مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٨) العدد (٤) ٢٠٢٤
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. ٢٠٢٤)٤(.No (٨)

مقارنة بين خوارزميات التعلم الآلي والعميق في الكشف عن هجمات الويب الشاذة بالنسبة لمجموعة بيانات صغيرة الحجم وتتصف باللا خطية

د. م. محمد مازن المصطفى *

م. ضحى يحي محمد **

(تاريخ الإيداع ٢٠٢٣/١١/١٦ . قُبِل للنشر في ٢٠٢٤/٣/١٤)

□ ملخّص□

أصبح التعلم العميق مركز اهتمام الباحثين في الآونة الأخيرة، حيث يتميز بقدرته على التعامل مع البيانات ذات الأبعاد العالية والتي تتصف باللا خطية. في حين تفوقت خوارزميات التعلم الآلي على التعلم العميق بالنسبة لمجموعة البيانات صغيرة الحجم. في بحثنا هذا قمنا بتطبيق ثمان نماذج تعلم آلي وسبع نماذج تعلم عميق على مجموعة البيانات (NSL-KDD) صغيرة الحجم وذات الأبعاد العالية والتي تتصف باللا خطية، من اجل نماذج التعلم الآلي سجلنا أعلى دقة ٨٤,٧٩% وأعلى أداء ٩٩,٩٩%، أما بالنسبة لنماذج التعلم العميق سجلنا أعلى دقة ٨١,٢٨% وأعلى أداء ٩٧,٢٩%. نتائج تجاربنا أكدت تفوق خوارزميات التعلم الآلي على خوارزميات التعلم العميق بالنسبة لمجموعة البيانات NSL-KDD()، وبالتالي تفوق خوارزميات التعلم الآلي على العميق بالنسبة لمجموعة البيانات الصغيرة والتي تتصف باللا خطية وذات أبعاد عالية.

الكلمات المفتاحية: التعلم الآلي، التعلم العميق، كشف هجمات الويب

* مدرس _ هندسة برمجيات_ الجامعة الافتراضية السورية .
** طالبة ماجستير في برنامج علوم الويب الجامعة الافتراضية السورية، حاصلة على إجازة في الهندسة المعلوماتية قسم البرمجيات ونظم المعلومات جامعة تشرين.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٨) العدد (٤) ٢٠٢٤
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (8) No. (4) 2024

Comparing machine learning and deep learning algorithms for detecting anomalous web attacks for a small-sized and non-linear dataset.

Prof. Eng. Mohammadmazen AlMustafa *
Eng. Douha yhea mohammad **

(Received 16/11/2023 . Accepted 14/3/2024)

□ ABSTRACT □

Recently, researchers have focused their attention on deep learning techniques, because its ability to deal with high-dimensional and non-linear data. On the other hand, machine learning algorithms outperformed deep learning algorithm for small size data set. In our research we applied eight machine learning algorithms and seven deep learning algorithms to the NSL-KDD dataset, this data set characterized by nonlinearity, high dimensions, and small size. We got the best value for the machine learning algorithms ٨٤,٧٩% for accuracy and ٩٩,٩٩% for performance, for deep learning algorithm, we got ٨١,٢٨% for accuracy and ٩٧,٢٩٪ for performance. From the results of our experiments, we proven superiority machine learning algorithms on deep learning algorithm for NSL-KDD dataset, thus outperforms the deep learning algorithms for the small-sized, high-dimensional, and non-linear dataset.

Keywords: machine learning, deep learning, detect web attacks.

* professor of software engineering, address Damascus.

** Master's student in Web Science at the Syrian Virtual University, address tartous.

المقدمة:

ازداد استخدام الإنترنت وتطورت التقنيات التي تعتمد عليه، وأصبحت المواقع وتطبيقات الإنترنت هدفاً رئيساً للهجمات الإلكترونية، وكان لابد من التركيز على حماية الشبكة للحفاظ على المعلومات الحساسة ومنع انتهاك الخصوصية. جذب مجال الكشف عن هجمات الويب اهتمام المبرمجين والمطورين وعلماء الحاسب بشكل كبير نظراً لكون الهجمات تشكل عائق في وجه الانفتاح المعلوماتي وخطر على المعلومات الحساسة والأنظمة المبنية على شبكات الويب. تنوعت أساليب الكشف عبر التاريخ، في البداية تم استخدام الأساليب غير التقليدية والطرق المبنية على القواعد ولكن لم تكن نتائجها مرضية، وقام الباحثون بدراسة تقنيات الذكاء الاصطناعي مثل خوارزميات التعلم الآلي والتعلم العميق التي تفوقت على الأساليب الاخرى من حيث قدرتها على الكشف التلقائي للأنماط الجديدة والتعامل مع الهجمات التي لم يتم اكتشافها، ولكن تباينت الآراء حول تحديد تقنية الذكاء الاصطناعي الأفضل للكشف عن هجمات الويب المختلفة.

