SJST
sjst.scst.edu.ly

Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)

مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

تأثير اختيار الميزات على أداء خوارزميات التعلم الآلي في كشف التسلل باستخدام مجموعتي بيانات NSL-KDD وUNSW-NB15

الصادق علي محمد العربي, عبيد بلقاسم زايد التقازي, سليمان عبدالجليل العارف المودي

المعهد العالي للعلوم والتقنية الزهراء, ليبيا

الملخص

تهدف هذه الورقة الى تحليل تأثير اختيار الميزات (Feature Selection) على أداء أنظمة كشف التسلل (IDS) المعتمدة على التعلم الآلي باستخدام مجموعتي بيانات قياسيتين هما NSL-KDD وUNSW-NB15. وقد تم تقييم ثلاث خوارزميات تعلم آلي شائعة وهي شجرة القرار (Decision Tree)، والغابات العشوائية (Random Forest)، وأقرب الجيران (K-Nearest Neighbors - KNN) وتم قياس الأداء بإستخدام عدة مقاييس تقييم تشمل الدقة (Accuracy), الدقة التنبؤية (Precision), الاسترجاع (Recall), والمعدل التوافقي (F1-score) بالإضافة الى زمن الاستدلال (Inference time) وزمن التدريب كمعايير لقياس الكفاءة. وقد اعتمدت الدراسة على منهجية اختيار ميزات باستخدام مربع كاي (Chi-Square) لاستبعاد الميزات الزائدة وتقليل الابعاد.

أظهرت النتائج التجريبية أن اختيار الميزات نجح في الحفاظ على الاستقرار والدقة العالية للنماذج وتحسين الكفاءة الحسابية, حيث حققت خوارزمية الغابات العشوائية دقة 99.78% على بيانات NSL-KDD و 97.63% على بيانات UNSW-NB15 اما خوارزمية KNN فقد حققت انخفاض كبير في زمن الاستدلال بعد تقليل عدد الميزات, حيث انخفض من 34.77% الى 7.84%, وفي الغابات العشوائية على مجموعة بيانات UNSW-NB15 حيث انخفض زمن التدريب من 24 ثانية الى 5.6 ثانية, مما يعزز كفاءة النموذج في أنظمة كشف التسلل الزمن الحقيقي. تؤكد النتائج أن اختيار الميزات لا يساهم فقط في تحسين دقة التصنيف، بل يقلل أيضًا من التعقيد الحسابي، مما يجعل أنظمة كشف التسلل (IDS) أكثر كفاءة وقابلية للتطبيق في البيئات الشبكية واسعة النطاق.

الكلمات المفتاحية:

أنظمة كشف التسلل(ISD)، اختيار الميزات، التعلم الآلي، بيانات NSL-KDD ، بيانات UNSW-NB15، الغابات العشوائية، شجرة القرار, أقرب الجيران، الأمن السيبراني.

Abstract

This paper aims to analyse the impact of Feature Selection on the performance of machine learning-based Intrusion Detection Systems (IDS) using two benchmark datasets: NSL-KDD and UNSW-NB15. Three widely used machine learning algorithms were evaluated: Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN). Performance was measured using several evaluation metrics, including Accuracy, Precision, Recall, and F1-score, in addition to Inference Time and Training Time as efficiency indicators. The study adopted a feature selection methodology based on the Chi-Square test to eliminate redundant features and reduce data dimensionality.

Experimental results showed that feature selection successfully maintained high model stability and accuracy while improving computational efficiency. Random Forest achieved an accuracy of 99.78% on the NSL-KDD dataset and 97.63% on the UNSW-NB15 dataset. In contrast, the KNN algorithm achieved a significant reduction in inference time after feature reduction, decreasing

SJST
sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

from 34.77 seconds to 7.84 seconds. Furthermore, for Random Forest on the UNSW-NB15 dataset, the training time decreased from 24 seconds to 5.6 seconds, enhancing the model's efficiency for real-time intrusion detection systems. The results confirm that feature selection not only improves classification accuracy but also reduces computational complexity, making IDS more efficient and scalable for large-scale network environments.

Keywords: *Intrusion Detection Systems (IDS), Feature Selection, Machine Learning, NSL-KDD Dataset, UNSW-NB15 Dataset, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Cybersecurity*

Submitted: 14/03/2026 Accepted: 18/05/2026

1- المقدمة

أصبحت أنظمة كشف التسلل (Intrusion Detection Systems - IDS) جزءًا أساسيًا في بنية أمن الشبكات الحديثة، نظرًا للزيادة المستمرة في الهجمات السيبرانية وتعقيد أساليبها. ومع التطور السريع في تقنيات الهجوم، لم تعد الحلول التقليدية كافية، مما دفع إلى الاعتماد بشكل متزايد على خوارزميات التعلم الآلي (Machine Learning) القادرة على اكتشاف الأنماط غير الطبيعية في حركة الشبكة والتكيف مع التهديدات الجديدة.

تعتمد فعالية نماذج التعلم الآلي في أنظمة كشف التسلل بشكل كبير على جودة وتمثيل البيانات المدخلة، حيث تؤثر الميزات غير المهمة أو الزائدة سلبًا على أداء النماذج من حيث الدقة والكفاءة الحسابية. لذلك، يُعد اختيار الميزات (Feature Selection) خطوة حاسمة في تقليل أبعاد البيانات، وتحسين قدرة النموذج على التعميم، وتقليل التعقيد الحسابي.

ورغم تعدد الدراسات التي تناولت تقنيات اختيار الميزات في مجالIDS ، إلا أن معظمها اعتمد على مجموعة بيانات واحدة فقط، مما يحد من إمكانية تعميم النتائج على بيئات شبكية مختلفة. بالإضافة إلى ذلك، ركزت العديد من الدراسات على تحسين دقة التصنيف فقط، دون تحليل شامل لتأثير تقليل الأبعاد على زمن التدريب وزمن الاستدلال، وهما عاملان أساسيان في تطبيقات الزمن الحقيقي.

في هذا السياق، تهدف هذه الدراسة إلى تقديم تحليل شامل لتأثير اختيار الميزات على أداء خوارزميات التعلم الآلي في أنظمة كشف التسلل باستخدام مجموعتي بيانات معياريتين هما NSL-KDD و UNSW-NB15. وتم اختيار هذه المجموعات لتمثيل بيئتين مختلفتين، حيث تمثل NSL-KDD بيئة تقليدية أكثر تنظيماً، بينما تعكس UNSW-NB15 بيئة شبكية حديثة وأكثر تعقيدًا وواقعية.

