مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (١) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (1) 2025

دراسة تأثير استخلاص الميزات باستخدام نهج التعلم الجماعي للكشف عن هجمات الأمن السيبراني

* م. ريم مالك إبراهيم

(تاريخ الإيداع ٢٠٢٤/١١/٦ . قُبِل للنشر في ٢٠٢٥/١/١٤)

□ ملخّص □

أصبح الأمن السيبراني من المجالات البالغة الأهمية في العصر الرقمي، حيث يؤدي الاعتماد المتزايد على الأنظمة القائمة على الانترنت وانتشار أجهزة انترنت الأشياء (IOT) إلى تعريض الأفراد والمؤسسات للهجمات الإلكترونية . تجاوز ظهور الهجمات الإلكترونية المتطورة تدابير الأمن التقليدية مما يجعل من الضروري تطوير أدوات متقدمة للكشف عن هذه التهديدات والتخفيف عنها. في الوقت الحالي يتم استخدام الأنظمة الخبيرة وخوارزميات التعلم الآلي على نطاق واسع في مجال كشف الاختراقات على الشبكة.

في هذا البحث تم كشف هجمات الأمن السيبراني الممندرجة ضمن قاعدة البيانات CSE-CIC-IDS2018 باستخدام النهج الجماعي في عملية استخلاص الميزات و باستخدام أربع خوارزميات تصنيف وهي Desion Tree(DT) و Random Forest (RF) و Naïve Bayes (NB) و Regression Linear(RL)، توصل البحث إلى أن استخدام النهج الجماعي أعطى أفضل قيم من أجل مقاييس الأداء Auc و F1-Score بدلا من استخدام تقنية استخلاص واحدة ، كما أعطت خوارزمية التصنيف RF أعلى القيم بالمقارنة مع خوارزميات التصنيف المستخدمة الأخرى وتقنيات استخلاص الميزات المفردة المستخدمة في الدراسات السابقة.

الكلمات المفتاحية: الأمن السيبراني، استخلاص الميزات ، خوارزميات التصنيف ، النهج الجماعي، مصفوفة الدقة.

* ماجستير في قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس–سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (١) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (1) 2025

Studying the effect of feature extraction using ensemble learning approach for detecting cybersecurity attacks

Eng.Reem Malek Ibrahim *

(Received 6/11/2024 . Accepted 14/1/2025)

□ ABSTRACT □

Cybersecurity has become a critical area in the digital age, as the increasing reliance on Internet-based systems and the proliferation of Internet of Things (IOT) devices expose individuals and organizations to cyberattacks. The emergence of sophisticated cyberattacks has bypassed traditional security measures, making it necessary to develop advanced tools to detect and mitigate these threats. Expert systems and machine learning algorithms are currently widely used in the field of network intrusion detection.

In this research, cybersecurity attacks included in the CSE-CIC-IDS2018 database were detected using an ensemble approach in the feature extraction process and using four classification algorithms, namely Desion Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Regression Linear (RL). The research concluded that using the ensemble approach gave the best values for the performance metrics Auc and F1-Score instead of using one extraction technique, and the RF classification algorithm gave the highest values compared to other classification algorithms and single feature extraction techniques used in previous studies.

Key Words: Cybersecurity, Feature Extraction, Classification Algorithms, Ensemble Approach, Confusion Matrix.

* Master of Information Technology Engineering Department, Faculty of Information and Communication Technology Engineering, Tartous University.

١. المقدمة

الأمن السيبراني هو مجال يركز على حماية الأنظمة والشبكات والبيانات من الهجمات الرقمية. يهدف إلى ضمان سلامة المعلومات وسرية البيانات وتوافر الخدمات. تكمن أهمية الأمن السيبراني في حماية البيانات الحساسة مثل البيانات الشخصية والمالية، بالإضافة الى تأمين الشبكات من خلال منع الوصول غير المصرح به إلى الشبكات.

توجد أنواع متعددة من التهديدات التي تؤثر على الأمن السيبراني، مثل البرمجيات الخبيثة بما في ذلك الفيروسات والديدان وبرامج الفدية، بالإضافة إلى الهجمات الإلكترونية مثل هجمات حجب الخدمة DDoS و هجمات التصيد الاحتيالي، كما تشمل التهديدات الهجمات الداخلية التي قد تأتي من موظفين أو متعاقدين غير موثوق بهم. لذلك، من الضروري تنفيذ استراتيجيات متعددة لحماية البيانات ومنع التهديدات بمختلف أشكالها. تشمل هذه الاستراتيجيات تشفير البيانات لحماية المعلومات أثناء النقل والتخزين، واستخدام جدران الحماية لمنع الوصول غير المصرح به إلى الشبكة، وبرامج مكافحة الفيروسات للكشف عن البرمجيات الخبيثة وإزالتها، بالإضافة إلى التحديثات المنتظمة لضمان أن الأنظمة والبرامج تبقى محدثة وآمنة.

يمكن لخوارزميات التعلم الآلي المدربة بكفاءة على مجموعات البيانات اكتشاف التطفل وحركة مرور الشبكة القادرة على تعريض نظام المعلومات للحظر. إذ تتميز هذه الخوارزميات عادةً بتفوقها على الأساليب الإحصائية التقليدية في مهام التصنيف ومع ذلك، قد لا تكون إعدادات العتبة لبعض الخوارزميات مناسبة للبيانات غير المتوازنة، مما يؤدي إلى عدم فعاليتها في التمييز بين فئات الأغلبية والأقلية في بيئات غير متوازنة بشكل كبير. لذا، من الضروري استخدام مقاييس تحمي من هذه النتيجة. في هذا البحث، تم استخدام مقياسي درجة F1 و (AUC)، اللذين يعدان مناسبين لتقييم أداء المصنفات على مجموعات البيانات. ومن المهم ملاحظة أن اختلال التوازن بين الفئات يكون أكثر وضوحًا في البيانات الضخمة، حيث يكون عدد حالات فئة الأغلبية مرتفعًا بشكل غير متناسب في تلك البيئة.

قد تواجه الأساليب التقليدية تحديات في التعامل مع كميات كبيرة من البيانات، وتنوع تنسيقاتها، وسرعة تدفقها من مصادر متعددة، بالإضافة إلى التناقضات في تدفق البيانات، وتصنيف المعلومات المهمة، وربط البيانات وتحويلها. في هذا البحث يعتمد أداء المصنف على تدريب واختبار عدة خوارزميات، وهي: شجرة القرار DT والغابة العشوائية RF وخوارزمية NB، والانحدار اللوجستي .LRتم اختيار هذه الخوارزميات نظرًا لتغطيتهم الجيدة لمجموعة متنوعة من نماذج التعلم الآلي واعتبارهم موثوقين من حيث الأداء [١].

كما تم تطبيق عملية تنظيف البيانات واستخلاص الميزات قبل البدء بعملية التصنيف والتقييم ، في هذا البحث تم استخدام نهج التعلم الجماعي في عملية استخلاص الميزات بهدف تحسين أداء المصنف، تساهم عملية استخلاص الميزات في توضيح البيانات بشكل أكبر بالإضافة الى التقليل من متطلبات الحساب.

