A Survey Current Datasets used for Intrusion Detection Using Machine Learning

Dr. Mohammed Hejazieh *

Ammar Moustafa **

(Received 29 / 6 / 2022. Accepted 18 / 9 / 2022)

□ ABSTRACT □

Cyberattacks in today's digital age cause the loss of sensitive data and a huge financial loss for enterprises and countries. Therefore, the role of the cyber security expert is very important to protect data from increased and new attacks. Researchers focus on anomaly-based intrusion detection systems to detect these unknown attacks and machine learning algorithms play a vital role in this process because they detect attacks accurately. Data sets currently used in intrusion detection systems suffer from a clear lack of real network threats, attack representation, and include a large number of abandoned threats, which limit the accuracy of detection within the current intrusion detection systems' methods of machine learning, which make them unable to trace increasing and new attacks in cloud environments, containers. This research paper aims to combine classification and analysis of existing data sets in order to improve the creation new data sets that simulate the actual reality of the network's real data. This will improve the efficiency of the next generation of intrusion detection systems and reflect network threats more accurately.

Keywords: Intrusion Detection system, Machine Learning, Cyber Attacks, UNSW-NB15 Data Set, CICIDS2018 Data Set, CICIDS2017 Data Set, DARPA Data Set, NSL-KDD Data Set, KDD'99Data Set, ADFA-IDS Data Set

* **Professor, Computers and Automated Control, Faculty of Mechanical & Electrical Engineering, Tishreen University, Lattakia, Syria.**

** **PhD Student, Computers and Automated Control Department, Faculty Mechanical & Electrical Engineering, Tishreen University, Lattakia, Syria, ammar.c.moustafa@tishreen.edu.sy**

دراسة استقصائية لمجموعات البيانات المستخدمة حاليا ضمن أنظمة كشف الاقتحام المستندة الى تقنيات تعلم الآلة

د. محمد حجازية *

عمار مصطفى **

(تاريخ الإيداع 29 / 6 / 2022. قُبِل للنشر في 18/ 9 / 2022)

□ ملخّص □

تتسبب الهجمات الإلكترونية في العصر الرقمي الراهن في فقدان البيانات الحساسة وخسارة مالية فادحة للمؤسسات والدول. لذلك، فإن دور خبير الأمن السيبراني مهم جدًا لحماية البيانات من الهجمات المتزايدة والجديدة. يركز الباحثون على نظم اكتشاف الاقتحام القائم على الشذوذ لاكتشاف تلك الهجمات الغير المعروفة وتلعب خوارزميات التعلم الآلي دورًا حيويًا في هذه العملية لأنها تكتشف الهجمات بدقة. تعاني مجموعات البيانات المستخدمة حاليا في أنظمة كشف الاقتحام نقصًا واضحًا في تهديدات الشبكة الحقيقية، وتمثيل الهجوم، وتتضمن عددًا كبيرًا من التهديدات المهجورة، والتي تحد من دقة الاكتشاف ضمن مناهج أنظمة كشف الاقتحام الحالية لتعلم الآلة، مما يجعلها غير قادرة على مواكبة الهجمات المتزايدة والجديدة في البيئات السحابية والحاويات البرمجية. تهدف هذه الورقة البحثية الى الجمع بين التصنيف وتحليل مجموعات البيانات الحالية من أجل تحسين إنشاء مجموعات بيانات جديدة مستقبلية تحاكي الواقع الفعلي لبيانات الشبكة الحقيقية. مما سيؤدي ذلك إلى تحسين كفاءة الجيل القادم من أنظمة كشف الاقتحام ويعكس تهديدات الشبكة بشكل أكثر دقة.

الكلمات المفتاحية :أنظمة كشف الاقتحام، تعلم الآلة، الهجوم السيبراني، مجموعة البيانات UNSW-NB15، مجموعة البيانات CICIDS2018، مجموعة البيانات CICIDS2017، مجموعة البيانات DARPA، مجموعة البيانات UNSW-NB15، مجموعة البيانات NSL-KDD، مجموعة البيانات KDD'99، مجموعة البيانات ADFA-IDS

* **أستاذ، قسم هندسة الحاسبات والتحكم الالي، كلية الهندسة الكهربائية والميكانيكية، جامعة تشرين، اللاذقية، سورية.**

** **طالب دكتوراه، قسم هندسة الحاسبات والتحكم الالي، كلية الهندسة الكهربائية والميكانيكية، جامعة تشرين، اللاذقية، سورية.**

بريد الكتروني : ammar.c.moustafa@tishreen.edu.sy

مقدمة:

اعتمد الباحثون على مجموعات البيانات المعيارية لتقييم نتائجهم عند استخدام أنظمة كشف الاقتحام ضد الهجمات، ومع ذلك تفتقر مجموعات البيانات المتاحة حاليًا إلى الخصائص الفعلية لحركة مرور ضمن الشبكة، وهذا السبب جعل معظم أنظمة كشف الاقتحام القائمة على الشذوذ غير قابلة للتطبيق في بيئات العمل الحالية [25]. علاوة على ذلك، فإن أنظمة كشف الاقتحام غير قادرة على التكيف مع التغيرات المستمرة التي تطرأ على الشبكات (أي العقد الجديدة، تغيير في أحمال المرور، تغيير الطوبولوجيا، وما إلى ذلك). إن هذه التغيرات، تجعل الاعتماد على مجموعات البيانات القديمة فقط لا يساعد في تطوير أنظمة كشف الاقتحام. يجب أن تأخذ عملية إنشاء مجموعات بيانات جديدة في الاعتبار حقيقة التغيير المستمر هذه. على سبيل المثال، اقتراح توليد مجموعة بيانات قياسية بوظائف قابلة للتمديد، من شأنه أن يزيل عبء إنشاء مجموعات بيانات من الصفر. يمكن أن تكون مجموعات البيانات إما حقيقية (أي مسجلة من إعدادات الشبكة) أو اصطناعية (أي محاكاة حركة المرور أو حقنها). يمكن استخدام حقن الهجوم الاصطناعي إما لإدخال هجمات على مجموعة بيانات موجودة أو موازنة فئات الهجوم الموجودة في مجموعة البيانات.

