UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE COMPUTACIÓN

TEMA:
REVISIÓN SISTEMÁTICA DE LA LITERATURA RELACIONADA CON CIBERSEGURIDAD APOYADA CON ANÁLISIS DE BIG DATA PARA ACTIVIDADES DE RED TEAM

Trabajo de titulación previo a la obtención del Título de:
Ingeniero en Ciencias de la Computación

AUTORES:
BRYAN STEEVEN QUEZADA HERRERA
DEBBY MELANY LEÓN YAGUANA

TUTOR:
JOSE LUIS AGUAYO MORALES

Quito, Ecuador
2022

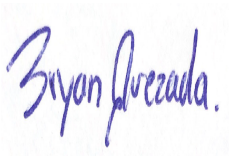
CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Bryan Steeven Quezada Herrera con documento de identificación N° 1725901431 y Debby Melany León Yaguana con documento de identificación N° 1724947757; manifestamos que:

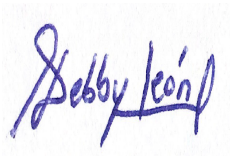
Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 13 de Septiembre del año 2022

Atentamente,



Bryan Steeven Quezada Herrera
1725901431



Debby Melany León Yaguana
1724947757

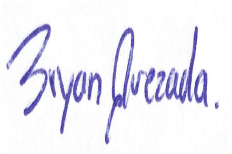
CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

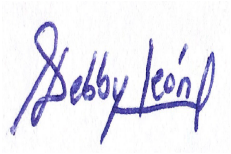
Nosotros, Bryan Steeven Quezada Herrera con documento de identificación N° 1725901431 y Debby Melany León Yaguana con documento de identificación N° 1724947757, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Artículo Académico: ”Revisión Sistemática de la Literatura relacionada con Ciberseguridad apoyada con Análisis de Big Data para Actividades de Red Team”, el cual ha sido desarrollado para optar por el título de: Ingeniero en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 13 de Septiembre del año 2022

Atentamente,





Bryan Steeven Quezada Herrera
1725901431

Debby Melany León Yaguana
1724947757

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Jose Luis Aguayo Morales con documento de identificación N° 1709562597, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: REVISIÓN SISTEMÁTICA DE LA LITERATURA RELACIONADA CON CIBERSEGURIDAD APOYADA CON ANÁLISIS DE BIG DATA PARA ACTIVIDADES DE RED TEAM, realizado por Bryan Steeven Quezada Herrera con documento de identificación N° 1725901431 y Debby Melany León Yaguana con documento de identificación N° 1724947757, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 13 de Septiembre del año 2022

Atentamente,

Ing. Jose Luis Aguayo Morales. MSc
1709562597

Revisión Sistemática de la Literatura relacionada con ciberseguridad apoyada con Análisis de Big Data para actividades de Red Team

1st Bryan Steeven Quezada Herrera
Universidad Politécnica Salesiana
Quito - Ecuador
bquezadah@est.ups.edu.ec

2nd Debby Melany León Yaguana
Universidad Politécnica Salesiana
Quito - Ecuador
dleony@est.ups.edu.ec

3rd Jose Luis Aguayo Morales
Universidad Politécnica Salesiana
Quito - Ecuador
jaguayo@ups.edu.ec

***Resumen*—Esta investigación revisó la literatura existente durante el período 2017 a 2021 sobre ciberseguridad respaldada por análisis de Big Data, atestiguando el uso de técnicas para la detección de ciberataques. Las metodologías de estudio utilizadas fueron Mapeo Sistemático y Revisión Sistemática de Literatura en artículos científicos, cuyo objetivo es sugerir algunas actividades del equipo rojo para verificar la ciberseguridad. En cada artículo académico se identificaron anomalías de ciberseguridad, técnicas de análisis de Big Data, algoritmos de clasificación, métricas y resultados de desempeño de los algoritmos. Durante la investigación se identificaron 13 anomalías de ciberseguridad, 10 técnicas de Análisis de Big Data, 16 algoritmos de clasificación y 15 métricas de desempeño. Las anomalías de ciberseguridad con mayor incidencia en los estudios fueron Phishing con 41,94%, Ataques de Red con 12,9%, Malware con 9,68% e Ingeniería Social con 6,45% cada una. Finalmente, para comprobar la ciberseguridad con análisis Big Data, se recomienda un análisis Red Team sobre las anomalías de mayor incidencia utilizando armas de código abierto, como: SET, METASPLOIT, DLL y OWASP ZAP.**

***Palabras Clave*—Big Data, Ciberseguridad, Red Team, Mapeo Sistemático, Revisión Sistemática de Literatura.**

***Abstract*—This research reviewed the existing literature during the period 2017 to 2021 on Cybersecurity supported by Big Data Analysis, evidencing the use of techniques to detect cyberattacks. The study methodologies used were Systematic Mapping and Systematic Literature Review in scientific papers, in order to suggest some Red Team activities to check the Cybersecurity. In each academic article were identified Cybersecurity anomalies, Big Data Analysis techniques, classification algorithms, metrics and performance results of the algorithms. During the investigation were identified 13 Cybersecurity anomalies, 10 Big Data Analysis techniques, 16 classification algorithms and 15 performance metrics. The Cybersecurity anomalies with the highest incidence in the studies were Phishing with 41.94%, Network Attacks with 12.9%, Malware with 9.68% and Social Engineering with 6.45% each one. Finally, to check Cybersecurity with Big Data Analysis, its recommend a Red Team analysis on the highest incidence anomalies using open source weapons, like: SET, METASPLOIT, DLL and OWASP ZAP.**

***Keywords*—Big Data, Cybersecurity, Red Team, Systematic Mapping, Systematic Literature Revision.**

I. Introducción

En la actualidad las compañías se fortalecen con nuevas herramientas tecnológicas que, además de asegurar la información, proporcionan algunos beneficios como: forjar el crecimiento de las organizaciones, minimizar tiempos de reacción en los procesos, resolución de futuros problemas, priorización y optimización de los canales de comunicación en apoyo de la toma de decisiones beneficiosas para la empresa. Sin embargo, la falta de protocolos de seguridad trae consigo que sean víctimas de constantes ciberataques [1]. Es por esa razón que se ha catalogado como necesidad primordial a la ciberseguridad, cuyos propósitos son asegurar y resguardar la información y datos de las organizaciones en contra de cualquier tipo de ataque tecnológico ocasionado por delincuentes cibernéticos.

En 2021, el informe de Investigaciones sobre Violación de Datos de Verizon (DBIR) expuso los incidentes de ciberseguridad del año, la edición incluye 88 países, 83 colaboradores, 79.635 incidentes y 5.258 violaciones de datos, convirtiéndolo en un informe global con información valiosa, en el año 2021 el informe disminuyó el número de incidentes analizados, sin embargo, se mostró un aumento en la cantidad de filtraciones de datos en comparación con el año 2020, con una diferencia de 1.308 filtraciones confirmadas, dando un total de 5.258 filtraciones en el año 2021. El resumen de violación de datos obtuvo los siguientes resultados: el 85% de las infracciones involucraron un elemento humano, el 13% de los incidentes que no son Ataque de Denegación de Servicios involucraron Ransomware y el 3% de las filtraciones involucraron la explotación de vulnerabilidades[2].

Con el avance tecnológico, la ciberseguridad es más propensa a nuevos ataques cibernéticos tales como: malware, phishing, robo de credenciales, suplantación de identidad, denegación de servicio, ataque de red de protocolo, etc. Por este motivo se emplea el uso de la analítica de Big Data para verificar fuentes de datos grandes y erradicar estos altercados, creando una falsa seguridad [3].

Esta investigación complementó la información existente en el período del 2017 al 2021 sobre la ciberseguridad apoyada con técnicas de Big Data, problema que genera inseguridad remanente que es evidenciada por el equipo de Red Team.