١-١ الدراسات المرجعية:

قامت العديد من الأبحاث باستخدام تقنيات الذكاء الاصطناعي في عملية كشف الاختراقات على مجموعات بيانات معيارية فيما يلي بعض من الدراسات في هذا المجال:

في الدراسة [1] تم استخدام نهج التجميع لمقارنة ٦ خوارزميات تصنيف هي:

DT, NB, RF, LR, Gradiant Boosting(GB), quadratic Discrimingant

في مرحلة المعالجة الأولية تم إزالة العينات ذات القيم المفقودة واللانهائية وإزالة السجلات المكررة، ثم تم تقسيم مجموعة البيانات الى مجموعتين للتدريب والاختبار بنسبة ٨٠ الى ٢٠، تم استخلاص الميزات باستخدام معامل الارتباط وchi-squared. أعطت كل من الخوارزميات التالية أعلى دقة تصنيف DT بقيمة 97.10على التوالي، وبقيمة AUC وصلت إلى 0.94.

في حين تم استخدام مصنف مشفر ذاتي يقوم بترميز البيانات بطريقة تؤدي الى تقليل الأبعاد [2]، في مرحلة المعالجة المسبقة تم استبدال قيم ال infinityو NANبقيمة الصفر، أما عملية استخلاص الميزات تمت باستخدام تقنية RF. تم تقييم النموذج المقترح باستخدام مقياس AUC، تمت مقارنة النتائج مع مشفر تلقائي آخر يدعى Kitnet، وأظهرت النتائج أن وقت اكتشاف الهجمات باستخدام النموذج المقترح كان أسرع من النموذج الآخر.

قامت دراسة أخرى باستخدام نهج أخذ العينات الناقصة واختيار الميزات المضمنة مع مصنف LightGBM، أثناء مرحلة تنظيف البيانات تمت إزالة القيم المفقودة والميزات غير المفيدة كما تم ترميز الهجمات بأرقام صحيحة. تم تقييم ستة خوارزميات تصنيف بالإضافة الى خوارزمية LightGMB، هذه الخوارزميات هي SVM, MLP, Adaboost,RF,CNN, NB وكانت نسبة تدريب البيانات الى اختبارها 70الى 30. تم اجراء عملية اختيار الميزات باستخدام XGBoost، وكان أداء خوارزمية LightGBM هو الأفضل حيث بلغت دقة الكشف 98.37 كما حققت هذه الخوارزمية ثاني أسرع وقت تدريب بين المصنفات المذكورة بعد مصنف CNN. [3]

كما تم استخدام مجموعتي البيانات CSE-CIS-IDS (2017) وCSE-CIS-IDS (2018) وعدة خوارزميات تصنيف لتقييم الأداء من بينها DT, XGBoost, KNN, Bag, LR وغيرها من الخوارزميات. أظهرت النتائج أن المصنفات القائمة على الشجرة حققت أفضل أداء واحتلت XGBoost المرتبة الأولى مع العديد من القيم المثالية ل F1 وAUC. وبمقارنة كلا قاعدتي البيانات توصل الباحثون إلى أن النموذج المستخدم من أجل مجموعة البيانات (2017) لا يمكن تعميمه على مجموعة البيانات الأخرى (2018).[4].

قامت دراسة أخرى بتقييم خمسة خوارزميات تصنيف على مجموعتي البيانات CSE-CIC-IDS2018 وISOT-HTTP-BoTnet لتحديد أفضل مصنف لشبكات الروبوتات. هذه الخوارزميات هي RF,DT,KNN,SVM,NB وعملية اختيار الميزات تمت باستخدام تقنية RF كما تم استخدام خوارزمية البحث الشبكي لتحسين الأداء، سجلت كل من RF أعلى دقة بالنسبة لكل من قاعدتي البيانات. [5]

٢. أهمية البحث وأهدافه:

يهدف هذا البحث إلى دراسة تأثير استخدام النهج الجماعي في عملية استخلاص السمات على دقة كشف الهجمات ضمن قاعدة البيانات CSE-CIC-IDS2018، يقدم البحث مساهمة جديدة في مجال كشف هجمات الأمن السيبراني من خلال استخدام النهج الجماعي عوضاً عن استخدام تقنية مفردة في عملية استخلاص السمات. تكمن أهمية البحث من خلال تطبيق نهج التعلم الجماعي الذي يجمع بين التنبؤات من نماذج متعددة وبالتالي الاستفادة من نقاط القوة في عدة خوارزميات مما يقلل التحيز والتباين وبالتالي الحصول على تنبؤات أكثر موثوقية بالإضافة إلى تحديد أكثر السمات ارتباطاً بالهدف مما ينعكس إيجاباً على دقة الكشف ومقاييس مصفوفة الدقة المستخدمة في تقييم خوارزميات التصنيف.

٣. طرائق البحث ومواده

٣-١ مجموعة البيانات المستخدمة Dataset (CSE-CIC-IDS2018)

هي مجموعة بيانات تستخدم لاكتشاف حالات التطفل تحوي على حالات طبيعية وحالات غير طبيعية لحركة مرور شبكة. تم تطويرها من قبل مركز أبحاث الأمن السيبراني CIC في كندا، وتُستخدم بشكل واسع في أبحاث الأمن السيبراني، وخاصة في مجال اكتشاف التسلل. [3]

تحتوي على٨٤ ميزة مختلفة. هذه الميزات تشمل معلومات حول حركة المرور في الشبكة، مثل: وقت الاتصال Connection Time، بروتوكول النقل Transport Protocol، حجم الحزمة Packet Size، عدد الحزم Number of Packets، معدل نقل البيانات Data Transfer Rate، نوع الهجوم Attack Type وغيرها من الميزات التي تساعد في تحليل حركة المرور وكشف التسلل. وهي موضحة في الجدول التالي:

الجدول (١): سمات قاعدة البيانات

No.	Feature	No.	Feature	No.	Feature
1	Flow ID	29	Fwd IAT Std	57	ECE Flag Count
2	Source IP	30	Fwd IAT Max	58	Down/Up Ratio
3	Source Port	31	Fwd IAT Min	59	Average Packet Size
4	Destination IP	32	Bwd IAT Total	60	Avg Fwd Segment Size
5	Destination Port	33	Bwd IAT Mean	61	Avg Bwd Segment Size
6	Protocol	34	Bwd IAT Std	62	Fwd Avg Bytes/Bulk
7	Time stamp	35	Bwd IAT Max	63	Fwd Avg Packets/Bulk
8	Flow Duration	36	Bwd IAT Min	64	Fwd Avg Bulk Rate
9	Total Fwd Packets	37	Fwd PSH Flags	65	Bwd Avg Bytes/Bulk
10	Total Backward Packets	38	Bwd PSH Flags	66	Bwd Avg Packets/Bulk
11	Total Length of Fwd Pck	39	Fwd URG Flags	67	Bwd Avg Bulk Rate
12	Total Length of Bwd Pck	40	Bwd URG Flags	68	Subflow Fwd Packets
13	Fwd Packet Length Max	41	Fwd Header Length	69	Subflow Fwd Bytes
14	Fwd Packet Length Min	42	Bwd Header Length	70	Subflow Bwd Packets
15	Fwd Pck Length Mean	43	Fwd Packets/s	71	Subflow Bwd Bytes
16	Fwd Packet Length Std	44	Bwd Packets/s	72	Init_Win_bytes_fwd
17	Bwd Packet Length Max	45	Min Packet Length	73	Act_data_pkt_fwd
18	Bwd Packet Length Min	46	Max Packet Length	74	Min_seg_size_fwd
19	Bwd Packet Length Mean	47	Packet Length Mean	75	Active Mean
20	Bwd Packet Length Std	48	Packet Length Std	76	Active Std
21	Flow Bytes/s	49	Packet Len. Variance	77	Active Max
22	Flow Packets/s	50	FIN Flag Count	78	Active Min
23	Flow IAT Mean	51	SYN Flag Count	79	Idle Mean
24	Flow IAT Std	52	RST Flag Count	80	Idle Packet
25	Flow IAT Max	53	PSH Flag Count	81	Idle Std
26	Flow IAT Min	54	ACK Flag Count	82	Idle Max
27	Fwd IAT Total	55	URG Flag Count	83	Idle Min
28	Fwd IAT Mean	56	CWE Flag Count	84	Label