1- الخصائص التي يجب أن تتمتع بها مجموعات البيانات:

ذكر الباحث في [25] أنه لكي يتم التعامل مع مجموعة بيانات، يجب أن تغطي المجموعة الأمور التالية:

- تكوين الشبكة Network Configuration : يشير إلى أن لديه معرفة كاملة بطوبولوجيا الشبكة حول كيفية توصيل أجهزة الشبكات في بيئة الاختبار بحيث يتم التقاط سيناريوهات الهجوم الواقعية
- حركة مرور الشبكة Network Traffic : يشير إلى التقاط جميع حزم الشبكة من المضيف والوجهة وجدار الحماية وتطبيقات الويب لتحليل التدفق وتوليد مجموعة البيانات.
- مجموعة البيانات الموسومة Labeled Dataset: يشير إلى وضع علامات على حالات البيانات التي تم التقاطها من حركة مرور الشبكة للحصول على فهم كامل لتفاعل الشبكة.
- تفاعل الشبكة Network Interaction : يشير إلى وجود السجل الكامل للاتصالات الشبكية داخل وخارج الشبكة.
- التقاط حركة المرور Capturing the Traffic: يشير إلى التقاط حركة مرور الشبكة الوظيفية وغير الوظيفية لقياس DR و FPR من IDS
- البروتوكولات Protocols: يجب أن تتضمن مجموعة البيانات المثالية جميع الاتصالات باستخدام بروتوكولات مختلفة سواء كانت طبيعية أو ضارة
- الهجمات Attacks : يجب أن تتكون مجموعة البيانات من فئات هجمات واسعة النطاق ومحدثة
- المزايا Features : يجب أن تحتفظ مجموعة البيانات بمجموعة كاملة من الميزات المحددة جيدًا لتصنيف الهجوم
- عدم التجانس Heterogeneity : يجب جمع مجموعة البيانات من مصادر مختلفة لتغطية جميع تفاصيل الإجراء المتبع للكشف عن الهجمات
- بيانات وصفية Metadata : يجب أن تحتوي مجموعة البيانات على وثائق مناسبة تصف بيئة الاختبار والبنية التحتية لنظام الهجوم والبنية التحتية لنظام الضحايا والسيناريوهات المتبعة في الهجوم

يلخص الجدول (1) مجموعات البيانات المتاحة وقد تم تصنيفها على أساس المجال الذي تنتمي إليه. علاوة على ذلك، يتم عرض الهجمات الموجودة في كل منها.

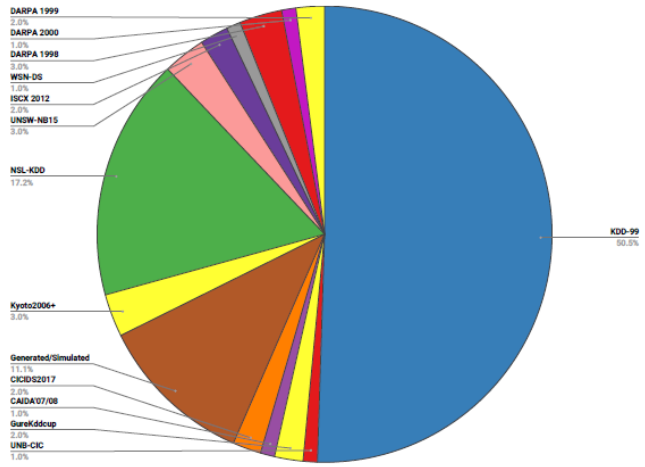
الجدول (1): مجموعات البيانات والهجمات المحتواة ضمنها

Year	DataSet	Normal	DoS	DDoS	Probe	U2R	R2L	Infiltrating/Scanning	SSH	Heartbleed	Brute Force	XSS	Sql Injection	Botnet	Network and Host Events	Port Scan	Meterpreter
2018	CICIDS2018	√	√	√				√	√	√	√	√	√	√		√	
2017	CICIDS2017	√	√	√				√	√	√	√	√	√	√		√	
2017	CIC DoS dataset	√	√														
2013 2017	ADFA-IDS	√							√								√
2017	Unified Network Dataset	√													√		
2016	DDoSTB	√		√													
2015	Booters			√													
2015	TUIDS Intrusion	√	√		√												
2014	Botnet dataset	√												√			
2012	STA2018	√	√	√				√	√								
2010	ISCXIDS2012	√	√	√				√	√					√			
2007	CAIDA DDoS			√													
1999	NSL-KDD	√	√		√	√	√										
1999	KDD'99	√	√		√	√	√										
1998 1999 2000	DARPA	√	√		√	√	√										

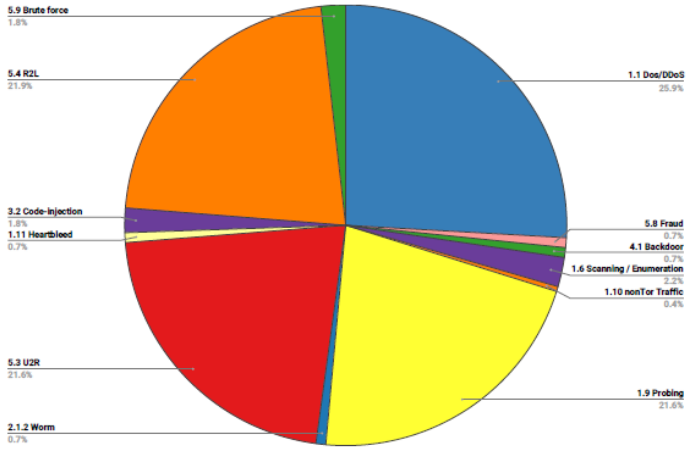
يبين الشكل (1) توزع مجموعات البيانات المستخدمة في البحوث التي تمت في العقد الماضي. و لوحظ بنتائج تلك البحوث استخدام 11٪ فقط من أنظمة كشف الاقتحام المذكورة مجموعات بيانات مولدة أو محاكاة.