Para tener una visión clara respecto al empleo del análisis de datos en ciberseguridad se aplicaron las estrategias de Mapeo Sistemático y Revisión Sistemática de Literatura (cuyas siglas

en inglés son SM y SLR respectivamente). El SM (Systematic Mapping) se conforma de tres etapas o secciones, mientras que la SLR (Systematic Literature Revision) se utilizó para identificar y relacionar los estudios primarios de la investigación.

Las principales aportaciones que generó este trabajo se mencionan a continuación:

i) Se realizó el SM y la SLR basándose en una clasificación para estructurar la literatura existente en un conjunto para el estudio de Big Data en ciberseguridad.
ii) Se han identificado técnicas especiales de análisis de big data (ABD) para reconocer anomalías de ciberseguridad.
iii) Esta investigación mostro a la ciberseguridad apoyada en el estudio de Big Data y recomendo actividades de Red Team.
iv) Al analizar la ciberseguridad, la investigación extrajo las anomalías más frecuentes y postuló diversos ataques de Red Team que podrían burlar a las técnicas antes mencionadas.

II. METODOLOGÍA

En la presente investigación se emplearon los métodos mencionados en párrafos anteriores. El SM permite la identificación, clasificación y sintetización de las investigaciones sobre el análisis de Big Data en ciberseguridad [4]. Por su parte, la revisión sistemática, mediante un proceso de recopilación y canalización crítica de estudios, permitió proporcionar un resumen completo de la literatura existente y verificar cuál es el estado actual de las investigaciones de ciberseguridad con relación a Big Data [5] en el período 2017 a 2021. Gracias a la Revisión Sistemática se define a la ciberseguridad como la manera de proteger, salvaguardar y mantener segura la información obtenida y almacenada en el ciberespacio [6]. Big Data es la compilación de un número significativo de datos, que se manejan a través de sistemas especializados para magnificar el rendimiento y la capacidad de los mismos, con el objetivo de obtener beneficios de dicha información en la toma de decisiones [7].

En la metodología del estado del arte se implementaron tres etapas que se organizaron de la siguiente forma: En la etapa 1 se definió al método PICO (ver Tabla I) que ayudo a considerar una estrategia de búsqueda bibliográfica y específica. Además, en esta etapa se mencionaron los criterios de elegibilidad utilizados en el estado del arte. En la etapa 2 se establecieron dos subsecciones para plantear las preguntas de investigación y determinar las tácticas de búsqueda que se aplicaron en los repositorios mencionados y así conseguir estudios correctamente alineados al propósito de este estudio. Finalmente, en la etapa 3 se realizó una síntesis de todos los datos recopilados para clasificar las diferentes categorías de los estudios reportados y conseguir una SM y SLR efectivos.

A. Etapa 1

En esta etapa se definió al método PICO como una táctica de investigación bibliográfica que aporta a determinar los criterios de elegibilidad de trabajos, estudios, investigaciones y artículos [8], con esto en consideración se definieron los elementos PICO. (Ver Tabla I).

TABLE I: ESTRUCTURACIÓN DEL MÉTODO PICO

PICO	**Considerations**
Population (P)	Ciberseguridad apoyada en Big Data.
Intervention (I)	Técnicas de análisis de Big Data usadas en ciberseguridad.
Comparision (C)	Estudios que publicaron análisis de Big Data utilizados en ciberseguridad.
Outcomes (O)	Actividades de Red Team que vulneren la ciberseguridad apoyada con Big Data.

Posteriormente, se instauraron las técnicas utilizadas en la estructura de cadenas de búsqueda. (ver Tabla II). Para obtener dicha cadena se emplearon expresiones booleanas adecuadas y conocidas tales como “AND” u “OR”, incluso aquellas que organizaron esto de la siguiente manera: (”Cybersecurity” OR ”Big Data” OR ”Red Team”) AND (”Cyber Safety” OR ”IT Security” OR ”Attack Protection” AND ”Data Science” OR ”Big Data Analytics” AND ”Ethical Hacking” OR ”Social Engineering” OR ”Pentest”).

TABLE II: TÉRMINOS PARA UTILIZACIÓN DE BÚSQUEDA

Términos	**Términos Semejantes**
Cybersecurity	Cyber Safety, IT Security, Attack Protection
Big Data	Data Science, Big Data Analytics
Red Team	Ethical Hacking, Social Engineering, Pentest

*1) **Criterios de elegibilidad**:* Para elegir los estudios principales, se utilizaron los criterios de inclusión y exclusión especificados en la Tabla III que se describen a continuación:

- **Criterios de inclusión:** por definición son aquellas tipologías temporales y geográficas que componen una población de estudio [9], dentro de esto se consideraron como inclusión a artículos en formato digital online, estudios primaros, artículos con total acceso, etc. (Ver Tabla III)
- **Criterios de exclusión:** son aquellas tipologías de los sujetos que pueden obstruir con la eficacia de la información [9], por ello se consideraron como exclusión a estudios fuera del periodo de inclusión, literatura gris, estudios de diferente idioma, etc. (Ver Tabla III)

B. Etapa 2.

*1) **Preguntas de investigación**:* El eje primordial de esta investigación fue describir en la actualidad al análisis de Big Data en ciberseguridad, para lo cual se plantearon una serie de interrogantes de la investigación, para el SM se etiquetan con: (SMP#) y para SLR se etiquetan con: (SLRP#), a continuación se enumeran las preguntas que se generaron durante la investigación:

- *SMP1: ¿Existe una clasificación para el estudio de ciberseguridad apoyada con análisis de Big Data y actividades de Red Team?*

TABLE III: CRITERIOS DE ELEGIBILIDAD

Tipo	Inclusión	Exclusión
Mapeo Sistemático	Artículos en formato digital online, indexados y publicados en Scopus, IEEE, Science Direct y Web of Science en el período 2017 a 2021	Estudios publicados antes del 2017
	Idioma de publicación (inglés)	Estudios publicados en idiomas distintos al inglés.
	Estudios primarios como conferencias (avanzados) y revistas.	Estudios secundarios como revisiones, editoriales, comentarios y libros.
	Artículos con acceso al resumen y al texto completo.	Literatura gris (folletos, editoriales, artículos de opinión) y artículos científicos que solicitan un pago extra.
Revisión Sistemática de la Literatura	Artículos que contengan palabras claves incluidas en el título y resumen.	Artículos que no contengan palabras claves o términos semejantes y artículos duplicados.
	Estudios de ciberseguridad relacionados con Big Data o Data Science.	Estudios que no relacionen ciberseguridad con Big Data o Data Science o estudios sobre cómo proteger Big Data.
	Estudios que contemplen las actividades de Red Team, pentest o ethical hacking.	Estudios que no desarrollen las actividades de Red Team, pentesting o ethical hacking.