تشمل قاعدة البيانات عدة أنواع من الهجمات منها ما هو قائم على الشبكة (مثل Dosو DDos) ومنها ما هو قائم على التطبيقات (مثل SQL Injection) بالإضافة الى بعض الهجمات المتقدمة (مثل Brute Force)، نوضح فيما يلي الهجمات بالتفصيل:

١. **هجمات الحرمان من الخدمة Dos:** تهدف إلى جعل الخدمة غير متاحة للمستخدمين الشرعيين عن طريق إغراق الخادم بطلبات زائدة.

٢. **الهجمات الموزعة للحرمان من الخدمة DDos:** مشابهة لهجمات DoS، ولكنها تُنفذ من عدة مصادر في وقت واحد، مما يجعل التصدي لها أكثر صعوبة.

٣. **الهجمات بالقوة الغاشمة Brute Force:** تستهدف كلمات المرور عن طريق تجربة جميع الاحتمالات الممكنة للوصول إلى حسابات المستخدمين.

٤. **هجمات الويب:** تشمل هجمات مثل SQL Injection و(XSS) Cross-Site Scripting، والتي تستهدف تطبيقات الويب.

٥. **التسلل Infiltration:** محاولة الوصول غير المصرح به إلى الشبكة أو النظام بهدف سرقة البيانات أو تنفيذ تعليمات ضارة.

٦. **الهجمات على الشبكات Network Attacks:** تشمل مجموعة من الهجمات التي تستهدف البنية التحتية للشبكة، مثل هجمات ARP Spoofing.

تساعد هذه الأنواع من الهجمات الباحثين والمطورين في فهم التهديدات المختلفة وتحسين أنظمة كشف التسلل من خلال تدريب النماذج على بيانات واقعية تحتوي على سيناريوهات مختلفة. فيما يلي جدول يوضح النسبة المئوية لعدد سجلات الهجمات في قاعدة البيانات المستخدمة:

الجدول (٢): النسبة المئوية لعدد سجلات الهجمات في قاعدة البيانات المستخدمة

اسم الهجوم	النسبة المئوية لعدد سجلات الهجوم
Benign	83%
DDos	8%
Dos	4%
Blute force	2.3%
Botnet	1.7%
Infilitration	0.9%
Web Attack	0.1%

من ميزات قاعدة البيانات أنها تحوي بيانات طبيعية بالإضافة إلى بيانات الهجمات مما يسمح بتقييم أداء أنظمة كشف التسلل كما أنها تتضمن بيانات من عدة بروتوكولات مثل TCP وUDP وICMP، مما يعكس سيناريوهات حقيقية بالإضافة إلى وجود وثائق مفصلة تشرح كيفية استخدام البيانات وكيفية إعداد بيئة الاختبار. تستخدم قاعدة البيانات للعديد من الأغراض منها تدريب نماذج التعلم الآلي واختبار أنظمة كشف التسلل وبشكل خاص في أبحاث الأمن السيبراني لفهم سلوك الهجمات وتحليلها.

٣-٢ نهج التعلم الجماعي:

هو تقنية في مجال تعلم الآلة تهدف إلى تحسين أداء النماذج من خلال دمج عدة نماذج، إذ يعتمد هذا النهج على فكرة أن مجموعة من النماذج يمكن أن تحقق أداء أفضل من نموذج واحد وذلك عن طريق تقليل الأخطاء وزيادة الدقة.[7][6]

الأنواع الرئيسية لنهج التعلم الجماعي:

١- **التصويت Voting:** يتم استخدام عدة نماذج مثل أشجار القرار ويتم اتخاذ القرار النهائي بناء على الأكثر تصويتا من قبل النماذج المختلفة.

٢- **التعلم بالتحسين Boosting:** يتم بناء نماذج جديدة تعتمد على الأخطاء التي ارتكبتها النماذج السابقة ومن الأمثلة عليه خوارزمية AdaBoost وGradient Boosting.

٣- **التعلم بالتكديس Stacking:** يتم تدريب عدة نماذج على مجموعة البيانات الأصلية ثم يتم استخدام مخرجات هذه النماذج كمدخلات لنموذج آخر يسمى المستوى الثاني لتقديم التنبؤ.

٤- **التعلم بالمتوسط Bagging:** يتم تدريب عدة نماذج على عينات مختلفة من البيانات باستخدام تقنية أخذ العينات مع الاستبدال، ثم يتم حساب المتوسط أو التصويت من هذه النماذج مثال على ذلك Random Forest.

يوجد العديد من الميزات والفوائد التي نحصل عليها بتطبيق نهج التعلم الجماعي وهي زيادة الدقة إذ تكون النماذج المجمعة غالبا أكثر دقة من النماذج الفردية، كما أنه يساهم في تقليل التباين في التنبؤات بالإضافة إلى أنه يجعل النموذج أكثر استقراراً وأقل عرضة للتقلبات الناتجة عن بيانات التدريب.

٣-٣ استخلاص الميزات:

تقنيات استخلاص الميزات تُقسم إلى عدة أنواع بناءً على طريقة عملها وأهدافها. فيما يلي الأنواع الرئيسية لتقنيات استخلاص الميزات:[9][8]

١. **تقنيات الفلتر Filter Methods** : تعمل هذه الأساليب على تقييم الميزات بشكل مستقل عن النموذج المستخدم، وتختار الميزات بناءً على معايير إحصائية معينة. يندرج تحت هذا النوع عدة تقنيات هي : اختبار الارتباط Correlation Tests ،اختبار كاي-تربيع Chi-Squared ،اختبار ANOVA .

٢. **تقنيات التغليف Wrapper Methods** : تقوم هذه الأساليب بتقييم مجموعة من الميزات بناءً على أداء نموذج معين. يتم استخدام النموذج لتحديد أي مجموعة من الميزات تعطي أفضل أداء. فيما يلي بعض التقنيات المندرجة تحت هذا النوع : الاختيار المتكرر Recursive Feature Elimination – RFE ،البحث العشوائي Random Search .

