SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.03 NO. 2 2024
DOI: 10.51984/SUCP.V3I2.3327

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

دراسة مقارنة حول مجموعات البيانات القياسية Standard Datasets لكشف التسلل في شبكات إنترنت الأشياء IoTs

*فاطمة عبد النبي حسن [1] و غزلان محمد الزروق [2]

[1] قسم الهندسة الكهربائية والإلكترونية، كلية الهندسة ، جامعة وادي الشاطئ ، ليبيا
[2] قسم الهندسة الطبية، كلية الهندسة، جامعة وادي الشاطئ، وادي الشاطئ، ليبيا

الملخص

في السنوات الأخيرة، ساهم الانتشار الواسع لتطبيقات انترنت الأشياء Internet of Things (IoT) في تطوير المدن الذكية، ولكن مع نمو شبكات المدن الذكية، يزداد خطر تهديدات وهجمات الأمن السيبراني. وعلى الرغم من انتشار العديد من أليات الأمان من تقنيات التشفير وجدران الحماية، إلّا أنه من المستحيل تجنب الهجمات المختلفة على شبكات إنترنت الأشياء. ولمعالجة هذه المشكلة، تم استخدام التعلم الآلي كأداة فعالة للكشف عن الهجمات. وهذا يتم من خلال تطبيق عدد من خوارزميات التصنيف الخاضعة للإشراف على مجموعة البيانات Dataset. تستعرض هذه الدراسة بعض مجموعات البيانات الشائعة لكشف التسلل إلى الشبكات بشكل عام وشبكات إنترنت الأشياء بشكل خاص، وأهمها: KDD Cup '99، Kyoto2006+، NSL-KDD، UNSW-NB15، CIC-IDS 2017، CSE-CIC-IDS 2018. بالإضافة إلى المقارنة بينها بناءً على عدد الميزات في كل منها، وجود هجمات حديثة، عدد السجلات الكلي وعدد فئات الهجوم. وفي اخر هذه الورقة تم استعراض أهم الدراسات السابقة التي تناولت تطبيق بعض خوارزميات التعلم الآلي على مجموعة بيانات قيد الدراسة وتلخيص مؤشرات الأداء والتي من ضمنها الدقة وزمن التدريب للخوارزميات.

الكلمات المفتاحية:

التعلم الآلي
التعلم الإشرافي
خوارزميات التعلم
مجموعة البيانات القياسية
مؤشرات الأداء

A Comparative Study on Standard Datasets for Intrusion Detection in Internet of Things Networks

* Fatimah A. Hasan [a], Guzlan M. Miskeen [b]

[a] Department of Electrical and Electronic Engineering, Faculty of Engineering, Wadi Al-shati University, wadi Al-shati, Libya
[b] Department of Medical Engineering, Faculty of Engineering, Wadi Al-shati University, wadi Al-shati, Libya

Keywords:
Learning algorithms
Machine learning
Performance indicators
Standard datasets
Supervised learning

ABSTRACT

In recent years, the widespread use of Internet of Things (IoT) applications has contributed to the development of smart cities, but with the growth of smart city networks, the risk of cybersecurity threats and attacks increases. Despite the spread of many security mechanisms such as encryption technologies and firewalls, it is impossible to avoid various attacks on IoT networks. To address this problem, machine learning has been used as an effective tool to detect attacks. This is done by applying a number of supervised classification algorithms to a dataset. This study reviews some of the common datasets for detecting intrusions into networks in general and IoT networks in particular, the most important of which are: KDD Cup '99, Kyoto2006+, NSL-KDD, UNSW-NB15, CIC-IDS 2017, CSE-CIC-IDS 2018. In addition to comparing them based on the number of features in each, the presence of recent attacks, the total number of records and the number of attack categories. At the end of this paper, the most important previous studies that dealt with the application of some machine learning algorithms on the data set under study were reviewed and the performance indicators were summarized, including accuracy and training time of the algorithms.

i. المقدمة

في السنوات الأخيرة، ساهم الانتشار الواسع لتطبيقات إنترنت الأشياء Internet of Things (IoT) إلى استخدامها بشكل كبير في المدن الذكية حول العالم. يتكون إطار المدينة الذكية من الطبقات الثلاثة التالية: الطبقة الطرفية Terminal Layer، وطبقات الضباب Fog Layer، وطبقة

*Corresponding author:
E-mail addresses: f.abdalnaby@wau.edu.ly ,(G. M. Miskeen) guz.miskeen@wau.edu.ly
Article History : Received 11 June 2024 - Received in revised form 28 September 2024 - Accepted 06 October 2024

السحابة Cloud Layer [1]. تحتوي الطبقة السحابية على موارد تخزين (مثل الخوادم والأجهزة الافتراضية) لتخزين كمية كبيرة من البيانات، وتعمل طبقة الضباب كجسر بين أجهزة الطبقة الطرفية وطبقة السحابة وتعتبر طبقة الضباب أكثر فعالية في تحديد الهجمات السيبرانية Cyber Attacks المختلفة من الطبقة السحابية المركزية، وتتكون الطبقة الطرفية من مجموعة من أجهزة الاستشعار المثبتة داخل المدينة لجمع البيانات [1]. ومن الأسباب التي تجعل شبكات إنترنت الأشياء IoT Networks عرضه للهجمات [1]:

1. تتمتع معظم أجهزة إنترنت الأشياء بموارد محدودة (على سبيل المثال قوة معالجة وذاكرة صغيرة).
2. ترتبط أجهزة إنترنت الأشياء ببروتوكولات مختلفة ويتسبب العدد المتزايد من أجهزة إنترنت الأشياء في حدوث تأخير في المراكز السحابية.
3. في بعض الأحيان تكون أجهزة إنترنت الأشياء غير مراقبة مما يجعل من الممكن وصول المتسللين إليها.
4. الجزء الأكبر من اتصالات البيانات يكون لا سلكيًا مما يعرضها للتصنت.

وعلى الرغم من انتشار العديد من أليات الأمان التقليدية من تقنيات التشفير وجدران الحماية إلّا أنه من المستحيل تجنب الهجمات المختلفة على شبكات إنترنت الأشياء، نتيجة لذلك يعد نظام كشف التسلل Intrusion Detection System (IDS) أحد الأليات الفعالة لحماية شبكات إنترنت الأشياء IoT networks [2]. وللحفاظ على مبادئ الأمان الثلاثة لأنظمة المعلومات وهي السرية Confidentiality والنزاهة Integrity والتوافر Availability وتعرف هذه الصفات الثلاثة باسم "CIA Triad" [3].

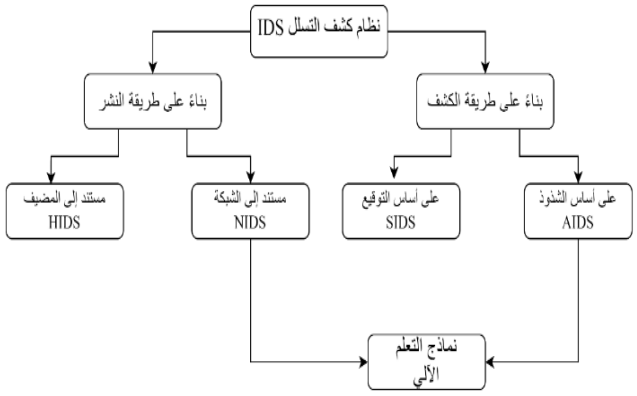
أ- نظام كشف التسلل Intrusion Detection System

نظام كشف التسلل Intrusion Detection System (IDS) هو ألية أمنية تراقب باستمرار حركة مرور الشبكة للكشف عن أي سلوك مشبوه ينتهك سياسة الأمان [2]. وأول من استخدم مصطلح نظام كشف التسلل هو James Anderson في أواخر السبعينات واوائل الثمانينات [4]. تسعى عملية تحليل حركة مرور الشبكة لتحديد النشاط الضار بعملية كشف التسلل ويسمى النظام الذي يقوم بأتمتة هذه العملية بنظام كشف التسلل IDS، ويشير كشف التسلل الى فعل الكشف عن الإجراءات التي تحاول المساس بسرية المصدر ونزاهته وتوفره [5]. حيث تشير السرية Confidentiality إلى منع وقوع المعلومات في ايدي غير مصرح بها، أما النزاهة Integrity هي ضمان منع التعديلات البيانات إلّا من خلال ألية معتمدة، تتضمن النزاهة حماية البيانات من الأنواع التالية من التعديلات غير المصرح بها مثل [6]:

1. قيام المستخدمين غير المصرح لهم بتغيير البيانات مثل (اختراق المتسلل لقاعدة البيانات وتغيير السجلات).
2. قيام المستخدمين المصرح لهم بإجراء تغيرات غير مصرح بها على البيانات.
3. تغيير البيانات من خلال ألية غير مناسبة.

أما التوافر Availability فهو قدرة المستخدمين المصرح لهم بالوصول الى البيانات لأغراض مشروعة [6].

هناك أنواع عدة من أنظمة كشف التسلل والتي يمكن تقسيمها بناءً على (طريقة وضعها في النظام أو ما يعرف بطريقة النشر وعلى أساس طرق الكشف) وتم توضيحها في الشكل 1.



شكل 1: أنواع نظام كشف التسلل IDS [2].

من منظور كشف التسلل القائم على طرق النشر Deployment-based IDS يتم تصنيف نظام كشف التسلل بشكل فرعي على أنه قائم على المضيف Host-based Intrusion Detection System (HIDS) أو نظام كشف تسلل قائم على الشبكة Network Intrusion Detection System (NIDS). ويتم نشر HIDS على مضيف المعلومات الفردي وتتمثل مهمته في مراقبة جميع الأنشطة من هذا المضيف الفردي أو المسح بحثًا عن انتهاكات سياسة الأمان والأنشطة المشبوهة، العيب الرئيسي هو نشره على جميع الأجهزة المضيفة التي تتطلب حماية من التطفل مما يؤدي إلى زيادة عبء المعالجة لكل عقدة ويؤدي في النهاية إلى تدهور أداء نظام كشف التسلل [2]. وفي المقابل يتم نشر نظام كشف التسلل المستند إلى الشبكة NIDS على الشبكة بهدف حماية الأجهزة والشبكة بأكملها من عمليات التطفل ويقوم نظام كشف التسلل المستند إلى الشبكة بمراقبة حركة مرور الشبكة باستمرار والمسح بحثًا عن الانتهاكات الأمنية المحتملة [2].

ومن منظور طريقة كشف التسلل يتم تقسيم IDS إلى كشف التسلل القائم على التوقيع Signature-based Intrusion Detection System (SIDS) ونظام كشف التسلل القائم على الشذوذ Anomaly-based Intrusion Detection System (AIDS). يعتمد نظام كشف التسلل القائم على التوقيع SIDS والمعروف ايضًا بكشف التسلل سيئ الاستخدام على فكرة تحديد التوقيع لأنماط الهجوم ويتم تخزين هذه التوقيعات في قاعدة بيانات التوقيع ويتم مطابقة أنماط البيانات مع هذه التوقيعات المخزنة للكشف عن الهجوم. ومن مميزاته كفاءة الكشف العالية عن الهجمات المعروفة بسبب توفر التوقيع لتلك الهجمات ومن ناحية أخرى تفتقر هذه الطريقة إلى القدرة على اكتشاف الهجمات الجديدة بسبب غياب أنماط التوقيع وايضًا يتم الاحتفاظ بقاعدة بيانات ضخمة للتوقيعات ومقارنتها بحزم البيانات بحثًا عن عملية الاقتحام المحتملة مما يجعلها نهجًا مستهلكًا للموارد [2].

