Journal of Education and Science (ISSN 1812-125X)

www.edusj.mosuljournals.com

Design and Implement Machine Learning Tool for Cyber Security Risk Assessment

O. I. Sheet **L M. Ibraheem**

Department of Software Engineering, College of Computer Sciences & Mathematics, University of Mosul, Mosul. Iraq

Article information

Article history:
Received: December 31, 2022
Accepted: March 15, 2023
Available online: June 01, 2023

Keywords:
Machine Learning (ML)
Risk Assessment
Light Gradient Boosting
CatBoost
Multi-Layer Perceptron (MLP).

Correspondence:
L M. Ibraheem
laheeb_alzubaidy321966@uomosul.edu.iq

Abstract

Cyber-attacks have increased in number and severity, which has negatively affected businesses and their services. As such, cyber security is no longer considered merely a technological problem, but must also be considered as critical to the economy and society. Existing solutions struggle to find indicators of unexpected risks, which limits their ability to make accurate risk assessments. This study presents a risk assessment method based on Machine Learning, an approach used to assess and predict companies' exposure to cybersecurity risks. For this purpose, four algorithm implementations from Machine Learning (Light Gradient Boosting, AdaBoost, CatBoost, Multi-Layer Perceptron) were implemented, trained, and evaluated using generative datasets representing the characteristics of different volumes of data (for example, number of employees, business sector, and known vulnerabilities and externel advisor). The quantitative evaluation conducted on this study shows the high accuracy of Machine Learning models and Especially Multi-Layer Perceptron was the best accuracy when working compared to previous work.

DOI:10.33899/edusj.2023.137554.1307,

1. المقدمة

أدت العولمة والتقنيات الذكية والرقمية إلى تصعيد الجريمة السيبرانية، حيث أنه مجال ناشئ للبحث والصناعة ، فقد سلط الضوء على أهمية أنظمة الدفاع القوية للأمن السيبراني على مستوى الشركات والمستويات الوطنية. تشير التقديرات إلى أن تأثيرات عدم كفاية الأمن السيبراني كلفت الاقتصاد العالمي 945 مليار دولار أمريكي [1]. في عام 2020 فأصبحت تشكل نقاط ضعف الأمن السيبراني مخاطر كبيرة على الشركات ، بما في ذلك انقطاع الأعمال وانتهاك الخصوصية والخسائر المالية[2]. على الرغم من الأهمية المتزايدة للاقتصاد الدولي، فما زال توفر البيانات حول المخاطر الإلكترونية محدوداً، والسبب يعود إلى أنه خطر ناشئ ومتطور لذلك ، فإن مصادر البيانات التاريخية محدودة[3] .قد يرجع ذلك أيضًا إلى حقيقة أن المؤسسات التي تم اختراقها بشكل عام لا تنشر الأحداث[4]. يشكل نقص البيانات تحديات للعديد من المجالات ومنها إدارة وتقييم مخاطر الأمن السيبراني[5].

الأمن السيبراني هو مجموعة من التدابير الأمنية التي يمكن استخدامها لحماية الفضاء السيبراني وأصول المستخدم من الوصول غير المرغوب فيه والاعتداءات[6]. يتيح تقييم المخاطر الى معرفة نقاط ضعف الانظمة واتخاذ خطوات لمعالجتها لهذا يتم تقييم مخاطر الامن السيبراني حتى نقلل من الضرر الذي تسببه الهجمات الالكترونية [7]. من التهديدات التي تستهدف الامن السيبراني هي البرامج الضارة ((Malicious Software (Malware)[8,9]، التصيد ([Phishing) [10]. رفض الخدمة(DoS)Denial of Service) / رفض الخدمة الموزع (DDoS) (Distributed Denial of Service /)[11] ، الهندسة الاجتماعية [10] , الهجوم المستمر المتقدم (Advanced Persistent Attack (APT)) [8] ، هجوم الشم (Sniffer-Attack) [8].

تقييم المخاطر هي عملية مهمة في إدارة المخاطر لنظام معلومات الشبكة. يتم إجراؤه على أساس احتمالية التهديد وتأثيره ، عادة تعتمد طرق تقييم المخاطر هذه على عوامل ذاتية مثل تحقيق الخبراء النوعي إذ يتم يتم اجراء تقييم للمخاطر كميا على اساس عدد الفئات ونوعيا لتقييم كل خطر على حدة [12]. ويساهم التعلم الالي في عملية تقييم المخاطر للامن السيبراني الذي يعتبر احد فروع الذكاء الاصطناعي التي تتخصص في مجال واسع من تحليل البيانات الذي يسمح للخوارزمية بالتعلم من البيانات للإجابة على سؤال أو إصدار قرار لحل المشكلات المعقدة [13]. وتعتبر نماذج التعلم الالي اداة مثالية في تقييم مخاطر الامن السيبراني حيث يتم تدريب النموذج باستخدام البيانات مما يسمح للنماذج بالكشف عن الأنماط من تلقاء نفسها ، مما يساعد في تحديات التجميع وتقليل الأبعاد للكشف عن هجمات البرامج الضارة والاحتيال غير المعروفة. ولذلك تم في هذا العمل استخدام طرق التعلم الالي , Light Gradient Boosting, AdaBoost, CatBoost، الشبكة العصبية الاصطناعية المدرك لتقييم مخاطر الامن السبراني.

فيما يلي كيفية تنظيم هذه الورقة. الفقره الثانية تتضمن البحوث السابقة. توضح الفقرة الثالثة العمل المقترح والخوارزميات المستخدمة في تقييم مخاطر الامن السيبراني، النتائج والمناقشة موضحة في الفقره الرابعة واخيرا تقدم الفقرة الخامسة الاستنتاجات وآفاق العمل المستقبلي.

2.الدراسات السابقة

الجدول (1) يوضح ملخص للبحوث والدراسات في مجال تقييم مخاطر الامن السيبراني وكما يلي:

الجدول 1: الدراسات السابقة

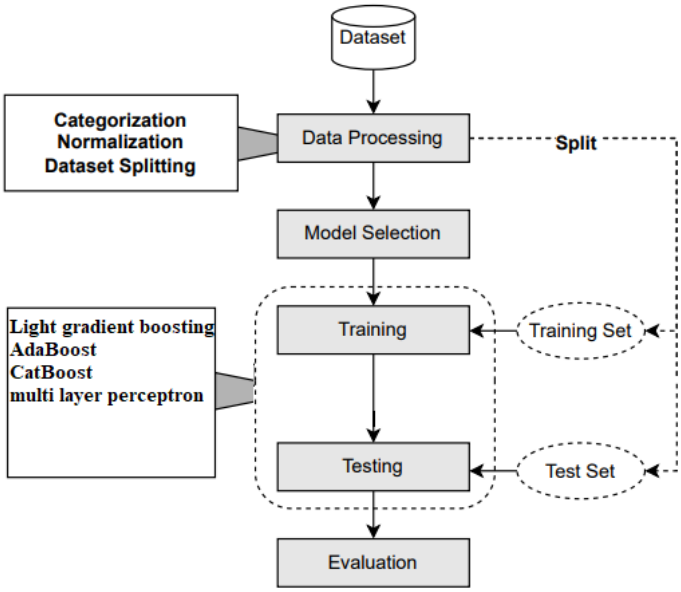
ت	البحوث	مجموعة البيانات	الخوارزميات المستخدمة	المقاييس	النتائج	ملاحظات
1	(Yuri Castro et al 2015 [16])	Stats19	1. Bayesian network (BN) 2. Decision tree(DT) 3. Multi-layer perceptron (MLP)	Accuracy	Accuracy for BN is 0.815 Accuracy for DT is 0.811 Accuracy for MLP is 0.813	بسبب الخصائص الكثيرة الموجودة بالبيانات كانت نسبة الدقة بالخوارزميات قليلة ولا يتم تقييم جميع المخاطر بشكل دقيق
2	(Qi Zhang et al 2017 [17])	attack evidence and anomaly evidence	1. Fuzzy Probability 2. Bayesian Network	Accuracy	أجريت تجارب على منصة محاكاة لمفاعل كيميائي مبسط. أظهرت عمليات المحاكاة التي أجريناها لـ 5000 مرة لكل سيناريو وقت حساب يبلغ حوالي 3 ثوانٍ لتقييم المخاطر .	هذا البحث يعتمد في تقييم المخاطر بشكل ديناميكي ولكن في حالة كانت البيانات غير دقيقة فان تقييم المخاطر ب fuzzy probability و Bayesian network يكون خاطئ .
3	(Zhao et al 2019 [18])	CIC2017 dataset	1. K-Nearest Neighbors (K-NN) 2. Random Forest (RF) 3. Iterative Dichotomiser 3 (ID3) 4. AdaBoost 5. ,Multi-layer perceptron (MLP) 6. Naïve Bayes (NB) 7. Quadratic Discriminant Analysis (QDA) 8. Lie Group	Precession	Accuracy for K-NN is 0.96 Accuracy for RF is 0.98 Accuracy for ID3 is 0.98 Accuracy for AdaBoost is 0.77 Accuracy for MLP is 0.77 Accuracy for NB is 0.88 Accuracy for QDA is 0.97 Accuracy for Lie Group is 0.83	يمكن لطريقة الحساب القائمة على Lie Group أن تحلل مخاطر أمن الشبكة كميًا وليس نوعيا وايضا يوجد تعقيد زمني بخوارزمية knn على الرغم من قوتها في التقييم .
4	(Franco et al 2020 [19])	بيانات توليدية (50,000 عينة)	1. Multi-layer perceptron (MLP) 2. Decision tree(DT) 3. Support vector machine (SVM) 4. K-Nearest Neighbors (K-NN)	Accuracy	Accuracy for MLP is 98.86 % Accuracy for DT is 92.64 % Accuracy for SVM is 99.03 % Accuracy for K-NN is 95.82 %	في خوارزمية MLP تم استخدام طبقتين مخفيتين كل طبقة تحتوي على 5 خلايا عصبية
5	**(**V. Sampath Kumar et al 2021 [20]**)**	بيانات توليدية (1209 عينة)	1. Naïve Bayes (NB) 2. Decision tree(DT)	Accuracy	Accuracy for NB is 0.828 Accuracy for DT is 0.942	دقة خوارزمية NBغير دقيقة خصوصا في حالة اذا لم تتنبأ بالخطر تعطيه صفراً وهذا الشي غير دقيق
6	(Maxim Kalinin et al 2021 [21])	بيانات توليدية (10000 عينة)	perceptron model and backpropagation linear discriminant analysis and logistic regression	Accuracy	Accuracy for perceptron model and backpropagation is 0.98 Accuracy for linear discriminant analysis and logistic regression is 0.84	احدى مشاكل logistic regression انه في حالة كانت الميزات كثيرة يؤدي الى فرط التجهيز
7	(van Haastrecht et al 2021 [22])	Enisa 2012	استخدام تقنية Self-Determination	1. Malicious URL Count 2. Awareness Training Score 3. Malware Infection Count 4. Malicious App Count	كانت الدقة على بيانات التصيد 87 % وعلى بيانات malware كانت الدقة 78 %	هذا البحث يستهدف بشكل خاص الشركات الصغيرة والمتوسطة ويعتمد في تقييم الخطر على 3 متغيرات وهي autonomy و competence و relatedness ولا يستهدف الشركات الكبيرة

في العمل المقترح يتم اقتراح خوارزمية Light Gradient Boosting لانه تم تطويرها حديثًا باستخدام إطار تعزيز التدرج لإنشاء شجرة القرار ولكونها عالجت مشاكل التصنيف والانحدار واقترحت خوارزمية AdaBoost لانه تحافظ على توزيع الاوزان حيث تمنح الاوزان الاعلى للمتغيرات الصلبة والاوزان الادنى

للمتغيرات السهلة وتم اختيار خوارزمية CatBoost لانها أداة جيدة للغة الآلة لحل البيانات غير المتجانسة والصاخبة والمتغيرات المعقدة حيث تستخدم أشجار القرار الثنائية كمتنبئات أساسية، وله خصائص قوية لتقليل ضبط المعلمة الفائقة ، وتقليل فرص فرط البيانات. فهو يجمع بين شجرة قرار تعزيز التدرج (Gradient Boosting Decision Tree) والميزات الفئوية ، ويركز على المتغيرات الفئوية ، ويتعامل مع مشاكل انحياز التدرج والتنبؤ واختيرت خوارزمية الشبكة العصبية الاصطناعية المتعددة الطبقات المدرك لكونها تعالج المشاكل المعقدة والعمليات الغير خطية وايضا تتوقع الاخراج بشكل سريع.

3.العمل المقترح

في هذه الفقرة توضحت خطوات العمل المقترح لتصميم وتنفيذ اداة لتقييم مخاطر الامن السيبراني باستخدام خوارزميات التعلم الآلي، حيث يتكون النظام المقترح في هذا البحث من خمس خطوات كما هو مبين في الشكل (1).



الشكل 1.مراحل العمل المقترح

يمكن شرح خطوات العمل باختصار على النحو التالي:

1. توليد مجموعة بيانات خاصة بتقييم المخاطر بالاعتماد على الخصائص التي ذكرت في الجدول (2).
2. المعالجة المسبقة للبيانات هي الخطوة الثانية حيث تم تصنيف بعض الخصائص وتحويل القيم من كلمة الى رقم ومن ثم استخدام تقنية تطبيع البيانات لقياس وتعديل البيانات في النطاق [0 و 1] عبر (طريقة Max-Min method).
3. الخطوة الثالثة هي تقسيم مجموعة البيانات المستخدمة في هذا العمل إلى 80% لتدريب الخوارزميات و20% للاختبار.
4. الخطوة الرابعة هي تطبيق اربع خوارزميات من تقنيات التعلم الالي على مجموعة البيانات(خطوات التدريب و الاختبار).
5. الخطوة الاخيرة هي عملية تقييم للنماذج التي تم تدريبها.

3.1 مجموعة البيانات المستخدمة

المرحلة الأكثر أهمية هي مرحلة جمع البيانات. إذ يتم جمع البيانات من أجهزة الاستشعار أو مصادر مختلفة أخرى وتخزينها لمزيد من المعالجة في هذه المرحلة. ومع ذلك ، في مجال تقييم مخاطر الأمن السيبراني ، لا تفصح الشركات عن أية معلومات على الإطلاق أو في بعض الحالات تنشر تقارير مختلفة غالبًا ما تكون غير كاملة ويصعب استخراج نتائج ذات مغزى منها [19]. تم في هذا العمل تنفيذ نهج مولد البيانات لاستخدامها في عملية التدريب على الخوارزميات. وحددت المعلمات التالية لاستخدامها كأساس لهذا العمل:

a) الارباح (Revenue) : هو الدخل الناتج عن الأنشطة والعمليات التجارية العادية ، وفي معظم الحالات يستخدم أيضًا لتصنيف الأعمال من خلال توفير مقياس لتحديد أحجامها.

b) استثمارات الأمن السيبراني (Cybersecurity Investments) : قد يكون لدى الشركات استراتيجيات استثمار في مجال الأمن السيبراني لضمان سلامة مستوى الدفاع. يجب أخذ هذا النوع من المعلومات في الاعتبار أثناء تقييم مخاطر الأمن السيبراني ، حيث يكون له تأثير على احتمال استهدافه بهجوم إلكتروني.