٣. **تقنيات التضمين Embedded Methods** : تجمع هذه الأساليب بين ميزات الفلتر والتغليف، حيث يتم اختيار الميزات أثناء تدريب النموذج نفسه. من بين تقنيات هذا النوع : شجرة القرار Decision Trees ، الانحدار اللوجستي معL1 (Lasso) Regularization .

٤. **تقنيات التعلم العميق:** تستخدم هذه الأساليب الشبكات العصبية لاستخلاص الميزات، حيث يمكن أن تتعلم الشبكة بشكل تلقائي الميزات الهامة من البيانات. مثل : التمثيلات المتعمقة Deep Representationsاستخدام الشبكات العصبية لاستخراج ميزات معقدة من البيانات.

٥. **تقنيات التحويل Transformation Methods:** تقوم هذه الأساليب بتحويل الميزات إلى شكل جديد يمكن أن يكون أكثر فائدة للنموذج.ولها عدة أنواع وهي : تحليل المكونات الرئيسية PCA ، تحليل التباين الخطي LDA .

٦. **تقنيات التجميع Clustering Methods** : تُستخدم لتجميع البيانات في مجموعات، مما يمكن أن يساعد في تحديد الميزات الأكثر أهمية بناءً على التجمعات.مثل K-Means Clustering

٧. **التقنيات القائمة على الشجرة:** تعتمد على شجرة القرار أو الغابات العشوائية لتحديد الأهمية النسبية لكل ميزة.

كل نوع من تقنيات استخلاص الميزات له مزاياه وعيوبه، واختيار التقنية المناسبة يعتمد على طبيعة البيانات والمشكلة المستهدفة. قد يكون من المفيد استخدام مجموعة من التقنيات للحصول على أفضل النتائج.

٤-٣ تقنيات تعلم الآلة:

التعلم الآلي هو فرع من فروع الذكاء الاصطناعي يهدف إلى إنشاء نظم يمكنها التعلم والتكيف من خلال البيانات بدلاً من البرمجة الصريحة. يعتمد التعلم الآلي على تقنيات وخوارزميات مختلفة لتحليل البيانات، استخراج الأنماط والتنبؤ بالسلوك المستقبلي. الذي يسمح لأنظمة الكمبيوتر بالتعلم مباشرة من الأمثلة والبيانات والخبرة، ولديه العديد من الخوارزميات[10][11].

أنواع التعلم الآلي:

١. **التعلم تحت الإشراف Supervised Learning** : يعتمد هذا النوع على وجود بيانات مدخلة مع نتائج معروفة، حيث يهدف إلى فهم العلاقة بين المدخلات والمخرجات. من الأمثلة عليه: الانحدار الخطي، شجرة القرار، والشبكات العصبية.

٢. **التعلم بدون إشراف Unsupervised Learning**: يعمل هذا النوع على بيانات غير مصنفة، حيث لا توجد نتائج معروفة، ويهدف إلى اكتشاف الأنماط أو التجمعات داخل البيانات. تشمل أمثلته: تحليل التجميعClustring وتقليل الأبعاد.

٣. **التعلم المعزز Reinforcement Learning**: يعتمد هذا النوع على مفهوم المكافآت والعقوبات، حيث يتعلم النموذج من خلال التجربة والخطأ لتحقيق أهداف محددة. يُستخدم بشكل واسع في مجالات مثل الألعاب والروبوتات.

فيما يلي مجموعة من خوارزميات التعلم الآلي المستخدمة في تصنيف وتحليل البيانات في هذا البحث:[13][12]

١- **المصنفNaïve Bayes** : خوارزمية نايف بايز هي طريقة بسيطة وفعالة لتصنيف البيانات، تعتمد على نظرية بايز. تعتبر "نايف" (بسيطة) لأنها تفترض استقلالية الميزات (المتغيرات) عن بعضها البعض. تتميز هذه الخوارزمية بكونها سريعة وسهلة التنفيذ، وتعمل بكفاءة مع مجموعات البيانات الكبيرة.ومع ذلك، من عيوبها أنها تفترض أن الميزات مستقلة عن بعضها، وهو ما قد لا يكون صحيحًا في العديد من الحالات، مما قد يؤثر سلبًا على دقة النموذج. كما أنها تواجه تحديات مع الميزات النادرة.

٢- **أشجار القرار Decision Tree** : شجرة القرار هي نموذج من نماذج التعلم الآلي تستخدم في مهام التصنيف والانحدار. تمثل البيانات بشكل شجري، حيث يتم تقسيم البيانات إلى مجموعات فرعية بناءً على ميزات معينة، مما يسهل اتخاذ القرارات. تتميز هذه الخوارزمية بسهولة الفهم كما أنها تتمتع بمرونة تجعلها مناسبة لمهام التصنيف والانحدار. ومع ذلك قد تتعلم الشجرة من التفاصيل غير المهمة من البيانات، مما يؤدي إلى أداء ضعيف عند التعامل مع بيانات جديدة.

٣- **الغابة العشوائية Random Forest** : هي تقنية تعلم آلي تستخدم لتصنيف البيانات أو التنبؤ بها، وهي تعتمد على إنشاء مجموعة من أشجار القرار (Decision Trees) وتجمع نتائجها لتحسين الدقة وتقليل الإفراط في التخصيص.

تتميز الغابات العشوائية بدقة عالية. ومع ذلك، تواجه بعض التحديات، مثل زيادة التعقيد مقارنةً بشجرة قرار واحدة.

٤- **الانحدار اللوجستي Logistic Regression:** هو نموذج إحصائي يُستخدم لتوقع نتائج ثنائية (ثنائية التصنيف)، مثل "نعم" أو "لا"يعتمد هذا النموذج على دالة لوجستية لتحويل المخرجات الخطية إلى احتمالات.

تتميز هذه الخوارزمية ببساطة الفهم كما أنها تعمل بشكل جيد عندما تكون العلاقة بين الميزات والنتيجة خطية في حين أنها تفترض أن العلاقة بين الميزات والنتيجة خطية، مما قد لا يكون صحيحًا في بعض الحالات.

٣-٥ مصفوفة الارتباك Confusion Matrix :

هي أداة تستخدم في تقييم أداء نماذج التعلم الآلي، وتُستخدم بشكل خاص في مجالات مثل تصنيف البيانات والكشف عن الهجمات. مصفوفة الارتباك تعطي فكرة واضحة عن كيفية أداء النموذج في التعرف على الحالات الإيجابية والسلبية، وتُستخدم كأداة هامة في تقييم جودة نتائج النماذج التعليمية.تتكون مصفوفة الارتباك من أربعة قسم رئيسية:

١. **True Positive (TP)** : يُمثل عدد الحالات التي تم تصنيفها بشكل صحيح كـ "إيجابية" من قبل النموذج.
٢. **True Negative (TN)**: يُمثل عدد الحالات التي تم تصنيفها بشكل صحيح كـ "سلبية" من قبل النموذج.
٣. **False Positive (FP)** : يُمثل عدد الحالات التي تم تصنيفها بشكل خاطئ كـ "إيجابية" من قبل النموذج.
٤. **False Negative (FN)** : يُمثل عدد الحالات التي تم تصنيفها بشكل خاطئ كـ "سلبية" من قبل النموذج.