- *SMP2: ¿Cuál es la repartición de los estudios durante el período 2017 a 2021?*
- *SLRP1: ¿Cuáles son las principales vulnerabilidades de ciberseguridad que ponen en riesgo a la red de datos?*
- *SLRP2: ¿Cuáles son las técnicas de análisis de Big Data destacadas para el descubrimiento de anomalías?*
- *SLRP3: ¿Qué métodos de Red Team son aplicables para burlar el estudio de Big Data?*
- *SLRP4: ¿Qué medidas de rendimiento de análisis de Big Data se utilizan para la detección de ataques?*

*2) **Estrategias de búsqueda**:* Para la recopilación de los diferentes artículos científicos, se empleó la cadena de búsqueda en cuatro repositorios específicos: Scopus, ScienceDirect, IEEE y Web of Science (Ver Tabla IV), considerando que existan suficientes artículos que cumplan con los estándares para elaborar un análisis de SM y SLR. A continuación, se mencionan los diferentes filtros que se aplicaron para la recopilación de información:

- **Filtro 1.** Se compilaron 3273 análisis principales, en los que se emplearon los cuatro primeros criterios de inclusión y exclusión (ver Tabla III). En dicho filtro se aplicó el rango de búsqueda en el periodo 2017 al 2021, el idioma de publicación principal y estudios como conferencias y revistas. Finalmente, en este filtro se consideraron únicamente a los artículos que fueron de acceso libre y disponibles en su totalidad, como resultado se obtuvieron 1425 estudios.
- **Filtro 2.** Con ayuda del aplicativo EndNote x9, se procedió a filtrar la información mediante un ingreso de la cadena de búsqueda en título y abstract. Además, se consideró excluir a los artículos duplicados y gracias a este proceso la cantidad de estudios totales disminuyó considerablemente a 153.
- **Filtro 3.** En el tercer filtro se procedió con la lectura individual de resúmenes en los estudios recolectados, con el fin de localizar los temas importantes que sean relativos al caso tratado, esto quiere decir, que se consideraron los artículos enfocados en ciberseguridad relacionados con Big Data, obteniendo así 86 estudios.
- **Filtro 4.** En el cuarto filtro se procedió con la lectura total de los artículos recolectados, esto con la misión de aplicar la SLR y así detectar los aspectos más relevantes de los mismos para ser tabulados e interpretados, como resultado final se obtuvieron 28 estudios principales.

C. Etapa 3.

En este punto se extrajo información de los diferentes artículos recolectados y escogidos en la etapa 2, luego se estructuró y sintetizó la indagación de estudios, lo que dio como resultado una taxonomía o clasificación que se puede observar en la Fig. 1. En esta se describió la clasificación de ataques frecuentes de ciberseguridad, técnicas de análisis de datos junto a algoritmos y métricas. Finalmente, se categorizó un análisis de actividades de Red Team que podrían burlar a estas técnicas mencionadas anteriormente.

1) Anomalías de ciberseguridad: Las anomalías son intentos intencionales de acceso a cualquier sistema informático cuyo objetivo principal es interrumpir los recursos de estos para obtener información confidencial, mediante la implantación de software malicioso [10].

TABLE V: ANOMALÍAS DE ciberseguridad

Etq	Anomalía	Referencia
A01	Phishing	[11][12][13][14][15][16][17][18][19][20][21][22][23]
A02	Network Attacks	[24][25][26][27]
A03	Social Engineering	[28][29]
A04	Malware	[30][31][32]
A05	Software vulnerability	[33]
A06	Password Attack	[34]
A07	Cloud Computing Threats	[35]
A08	Web Aplication Attack	[13]
A09	Advanced Persistent Threats	[36]
A10	Repeated Frauds	[36]
A11	Attack Signature	[37]
A12	Cyber Domain Attacks	[38]
A13	Malicious Proxy Servers	[31]

TABLE IV: Cadena de búsqueda y filtros en cada repositorio

Repositorio	Cadena de búsqueda	Tipo de artículo	Resultados de cadena	Filtro 1	Filtro 2	Filtro 3
IEEE	(”Document Title”: ”Cybersecurity” OR ”Big Data” OR ”Red Team”) AND ”Abstract”: ”Cyber Safety” OR ”IT Security” OR ”Attack Protection” AND ”Data Science” OR ”Big Data Analytics” AND ”Ethical Hacking” OR ”Social Engineering” OR ”Pentest”.	Revistas y Conferencias	1608	600	58	27
Web of Science	TI=(Cybersecurity OR Big Data OR Red Team) AND AB=(cyber safety OR it security OR attack protection AND data science OR Big Data analytics AND ethical hacking OR social engineering OR pentest)	Revistas	752	228	22	15
Science Direct	Title: Cybersecurity OR Big Data OR Red Team. Title, abstract, key-words:Cyber Safety OR IT Security OR Attack Protection AND Data Science OR Big Data Analytics AND Ethical Hacking OR Social Engineering OR Pentest	Revistas	444	310	30	21
Scopus	(ALL(”Cybersecurity” OR ”Cyber Safety” OR ”IT Security” OR ”Attack Protection”) AND ALL(”Big Data” OR ”Data Science” OR ”Big Data Analytics”) AND ALL(”Red Team” OR ”Ethical Hacking” OR ”Social Engineering” OR ”Pentest”))	Revistas y Conferencias	469	287	43	23
		TOTAL	3273	1425	153	86

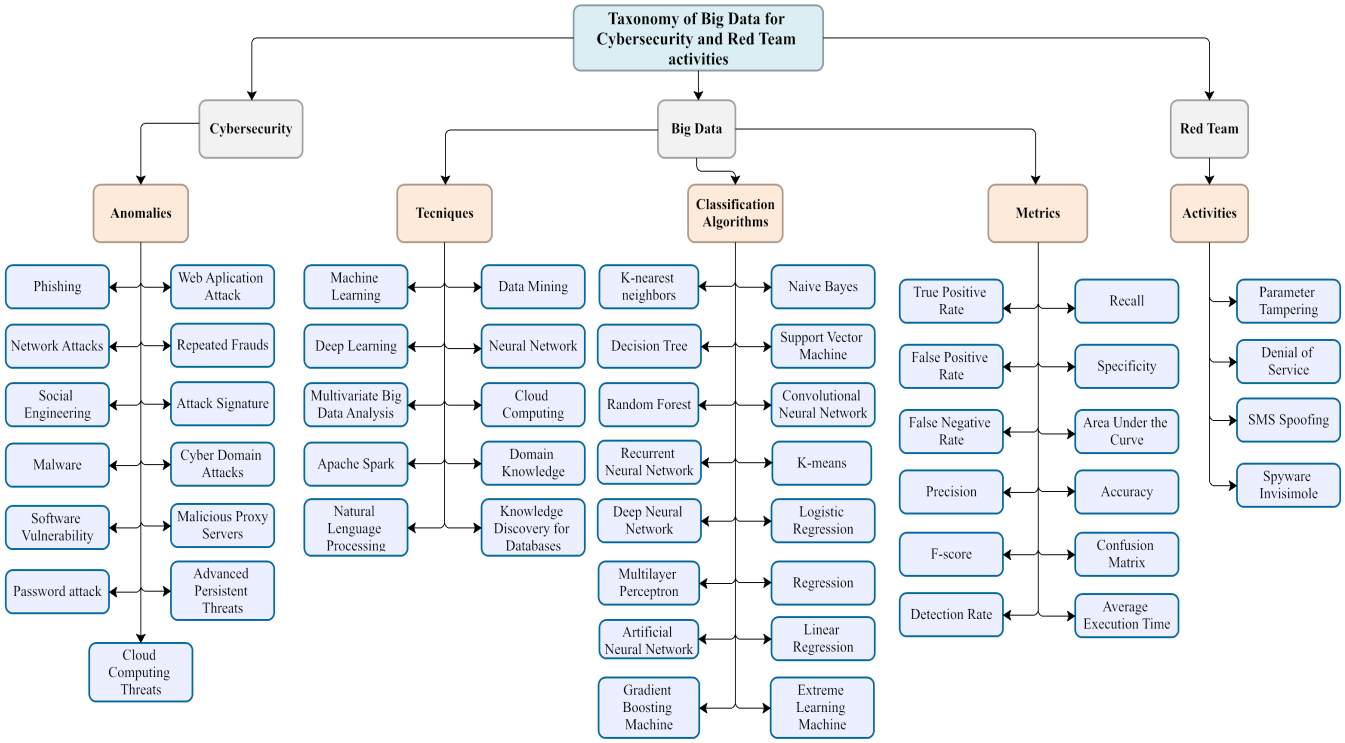


Fig. 1: Taxonomía de análisis de Big Data y actividades de Red Team en ciberseguridad

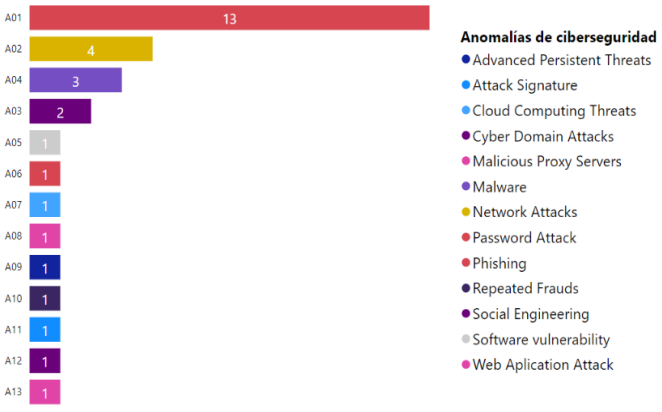


Fig. 2: Frecuencia de anomalías de ciberseguridad

Por medio de la SLR las anomalías que se identificaron en esta investigación se expusieron en la Tabla V y su recurrencia de aparición se refleja en la Fig. 2 contemplando así, las principales anomalías de ciberseguridad que afectan a un sistema informático perjudicando su funcionamiento.