c) عدد الموظفين ومستوى التدريب (Number of Employees and Training Level) : هي المعلومات المتعلقة بالعدد الفعلي للموظفين في الشركة وكذلك مستوى التدريب المقابل في مجال الأمن السيبراني (على سبيل المثال ، المعرفة الأساسية للأمن السيبراني والتدريب على التصيد) تمثل المعلومات السياقية الأساسية المطلوبة لتقييم المخاطر السيبرانية المحتملة ويتم قياس مستوى تدريب الموظف على أنه منخفض ومتوسط وعال.

d) الهجمات الإلكترونية الناجحة/ الفاشلة (Successful/Failed Cyberattacks) : تشير هذه المعلمة إلى عدد الهجمات الإلكترونية التي تعرضت لها الشركة بالفعل. يتضمن ذلك هجمات مختلفة على سبيل المثال ، DDoS والتصيد الاحتيالي التي استهدفت البنية التحتية للمؤسسة وأدت إما إلى خسارة مالية أو الإضرار بالسمعة وتؤخذ المحاولات الفاشلة في الاعتبار أيضًا.

e) نقاط الضعف المعروفة (Known Vulnerabilities) : للحصول على تقييم فعال وشامل للمخاطر، من الضروري الإبلاغ عن أي نقاط ضعف معروفة للبنية التحتية عادة ما تكون إدارة الثغرات الأمنية مسؤولية رئيسية لفريق أمن تكنولوجيا المعلومات في الشركات حيث تتضمن هذه المرحلة عادةً تقييم أي ثغرة أمنية موجودة في أنظمة المؤسسة والإبلاغ عنها هناك مجموعة متنوعة من الأدوات الشاملة المستخدمة لفحص الثغرات الأمنية ، مثل Nmap و Metasploit و OWASP. حاليًا يتم تحديد إجمالي عدد مواطن الضعف المعروفة أثناء عملية التوليد التركيبية.

f) المستشار الخارجي للأمن السيبراني (External Cybersecurity Advisor) : لزيادة تعزيز مرونتها الإلكترونية (أي القدرة على الاستعداد للهجمات الإلكترونية والاستجابة لها والتعافي منها) ، يتم تشجيع الشركات على تعيين مستشار خارجي للأمن السيبراني (CSA). أثناء مرحلة إنشاء البيانات التركيبية، يتم إنشاء قيمة ثنائية (إما نعم أو لا).

g) المخاطر (Risk): يمثل قيمة التقييم النوعي للمخاطر بناءً على المعايير التي تم إنشاؤها مسبقًا نظرًا لأن عملية توليد البيانات التركيبية مصممة لإنشاء سجلات تاريخية للشركات العاملة في صناعات قابلة للمقارنة، فقد يتم اشتقاق قيمة عمود المخاطر من تقنيات تقييم المخاطر النوعية الرسمية أو المصممة خصيصًا يمكن أن تفترض المخاطر المتولدة إحدى القيم التالية: منخفضة ومتوسطة ومرتفعة.

لتوليد المعلومات المذكورة أعلاه، تم عمل بعض الافتراضات. أولاً ، تم تحديد الحدود العليا والسفلى لكل عمود بحيث تقع كل قيمة تم إنشاؤها بشكل فعال في النطاق المحدد. يوضح الجدول (2) المتغيرات التي تم استخدامها في المعادلة (1)، ويقدم الجدول (3) نظرة عامة على الحدود المحددة بالإضافة إلى أمثلة على القيم لكل معلومات تم إنشاؤها. تُستخدم هذه السمات أيضًا كمدخلات لتعيين المخاطر وفقًا لما تقترحه المعادلة 1.

$$Computed\_risk=\frac{invested\_amount}{business\_value}+\frac{nr\_employees}{total\_employee}*map(employee\_training)+map(external\_adv)-\frac{succ\_attacks}{max\_attacks}-\frac{Known\_vuln}{maxKnown\_vuln} \quad (1)$$

الجدول (2): قيمة المتغيرات المستخدمة في المعادلة (1)

المتغير	قيمة المتغير
Invested_amount	المبلغ المستثمر
Business_value	تمثل القيمة الحقيقية للخطر
Nr_employee	تمثل عدد الموظفين في كل خطر
Total_employee	تمثل عدد الموظفين الكلي
Employ_training	تمثل عدد الموظفين المدربين
External_adv	تمثل الخبراء
Succ_attacks	عدد الهجمات الناجحة
Max_attacks	تمثل اكبر عدد من الهجمات
Known_vuln	تمثل الثغرات الموجودة في كل خطر
maxKnown_vuln	تمثل اكبر عدد من الثغرات
Computed_risk	تمثل قيمة الخطر الناتجة من المعادلة (1)

من المهم ملاحظة أن المخاطر لا تنشأ بشكل عشوائي، بدلاً من ذلك يتم حسابها بناءً على السمات التي تم إنشاؤها والموضحة في الجدول (3) باستخدام المعادلة (1). بالنسبة لعملية التعلم الخاضعة للإشراف، يجب تسمية مجموعة البيانات. نتيجة لذلك يتم تعيين مخرجات المخاطرة المحسوبة إلى فئة منخفضة أو متوسطة أو عالية. قد تكون عملية وضع العلامات اليدوية باهظة الثمن، نظرًا لأن مجموعة البيانات التي أنشئت تتضمن آلاف السجلات. لذلك، بناءً على القيمة الرقمية للمخاطر المحسوبة (x)، يتم تحديد نطاق التعيين. هذا يعني أنه يتم تصنيف كل قيمة مخاطرة محسوبة باستخدام النطاق كما هو محدد في المعادلة (2). بالاعتماد على هذه الخصائص ولدت 50000 عينة وكان الوقت المستغرق لتوليد هذه البيانات هو 20 دقيقة، يوضح الجدول (3) مدى الخصائص التي تولدت، والجدول (4) يوضح مثال عن البيانات التي تم توليدها.

$$map(x)=\begin{cases}0, & \text{if } x = Low\\ 1, & \text{if } x= Medium\\ 2, & \text{if } x= High\end{cases} \quad (2)$$

الجدول 3. نظرة عامة على الخصائص التي تم انشائها لمجموعة البيانات

Information	ID	Range
Revenue	business_value	0 to 5,000,000
Cybersecurity Investment	Invested_amount	0-30 % * Revenue
Successful Attacks	succ_attack	0 to 50
Failed Attacks	Fail_attack	0 to 50
Number of Employees	Nr_ employees	30 to 10,000
Employee Training	Employees_training	Low, Medium, or High
Known Vulnerabilities	Known_vuln	0 to 10
External Cybersecurity Advisor	External_adv	Yes or no
Risk	Successful Attacks	Successful Attacks

الجدول 4. مثال عن البيانات التوليدية

Invested_ amount	Successful Attacks	Failed Attacks	Business Value	Number of Employees	Employee Training	Known Vulnerabilities	External Advisor	Risk
818686	32	22	5006178	5908	LOW	4	NO	HIGH
114066	43	41	4964853	5581	MEDIUM	7	YES	MEDIUM
787223	15	37	5007265	4697	MEDIUM	3	YES	LOW
241955	10	45	4954403	7631	LOW	0	YES	MEDIUM
186203	49	47	4929074	8170	HIGH	3	NO	MEDIUM
1435301	44	12	4938481	2213	MEDIUM	9	NO	HIGH
760329	24	40	4978074	2421	LOW	0	NO	HIGH

3.2 معالجة البيانات

في عملية معالجة البيانات يتم اجراء خطوتين:

1. **اجراء عملية لتصنيف البيانات** حيث يتم في هذه المرحلة اجراء عملية تصنيف على 3 خصائص من مجموعة البيانات حيث تم التصنيف لخاصيتي Employee Training وRisk كما في الجدول (5) وصنفت على خاصية External Advisor كما في الجدول (6).