باستخدام هذه الأرقام، يُمكن حساب مقاييس أداء النموذج التالية، فيما يلي شرح كل مقياس مع صيغة حساب كل منها : [14]

١- **الانتقاء (Precision)** : يُستخدم لقياس مدى دقة التنبؤات الإيجابية التي قام النموذج بعملها. يُحسب الانتقاء عن طريق قسمة عدد الحالات الإيجابية التي تم تصنيفها بشكل صحيح (True Positive)على إجمالي عدد الحالات التي تم تصنيفها كإيجابية من قبل النموذج.صيغة الانتقاء:

$$Precision = TP/TP + FP \quad (١)$$

٢- **الاسترجاع (Recall)** : يُستخدم لقياس مدى قدرة النموذج على اكتشاف جميع الحالات الإيجابية الموجودة في البيانات. يُحسب الاسترجاع عن طريق قسمة عدد الحالات الإيجابية التي تم تصنيفها بشكل صحيح (True Positive)على إجمالي عدد الحالات الإيجابية الفعلية في البيانات.صيغة الاسترجاع:

$$Recall = TP/TP + FN \quad (٢)$$

٣- **القيمة المتوسطة المربعة للخطأ (F1-score):** تجمع بين الدقة Precision والاسترجاع Recallفي مقياس واحد يعكس كفاءة النموذج في التعرف على الحالات الإيجابية وتجنب الكثير من الخطأ:

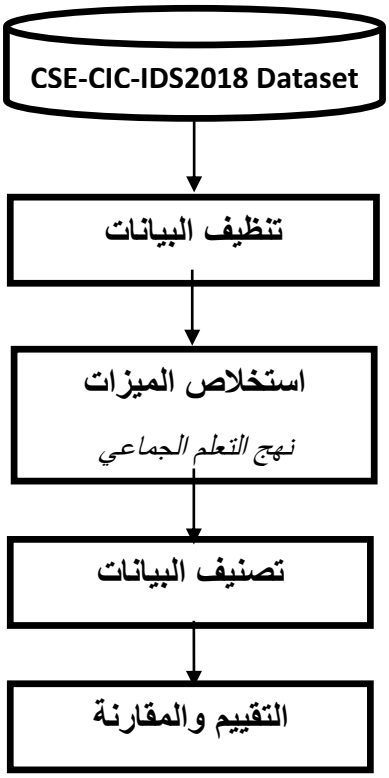
$$F1\text{-}score = 2 \times Precision \times Recall/Precision + Recall \quad (٣)$$

٤- **منحنى AUC-ROC** : اختصار ل Area Under the Curve-Receiver Operating Characteristic هو أداة تستخدم لتقييم أداء نماذج التصنيف ، إذا كان النموذج مثاليًا، فإن المنحنى سيكون قريبًا جدًا من الزاوية العلوية اليسرى، حيث يكون معدل الإيجابيات الحقيقية ١ ومعدل الإيجابيات الكاذبة ٠.يمثل المحتوى العلاقة بين معدل الإيجابيات الصحيحة ومعدل الإيجابيات الخاطئة

٤.النتائج والمناقشة

سيتمّ عرض النتائج العملية التي توصل إليها البحث . وفي سبيل التحقق من نسبة التصنيف وموثوقية المصنف المصمم تم استخدام قاعدة بيانات معيارية من أجل عملية التدريب والاختبار. وتمّ استخدام تقنيات مختلفة من أجل عملية التصنيف.

فيما يلي مخطط صندوقي يوضح مراحل العمل الرئيسية:



الشكل (١) : مخطط صندوقي يوضح مراحل العمل الرئيسية

المرحلة الأولى : تنظيف البيانات في البداية تم التأكد من عدم وجود خلايا فارغة أو مكررة في سجلات قاعدة البيانات ، بهدف حذفها لانها تؤثر على عملية التصنيف ، تم إيجاد ٥٩ سجل مكرر بكامل بياناته تمت تصفيتهم وازالتهم من قاعدة البيانات.

بالنسبة للسمات تم إيجاد ٨ حقول من السمات بقيم صفرية تم الاستغناء عنها، وهي كالتالي :

BWd_PSH_Flags,Bwd_URG_Flags,Fwd_Avg_Bytes_Bulk,Fwd_Avg_Packets_Bulk,Fwd_Avg_Bulk_Rate,Bwd_Avg_Bytes_Bulk,Bwd_Avg_Packets_Bulk,Bwd_Avg_Bulk_Rate

كما تم اسقاط حقل البروتوكول *Protocol* لان حقل DST يحوي قيم البروتوكول المكافئة لكل قيمة *Destination-Port* .

كما تم اسقاط حقل الطابع الزمني *TimeStamp* لأننا أردنا للمصنفات ألا تميز بين تنبؤات الهجوم بناء على الوقت ، أي انه يجب على المصنفات ان تكون قادرة على التمييز بين الهجمات بغض النظر عما اذا كانت كبيرة الحجم او بطيئة او خفية.

مرحلة اسقاط السمات تعتمد على الهدف من استخدام مجموعة البيانات ، بناء على ذلك يتم الإبقاء او اسقاط السمات، تم توضيح سبب اسقاط السمتين *Protocol* و *TimeStamp* من قاعدة البيانات المستخدمة في هذا البحث. بعد هذه العملية يصبح عدد السمات ٧٣ وسمة label التي تحدد فيما إذا كان السجل هجوم أم طبيعي.

تم ترميز الهجومات بالرقم واحد والسجلات الطبيعية بالرقم صفر،وبالتالي يمكن التمييز بين سجلات الهجوم والسجلات الطبيعية عند التصنيف.

المرحلة الثانية :استخلاص الميزات تم في هذا البحث الاعتماد على نهج التعلم الجماعي في عملية استخلاص الميزات، اذفي البداية تم استخدام **خمس** تقيات تصنيف لتوليد خمس قوائم من الميزات ثم معالجة القوائم الناتجة لاستخلاص الميزات منها .

خلال عملية التصنيف نختار على الأكثر أعلى ٢٠ ميزة مرتبة ،يعود السبب في اختيار الحد الأعلى هو ٢٠ ميزة هو أننا أردنا اختيار قائمة من الميزات طويلة بما يكفي بحيث يكون هناك فرصة جيدة للتصنيفات المختلفة بامتلاك عناصر مشتركة.