من الواضح أيضًا من خلال هذا التحليل أن معظم مجموعات البيانات تفتقر إلى خصائص الحياة الواقعية. ويُظهر الشكل أيضا استخدام KDD-99 كمجموعة بيانات مفضلة. قدم الباحثين في[3] و [5] تحليلاً شاملاً لعيوب مجموعة بيانات KDD "99. كذلك، يُظهر الجدول الزمني المقدم كلاً من نقاط النقد المختلفة وتحذير مختبر UCI بعدم استخدام مجموعة بيانات KDD Cup 99، مما يؤكد بشكل أكبر عيوب استخدام KDD Cup 99 في بحوث أنظمة كشف الاقتحام IDS الحالية. مجموعة البيانات الثانية الأكثر استخدامًا هي مجموعة بيانات DARPA [2] ، حيث فشلت في تمثيل الهجمات الحالية بدقة بسبب قدمها. علاوة على ذلك، تم الإبلاغ عن العديد من النتائج في عملية الكشف والتي لا لا يمكن تطبيقها في سيناريوهات العالم الحقيقي. يعرض الشكل (2) الهجمات التي اكتشفتها مختلف أنظمة كشف الاقتحام. ويتضح أن الهجمات الأربع المتاحة في مجموعة البيانات KDD-99 هي الأكثر تغطية، وهي على وجه التحديد DoS/DDoS، Probing، R2L، U2R. علاوة على ذلك هناك 12 هجوماً فقط مدرجة في الشكل (2) الذي

يُسلط الضوء على القيود المحتملة لهذه الأنظمة لمواجهة مجموعة واسعة من الهجمات وهجمات zero-day. لمعالجة اكتشاف هجمات zero-day ، هناك حاجة لبناء مجموعات بيانات قابلة للتمديد يمكن استخدامها لتدريب نماذج التعلم الآلي المستخدمة للكشف عن الشذوذ



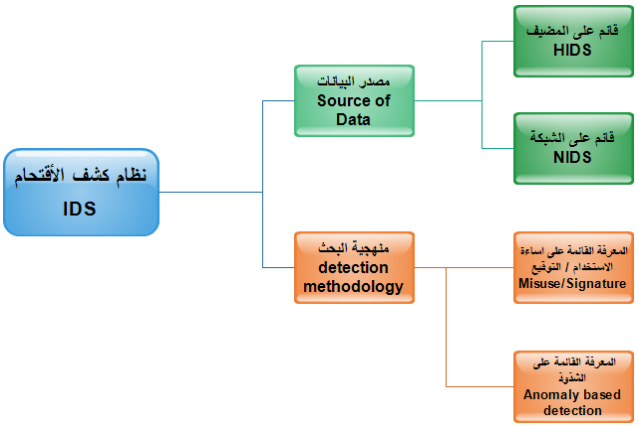
الشكل (1): توزع مجموعات البيانات المستخدمة في تقييم أنظمة كشف الاقتحام [33]



الشكل (2): النسبة المئوية للهجمات التي تم اكتشافها من خلال مختلف أنظمة كشف الاقتحام [33]

2- نظام كشف الاقتحام Intrusion Detection

يتم تصنيف نظام الكشف عن الاقتحام بناءً على مصدر البيانات ومنهجية الكشف. ويبين الشكل (3) تصنيف الكشف عن الاقتحام. ويوجد نوعين من هذه الأنظمة: النوع الأول القائم على المضيف HIDS إذا تتم مراقبة التسلل والاقتحام على المضيف أو الأجهزة على الشبكة، النوع الثاني: القائم على الشبكة NIDS، يتم رصد الاقتحام على الشبكة بأكملها. واستنادا إلى منهجية الكشف، يصنف نظام كشف الاقتحام إلى نهج قائم على التوقيع Signature وإساءة الاستخدام Misuse ، في نهج الكشف عن إساءة الاستخدام، يتم مطابقة السلوك غير الطبيعي للشبكة مع الأنماط المعروفة للهجمات المكتشفة. يحدد الاكتشاف القائم على الشذوذ حركة مرور الشبكة غير العادية.



الشكل (3) : تصنيف الكشف عن الاقتحام

3-1- أنظمة كشف الاقتحام القائمة على الشذوذ والمستندة على تقنيات تعلم الآلة:

التعلم الآلي هو عملية استخلاص المعرفة من كميات كبيرة من البيانات. تتكون نماذج تعلم الآلة من مجموعة من القواعد أو الأساليب أو «وظائف النقل» المعقدة التي يمكن تطبيقها للعثور على أنماط بيانات جديدة تكون مصدر للاهتمام، أو للتعرف على سلوك ما أو التنبؤ به. تم تطبيق تقنيات التعلم الآلي على نطاق واسع في مجال أنظمة كشف الاقتحام القائمة على الشذوذ AIDS عبر العديد من الخوارزميات والتقنيات مثل العنقدة Clustering والشبكات العصبية Neural Networks وقواعد الارتباط وأشجار القرار Decision Trees والخوارزميات الجينية Genetic والجار الاقرب Nearest Neighbor لاكتشاف المعرفة من مجموعات بيانات كشف الاقتحام

3- مناقشة حول مجموعات البيانات المستخدمة في بحوث الكشف عن الاقتحام:

استخدم معظم الباحثين مجموعات البيانات DARPA ، KDD Cup 1999 ، NSL-KDD ، UNSW-NB15 وكيوتو و CSCIDS 2017 للكشف عن الهجمات . تحتوي مجموعات البيانات المستخدمة للكشف عن الهجمات من قبل الباحثين على بيانات تدريبية Training Data وبيانات اختبار Testing Data.