الجدول 5. التصنيف على خاصية Employee Training وRisk

قيمة الخاصية قبل التصنيف	قيمة الخاصية قبل التصنيف
Low	0
Medium	1
High	2

الجدول 6. التصنيف على خاصية External Advisor

قيمة الخاصية قبل التصنيف	قيمة الخاصية قبل التصنيف
Yes	1
No	0

2. **تطبيع البيانات**

يعتبر تطبيع البيانات تقنية مهمة جدًا تُستخدم لتحسين أداء نظام التعلم الالي [23]. والسبب في ذلك هو أن بعض مجموعات البيانات (على سبيل المثال، مجموعة البيانات التوليدية) تتضمن ميزات ذات قيم ونطاقات ومقاييس مختلفة جدًا. في هذا البحث تم استخدمت تقنية Max- Min لتطبيع سمة البيانات في مجموعة البيانات ضمن النطاق [0-1] باستخدام المعادلة 3 [24] ، حيث$Z \epsilon [min, max]$ وz' ينتمي الى [0، 1]

$$z' = \frac{Z\text{-min}}{\text{max-min}} \qquad (3)$$

3.3 بيانات التدريب والاختبار

في هذه المرحلة من خطوات تصميم العمل المقترح يتم تقسيم البيانات المتولدة بنسبة 80% لبيانات التدريب وعددها 40,000 عينة وبيانات الاختبار لنسبة 20% وعددها 10,000 عينة.

3.4 تقنيات التعلم الالي المقترحة

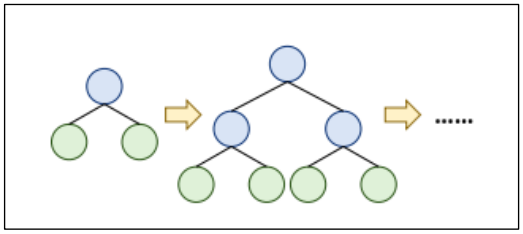
استخدم في هذا العمل اربع من خوارزميات التعلم الالي لتقييم مخاطر الامن السبراني وهي خوارزمية آلة تعزيز التدرج الخفيف Light gradient boosting machine(LGBM)، خوارزمية AdaBoost، خوارزميه CatBoost، خوارزمية الشبكة العصبية الاصطناعية المتعدد الطبقات المدرك.

1. **خوارزمية الة تعزيز التدرج الخفيف (Light gradient boosting machine (LGBM)):** هو نهج تعلم آلي قائم على الأشجار تم تطويره حديثًا باستخدام إطار تعزيز التدرج لإنشاء شجرة القرار، يمكن لهذا النهج معالجة كل من مهام الانحدار والتصنيف. يمكن أن يعالج أيضًا التحدي الرئيس الذي يواجه في مناهج التعلم الآلي التقليدية، أي التعقيدات الحسابية، والتي تستغرق وقتًا طويلاً للغاية. هي طريقة سريعة وموزعة وذات كفاءة عالية تقلل من عدد عينات البيانات والميزات حيث يتضمن نموذج LGBM ثلاث خطوات رئيسية عند إنشاء شجرة القرار[25].

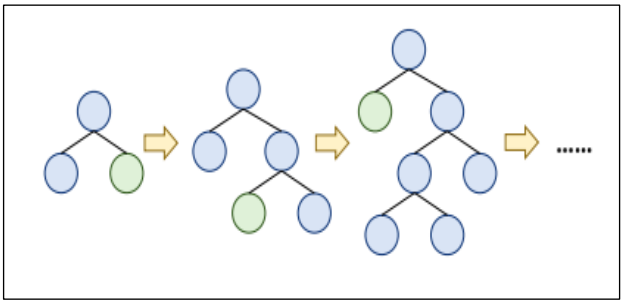
a. خوارزمية قائمة على المدرج التكراري: في هذه الخطوة يتم تحويل الميزات المستمرة إلى سلاسل مختلفة تُستخدم لإنشاء مخططات بيانية لمؤشر الميزات للعثور على أفضل نقطة تقسيم من الرسوم البيانية للميزات [25].

b. أخذ العينات على أساس التدرج من جانب واحد. يتم فرز عينات البيانات بترتيب تنازلي وفقًا لتدرجاتها، ويتم اختيار الجزء العلوي a منها كعينة مجموعة فرعية ذات تدرجات كبيرة. ثم يتم اختيار عينات b بشكل عشوائي من البيانات المتبقية كعينة فرعية ذات تدرجات صغيرة. يتم ضرب البيانات المأخوذة من العينات ذات التدرجات الصغيرة بمعامل الوزن $(1 - ab)$. وبالتالي، يتم تعلم المصنف الجديد وإنشاءه باستخدام البيانات حتى يتم التقارب بين البيانات المدربة [25].

c. تجميع الميزات الحصرية وفي هذه المرحلة ينشأ الرسم البياني ذو الحواف الموزونة، ويتوافق كل وزن مع إجمالي عدد التعارضات بين ميزتين. ثم يتم فرز الميزات بترتيب تنازلي وفقًا لدرجة كل ميزة (كلما زادت الدرجة، زاد التعارض مع النقاط الأخرى). أخيرًا، يتم التحقق من كل ميزة في التسلسل الذي تم فرزها، ويتم تعيينها إلى مجموعة مع تعارضات صغيرة أو يتم إنشاء مجموعة جديدة [25]. ولقد اثبتت هذه الطريقة فعاليتها وتعمل بشكل اسرع من الطرق التقليدية ، يوضح الشكل (2) نمو الاوراق ويوضح الشكل (3)مستوى نمو الاوراق.