تقنيات تصنيف البيانات المستخدمة هي LightGBM–CATBoost–GR–IG–CS كل تقنية من هذه التقنيات تعطي قائمة بالميزات الأكثر أهمية مرتبة كما في الجدول (٣) :

الجدول (٣): قائمة بالميزات الأكثر أهمية

XGBoost	CatBoost	CS	Information Gain	Gain Ratio
1.Fwd_Packets_s	**1**.Fwd_Packets_s	**1**.Fwd_Packets_s	**1**.Fwd_Packets_s	**1**.Fwd_Packets_s
2.Flow_Packets_s	**2**.Flow_Packets_s	**2**.Flow_Packets_s	**2**.Flow_Packets_s	**2**.Flow_Packets_s
3.Flow_lAT_Mean	**3**.Flow_lAT_Mean	**3**.Flow_lAT_Mean	**3**.Flow_lAT_Mean	**3**.Flow_lAT_Mean
4.Destination_Port	**4**.Destination_Port	**4**.Destination_Port	**4**.Destination_Port	**4**.Destination_Port
5.Bwd_Packets_s	**5**.Bwd_Packets_s	**5**.Bwd_Packets_s	**5**.Bwd_Packets_s	**5**.Bwd_Packets_s
6.Flow_Bytes_s	**6.**Flow_Bytes_s	**6.**Flow_Bytes_s	**6.**Flow_Bytes_s	**6.**Flow_Bytes_s
7.Fwd_lAT_Mean	**7.**Fwd_lAT_Mean	**7.**Fwd_lAT_Mean	**7.**Fwd_lAT_Mean	**7.**Fwd_lAT_Mean
8.Flow_lAT_Max	**8.**Flow_lAT_Max	**8.**Flow_lAT_Max	**8.**Flow_lAT_Max	**8.**Flow_lAT_Max
9.Fwd_Packet_Length_Mean	**9.**Fwd_Packet_Length_Mean	**9.**Fwd_Packet_Length_Mean	**9.**Fwd_Packet_Length_Mean	**9.**Fwd_Packet_Length_Mean
10.Avg_Fwd_Segment_Size	**10**.Avg_Fwd_Segment_Size	**10**.Avg_Fwd_Segment_Size	**10**.Avg_Fwd_Segment_Size	**10**.Avg_Fwd_Segment_Size
11.Packet_Length_Std	**11**.Packet_Length_Std	**11**.Packet_Length_Std	**11**.Packet_Length_Std	**11**.Packet_Length_Std
12. Fwd_Header_Length	**12.**Fwd_Header_Length	**12.**Fwd_Header_Length	**12.**Fwd_Header_Length	**12.**Fwd_Header_Length
13. Packet_Length_Variance	**13.**Packet_Length_Variance	**13.**Packet_Length_Variance	**13.**Packet_Length_Variance	**13.**Packet_Length_Variance
14.Total_Length_of_Fwd_Packets	**14**.Total_Length_of_Fwd_Packets	**14**.Total_Length_of_Fwd_Packets	**14**.Total_Length_of_Fwd_Packets	**14**.Total_Length_of_Fwd_Packets
15. Subflow_Fwd_Bytes	**15.**Subflow_Fwd_Bytes	**15.**Subflow_Fwd_Bytes	**15.**Subflow_Fwd_Bytes	**15.**Subflow_Fwd_Bytes
16. Packet_Length_Mean	**15.**Subflow_Fwd_Bytes	**16.**Packet_Length_Mean	**16.**Packet_Length_Mean	**16.**Packet_Length_Mean
17.Fwd_Packet_Le	**16.**Packet_Length	**17**.Fwd_Packet_Leng	**17**.Fwd_Packet_Leng	**17**.Fwd_Packet_Leng

ngth_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets **21.**Flow_lAT_Min **22.**Idel_Max **23.**Idel_Mean **24.**Idel_Min **25.**Active_Mean **26.**Active_Max **27.**Active_Min	_Mean **17**.Fwd_Packet_Length_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets **21.**Flow_lAT_Min **22.**Idel_Max **23.**Idel_Mean **24.**Idel_Min **25.**Active_Mean	th_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets **21.**Flow_lAT_Min **22.**Idel_Max **23.**Idel_Mean **24.**Idel_Min **25.**Active_Mean	th_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets **21.**Flow_lAT_Min **22.**Idel_Max **23.**Idel_Mean **24.**Idel_Min	th_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets

نشكل من القوائم الخمسة السابقة ثلاث قوائم تكون كما يلي :

القائمة١ : تحوي السمات المشتركة بين ٣ من التصنيفات (XGBoost,CATBoost,CS) كانت النتيجة وجود ٢٥ سمة مشتركة .

القائمة٢ : تحوي السمات المشتركة بين ٤ من التصنيفات (XGBoost,CATBoost,CS,Informaion Gain)وكانت النتيجة وجود ٢٤ سمة مشتركة.

القائمة ٣: تحوي السمات المشتركة بين ٥ من التصنيفات (XGBoost,CATBoost,CS,Information Gain,Gain Ratio) كانت النتيجة وجود ٢٠ سمة مشتركة .

ونضيف قائمة رابعة تحوي جميع سمات مجموعة البيانات المستخدمة بهدف المقارنة ونطلق عليها اسم **القائمة ٤** . في مايلي جدول (٤) يوضح السمات في كل من القوائم الثلاثة السابقة:

الجدول (٤) : السمات في كل من القوائم الثلاثة

القائمة ١	القائمة ٢	القائمة ٣
1.Fwd_Packets_s **2**.Flow_Packets_s **3**.Flow_lAT_Mean **4**.Destination_Port **5**.Bwd_Packets_s **6.**Flow_Bytes_s **7.**Fwd_lAT_Mean **8.**Flow_lAT_Max **9.**Fwd_Packet_Length_Mean **10**.Avg_Fwd_Segment_Size **11**.Packet_Length_Std **12.**Fwd_Header_Length **13.**Packet_Length_Variance **14**.Total_Length_of_Fwd_Packets **15.**Subflow_Fwd_Bytes **16.**Packet_Length_Mean **17**.Fwd_Packet_Length_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets **21.**Flow_lAT_Min **22.**Idel_Max	**1**.Fwd_Packets_s **2**.Flow_Packets_s **3**.Flow_lAT_Mean **4**.Destination_Port **5**.Bwd_Packets_s **6.**Flow_Bytes_s **7.**Fwd_lAT_Mean **8.**Flow_lAT_Max **9.**Fwd_Packet_Length_Mean **10**.Avg_Fwd_Segment_Size **11**.Packet_Length_Std **12.**Fwd_Header_Length **13.**Packet_Length_Variance **14**.Total_Length_of_Fwd_Packets **15.**Subflow_Fwd_Bytes **16.**Packet_Length_Mean **17**.Fwd_Packet_Length_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets **21.**Flow_lAT_Min **22.**Idel_Max	**1**.Fwd_Packets_s **2**.Flow_Packets_s **3**.Flow_lAT_Mean **4**.Destination_Port **5**.Bwd_Packets_s **6.**Flow_Bytes_s **7.**Fwd_lAT_Mean **8.**Flow_lAT_Max **9.**Fwd_Packet_Length_Mean **10**.Avg_Fwd_Segment_Size **11**.Packet_Length_Std **12.**Fwd_Header_Length **13.**Packet_Length_Variance **14**.Total_Length_of_Fwd_Packets **15.**Subflow_Fwd_Bytes **16.**Packet_Length_Mean **17**.Fwd_Packet_Length_Max **18.**Bwd_Packet_Length_Mean **19**.Avg_Bwd_Segment_Size **20.**Total_Length_of_Bwd_Packets

23. Idel_Mean **24.**Idel_Min **25.**Active_Mean	**23.**Idel_Mean **24.**Idel_Min	

نحصل في نهاية هذه المرحلة على ثلاث قوائم من السمات الأكثر أهمية في قاعدة البيانات بالاعتماد على نهج التعلم الجماعي موضحة في الجدول (٤) وبالتالي تصبح قاعدة البيانات جاهزة لعملية التصنيف والمقارنة .