ويعتمد نظام كشف التسلل القائم على الشذوذ AIDS والمسمى ايضًا بمعرفات الهوية المبنية على السلوك على فكرة تحديد ملف تعريف واضح للبيانات الطبيعية وأي انحراف عن هذا الملف الشخصي الطبيعي سيتم اعتباره سلوك شاذ أو غير طبيعي، ومن مميزات AIDS قدرته على اكتشاف الهجمات الجديدة وغير المعروفة ومع ذلك فإن عيبه الرئيسي هو ارتفاع معدل التنبيهات الخاطئة False Alarm Rate (FAR) حيث أنه من الصعب العثور على الحدود بين السلوك الطبيعي وغير الطبيعي لاكتشاف التسلل [2]. وفي الآونة تم نشر أنظمة كشف التسلل المستندة إلى التعلم الآلي كحلول محتملة لاكتشاف عمليات التسلل عبر الشبكة بطريقة فعالة [2] ، فمن حيث طريقة الكشف يتم استخدام نظام كشف التسلل القائم على الشذوذ

Anomaly-based IDS (AIDS) ونظام كشف التسلل إلى الشبكة Network-based IDS (NIDS) من حيث طريقة وضع أنظمة كشف التسلل في النظام.

حيث أن نظام كشف التسلل المستند إلى الشبكة هو طريقة تستخدم لتصنيف حركة مرور الشبكة على أنها ضارة أو عادية ويُظهر نظام كشف التسلل القائم على الشذوذ اكتشاف الهجمات الجديدة وغير المعروفة ونظرًا لأن هذه مشكلة تصنيف سيتم استخدام نماذج مختلفة للتعلم الآلي على نطاق واسع في أنظمة كشف التسلل [7].

ويتم تصنيف حركة مرور الشبكة بإستخدام الخوارزميات المتقدمة، وفي هذه المرحلة يأتي دور التعلم الآلي حيث تشكل خوارزميات التعلم الآلي أساس الأنظمة الذكية المستخدمة في مجال أمن الشبكات ويمكن التعبير عنها بشكل عام من خلال تحليل المشكلة التي تواجهها البرمجيات المبرمجة في شبكات الحاسوب بناءً على مجموعة بيانات محددة او تجارب سابقة [8].

تحتوي مجموعة البيانات المعيارية Benchmarked Datasets على أعمدة وتعرف بالسمات Features وصفوف تعرف بالسجلات Records وقد لا تكون بعض القيم موجودة في مجموعة البيانات أو قد تكون موجودة لكن مكررة [9]. لذلك يتم استخدام هندسة الميزات Feature Engineering على مجموعة البيانات لتحسين أداء نماذج التعلم الآلي المختلفة من خلال حذف أو تقدير القيم المفقودة واختيار أو استخراج السمات أو الميزات الأكثر أهمية من البيانات الأولية [10] (من خلال تطبيق اختبارات إحصائية مثل الارتباط وغيرها). كذلك يعتمد نظام كشف التسلل في التعلم الآلي بشكل كبير على هندسة الميزات لمعرفة المعلومات المفيدة من حركة مرور الشبكة [2].

ب- خوارزميات التعلم الآلي

التعلم الآلي هو مجموعة فرعية من الذكاء الاصطناعي ويتضمن جميع الأساليب والخوارزميات التي تمكن الآلات من التعلم تلقائيًا باستخدام النماذج الرياضية من أجل استخراج معلومات مفيدة من مجموعات البيانات الكبيرة [2]. خوارزميات التعلم الآلي الأكثر شيوعًا والمستخدمة في نظام كشف التسلل IDS هي شجرة القرار Decision Tree (DT)، والغابة العشوائية Random Forest (RF)، والجار الأقرب K-Nearest Neighbor (KNN)، وآلة ناقل الدعم Support Vector Machine (SVM)، والانحدار اللوجستي Logistic Regression (LR) [2].

1. **شجرة القرار Decision Tree (DT)**: هي إحدى خوارزميات التعلم الآلي الأساسية الخاضعة للإشراف والتي تُستخدم لتصنيف وانحدار مجموعة البيانات المحددة من خلال تطبيق سلسلة القرارات (أو القواعد Rules) يحتوي النموذج على بنية شجرة تقليدية مع العقد والفروع والأوراق. تمثل كل عقدة سمة أو ميزة، ويمثل الفرع قرارًا أو قاعدة بينما تمثل كل ورقة نتيجة محتملة أو تسمية فئة. تقوم خوارزمية DT تلقائيًا بتحديد أفضل الميزات لبناء شجرة وبعد ذلك إجراء عملية التقليم لإزالة الفروع غير ذات الصلة من الشجرة لتجنب الإفراط في التركيب. النماذج الأكثر شيوعاً في شجرة القرار هي CART وC4.5 وID3 والعديد من خوارزميات التعلم المتقدمة مثل Random Forest (RF) وXGBoost هي مصنوعة من أشجار القرار المتعددة [2].
2. **الغابة العشوائية Random Forest (RF):** تستخدم الغابة العشوائية على نطاق واسع لكل من التصنيف والانحدار وهي خوارزمية قائمة على شجرة القرار [11]. وعادةً ما تتمتع بدقة أفضل مقارنة بشجرة القرار، تتكون خوارزمية الغابة العشوائية من العديد من أشجار القرار التي تحتوي على تباين عالي وانحياز منخفض وتؤدي إلى مخرجات غير مرغوب فيها [12]. ويتم في الغابة العشوائية استخدام عينات عشوائية لإنشاء أشجار القرار ومن تم يتم التنبؤ من كل شجرة والعثور على الحل الأفضل من خلال طريقة التصويت [11].
3. **الجار الأقرب K-Nearest Neighbor (KNN)**: هي واحدة من أبسط خوارزميات التعلم الآلي الخاضعة للإشراف والتي تستخدم فكرة "تشابه الميزات" للتنبؤ بالفصل لعينة بيانات معينة. ويحدد العينة بناءً على جيرانها عن طريق حساب بعدها عن الجيران. في خوارزمية KNN، تؤثر المعلمة k على أداء النموذج فإذا كانت قيمة k صغيرة جدًا، فقد يكون النموذج عرضة للتركيب الزائد، في حين أن التحديد الكبير جدًا لقيمة k قد يؤدي إلى سوء تصنيف العينة [2].
4. **آلة ناقل الدعم Support Vector Machine (SVM)**: هي عبارة عن خوارزمية تعلم آلي خاضعة للإشراف وتستخدم لحل المشاكل الخطية وغير الخطية، بالنسبة للمشاكل غير الخطية يتم استخدام وظائف النواة Kernel Functions للحصول على أقصى مستوى هامشي مثالي والذي يعمل كحدود قرار باستخدام متجهات الدعم [2].
5. **الانحدار اللوجستي Logistic Regression (LR)**: يتم استخدامه لحل مشاكل التصنيف الثنائية ومتعددة الفئات ويتم التنبؤ باحتمالية وقوع حدث ما من خلال إعطاء البيانات المناسبة للدالة اللوجستية. وتقع مخرجات هذه الدالة بين 0 و1. والقيمة المتوسطة لها 0.5 وتعتبر العتبة بين الفئة 1 والفئة 0، يستخدم الانحدار اللوجستي المتغيرات الثنائية حيث يعتبر الخرج الأكبر من 0.5 فئة 1 وإذا كان الخرج أقل من 0.5 فإنه يعتبر فئة 0 [11].

ii. مقاييس الأداء

يتم تحديد أي من خوارزميات التصنيف أفضل من خلال مقاييس الأداء وأبرزها الدقة Accuracy والاستدعاء Recall وF-Measure ولحساب هذه المقاييس نحتاج إلى مصفوفة التشتت Confusion Matrix التي تظهر العلاقة بين السجلات المصنفة بشكل جيد والسجلات المصنفة بشكل خاطئ [4]. الجدول 1 يوضح استخدام مصفوفة التشتت Confusion Matrix العامة في التقييم، يمكن شرح المصطلحات الموجودة في مصفوفة التشتت على النحو التالي [4]:

إيجابي صحيح (TP): عدد السجلات التي تم اكتشافها بشكل صحيح كفئة عادية.

إيجابي خاطئ (FP): عدد السجلات التي لم يتم اكتشافها بشكل صحيح كفئة عادية.

سلبي خاطئ (FN): عدد السجلات التي لم يتم اكتشافها بشكل صحيح كفئة هجوم.

سلبي صحيح (TN): عدد السجلات التي تم اكتشافها بشكل صحيح كفئة هجوم.

جدول 1: مصفوفة التشتت [4].

Confusion Matrix		Predicted Value	
		Normal	Attack
Actual Value	Normal	TP	FN
	Attack	FP	TN

وباستخدام مصفوفة التشتت Confusion Matrix يمكن تحديد الفئات التي يواجه النموذج صعوبة في تصنيفها وتحديد أدائها [13]. وتعتمد معظم مقاييس الأداء على مصفوفة التشتت المستخدمة لتقييم الأداء، حيث توضح القيم الموجودة فيها أداء خوارزمية التصنيف، يتم عرض مقاييس الأداء المتبعة لتقييم نظام كشف التسلل كالتالي [4]:

1. الدقة Accuracy

وهي النسبة بين السجلات المصنفة بشكل صحيح على كافة سجلات مجموعة البيانات ويتم حسابها كما في المعادلة (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. الاستدعاء Recall

وهو نسبة التسميات الصحيحة True Labels التي تم تصنيفها بشكل صحيح على جميع التسميات الايجابية Positive Labels ويتم حسابه كما في المعادلة (2):

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

وهو يعرف أيضابمعدل الكشف Detection Rate (DR) أو الحساسية Sensitivity (وهو من ضمن مقاييس تستخدم في تقييم نظام كشف التسلل) أو المعدل الإيجابي الحقيقي True Positive Rate، وكلما ارتفعت قيمته كان ذلك أفضل [14].

3. الدقة أو الضبط Precision

نسبة التسميات الصحيحة التي تم تصنيفها على جميع التسميات ويتم حسابه كما في المعادلة (3):

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4. F-Measure

وهو المتوسط التوافقي لـ Recall وPrecision ويتم حسابه كما في المعادلة (4):

$$F - Measure = 2\left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (4)$$

5. معدل التنبيه الكاذب False Alarm Rate (FAR)

ويعرف ايضًا باسم المعدل الإيجابي الكاذب False Positive Rate (FPR) ويتم تعريفه على أنه نسبة التنبؤ السلبي والتي تعتبر إيجابيات خاطئة (شذوذ Anomaly) لجميع التنبؤات السلبية وكلما انخفضت قيمته كان ذلك أفضل ويتم حسابه كما في المعادلة (5) [14]:

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad (5)$$

iii. تقنيات اختيار الميزة

يتم استخدام تقنيات اختيار الميزات Features Selection لمواجهة الأبعاد العالية في مجموعة البيانات الكبيرة لأنه ليست كل البيانات الموجودة في الشبكة ذات صلة وقد تؤثر على أداء نظام كشف التسلل [15]. وفي اختيار الميزات يتم تحديد الميزات ذات الصلة من مجموعة البيانات وإزالة الميزات الأقل أهمية والتي لا تساهم كثيرًا في المتغير المستهدف من أجل تحسين أداء كشف التسلل بالإضافة إلى تقليل وقت التدريب للنموذج [16]، ويقدم الجدول 2 وصف مختصر للطرق الثلاثة الرئيسية لاختيار الميزات.