3-1- مجموعة التدريب KDD Cup1999

تم جمع مجموعة بيانات KDD Cup1999 بناءً على برنامج وكالة مشاريع البحوث الدفاعية المتقدمة DARPA للكشف عن الاقتحام و هي وكالة بحث وتطوير تابعة للولايات المتحدة ، والذي يمكن الوصول إليها عبر MIT Lincoln Lab، تم بدء العمل بـ 5000 سجل لمجموعة البيانات الإجمالية، بعد ذلك 25000 سجل حتى بلغ عدد السجلات 300000 سجل.

تتناول مجموعة بيانات KDD Cup1999 متطلبات الاستخدام المناسب للبيانات من أجل تقييم نظام كشف الاقتحام. تم بناء مجموعة البيانات هذه كمجموعة بيانات محاكاة في عام 1998 ومنذ ذلك الحين تم استخدامها بشكل مكثف في مجالات التنقيب عن البيانات وتعلم الآلة [6] .

تحتوي مجموعة بيانات KDD Cup1999 على بيانات التدريب والاختبار وتحتوي على 41 ميزة مصنفة إلى ميزات أساسية وحركة المرور والمحتوى [24]. وعلى ما يقرب من خمسة ملايين بيانات أولية، تمثل بيانات الهجوم منها حوالي 80٪. وتصنف هذه البيانات إلى فئة واحدة «طبيعية» وأربع فئات «هجوم» وهي حجب الخدمة (DOS)، المستخدم المحلي إلى المستخدم الجذر (U2R)، التحقيق في الهجمات (Probe)، والمستخدم البعيد إلى المستخدم المحلي (R2L). تم تضمين ما مجموعه 22 نوعًا من الهجمات في مجموعة بيانات KDD Cup1999 ، مع كون كل هجوم جزءًا من

الفئات المذكورة أعلاه. كما تم الكشف عن العيوب المتأصلة في مجموعة بيانات KDD Cup1999 من خلال العديد من التحليلات الإحصائية التي أثرت على دقة اكتشافه للعديد من IDS من قبل الباحثين، المشكلة الأكثر أهمية في مجموعة بيانات KDD هي أنه يحتوي على عدد كبير من السجلات المتماثلة .حيث أظهرت النتائج أن حوالي 78٪ سجلات مكررة في مجموعة بيانات التدريب و75٪ سجلات اختبار، قد يؤدي العدد الكبير من السجلات المتكررة إلى أن تكون خوارزميات التعلم جزئية أي أن الخوارزمية ستتوقف عن تعلم السجلات غير المتكررة. قد تكون هذه السجلات ضارة بالشبكة مثل U2R، R2L إلخ.

3-2- مجموعة التدريب NSL-KDD

تحتوي مجموعة بيانات NSL-KDD على أهم سجلات مجموعة بيانات KDD Cup1999 وتصنف خصائص بياناتها في عدة مجموعات [7]. لا تتضمن مجموعة بيانات NSL-KDD سجلات غير ذات صلة في مجموعة التدريب وبالتالي تقلل من حجم البيانات بحذف السجلات المكررة، لذلك لن تكون المصنفات جزئية تجاه المزيد من السجلات المتكررة وبالتالي تحسين وظيفة خوارزميات تعلم الآلة. تتمتع مجموعة بيانات NSL-KDD بالمزايا التالية مقارنة بمجموعة بيانات KDD الأصلية [17] :

- لا يتضمن سجلات زائدة عن الحاجة في مجموعة التدريب، لذلك لن ينحاز المصنفون إلى سجلات أكثر تكرارًا.
- لا توجد سجلات مكررة في مجموعات الاختبارات المقترحة، وبالتالي، فإن أداء المصنفات لا ينحاز إلى الأساليب التي لديها معدلات كشف أفضل في السجلات المتكررة.
- عدد السجلات في مجموعات التدريب والاختبار مقبول، الامر الذي يجعل من الممكن إجراء التجارب على المجموعة الكاملة دون الحاجة إلى اختيار جزء صغير بشكل عشوائي. وبالتالي، ستكون نتائج تقييم مختلف الأعمال البحثية متسقة وقابلة للمقارنة.

تحتوي كل حالة في مجموعة التدريب على جلسة اتصال واحدة مقسمة إلى أربع مجموعات، مثل الميزات الأساسية من اتصال الشبكة والميزات المتعلقة بالمحتوى والميزات المتعلقة بالزمن وميزات حركة المرور المستندة إلى المضيف. يتم وسم كل حالة إما بأنها عادية أو هجوم [26] . تم تجميع فئات الهجوم الموجودة في مجموعة بيانات NSL-KDD في أربع فئات كما هو الحال في مجموعة البيانات KDD Cup1999

كشف المزيد من التحليل لمجموعة بيانات التدريب KDDTrain + عن إحدى الحقائق المهمة جدًا حول متجهات شبكة فئة الهجوم كما هو موضح في الجدول (2). حيث أن معظم الهجمات التي يقوم بها المهاجم تستخدم مكدس بروتوكول TCPعن طريق استغلال شفافية وسهولة استخدام بروتوكول TCP من قبل المهاجمين لشن هجمات قائمة على الشبكة وعلى أجهزة الحاسب للضحايا [25]

الجدول (2): البروتوكولات المستخدمة من قبل فئة هجومات مختلفة

فئات الهجوم				
U2R	R2L	PROBE	DOS	البروتوكول
49	995	5857	42188	TCP
3	0	1664	892	UDP
0	0	4135	2847	ICMP