2) Técnicas de ABD: Durante la aplicación de la SLR se encontraron algunas técnicas descriptivas para ABD centradas en el descubrimiento de nuevas relaciones o hechos que antes se desconocían y así poder alcanzar los objetivos planteados en las diversas investigaciones recopiladas [39].
Las técnicas de análisis encontradas en esta investigación se encuentran expuestas en la Tabla VI y su frecuencia de

aparición se refleja en la Fig. 3.

TABLE VI: TÉCNICAS DE ANÁLISIS DE Big Data

Etq	Técnica de análisis	Referencia
T01	Machine Learning (ML)	[30][35][37][26][11][12] [31][33][13][14][15][16] [32][18][19][20][21][22] [23][29][34][27]
T02	Data Mining (DM)	[14][15]
T03	Deep Learning (DL)	[38][17][21]
T04	Neural Networks (NN)	[24][28][22]
T05	Multivariate Big Data Analysis (MBDA)	[36][25]
T06	Apache Spark (AS)	[26]
T07	Knowledge Discovery for Databases (KDD)	[15]
T08	Natural Language Processing (NLP)	[29]
T09	Domain Knowledge (DK)	[31]
T10	Cloud Computing (CC)	[36][35]

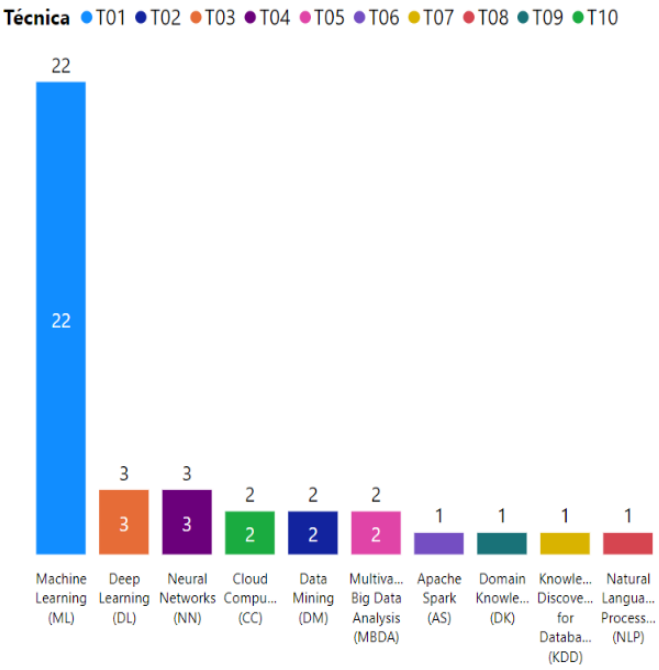


Fig. 3: Frecuencia de técnicas de ABD

3) Algoritmos de Clasificación: Para el ABD, la clasificación es el primer paso para categorizar y separar conjuntos de datos, esta se implementó en función de la regla de aprendizaje que identifica un modelo, para una estrecha relación entre un conjunto de atributos y la etiqueta [40]. En esta investigación, los algoritmos encontrados con la aplicación de la SLR y mostrados en la Tabla VII, contribuyeron a la organización de la taxonomía presentada y en la Fig. 4 se evidenció la frecuencia con la que se utilizaron los algoritmos.

TABLE VII: ALGORITMOS DE CLASIFICACIÓN

Etq	Algoritmo	Referencia
AL01	Naive Bayes (NB)	[30][37][14][18][27]
AL02	K-nearest neighbors (KNN)	[30][37][26] [21][23][29] [34]
AL03	Decision Tree (DT)	[30][11][12][28][14][17] [21][23][34]
AL04	Support Vector Machine (SVM)	[30][25][37][38][31][33] [13][16][32][17][19][20] [21][22][34]
AL05	Random Forest (RF)	[30][36][11][12][28] [33][13][15][17][18][21] [22][23][29][34][27]
AL06	Convolutional Neural Network (CNN)	[24][33][32][17]
AL07	Extreme Learning Machine (ELM)	[38]
AL08	Recurrent Neural Network (RNN)	[33]
AL09	Logistic Regression (LR)	[37][13][19][21][23][34] [27]
AL10	K-means (KM)	[35][14]
AL11	Gradient Boosting Machine (GBM)	[15][21][23]
AL12	Deep Neural Network (DNN)	[16]
AL13	Multilayer Perceptron (MLP)	[37][32][18][29][34]
AL14	Regresion (R)	[18]
AL15	Artificial Neural Network (ANN)	[37][19][21]
AL16	Linear Regression (LIR)	[22]

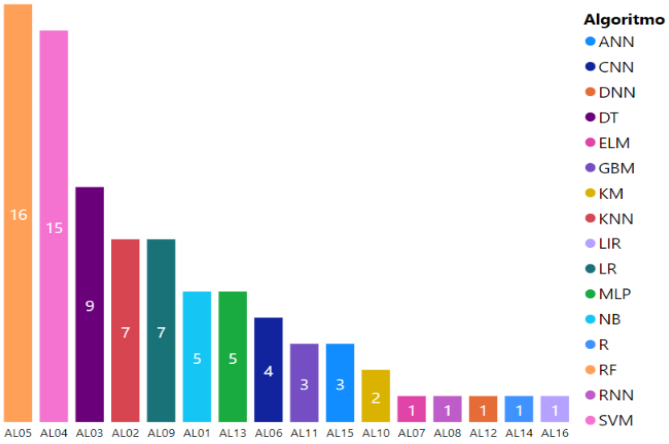


Fig. 4: Frecuencia de algoritmos de clasificación

4) Métricas de Rendimiento: Durante la aplicación de la SLR se descubrieron algunas medidas de rendimiento con las que los autores evalúan la efectividad y precisión de los algoritmos de clasificación de Big Data, se utilizaron varios modelos matemáticos para estas mediciones, que se describen en la siguiente sección. Además, en la Fig. 5 se verifica la frecuencia del uso de estas medidas de rendimiento en los estudios recolectados:

- **Confusion Matrix**: es un resumen de los resultados predichos del modelo de clasificación, se obtiene resumiendo el recuento total de predicciones clasificadas correcta e incorrectamente en función de cada clase. Es necesario derivar los siguientes valores antes de diseñar la matriz de confusión:

  - **True Positive (TP)** Los valores positivos verdaderos se refieren al número de instancias que han sido clasificadas correctamente por el modelo.
  - **True Negative (TN)** Los valores negativos verdaderos son la cantidad de casos negativos que fueron clasificados efectivamente por el modelo.
  - **False Positive Rate (FP)** El valor de los falsos positivos es el número de instancias negativas etiquetadas incorrectamente como instancias positivas.
  - **False Negative (FN)** El valor de los falsos negativos es el número de instancias positivas etiquetadas incorrectamente como instancias negativas.
- **False Positive Rate (FPR)** Denominado Fall-Out, es la proporción de instancias negativas clasificadas incorrectamente como instancias positivas. En términos más sencillos, es la probabilidad de que se produzcan falsas alarmas. El FPR se calcula mediante la siguiente ecuación:

$$FalsePositiveRate = \frac{FP}{TN + FP} \tag{1}$$

- **True Positive Rate (TPR)** Es una medida de la proporción de casos positivos en los datos que se identifican correctamente como tales. Para el cálculo de la siguiente métrica usamos la siguiente ecuación:

$$TruePositiveRate = \frac{TP}{TP + FN} \tag{2}$$

- **False Negative Rate (FPR)** Se refiere a la proporción de muestras clasificadas incorrectamente respecto al número de muestras positivas. El FNR se obtiene con la siguiente ecuación:

$$FalseNegativeRate = \frac{FN}{TP + FN} \tag{3}$$

- **Precision (P)** Es la proporción de positivos predichos que son positivos reales. Se aplica en diversas áreas, como el aprendizaje automático, la minería de datos, etc. Esta métrica se calcula mediante la siguiente ecuación:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- **F-Score (F1)** Corresponde a la media de la precisión y la recuperación. Esta métrica se calcula utilizando la siguiente ecuación:

$$F1 = 2 * \frac{P * R}{P + R} \tag{5}$$

- **Accuracy (AC)** Esta medida puede describirse como la eficacia general del modelo de clasificación. Para el cálculo de la misma, empleamos la siguiente ecuación:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

- **Specificity (S)** Esta métrica describe la eficacia del modelo de clasificación, para identificar las etiquetas negativas. Este valor se obtiene mediante la siguiente fórmula:

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

- **Recall (R)** Denominado TPR, se refiere a la proporción de casos reales positivos que se han predicho como positivos. El recall puede calcularse mediante la siguiente ecuación:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

- **Area under the curve (AUC)** Es una herramienta estadística utilizada para valorar la exactitud en la predicción de eventos binarios. Cuanto mayor sea el AUC, mejor será el rendimiento del modelo a la hora de distinguir entre las clases positivas y negativas.

$$AreaUndertheCurve = \frac{SE + SP}{2} \tag{9}$$

TABLE VIII: MÉTRICAS DE RENDIMIENTO

Etq	Métrica	Referencia
M01	True Positive Rate (TPR)	[30][26][38][14][20]
M02	False Positive Rate (FPR)	[30][38][27]
M03	False Negative Rate (FNR)	[30]
M04	Precision (P)	[30][24][36][26][11][31] [13][32][17][18][23][29] [34]
M05	F-Score (F1)	[30][24][36][26][11][28] [13][32][17][23][34]
M06	Accuracy (AC)	[30][24][36][35][11][12] [28][13][15][32][17][19] [20][21][22][29][34][27]
M07	Recall (R)	[24][36][36][26][11][13] [17][23][34]
M08	Area under the curve (AUC)	[28][15][16][29][34]
M09	True Positive (TP)	[33][16][20]
M10	False Positive (FP)	[33]
M11	True Negative (TN)	[33][16][20]
M12	False Negative (FN)	[33][20]
M13	Detection Rate (DR)	[26][27]
M14	Average execution time (AET)	[24][26]
M15	Specificity (S)	[25]

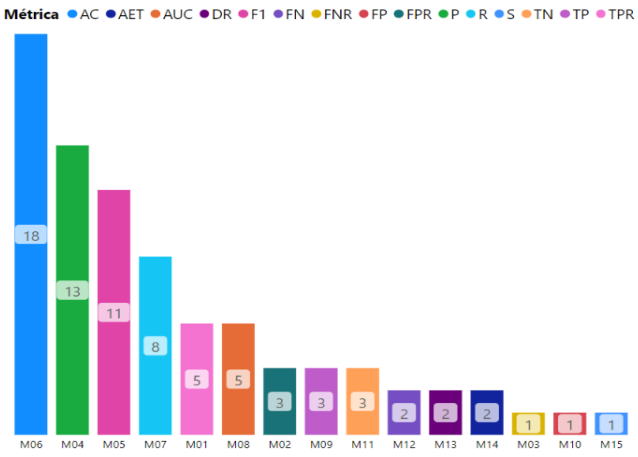


Fig. 5: Frecuencia de métricas de rendimiento en análisis de Big Data

III. RESULTADOS Y DISCUSIÓN

A. SMP1: ¿Existe una clasificación para el estudio de ciberseguridad apoyada con análisis de Big Data y actividades de Red Team?

En este artículo se propuso una taxonomía para el estudio de ciberseguridad apoyada con análisis de Big Data y actividades de Red Team. Dicha taxonomía se clasificó en tres categorías que son: Big Data, Ciberseguridad y Red Team. Además, sé sub clasifican en Anomalías, Técnicas, Algoritmos de clasificación, Métricas y actividades de Red Team, respectivamente. Debido a los limitados estudios que se han realizado con la combinación de Big Data y ciberseguridad con actividades de Red Team, se identificó la relación entre los tres dominios, basándose en estudios experimentales que se utilizaron para ciberseguridad con una combinación de tecnologías de Big Data o de Red Team. La taxonomía derivada se ha ilustrado en Fig. 1.

B. SMP2: ¿Cuál es la repartición de los estudios durante el período 2017 a 2021?

La repartición de los estudios considerados en revistas y conferencias se realizó así: en el periodo 2017 al 2021 se encontró un porcentaje de 53.57% de publicaciones en revistas, mientras que por el lado de las conferencias, el porcentaje de publicaciones fue del 46.43%. Por lo tanto, se comprobó una mayoría de publicaciones en revistas en todo el conjunto de estudios investigados. (Ver Fig. 6.)

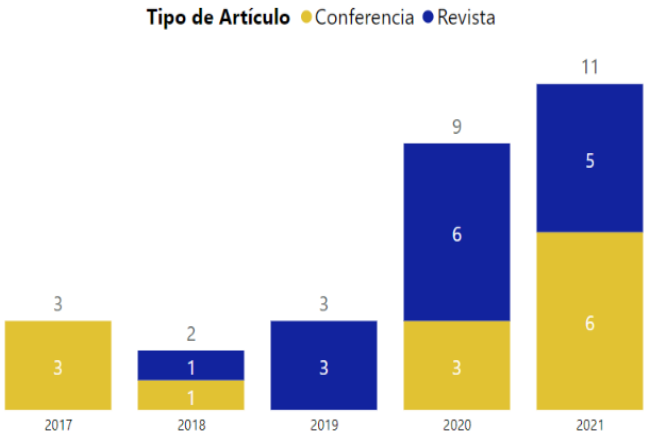


Fig. 6: Repartición de estudios

C. SLRP1: ¿Cuáles son las principales vulnerabilidades de ciberseguridad que ponen en riesgo a la red de datos?

En los estudios que se utilizaron para esta investigación se identificaron diversas anomalías de ciberseguridad como se aprecia en la Tabla V, del mismo modo se verificó la frecuencia de aparición de dichas anomalías en la Fig 2. En la fig 7 se presentaron las principales anomalías que afectan a la seguridad integral de los datos masivos:

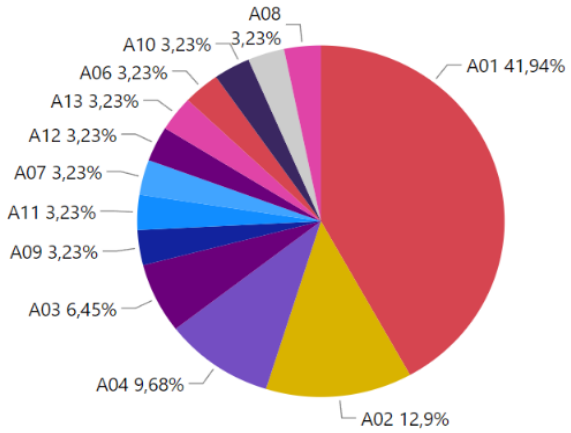


Fig. 7: Principales anomalías de ciberseguridad

Mediante esta interpretación se pudieron verificar cuáles son las principales anomalías, que se detallan a continuación:

- Phishing (A01): En la Fig. 7 se verificó que esta anomalía obtuvo un 41.94% del total analizado. En los papers [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]; se identificaron varios parámetros que son de análisis para la detección de phishing, sin embargo, se determinó de manera general que la forma más factible de identificar un ataque de este tipo es a través del análisis de URL. En estos papers en su mayoría se utiliza la técnica de ML para la recopilación y extracción de dichos parámetros.
- Network Attack (A02): En la Fig. 7 se verificó que esta anomalía obtuvo un 12.9% del total analizado. En los papers [24], [25], [26], [27]; mediante el uso de ABD se identificó que para la detección y prevención de este ataque se requiere el uso de un sistema de detección de intrusos.
- Social Engineering (A03): En la Fig. 7 se comprobó que esta anomalía obtuvo un 6.45% del total analizado. En [28] se empleó MLP para identificar dicha anomalía directamente sobre chats en línea. En [29] se revisó el empleo de NLP para identificar la anomalía aplicada a un sistema telefónico en tiempo real enfocado en la detección de firmas de voz.
- Malware (A04): En la Fig. 7 se verificó que esta anomalía obtuvo un 9.68% del total analizado. En [30] se desarrolló un marco de clasificación de malware, con el fin comprender el papel de este. En dicho estudio se trata al Spyware como uno de los principales ataques que almacena el historial de navegación en archivos. En [30] se utiliza Cyber Threat Intelligence (CTI) para encontrar proxys maliciosos en un conjunto de datos. Finalmente, en [32] se emplea Machine Learning con la finalidad de discernir familias de virus, troyanos y programas potencialmente no deseados.

D. SLRP2: ¿Cuáles son las técnicas de análisis de Big Data destacadas para la detección de anomalías?

En los estudios recopilados se identificaron diversas técnicas para el ABD en ciberseguridad, como se aprecia en la Tabla VI, del mismo modo se verificó la frecuencia de uso de dichas técnicas en la Fig. 3. A continuación, en la Fig. 8 se representa el porcentaje de utilización de cada una de las técnicas.

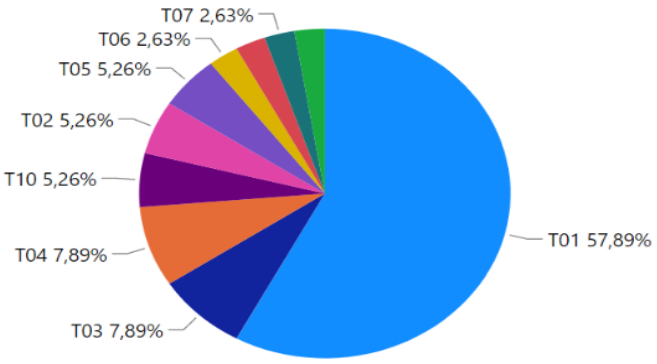


Fig. 8: Técnicas destacadas de análisis de Big Data.

De este modo se verificaron cuáles son las técnicas de ABD destacadas para la detección de ataques de ciberseguridad, que se exponen a continuación:

- Machine Learning (T01): Esta es la técnica de mayor uso en los estudios con un 57.89% como se puede apreciar en la Fig. 8. El aprendizaje automático o Machine Learning tiene como objetivo principal permitir que un sistema aprenda del pasado o del presente y utilizar dicho conocimiento para realizar predicciones o decisiones sobre eventos futuros desconocidos [41]. En [35] se empleó el ML en la nube para aumentar la seguridad y mejorar la velocidad de transmisión de datos. En [11] se propuso una detección de ataques de phishing basada en el aprendizaje automático para una verificación precisa. En [31] se abordó al ML para identificar mensajes relevantes de ciberseguridad en foros y extraer automáticamente información sobre diversas amenazas cibernéticas como credenciales filtradas, malware, servidores proxy maliciosos, etc.
- Deep Learning (T03): Esta técnica obtuvo un porcentaje de utilización del 7.89%. Deep Learning es un subconjunto del aprendizaje automático que tiene tres técnicas de aprendizaje que son: el aprendizaje supervisado, el semi-supervisado y no supervisado. Esta técnica consiste en muchas capas de redes neuronales artificiales que cada una de ellas contiene algunas neuronas con funciones de activación que pueden ser utilizadas para producir salidas no lineales [42]. En [21] se empleó al DL para la detección de páginas web de ataques cibernéticos, esto para predecir si una URL es legítima o es una suplantación de identidad. En [17] se empleó el enfoque de DL para la detección y clasificación de alta precisión para distinguir los sitios web genuinos de los de phishing.
- Neural Network (T04): Como ultima técnica destacada se considera a la Neural Network con un 7.89%. NN es un procesador distribuido paralelo compuesto por un gran número de unidades de procesamiento de datos que adquiere y ajusta continuamente los pesos de las neuronas interconectadas [43]. En [28] se emplearon las NN para procesar un diálogo y luego crear un conjunto de datos que pueden ser utilizados para la clasificación y así poder detectar ataques de ingeniería social de manera satisfactoria. En [24] emplearon NN para extraer características influyentes de los datos de intrusión y así poder evitar el sobre ajuste en las conexiones recurrentes.

E. SLRP3: ¿Qué métodos de Red Team son aplicables para burlar el estudio de Big Data?

En los estudios recopilados se hallaron diversas anomalías de seguridad. Cabe mencionar que el Red Team, es la concurrencia de operaciones que permiten evidenciar riesgos y peligros en un sistema informático. Estas actividades se basan en retos, metas y proyectos que atacan al sistema con el objetivo de obtener errores que se puedan registrar y mejorar [44]. Con esta premisa se postularon diversas actividades de Red Team para burlar a las técnicas de ABD, las cuales se mencionan a continuación:

- Una vez identificada la anomalía (A01) en los papers mencionados con anterioridad, se identificó como equipo de Red Team un ataque de tipo Parameter Tampering que busca vulnerabilidades en los URL. Este ataque aborda el valor de ciertos parámetros que se intercambian entre el cliente y el servidor con fines maliciosos; se usa la herramienta OWASP ZAP para capturar la petición a través del navegador antes enviarla al servidor, con el fin de modificar un parámetro y la configuración del proxy para que todo el tráfico se reenvíe a otro puerto [45].
- Una vez identificada la anomalía (A02) se comprobó como equipo de Red Team una ofensiva de Denegación de Servicio (DoS) con el uso de METASPLOIT. Esta herramienta, al ser un tipo de SYN Flood permite el envío masivo de solicitudes de conexiones TCP con el objetivo de evitar el trabajo del IDS [46].
- Una vez identificada la anomalía (A03) se verificó como equipo de Red Team a un ataque de SMS Spoofing. En donde a través del framework de SET se suplanta la identidad de un número telefónico. Mediante la opción Perform Spoofing Attack se ingresa al número telefónico al que se tiene planeado el ataque. Finalmente, este ataque otorga el acceso a un mensaje de texto genérico que podría ser editado y enviado mediante proveedores propios [47].
- Una vez identificada la anomalía (A04) se verificó como equipo de Red Team un Spyware llamado Invisimole que se compone de dos módulos con múltiples funcionalidades para extraer información. Este tipo de ataque se encuentra empaquetado en una DLL que permite obtener persistencia secuestrando una DLL legítima. Los dos módulos ejecutados por el malware introducen puertas

traseras para ejecutar tareas de espionaje o hacer cambios en el sistema y permiten al operador recopilar toda la información posible [48].

F. SLRP4: ¿Qué medidas de rendimiento de análisis de Big Data se utilizan para la detección de ataques?

En los estudios recopilados se identificaron diversas medidas de rendimiento de ABD para ciberseguridad, como se aprecia en la Tabla VIII, del mismo modo se puede verificar la recurrencia de uso de dichas métricas en la Fig 5. Posteriormente, en la Tabla IX se encuentra una síntesis de las métricas de rendimiento y sus resultados más relevantes, como una compilación de lo realizado durante la investigación.