جدول 2: وصف مختصر لطرق اختيار الميزات [15].

طريقة اختيار الميزة	مميزاتها	عيوبها	أمثلة عليها
طريقة التصفية Filter Method	سريعة وكفاءتها الحسابية عالية و أقل عرضة للتجهيز الزائد.	قد تفشل في العثورعلى الميزات ودقتها أقل من طريقة التغليف.	ارتباط بيرسون Pearson Correlation. مربع كاي Chi- Square.
طريقة التغليف Wrapper Method	عالية الدقة.	مكلفة حسابيًا وبطيئة بسبب التكرار، عرضة لمشاكل التجهيز الزائد Overfitting وتعتمد على نماذج التعلم الآلي في اختيار الميزات.	التحديد الأمامي Forward Selection. الحذف العكسي Backward Elimination. إزالة الميزة التكرارية Recursive Feature Elimination.
الطريقة المضمنة Embedded Method	تكلفة إضافية منخفضة	مثل عيوب طريقة التغليف.	Lasso, Elastic Net

ويوجد تقنيات أخرى لاختيار الميزات وتعرف **بالأساليب القائمة على التعلم** Learning-based-Methode وتشمل بعض خوارزميات التعلم غير خاضع للاشراف Unsupervised Learning Algorithmes وخوارزميات التعلم شبة خاضعة للاشراف Semi Supervised Learning Algorithmes وخوارزميات التعلم الخاضعة للاشراف Supervised Learning Algorithmes وخوارزميات التعلم الجماعي Ensemble Learning Algorithmes وتعد تقنيات التجميع Clustring Techniques مثالاً اساسيًا للتعلم غير خاضع للاشراف ويمكن أن يوفر اختيار الميزات بواسطة التعلم غير خاضع للاشراف وصفًا وموثوقية أفضل للبيانات، كذلك تزعم التوصيات الموجودة في دراسة [17] أن تقنيات اختيار الميزات المستندة إلى التعلم الجماعي تتفوق على التقنيات الأخرى وخاصة مصنفات الأشجار الإضافية Extra Tree Classifier [18] [19].

وبالرجوع إلى تقنيات اختيار الميزة التي تم تقديمها في الجدول السابق نلاحظ أنه لا تتضمن أساليب التصفية نموذج للتعلم الآلي لتحديد ما إذا كانت الميزة جيدة أم سيئة بينما تستخدم أساليب التغليف نموذج التعلم الآلي وتدريب الميزة لتحديد ما إذا كانت ضرورية أم لا، بالإضافة إلى ذلك تعد طرق التصفية أسرع بكثير مقارنةً بطرق التغليف لأنها لا تتضمن تدريب النماذج ومن ناحية أخرى تعد طرق التغليف مكلفة من الناحية الحسابية وفي حالة البيانات الضخمة لا تعد طرق التغليف هي الطريقة الأكثر فعالية لإختيار الميزات التي يجب مراعاتها [16].

iv. تقسيم البيانات

عند تنفيذ نماذج التعلم الآلي لأغراض التنبؤ أو التصنيف من المهم تقسيم البيانات بشكل صحيح لتقييم أداء النماذج ولتجنب أوجه القصور في

البيانات يتم تقسيم البيانات إلى مجموعتين وهما مجموعة التدريب ومجموعة الاختبار، حيث تستخدم مجموعة التدريب لتدريب النماذج وتستخدم مجموعة الاختبار لتقييم أداء النموذج النهائي. ولا توجد قاعدة واضحة حول حجم مجموعات التدريب والاختبار ومع ذلك عادة ما تكون مجموعة التدريب هي الجزء الأكبر من البيانات حيث يجب تدريب نماذج التعلم الآلي على مجموعة كمية كبيرة من البيانات لتكون فعالة [9]. يعتبر تقسيم مجموعة البيانات بشكل عشوائي بنسبة 80% تدريب و20% اختبار أفضل طريقة لتجنب مشكلة الافراط في الملائمة Overfitting حيث يحفظ النموذج البيانات بدلاً من التعلم منها [20].

٧. مجموعات البيانات المعيارية لنظام كشف التسلل الى الشبكة

مجموعات البيانات الشائعة لكشف التسلل إلى الشبكة هي:

1. KDD Cup '99

تم جمع هذه البيانات في مختبر لينكولين بمعهد ماساتشوستس للتكنولوجيا(Massachusetts Institute of Technology (MIT منذ عام 1999 [21]. وهي احدى من مجموعات البيانات الأكثر شيوعًا والأكثر استخدامًا في نظم IDS وتحتوي على ما يقارب 5 مليون سجل للتدريب و2 مليون سجل للاختبار، ويحتوي كل سجل على 41 ميزة أو سمة مختلفة ويتم تصنيفها على انها عادية Normal أو هجوم Attack، يتم تصنيف الهجمات إلى أربع فئات مختلفة مثل هجوم رفض الخدمة Denial of Service Attack (DoS) وهجوم الاستجواب Probe Attack وهجوم البعيد إلى المحلي Remote to Local Attack (R2L) وهجوم المستخدم إلى الجذر User to Root Attack (U2R) [2]. وتم وصف هذه الفئات ادناه [21]:

- **هجوم رفض الخدمة Denial of Service Attack (DoS):** وفيه لا يسمح المهاجم للمستخدمين الشرعيين بالوصول إلى موارد الحوسبة او يقوم بتحميلها بشكل زائد بحيث لا يمكن معالجة الطلبات في الوقت الفعلي.
- **هجوم الاستجواب Probe Attack:** يقوم المهاجم بجمع معلومات حول النظام أو شبكة الكمبيوتر للعثور على نقاط الضعف المعروفة عن طريق مسح Scan جهاز أو جهاز الشبكة من أجل تحديد نقاط الضعف التي قد تم استغلالها فيما بعد من أجل اختراق النظام.
- **هجوم البعيد إلى المحلي(R2L) Remote to Local Attack:** وفيه المهاجم ليس لديه حساب مستخدم على جهاز الضحية وبالتالي يحاول الوصول الى النظام البعيد دون أن يكون لديه حساب.
- **هجوم المستخدم إلى الجذر(U2R) User to Root Attack:** وفيه يستكشف المهاجم نقاط الضعف من أجل الوصول إلى امتيازات المسؤول (الوصول الجذري إلى النظام) يبدأ المهاجم في النظام بإستخدام حساب المستخدم العادي ويبحث عن نقاط الضعف من أجل الحصول على امتيازات المستخدم الفائقة.

نظرًا لأن مجموعة بيانات KDD Cup ' 99 عبارة عن محاكاة لحركة مرور الشبكة فيوجد بها عدد كبير من السجلات الزائدة في مجموعة التدريب والسجلات المكررة في مجموعة الاختبار مما يمنع تصنيف السجلات الأخرى التي ليست زائدة عن الحاجة [21]. فعند تحليل مجموعات التدريب والاختبار وجد أن حوالي 78% و 75% من السجلات مكررة في مجموعتي التدريب والاختبار على التوالي وستؤدي هذه الكمية الكبيرة من السجلات الزائدة في مجموعة التدريب إلى انحياز خوارزميات التعلم نحو السجلات الأكثر تكرارًا وبالتالي منعها من تعلم السجلات غير المكررة والتي عادة ما تكون أكثر ضررًا للشبكات مثل هجمات المستخدم إلى الجذر User to Root (U2R)، ومن ناحية أخرى سيؤدي وجود سجلات مكررة في مجموعة الاختبار الى انحياز نتائج التقييم للطرق التي تتمتع بمعدلات اكتشاف افضل على السجلات المتكررة [22]. تحتوي مجموعة بيانات KDD Cup '99 على 24 نوع من الهجمات التدريبية مع 14 نوع إضافي في بيانات الاختبار [22]. الجدول 3 يوضح احصائيات السجلات الزائدة في مجموعة بيانات KDD Cup '99 (تدريب واختبار).

جدول 3: احصائيات السجلات الزائدة في مجموعة بيانات KDD Cup '99 (تدريب واختبار)[22].

الفئات	السجلات الأصلية	السجلات المميزة	معدل التخفيض	نسبة الفئة الى حجم مجموعة البيانات
مجموعة بيانات التدريب				
الهجمات (Attacks)	3,925,650	262,178	%93.32	%80.14
الحالات العادية (Normal)	972,781	812,814	%16.44	%19.86
الإجمالي	4,898,431	1,074,992	%78.05	%100
مجموعة بيانات الاختبار				
الهجمات (Attacks)	250,436	29,378	%88.26	%80.52
الحالات العادية (Normal)	60,591	47,911	%20.92	%19.48
الإجمالي	311,027	77,289	%75.15	%100

بعد تخفيض عدد سجلات الاختبار من 2 مليون إلى 311,027 أصبح اجمالي السجلات التدريب مع الاختبار 5,209,458 سجل [22].

2. Kyoto2006+

تم انشاء مجموعة البيانات هذه من سجلات حركة مرور الشبكة والتي تم الحصول عليها عن طريق نشر مصائد مخترقي الشبكات Honeypots وأجهزة استشعار الشبكة المظلمة Darknet Sensors وخوادم البريد الإلكتروني Email Servers ومستكشف الويب Web Crawler وغيرها من التدابير الأمنية للشبكة من قبل جامعة كيوتو، تتضمن أحدث مجموعة بيانات سجل حركة المرور من عام 2006 إلى عام 2015 ويحتوي كل سجل على 24 ميزة إحصائية 14 منها مستمدة من مجموعة بيانات KDD Cup '99 بينما العشر ميزات المتبقية هي ميزات إضافية [2]. خلال فترة المراقبة كانت هناك 50,033,015 جلسة عادية و43,043,225 جلسة هجوم و 425,719 جلسات تتعلق بهجمات غير معروفة [21].

3. NSL-KDD

لأن مجموعة بيانات KDD Cup '99 تحتوي على سجلات زائدة بنسبة 78%

وسجلات مكررة بنسبة 75% مما يمنع تصنيف السجلات الأخرى ولإصلاح هذه المشكلة تم انشاء مجموعة بيانات جديدة وهي NSL- KDD [21]. NSL-KDD عبارة عن مجموعة بيانات مقدمة لتسوية العديد من الأمور المتأصلة في مجموعة بيانات KDD Cup'99، وعلى الرغم من أن هذا الإصدار من مجموعة بيانات KDD يواجه قدرًا كبيرًا من المشكلات التي ذكرها McHugh حيث أنها لا تمثل بشكل فعال الشبكات الحقيقية الموجودة في الحياة العملية وبسبب عدم وجود مجموعات بيانات عالمية لمعرفات الشبكة يتم استخدامها على نطاق واسع بكفاءة كمجموعة بيانات مرجعية. علاوة على ذلك، فإن حجم سجلات التدريب والاختبار لمجموعة بيانات NSL-KDD مناسب وعدد فئات الهجوم في مجموعة بيانات NSL-KDD أربع فئات هي هجوم رفض الخدمة Denial of Service Attack(DoS) وهجوم الاستجواب Probe Attack وهجوم البعيد إلى المحلي Remote to Local Attack(R2L) وهجوم المستخدم إلى الجذر User to Root Attack (U2R)[23]. وتم تقديم وصف هذه الفئات في مجموعة البيانات السابقة KDD Cup '99.