الشكل2. نمو الاوراق

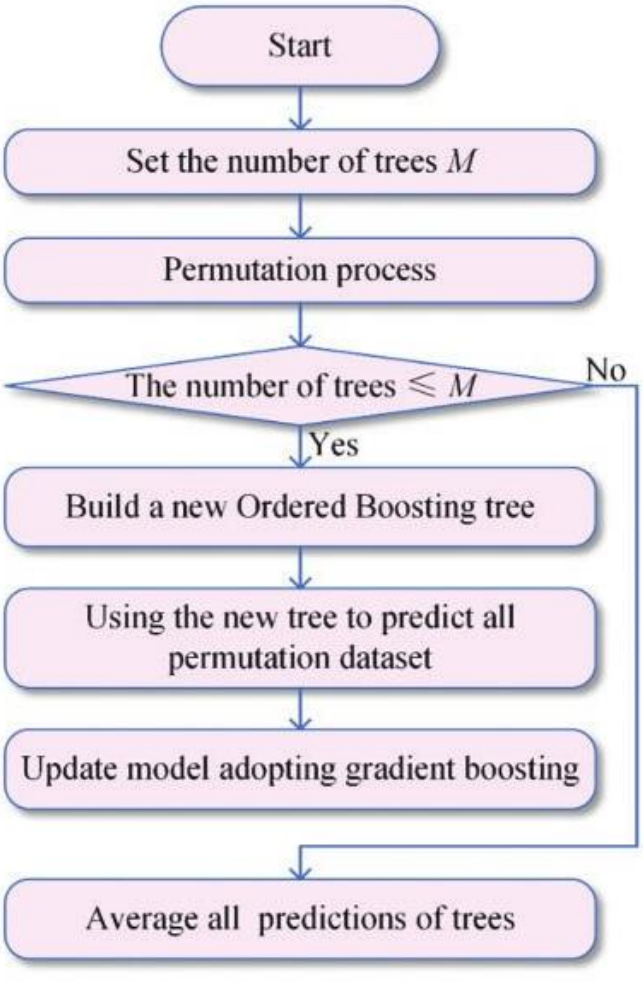


الشكل 3. مستوى نمو الاوراق

3. **خوارزمية AdaBoost :** يحافظ AdaBoost على توزيع الاوزان (weight) بحيث يتم توزيع الاوزان على العينات بشكل موحد, تستدعي AdaBoost خوارزمية Component Learn بشكل متكرر في سلسلة من الدورات ,في كل دورة يوفر AdaBoost عينات تدريبية مع توزيع وزن على كل مكونات التعلم بعد ذلك تقوم Component Learn بتدريب المصنف، ثم تحدث الاوزان بعد كل دورة وفقا لنتائج التنبؤ على عينات التدريب، العينات السهلة المصنفة بشكل صحيح تحصل على اوزان اقل والعينات الصلبة التي تم تصنيفها بشكل خاطئ تحصل على اوزان اعلى وبالتالي يركز AdaBoost على العينات ذات الاوزان الاعلى والتي تبدو اصعب بالنسبة لمكونات التعلم، تستمر هذه العملية لكل الدورات واخيرا، يجمع AdaBoost جميع فئات المكونات بشكل خطي في فرضية نهائية واحد، وتتمثل الخاصية النظرية المهمة لـAdaBoost أنه إذا كانت فئات المكون لديها دقة أفضل من النصف بقليل فإن خطأ التدريب في الفرضية النهائية ينخفض إلى الصفر أضعافًا مضاعفة بسرعة وهذا يعني أن مصنفات المكونات يجب أن تكون أفضل بقليل من المصنفات العشوائية [26].الخطوات الاتية تمثل خوارزمية AdaBoost [27]:

1. الخطوة الاولى : ادخال مجموعة البيانات D حيث D تمثل {(a1,c1),(a2,c2)….(an, cn)} ، ثابت التعلم الاساسي (L) ، عدد مرات التعلم (T).
2. الخطوة الثانية : تهيئة الاوزان (w) لجميع عينات التدريب حيث $w_i = 1/N$ ، $i=1,2,\ldots,N$
3. الخطوة الثالثة : تنفيذ الخطوات التالية من t=1 الى T
   - تدريب ثابت التعلم الاساسي h_t من D الى D_t باستخدام الاوزان w_i حيث $h_t=L(D,D_t)$
   - حساب الخطأ لل h_t $$err=\frac{\sum_{i=1}^{N} w_i I(h_t(a_i)\neq C_i)}{\sum_{i=1}^{N} w_i} \quad (4)$$
   - حساب الوزن لل h_t $$\alpha_t= \log\left(\frac{1-err_t}{err_t}\right) \quad (5)$$
   - $$W_i=w_i * \exp[\alpha_t I(h_t(a_i)\neq c_i) \quad (6)$$
4. الخطوة الرابعة : $$H(a)=Sign\sum_{t=1}^{T} \alpha t\, h t(a) \quad (7)$$

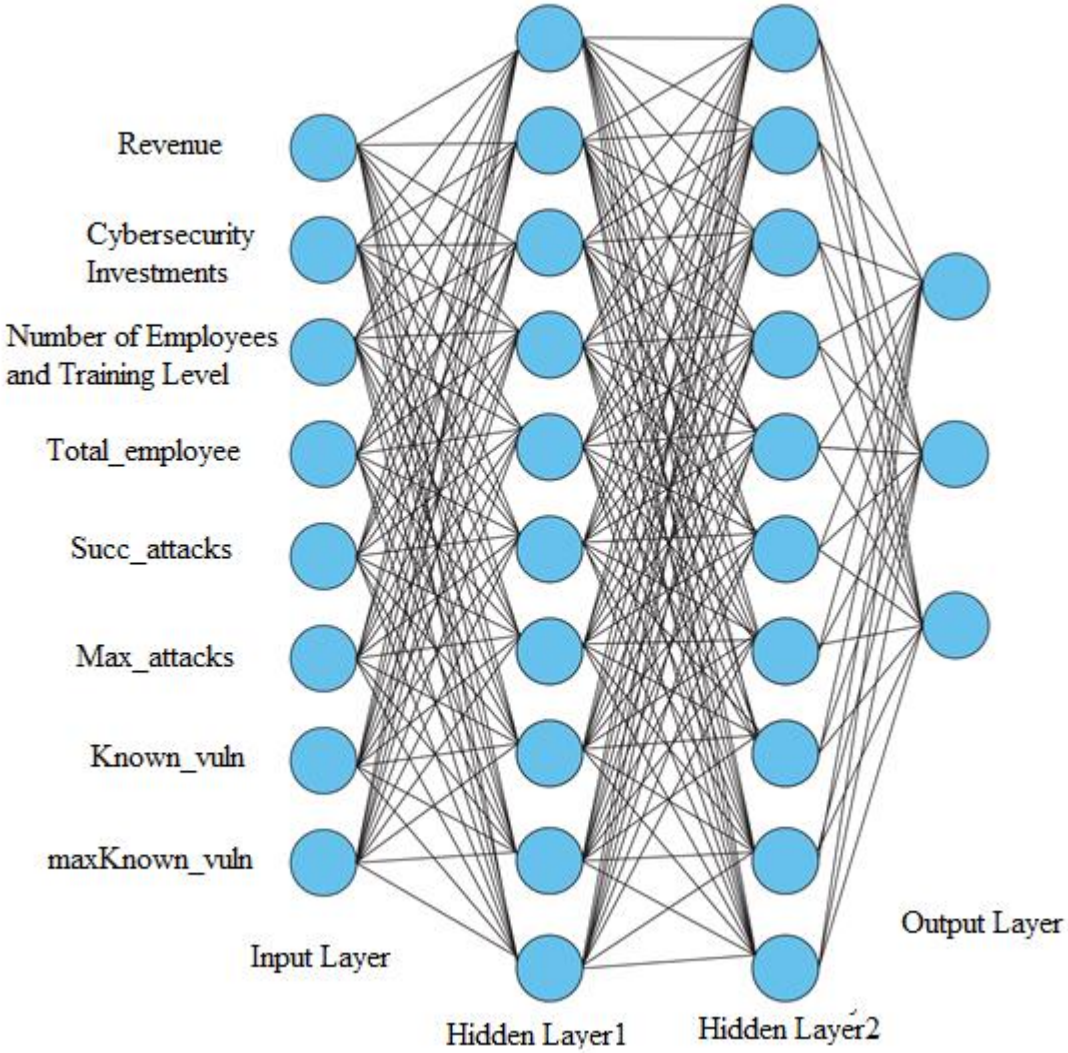
3. خوارزمية CatBoost: تُستخدم خوارزمية CatBoost كأداة لغة الآلة لتدريب مجموعات البيانات على تصنيف الأخطاء لتحسين أدائها، وسهولة الاستخدام، والمعالجة التلقائية للميزات الفئوية على تقنيات لغة الماكينة الأخرى (على سبيل المثال: (personal component analysis (PCA). كما أنه لا يتطلب معالجة مسبقة صريحة للبيانات لتحويل جميع فئات بيانات الأعطال إلى أرقام. تعد CatBoost أداة جيدة للغة الآلة لحل البيانات غير المتجانسة والصاخبة والمتغيرات المعقدة. تستخدم أشجار القرار الثنائية كمتنبئات أساسية، وله خصائص قوية لتقليل ضبط المعلمة الفائقة ، وتقليل فرص فرط البيانات. فهو يجمع بين شجرة قرار تعزيز التدرج (Gradient Boosting Decision Tree) والميزات الفئوية ، ويركز على المتغيرات الفئوية ، ويتعامل مع مشاكل انحياز التدرج والتنبؤ [28].
يوضح الشكل (4) مخطط لخوارزمية CatBoost :