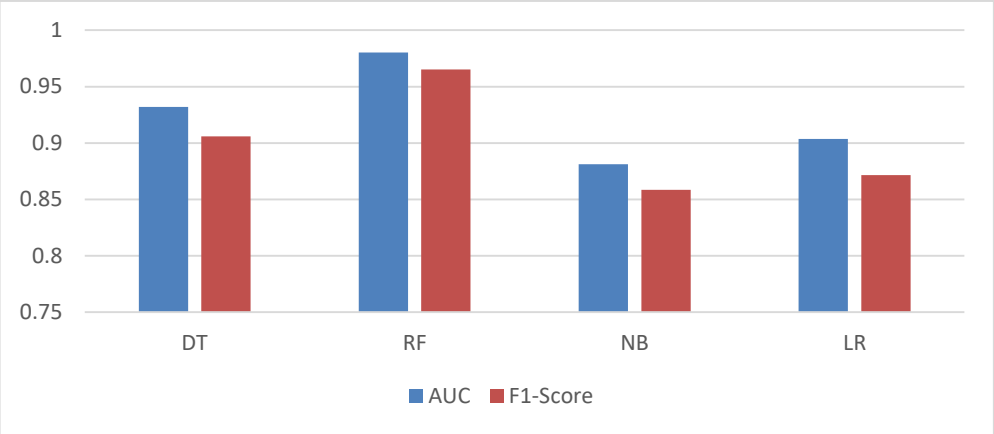
المرحلة الثالثة : التصنيف والمقارنة بعد عمليتي تنظيف البيانات واستخلاص الميزات سوف نخضع كل قائمة من القوائم الأربع التي حصلنا عليها الى خوارزمية تصنيف وبعد ذلك ستتم مقارنة النتائج بهدف دراسة تأثير السمات على دقة التصنيف من أجل جميع قوائم السمات._فيما يلي جدول يوضح قيمة AUC و F1 من أجل جميع القوائم :

الجدول (٥) : قيمة AUC و F1 من أجل جميع القوائم

القائمة/الخوارزمية	القائمة الأولى		القائمة الثانية		القائمة الثالثة		القائمة الرابعة	
	AUC	F1-Score	AUC	F1-Score	AUC	F1-Score	AUC	F1-Score
DT	0.9208	0.8952	0.9255	0.9041	0.9321	0.9058	0.9013	0.8923
RF	0.9613	0.9256	0.9729	0.9387	**0.9802**	**0.9651**	0.9654	0.9202
NB	0.8398	0.8039	0.8589	0.8133	0.8812	0.8584	0.8244	0.8001
LR	0.8754	0.8382	0.8874	0.8564	0.9036	0.8714	0.8728	0.8225

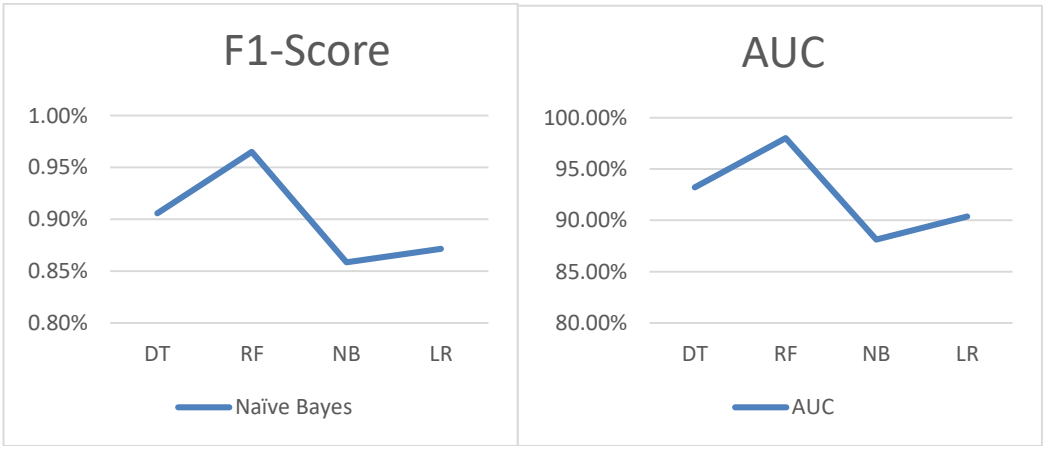
من الجدول (٥) نجد أن قيمة كل من AUC و F1 كانت الأعلى من أجل القائمة الثالثة ومن أجل جميع خوارزميات التصنيف، كما أننا حصلنا على قيم جيدة من أجل القائمة الرابعة التي تتضمن جميع السمات في حين لم تكن الأعلى بالمقارنة مع قيم القوائم الثلاثة التي تتضمن السمات الأكثر ارتباطا . بمقارنة القيم من أجل قوائم السمات الثلاثة نجد أن القيم جميعها جيدة ومقبولة لكن قائمة السمات الثالثة أعطت أعلى القيم وبالتالي يمكن اعتمادها من أجل عملية التصنيف والاستغناء عن السمات الأخرى.

فيما يلي مخطط البياني يوضح قيم AUC و F1 من أجل جميع خوارزميات التصنيف ومن أجل قائمة السمات الثالثة التي أعطت أعلا دقة :



الشكل (٢): مخطط البياني يوضح قيم AUC و F1 من أجل جميع خوارزميات التصنيف

نلاحظ من الشكل البياني (٢) أن قيم AUC و F1-Score في خوارزمية RF هي الأعلى ،وأنه يمكن اعتماد قائمة السمات الثالثة بأنها القائمة النهائية لقاعدة البيانات بعد عملية استخلاص الميزات كونها أعطت أعلى دقة مقارنة بالقيم الناتجة عن استخدام القوائم الأخرى. فيما يلي المخطط البياني (٣) يوضح النسب المئوية لقيم AUC و F1 من أجل جميع خوارزميات التصنيف المستخدمة



الشكل (٣):النسبة المئوية لقيم AUC و F1 من أجل جميع خوارزميات التصنيف

يبين الشكل (٣) أن القيم من أجل خوارزمية RF أعطت أعلى قيمة ،في حين أن أقل قيم كانت باستخدام مصنف Naïve Bayes بالمقارنة مع الخوارزميات المستخدمة الاخرى .وبالمقارنة مع الدراسات السابقة التي استخدمت نفس مجموعة البيانات ونفس معايير تقييم الأداء نجد أن الآلية المتبعة في تصنيف السجلات في هذا البحث أعطت أفضل القيم كما هو موضح في الجدول (٧) :

