulados y con datasets públicos, con escasa demostración de desempeño en escenarios corporativos reales y cobertura limitada de factores críticos de adopción, como escalabilidad, latencia e integración con plataformas empresariales (p. ej., Office 365/Google Workspace).

Asimismo, la literatura sugiere que el phishing debe entenderse no solo como robo de credenciales, sino como vector inicial de incidentes de mayor impacto (p. ej., ransomware), por lo que la evaluación no debería limitarse a métricas de clasificación aisladas. La principal limitación identificada es la falta de estudios que reporten métricas operacionales y costos de despliegue en producción. En consecuencia, trabajos futuros deberían priorizar validación empírica en entornos reales con métricas como latencia y TCO, integración específica vía APIs corporativas y métricas que capturen el impacto extendido de estas soluciones en la contención de amenazas encadenadas.

Referencias

- [1] C. Opara, P. Modesti, and L. Golightly, "Evaluating spam filters and stylometric detection of AI-generated phishing emails," *Expert Syst. Appl.*, vol. 276, 2025, Art. no. 127044, doi: 10.1016/j.eswa.2025.127044.
- [2] S. S. Shafin, "An explainable feature selection framework for web phishing detection with machine learning," *Data Sci. Manag.*, vol. 8, pp. 127–136, 2025, doi: 10.1016/j.dsm.2024.08.004.
- [3] H. Mun, J. Park, Y. Kim, B. Kim, and J. Kim, "PhiShield: An AI-based personalized anti-spam solution with third-party integration," *Electronics*, vol. 14, no. 8, Art. no. 1581, 2025, doi: 10.3390/electronics14081581.
- [4] N. Odeh, D. Eleyan, and A. Eleyan, "Enhancing web security through machine learning-based detection of phishing websites," *Int. J. Comput. Netw. Inf. Secur.*, vol. 17, no. 1, pp. 39–56, 2025, doi: 10.5815/ijcnis.2025.01.04
- [5] R. Purnamadewi *et al.*, "Enhancing detection of zero-day phishing email attacks in the Indonesian language using deep learning algorithms," *Bull. Electr. Eng. Inform.*, vol. 14, no. 1, 2025, doi: 10.11591/eei.v14i1.8759
- [6] Y. Yuan, G. Apruzzese, and M. Conti, "Beyond the west: Revealing and bridging the gap between Western and Chinese phishing website detection," *Comput. Secur.*, vol. 148, 2025, Art. no. 104115, doi: 10.1016/j.cose.2024.104115.
- [7] M. Murhej and G. Nallasivan, "Multimodal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-based EAI-SC-LSTM," *Frontiers in Communications and Networks*, vol. 6, Art. no. 1587654, 2025, doi: 10.3389/frcmn.2025.15876542025
- [8] M. Murhej and G. Nallasivan, "Multimodal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-BASED EAI-SC-LSTM," *Front. Commun. Netw.*, vol. 6, 2025, Art. no. 1587654, doi: 10.3389/frcmn.2025.1587654.
- [9] X. Chen, "Hybrid Phishing Detection Using Stochastic Gradient Descent and Naïve Bayes Optimized with the Mayfly Algorithm," *IJCAI*, vol. 49, no. 21, May 2025.
- [10] L. Al-Shalabi and Y. H. Jazyah, "Phishing detection using clustering and machine learning," *IAES Int. J. Artif. Intell.*, vol. 13, no. 4, pp. 4526–4536, Dec. 2024, doi: 10.11591/ijai.v13.i4.pp4526-4536.
- [11] R. Jayaprakash, K. Natarajan, J. A. Daniel, C. V. Chinnappan, J. Giri, H. Qin, and S. Mallik, "Heuristic machine learning approaches for identifying phishing threats across web and email platforms," *Front. Artif. Intell.*, vol. 7, 2024, Art. no. 1414122, doi: 10.3389/frai.2024.1414122.
- [12] K. S. Jishnu and B. Arthi, "Real-time phishing URL detection framework using knowledge distilled ELECTRA," *Automatika*, vol. 65, no. 4, pp. 1621–1639, 2024, doi: 10.1080/00051144.2024.2415797.
- [13] E. Sangra, R. Agrawal, P. R. Gundalwar, K. Sharma, D. Bangri, and D. Nandi, "Malicious website detection using random forest and Pearson correlation for effective feature selection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, 2024, doi: 10.14569/IJACSA.2024.0150876.
- [14] G. Alkhodhairi and K. Saleem, "Machine learning algorithm for detecting suspicious email messages using Natural Language Processing (NLP)," *Alexandria Eng. J.*, vol. 128, pp. 153–165, Sep. 2025, doi: 10.1016/j.aej.2025.04.067