الجدول (3): إحصاءات السجلات الزائدة عن الحاجة في مجموعة التدريب KDD Cup1999

	السجلات الأصلية	سجلات متميزة	معدل التخفيض
هجوم	3,925,650	262,178	93.32%
بيانات طبيعية	972,781	812,814	16.44%
العدد الإجمالي	4,898,431	1,074,992	78.05%

الجدول (4): إحصاءات السجلات الزائدة عن الحاجة في مجموعة اختبار KDD Cup1999

	السجلات الأصلية	سجلات متميزة	معدل التخفيض
هجوم	250,436	29,378	88.26%
بيانات طبيعية	60,591	47,911	20.92%
العدد الإجمالي	311,027	77,289	75.15%

3-3- مجموعة التدريب UNSW-NB15

تم إصدار مجموعة البيانات UNSW-NB15 في عام 2015 بواسطة أداة IXIA PerfectStorm في مختبر Cyber Range في جامعة نيو ساوث ويلز كانبيرا لتوليد مزيج من الأنشطة الطبيعية الحديثة الحقيقية وسلوكيات الهجوم الاصطناعية المعاصرة [3]. تتكون مجموعة البيانات هذه من أنشطة شبكة حديثة بلغت 2,540,044 سجل (طبيعية وغير طبيعية). تم جمع هذه السجلات بواسطة مولد حركة بيانات IXIA وباستخدام ثلاثة خوادم افتراضية. تم تكوين خادمين لتوزيع حركة المرور العادية للشبكة والثالث تم تكوينه لتوليد حركة مرور غير طبيعية للشبكة.

تم استخراج 49 ميزة مع أصناف موسومة بما في ذلك الميزات القائمة على الحزم والقائمة على التدفق من حزم الشبكة المستخرجة بواسطة أدوات Argus و Bro-IDS. يتم استخراج الميزات القائمة على الحزمة من رأس الحزمة وحمولتها (تسمى أيضًا بيانات الحزمة). في المقابل، يتم إنشاء الميزات القائمة على التدفق باستخدام تسلسل الحزم، من المصدر إلى الوجهة، تعد الوجهة وطول الحزم وأزمنة الوصول بين الحزم من أهم الخصائص في تركيبة الميزات القائمة على التدفق، وتصنف الميزات إلى ثلاث مجموعات، وهي الأساسية (6 إلى 18) والمحتوى (19 إلى 26) والزمن (27 إلى 35). بينما تم تصنيف الميزات من 36 إلى 40 ومن 41 إلى 47 على أنها ميزات عامة الغرض وميزات اتصال، على التوالي.تشمل مجموعة البيانات UNSW-NB15 على تسعة أنواع هجوم مقارنة بـ 14 نوعًا من الهجوم في مجموعة بيانات KDD 99"، هناك 221.876 سجل يحمل سمة طبيعي و 321.283 سجل يحمل سمة الهجوم في العدد الإجمالي للسجلات.

الجدول (5): عدد سجلات التدريب والاختبار لكل صنف ضمن مجموعة التدريب UNSW-NB15

الصنف	سجلات التدريب	سجلات الاختبار
Normal	56,000	37,000
Analysis	2,000	677
Backdoor	1,746	583
DoS	12,264	4,089
Exploits	33,393	11,132
Fuzzers	18,184	6,062
Generic	40,000	18,871
Reconnaissance	10,491	3,496
Shellcode	1,133	378
Worms	130	44
المجموع الكلي للسجلات	175,341	82,332

تعتبر مجموعة البيانات UNSW-NB15 مفيدة لأنظمة كشف الاقتحام القائمة على الشبكة وتعتبر مجموعة بيانات مرجعية للباحثين في هذا الإطار.

3-4- مجموعة بيانات ADFA-LD , ADFA-WD NGIDS-DS

تم إنشاء مجموعة البيانات هذه في الجيل التالي من البنية التحتية للمركز الأسترالي للأمن السيبراني (Australian Centre OF Cyber Security :ACCS) في جامعة نيو ساوث ويل (UNSW) أكاديمية قوة الدفاع الأسترالية (Australian Defence Force Academy :ADFA)، تحتوي مجموعات البيانات على سجلات من نظم التشغيل لينكس وويندوز، وهي مستمدة من تقييم أنظمة كشف الاقتحام القائمة على المضيف. تم استخدام Ubuntu Linux الإصدار 11.04 كنظام تشغيل مضيف لبناء ADFA-LD. بعض حالات الهجوم في ADFA-LD مستمدة من برامج ضارة جديدة وهجمات Zero-Day، مما يجعل مجموعة البيانات هذه مناسبة لتسليط الضوء على الاختلافات بين نهج الأنظمة القائمة على التوقيع SIDS و القائمة على الشذوذ AIDS لكشف الاقتحام. يبين الجدول (6) بعض ميزات ADFA-LD مع نوع ووصف كل ميزة. تحتوي مجموعة البيانات NGIDS-DSعلى أنشطة الشبكة الطبيعية والغير طبيعية لمضيف (LINUX) ،تتألف NGIDS-DS من أربعة أنواع رئيسية من الملفات لتقييم أداء أنظمة كشف الاقتحام في المستقبل. يعتبر النوع الأول من الملفات بمثابة ملفات سجل المضيف، يحتوي مجلد سجلات المضيف على 99 ملف csv. النوع الثاني من الملفات يعتبر ملف سجل الشبكة واسمه NGIDS.pcap. يحتوي الملف الثالث على معلومات الحقيقة الأساسية واسمهground_truth.csv.اسم الملف الرابع feature_descr.csv،ويحمل معلومات السمات لسجلات المضيف[10].