تحتوي مجموعة بيانات NSL-KDD على الاتي مقارنة بـ KDD الأصلية [23]:

- لا توجد سجلات زائدة في مجموعة بيانات التدريب وبالتالي لن يتأثر المصنف بالعديد من السجلات المتكررة.
- لا يوجد أي سجل متكرر في مجموعة الاختبار لذلك لا يتماشى أداء المتعلمين مع الاستراتيجيات التي لديها معدلات اكتشاف أعلى في السجلات المتكررة.
- يرتبط عدد السجلات المختارة من كل مجموعة على مستوى المشكلة بشكل عكسي بكمية السجلات في مجموعة بيانات KDD الأصلية.
- أعداد السجلات في مجموعة بيانات التدريب ومجموعة بيانات الاختبار مرضية مما يجعل من المعقول اجراء التجارب على السجلات بأكملها وعدم اختيار قدر صغير منها.

بالإضافة الى ذلك مجموعة بيانات NSL-KDD لديها 41 ميزة تصنف على أنها عادية أو هجوم [23]. تتكون مجموعة بيانات التدريب من 21 هجوم مختلف من أصل 37 هجوم موجود في مجموعة بيانات الاختبار لمجموعة بيانات NSL-KDD. الهجمات المعروفة هي تلك الموجودة في مجموعة التدريب في حين أن الـ 16 الإضافية متاحة فقط في مجموعة الاختبار يتم تجميع أنواع الهجمات في 4 فئات وهي DOS، Probe ، R2L، U2R [21]. الجدول 4 يوضح توزيع الهجمات والحالات العادية في مجموعة بيانات NSL-KDD (تدريب واختبار).

جدول 4: توزيع الهجمات والحالات العادية في مجموعة بيانات-NSL KDD(تدريب واختبار)[24].

الفئات	السجلات الأصلية	نسبة الفئة الى حجم مجموعة البيانات
مجموعة بيانات التدريب		
عادي (Normal)	67,343	%53.46
الهجوم (Attack)	58,630	%46.54
الإجمالي	125,973	%100
مجموعة بيانات الاختبار		
عادي (Normal)	9711	%43.07
الهجوم (Attack)	12,833	%56.93
الإجمالي	22,544	%100

4. UNSW-NB15

تم انشاء مجموعة البيانات هذه بواسطة المركز الأسترالي للأمن السيبراني في 2015 وهي تحتوي على ما يقارب من مليوني سجل و49 ميزة التي يتم استخلاصها بإستخدام أدوات Bro-IDS وأدوات Argus وبعض الخوارزميات المطوّرة حديثًا تحتوي مجموعة البيانات هذه على فئات الهجمات المسماة [2]: Fuzzers, Port Scan, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms.

وتغطي مجموعة بيانات UNSW-NB15 كشف التسلل في أجهزة إنترنت الأشياء [1]. كانت فترة المحاكاة لمجموعة البيانات هذه 16 ساعة في 22 يناير 2015 و15 ساعة في 17 فبراير 2015 لالتقاط 100 جيجابايت. مجموعة بيانات UNSW-NB15 تتكون من 2,540,044 من السجلات التي تم تخزينها في 4 ملفات CSV، ويبلغ اجمالي عدد الهجمات في مجموعة البيانات 321,283 والحالات العادية 2,218,761 سجل ويمثل حجم حزم المعلومات العادية 87.35% من حجم مجموعات البيانات في حين تمثل حزم معلومات الهجوم 12.65% [25]. مجموعة بيانات التدريب لـ UNSW-NB15 الأكثر استخدامًا تتضمن 175,341 سجل ومجموعة بيانات UNSW-NB15 الخاصة بالاختبار بها 82,332 سجل [18]. وتم وصف فئات الهجوم كالاتي [26]:

- **هجوم التشويش Fuzzers Attack**: هي عملية تلقائية للعثور على أخطاء برمجية قابلة للاختراق عن طريق تغذية التباديل المختلفة للبيانات بشكل عشوائي في برنامج مستهدف حتى تكتشف أحد تلك التباديل عن ثغرة أمنية.
- **هجوم فحص المنفذ Port Scan Attack:** نوع عام لوصف فحص المنفذ لتحديد المنافذ المفتوحة في شبكة الكمبيوتر من خلال ارسال طلبات اتصال إلى مجموعة متنوعة من المنافذ على جهاز الكمبيوتر.
- **هجوم الوصول الخفي Backdoors Attack:** هو نوع من البرامج الضارة الذي ينفي المصادقة العادية لمنح الوصول عن بعد إلى الموارد مثل قواعد البيانات وخوادم الملفات.
- **هجوم رفض الخدمة Denial of Service Attack (DoS)**: تم التطرق له سابقًا.
- **هجوم استغلال الثغرات Exploits Attack:** هو نوع من التعليمات البرمجية التي تستغل ثغرة أمنية في البرامج أو ثغرة أمنية غالبًا ما يتم دمجها في البرامج الضارة مما يسمح له بالانتشار.
- **الهجوم العام Generic Attack:** هو هجوم تصادمي على المفاتيح السرية للشفرات.
- **هجوم الاستكشاف Reconnaissance Attack:** وهو مجموعة من التقنيات البسيطة التي تجمع معلومات حول الشبكة أو الخادم المستهدف.
- **هجوم كود القشرة Shellcode Attack:** مجموعة من التعليمات أو البيانات التي يتم إدخالها وتنفيذها بواسطة برنامج معيب ويتلاعب مباشرة بالسجلات ووظائف البرامج.
- **هجوم الديدان Worms Attack:** وهو عبارة عن تعليمات برمجية ضارة ذاتية النسخ وتستهلك الكثير من ذاكرة النظام وعرض النطاق الترددي للشبكة ويقلل من توافر الأنظمة.

الخصائص الرئيسية لمجموعة بيانات UNSW-NB15 هي مزيج من السلوكيات الطبيعية الحديثة والحقيقية وأنشطة الهجوم الاصطناعية [27]. الجدول 5 يوضح توزيع الهجمات والحالات العادية في مجموعة بياناتUNSW-NB15.

جدول 5: توزيع الهجمات والحالات العادية في مجموعة بيانات -UNSW NB15 [25].

الفئة	عدد السجلات	نسبة الفئة الى حجم مجموعة البيانات
عادي (Normal)	2,218,761	%87.35
الهجوم (Attack)	321,283	%12.65
الإجمالي	2,540,044	%100

5. CIC-IDS 2017

تم انشاء مجموعة البيانات هذه من قبل المعهد الكندي للأمن السيبراني Canadian Institute of Cyber Security (CIC) في عام 2017 وهي تحتوي على التدفقات العادية والهجمات الواقعية المحدثة، يتم تحليل حركة مرور الشبكة بواسطة CIC FlowMeter بإستخدام المعلومات المستندة إلى الطوابع الزمنية وعناوين IP المصدر والوجهة والبروتوكولات والهجمات، علاوة على ذلك يتضمن CIC-IDS2017 سبع فئات هجوم وهي [2]: Brute Force, Heartbleed, Botnet, DoS, DDoS, Web Attack, Infiltration.

تتضمن مجموعة بيانات CIC-IDS 2017 اجمالي 2,830,743 سجل و 78 ميزة أي مجموعة بيانات ذات أبعاد عالية وهي مجموعة بيانات غير متوازنة لان نسبة الحالات العادية تفوق نسبة الهجوم [28]. يتم وصف الخصائص الرئيسية لفئات الهجمات المختلفة أدناه [29]:

- **الهجوم العام Brute Force Attack:** تم التطرق اليه في مجموعة بيانات UNSW-NB15.
- **هجوم ثغرة النزيف القلبي Heartbleed Attack:** يتم تنفيذه عن طريق ارسال طلب نبضات قلب مشوه من أجل تحفيز استجابة الضحية.
- **شبكة الأتمتة Botnet:** تُستخدم شبكات أجهزة الكمبيوتر المسروقة (المسماة الزومبي) الخاضعة لسيطرة مجرمي الانترنت المستخدمة لتنفيذ العديد من عمليات الاحتيال وحملات البريد العشوائي والهجمات الإلكترونية.
- **هجوم رفض الخدمة Denial of Service Attack (DoS):** تم التطرق اليه سابقًا.
- **هجوم رفض الخدمة الموزع Distributed Denial of Service Attack (DDoS):** نتيجة الأنظمة المخترقة المتعددة التي تغمر النظام المستهدف عن طريق توليد حركة مرور كبيرة على الشبكة، مما يؤدي إلى تجاوز عرض النطاق الترددي أو موارد شبكة الضحية.
- **هجوم الويب Web Attack:** يستهدف نقاط الضعف في مواقع الويب للوصول غير المصرح به أو الحصول على معلومات سرية أو تقديم محتوى ضار او تغيير موقع الويب في حقن SQL يستخدم المهاجم سلسلة من أوامر SQL لإجبار قاعدة البيانات على الرد في البرمجة النصية عبر مواقع (XSS)، يجد المهاجمون إمكانية حقن البرنامج النصي عندما لا يتم اختبار الكود بشكل صحيح، وفي Brute Force over HTTP يحاولون على قائمة كلمات المرور للعثور على كلمة مرور المسؤول.
- **هجوم التسلل Infiltration Attack:** يتم تطبيقه عادةً من خلال استغلال البرامج الضعيفة الموجودة في جهاز الكمبيوتر المستهدف.

الجدول 6 يوضح توزيع الهجمات والحالات العادية في مجموعة بيانات CIC-IDS2017.

جدول 6: توزيع الهجمات والحالات العادية في مجموعة بيانات -CIC IDS2017 [28].

الفئات	السجلات الأصلية	نسبة الفئة الى حجم مجموعة البيانات
الهجوم (Attack)	557,646	%19.70
عادي (Normal)	2,273,097	%80.30
الإجمالي	2,830,743	%100

6. CSE-CIC-IDS2018

تم انشاء مجموعة البيانات هذه بشكل مشترك بواسطة مؤسسة أمن الاتصالات Communications Security Establishment (CSE) والمعهد الكندي للأمن السيبراني The Canadian Institute of Cyber Security (CIC) وتحتوي على تمثيل مجرد للأحداث المختلفة وبالنسبة لجيل مجموعة البيانات يتم دمج كل هذه الملفات الشخصية مع مجموعة فريدة من الميزات وتتضمن سبع فئات هجوم مثل الموجودة في مجموعة بيانات CIC-IDS 2017 [2]. الجدول 7 يوضح توزيع الهجمات والحالات العادية في مجموعة بيانات CSE-CIC-IDS 2018.

جدول 7: توزيع الهجمات والحالات العادية في مجموعة بيانات -CSE CIC-IDS 2018[25].