تعتمد الدراسة على ثلاث خوارزميات تعلم آلي شائعة هي شجرة القرار(Decision Tree (DT ، و الغابات العشوائية Random Forest (RF)، وأقرب الجيران K-Nearest Neighbors (KNN)، وذلك لتغطية نماذج تمثل أساليب مختلفة في التعلم: النماذج الشجرية، النماذج التجميعية، والنماذج المعتمدة على المسافة. ويتم تقييم الأداء باستخدام مقاييس الدقة (Accuracy)، والدقة التنبؤية (Precision) ، والاسترجاع (Recall) ، والمعدل التوافقي (F1-score) ، بالإضافة إلى زمن التدريب (Training Time) وزمن الاستدلال (Inference Time) لقياس الكفاءة التشغيلية.

تعتمد منهجية الدراسة على تقليل الأبعاد باستخدام أسلوب اختيار ميزات يعتمد على اختبار مربع كاي Chi-Square ، بهدف تحديد أكثر الميزات تأثيرًا في عملية التصنيف وإزالة الميزات غير المهمة أو المتكررة. وتم تطبيق هذه المنهجية بشكل مستقل على كل مجموعة بيانات لضمان التكيف مع خصائصها المختلفة.

وبناءً على ذلك، تسعى هذه الدراسة للإجابة على السؤال البحثي التالي:

هل يساهم اختيار الميزات في تحسين التوازن بين دقة التصنيف والكفاءة الزمنية لخوارزميات التعلم الآلي في أنظمة كشف التسلل عند استخدام مجموعتي بيانات مختلفتين NSL-KDD) و UNSW-NB15) مقارنة باستخدام جميع الميزات الأصلية؟

2. الدراسات السابقة

- دراسة Yin, Y. et al. (2022) استخدم طريقة هجينة لاختيار الميزات (IGRF-RFE) لكشف التسلل باستخدام مجموعة بيانات UNSW-NB15, وقام دمج معلومات الكسب (Information Gain) مع الغابات العشوائية (Random Forest)

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

وإزالة الميزات التكرارية (RFE), وأظهرت النتائج ان تقليل عدد الميزات إلى حوالي 23 ميزة يحسن دقة نماذج كشف التسلل بشكل ملحوظ وتحسين الأداء العام على بيانات UNSW-NB15.

- قام Kasongo & Sun. 2020 بتحليل أداء أنظمة كشف التسلل باستخدام اختيار الميزات على مجموعة بيانات UNSW-NB15, وتأثير اختيار الميزات على أداء خوارزميات التعلم الآلي في IDS, وأظهرت النتائج تحسين الدقة وتقليل الأبعاد وتحسين كفاءة النماذج في معالجة البيانات.

- دراسة Cheng et al. وآخرون 2024 بعنوان اختيار ميزات متعدد الأهداف لأنظمة كشف التسلل والهدف هو تحسين اختيار الميزات بناءً على أكثر من هدف مثل: زيادة الدقة, تقليل عدد الميزات وتحسين معدل الكشف, وأظهرت النتائج تحسين التعميم عبر أكثر من مجموعة بيانات مثل NSL-KDD وUNSW-NB15.

- دراسة Jouhari et al. وآخرون (2024) بعنوان نظام كشف تسلل فعال باستخدام Chi-Square مع التعلم العميق تهدف لدمج اختيار الميزات الإحصائي (Chi-Square) مع نماذج التعلم العميق وأظهرت النتائج دقة عالية وصلت إلى حوالي 97%, تقليل زمن الاستجابة وتحسين الأداء في الشبكات الحديثة.

- دراسة Emirmahmutoğlu & Atay, (2025) بعنوان إطار عمل يعتمد على اختيار الميزات لتحسين أنظمة كشف التسلل باستخدام التعلم الآلي لغرض بناء إطار شامل يدمج اختيار الميزات مع خوارزميات التعلم الآلي وأظهرت النتائج تحسين الأداء عبر عدة مجموعات بيانات وتعزيز قابلية التعميم للنماذج.

- دراسة Ali et al. وآخرون (2023) بعنوان كشف التسلل باستخدام التعلم الآلي مع اختيار الميزات التجميعي لغرض استخدام طرق اختيار ميزات تعتمد على النماذج التجميعية (Ensemble) وأظهرت النتائج تحسين دقة التصنيف وتقليل الخطأ في الكشف و كان الأفضل أداءً Random Forest.

- دراسة Alqahtani et al. وآخرون (2023) بعنوان مقارنة أنظمة كشف التسلل باستخدام التعلم الآلي على NSL-KDD وUNSW-NB15 لدراسة مقارنة بين مجموعتي بيانات وكانت النتائج UNSW-NB15 أكثر واقعية وحداثة وان اختيار الميزات يحسن الأداء في كلا البيانات.

- دراسة Khan et al. وآخرون (2024) بعنوان نظام كشف تسلل خفيف للإنترنت الأشياء باستخدام اختيار الميزات لغرض تصميم نظام كشف تسلل (IDS) خفيف مناسب للأنظمة ذات الموارد المحدودة وتوصلت النتائج الى تقليل زمن الاستجابة وتحسين الأداء في البيئات الزمنية الحقيقية.

- دراسة لـ Zhang et al وآخرون (2023) بعنوان طريقة هجينة لاختيار الميزات لكشف الشذوذ في الشبكات وتهدف لدمج عدة طرق لاختيار الميزات لتحسين الكشف, وأكدت النتائج تحسين معدل الكشف وتقليل عدد الميزات بشكل كبير.

خلصت الدراسات الحديثة الى أن اختيار الميزات (Features selection) يمثل عنصرًا أساسيًا في تحسين أداء خوارزميات التعلم الآلي في أنظمة كشف التسلل (IDS)، حيث بينت دراسات مثل Yin وآخرون (2022) وAli وآخرون (2023) أن تقليل الأبعاد يؤدي إلى تحسين الدقة وتقليل زمن المعالجة. كما أكدت دراسات أخرى مثل Alqahtani وآخرون (2023) أهمية استخدام أكثر من مجموعة بيانات لضمان قابلية تعميم النتائج.