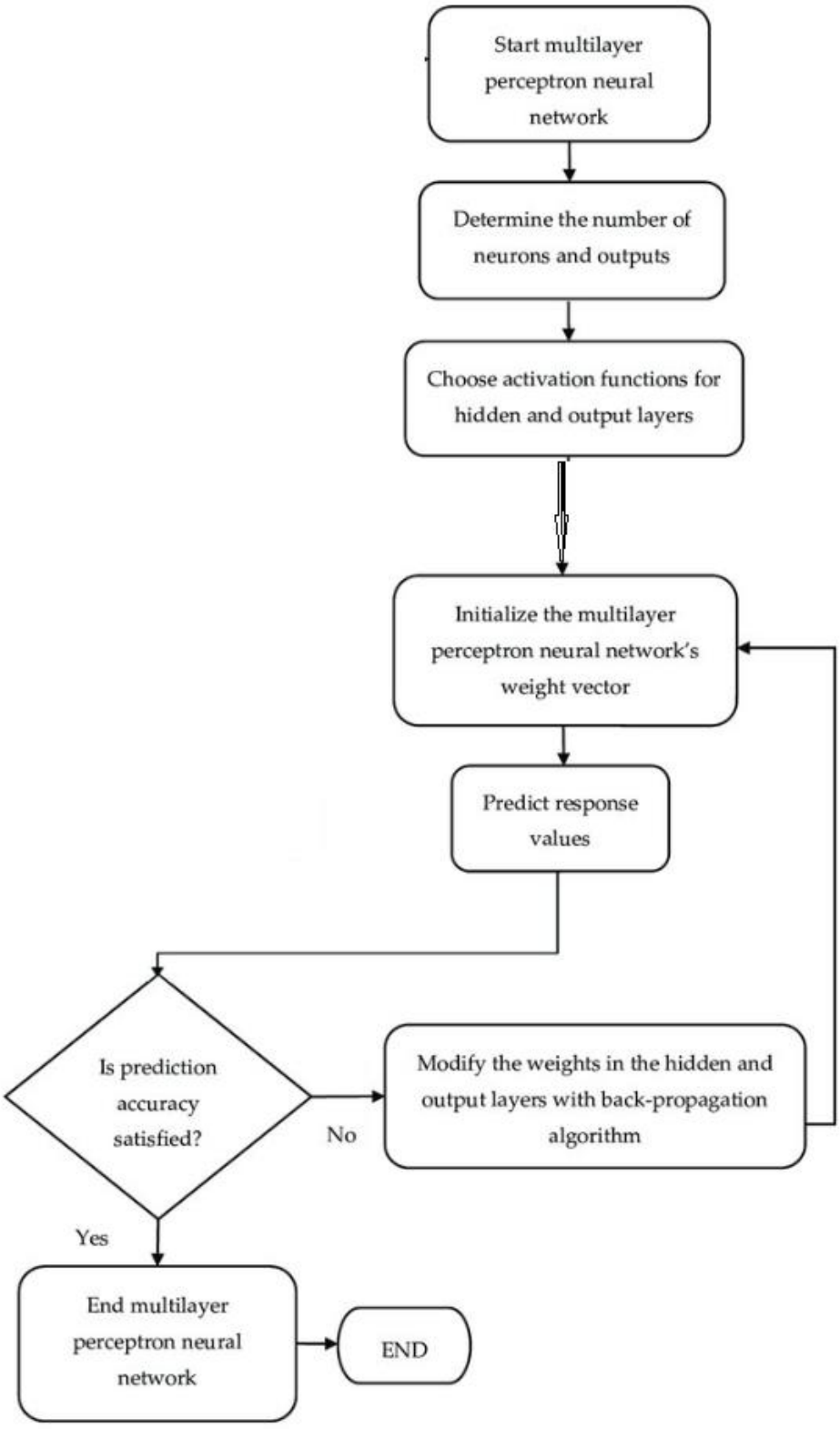
الشكل 4 .مخطط لخوارزمية CatBoost [29]

4. خوارزمية الشبكة العصبية الاصطناعية المتعددة الطبقات المدرك Multi Layer Perceptron (MLP)
هي خوارزمية تصنيف وهي فئة من وحدة التغذية الامامية ، فهي ترث خصائص الشبكات العصبية الاصطناعية ، مثل طبقة الإدخال ، والطبقة (الطبقات) المخفية، وطبقة الإخراج ، ووظائف الإدراك والتفعيل. يعطي الشكل (5) تمثيلًا مرئيًا مبسطًا لنموذج MLP الذي تم إنشاؤه في هذا البحث والشكل (6) يوضح سير العمل في

هذه الخوارزمية . تتوافق كل عقدة في طبقة الإدخال مع ميزة محددة لمجموعة البيانات التي أنشئت. يحتوي نموذج MLP على عدد إجمالي من طبقتين مخفيتين مع عشر خلايا عصبية لكل منهما. يعد اختيار أفضل المعلمات لشبكة ANN مهمة صعبة للغاية. من ناحية أخرى، تم تحديد طبقة المخرجات بناءً على فئات مخرجات النموذج (أي منخفضة ومتوسطة وعالية). لذلك فهو يتكون من ثلاث خلايا عصبية تمثل كل حالة تصنيف ممكنة. خلال مرحلة التدريب، يستخدم MLP تقنية تسمى backpropagation، تنشر ANN بيانات الإدخال للأمام عبر الخلية العصبية باتجاه طبقة الإخراج حيث يحدث التنبؤ، تشير خوارزمية الانتشار العكسي إلى عملية نشر المعلومات حول خطأ التنبؤ مرة أخرى من طبقة المخرجات عبر الشبكة بأكملها من أجل ضبط الأوزان وتحسين الدقة [30].



الشكل 5. معمارية الشبكة العصبية الاصطناعية المتعددة الطبقات المدرك



الشكل 6. مخطط للشبكة العصبية الاصطناعية المتعددة الطبقات المدرك[31]

3.5. مقاييس التقييم

ان الغاية من المقاييس معرفة مدى دقة خوارزميات التعلم الالي في تقييم المخاطر وذلك من خلال المقارنة بالتقييم الحاصل بالبيانات مع التقييم المتوقع من خلال الخوارزميات باستخدام مقاييس التقييم وهناك العديد من مقاييس التقييم وقد استخدمت خمسة مقاييس وهي الاكثر شيوعا واستخداما ولغرض المقارنة مع اعمال سابقة تم استخدام نفس المقاييس، وان الرموز المستخدمة في المعادلات موضحة بالشكل الاتي:

1. True Positive (TP): هو التوقع الايجابي الحقيقي للخطر.
2. True Negative (TN): هو التوقع السلبي الحقيقي للخطر.
3. False Positive (TP): هو التوقع الايجابي الكاذب للخطر.
4. False Negative (FN): هو التوقع السلبي الكاذب للخطر.

3.5.1 الدقة (Accuracy)

هي العدد الكلي للتنبؤات الصحيحة مقسوما على العدد الكلي للتنبؤات التي تم اجراؤها على مجموعة البيانات [32] ، ادق دقة هي 1 بينما اقل دقة هي 0 ، والتي يمكن حسابها عن طريق المعادلة التالية :

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

3.5.2 الدقة (Precession)

هي نسبة التوقعات الايجابية الصحيحة (TP) مقسوما على العدد الكلي للتوقعات الايجابية [32] وافضل دقة هي 1 واسوأ دقة هي 0 ، والتي يمكن حسابها عن طريقة المعادلة التالية :

$$\text{precession}=\frac{TP}{TP+FP} \quad (9)$$

3.5.3 معدل الاستدعاء (Recall)

هي نسبة التوقعات الايجابية الصحيحة (TP) مقسوما على التوقعات الايجابية الصحيحة مع التوقعات السلبية الكاذبة [32] وكما في المعادلة التالية :

$$\text{Recall}=\frac{TP}{TP+FN} \quad (10)$$

3.5.4 درجة F1- (F1-score)

هو المتوسط التوافقي للدقة والاستدعاء [2] والتي يمكن حسابها عن طريق المعادلة التالية:

$$\text{F1-score}=2*\frac{\text{precession}*\text{Recall}}{\text{precession}+\text{Recall}} \quad (11)$$

3.5.5 مصفوفة الارتباك (Confusion Matrix)

هو عبارة عن جدول يعرض فيه نتائج التوقع للتصنيف حيث يلخص فيه قيم التوقع الصحيحة والخاطئة من خلال المقارنة مع قيم التدريب والتي يتم وصفها بالصواب والخطأ مع قيم التنبؤ التي يتم وصفها بالايجابية والسلبية [33]. في هذا البحث استخدمت مصفوفة الارتباك لمعرفة تقييم الخطر حيث يوضح الجدول (7) الاتي نموذج مصفوفة الارتباك.