ركز الباحثون في السنوات الأخيرة في دراساتهم على مجال كشف الشذوذ وقاموا باقتراح طرق عديدة، وكان لمجال الكشف عن هجمات الويب الاهتمام الأكبر وهذا ما أكدته الدراسة [٩]، حيث قارنوا فيها عدد الأوراق البحثية المنشورة في مجالات الكشف عن الشذوذ المختلفة بين عامي ٢٠٠٠ وعام ٢٠٢٠. في عام ٢٠١٨ قامت الدراسة [٥] باختبار خوارزميات التعلم العميق وتحديداً خوارزمية الشبكة العصبية التلافيفية بأنواعها البسيطة والمتوسطة والعميقة، واثبتوا تفوق الشبكة العصبية التلافيفية البسيطة على الشبكتين العصبتين التلافيفيتين المتوسطة والعميقة. في عام ٢٠١٩ قامت الدراسة [٨] بإثبات تجريبياً تفوق خوارزميات التعلم العميق على الآلي بالنسبة لمجموعة البيانات التي تتصف باللا خطية وذات الأبعاد العالية، وفي العام نفسه قام الباحثون في الدراسة [٦] بحل مشكلة الكشف بالشبكة الحقيقة التي تحتوي على الأقل أكثر من حركتي مرور. في عام ٢٠٢٠ قامت الدراسة [٤] بتقييم شامل لإثني عشر خوارزمية تعلم آلي وعميق على مجموعة بيانات صغيرة الحجم حيث اثبتوا من خلال هذه الدراسة تفوق خوارزميات التعلم الآلي على العميق من أجل مجموعة البيانات صغيرة الحجم، وقام الباحثون في الدراسة [٢] بتصميم نظام كشف ذكي باستخدام تقنيات التعلم العميق قادر على اكتشاف الاختراقات المختلفة للشبكة، وتقييم أداء الحل المقترح باستخدام مصفوفات تقييم مختلفة وقدم مقارنة بين نتائج الحل المقترح للعثور على أفضل نموذج لنظام اكتشاف اختراق الشبكة. في عام ٢٠٢١ قدمت الدراسة [١١] نهج للكشف عن الشذوذ في شبكة الويب بشكل تلقائي وللاختيار الأمثل للمتغيرات. في عام ٢٠٢٢ اقترحت الدراسة [١] نهج للكشف عن هجمات الويب، ويقوم باستخلاص الميزات بشكل تلقائي، حيث يقدم مقياس جديد يدعى (earliness) لتقييم مدى قدرة النهج على اكتشاف الهجمات في وقت مبكر، وفي دراسة أخرى [١٠] قاموا بدراسة دقة الكشف عن هجمات الويب باستخدام تقنيات التعلم العميق، وبالتحديد خوارزمية الشبكة العصبية التلافيفية أحادية البعد. في عام ٢٠٢٣ قامت الدراسة [١٢] باختبار ٦ خوارزميات تعلم عميق وقاموا بإجراء التجارب من أجل خمس طبقات مخفية لكل خوارزمية وعدد عصبونات مختلف واختبروا الخوارزميات بالنسبة للتصنيف الثنائي والتصنيف المتعدد، وفي دراسة أخرى [٧] قاموا بدراسة الكشف عن هجمات الويب في الوقت الحقيقي بشكل ديناميكي، اعتماداً على بناء نموذج تعلم عميق هجين، يجمع بين نموذج الشبكة العصبية التلافيفية ونموذج شبكة محددة البرمجيات (software defined network).

في دراستنا هذه أجرينا مقارنة أداء خوارزميات التعلم الآلي والعميق على مجموعة البيانات (NSL–KDD) صغيرة الحجم، التي تتصف باللا خطية وذات الأبعاد العالية.