الفئات	السجلات الأصلية	نسبة الفئة إلى حجم مجموعة البيانات
عادي(Normal)	2,856,035	%63.11
الهجوم (Attack)	1,669,364	%36.89
الإجمالي	4,525,399	%100

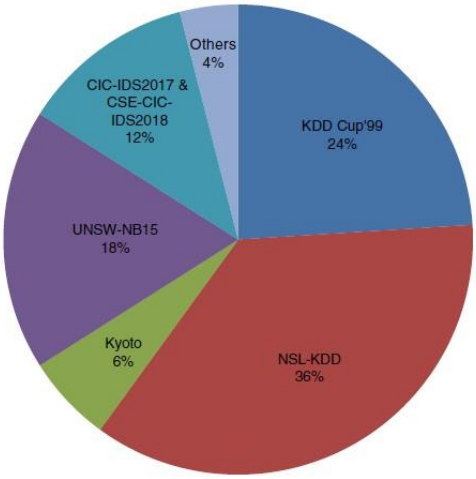
vi. مقارنة بين مجموعات البيانات المعيارية لنظام كشف التسلل إلى الشبكة

الجدول 8 يوضح مقارنة بين مجموعات البيانات المعيارية لنظام كشف التسلل إلى الشبكة من حيث: السنة ومكان التطوير ومدة الجمع بالإضافة إلى اتاحتها للعامة واحتوائها على الهجمات الحديثة وعدد الميزات وعدد السجلات وكذلك عدد فئات الهجوم في كل مجموعة بيانات.

جدول 8: مقارنة بين مجموعات البيانات المعيارية لنظام كشف التسلل إلى الشبكة.

مجموعة البيانات	السنة	طورت بواسطة	مدة جمعها	هجمات حديثة	عدد الميزات	عدد السجلات الكلي	عدد فئات الهجوم	IoT
KDD Cup '99	1999 [4]	جامعة كاليفورنيا [30]	7 أسابيع [4]	× [4]	41 [2]	5,209,458 سجل [22]	4 (DoS, Prob, U2R, R2L) [2]	× [1]
Koto 2006+	2009 [4]	جامعة كيوتو [30]	3 سنوات [4]	× [4]	24 [21]	93,076,270 سجل [25]	2 هجمات (معروفة وغير معروفة) [2]	× [1]
NSL-KDD	2009 [4]	جامعة كاليفورنيا [30]	31 ساعة [4]	× [4]	41 [2]	148,517 سجل [24]	4 (DoS, Prob, U2R, R2L) [2]	× [1]
UNSW-NB15	2015 [4]	المركز الأسترالي للأمن السيبراني [2]	31 ساعة [4]	✓ [4]	49 [2]	2,540,044 سجل تم تخزينها في 4 ملفات CSV [25]	9 (Backdoor, DoS, Exploits, Fuzzers, Generic, Port Scan, Reconnaissance, Shellcode, Worms) [2]	✓ [1]
CIC-IDS 2017	2017 [4]	المعهد الكندي للأمن السيبراني [30].	5 أيام [4]	✓ [4]	78 [28]	2,830,743 سجل [28]	7 (Burte Force, Heartbleed, Botnet, Dos, DDoS, Web Attack, Infiltration) [2]	✓ [1]
CSE-CIC-IDS2018	2018 [4]	المعهد الكندي للأمن السيبراني [30]	10 أيام [4]	✓ [4]	80 [30]	4,525,399 سجل [25]	7 (Burte Force, Heartbleed, Botnet, Dos, DDoS, Web Attack, Infiltration) [2]	✓ [1]

لا شكّ أن مجموعات البيانات المعيارية Benchmarked Datasets عنصرًا مهمًا وتستخدم لاختبار أداء المنهجية المقترحة [2]. ويظهر الشكل 2 تحليل استخدام مجموعات البيانات العامة، حيث أنه تم استخدام مجموعة بيانات UNSW-NB15 [31] بنسبة 18% من المرات في الأبحاث في الفترة من2019 إلى 2023 وهي تعتبر ثالث أعلى نسبة مقارنة بمجموعات البيانات الأخرى. وايضاً من خلال الجدول 8 تظهر مزايا مجموعة البيانات UNSW- NB15 مقارنةً بالمجموعات الأخرى.



شكل 2: تحليل استخدام مجموعات البيانات العامة [2].

فبمقارنة مجموعة البيانات UNSW-NB15 بمجموعات البيانات الأخرى مثل KDD Cup '99 وNSL-KDD نجد أن مجموعة بيانات KDD Cup '99 ونسختها المحسنة NSL-KDD هما الأكثر استخدامًا على نطاق واسع في نظام كشف التسلل إلى الشبكة NIDS ومع ذلك أشارت كثير من الدراسات إلى أن مجموعات البيانات هذه لا تعكس أداء مخرجات كشف التسلل إلى الشبكة بشكل جيد بسبب بعض المشكلات الأتية [32]:

1. الهجمات الحديثة ذات الأثر المنخفض مفقودة لذلك ستقترب هجمات التجسس والتسلل من السلوكيات الطبيعية بمرور الوقت.
2. منذ انشاء مجموعات البيانات هذه قبل عشرين عامًا هناك فرق كبير بين التدفق الطبيعي الذي تحدده مجموعات البيانات القياسية والتدفق الطبيعي الحالي.
3. يختلف توزيع مجموعات بيانات التدريب والاختبار المعيارية الحالية باختلاف أنواع البيانات مما سيودي إلى انحراف المصنف وانخفاض الدقة.
4. بالإضافة إلى ذلك تعتبر النسخة المحسنة من KDD Cup '99 وهي NSL-KDD ليست تمثيل مثالي للشبكات الحقيقية الموجودة.

واستجابةً لهذ التحديات تم انشاء مجموعة بيانات UNSW-NB15 في عام 2015 والتي تولد مزيجًا من السلوكيات الطبيعية الحديثة والحقيقية وبيئة الهجوم الشامل الحالية وبالمقارنة بينها وبين مجموعات بيانات KDD القديمة والمحسنة نجد أن مجموعة البيانات UNSW-NB15 تتضمن أنواعًا غير طبيعية وأكثر حداثة وسلوكيات عادية تم التقاطها بمرور الوقت إلى جانب انها تتضمن ميزات متعمقة لحركة مرور الشبكة [32].

وبمقارنتها مع مجموعة بيانات Kyoto 2006+، نجد أن مجموعة بيانات Kyoto 2006+ قديمة نوعًا ما ولا تحتوي على هجمات حديثة وكذلك لا تحتوي على معلومات حول هجمات معينة [21].

من الواضح أن النموذج الذي سيتم تدريبه والتحقق منه باستخدام أحدث مجموعات البيانات مثل UNSW-NB15 سيكون أداؤه أفضل نسبيًا من النموذج الذي يتم تدريبه والتحقق منه باستخدام مجموعات بيانات قديمة في العالم الحقيقي [2].

أما مجموعات البيانات الحديثة الأخرى وهي CIC-IDS2017 و CSE-CIC-IDS2018 هناك قيود حول استخدامها ومنها [30]:

1. يتم تخزين عينات البيانات الناتجة عن تحليل تدفق الشبكة في ملفات وتعد معالجة هذه الملفات مهمة شاقة للغاية حيث تحتوي هذه الملفات على عدد كبير من مثيلات البيانات في كل ملف.
2. يمكن دمج الملفات الموجودة في مجموعة البيانات لتشمل كل تصنف من تسميات الهجوم للمعالجة لكن الجمع بين مثيلات كل نوع من أنواع الهجوم يزيد من حجم مجموعة البيانات مما يؤدي إلى مزيد من وقت الحوسبة والمعالجة.

3. تتكون مجموعات البيانات هذه من سجلات مفقودة وسجلات مكررة.
4. مجموعات البيانات هذه عرضه لمشكلة التوازن الذي يؤدي إلى انخفاض الدقة وارتفاع معدل الإيجابيات الكاذبة False Positive Rate (FPR) للنظام.

وبذلك نجد أن هذه البيانات المعقدة تحتاج إلى الكثير من الوقت في الحوسبة والمعالجة وتحليل هذه البيانات يعتمد على استخدام التعلم العميق Deep Learning (DL) [2].

vii. الدراسات السابقة

فيما يلي سرد لأهم الدراسات التي تناولت خوارزميات التصنيف المختلفة التي تم استخدامها لتصنيف مجموعة بيانات UNSW-NB15 بشكل خاص.

1. كشف التسلل إلى الشبكة استنادًا إلى خوارزمية SVM لمجموعة بيانات UNSW-NB15

في 2019 قام الباحثان Jing و Chen باستخدام خوارزمية آلة المتجهات الداعمة SVM على مجموعة بيانات UNSW-NB15، وفي هذه الدراسة يوجد 257,673 سجل لمجموعة البيانات هذه بما في ذلك 175,341 سجل للتدريب و82,332 سجل للاختبار ولا تحتوي هذه السجلات على تكرار لضمان موثوقية تقييمات نظام كشف التسلل إلى الشبكة NIDS، وتحتوي كلتا مجموعتي البيانات على 44 ميزة. وفي هذه الدراسة لم يتم اختيار الميزات وبعد ذلك استخدمت خوارزمية SVM مع التصنيفين وتم التقييم من خلال الدقة Accuracy ومعدل الإيجابيات الخاطئة False Positive Rate (FPR) وكانت النتائج جيدة، فبالنسبة للتصنيف الثنائي كانت دقة خوارزمية SVM هي 85.99% ومعدل الإيجابيات الخاطئة وصل إلى 16.50%، كما حققت نتائج التصنيف المتعدد بإستخدام خوارزمية SVM دقة بنسبة 75.77% ومعدل إيجابيات خاطئة يساوي 3.04% [32].

2. استخدام تقنيات التعلم الآلي لتحديد الهجمات الإلكترونية النادرة على مجموعة بيانات UNSW-NB15

استخدم Bagui وآخرون في2019 تقنيات اختيار الميزات المختلطة Hybrid Feature Selection وتقنيات التصنيف لتصنيف الهجمات السيبرانية في مجموعة بيانات UNSW-NB15، تقنيات التصنيف التي قاموا باستخدامها الأولى كانت تقنية احتمالية وتسمى بايز البسيط Naïve Bayes (NB) والثانية على أساس أشجار القرار وهي Java 48 (J48) وهي تعتمد على مجموعة من القرارات البسيطة لأتخاذ القرار النهائي وتم اجراء التصنيف بإستخدام أداة Weka. قاموا اولاً باختيار عينة مكونة من 8000 سجل بشكل عشوائي من مجموعة بيانات UNSW-NB15 بعد ذلك قاموا باختيار الميزة بإستخدام مزيج من K-التجمع K-Means Clustering وهي تتبع طريقة اختيار الميزة القائمة على التعلم الآلي. واختيار الميزة على أساس الارتباط Correlation-based Feature Selection (CFS) وهو يتبع طريقة الترشيح Filter Method، تم استخدام التقنيتين كتقنية مختلطة ومن خلال هذا النموذج الهجين تم التوصل إلى مجموعة فرعية من الميزات وعددها 29 ميزة ومن تم قامت الدراسة باستخدام خوارزميات التصنيف لتحديد الدقة Accuracy ومعدل اكتشاف الهجمات Attack Detection Rate (ADR) ومعدل الإنذارات الخاطئة False Alarm Rate (FAR). وأظهرت النتائج أن طريقة اختيار الميزات المختلطة مع مصنف NB أدت إلى تحقيق دقة عالية بنسبة 82.07% ومن دون اختيار الميزات كانت دقة المصنف NB هي 69.32%، ولم يكن هناك تحسن في معدل اكتشاف الهجمات ADR مع اختيار الميزات لمصنف NB حيث كان 75.57% ومن دون اختيار الميزات كان يساوي 92.78%، ولكن كان هناك تحسن في معدل الإنذارات الخاطئة FAR حيث كان معدل الإنذارات الخاطئة لمصنف NB قبل اختيار الميزة 26.89% وبعد اختيار الميزة أصبح 17.92%. اما في مصنف J48 فإن اختيار الميزة لم يكن له تأثير كبير على دقة التصنيف حيث كانت دقة هذا المصنف قبل اختيار الميزات تساوي 98.01% وبعد اختيار الميزات قلت إلى 96.63%، ومع ذلك كان معدل الكشف ADR منخفض مع اختيار الميزات حيث يبلغ 21.56% وبدون اختيار الميزات كان مرتفع ويصل إلى 46.18%، اما معدل الإنذارات الخاطئة FAR مع وبدون اختيار الميزات كان منخفض حيث يساوي 1.98% بدون اختيار الميزات و3.31% مع اختيار الميزة. وبشكل عام فإن خوارزمية J48 كان أدائها أفضل من NB من حيث الدقة المرتفعة ومعدل الإنذارات الخاطئة المنخفض مع أو بدون اختيار الميزة [33].