الجدول 7. مصفوفة الارتباك

	Prediction		
Actual	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	False Positive (FP)

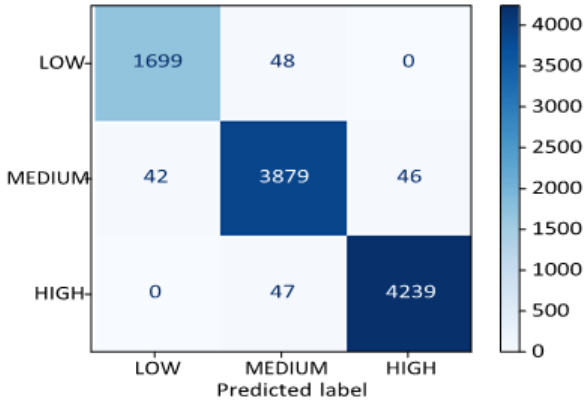
4. مناقشة النتائج

الجدول (8) يوضح نتائج تنفيذ الخوارزميات على البيانات التي تم توليدها لتقييم مخاطر الامن السيبراني

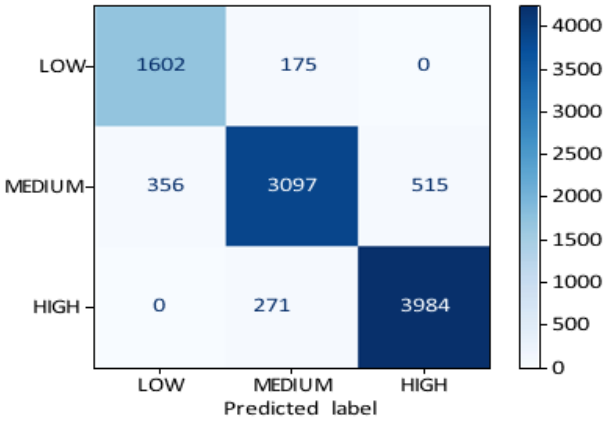
الجدول 8. نتائج تنفيذ الخوارزميات على البيانات التوليدية

	Accuracy	Precession	Recall	F1-score
LGBM	98.17 %	98 %	98 %	98 %
AdaBoost	86.83 %	86 %	87.33 %	86.33 %
CatBoost	99.01 %	99 %	99 %	99 %
MLP	99.45 %	99.33 %	99.33 %	99.33 %

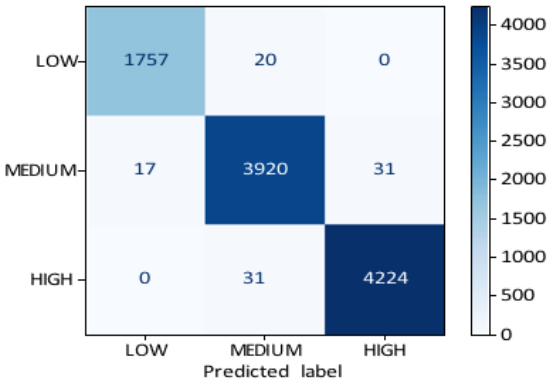
تم اجراء تقييم كمي للبيانات عن طريق مقياس مصفوفة الارتباك للبيانات التي تم توليدها وعمل مقارنة بين النماذج حيث يمثل الشكل (7) مصفوفة الارتباك لخوارزمية LGBM والشكل (8) يمثل مصفوفة الارتباك لخوارزمية AdaBoost والشكل (9) يمثل مصفوفة الارتباك لخوارزمية CatBoost والشكل (10) يمثل مصفوفة الارتباك لخوارزمية MLP.



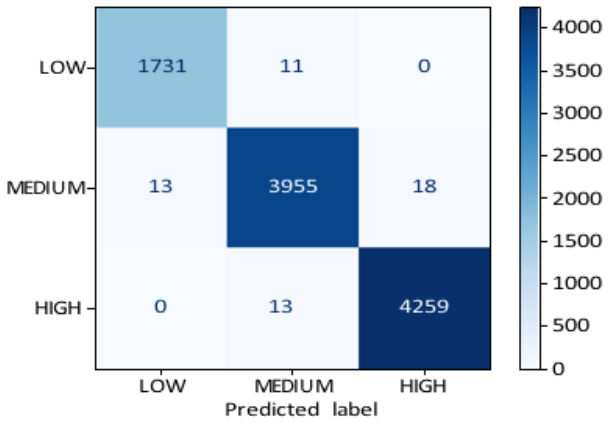
الشكل 7. مصفوفة الارتباك لخوارزمية LGBM



الشكل 8. مصفوفة الارتباك لخوارزمية AdaBoost



الشكل 9. مصفوفة الارتباك لخوارزمية CatBoost



الشكل 10. مصفوفة الارتباك لخوارزمية MLP

يعرض الجدول (9) مقارنة العمل المقترح مع الاعمال السابقة وذلك لاستخدام نفس قاعدة البيانات التوليدية .

الجدول 9. مقارنة النتائج مع الاعمال السابقة

	Algorithm	Accuracy
Proposed Work	LGBM	98.17 %
	AdaBoost	86.83 %
	CatBoost	99.01 %
	MLP	99.45 %
M. Franco [19]	MLP	98.86 %
	DT	92.64 %
	SVM	99.03 %
	K-NN	95.82 %

نلاحظ من الجدول (9) انه نتيجة خوارزمية MLP المقترحة افضل من الدراسة السابقة[19] والسبب يعود الى استخدام 10 خلايا عصبية في كل طبقة مخفية اما في الدراسة السابقة فقد استخدمت 5 خلايا عصبية .

5. الاستنتاجات

في هذه الدراسة تم توليد 50,000 عينة بالاعتماد على 8 متغيرات وتم تقييمها بالاعتماد على المعادلة (1) وتقسيمها الى 3 فئات واطي ومتوسط وعال واستخدام اربع خوارزميات من تقنيات التعلم الالي لتدريب البيانات واختبارها, حيث استخدمت خوارزمية Adaboost وكانت دقة هذه الخوارزمية % 86.83 واستخدام خوارزمية LGBM وكانت دقة الخوارزمية % 98.17 واستخدام خوارزمية CatBoost وكانت دقة الخوارزمية % 99.01 واستخدام خوارزمية MLP وكانت دقة الخوارزمية % 99.45 نستنتج من هذا البحث ان افضل خوارزمية لتقييم المخاطر هيMLP لحصولها على أعلى دقة في تقييم المخاطر ونقترح بالمستقبل استخدام التعلم العميق مع البيانات التوليدية ومقارنتها مع نتائج التعلم الآلي في عملية تقييم المخاطر .

شكر وتقدير

الشكر والتقدير الى جامعة الموصل / كلية علوم الحاسوب والرياضيات على مرافقهم التي ساعدت في تحسين جودة هذا العمل.

المصادر

1. Maleks Smith, Z., E. Lostri, and J.A. Lewis. "The hidden costs of cybercrime".2020 available from:https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf. [Accessed 16 May 2021].
2. B. Sheehan, F. Murphy, M. Mullins, and C. Ryan, 'Connected and autonomous vehicles: A cyber-risk classification framework', Transp. Res. Part A Policy Pract., vol. 124, pp. 523–536, Jun. 2019. https://doi.org/10.1016/j.tra.2018.06.033.
3. C. Biener, M. Eling, and J. H. Wirfs, 'Insurability of cyber risk: An empirical analysis', Geneva Pap. Risk Insur. Issues Pract., vol. 40, no. 1, pp. 131–158, Jan. 2015.https://doi. org/10.1057/gpp.2014.19
4. M. Eling and W. Schnell, 'What do we know about cyber risk and cyber risk insurance?', J. Risk Finance, vol. 17, no. 5, pp. 474–491, Nov. 2016. https://doi.org/10.1108/jrf-09-2016-0122.