قمنا بتنظيم ما تبقى من الورقة كما يلي: القسم الثاني يتناول أهمية البحث ومشكلته وأهدافه، القسم الثالث يتضمن منهجية البحث، القسم الرابع يحتوي النتائج ومناقشة تلك النتائج، القسم الخامس يتضمن الاستنتاجات والتوصيات، أما القسم الأخير يتضمن المصادر والمراجع.

أهمية البحث وأهدافه:

يعتبر الكشف عن هجمات الويب جزء أساسي من استراتيجيات الأمان، التي يجب ان تتوفر في أي مؤسسة أو موقع ويب، وذلك للحفاظ على سلامة المستخدمين وحماية البيانات وسمعة المؤسسة وتجنب الخسائر المالية. تنوعت أساليب الكشف عن هجمات الويب، حيث تفوقت تقنيات التعلم الآلي والعميق على الأساليب التقليدية في الكشف بسبب قدرتهم على تحليل البيانات واكتشاف الأنماط غير المعتادة وغير المعروفة لهجمات الويب [١١]، وبالتالي يعزز استخدام تقنيات التعلم الآلي والعميق في الكشف عن هجمات الويب من قدرة المؤسسات ومواقع الويب على التصدي للتهديدات الأمنية وحماية بياناتها. تتركز مشكلة البحث في دراسة أداء طرق التعلم الآلي والتعلم العميق في الكشف عن هجمات الويب ومقارنتها مع الطرق الحالية بهدف الوصول الى أفضل مستوى أمان ممكن. أما الهدف من هذا البحث هو القيام بدراسة تجريبية لخوارزميات التعلم الآلي والعميق على مجموعة بيانات صغيرة الحجم وذات ابعاد عالية وتتصف باللا خطية، حيث توصلنا من الدراسات السابقة الى تفوق خوارزميات التعلم الآلي على العميق بالنسبة لمجموعة البيانات صغيرة الحجم [٤] ، وتفوق خوارزميات التعلم العميق على الآلي بالنسبة لمجموعة البيانات ذات الأبعاد العالية والتي تتصف باللا خطية الدراسة [٨]، كما أن عدد عينات مجموعة البيانات NSL–KDD)) أقل من عدد عينات مجموعات البيانات المستخدمة في الدراسة [٤] التي صنفت في الدراسة السابقة على أنها صغيرة الحجم، وبالتالي يمكن وصف مجموعة البيانات ((NSL–KDD بأنها صغيرة الحجم، وفي الدراسة [٨] اعتبرت مجموعة البيانات (KDD–NSL) بأنها تتصف باللا خطية وذات أبعاد عالية. فكان لابد من دراسة أداء خوارزميات التعلم الآلي والعميق على مجموعة البيانات (NSL–KDD) للوصول إلى الخوارزمية ذات الأداء الأفضل.

منهجية البحث:

قمنا في بحثنا هذا بإجراء التجارب على مجموعة البيانات **NSL–KDD**))، بداية قمنا بتنظيف وإجراء معالجة أولية لمجموعة البيانات لتصبح جاهزة لتطبيق نماذج التعلم الآلي والعميق عليها، ومن ثم قمنا بتطبيق تلك النماذج ومقارنة نتائج تلك التجارب للوصول إلى النموذج الذي يحقق أعلى دقة. تعتمد منهجية هذا البحث على أربع خطوات أساسية: الخطوة الأولى توصيف مجموعة البيانات التي سنقوم بإجراء التجارب عليها، الخطوة الثانية القيام بعملية المعالجة الأولية للبيانات لتصبح جاهزة لاستخدامها في إحدى نماذج التعلم الآلي أو العميق، الخطوة الثالثة تتضمن بناء نماذج التعلم الآلي وتطبيقها على مجموعة البيانات (**NSL–KDD**) بعد إجراء عملية المعالجة الأولية لها، الخطوة الرابعة تتضمن بناء نماذج تعلم عميق وتطبيقها على مجموعة البيانات نفسها.