3. AD-IoT: الكشف عن الحالات الشاذة للهجمات الإلكترونية لإنترنت الأشياء للمدينة الذكية باستخدام التعلم الآلي

لمعالجة تهديدات الأمن السيبراني لإنترنت الأشياء في المدينة الذكية اقترح Alrashdi وآخرون في 2019 نظام الكشف عن شذوذ-إنترنت الأشياء Anomaly Detection – Internet of Things (AD-IoT) لاكتشاف الهجمات الإلكترونية لإنترنت الأشياء أو الأنشطة غير العادية في حركة مرور شبكة إنترنت الأشياء من شبكات الضباب الموزعة عبر المدينة الذكية بدلًا من الكشف عن الكمية الهائلة من التخزين السحابي للمدينة لتحديد السلوكيات الطبيعية وغير الطبيعية. خوارزمية التعلم الآلي التي تم استخدامها لتقييم هذا النظام هي خوارزمية الغابة العشوائية RF بإستخدام مجموعة بيانات UNSW-NB15، لم توضح هذه الدراسة عدد السجلات المستخدمة في التدريب والاختبار لكنها اظهرت أنه تم تقليل عدد الميزات إلى 12 ميزة بإستخدام مصنف الأشجار الإضافية Extra Trees Classifier (ETC) وهو يتبع طرق اختيار الميزة القائمة على التعلم الجماعي Ensemble Learning وكانت نسبة الضبط Precision والاستدعاء Recall و-F1 Score لخوارزمية RF جميعها تساوي 98% باستخدام الميزات المختارة [34].

4. نظام كشف التسلل بإستخدام اختيار الميزات مع خوارزميات التعلم الآلي للتجميع والتصنيف في مجموعة بيانات UNSW-NB15

في 2020 استخدم Hammad وآخرون أربع خوارزميات مختلفة لتصنيف الهجمات الإلكترونية في مجموعة بيانات UNSW-NB15 هذه الخوارزميات هي بايز البسيط NB والغابة العشوائية RF وخوارزمية C4.5 لبناء شجرة القرار J48 والتصنيف البسيط Zero R (ZR)، ايضًا تم استخدام خوارزمية التجميع K-Means وتعظيم التوقعات Expectation Maximization (EM) لتجميع مجموعة بيانات UNSW-NB15 في مجموعتين اعتمادًا على هجوم السمة المستهدفة أو حركة مرور الشبكة العادية. ولتطوير مجموعة فرعية من الميزات تم اختيار الميزات على أساس الارتباط Correlation-based Feature Selection (CFS) وعدد الميزات التي تم اختيارها هي 8 ميزات، تحتوي نسبة 10% من مجموعة البيانات على إجمالي 257,673 سجل وباستخدام تقنية التحقق المتقاطع Cross Validation Method تم استخدام 231,908 سجل للتدريب و25,765 سجل للاختبار، تم بعد ذلك قامت الدراسة بإستخدام تقنيات التصنيف والتجميع المذكورة وبعد ذلك تم تقييم كفاءة النموذج عن طريق حساب الدقة Accuracy والاستدعاء Recall والضبط Precision وF-Measure ومعدل الإيجابيات الخاطئة

False Positive Rate (FPR). الطرق المستخدمة في هذه الدراسة قدمت أداة فعالة لتحليل كشف التسلل في الشبكات الكبيرة وأظهرت النتائج أن الخوارزميتان RF، J48 حققت أفضل نتيجة دقة، بنسبة 97.59% لخوارزمية RF و93.78% بواسطة خوارزمية J48 عند استخدام تقنية CFS لاختيار الميزات [35].

5. تحسين أداء أنظمة كشف التسلل إلى الشبكات القائمة على التعلم الآلي في مجموعة بيانات UNSW-NB15.

في 2021 قدم Moualla وآخرون معرفات شبكة جديدة تلعب دورًا مهمًا في أمن الشبكات وتواجه الهجمات الإلكترونية الحالية على الشبكة بإستخدام مجموعة بيانات معيارية وهي UNSW-NB15، النظام الذي تم استخدامه في هذه الدراسة عبارة عن معرفات شبكة متعددة الطبقات قائمة على التعلم الآلي وقابلة للتطوير ديناميكيًا ويتكون من عدة مراحل تعتمد على التعلم الآلي الخاضع للإشراف. ومن أجل الحد من سوء التصنيف أول ما قاموا به لحل مشكلة الفئات غير المتوازنة هو استخدام تقنية الافراط في أخذ عينات الأقلية الاصطناعية Synthetic Minority Oversampling Technique (SMOTE) وهي طريقة شائعة جدًا لإعادة أخذ العينات خاصةً عندما تهيمن بعض الفئات على فئات على فئات أخرى، بعد هذه التقنية تم إجراء المعالجة المسبقة المتمثلة في تنظيف البيانات Data Cleaning وتحويل الميزات الأسمية إلى قيم رقمية عن طريق الترميز السريع One-Hot Encoding والقيم الرقمية التي تم الحصول عليها تمت تسويتها عن طريق استراتيجية Z-Score Normalization التي تطرح متوسط الميزات من كل ميزة وتقسم الفرق على الانحراف المعياري للميزات، تم بعد ذلك تم اختيار الميزات المهمة عن طريق مصنف الأشجار العشوائية للغاية Extremely Randomized Trees أو ما يعرف بمصنف الأشجار الإضافية Extra Trees Classifier (ETC) وهي تتبع طريقة اختيار الميزة القائمة على التعلم الجماعي Ensemble Learning حسب معيار Gini الذي يقيس الاحتمال غير الصحيح لميزة معينة عند اختيارها عشوائيًا وتتراوح قيمته من 0 إلى 1 وكلما انخفضت القيمة زادت أهمية الميزة ذات الصلة، تم اختيار 14 ميزة بواسطة هذه التقنية. وكان عدد سجلات التدريب 175,341 سجل أي 80% من مجموعة البيانات للتدريب وعدد سجلات الاختبار 82,332 سجل أي 20% من مجموعة البيانات للاختبار، وتم استخدام خوارزمية آلة التعلم المتطرفة Extreme Learning Machine (ELM) لاكتشاف الهجمات بشكل منفصل واحد مقابل الكل "كمصنف ثنائي" لكل منهما ومن تم أصبحت مخرجات ELM مدخلات لطبقة متصلة بالكامل من أجل تعلم جميع مجموعاتها وتليها طبقة الانحدار اللوجستي لاتخاذ قرارات سهلة لجميع الفئات. ومن أجل تقييم الأداء تم استخدام الدقة Accuracy والضبط Precision والاستدعاء Recall ومعدل الإنذارات الكاذبة False (FAR) Alarm Rate وخصائص المستقبل التشغيلية Receiver Operating Characteristic (ROC) ومنحنيات الاسترجاع والتدقيق Precision-Recall Curves (PRC). وأظهرت النتائج أن مصنف ELM أدى أفضل الأداء وكانت الدقة 98.19% ومعدل الإنذارات الكاذبة FAR يساوي 0.216% بإستخدام اختيار الميزات [18].

6. تحليل خوارزميات تعلم الآلة مع اختيار الميزات لاكتشاف التسلل بإستخدام مجموعة بيانات UNSW-NB15.

قاما الباحثان Kocher وKumar في2021 باستخدام مجموعة بيانات UNSW-NB15 لتدريب مصنفات التعلم الآلي، وتم استخدام المصنفات الأتية وهي K-الجيران الأقرب KNN وهبوط التدرج العشوائي Stochastic Gradient Descent (SGD) التي تقوم بتحديث النموذج باستخدام عينات عشوائية من البيانات التدريبية في كل خطوة مما يجعل عملية التدريب أسرع وتقلل من الحسابات الكبيرة وكذلك تم استخدام الغابة العشوائية RF والانحدار اللوجستي LR وبايز البسيط NB، وتم استخدام بيانات التدريب من UNSW-NB15 والتي تحتوي على 175,341 سجل واولاً قبل كل شيء تمت معالجة القيم الخالية الموجودة في مجموعة البيانات وبعدها تم تحويل البيانات الفئوية إلى بيانات رقمية بإستخدام تشفير السمة Label Encoder واستخدام التشفير السريع One- Hot Encoding لفصل العلاقة بين القيم التي تم الحصول عليها من خلال تشفير التسمية وتم بعد ذلك استخدام المقياس القياسي Standard Scaler لتوحيد القيم على مقياس واحد. بعد المعالجة المسبقة تم تحديد الميزات باستخدام مربع Chi-Square وهي تقنية اختيار الميزة القائمة على التصفية أو الترشيح A Filter-based Feature Selection، وبشكل عام تحتوي مجموعة بيانات UNSW-NB15 على 49 سمة أو ميزة وبعد تنفيذ اختيار الميزات تم تحديد 23 ميزة مهمة، تم بعد ذلك تم فصل البيانات التي تمت معالجتها مسبقًا إلى 80% من البيانات للتدريب و20% من البيانات للاختبار وتم استخدام المصنفات السابقة الذكر لبناء النماذج ومن تم التنبؤ بتسميات بيانات الاختبار بإستخدام هذه النتائج وتم اجراء مقارنة بين السمات الفعلية والمتوقعة، مقاييس الأداء التي تم استخدامها للتقييم هي الدقة Accuracy والاستدعاء Recall و F1-Score والمعدل الإيجابي الحقيقي True Positive Rate (TPR) والمعدل الإيجابي الخاطئ False Positive Rate (FPR) مع وبدون تقنية اختيار الميزات. وأظهرت النتائج أن مصنف RF أفضل من المصنفات الأخرى حيث تبلغ دقته مع جميع الميزات 99.57% و99.64% مع الميزات المحددة [11].