3. المساهمة العلمية

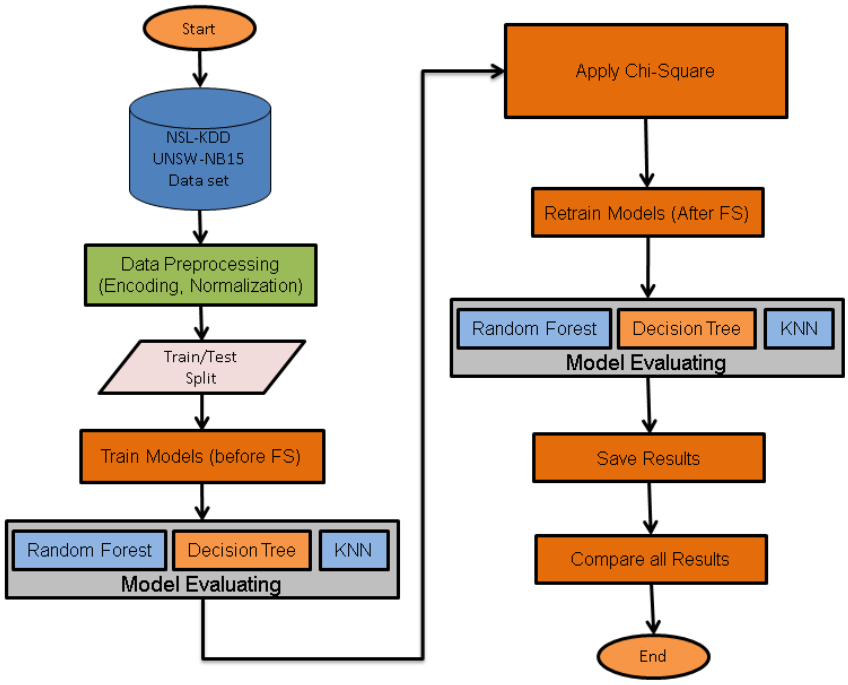
تقدم هذه الدراسة مساهمة علمية في مجال أنظمة كشف التسلل (IDS) المعتمدة على التعلم الآلي من خلال تطوير إطار تقييم شامل ومتعدد المجموعات البيانية يهدف إلى تحليل أثر اختيار الميزات على كل من دقة التصنيف والكفاءة الحسابية بشكل متزامن، وذلك باستخدام مجموعتي بيانات معياريتين تمثلان بيئات شبكية مختلفة NSL-KDD وUNSW-NB15 وتكمن المساهمة المنهجية الأساسية في اعتماد خط معالجة بيانات خالٍ من تسرب المعلومات(Leakage-Free Pipeline) ، حيث يتم تطبيق اختبار Chi-Square لاختيار الميزات بعد عملية تقسيم البيانات، وبشكل حصري على مجموعة التدريب فقط، مع إسقاط نفس التحويل على بيانات الاختبار، مما يعزز موثوقية التقييم ويضمن قابلية تعميم النتائج في البيئات الواقعية.

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

كما تقدم الدراسة تحليلًا مقارنًا لسلوك خوارزميات تعلم آلي مختلفة (Decision Tree ، Random Forest، KNN) تجاه تقليل الأبعاد، حيث تُظهر النتائج أن النماذج التجميعية تتمتع باستقرار أعلى تحت تقليل الميزات، بينما تُظهر النماذج المعتمدة على المسافة حساسية أكبر لهذا التقليل مع تحسن ملحوظ في الكفاءة الزمنية.
وأخيرًا، تؤكد الدراسة أن اختيار الميزات لا يُعد مجرد خطوة لتحسين الدقة، بل يمثل عاملًا حاسمًا في تقليل التعقيد الحسابي وتحسين زمن الاستجابة، مما يجعله عنصرًا أساسيًا في تصميم أنظمة كشف التسلل الفعالة في التطبيقات الزمنية الحساسة والبيئات الشبكية واسعة النطاق.

4. المنهجية

تم في هذه الورقة البحثية اتباع المنهجية العامة الموضحة بالشكل رقم 1 والذي يوضح سير العمل الكامل لنظام كشف التسلل (IDS) وتقييم النظام قبل وبعد اختيار الميزات (Feature selection).



شكل 1: المنهجية العامة لسير العمل الكامل لـ (IDS) قبل وبعد اختيار الميزات

4.1- مجموعات البيانات

تم استخدام مجموعتين من البيانات:
NSL-KDD هي مجموعة بيانات معيارية تُستخدم في أبحاث كشف التسلل (IDS) في الشبكات, وتحتوي على خصائص لحركة المرور الشبكية مع تصنيف كل اتصال إلى طبيعي أو هجوم, وتُستخدم كذلك لتقييم ومقارنة خوارزميات التعلم الآلي في مجال الأمن السيبراني.
UNSW-NB15: مجموعة بيانات حديثة تمثل حركة شبكة واقعية تحتوي على أنواع هجمات متقدمة, وتم الحصول على مجموعة بيانات UNSW-NB15 من المصدر الرسمي لجامعة University of New South Wales، وهي مجموعة بيانات حديثة تعكس بيئة شبكية واقعية وتحتوي على أنواع متعددة من الهجمات السيبرانية.

SJST
sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

4.2- الخوارزميات المستخدمة

تم استخدام ثلاث خوارزميات تعلم آلي شائعة في أنظمة كشف التسلل، وهي: Decision Tree(DT)، Random Forest(RF)، وK-Nearest Neighbors (KNN). تم اختيار هذه الخوارزميات لأنها تمثل ثلاث مدارس مختلفة في التعلم الآلي: النماذج الشجرية، النماذج التجميعية، والنماذج المعتمدة على المسافة.

1. شجرة القرار (DT)

تُعد شجرة القرار من خوارزميات التعلم الخاضع للإشراف (Supervised Learning)، وتعمل عن طريق تقسيم البيانات بشكل هرمي إلى عقد (Nodes) بناءً على القيم الأكثر تأثيرًا في التمييز بين الفئات, وتعتبر سهلة الفهم والتفسير ولا تحتاج إلى مقياس موحد للبيانات وسريعة في التدريب.

2. الغابات العشوائية (RF)

تُعتبر Random Forest من خوارزميات التعلم التجميعي (Ensemble Learning)، وتعتمد على بناء مجموعة أشجار قرار باستخدام عينات عشوائية من البيانات (Bootstrapping), و يتم اتخاذ القرار النهائي بناءً على التصويت (Majority Voting), وتعتبر ذات دقة عالية مقارنة بشجرة القرار ومقاومة لمشكلة overfitting وتعمل جيدًا مع البيانات عالية الأبعاد وتُستخدم Random Forest بشكل واسع في كشف التسلل لأنها قادرة على التعامل مع الأنماط المعقدة في بيانات الشبكة وتحقيق توازن بين الدقة والاستقرار.

3. خوارزمية أقرب الجيران (KNN)

تُعد KNN من خوارزميات التعلم البسيط (Lazy Learning)، حيث لا يتم بناء نموذج فعلي أثناء التدريب، بل يتم تخزين البيانات فقط , وهي بسيطة وسهلة التطبيق ولا تحتاج إلى تدريب فعلي وتكون فعالة في البيانات الصغيرة, تُستخدم KNN في تصنيف حركة الشبكة، لكنها تتأثر سلبًا عند استخدام عدد كبير من الميزات، لذلك تستفيد بشكل كبير من اختيار الميزات (Feature Selection).

4.3- اختيار الميزات

تم استخدام اختبار مربع كاي (Chi-square) لاختيار الميزات لأنه يقيّم بكفاءة درجة الاعتماد بين كل ميزة والفئة المستهدفة. كما أنه منخفض التكلفة حسابيًا، وملائم جدًا لمشكلات التصنيف ذات الأبعاد العالية، ويُعد فعالًا بشكل خاص مع مجموعات بيانات كشف التسلل التي تحتوي على سمات فئوية وعددية في آنٍ واحد خاصةً مع بيانات مثل NSL-KDD وUNSW-NB15, وقد تم استخدامه بعد عملية تقسيم البيانات الى مجموعة تدريب ومجموعة اختبار وذلك لضمان عدم حدوث تسرب للبيانات (Data Leakage).