توصيف مجموعة البيانات NSL(KDD-):

قدمت الدراسة **[٣]** توصيف لمجموعة البيانات **(NSL–KDD)**، حيث يحتوي كل سجل في مجموعة البيانات هذه على ٤١ ميزة دخل وعمود خرج، وقام الباحثون في هذه الدراسة بتصنيف ميزات الدخل إلى الميزات الأساسية، الميزات المتعلقة بالمحتوى، الميزات الزمنية لحركة المرور، ميزات المضيف. ويمثل عمود الخرج بقيمتين رئيسيتين إما طبيعي ((**Normal** تدل على أن الطلب ليس هجوماً، او شذوذ **anomaly**)) وتدل على أن الطلب يمثل هجوم على الشبكة.

تنقسم مجموعة البيانات **NSL–KDD**)) إلى عينات تدريب يبلغ عددها ١٢٥٩٧٣ سجل وعينات اختبار يبلغ عددها ٢٢٥٤٤ سجل وتبلغ النسبة المئوية للعينات الشاذة في عينات التدريب ٥٣,٤٦% وفي عينات الاختبار ٥٦,٩٢%، وتبلغ النسبة المئوية للعينات الطبيعية في عينات التدريب ٤٦,٥٤% وفي عينات الاختبار ٤٣,٠٨%، مما سبق نستنتج أنه تتناسب تقريبياً عدد العينات الطبيعة والشذوذ بالنسبة لمجموعتي بيانات التدريب والاختبار، ومنه تعتبر مجموعة البيانات **(NSL–KDD)** متوازنة (Balanced). يعتبر توازن مجموعة البيانات مهما جداً في تحليل البيانات وتعلم الآلة، حيث يؤدي عدم وجود توازن في مجموعة البيانات إلى تحيز في عملية التعلم.

المعالجة الأولية للبيانات:

تتكون المعالجة الأولية للبيانات من مجموعة من الخطوات التي تطبق على البيانات، لتحسين جودتها، وتوحيد تنسيقها وتحويلها إلى تمثيل يمكن فهمه واستخدامه بشكل فعال. تتضمن هذه الخطوات تقسيم واصفات الدخل إلى واصفات عددية وفئوية، حيث قمنا بتقسيم واصفات مجموعة البيانات (NSL-KDD) إلى قسمين: القسم الأول يمثل بمجموعة بيانات عددية وتشمل الواصفات العددية والثنائية، أما القسم الثاني يمثل بمجموعة الواصفات الفئوية.

قمنا بتطبيق خوارزمية)one-hot encoding) على الواصفات الفئوية، واستخدامنا خوارزمية (standard scaler) على الواصفات العددية، وطبقنا خوارزمية (label encoding) على عمود الخرج، وأخيراً قمنا بحفظ البيانات الناتجة عن المعالجة الأولية، لاستخدامها لاحقاً في عمليات التدريب والاختبار.

آلية الكشف عن الهجمات:

تستخدم مجموعة البيانات (NSL-KDD) في تقييم أداء نظم الكشف عن اختراق الشبكة، حيث تحتوي هذه المجموعة على عدد كبير من السجلات، التي تمثل الأنشطة الشبكية، بما في ذلك الأنشطة الطبيعية والهجمات، حيث تنتمي الهجمات في مجموعة البيانات هذه إلى إحدى الفئات التالية (DOS, Probing, U2R, R2L)، كما تنوعت طرق الكشف عن هجمات الويب منها الطرق التقليدية والطرق المبنية على القواعد والتعلم الآلي والتعلم العميق، قمنا بدراستنا هذه ببناء نماذج تعلم آلي وعميق وتدريبها على بيانات التدريب الخاصة بمجموعة البيانات (NSL–KDD)، واختبار أداء النماذج في تحديد ما إذا كانت الأنشطة في مجموعة البيانات تشير إلى هجوم أم لا، وذلك من خلال تطبيق النموذج المدرب على بيانات الاختبار الخاصة بمجموعة البيانات

السابقة، ومن ثم مقارنة نتائج التنبؤ مع القيم الصحيحة في مجموعة الاختبار حيث استخدمنا لذلك المقاييس التالية(accuracy, precision, recall, F1score).

التعلم الآلي:

يعتبر التعلم الآلي فرع من فروع الذكاء الاصطناعي، ويعتمد على دراسة التقنيات الإحصائية والحسابية في التنقيب عن البيانات، وتحليلها لكشف الأنماط المتكررة او القواعد، واستخدام هذه الأنماط والقواعد لاتخاذ القرارات وتحقيق مهام محددة بدون برمجة صريحة. اعتمدنا في تجاربنا على ثمان خوارزميات تعلم الي قائم على الاشراف وهي كما يلي:

١. خوارزمية آلة المتجه الداعم)support vector machine): تعتمد على إيجاد أفضل مستوى فاصل (separating hyperplane)، يفصل بين مجموعة البيانات)Class).