- [15] H. Mun, J. Park, Y. Kim, B. Kim, and J. Kim, "PhiShield: An AI-Based Personalized Anti-Spam Solution with Third-Party Integration," *Electronics*, vol. 14, no. 8, 2025, doi: 10.3390/electronics14081581.
- [16] A. Alhuzali, A. Alloqmani, M. Aljabri, and F. Alharbi, "In-depth analysis of phishing email detection: Evaluating the performance of machine learning and deep learning models across multiple datasets," *Applied Sciences*, vol. 15, no. 6, p. 3396, 2025, doi: 10.3390/app15063396.
- [17] S. Hatkar, K. Rout, A. Lad, and G. Swathika, "An integrated UTM solution for modern cybersecurity: Combining deep inspection, ML, and policy automation," *IEEE Access*, vol. 2025, pp. 1–12, 2025, doi: 10.1109/ACCESS.2025.3618281.
- [18] R. Meléndez, M. Ptaszynski, and F. Masui, "Comparative investigation of traditional machine-learning models and transformer models for phishing email detection," *Electronics*, vol. 13, no. 24, 2024, doi: 10.3390/electronics13244877.
- [19] L. Al-Shalabi and Y. H. Jazyah, "Phishing detection using clustering and machine learning," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 4, pp. 4526–4536, 2024, doi: 10.11591/ijai.v13.i4.pp4526-4536.
- [20] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification," *Electronics*, vol. 13, no. 11, 2024, doi: 10.3390/electronics13112034.
- [21] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models," *Sensors*, vol. 24, no. 7, 2024, doi: 10.3390/s24072077.
- [22] Z. Diko and K. Sibanda, "Comparative Analysis of Popular Supervised Machine Learning Algorithms for Detecting Malicious Universal Resource Locators," *Journal of Cyber Security and Mobility*, vol. 13, no. 5, 2024, doi: 10.13052/jcsm2245-1439.13513.
- [23] A. S. Saleem Raja, G. Pradeepa, S. Mahalakshmi, and M. S. Jayakumar, "Natural language based malicious domain detection using machine learning and deep learning," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 23, no. 2, pp. 304–312, 2023, doi: 10.17586/2226-1494-2023-23-2-304-312.
- [24] H. J. Alshahrani *et al.*, "Improved Fruitfly Optimization with Stacked Residual Deep Learning Based Email Classification," *Intelligent Automation and Soft Computing*, 2023, doi: 10.32604/iasc.2023.034841.
- [25] Y. al-Tamimi and M. Shkoukani, "Employing cluster-based class decomposition approach to detect phishing websites using machine learning classifiers," *International Journal of Data and Network Science*, 2023, doi: 10.5267/ijdns.2022.10.002.
- [26] K. Ji and Y. Kwon, "New Spam Filtering Method with Hadoop Tuning-Based MapReduce Naïve Bayes," *Computer Systems Science and Engineering*, 2023, doi: 10.32604/csse.2023.031270
- [27] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Expert Systems with Applications*, 2022, doi: 10.1016/j.eswa.2022.117193
- [28] M. Somesha and A. R. Pais, "Classification of Phishing Email Using Word Embedding and Machine Learning Techniques," *Journal of Cyber Security and Mobility*, 2022, doi: 10.13052/jcsm2245-1439.1131.
- [29] A. Butnaru, A. Mylonas, and N. Pitropakis, "Towards lightweight URL-based phishing detection," *Future Internet*, 2021, doi: 10.3390/fi13060154.
- [30] L. Gallo, A. Maiello, A. Botta, and G. Ventre, "2 Years in the anti-phishing group of a large company," *Computers & Security*, 2021, doi: 10.1016/j.cose.2021.102259.
- [31] A. Karim, S. Azam, B. Shanmugam, and K. Krishnan, "An Unsupervised Approach for Content-Based Clustering of Emails into Spam and Ham through Multilingual Feature Formulation," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3116128.
- [32] M. A. Mohammed, D. A. Ibrahim, and A. O. Salman, "Adaptive intelligent learning approach based on visual anti-spam email model for multi-natural language," *Journal of Intelligent Systems*, 2021, doi: 10.1515/jisys-2021-0045.
- [33] G. Mohamed, J. Visumathi, M. Mahdal, J. Anand, and M. Elangovan, "An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approach," *Processes*, 2022, doi: 10.3390/pr10071356.
- [34] S. Jamal, H. Wimmer, and I. H. Sarker, "An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach," *Security and Privacy*, 2024, doi: 10.1002/spy2.402.