الجدول (6): ميزات مجموعة البيانات ADFA-LD [32]

Name	Type	Description
srcip	nominal	Source IP address
sport	integer	Source port number
dstip	nominal	Destination IP address
dsport	integer	Destination port number
proto	nominal	Transaction protocol
state	nominal	Indicates to the state and its dependent protocol
dur	Float	Record total duration
sbytes	Integer	Source to destination transaction bytes
dbytes	Integer	Destination to source transaction bytes
sttl	Integer	Source to destination time to live value
dttl	Integer	Destination to source time to live value
sloss	Integer	Source packets retransmitted or dropped
dloss	Integer	Destination packets retransmitted or dropped
service	nominal	http, ftp, smtp, ssh, dns, ftp-data ,irc and (-) if not much used service
Sload	Float	Source bits per second
Dload	Float	Destination bits per second
Spfcts	integer	Source to destination packet count
Dpkts	integer	Destination to source packet count
swin	integer	Source TCP window advertisement value
dwin	integer	Destination TCP window advertisement value
stcpb	integer	Source TCP base sequence number
dtcpb	integer	Destination TCP base sequence number
smeansz	integer	Mean of the how packet size transmitted by the src
dmeansz	integer	Mean of the how packet size transmitted by the dst
trans_depth	integer	Represents the pipelined depth into the connection of http request response transaction
resbdvlen	integer	Actual uncompressed content size of the data transferred from the server's http service

تم تصنيف ADFA-LD و ADFA-WD و ADFA-WDSAA على بيانات تحتوي على ثلاث مجلدات بيانات مختلفة

- بيانات التدريب Training data (تحتوي على تتبع للبيانات الطبيعية فقط)
- بيانات التحقق Validation (تحتوي على تتبع للبيانات الطبيعية فقط)
- بيانات الهجوم Attack data (تحتوي على تتبع للبيانات الغير طبيعية (هجوم) فقط)

في حالة تصميم نظام الكشف عن الاقتحام القائم على الشذوذ يمكننا استخدام بيانات التدريب في مرحلة التدريب مع سمة طبيعي، في حال استخدام طرق تعلم الآلة الخاضعة للإشراف وفي مرحلة الاختبار يمكننا استخدام جميع بيانات الهجوم وبيانات التحقق من الصحة لقياس معدل الكشف (Data Rate: DR)، المعدل الإيجابي الخاطئ (False Positive Rate:FPR)، المعدل السلبي الخاطئ (False Negative Rate :FNR) ومعدل الإنذار الخاطئ (False Alert Rate: FAR). في حالة تصميم نظام الكشف عن الاقتحام القائم على التوقيع، يتم استخدام بيانات التدريب الطبيعية باستخدام سمة طبيعي وبعض بيانات الهجوم (من بيانات هجوم الاختبار) مع سمة الهجوم أثناء مرحلة التدريب واستخدام طرق تعلم الآلة الخاضعة للإشراف. خلال مرحلة الاختبار يتم استخدام بقية بيانات الهجوم التي لا تستخدم في التدريب وجميع بيانات التحقق الطبيعية، لقياس DR و FPR و FNR و FAR

تتضمن مجموعة البيانات ADFA-LD أيضًا استدعاءات النظام لأنواع مختلفة من الهجمات. توفر مجموعة بيانات ويندوز ADFA (ADFA-WD) مجموعة بيانات ويندوز حديثة لتقييم أنظمة كشف الاقتحام القائمة على المضيف. يوضح الجدول (7) عدد الاستدعاءات المطلوبة لكل فئة من فئات AFDA-LD و AFDA-WD الجدول (8) يصف تفاصيل كل فئة هجوم في مجموعة بيانات ADFA-LD. يسرد الجدول 11 ناقلات وتأثيرات ADFA-WD [32]

الجدول (7): عدد استدعاءات النظام ضمن فئات مختلفة لمجموعة البيانات AFDA-LD و AFDA-WD

ADFA- LD			ADFA-WD	
Dataset	Traces	System Calls	Traces	System Calls
Training data	833	308,077	355	13,504,419
Validation data	4372	2,122,085	1827	117,918,735
Attack data	746	317,388	5542	74,202,804
Total	5951	2,747,550	7724	205,625,958

الجدول (8): أصناف الهجمات ضمن مجموعة البيانات AFDA-LD

Attack	Payload	Vector	Count
Hydra-FTP	Password brute force	FTP by Hydra	162
Hydra-SSH	Password brute force	SSH Hydra	176
Adduser	Add new super user	Client-side poisoned executable	91
Java-Meterpreter	Java based Meterpreter	TIkiWiki vulnerability exploit	124
Meterpreter	Linux Meterpreter Payload	Client side poisoned executable	75
Webshell	C100 Webshell	PHP remote file inclusion vulnerability	118

3-5- مجموعة التدريب CICIDS2017 و CSE-CIC-IDS-2018

قدم المعهد الكندي للأمن السيبراني مجموعتي بيانات تدريب حديثة تسمى CICIDS2017 و CSE-CIC-IDS-2018 مخصصة لأنظمة كشف الاقتحام القائمة على الشذوذ، تتكون مجموعتي التدريب من أحدث التهديدات الأمنية والميزات. تم الحصول على مجموعة التدريب CICIDS2017 على مدى خمسة أيام تتضمن حركة بيانات طبيعية وهجمات مختلفة للمعهد الكندي للأمن السيبراني ودمجها وتقسيمها عبر أكثر من ثمانية ملفات مختلفة بينما تم الحصول على مجموعة التدريب CSE-CIC-IDS-2018 على مدى عشرة أيام وتقسيمها لتسعة ملفات مختلفة. تسرد

مجموعة بيانات KDD CUP 99 و NSL-KDD بعض فئات الهجوم بينما تسرد مجموعات البيانات CIC-IDS-2017 CSE-CIC-IDS-2018 مجموعة جديدة من الهجمات الناتجة عن ميزات حركة المرور الحقيقية للشبكة مثل حجب الخدمة الموزع، وحجب الخدمة، والقوة الغاشمة، وحقن SQL، و Botnet، تم تصنيف مجموعات البيانات هذه على حالات بها أكثر من 80 ميزة. تم عرض خصائص مجموعة البيانات CICIDS2017 و CSE-CIC-IDS-2018 وحالة التصنيف المفصلة في الجدول (9) والجدول (10) على التوالي.