٢. خوارزمية الانحدار اللوجستي)Logistic Regression): تستخدم لتصنيف البيانات الى فئات)Class).

٣. خوارزمية شجرة القرار (Decision Tree): تقوم الخوارزمية بتحليل البيانات وتقسيمها الى فئات باستخدام سلسلة من القرارات المستندة الى القيم الموجوده في مجموعة من المتغيرات، تمثل هذه القرارات في شكل هيكل شجري.

٤. خوارزمية الغابة العشوائية (Random forest): تستند الى مجموعة من الأشجار القرارية (Decision Tree)، تعتمد على بناء أشجار قرارية مستقلة ومتنوعة، ثم يتم دمج قراراتها للوصول الى التنبؤ النهائي.

٥. خوارزمية الجيران الأقرب (K-Nearest Neighbors): تعمل على مبدأ ان العينات المتشابهة يجب ان يكون لها تصنيف مشابه، وتعتمد الخوارزمية على حساب المسافة بين العينة المراد تصنيفها وبين العينات التدريبية.

٦. خوارزمية (CatBoost): يعتمد على تقنية التعزيز مرتبة ((ordered boosting، تدمج نماذج ضعيفة متعددة لإنشاء نموذج تنبؤي اقوى يستخدم للتصنيف (classification).

٧. خوارزمية)AdaBoost): هي خوارزمية تعتمد على تقنية التعزيز (boosting)، تقوم بتدريب سلسلة من النماذج الضعيفة، بشكل متكرر بناءً على عينات التدريب الموزونة، يتم تعديل أوزان العينات التدريبية بناء على أداء النماذج السابقة.

٨. خوارزمية)XGBoost): يعتمد على تقنية التعزيز (boosting)لإضافة الأشجار إلى المجموعة بشكل متكرر، تقوم كل شجرة لاحقة بتصحيح أخطاء الأشجار السابقة.

إن جميع الخوارزميات السابقة تحتوي على بارامترات فائقة التي تتحكم بسلوكها، قمنا باستخدام خوارزمية البحث الشبكي وخوارزمية التقاطع المتبادل لاختيار النموذج الأفضل وتحسين البارامترات الفائقة وذلك بالنسبة لكل خوارزمية.

التعلم العميق:

يعتبر التعلم العميق فرع من التعلم الآلي، يركز على الشبكات العصبية الاصطناعية والخوارزميات المستوحاة من بنية ووظيفة الدماغ البشري، لمعالجة البيانات واستخلاص الأنماط والمعلومات المعقدة باستخدام طبقات متعددة من الخلايا العصبية الاصطناعية المترابطة. قمنا في بحثنا هذا بإجراء التجارب على سبع خوارزميات تعلم عميق، وهي كما يلي:

- الشبكة العصبية التلافيفية Neural Network (Convolutional): تتكون من طبقة التلافيف (convolutional layer) وطبقة التجميع (pooling layer)، وطبقة الارتباط الكامل (dense layer)، وطبقة التسريب (Drop)، يتميز هذا النموذج بالقدرة على التقاط المعلومات المكانية، ويتمتع بخاصية لمشاركة المعلمات، حيث يتم استخدام مجموعة الأوزان نفسها عبر مواقع مكانية مختلفة.
- الشبكة العصبونية التكرارية (Recurrent neural networks): ظهرت نتيجة الحاجة إلى التعامل مع البيانات التسلسلية والبيانات التي تحتوي على عناصر زمنية، تعتمد القيمة الحالية على القيم السابقة، وتستخدم الشبكة العصبونية التكرارية الذاكرة لحفظ العلاقة والتبعية بين قيم البيانات المتتالية، ويمكن تدريب هذه الشبكات لتعلم الأنماط المتكررة في التسلسلات، ولها عدة أنواع:

❖ الشبكة العصبونية التكرارية ذات الذاكرة طويلة وقصيرة المدى (long short term memory network): تعتمد على التبعية حيث أن مخرجات الخطوة السابقة تستخدم كمدخلات للخطوة الحالية، وتنقسم خوارزمية (LSTM) لعدة أنواع منها:

١. الشبكة العصبونية التكرارية ذات الذاكرة طويلة وقصيرة المدى البسيطة ((Vanilla LSTM: هو نموذج (LSTM) يحتوي على طبقة مخفية واحدة من وحدات (LSTM)، وطبقة إخراج تستخدم للتنبؤ.