يساعد مربع كاي على حذف الميزات الأقل أهمية قبل تدريب النماذج، مما يؤدي إلى تقليل زمن التدريب , تقليل زمن التنبؤ وتقليل استهلاك الذاكرة.

وقد تم اختيار أفضل 40 ميزة بالنسبة لبيانات NSL-KDD باستخدام اختبار Chi-Square لتحليل أهمية الميزات، وقد تم ترتيب الميزات بناءً على درجة تأثيرها في عملية التصنيف، ثم اختيار الميزات الأعلى أهمية وبناء على مقياس افضل دقة كلية (Accuracy) وافضل مقياس (F1-Score) كما هي موضحة بالجدول1, ويعد الانخفاض البسيط في عدد الميزات بعد تطبيق اختيار الميزات على مجموعة بيانات (NSL-KDD من 43 إلى 40 ميزة) نتيجة منطقية لطبيعة البيانات نفسها، حيث تُعتبر NSL-KDDمجموعة بيانات منظمة ونظيفة نسبيًا، وقد تم تطويرها أصلًا لمعالجة مشكلات التكرار وعدم التوازن الموجودة في مجموع KDD’99. لذلك فإن معظم الميزات الموجودة فيها تمتلك قيمة تمييزية مهمة تساعد في التفريق بين حركة المرور الطبيعية والهجمات الإلكترونية. ولهذا السبب احتفظت خوارزمية اختيار الميزات بمعظم الخصائص، وحذف عدد محدود فقط من الميزات ذات التأثير الضعيف أو المتكرر.

SJST
sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

جدول1: يوضح افضل اختيار للميزات بناء على اختبار مربع كاي لبيانات NSL-KDD

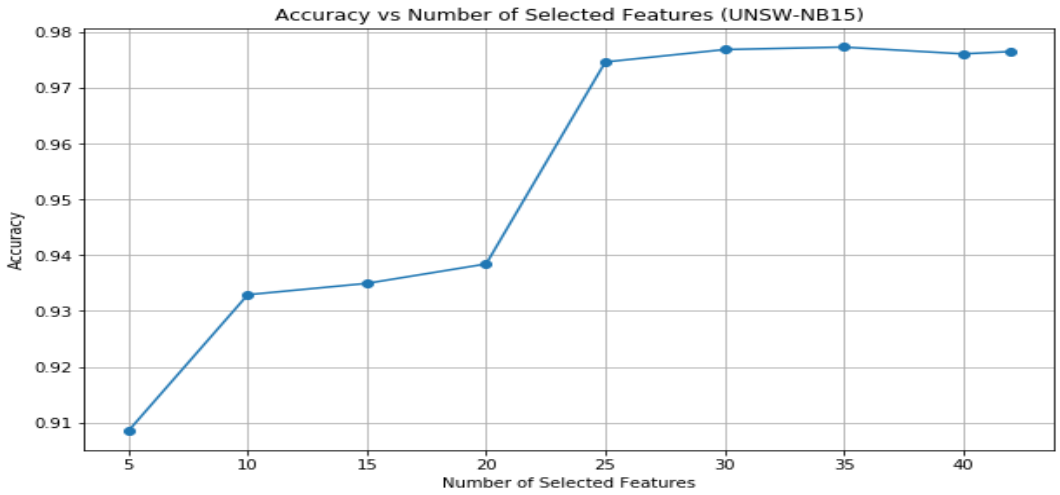
Accuracy	F1-Score	Training time (s)	Inference time (s)	No. of features
0.838934	0.847874	1.796103	0.298017	5
0.971105	0.968864	3.459198	0.406023	10
0.982642	0.981269	3.345191	0.396023	15
0.995290	0.994936	3.787217	0.432024	20
0.996560	0.996301	4.230242	0.409023	25
0.997036	0.996813	3.981228	0.409023	30
0.997486	0.997297	4.008229	0.427024	35
0.997857	**0.997695**	**4.968284**	**0.430025**	**40**
0.997751	0.997581	4.596263	0.408023	41

ونظرًا لاختلاف بنية وخصائص مجموعات البيانات المستخدمة (NSL-KDD وUNSW-NB15)، تم تطبيق عملية اختيار الميزات (Feature Selection) بشكل مستقل على كل مجموعة بيانات.
ولذلك تم تطبيق عملية اختيار الميزات بشكل مستقل على مجموعة بيانات UNSW-NB15 باستخدام نفس المنهجية المعتمدة على اختبار Chi-Square. وبعد تقييم أهمية الميزات، تم اختيار أفضل 35 ميزة تمثل الخصائص الأكثر تأثيرًا في عملية التصنيف وبناء على مقياس افضل دقة كلية (Accuracy) وافضل مقياس (F1-Score) كما هو مبين بالجدول2 والشكل2.

جدول2: يوضح افضل اختيار للميزات بناء على اختبار مربع كاي لبيانات UNSW-NB15

Accuracy	F1-Score	Training time (s)	Inference time (s)	No. of features
0.90854	0.91572	4.96081	0.468	5
0.93291	0.93799	5.69401	0.468	10
0.93494	0.93984	5.21041	0.468	15
0.93842	0.94322	4.77361	0.468	20
0.97462	0.97683	5.60041	0.468	25
0.97684	0.97888	4.75801	0.468	30
0.97729	**0.97928**	**4.61761**	**0.468**	**35**
0.97607	0.97818	5.81881	0.468	40
0.97648	0.97855	5.41321	0.468	42

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198



شكل 2: اختيار الميزات بناء على أعلى Accuracy

وقد تم استخدام هذه الميزات في تدريب النماذج ومقارنتها مع الأداء باستخدام جميع الميزات الأصلية، وذلك لتقييم تأثير تقليل الأبعاد على الدقة وزمن الاستدلال.

يُلاحظ أن النسبة المئوية لتقليل الميزات اختلفت بين مجموعتي البيانات (من 43 إلى 40 في NSL-KDD، ومن 45 إلى 35 في UNSW-NB15). يعود ذلك إلى طبيعة البيانات؛ فمجموعة NSL-KDD قديمة وتم بالفعل تنقيتها واختيار ميزاتها بعناية عند إنشائها، مما يجعل أغلب ميزاتها ذات أهمية فعلية. أما مجموعة بيانات UNSW-NB15 فهي تحتوي على ميزات حديثة أكثر تعقيداً يصاحبها قدر كبير من التكرار، مما سمح بتقليل أكبر. يثبت هذا أن المنهجية المقترحة تتكيف مع طبيعة البيانات بدلاً من حذف الميزات بشكل عشوائي ثابت.