7. تأثيرات اختيار الميزة والتسوية على كشف التسلل إلى الشبكة

في 2023 قام الباحثان Umar وZhanfang بإجراء تحليل معمق لتأثيرات اختيار الميزة Feature Selection والتسوية Normalization على نماذج لنظام كشف التسلل IDS المختلفة والمبنية بإستخدام مجموعتي بيانات كشف تسلل وهما UNSW-NB15 وNSL-KDD وخمس خوارزميات تعلم آلي مختلفة وهي آلة المتجهات الداعمة SVM والشبكة العصبية الاصطناعية Artificial Neural Network (ANN) وK-الجار الأقرب KNN والغابة العشوائية RF وبايز البسيط NB، تم استخدام أدوات مثل Excel, Weka, Python لتحليل البيانات واستكشافها وJupyter Notebook كبيئة تنفيذ لبايثون. وفي المعالجة المسبقة من نوع التحويل تحديدًا (التشفير والتميز والتسوية) اولاً في التشفير تم تحويل الميزات الفئوية إلى ميزات رقمية بإستخدام التشفير الثنائي One-Hot Encoding او ما يعرف بتشفير dummy ونظرًا لأنه يزيد من ابعاد مجموعة البيانات لتجنب فقدان بعض الميزات الأساسية فقد تم اجراءه بعد عملية اختيار الميزات وتسويتها، ثانيًا في التسوية فقد تم استخدام الحد الأدنى والأقصى Min-Max على مجموعتي البيانات. وفي اختيار الميزات تم استخدام شجرة القرار المستندة إلى الغلاف Wrapper-based Decision Tree وبعد التشفير زادت ابعاد مجموعتي البيانات وتم استخدام الميزات المشفرة النهائية فقط في تدريب النماذج وتقييمها، في مجموعة بيانات UNSW-NB15 قبل التشفير كان عدد كل الميزات 42 ميزة تم اختيار 20 ميزة وبعد التشفير أصبحت كل الميزات 194 ميزة

والميزات المختارة أصبحت 172 ميزة اما مجموعة بيانات NSL-KDD قبل التشفير كان عدد الميزات 41 ميزة وتم اختيار 19 ميزة وبعد اجراء التشفير أصبح عدد الميزات 122 والميزات المختارة أصبح عددها 98 ميزة. واثناء استخدام مجموعة بيانات UNSW-NB15 تم استخدام مجموعة التدريب فقط لكل من التدريب والاختبار والتي تحتوي على 175,341 سجل، وكذلك عند استخدام مجموعة بيانات NSL-KDD تم استخدام مجموعة التدريب فقط لكل من التدريب والاختبار والتي تحتوي على 125,973 سجل، وتم تقسيم مجموعات البيانات التي تم استخدامها إلى 67% تدريب و33% اختبار، بعد ذلك تم تدريب خوارزميات التعلم الآلي التي تم اختيارها وتقييم أدائها، ووجدت هذه الدراسة أن التسوية أكثر أهمية من اختيار الميزات في تحسن الأداء والوقت الحسابي للنماذج في كلتا مجموعتي البيانات بينما فشل اختيار الميزة في UNSW-NB15 في تقليل الوقت الحسابي للنماذج وفي حالة النماذج المبنية باستخدام NSL-KDD فإن ذلك يقلل من أدائها، كشفت هذه الدراسة ايضًا أنه بالمقارنة مع مجموعة بيانات UNSW-NB15 فأن مجموعة بيانات NSL-KDD أقل تعقيدًا وغير مناسبة لبناء نماذج نظام كشف تسلل حديثة وموثوقة وعلاوة على ذلك تم تحقيق أفضل أداء في مجموعتي البيانات بإستخدام الغابة العشوائية RF بدقة تبلغ 99.75% في مجموعة بيانات NSL-KDD و98.51% في مجموعة بيانات UNSW-NB15 عند تطبيق اختيار الميزة، وقبل اختيار الميزة كانت دقة RF في مجموعة NSL-KDD 98.81% وفي UNSW-NB15 95.74% وتعمل NSL-KDD بشكل أفضل دون تطبيع وعلى العكس من ذلك يعمل UNSW-NB15 بشكل أفضل مع التطبيع وتم اجراء التطبيع على مجموعات البيانات الكاملة [14].

viii. مقارنة بين الدراسات السابقة

من المعروف أنه عند استخدام كل الميزات تكون دقة النموذج عالية لكنه يستغرق وقت أطول للتدريب ومع اختيار الميزات الهامة واستبعاد الميزات التي يكون ارتباطها ضعيف بالهدف يقل وقت التدريب مع الحفاظ على الدقة وعلى الرغم من أن استبعاد الميزات غير الهامة قد يؤدي إلى انخفاض الدقة بنسبة أقل من 1% إلاّ أنه يساهم في تدريب نماذج التعلم الآلي مع الميزات المهمة فقط وهذا بدوره يؤدي إلى تحسين أداء النماذج من خلال الحفاظ على الدقة وتقليل وقت التدريب.

في دراسة [32] تم استخدام كل الميزات مع خوارزمية آلة المتجهات الداعمة SVM في التصنيفين الثنائي والمتعدد حيث كانت دقة الخوارزمية في التصنيف الثنائي 85.99% وفي التصنيف المتعدد كانت الدقة 75.77% وعلى الرغم من أنه تم استخدام كل الميزات إلاّ أن الدقة في التصنيفين كانت منخفضة ربما يكون السبب وراء هذا أنه كانت مجموعة البيانات المستخدمة كبيرة حيث بلغ اجمالي السجلات 257,673 ولم يتم تقسيمها بشكل مناسب حيث تم تخصيص 68% من السجلات للتدريب ومن المفترض أن يكون التقسيم المثالي 80% للتدريب و20% للاختبار وخاصةً في مجموعات البيانات الكبيرة لتحقيق توازن أفضل بين تدريب النموذج واختباره.

وفي دراسة [33] أظهرت النتائج أن طريقة اختيار الميزات المختلطة مع خوارزمية NB كانت قادرة على تحسين دقة التصنيف وأوضحت هذه الدراسة أن سبب ذلك يرجع إلى أن خوارزمية NB لا تعمل بشكل جيد في حالة وجود ميزات زائدة أو غير ذات صلة وبالتالي فإن اختيار الميزات مهم جدًا عند استخدام هذه الخوارزمية، بينما كانت دقة خوارزمية J48 أفضل من NB مع وبدون اختيار الميزات.

وفي [34] تم اقتراح نظام الكشف عن شذوذ-إنترنت الأشياء Anomaly Detection – Internet of Things (AD-IoT) لاكتشاف الهجمات الإلكترونية أو الأنشطة غير العادية في حركة مرور شبكة إنترنت الأشياء وأظهرت النتائج أن خوارزمية RF المستخدمة والتي اعتمد عليها هذا النظام حققت دقة تصنيف عالية في هذه الدراسة باستخدام الميزات المختارة.

أظهرت النتائج في دراسة [35] أن الخوارزميتان RF، J48 حققتا أفضل نتيجة دقة، بنسبة 97.59% لخوارزمية RF و93.78% بواسطة خوارزمية J48 عند استخدام تقنية CFS لاختيار الميزات، وتم استخدام مجموعة كبيرة من السجلات لمجموعة بيانات UNSW-NB15 وتبلغ 257,673 سجل وتم تقسيمها إلى 90% تدريب و10% اختبار وقد يكون هناك عدم توازن في الفئات لهذا كانت الدقة منخفضة نوعًا ما.

بينما في دراسة [18] التي تم فيها استخدام خوارزمية آلة التعلم المتطرفة Extreme Learning Machine (ELM) وأظهرت النتائج أن مصنف ELM أدى أفضل الأداء من حيث الدقة ومعدل الإنذارات الكاذبة وغيرها من المقاييس الأخرى التي تم استخدامها من أجل تصنيف الهجمات والحفاظ على أمن الشبكات.

في دراسة [11] تم تحديد الميزات المهمة بإستخدام مربع كاي Chi-Square وكان عددها 23 ميزة وكانت دقة خوارزمية الغابة العشوائية RF في التصنيف الثنائي مع 23 ميزة 99.64% وكانت دقة هذه الخوارزمية مع كل الميزات 99.57% السبب في ارتفاع الدقة بعد تقليل عدد الميزات يرجع إلى التقنية المستخدمة في اختيار الميزات وهي مربع كاي Chi-Square الذي يعمل على تقليل التباين والضوضاء في البيانات مما يساعد على تحسين الدقة في معظم الخوارزميات.

في دراسة [14] تم اختيار الميزات من مجموعة التدريب فقط بإستخدام طريقة التغليف Wrapper Method وكان عدد الميزات التي تم اختيارها 20 ميزة بعد ذلك تم تسوية البيانات واجراء التشفير السريع One-Hot Encoding وكانت دقة خوارزمية الغابة العشوائية RF في التصنيف الثنائي مع كل الميزات التي عددها 41 ميزة هي 95.74% ومع اختيار الميزات كانت دقة هذه الخوارزمية 98.51% لكنها استغرقت وقت تنفيذ طويل والسبب يرجع إلى أن طريقة التغليف بطيئة بسبب التكرار بالإضافة إلى أنها عملية مكلفة حسابيًا لكنها تحقق دقة عالية عند استخدامها. كما أشارت هذه الدراسة إلى أن مجموعة بيانات UNSW-NB15 أكثر حداثة وموثوقية لبناء نماذج كشف التسلل مقارنة بمجموعات البيانات القديمة الأخرى.

وعلى الرغم من أن كل هذه الدراسات السابقة استخدمت مجموعة بيانات UNSW-NB15 إلّا انها اختلفت في تقنيات اختيار الميزات وخوارزميات التعلم الآلي المستخدمة وكذلك عدد سجلات التدريب والاختبار، لذلك لا يمكن القول بأن هناك دراسة أفضل من الأخرى لأن جميع النتائج التي تم عرضها والتقنيات التي تم استخدامها فعالة ومثالية لكشف وتصنيف الهجمات في شبكات إنترنت الأشياء. والجدول 9 يوضح مقارنة شاملة بين هذه الدراسات السابقة.

جدول 9: مقارنة شاملة بين الدراسات السابقة.

المؤلفين والسنة	مجموعة البيانات ولغة البرمجة والاداة المستخدمة.	ML	نوع التصنيف		نسبة التقسيم	تقنية اختيار الميزة وعدد الميزات المختارة	الخوارزمية المستخدمة	النتائج المتحصل عليها			
			ثنائي	متعدد							
Jing & Chen 2019 [32]	مجموعة البيانات UNSW-NB15 لغة البرمجة Java	✓	✓	✓	68% تدريب. 32% اختبار.	لم يتم اختيار الميزات.	SVM	نتائج التصنيف الثنائي لخوارزمية Support Vector Machine		نتائج التصنيف المتعدد لخوارزمية Support Vector Machine	
								85.99		75.77	
								15.26		-	
								16.50		3.04	
								-		50.90	
Bagui & et al 2019 [33]	مجموعة البيانات UNSW-NB15 الأداة Weka	✓		✓	-	التجميع K-Means Clustering والارتباط Correlation، تم استخدام التقنيتين كتقنية مختلطة وعدد الميزات المختارة هي 29 ميزة.	NB J48	من دون اختيار الميزات		مع اختيار الميزات	
								NB	J48	NB	J48
								69.32	98.01	82.07	96.68
								92.78	46.18	75.57	21.56
								26.89	1.98	17.92	3.31
Alrashdi & et al 2019 [34]	مجموعة البيانات UNSW-NB15 لغة البرمجة Python	✓	✓		-	مصنف الأشجار الإضافية Extra Trees Classifier عدد الميزات المختارة 12 ميزة.	RF	نتائج خوارزمية RF باستخدام 12 ميزة			
								98			
								98			
								98			
Hammad & et al 2020 [35]	مجموعة البيانات UNSW-NB15	✓	✓		90% تدريب. 10% اختبار.	على أساس الارتباط Correlation-based Feature Selection (CFS) وعدد الميزات المختارة 8 ميزات.	NB RF J48 ZR	باستخدام تقنية التحقق المتقاطع وتقنية اختيار الميزات على أساس الارتباط			
								NB	J48	RF	ZR
								76.04	93.78	97.60	68.06
								14.9	12.9	3	68.1
								76	93.8	97.6	63.1
								83.4	94.2	97.6	68.1
								76.8	93.6	97.6	68.1