٢. الشبكة العصبونية التكرارية ذات الذاكرة طويلة وقصيرة المدى المكدسة (Stacked LSTM): هو نموذج)LSTM) يتكون من طبقات مخفية متعددة واحدة فوق الأخرى.

٣. الشبكة العصبونية التكرارية ذات الذاكرة طويلة وقصيرة المدى ثنائية الاتجاه (Bidirectional LSTM): هو نموذج (LSTM) يسمح بمعالجة المعلومات في الاتجاهين الأمامي والخلفي، حيث يلتقط المعلومات من السياقات الماضية والمستقبلية في وقت واحد.

٤. الشبكة العصبونية التكرارية ذات الذاكرة طويلة وقصيرة المدى المكدسة وثنائية الاتجاه (stacked LSTMBidirectional): تعتمد على تكديس طبقات متعددة من وحدات LSTM() فوق بعضها البعض في الاتجاهين الأمامي والخلفي.

❖ الشبكة العصبونية التكرارية ذات البوابات (Gated Recurrent Unit): نوع من الشبكة العصبية التكرارية (RNN)، وتتميز بانها بسيطة وتتناسب مع المسائل التي تتضمن تسلسل زمني حيث التسلسل التالي متعلق بما قبله، وتعد الشبكات التكرارية ذات البوابات (Gated Recurrent unite) ابسط من الشبكات (LSTM) وتلائم المسائل ذات البيانات صغيرة نسبيا.

- نموذج التدريب الهجين ((CNN-LSTM: هو نوع من نماذج الشبكة العصبية العميقة، يجمع بين الشبكات العصبية التلافيفية (CNN) وشبكات الذاكرة طويلة وقصيرة المدى (LSTM)، حيث يستخدم خرج الشبكة العصبية التلافيفية (CNN) كدخل للشبكة الذاكرة طويلة وقصيرة المدى (LSTM).

قمنا بتحسين البارامترات الفائقة المستخدمة في الشبكات العصبية السابقة، باستخدام خوارزمية (Hyperband tuning) .

النتائج:

قمنا بإجراء التجارب على مجموعة البيانات ((NSL-KDD، قمنا بدايةً بتنظيف مجموعة البيانات هذه واجراء عملية معالجة أولية للبيانات الفئوية والعددية، حيث أجرينا التجارب على منصة (google Colab) باستخدام لغة برمجة (python)، وقمنا بتصدير الملفات الناتجة عن المعالجة الأولية لمجموعة البيانات الى ملف خاص موجود في موقع التخزين السحابي (google drive) بين تاريخي ٢٠٢٣/٥/١٦ وتاريخ ٢٠٢٣/١٠/٩ في مدينة طرطوس في الجمهورية العربية السورية، بعد ذلك قمنا بتطبيق ثمان نماذج تعلم الي على مجموعة البيانات المعالجة فكانت النتائج كما في الجدول(١).

الجدول (١): نتائج تطبيق نماذج التعلم الآلي على مجموعة البيانات NSL-KDD

algorithm	accuracy	precision	Recall	F1_score	AUC
Logistic Regression	75.05%	61.27%	92.32%	73.65%	78.44%
SVM	84.04%	75.41%	95.7%	84.34%	85.17%
Decision Tree	79.51%	67.73%	94.85%	79.01%	81.94%
Random Forest	77.04%	61.71%	96.81%	75.37%	81.3%
XGBoost	79.92%	66.88%	96.88%	79.13%	82.91%
CatBoost	84.79%	76.88%	95.53%	85.18%	85.63%
AdaBoost	80.17%	68.53%	95.23%	79.7%	82.44%
KNN	77.77%	63.06%	96.76%	76.35%	81.67%

من الجدول السابق توصلنا الى ان خوارزمية (CatBoost) حققت أفضل دقة قدرها ٨٤,٧٩% مقارنة مع خوارزميات التعلم الآلي الأخرى. قمنا بعد ذلك بتطبيق سبع نماذج خوارزميات تعلم عميق على نفس مجموعة البيانات، فكانت النتائج كما في الجدول (٢).