وبالرغم من اختلاف الميزات بين مجموعتي البيانات، إلا أن كلا المجموعتين تحتوي على خصائص تعبر عن سلوك الشبكة مثل حجم البيانات ومعدلات الاتصال، مما يسمح بإجراء مقارنة عادلة بين أداء النماذج.

4.4- النتائج وتحليل المقارنة (Results & Comparison)

تم تنفيذ التجارب وفق السيناريو التالي لكل مجموعة بيانات:

1- إستخدام بيانات NSL-KDD

تم تحميل بيانات NSL-KDD فكان عدد العينات الكلي 125973 عينة وتم تقسيم هذه البيانات بنسبة 70% من اجمالي العينات لغرض التدريب و 30% لغرض الاختبار وبذلك يكون عدد عينات التدريب والاختبار في هذه البيانات هي:

عدد عينات التدريب : 88181 وعدد عينات الاختبار : 37792

وتم تطبيقهما على مرحلتين وهما:-

المرحلة الاولى: استخدام جميع الميزات وعددها 43 ميزة (Baseline Model) وكانت النتائج كما هي مبينة بالجدول3

جدول3 : نتائج تقييم النماذج الثلاثة قبل اختيار الميزات

Model	Accuracy	Precision	Recall	F1-Score	Training Time	Inference Time
RF	0.99672	0.99693	0.99602	0.99647	1.43108	0.02500
DT	0.99775	0.99858	0.99659	0.99758	4.42825	0.42002
KNN	0.99495	0.99426	0.99488	0.99457	5.24330	36.25107

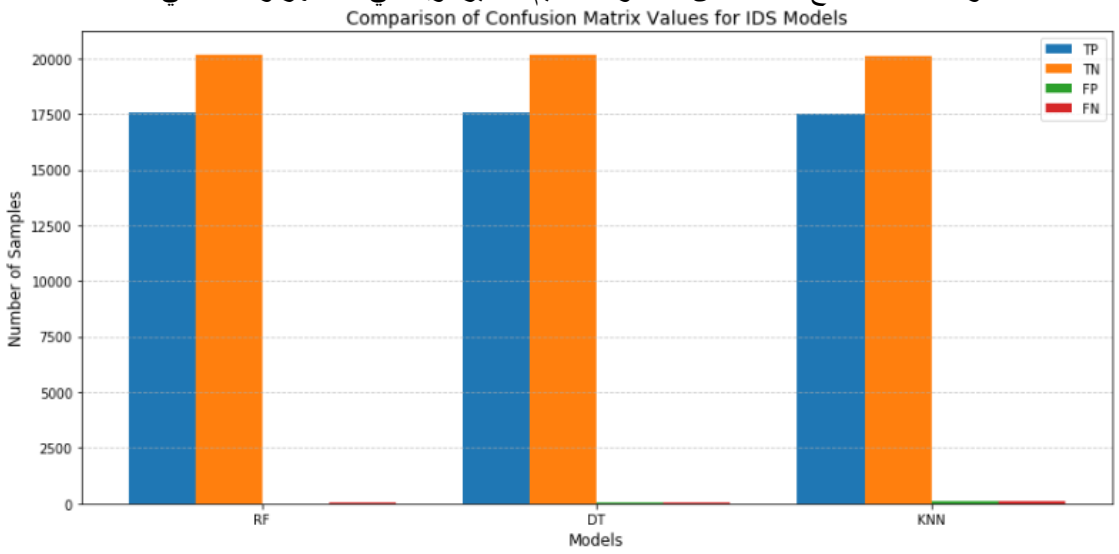
sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

1- الغابات العشوائية (RF)
Confusion Matrix
[101 20102]
[17560 29]

2- شجرة القرار (DT)
:Confusion Matrix
[11 20192]
[17549 40]

3- أقرب الجيران (KNN)
Confusion Matrix
[32 20171]
[17499 90]

ويمثل هذه المصفوفات الشكل3 مع ملاحظة ان FP و FN قيم صغيرة وبالتالي لا تظهر واضحة في الشكل.



شكل 3: يمثل مصفوفات الارتباك للنماذج الثلاثة قبل اختيار الميزات بيانات NSL-KDD

المرحلة الثانية: استخدام الميزات المختارة وعددها 40 ميزة (Optimized Model) وكانت النتائج كما هي مبينة بالجدول4

جدول4 : نتائج تقييم النماذج الثلاثة بعد اختيار الميزات

Model	Accuracy	Precision	Recall	F1-Score	Training Time (s)	Inference Time (s)
RF	0.99786	0.99858	0.99682	0.99770	5.49131	0.40602
DT	0.99669	0.99693	0.99596	0.99644	1.87711	0.02200
KNN	0.99219	0.99125	0.99198	0.99162	2.00611	7.84145

1- الغابات العشوائية
Confusion Matrix
[25 20178]
[17533 56]

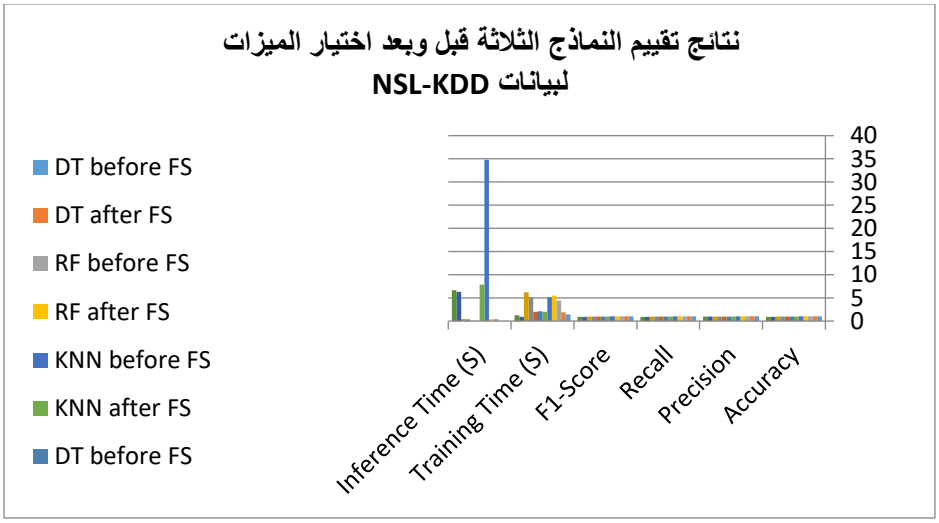
2- شجرة القرار
Confusion Matrix
[54 20149]
[17518 71]

3- أقرب الجيران
Confusion Matrix
[101 20102]
[17499 90]

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

جدول5 : يوضح نتائج تقييم النماذج الثلاثة قبل وبعد اختيار الميزات.