5. G. Falco et al., ‘Cyber risk research impeded by disciplinary barriers’, Science, vol. 366, no. 6469, pp. 1066–1069, Nov. 2019.
6. Y.-Y. Leong and Y.-C. Chen, ‘Cyber risk cost and management in IoT devices-linked health insurance’, Geneva Pap. Risk Insur. Issues Pract., vol. 45, no. 4, pp. 737–759, Oct. 2020.https://doi. org/10.1057/s41288-020-00169-4
7. UNODC," Vulnerability disclosure",2019, Available from: https://www.unodc.org/e4j/ar/ cybercrime/module-9/key-issues/vulnerability-disclosure.html
8. Chio.C, Machine learning and security: Protecting systems with data and algorithms. O’Reilly Media, Inc, 2018.
9. Kumar, G., Saini, D., and Cuong, N., Cyber Defense Mechanisms Security, Privacy, and Challenges, CRC Press, 2020.
10. Coombs, T., Artificial Intelligence & Cybersecurity for dummies, John Wiley & Sons, 2018.
11. "Cisco Annual Internet Report (2018–2023) White Paper," 2021.
12. M. S. Ben Mahmoud, N. Larrieu and A. Pirovano, "A Risk Propagation Based Quantitative Assessment Methodology for Network Security - Aeronautical Network Case Study," *2011 Conference on Network and Information Systems Security*, pp. 1-9,2011, doi: 10.1109/SAR-SSI.2011.5931372.
13. Kelleher, J., Namee, B., and D'arcy, A., Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, MIT press, 2020.
14. Parisi, A., Hands-On Artificial Intelligence for Cybersecurity: Implement smart AI systems for preventing cyber-attacks and detecting threats and network anomalies, Packt Publishing Ltd, 2019.
15. Konar, A., Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain, CRC press, 2018.
16. Y. Castro and Y. J. Kim, ‘Data mining on road safety: factor assessment on vehicle accidents using classification models’, Int. J. Crashworthiness, vol. 21, no. 2, pp. 104–111, Mar. 2016.
17. Q. Zhang, C. Zhou, Y.-C. Tian, N. Xiong, Y. Qin, and B. Hu, ‘A fuzzy probability Bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems’, IEEE Trans. Industr. Inform., vol. 14, no. 6, pp. 2497–2506, Jun. 2018.
18. X. Zhao, Q. Chen, J. Xue, Y. Zhang, and J. Zhao, ‘A method for calculating network system security risk based on a lie group’, IEEE Access, vol. 7, pp. 70610–70623, 2019.
19. M. Franco, E. Sula, B. Rodrigues, E. Scheid, and B. Stiller, ProtectDDoS: A Platform for Trustworthy Offering and Recommendation of Protections; International Conference on Economics of Grids, Clouds, Software and Services, vol. 2020. Izola, Slovenia, pp. 1–12, 2020.
20. V. S. Kumar and V. L. Narasimhan, ‘Using deep learning for assessing cybersecurity economic risks in virtual power plants’, in 2021 7th International Conference on Electrical Energy Systems (ICEES), Chennai, India, 2021, doi: 10.1109/ICEES51510.2021.9383723.
21. M. Kalinin, V. Krundyshev, and P. Zegzhda, ‘Cybersecurity risk assessment in smart city infrastructures’, Machines, vol. 9, no. 4, p. 78, Apr. 2021.
22. M. van Haastrecht, I. Sarhan, A. Shojaifar, L. Baumgartner, W. Mallouli, and M. Spruit, ‘A Threat-Based Cybersecurity Risk Assessment Approach Addressing SME Needs’, in Proceedings of the 16th International Conference on Availability, Reliability and Security, Vienna, Austria, 2021.
23. Z. G. Chen, H. S. Kang, S. N. Yin, S. R. Kim, “Automatic Ransomware Detection and Analysis Based on Dynamic API Calls Flow Graph,” Proceedings of the International Conference on Research in Adaptive and Convergent Systems, 2017.
24. Abdullah, Mohammed Hamid Abdulraheem. Designing Deep Learning Based Network Intrusion Detection System for Software Defined Network. Diss. University of Mosul, 2020.
25. G. Ke et al., ‘LightGBM: a highly efficient gradient boosting decision tree’, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA pp. 3149–3157, 2017.
26. R. E. Schapire and Y. Singer, ‘Improved boosting algorithms using confidence-rated predictions’, in Proceedings of the eleventh annual conference on Computational learning theory - COLT’ 98, Madison, Wisconsin, United States, 1998.
27. P. Bahad and P. Saxena, ‘Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics’, in International Conference on Intelligent Computing and Smart Communication 2019, 2020, pp. 235–244.
28. Y. Zhang, Z. Zhao, and J. Zheng, ‘CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China’, Journal of Hydrology, vol. 588, p. 125087, Sep. 2020.

29. Armaghani, Liu, Z., Fakharian ,D. J., D , P., Li., V. Ulrikh, D., N.,Orekhova, N., & Khedher, K. M, Rock strength estimation using several tree-based ML techniques. *CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES*, vol *133 no* 3,pp. 799-824,2020.
30. Thomas Wood: What is Backpropagation? https://deepai.org/machine-learning-glossary-and-terms/backpropagation, [Accessed: 21-Dec-2022].
31. D. K. Agustika, N. A. Ariyanti, I. N. K. Wardana, D. D. Iliescu, and M. S. Leeson, 'Classification of chili plant origin by using multilayer perceptron neural network', in *2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, (pp. 365-369). Semarang, Indonesia, 2021.
32. Ž. Đ. Vujovic, 'Classification model evaluation metrics', Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, 2021.
33. A. K. Santra and C. J. Christy, 'Genetic algorithm and confusion matrix for document clustering', International Journal of Computer Science Issues (IJCSI), vol. 9, no. 1, 2012.

تصميم وتنفيذ أداة التعلم الآلي لتقييم مخاطر الأمن السيبراني

عمر ابراهيم شيت، لهيب محمد ابراهيم

قسم البرمجيات، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة

زادت الهجمات الإلكترونية من حيث العدد والشدة ، مما أثر سلبيا على الأعمال التجارية وخدماتها. على هذا النحو ، لم يعد اعتبار الأمن السيبراني مجرد مشكلة تكنولوجية ، ولكن يجب أيضًا اعتباره أمرًا بالغ الأهمية للاقتصاد والمجتمع. تكافح الحلول الحالية للعثور على مؤشرات للمخاطر غير المتوقعة ، مما يحد من قدرتها على إجراء تقييمات دقيقة للمخاطر. تقدم هذه الدراسة طريقة لتقييم المخاطر بالاعتماد على التعلم الآلي ((Machine Learning(ML) ، وهو نهج يستخدم للتقييم والتنبؤ بمدى تعرض الشركات لمخاطر الأمن السيبراني. لهذا الغرض ، تم تنفيذ أربع خوارزميات من خوارزميات التعلم الالي -light gradient boosting) AdaBoost-CatBoost -Multi Layer Perceptron(MLP)) وتدريبها وتقييمها باستخدام مجموعات بيانات توليدية تمثل خصائص أحجام مختلفة من البيانات (على سبيل المثال ، وعدد الموظفين وقطاع الأعمال ونقاط الضعف المعروفة وخبراء الامن السيبراني). يُظهر التقييم الكمي الذي أجري على هذا الدراسة الدقة العالية لنماذج ML وخصوصا MLP كانت افضل دقة عند عمل مقارنة مع الاعمال السابقة.