الجدول (٢): نتائج تطبيق نماذج التعلم العميق على مجموعة البيانات ((NSL-KDD

algorithm	accuracy	precision	Recall	F1_score	AUC
CNN	%٨٠,٦٣	%٨٤,٨٦	٨٠,٤٢ %	%٨٢,٣٥	%٨٩,٥٦
Vanilla LSTM	%٨١,٢٨	%٩٢,٢٦	٧٣,٢٨ %	%٨١,٤١	%٩٠,٤٩
Stacked LSTM	%٧٥,٩١	%٩٤,٠٣	%٦١,٨	%٧٤,١١	%٨٨,٢
Bidirectional LSTM	%٧٥,٨٤	%٩٢,١٨	٦٣,٠٥ %	%٧٤,٥٢	%٨٨,٨١
Stacked Bidirectional LSTM	%٧٨,٠١	%٩٣,٨٤	%٦٥,٨	%٧٧,٠٤	%٨٩,٢٣
GRU	%٧٥,٠٦	%٩٢,١٤	٦١,٥٣ %	%٧٣,٣٨	%٨٧,٦٢
CNN-LSTM	%٧٩,١٤	%٩٦,٧٢	٦٥,٦٥ %	%٧٧,٩٣	%٩٠,١١

من الجدول (٢) نجد ان خوارزمية ((Vanilla LSTM حققت أعلى دقة مقارنة مع خوارزميات التعلم العميق الأخرى بقيمة قدرها ٨١,٢٨%. أما الجدول (٣) يحتوي على مقارنة بين نتائج تجارب خوارزميات التعلم الآلي والعميق التي قمنا بها، وذلك من حيث الدقة والأداء.

الجدول (٣): مقارنة الخوارزميات الأفضل أداء ودقة بين خوارزميات التعلم الآلي وخوارزميات التعلم العميق

	Accuracy		precision	
	Best algorithm	value	Best algorithm	value
Machine Learning	CatBoost	84.79%	adaboost	99.99%
Deep Learning	LSTM	81.28%	Bidirectional LSTM	97.29%

من الجدول (٣) نستنتج تفوق خوارزميات التعلم الآلي على خوارزميات التعلم العميق من حيث الدقة والأداء، كما أن خوارزميات التعلم العميق تحتاج الى تكلفة اعلى من حيث زمن التدريب والذاكرة. قمنا بمقارنة أفضل النتائج المستخلصة من الدراسات السابقة التي تعتمد على مجموعة البيانات (NSL-KDD)، مع أفضل

الدراسة	الخوارزمية	الدقة
الدراسة [٥]	Shallow CNN	%٧٩,٤٤
الدراسة [٨]	LSTM-seq٢seq	%٩٩
الدراسة [٢]	CNN	%٩٧,٠١
تجاربنا	CatBoost	84.79%

النتائج التي توصلنا إليها من خلال تجاربنا لنفس مجموعة البيانات كما في الجدول (٤).

الجدول (٤): مقارنة تجاربنا مع الدراسات السابقة التي تستخدم مجموعة البيانات (NSL-KDD)

قمنا بمقارنة أفضل نتيجة توصلنا اليها بنتائج الدراسات السابقة بالنسبة لمجموعة البيانات ((NSL-KDD، كما في الجدول (٤)، بالنسبة للدراسة **[٥]** تفوقت نتائجنا على نتائج هذه الدراسة، في حين تفوقت الخوارزميات المستخدمة في الدراستين **[٨]**، [٢] على نتائج الخوارزميات المستخدمة في تجاربنا.

الاستنتاجات والتوصيات:

ومما سبق وجدنا أن مجموعة البيانات ((NSL-KDD تتصف بأنها عالية الأبعاد وباللا خطية بالإضافة إلى صغر حجمها، كما توصلنا من الدراسة [٨] إلى تفوق خوارزميات التعلم العميق على الآلي بالنسبة لمجموعة البيانات عالية الابعاد وغير الخطية فقط، ومن الدراسة [٤] تفوق خوارزميات التعلم الآلي على العميق بالنسبة لمجموعة البيانات صغيرة الحجم فقط. لم يتم في الأبحاث السابقة مقارنة بين خوارزميات التعلم الآلي والعميق فيما يتعلق بمجوعة البيانات صغيرة الحجم وذات الأبعاد العالية وغير الخطية.