Dataset	Model	Stage	Accuracy	Precision	Recall	F1-Score	Training Time SEC	Inference Time SEC
NSL-KDD	DT	Before FS	0.99672	0.99693	0.99602	0.99647	1.43108	0.02500
NSL-KDD	DT	After FS	0.99669	0.99693	0.99596	0.99644	1.87711	0.02200
NSL-KDD	RF	Before FS	0.99775	0.99858	0.99659	0.99758	4.42825	0.42002
NSL-KDD	RF	After FS	0.99786	0.99858	0.99682	0.99770	5.49131	0.40602
NSL-KDD	KNN	Before FS	0.99495	0.99426	0.99488	0.99457	5.11681	34.77186
NSL-KDD	KNN	After FS	0.99219	0.99125	0.99198	0.99162	2.00611	7.84145



شكل 4: مقارنة بين النماذج الثلاثة قبل وبعد اختيار الميزات لبيانات NSL-KDD

2- استخدام بيانات UNSW-NB15

تم تحميل بيانات UNSW-NB15 فكان عدد العينات الكلي 82332 عينة وتم تقسيم هذه البيانات بنسبة 70% من اجمالي العينات لغرض التدريب و 30% لغرض الاختبار وبذلك يكون عدد عينات التدريب والاختبار في هذه البيانات هي:

عدد عينات التدريب : 57632 وعدد عينات الاختبار : 24700

وتم تطبيقهما على مرحلتين وبنفس المنهجية التي اتبعت مع بيانات NSL-KDD وهما:-

المرحلة الاولى: استخدام جميع الميزات وعددها 45 ميزة (Baseline Model) وكانت النتائج كما هي مبينة بالجدول6

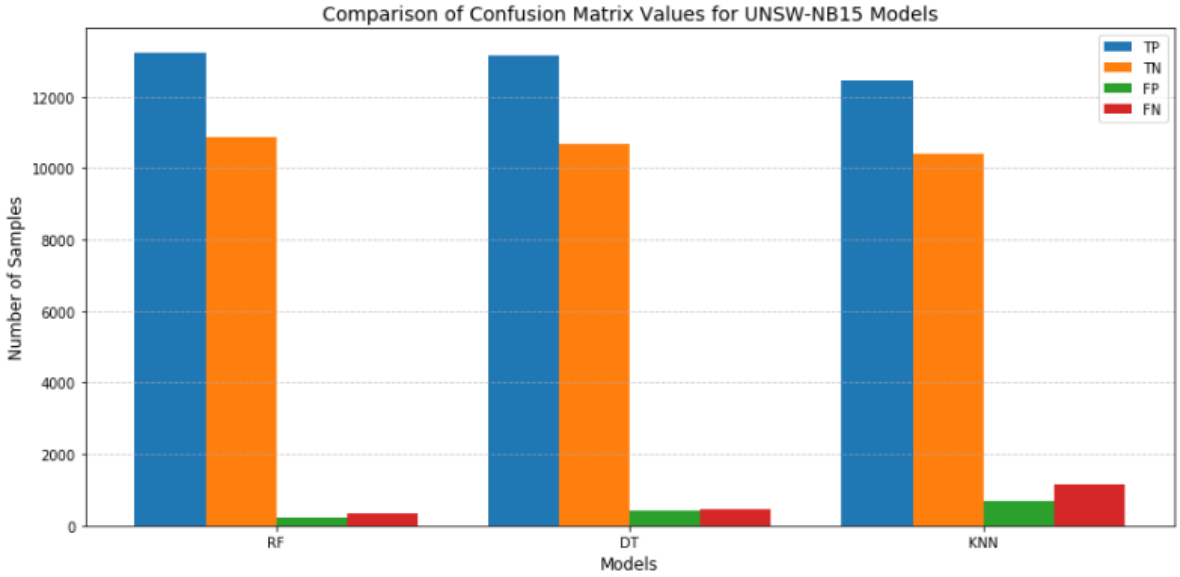
جدول6 : نتائج تقييم النماذج الثلاثة قبل اختيار الميزات

Model	Accuracy	Precision	Recall	F1-Score	Training Time	Inference Time
RF	0.97636	0.98165	0.97529	0.97846	5.09029	0.40702
DT	0.96308	0.96743	0.96544	0.96644	2.08812	0.02500
KNN	0.91130	0.94115	0.89485	0.91742	0.92405	6.33936

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

1- الغابات العشوائية	2- شجرة القرار	3- أقرب الجيران
Confusion Matrix	Confusion Matrix	Confusion Matrix
[234 10866]	[433 10667]	[704 10396]
[13250 350]	[13155 445]	[12465 1135]

ويمثل هذه المصفوفات الشكل5 مع ملاحظة ان FP و FN قيم صغيرة نسبياً وبالتالي وهي تظهر بوضح في الشكل5.



شكل 5: يمثل مصفوفات الارتباك للنماذج الثلاثة قبل اختيار الميزات بيانات UNSW-NB15

المرحلة الثانية: اد تطبيق اختبار مربع كاي (Chi-Square) لترتيب الميزات وفقاً لمدى ارتباطها بالمتغير الهدف. وبعد ترتيب الميزات تنازلياً، تم اختيار أفضل 35 ميزة ذات أعلى القيم.

وقد تم تحديد القيمة الحدية الفعلية لاختيار الميزات عند ($\chi^2 \geq 9.03$)، والتي تمثل أصغر قيمة لمربع كاي ضمن مجموعة الميزات المختارة، مما يضمن الاحتفاظ بالميزات الأكثر تأثيراً على عملية التصنيف.

جدول7 : نتائج تقييم النماذج الثلاثة بعد اختيار الميزات

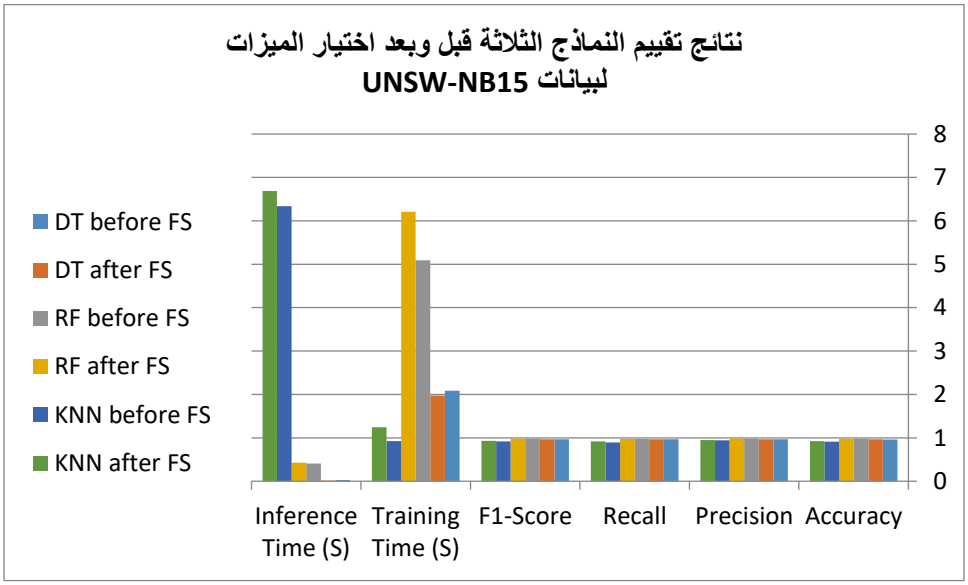
Model	Accuracy	Precision	Recall	F1-Score	Training Time	Inference Time
RF	0.97636	0.98265	0.97426	0.97844	6.20635	0.42602
DT	0.96445	0.96813	0.96728	0.96771	1.97311	0.01400
KNN	0.92555	0.94654	0.91654	0.93130	1.24507	6.68538