*Corresponding author:
E-mail addresses: f.abdalnaby@wau.edu.ly ,(G. M. Miskeen) guz.miskeen@wau.edu.ly
Article History : Received 11 June 2024 - Received in revised form 28 September 2024 - Accepted 06 October 2024

	نتائج التصنيف المتعدد لخوارزمية ELM مع استخدام تقنية اختيار الميزة					ELM	مصنف الأشجار الإضافية Extra Trees Classifier حسب معيار Gini وعدد الميزات المختارة 14 ميزة.	80% تدريب. 20% اختبار.	✓		✓	مجموعة البيانات UNSW-NB15 لغة البرمجة Python	Moualla & et al 2021 [18]
Accuracy (%)	98.19												
TPR (%)	94.79												
FPR (%)	0.216												
F1-Score (%)	96.08												
	LR	NB	RF	SGD	KNN	KNN SGD RF LR NB	مربع كاي Chi-Square، عدد الميزات المهمة التي تم اختيارها 23 ميزة.	80% تدريب. 20% اختبار.		✓	✓	مجموعة البيانات UNSW-NB15 لغة البرمجة Python	Kocher & Kumar 2021 [11]
	مع كل الميزات وبعد اختيار الميزات على التوالي												
Accuracy (%)	98.42 98.17	76.59 75.16	99.57 99.64	98.16 97.99	98.28 98.90								
Precision (%)	98 98	100 100	100 100	98 98	99 99								
Recall (%)	99 100	69 67	100 100	100 100	99 99								
F1-Score (%)	99 99	82 80	100 100	99 99	99 99								
MSE (%)	1.5 1.8	23.4 24.8	0.4 0.3	1.8 2	1.7 1.1								
TPR (%)	99.4 99.6	69.3 67.4	99.7 99.8	98.3 99.7	98.9 99.3								
FPR (%)	4.8 6.3	0.7 2.5	0.9 0.9	4.3 7.3	3.9 2.5								
	مجموعة بيانات UNSW-NB15 مع كل الميزات وبعد اختيار الميزات على التوالي					SVM ANN K-NN RF NB	شجرة القرار Decision Tree التي تتبع طريقة التغليف Wrapper Method، ومن مجموعة بيانات UNSW-NB15 تم اختيار 20 ميزة.	67% تدريب. 33% اختبار.		✓	✓	مجموعتين بيانات وهما: UNSW-NB15 NSL-KDD الأداة Jupyter Notebook لغة البرمجة بايثون Python	Umar & Zhanfang 2023 [14]
	ANN	SVM	KNN	RF	NB								
Accuracy (%)	94.62 94.32	93.67 93.56	93.81 95.8	95.74 98.51	48.08 48.11								
DR (%)	97.54 98.48	99.63 99.54	96.24 97.28	97.84 99.17	23.91 23.76								
FAR (%)	11.64 14.56	19.14 19.95	11.42 7.36	8.77 2.89	0.02 0.01								
Time	5.81 5.43 (m)	86.89 170.6 (m)	8.38 12.46 (m)	0.54 0.56 (m)	2.89 4.8 (s)								

*Corresponding author:
E-mail addresses: f.abdalnaby@wau.edu.ly ,(G. M. Miskeen) guz.miskeen@wau.edu.ly
Article History : Received 11 June 2024 - Received in revised form 28 September 2024 - Accepted 06 October 2024

ix. الخلاصة

قدمنا في هذه الدراسة نظرة عامة على تقنيات التعلم الآلي الخاضعة للإشراف لأنظمة كشف التسلل بالشبكات (NIDS) ومنهجيات الكشف المتميزة بالإضافة إلى المصنفات الخاصة بأنظمة كشف التسلل (IDS). كذلك تمت المقارنة بين مجموعات البيانات المعيارية لأنظمة كشف التسلل وإبراز أحدث مجموعة البيانات وهي UNSW-NB15 وتم استخدامها في الدراسات السابقة التي تم عرضها في هذه الورقة وذلك لمزاياها المذكورة، وتبين أن مجموعات البيانات مثل KDDCUP'99 وNSL-KDD لا تعطي تمثيلًا شاملًا للتوجه الحديث لسيناريوهات حركة مرور الشبكة والهجوم. تعتبر UNSW-NB15مجموعة بيانات مرجعية حديثة لـ NIDS مقارنة بمجموعات البيانات القديمة التي تظهر عددًا محدودًا من الهجمات ومعلومات الحزم التي لم يعد استخدامها فعالًا نظرًا لعدم حداثتها. وبعد ذلك تم التطرق إلى أهم وأحدث الأبحاث والدراسات السابقة وماهي الأساليب التي تمت تجربتها. وتوصي الدراسة إلى استخدام المزيد من مجموعات البيانات الخاصة بكشف التسلل إلى شبكات انترنت الأشياء مثل مجموعة بيانات CIC-IDS2017 وغيرها من مجموعات البيانات الحديثة الأخرى.

قائمة المراجع

[1]- M. M. Rashid, J. Kamruzzaman, M. M. Hassan, T. Imam, and S. Gordon, "Cyberattacks detection in iot-based smart city applications using machine learning techniques," *International Journal of environmental research and public health*, vol. 17, no. 24, pp. 9347, 2020.

[2]- Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, pp. e4150, 2021.

[3]- M. Sarhan, S. Layeghy, and M. Portmann, "Feature analysis for machine learning-based IoT intrusion detection," *arXiv preprint arXiv:2108.12732*, 2021.

[4]- R. Alshamy, and M. Ghurab, "A review of big data in network intrusion detection system: Challenges, approaches, datasets, and tools," *Journal of Computer Sciences and Engineering*, vol. 8, no. 7, pp. 62-74, 2020.

[5]- M. Leon, T. Markovic, and S. Punnekkat, "Comparative Evaluation of Machine Learning Algorithms for Network Intrusion Detection and Attack Classification," in 2022 international joint conference on neural networks (IJCNN), 2022, pp. 01-08.

[6]- M. G. Solomon, and M. Chapple, *Information security illuminated*: Jones & Bartlett Publishers, 2004.

[7]- S. Sriram, K. Simran, R. Vinayakumar, S. Akarsh, and K. Soman, "Towards evaluating the robustness of deep intrusion detection models in adversarial environment," in International Symposium on Security in Computing and Communication, 2019, pp. 111-120.

[8]- E. Alpaydin, *Introduction to machine learning*: MIT press, 2020.

[9]- د. ع. طعيمة, تعلم الآلة وعلم البيانات : الأساسيات والمفاهيم والخوارزميات والادوات.2022 ,p. pp. 465 ,

[10]- A. Sugandhi. "Feature Engineering for Machine Learning," Sep 5th, 2023.[online]. Available: https://www.knowledgehut.com/blog/data-science/feature-engineering-for-machine-learning. [Accessed:Oct 3rd 2023].

[11]- G. Kocher, and G. Kumar, "Analysis of machine learning algorithms with feature selection for intrusion detection using UNSW-NB15 dataset," *Available at SSRN 3784406*, 2021.

[12]- N. Kaur, M. Bansal, and S. S. Sran, "Scrutinizing attacks and evaluating performance appraisal parameters via feature selection in intrusion detection system," 2021.

[13]- M. A. Arif. "Confusion Matrices and Classification Reports: A Guide to Evaluating Machine Learning Models," Apr 3th, 2023.[online]. Available: https://smuhabdullah.medium.com/confusion-matrices-and-classification-reports-a-guide-to-evaluating-machine-learning-models-385496cf7cee. [Accessed: Feb20th,2024].

[14]- M. A. Umar, and C. Zhanfang, "Effects of Feature Selection and Normalization on Network Intrusion Detection," *Authorea Preprints*, 2023.

[15]- M. A. Siddiqi, and W. Pak, "Optimizing filter-based feature selection method flow for intrusion detection system," *Electronics*, vol. 9, no. 12, pp. 2114, 2020.

[16]- S. Yemulwar. "Feature Selection Techniques," Sep 27th, 2019.[online]. Available: https://medium.com/analytics-vidhya/feature-selection-techniques-2614b3b7efcd. [Accessed: Jan 8th,2024].

[17]- P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3-42, 2006.

[18]- S. Moualla, K. Khorzom, and A. Jafar, "Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1-13, 2021.

[19]- G. Chandrashekar, and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.

[20]- A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *Ieee Access*, vol. 8, pp. 165130-165150, 2020.

[21]- D. D. Protić, "Review of KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets," *Vojnotehnički glasnik/Military Technical Courier*, vol. 66, no. 3, pp. 580-596, 2018.

[22]- M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE symposium on computational intelligence for security and defense applications, 2009, pp. 1-6.

[23]- E. Hassan, M. Saleh, and A. Ahmed, "Network intrusion detection approach using machine learning based on decision tree algorithm," *Journal of Engineering and Applied Sciences*, vol. 7, no. 2, pp. 1, 2020.

[24]- M. Latah, and L. Toker, "Towards an efficient anomaly-based intrusion detection for software-defined networks," *IET networks*, vol. 7, no. 6, pp. 453-459, 2018.

[25]- M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy, and S. Othman, "A detailed analysis of benchmark datasets for network intrusion detection system," *Asian Journal of Research in Computer Science*, vol. 7, no. 4, pp. 14-33, 2021.

[26]- S. Meftah, T. Rachidi, and N. Assem, "Network based intrusion detection using the UNSW-NB15 dataset," *International Journal of Computing and Digital Systems*, vol. 8, no. 5, pp. 478-487, 2019.

[27]- N. Moustafa, and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in 2015 military communications and information systems conference (MilCIS), 2015, pp. 1-6.

[28]- N. Elmrabit, F. Zhou, F. Li, and H. Zhou, "Evaluation of machine learning algorithms for anomaly detection," in 2020 international conference on cyber security and protection of digital services (cyber security), 2020, pp. 1-8.

[29]- M. Rodríguez, Á. Alesanco, L. Mehavilla, and J. García, "Evaluation of Machine Learning Techniques for Traffic Flow-Based Intrusion Detection," *Sensors*, vol. 22, no. 23, pp. 9326, 2022.

[30]- A. Thakkar, and R. Lohiya, "A review of the advancement in intrusion detection datasets," *Procedia Computer Science*, vol. 167, pp. 636-645, 2020.

[31]- UNSW-NB15 Dataset.[online]. Available: https://github.com/abhinav-bhardwaj/IoT-Network-Intrusion-Detection-System-UNSW-NB15/tree/master/datasets. [Accessed: Jan 7th, 2024].

[32]- D. Jing, and H.-B. Chen, "SVM based network intrusion detection for the UNSW-NB15 dataset," in 2019 IEEE 13th international conference on ASIC (ASICON), 2019, pp. 1-4.

[33]- S. Bagui, E. Kalaimannan, S. Bagui, D. Nandi, and A. Pinto, "Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset," *Security and Privacy*, vol. 2, no. 6, pp. e91, 2019.

[34]- I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy, and H. Ming, "Ad-iot: Anomaly detection of iot cyberattacks in smart city using machine learning," in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0305-0310.

[35]- M. Hammad, W. El-Medany, and Y. Ismail, "Intrusion detection system using feature selection with clustering and classification machine learning algorithms on the unsw-nb15 dataset," in 2020 international conference on innovation and intelligence for informatics, computing and technologies (3ICT), 2020, pp. 1-6.