اعتماداً على نتائج التجارب التي قمنا بها الموضحة في الجدول (١)، الجدول (٢)، الجدول (٣) توصلنا الى تفوق خوارزميات التعلم الآلي على خوارزميات التعلم العميق بالنسبة لمجموعة البيانات ((NSL-KDD. وبناء على ما توصلنا اليه من استنتاجات اعلاها بان مجموعة البيانات ((NSL-KDD تتصف بانها عالية الابعاد وغير الخطية بالإضافة الى صغر حجمها ومن تجاربنا نستنتج تفوق خوارزميات التعلم الآلي على العميق بالنسبة لمجموعة البيانات صغيرة الحجم والتي تتصف بانها ذات أبعاد عالية وغير الخطية.

في المستقبل، نسعى إلى إجراء التجارب على مجموعة أكبر من خوارزميات التعلم العميق الهجينة من أجل مجموعة البيانات (NSL-KDD) ومقارنة نتائج هذه التجارب مع نتائج تطبيق هذه الخوارزميات على مجموعة بيانات كبيرة الحجم.

المراجع

[1] Ahmad, T., Truscan, D., Vain, J., & Porres, I. (2022), *Early detection of network attacks using deep learning*. *ahmad2022early*, pp. 30--39, Retrieved from https://ieeexplore.ieee.org/abstract/document/9787923

[2] Al-Emadi, S., Al-Mohannadi, A., & Al-Senaid, F. (2020), *Using deep learning techniques for network intrusion detection*. *2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT)*, pp. 171--176.

[3] Dhanabal, L., & Shantharajah, S. (2015), *A study on NSL-KDD dataset for intrusion detection system based on classification algorithms*. *International journal of*

advanced research in computer and communication engineering, *6*, pp. 446--452. Retrieved from https://e-tarjome.com/storage/btn_uploaded/2019-07-13/1563006133_9702-etarjome-English.pdf

[4] Elmrabit, N., Zhou, F., Li, F., & Zhou, H. (2020), *Evaluation of machine learning algorithms for anomaly detection*. *2020 international conference on cyber security and protection of digital services (cyber security)*, pp. 1--8. Retrieved from https://ieeexplore.ieee.org/abstract/document/9138871

[5] Kwon, D., Natarajan, K., Suh, S. C., Kim, H., & Kim, J. (2018), *An empirical study on network anomaly detection using convolutional neural networks*. *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1595--1598. https://ieeexplore.ieee.org/abstract/document/8416441

[6] Lin, P., Ye, K., & Xu, C.-Z. (2019), *Dynamic network anomaly detection system by using deep learning techniques*. *Cloud Computing--CLOUD 2019: 12th International Conference, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25--30, 2019, Proceedings 12*, pp. 161--176. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-23502-4_12

[7] Liu, H., & Wang, H. (2023), *Real-Time Anomaly Detection of Network Traffic Based on CNN*. Symmetry, p. 1205. Retrieved from https://www.mdpi.com/2073-8994/15/6/1205

[8] Malaiya, R. K., Kwon, D., Suh, S. C., Kim, H., Kim, I., & Kim, J. (2019), *An empirical evaluation of deep learning for network anomaly detection*. *I*EEE Access, pp. 140806--140817. https://ieeexplore.ieee.org/abstract/document/8846674

[9] Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021), *Machine learning for anomaly detection: A systematic review*. *Ieee Access*, pp. 78658--78700. Retrieved from https://ieeexplore.ieee.org/iel7/6287639/6514899/09439459.pdf

[10] Qazi, E. U., Almorjan, A., & Zia, T. (2022), *A one-dimensional convolutional neural network (1d-cnn) based deep learning system for network intrusion detection*. *Applied Sciences*, 16, p. 7986. Retrieved from https://www.mdpi.com/2076-3417/12/16/7986

[11] Tekerek, A. (2021, *A novel architecture for web-based attack detection using convolutional neural network*. *Computers \& Security*, p. 102096. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0167404820303692

[12] Wang, Y.-C., Houng, Y.-C., Chen, H.-X., & Tseng, S.-M. (2023), *Network Anomaly Intrusion Detection Based on Deep Learning Approach*. Sensors, p. 2171. Retrieved from https://www.mdpi.com/1424-8220/23/4/2171