1- الغابات العشوائية	2- شجرة القرار	3- أقرب الجيران
Confusion Matrix	Confusion Matrix	Confusion Matrix
[234 10866]	[433 10667]	[573 10527]
[13250 350]	[13155 445]	[12435 1165]

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

جدول8 : يوضح نتائج تقييم النماذج الثلاثة قبل وبعد اختيار الميزات

Dataset	Model	Stage	Accuracy	Precision	Recall	F1-Score	Training Time (S)	Inference Time (S)
UNSW-NB15	DT	Before FS	0.96308	0.96743	0.96544	0.96644	2.08812	0.02500
UNSW-NB15	DT	After FS	0.96445	0.96813	0.96728	0.96771	1.97311	0.01400
UNSW-NB15	RF	Before FS	0.97636	0.98165	0.97529	0.97846	5.09029	0.40702
UNSW-NB15	RF	After FS	0.97636	0.98265	0.97426	0.97844	6.20635	0.42602
UNSW-NB15	KNN	Before FS	0.91130	0.94115	0.89485	0.91742	0.92405	6.33936
UNSW-NB15	KNN	After FS	0.92555	0.94654	0.91654	0.93130	1.24507	6.68538



شكل 6: مقارنة بين النماذج الثلاثة قبل وبعد اختيار الميزات لبيانات UNSW-NB15

5. النتائج والمناقشة

تم تنفيذ التجارب على مجموعتي البيانات NSL-KDD و UNSW-NB15بهدف تقييم تأثير اختيار الميزات (Feature Selection)على أداء خوارزميات التعلم الآلي المستخدمة في أنظمة كشف التسلل (IDS). وقد شملت التجارب ثلاث خوارزميات رئيسية وهي Decision Tree (DT)، والغابات العشوائية Random Forest (RF)، واقرب الجيران K-Nearest Neighbors (KNN). وتم تقييم النماذج قبل وبعد تطبيق اختيار الميزات باستخدام مقاييس Accuracy و Precision و Recall وF1-Score، بالإضافة إلى زمن التدريب (Training Time) وزمن الاستدلال (Inference Time).

أظهرت النتائج أن تأثير اختيار الميزات يختلف باختلاف طبيعة الخوارزمية ومجموعة البيانات المستخدمة، إلا أن جميع النماذج استفادت بدرجات متفاوتة من تقليل عدد الميزات، سواء من ناحية الكفاءة الحسابية أو الحفاظ على استقرار الأداء.

أولاً: نتائج مجموعة بيانات NSL-KDD

أظهرت نتائج مجموعة بيانات NSL-KDD أن خوارزمية Random Forest حققت أعلى أداء بين جميع النماذج، حيث ارتفعت الدقة من 99.775% قبل اختيار الميزات إلى 99.786% بعد اختيار الميزات، كما ارتفع مقياس F1-Score من

SJST
sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

99.758% إلى 99.770%. وفي الوقت نفسه، انخفض زمن الاستدلال من 0.420 ثانية إلى 0.406 ثانية، مما يشير إلى أن تقليل عدد الميزات ساهم في تحسين الكفاءة دون التأثير سلبًا على دقة التصنيف.
أما خوارزمية Decision Tree فقد حافظت على أداء مستقر للغاية بعد اختيار الميزات، حيث بقيت الدقة قريبة جدًا من قيمتها الأصلية مع انخفاض طفيف جدًا لا يُعد مؤثرًا عمليًا، في حين انخفض زمن الاستدلال من 0.025 ثانية إلى 0.022 ثانية، مما يدل على أن اختيار الميزات ساهم في تحسين سرعة النموذج مع الحفاظ على استقراره.
بالنسبة لخوارزمية KNN ، فقد كان تأثير اختيار الميزات أكثر وضوحًا من ناحية الكفاءة الحسابية، حيث انخفض زمن التدريب من 5.116 ثانية إلى 2.006 ثانية، كما انخفض زمن الاستدلال بشكل كبير من 34.77 ثانية إلى 7.84 ثانية. ورغم وجود انخفاض طفيف في الدقة من 99.495% إلى 99.219%، إلا أن التحسن الكبير في السرعة يُعد مهمًا جدًا في تطبيقات كشف التسلل الزمن الحقيقي، خاصةً أن خوارزمية KNN تعتمد بشكل مباشر على حساب المسافات بين العينات، وبالتالي تتأثر بشكل كبير بعدد الميزات المستخدمة.

تشير هذه النتائج إلى أن تقليل الأبعاد ساهم في تخفيض التعقيد الحسابي بشكل واضح، خصوصًا في الخوارزميات المعتمدة على المسافة مثلKNN ، بينما حافظت النماذج التجميعية مثل Random Forest على استقرارها العالي وقدرتها على التعميم.

ثانيًا: نتائج مجموعة بيانات UNSW-NB15

أظهرت نتائج مجموعة بيانات UNSW-NB15 سلوكًا مختلفًا نسبيًا بسبب الطبيعة الأكثر تعقيدًا وحداثة لهذه البيانات مقارنةً بـ NSL-KDD.وقد حققت خوارزمية Random Forest أفضل أداء أيضًا، حيث حافظت على دقة مرتفعة بلغت 97.636% قبل وبعد اختيار الميزات تقريبًا، مع استقرار واضح في قيم Precision و Recall و F1-Score. ورغم وجود زيادة طفيفة في زمن التدريب والاستدلال بعد اختيار الميزات، إلا أن الأداء العام بقي مستقرًا بدرجة عالية، مما يعكس قدرة Random Forest على التعامل مع البيانات عالية الأبعاد حتى بعد تقليل بعض الميزات.
أما خوارزمية Decision Tree فقد استفادت بشكل إيجابي من اختيار الميزات، حيث ارتفعت الدقة من 96.308% إلى 96.445%، كما تحسن مقياس F1-Score من 96.644% إلى 96.771%. بالإضافة إلى ذلك، انخفض زمن التدريب من 2.088 ثانية إلى 1.973 ثانية، وانخفض زمن الاستدلال من 0.025 ثانية إلى 0.014 ثانية، مما يعكس تحسنًا متوازنًا في كل من الأداء والكفاءة.
في المقابل، أظهرت خوارزمية KNN أكبر استفادة من عملية اختيار الميزات على مجموعة بيانات UNSW-NB15 ، حيث ارتفعت الدقة من 91.130% إلى 92.555%، كما تحسن مقياس F1-Score من 91.742% إلى 93.130%. الا ان التدريب ارتفع قليلا من 0.924 ثانية إلى 1.245 ثانية مع بقاء زمن الاستدلال ضمن نطاق مقبول نسبيًا مقارنة بطبيعة الخوارزمية. ويؤكد ذلك أن تقليل الميزات ساهم في إزالة الخصائص الأقل أهمية والتي كانت تؤثر سلبًا على حساب المسافات داخل نموذج.KNN