Además, se comprueba que la medida de rendimiento más usada de acuerdo a las investigaciones es el Acurracy (AC) con un porcentaje de 23.08%. Esta medida describe la eficacia general de un modelo de clasificación.

IV. CONCLUSIONES

Con la aplicación de SM en las fuentes bibliográficas recopiladas se obtuvo tres etapas, que consistieron en: i) estructuración del método PICO, definición de la cadena de búsqueda y definición de criterios de elegibilidad, ii) definición de preguntas de investigación y estrategias de búsqueda y iii) estructuración y sintetización de las investigaciones para obtener los principales artículos con resultados significativos.

Una vez aplicada la metodología SLR se realizó un análisis crítico de los estudios de ABD en ciberseguridad publicados en el periodo del 2017 al 2021 en cuatro repositorios específicos. En los artículos se identificaron diversas anomalías de ciberseguridad que afectan a la seguridad integral de los datos. Las principales anomalías utilizadas para esta investigación fueron A01 con 41.94%, A02 con 12.9%, A04 con 9.68% y A03 con 6.45%. Dichas anomalías fueron objeto de estudio para evaluar y recomendar actividades de Red Team que burlen a las técnicas de ABD que se trataron a lo largo de esta investigación.

Los resultados arrojados a través de SM y SLR son los siguientes: T01 con 57.89%, T03 con 7.89% y T04 con 7.89% del interés investigativo. Los algoritmos de clasificación con mayor frecuencia en los estudios fueron: AL05 con 19.75%, AL04 con 18.52% y AL03 con 11.11%. Además, la métrica de rendimiento más utilizada en los estudios fue el M06 con 23.08% que permite verificar la eficacia general de los algoritmos antes mencionados.

Finalmente, con los resultados del SM y SLR se hallaron las principales debilidades de la red de datos, mismas que el equipo de Red Team puede usar para verificar la seguridad remanente con los siguientes ataques: Parameter Tampering, Denegación de Servicio, SMS Spoofing y Spyware Invisimole.

TABLE IX: RESULTADO DE LAS MÉTRICAS EN EL ANÁLISIS DE Big Data PARA LA DETECCIÓN DE ANOMALÍAS

Ref	Técnica	Algoritmos	Resultado métricas
[30]	ML	NB, KNN, DT, SVM, RF	TPR= 0.996, FPR=0.005, FNR=0.004, P: 0.995, F1= 0.995, AC=99.5%, MCC= 0.991
[24]	NN	CNN	P=0.98, F1=0.98, AC=97.17%, R=0.98, AET=0.002383
[36]	CC, MBDA	RF	P=0.86, F1=0.87, AC=89%, R=0.89
[35]	ML, CC	KM	AC=91.7%
[25]	MBDA	SVM	S=99
[37]	ML	LR, NB, ANN, MLP, KNN, SVM	AC=93.05%
[26]	AS, ML	KNN	TPR=0.029, P=0.98, F1=0.98, R=0.98, AET=60635sec, DR=98.5
[11]	ML	DT, RF	P=0.9689, F1=0.5874, AC=96.96%, R=0.4216
[12]	ML	DT, RF	AC=95.0%
[28]	NN	DT, RF	F1=0.759, AC=92.40%, AUC=0.722
[38]	DL	SVM, ELM	TPR=88.81%, FPR=0.0006%
[31]	ML, DK	SVM	P=98.82%
[33]	ML	CNN, RF, SVM, RNN	TP=63.8%, FP=0.5, TN=34.9, FN=0.7
[13]	ML, NN	SVM, RF, LR	AC=98.65% P=0.9890, R=0.9881, F1=0.9886
[14]	ML, DM	KM, NB, DT	TPR=89.985%
[15]	ML, KDD, DM	RF, GBM	AC=93.70%, AUC=0.979
[16]	ML	SVM, DNN	AUC=0.8250, TN=0.7716, TC=0.9177
[32]	ML	MLP. CNN, SVM	P=88%, AC=0.89, F1=0.88
[17]	DL	CNN, SVM, RF, DT	AC=93.73%, P= 0.970, F1=0.976, R= 0.982
[18]	ML	NB, RF, R, MLP	P=90,08%
[19]	ML	SVM, LR, ANN	AC= 94.50%
[20]	ML	SVM	TPR=93.12%; TP=1059; FN=84; TN=16056; AC=99.50
[21]	ML, DL	RF, ANN, GBM, SVM, LR, KNN, DT	AC=98.72%
[22]	ML	SVM, RF, LIR	AC=99.33%
[23]	ML	RF, GBM, LR, KNN, DT	P=97.85, R=98.61, F1=98.23
[29]	ML, NLP	RF, KNN, MLP	AC=78.4%, P=79.7%, AUC=81.6%
[34]	ML	KNN, LR, DT, RF, SVM, MLP	AC=90.40%, P=1, R=0.206, F1=0.341, AUC=0.943
[27]	ML	RF, NB, LR	FPR= 0.51, AC=99.72%

REFERENCES

[1] R. Sabillon, J. Serra-Ruiz, V. Cavaller, and J. Cano, "A comprehensive cybersecurity audit model to improve cybersecurity assurance: The cybersecurity audit model (csam)," in *Proceedings - 2017 International Conference on Information Systems and Computer Science, INCISCOS 2017*, vol. 2017-November. United States: Institute of Electrical and Electronics Engineers Inc., Mar. 2018, pp. 253–259, 2nd International Conference on Information Systems and Computer Science, INCISCOS 2017 ; Conference date: 23-11-2017 Through 25-11-2017.

[2] G. Bassett, C. D. Hylender, P. Langlois, A. Pinto, and S. Widup, "Data breach investigations report," *Verizon DBIR Team, Tech. Rep*, pp. 6–8, 2021.

[3] T. T. Teoh, Y. Y. Nguwi, Y. Elovici, N. M. Cheung, and W. L. Ng, "Analyst intuition based hidden markov model on high speed, temporal cyber security big data," in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, July 2017, pp. 2080–2083.

[4] C. I. N. Corona and M. S. R. Montoya, "Mapeo sistemático de la literatura sobre evaluación docente (2013-2017)," *Educação e Pesquisa*, 2018. [Online]. Available: http://hdl.handle.net/11285/632774

[5] O. Barbosa and C. F. Alves, "A systematic mapping study on software ecosystems," in *IWSECO@ICSOB*, 2011.

[6] J. LeClair, K. M. Hollis, and D. M. Pheils, "Cybersecurity education and training and its reliance on steam," in *2014 IEEE Integrated STEM Education Conference*, March 2014, pp. 1–5.

[7] S. M. Shamsuddin and S. Hasan, "Data science vs big data @ utm big data centre," in *2015 International Conference on Science in Information Technology (ICSITech)*, Oct 2015, pp. 1–4.

[8] T. F. Frandsen, M. F. Bruun Nielsen, C. L. Lindhardt, and M. B. Eriksen, "Using the full pico model as a search tool for systematic reviews resulted in lower recall for some pico elements," *Journal of Clinical Epidemiology*, vol. 127, pp. 69–75, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895435620305692

[9] L. M. Connelly, "Inclusion and exclusion criteria," *Medsurg Nursing*, vol. 29, no. 2, p. 125, Mar 2020, copyright - Copyright Anthony J. Jannetti, Inc. Mar/Apr 2020; Última actualización - 2021-06-02. [Online]. Available: https://www.proquest.com/scholarly-journals/inclusion-exclusion-criteria/docview/2388933304/se-2?accountid=32861

[10] K. Sudar, P. Deepalakshmi, P. Nagaraj, and V. Muneeswaran, "Analysis of cyberattacks and its detection mechanisms," in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Nov 2020, pp. 12–16.