الجدول (٦):مقارنة مع دراسات سابقة

الدراسة	تقنية استخلاص الميزات	خوارزميات التصنيف المستخدمة	قيمة AUC
الدراسة [1]	معامل الارتباط ، chi-sqarted	DT,NB,RF,LR..	من أجلDT AUC=94
الدراسة [3]	XGBoost	SVM,MLP,RF,NB,LightGMB..	من أجلLightGBM AUC=95.33

الدراسة [5]	RF	DT,KNN,SVM,NB,RF	من أجلRF AUC=93.01
الدراسة الحالية	نهج التعلم الجماعي	DT,RF,NB,LR	من أجلRF AUC=98.02

من الجدول (٦) نلاحظ أن الدراسة الحالية أعطت أفضل قيمة لمقياس AUC بالمقارنة مع الدراسات السابقة التي استخدمت نفس مجموعة البيانات ونفس مقاييس الأداء وكان ذلك من أجل خوارزمية RF بقيمة AUC=98.02 . تتميز الدراسة الحالية أنها استخدمت نهج التعلم الجماعي بدلاً من استخدام تقنية استخلاص ميزات واحدة مما انعكس ايجاباً على دقة الكشف وقيمة Auc و F1 بشكل خاص كما انها قارنت بين أربع خوارزميات تصنيف.ومن أجل المصنفات الأربعة حصلنا على أفضل دقة كشف بالمقارنة مع الدراسات السابقة والتي استخدمت نفس المصنفات .في هذه الدراسة تم استخدام نهج التعلم الجماعي في مرحلة استخلاص الميزات وهذا ما لم تتطرق له الدراسات السابقة، إن تطبيق هذا النهج في مرحلة استخلاص الميزات يساعد في تقليل مجموعة السمات الأولية مع الاحتفاظ بالمعلومات المهمة مما يساعد في تحسين كفاءة ودقة النماذج من خلال التركيز على المعلومات الأكثر صلة فقط.

٤-الاستنتاجات والتوصيات:

تم في هذا البحث كشف هجمات الأمن السيبراني على مجموعة البيانات باستخدام نهج التعلم الجماعي وتم التصنيف باستخدام أربع خوارزميات تصنيف، من خلال هذا البحث تم التوصل إلى ما يلي:

- ✓ استخدام النهج الجماعي في عملية استخلاص الميزات كان له أثر إيجابي على دقة الكشف كونه أعطى دقة عالية بالمقارنة مع الخوارزميات الأخرى.
- ✓ توصل البحث إلى قيمة Auc كانت كما يلي 0.9036, 0.8812, 0.9802, 0.9321 من أجل خوارزميات DT,RF,NB,LR على التوالي.
- ✓ **تكمن أهمية البحث : في استخدام نهج التعلم الجماعي في عملية استخلاص الميزات بالمقارنة مع الدراسات السابقة ، إذ ينتج عن تجميع التنبؤات من نماذج متعددة وجمع عدة نماذج نتائج أفضل دقة من استخدام النماذج الفردية، كما يساعد في التعميم بشكل أفضل على البيانات الجديدة مما يجعلها مناسبة للتطبيقات في العالم الحقيقي.**
- ✓ **يمكّننا الجمع بين نتائج عدة خوارزميات وتقنيات استخلاص الميزات من الحصول على أكثر السمات أهمية و ارتباطاً بالهدف** على عكس استخدام تقنية استخلاص ميزات واحدة مما يزيد من دقة وأداء عملية التصنيف لأننا نتعامل مع السمات الأكثر أهمية في مجموعة البيانات والناتجة عن عدة تقنيات استخلاص ميزات وليس فقط تقنية واحدة.

يقترح البحث التعديلات والتطويرات التالية والتي قد تضيف أفكاراً جديدة في مجال الأمن السيبراني:

- ✓ استخدام تقنيات استخلاص ميزات أخرى وتطبيقها في مجال النهج الجماعي على نفس قاعدة البيانات أو على قواعد بيانات أخرى مختلفة.

- ✓ تطوير بنية المصنفات المعتمدة في مرحلة اتخاذ القرار واختبار النظام على مصنفات أخرى واستخدام مصنف هجين ناتج عن دمج وتكامل عدة مصنفات مفردة.
- ✓ تطوير البحث السابق بحيث يأخذ بالحسبان عامل الوقت.

المراجع

[1] Fitni QRS, Ramli K.(2020). Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems. In: 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT). IEEE; 2020. p. 118–24.

[2] Li X, Chen W, Zhang Q, Wu L.(2020). Building auto-encoder intrusion detection system based on random forest feature selection. Comput Secur. 2020;95:101851.

[3] Hua Y. (2020) An efficient traffic classification scheme using embedded feature selection and lightgbm. In: 2020 Information Communication Technologies Conference (ICTC). IEEE; 2020. p. 125–30.

[4] D'hooge L, Wauters T, Volckaert B, De Turck FF.(2020) Inter-dataset generalization strength of supervised machine learning methods for intrusion detection. J Inf Secur Appl. 2020;54:102564.

[5] Huancayo Ramos KS, Sotelo Monge MA, Maestre Vidal J. Benchmark-based reference model for evaluating botnet

detection tools driven by traffic-flow analytics. Sensors. 2020;20(16):4501.

[6] Sumaiya Thaseen, I., Poorva, B., & Ushasree, P. S. (2020). Network Intrusion Detection using Machine Learning Techniques. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). doi:10.1109/ic-etite47903.2020.148

[7] Waskle, S., Parashar, L., & Singh, U. (2020). Intrusion Detection System Using PCA with Random Forest Approach. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). doi:10.1109/icesc48915.2020.9155.

[8] Saini, Parvinder Singh; Behal, Sunny; Bhatia, Sajal," DDoS attacks detection using machine learning and deep learning techniques: analysis and comparison" . [IEEE 2020 7th International Conference on Computing for Sustainable Global

Development (INDIACom) - New Delhi, India (2020.3.12-2020.3.14)] 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom) Detection of DDoS Attacks using Machine Learning Algorithms. 16-21.doi:10.23919/INDIACom49435.2020.9083716,2023.

[9] Taher, K. A., Mohammed Yasin Jisan, B., & Rahman, M. M. (2019). Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection. 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST). doi:10.1109/icrest.2019.8644161

[10] Bilal.M; Ekhlas.K. G. 2021, "Intrusion Detection System for NSL-KDD dataset based on deep learning and recursive feature elimination," Engineering and Technology Journal, Vol. 39, No. 07, pp. 1069-1079.

[11] U. Islam et al.,2022. "Detection of distributed denial of service (DDoS) attacks in IoT based monitoring system of banking sector using machine learning models," Sustainability, vol. 14, no. 14, p. 8374, doi: 10.3390/su14148374.

[12] Z. Liu, L. Qian, and S. Tang,2022. "The prediction of DDoS attack by machine learning," in Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021), pp. 681–686, doi: 10.1117/12.2628658.

[13] Rawat.Sh;Srinivasan.A;Ravi.V;Ghosh.U.2020. Intrusion detection systems using classical machinelearning techniques vs integrated unsupervised feature learning and deep neural network.wileyonlinelibrary.com/journal/itl2.

[14] Abrar.I;Ayub.Z;Masoodi.F,2020. A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. Conference on Smart Electronics and Communication (ICOSEC 2020) IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9.