المقارنة العامة بين مجموعتي البيانات

من خلال النتائج، يمكن ملاحظة أن أداء النماذج على مجموعة بيانات NSL-KDD كان أعلى بشكل عام مقارنةً بمجموعة بيانات UNSW-NB15. ويعود ذلك إلى أن NSL-KDD تُعد مجموعة بيانات أكثر تنظيمًا وأقل تعقيدًا، بينما تمثل UNSW-NB15بيئة شبكية حديثة وأكثر واقعية تحتوي على هجمات معقدة وأنماط حركة مرور متنوعة، مما يجعل عملية التصنيف أكثر صعوبة.
كما أظهرت النتائج أن تأثير اختيار الميزات لا يقتصر فقط على تحسين الدقة، بل يمتد أيضًا إلى تقليل التعقيد الحسابي وتحسين سرعة الاستجابة، وهو عامل بالغ الأهمية في أنظمة كشف التسلل الزمن الحقيقي. وقد كان هذا التأثير أكثر وضوحًا في خوارزمية KNNبسبب اعتمادها المباشر على حساب المسافات، في حين أظهرت خوارزمية Random Forest قدرة عالية على الحفاظ على الاستقرار حتى بعد تقليل عدد الميزات.
بشكل عام، تؤكد النتائج أن اختيار الميزات يمثل خطوة أساسية لتحسين كفاءة أنظمة كشف التسلل المعتمدة على التعلم الآلي، حيث يساهم في تقليل الأبعاد وتحسين الأداء وتقليل الزمن الحسابي دون التأثير السلبي على دقة التصنيف، كما أن استخدام أكثر من مجموعة بيانات يعزز من موثوقية النتائج وقابليتها للتعميم في البيئات الشبكية المختلفة.

SJST
sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol 8, No.1. Jan. – May. 2026
Pages: 185 ~ 198

6. الخاتمة

تقدم هذه الدراسة تحليلًا شاملًا لتأثير اختيار الميزات على أداء أنظمة كشف التسلل المعتمدة على التعلم الآلي باستخدام مجموعتي بيانات معياريتين هما NSL-KDD وUNSW-NB15، وذلك عبر ثلاث خوارزميات مختلفة تمثل نماذج شجرية وتجميعية وتعتمد على المسافة.

أظهرت النتائج أن اختيار الميزات يمثل خطوة حاسمة في تحسين كفاءة أنظمة كشف التسلل، حيث أسهم في تقليل الأبعاد بشكل فعّال مع الحفاظ على مستويات عالية من الدقة، بل وتحسين الأداء في بعض الحالات. كما تبين أن الأثر الأكبر كان على الكفاءة الزمنية، خاصة في الخوارزميات المعتمدة على المسافة مثل KNN ، بينما حافظت النماذج التجميعية مثل Random Forest على استقرار عالٍ في الأداء قبل وبعد تقليل الميزات.

كذلك أكدت النتائج أن استخدام أكثر من مجموعة بيانات يعزز من موثوقية التقييم ويكشف اختلاف استجابة النماذج حسب طبيعة البيانات، حيث أظهرت UNSW-NB15 تحديًا أكبر مقارنة بـ NSL-KDD بسبب تعقيدها وواقعيتها.

بشكل عام، تثبت هذه الدراسة أن اختيار الميزات ليس مجرد خطوة لتحسين الدقة، بل عنصر أساسي في بناء أنظمة كشف تسلل فعالة وقابلة للتطبيق في البيئات الزمنية الحقيقية، حيث يجمع بين تحسين الأداء وتقليل التعقيد الحسابي وتعزيز قابلية التوسع.

وتقترح الدراسة مستقبلًا التوجه نحو دمج تقنيات اختيار ميزات أكثر تقدمًا مثل الأساليب متعددة الأهداف (Multi-objective) أو الأساليب القائمة على التعلم العميق لتحسين القدرة على التعميم في بيئات الشبكات الحديثة

المراجع

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

[2] R. Lippmann et al., “The 1999 DARPA off-line intrusion detection evaluation,” Computer Networks, vol. 34, no. 4, pp. 579–595, 2000.

[3] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[4] S. Chandrasekaran and R. Balasubramanian, “Performance evaluation of Decision Tree, Random Forest, and KNN for intrusion detection,” International Journal of Advanced Research in Computer Science, vol. 12, no. 6, 2021.

[5] S. Shafiq, K. Khurshid, and R. Akbar, “Optimized feature selection approach for machine learning-based intrusion detection systems,” Applied Computing and Informatics, 2020.

[6] S. M. Kasongo and Y. Sun, “Performance analysis of intrusion detection systems using feature selection method on UNSW-NB15 dataset,” Journal of Big Data, Springer, 2020.

[7] Y. Yin et al., “IGRF-RFE: A hybrid feature selection method for intrusion detection on UNSW-NB15,” Journal of Big Data, Springer, 2022.

[8] S. Dhal and C. Azad, “A comprehensive survey on feature selection in machine learning,” Applied Intelligence, 2021.

[9] S. Ali et al., “Machine learning-based intrusion detection using ensemble feature selection,” IEEE Access, 2023.

[10] M. Alqahtani et al., “Comparative analysis of intrusion detection systems using NSL-KDD and UNSW-NB15 datasets,” Computers & Security, Elsevier, 2023.

[11] M. A. Khan et al., “Lightweight intrusion detection system using feature selection for IoT networks,” Future Generation Computer Systems, 2024.

[12] Z. H. Cheng et al., “Multi-objective feature selection for intrusion detection systems,” arXiv preprint, 2024.

[13] M. Jouhari et al., “Efficient intrusion detection using Chi-square feature selection with deep learning,” arXiv preprint, 2024.
[14] E. Emirmahmutoğlu and Y. Atay, “Feature selection-driven machine learning framework for intrusion detection systems,” Peer-to-Peer Networking and Applications, Springer, 2025.
[15] H. Zhang et al., “Hybrid feature selection for network anomaly detection,” IEEE Transactions on Network Science and Engineering, 2023.
[16] Y. Saheed et al., “Hybrid autoencoder and PSO feature selection for intrusion detection systems,” Frontiers in Computer Science, 2023.