[11] M. Alam, D. Sarma, F. Lima, I. Saha, R.-E. Ulfath, and S. Hossain, "Phishing attacks detection using machine learning approach." Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1173–1179, cited By 16. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094854882&doi=10.1109%2fICSSIT48917.2020.9214225&partnerID=40&md5=fc202ee046c28896a06ec2a985be3535

[12] D. Brites and M. Wei, "Phishfry - a proactive approach to classify phishing sites using scikit learn." Institute of Electrical and Electronics Engineers Inc., 2019, cited By 3. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082305226&doi=10.1109%2fGCWkshps45667.2019.9024428&partnerID=40&md5=456b1ba56196e23f88ebfac27dec743c

[13] A. Hashim, R. Medani, and T. A. Attia, "Defences against web application attacks and detecting phishing links using machine learning," in *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, Feb 2021, pp. 1–6.

[14] Şentürk, E. Yerli, and Soğukpınar, "Email phishing detection and prevention by using data mining techniques," in *2017 International Conference on Computer Science and Engineering (UBMK)*, Oct 2017, pp. 707–712.

[15] S. Dangwal and A.-N. Moldovan, "Feature selection for machine learning-based phishing websites detection," in *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, June 2021, pp. 1–6.

[16] H. Shirazi, K. Haefner, and I. Ray, "Fresh-phish: A framework for auto-detection of phishing websites," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, Aug 2017, pp. 137–143.

[17] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," in *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*, March 2020, pp. 1–6.

[18] J. S. Mittapalli, S. Ojha, and S. T, "Phishing attack detection using python and machine learning," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, June 2021, pp. 531–536.

[19] F. Salahdine, Z. El Mrabet, and N. Kaabouch, "Phishing attacks detection a machine learning-based approach," in *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, Dec 2021, pp. 0250–0255.

[20] W. Niu, X. Zhang, G. Yang, Z. Ma, and Z. Zhuo, "Phishing emails detection using cs-svm," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, Dec 2017, pp. 1054–1059.

[21] K. Mridha, J. Hasan, S. D, and A. Ghosh, "Phishing url classification analysis using ann algorithm," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Sep. 2021, pp. 1–7.

[22] J. Stobbs, B. Issac, and S. M. Jacob, "Phishing web page detection using optimised machine learning," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Dec 2020, pp. 483–490.

[23] P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, "Robust ensemble machine learning model for filtering phishing urls: Expandable random gradient stacked voting classifier (erg-svc)," *IEEE Access*, vol. 9, pp. 150 142–150 161, 2021.

[24] M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, and G. Fortino, "A hybrid deep learning model for efficient intrusion detection in big data environment," *Information Sciences*, vol. 513, pp. 386–396, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519310382

[25] J. Camacho, J. M. García-Giménez, N. M. Fuentes-García, and G. Maciá-Fernández, "Multivariate big data analysis for intrusion detection: 5 steps from the haystack to the needle," *Computers Security*, vol. 87, p. 101603, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404818307909

[26] A. Abid and F. Jemili, "Intrusion detection based on graph oriented big data analytics," *Procedia Computer Science*, vol. 176, pp. 572–581, 2020, knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920318834

[27] K. Siddique, Z. Akhtar, M. A. Khan, Y.-H. Jung, and Y. Kim, "Developing an intrusion detection framework for high-speed big data networks: A comprehensive approach," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, vol. 12, no. 8, pp. 4021–4037, AUG 31 2018.

[28] M. Lansley, F. Mouton, S. Kapetanakis, and N. Polatidis, "Seader++: Social engineering attack detection in online environments using machine learning," *Journal of Information and Telecommunication*, vol. 4, no. 3, pp. 346–362, 2020, cited By 13. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091983570&doi=10.1080%2f24751839.2020.1747001&partnerID=40&md5=70443a0c50913ee993ba03cbc10fa131

[29] M. Lansley, S. Kapetanakis, and N. Polatidis, "Seader++ v2: Detecting social engineering attacks using natural language processing and machine learning," in *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Aug 2020, pp. 1–6.

[30] "Improving malware detection using big data and ensemble learning," *Computers Electrical Engineering*, vol. 86, p. 106729.

[31] I. Deliu, C. Leichter, and K. Franke, "Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec 2018, pp. 5008–5013.

[32] A. Walker, T. Das, R. M. Shukla, and S. Sengupta, "Friend or foe: Discerning benign vs malicious software and malware family," in *2021 IEEE Global Communications Conference (GLOBECOM)*, Dec 2021, pp. 01–06.

[33] G. Siewruk and W. Mazurczyk, "Context-aware software vulnerability classification using machine learning," *IEEE Access*, vol. 9, pp. 88 852–88 867, 2021.

[34] K. Lee and K. Yim, "Cybersecurity threats based on machine learning-based offensive technique for password authentication," *APPLIED SCIENCES-BASEL*, vol. 10, no. 4, FEB 2020.

[35] A. S. Mohammad and M. R. Pradhan, "Machine learning with big data analytics for cloud security," *Computers Electrical Engineering*, vol. 96, p. 107527, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790621004729

[36] C. A. Ardagna, V. Bellandi, E. Damiani, M. Bezzi, and C. Hebert, "Big data analytics-as-a-service: Bridging the gap between security experts and data scientists," *Computers Electrical Engineering*, vol. 93, p. 107215, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790621002081

[37] Y. Jiang and Y. Atif, "A selective ensemble model for cognitive cybersecurity analysis," *Journal of Network and Computer Applications*, vol. 193, p. 103210, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804521002125

[38] W. Yan, L. K. Mestha, and M. Abbaszadeh, "Attack detection for securing cyber physical systems," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8471–8481, Oct 2019.

[39] A. F. de Castro and A. G. de Oliveira, "Study on dengue cases using data analysis techniques: A case study in the state of pernambuco, brazil," in *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, June 2021, pp. 1–6.

[40] A. S. Rani and S. Jyothi, "Performance analysis of classification algorithms under different datasets," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, March 2016, pp. 1584–1589.

[41] R. A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I. A. Targio Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management*, vol. 45, pp. 289–307, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0268401218301658

[42] M. A. Amanullah, R. A. A. Habeeb, F. H. Nasaruddin, A. Gani, E. Ahmed, A. S. M. Nainar, N. M. Akim, and M. Imran, "Deep learning and big data technologies for iot security," *Computer Communications*, vol. 151, pp. 495–517, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366419315361

[43] T. Liu, H. Mei, Q. Sun, and H. Zhou, "Application of neural network in fault location of optical transport network," *China Communications*, vol. 16, no. 10, pp. 214–225, Oct 2019.

[44] J. Tang, G. Leu, and H. A. Abbass, *Computational Red Teaming*. IEEE, 2020, pp. 241–251. [Online]. Available: https://bibliotecas.ups.edu.ec:2095/document/8889944

[45] H. C., "Realizar un ataque parameter tampering," *Open Webinars*, Ago 2019. [Online]. Available: https://openwebinars.net/blog/realizar-un-ataque-parameter-tampering/

[46] G. J. Hurtado M., "Configuración de una herramienta open source para la detección de intrusos en redes wifi caso de estudio: Suricata ids," pp. 65–71, Dec 2019. [Online]. Available: https://dspace.unl.edu.ec/jspui/handle/123456789/22835

[47] V. J., "Uso de sms spoofing desde set," *Testpurposes*, Nov 2010. [Online]. Available: https://testpurposes.net/2010/11/20/uso-de-sms-spoofing-desde-set/

[48] S. F., "Invisimole: el malware espía que convierte tu ordenador en un sistema de vigilancia," *Una al Día*, Jun 2018. [Online]. Available: https://unaaldia.hispasec.com/2018/06/invisimole-el-malware-espia-que-convierte-tu-ordenador-en-un-sistema\-de-vigilancia.html