الجدول (9): الخصائص العامة لمجموعة البيانات CICIDS2017 و CSE-CIC-IDS-2018

اسم مجموعة التدريب	CICIDS2017	CSE-CIC-IDS-2018
نمط مجموعة التدريب	Multi class	Multi class
الفترة الزمنية	5	10
البنية التحتية للهجوم	4 PCs, 1 router, 1 switch	50 PCs
البنية التحتية للضحية	3 server, 1 firewall, 2 switches, 10 PCs	420 PCs, 30 servers
تاريخ الإصدار	2017	2018
مجموع عدد الحالات المتميزة	2830540	16,233,002
معدل حالات الهجوم	19.7 %	17%
عدد الميزات	80	80
عدد الأصناف المميزة	15	18

الجدول (10) حالة التصنيف لمجموعة بيانات CICIDs2017

سمات الصنف	عدد الحالات
BENIGN	2359087
DoS Hulk	231072
PortScan	158930
DDoS	41835
DoS GoldenEye	10293
FTP-Patator	7938
SSH-Patator	5897
DoS slowloris	5796
DoS Slowhttptest	5499
Bot	1966
Web Attack – Brute Force	1507
Web Attack – XSS	652
Infiltration	36
Web Attack – Sql Injection	21
Heartbleed	11

3-5-1-المشاكل التي تعاني منها مجموعة التدريب CICIDS2017

تحتوي مجموعة البيانات CICIDS2017 على عدد قليل من أوجه القصور ونقاط الضعف

- حضور مبعثر Scattered Presence

مجموعة البيانات CICIDS2017 مبعثرة عبر ثمانية ملفات. تعتبر معالجة الملفات الفردية مهمة شاقة. لذلك، تم دمج تلك الملفات لتشكيل ملف واحد يحتوي على ما مجموعه 3119345 حالة لجميع الملفات.

- حجم هائل من البيانات Huge Volume of Data

بعد الجمع بين جميع الملفات، أصبحت مجموعة البيانات المدمجة تحتوي على بيانات لجميع سمات الهجوم الحديثة المحتملة في مكان واحد. ولكن في الوقت نفسه، يصبح حجم مجموعة البيانات المدمجة ضخمًا. هذا الحجم الهائل من البيانات يستهلك المزيد من الحمل الزائد للتحميل والمعالجة.

- القيم المفقودة Missing Values

تحتوي مجموعة بيانات CICIDS2017 المجمعة على 288602 حالة بها سمات لصنف مفقود و 203 حالات بها معلومات مفقودة. تمت إزالة هذه الحالات غير المرغوب فيها لتشكيل مجموعة بيانات تحتوي على 2830540 حالة فريدة ، كما يشير الجدول (10) الى وجود اختلال في معدل الانتشار للأصناف ضمن مجموعة البيانات ، بلغ معدل انتشار فئة البيانات الحميدة والتي تمثل الأغلبية هو 83.34٪ بينما بلغ معدل انتشار فئة الأقلية 0.00039٪ (نزيف القلب heartbleed). إن مثل هذا الاختلاف الكبير في معدل الانتشار، يجعل المصنف يميل نحو فئة البيانات الحميدة، ويصبح الوضع أسوأ عندما يعتمد المصنف على عينة من مجموعة البيانات هذه، حيث هناك احتمال كبير ألا توجد حالات لعلامة هجوم معينة مثل «نزيف القلب» أو «هجوم الويب – حقن Sql» في عينة مجموعة البيانات. ونتيجة لذلك، سيفشل المصنف في اكتشاف مثل هذا الهجوم عند وصول حالة من نوع هذا الهجوم. وبالتالي قد يؤدي إلى انخفاض دقة النظام وارتفاع مستوى الإنذار الكاذب

الجدول (11): مقارنة بين مجموعات البيانات المستخدمة حاليا [31]

Dataset	Realistic Traffic	Label Data	Zero –Day Attack	Full Packet Capture	Year
DARPA	√	√	X	√	1998
KDD CUP 99	√	√	X	√	1999
CAIDA	√	X	X	X	2007
NSL-KDD	√	√	X	√	2009
UNSW-NB15	√	√	X	√	2015
ADFA-WD	√	√	√	√	2014
ADFA- LD	√	√	√	√	2014
CICIDS2017	√	√	√	√	2017
CSE-CIC-IDS-2018	√	√	√	√	2018

الجدول (12): مقارنة بين مجموعات البيانات المستخدمة حاليا من حيث عدد المزايا

Dataset	Number of Features
DARPA	41
KDD CUP 99	41
CAIDA	-
NSL-KDD	41
UNSW-NB15	49
ADFA-WD	26
ADFA- LD	26
CICIDS2017	84
CSE-CIC-IDS-2018	84

الاستنتاجات والتوصيات:

تستخدم معظم أنظمة اكتشاف الاقتحام المعتمدة على تعلم الآلة ومنهجية التعلم العميق مجموعات البيانات القياسية مثل KDD Cup 99 و NSL-KDD و UNSW-NB15 و CSCIDS 2017. تفتقر معظم هذه المجموعات إلى حركة بيانات شبكة الحقيقية. كما لا تُصرح معظم المنظمات عن حركة بيانات الشبكة الخاصة بها بسبب مشكلة السرية. لذلك، هناك طلب كبير على بيانات حركة بيانات الشبكة في الزمن الحقيقي. مجموعات البيانات التي تم مناقشتها سابقا لا تواكب

التطور الحاصل للهجمات الجديدة التي اعتمدت أساليب جديدة غير مألوفة سابقا، كما أن مجموعات البيانات الحالية لا تناقش ولا تتضمن الهجمات والتهديدات الخاصة بالحاويات البرمجية التي تم اعتمادها بشكل متسارع في الآونة الأخيرة بشكل كبير مما خلق تحديات كبيرة أمام أنظمة كشف الاقتحام ومجموعات البيانات المستخدمة حاليا والحاجة الى الية جديدة لتمثيل سلوكها بهدف الكشف عن التهديدات الغير معروفة. لا يمكن للباحثين تطوير IDS فعال إلا عندما يتم توفير سيناريو هجوم في الزمن الفعلي يتضمن هجمات مبتكرة.

References:

[1]. K. Kendall, "A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems," Master Thesis, pp. 12–26, 1999, doi: citeulike-article-id:9077111.
[2]. R. P. Lippmann et al., "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," Proc. - DARPA Inf. Surviv. Conf. Expo. DISCEX 2000, vol. 2, 2000, pp. 12–26, doi: 10.1109/DISCEX.2000.821506.
[3]. KDD Cup 1999. Available on: http://kdd.ics.uci.edu/ databases/ kddcup 99/kddcup99.html, Ocotber 2007
[4]. Lee, W., &Stolfo, S. J. A Framework for Constructing Features and Models for Intrusion Detection Systems (Vol. 3), 2001.
[5]. [1] M. Tavallaee and E. B. and W. a. G. A. A. Lu, "'A detailed analysis of the KDD CUP 99 data set,' in Computational Intelligence for Security and Defense Applications," no. Cisda, 2009, pp. 1–6.
[6]. O. Atilla and E. Hamit, "A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015," PeerJ, 2016, pp. 0–21, doi: 10.7287/PEERJ.PREPRINTS.1954V1.
[7]. NSL-KDD data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009
[8]. Canadian Institute for Cybersecurity, "Intrusion detection evaluation dataset (CICIDS2017)." 2017, (Accessed on 06/15/2018). [Online]. Available: http://www.unb.ca/cic/datasets/ids-2017.html
[9]. "CIC DoS dataset," 2017, (Accessed on 06/15/2018). [Online]. Available: http://www.unb.ca/cic/datasets/dos-dataset.html
[10]. G. Creech and J. Hu, "ADFA IDS dataset." March 2017, (Accessed on 05/13/2019). [Online]. Available: http://www.azsecure-data.org/
[11] . "Generation of a new IDS test dataset: Time to retire the KDD collection." in Proceedings of the Wireless Communications and Networking Conference (WCNC). IEEE, 2013, pp. 4487–4492.
[12]. M. J. M. Turcotte, A. D. Kent, and C. Hash, "Unified Host and Network Data Set." arXiv e-prints, pp. 1–16, August 2017.
[13]. S. Behal and K. Kumar, "Measuring the impact of DDoS attacks on web services-a realtime experimentation." International Journal of Computer Science and Information Security, vol. 14, no. 9, p. 323, 2016.
[14]. J. J. Santanna, R. van Rijswijk-Deij, R. Hofstede, A. Sperotto, M. Wierbosch, L. Z. Granville, and A. Pras, "Booters â˜A ˇT an analysis of DDoSas- a-service attacks." in 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), 2015, pp. 243–251.

[15]. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Towards generating real-life datasets for network intrusion detection." International Journal of Network Security, vol.17,no. 6, pp. 683–701, 2015.
[16]. C. for Applied Internet Data Analysis, "The CAIDA UCSD "DDoS attack 2007" dataset." 2007, (Accessed on 05/05/2020). [Online]. Available: https://www.caida.org/data/passive/ddos-20070804_dataset.xml
[17]. Canadian Institute for Cybersecurity, "NSL-KDD dataset." 2009, (Accessed on 06/15/2018) . [Online]. Available: http://www.unb.ca/cic/ datasets/nsl.html
[18]. S. Hettich and S. D. Bay, "The UCI KDD archive," Irvine, CA: University of California, Department of Information and Computer Science, 1999, (Accessed on 06/15/2018). [Online]. Available: http://kdd.ics.uci.edu
[19]. Lincoln Laboratory, "MIT Lincoln Laboratory: DARPA intrusion detection evaluation." 2000, (Accessed on 06/15/2018). [Online]. Available: https://www.ll.mit.edu/ideval/data
[20]. Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, Jaideep Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection", Computer Science Department, University of Minnesota 200 Union Street SE, Minneapolis, MN 55455, USA
[21]. http://en.wikipedia.org/wiki/Data_mining
[22]. http://nsl.cs.unb.ca/NSL-KDD/
[23]. http://www.cs.waikato.ac.nz/ml/weka/
[24]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)
[25]. S. Revathi, Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 12, December - 2013
[26]. Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013
[27]. Santosh Kumar Sahu Sauravranjan Sarangi Sanjaya Kumar Jena, " A Detail Analysis on Intrusion Detection Datasets", 2014 IEEE International Advance Computing Conference (IACC)
[28]. Sapna S. Kaushik, Dr. Prof.P.R.Deshmukh," Detection of Attacks in an Intrusion Detection System", International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011, 982-986
[29]. UNSW-NB15 DataSet for Network Intrusion Detection Systems. Available on: http://www.cybersecurity. unsw. adfa.edu.au /ADFA% 20NB15%20Datasets
[30]. N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Inf. Secur. J., vol. 25, no. 1–3, 2016,pp. 18–31, doi: 10.1080/19393555.2015.1125974.
[31]. Ansam Khraisat, Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman " Survey of intrusion detection systems: techniques, datasets and challenges" , Internet Commerce Security Laboratory, Federation University Australia, Mount Helen, Australia, (2019) 2:20, https://doi.org/10.1186/s42400-019-0038-7

[32].https://learn.saylor.org/mod/book/view.php?id=29755&chapterid=5445
[33]. H.HINDY ,D.BROSSET ,E.BAYNE," A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems", IEEE Department, University of Strathclyde, Glasgow, Scotland, UK, Digital Object Identifier 10.1109/ACCESS.2017.DOI