Syrian Arab Republic

Ministry of Higher Education and Scientific Research

Syrian Virtual University

الجمهورية العربية السورية

وزارة التعلم العالي والبحث العلمي

الجامعة الافتراضية السورية

اكتشاف الشذوذ في مجموعة البيانات الشبكية باستخدام تقنيات التعلم الآلي

Anomaly Detection In Network Dataset Using Machine Learning Techniques

تقديم:

حسين ابراهيم علي

إشراف

د.أكرم مرعي

بحث مقدم لنيل درجة الماجستير في الهندسة المعلوماتية

اكتشاف الشذوذ في مجموعة البيانات الشبكية باستخدام تقنيات التعلم الآلي

Anomaly Detection In Network Dataset Using Machine Learning Techniques

فهرس المحتويات

فهرس الاشكال

الملخص

إن اكتشاف الشذوذ في بيانات الشبكة باستخدام التعلم الآلي هو تقنية قوية لتحديد الأنماط أو السلوكيات غير العادية في حركة مرور الشبكة التي تنحرف عن المعايير المتوقعة. يستفيد هذا النهج من منهجيات التعلم الآلي المختلفة لتحليل تدفقات البيانات مما يتيح أنظمة الإنذار المبكر للبنية التحتية الحرجة وتعزيز أمان الشبكة. يمكن للطرق غير الخاضعة للإشراف مثل التجميع والخوارزميات القائمة على الكثافة تحديد الشذوذ دون بيانات مصنفة، بينما يمكن تدريب الأساليب الخاضعة للإشراف التي تنطوي على خوارزميات التصنيف ونماذج المجموعة والتعلم العميق على البيانات المصنفة لتصنيف حركة المرور على أنها طبيعية أو شاذة. ومع ذلك، لا تزال التحديات قائمة في اكتشاف الميزات الأساسية التي تؤثر في دقة اكتشاف الشذوذ والزمن اللازم لاكتشاف النقاط الشاذة. وعلى الرغم من هذه التحديات، يظل اكتشاف الشذوذ باستخدام التعلم الآلي عنصرًا أساسيًا في استراتيجيات أمن الشبكات وإدارتها الحديثة، ويتطور باستمرار لمعالجة التهديدات الناشئة وبيئات الشبكات المعقدة.

لذلك تم في هذا البحث دراسة تأثير الميزات على خوارزميات التعلم الالي وذلك باستخدام ثلاث خوارزميات لاختيار الميزات وهي Feature Importance, correlation, chi-square وثلاث مصنفات تعلم آلي وهي شجرة القرار و الغابة العشوائية وخوارزمية اقرب جار و وقد أظهرت النتائج ان الميزات التي تم اختيارها باستخدام خوارزمية FI هي الأفضل من باقي الخوارزميات حيث كان مصنف الغابة العشوائية وشجرة القرار هما الأفضل من باقي المصنفات بدرجة f1 وصلت ل 1.

الكلمات المفتاحية: اكتشاف الشذوذ، بيانات الشبكة، التعلم الالي، هندسة الميزات، تصنيف البيانات.

Abstract

Anomaly detection in network data using machine learning is a powerful technique for identifying unusual patterns or behaviors in network traffic that deviate from expected norms. This approach leverages various machine learning methodologies to analyze data streams, enabling early warning systems for critical infrastructure and enhancing network security. Unsupervised methods such as clustering and density-based algorithms can identify anomalies without labeled data, while supervised methods involving classification algorithms, ensemble models, and deep learning can be trained on labeled data to classify traffic as normal or anomalous. However, challenges remain in discovering the underlying features that affect the accuracy of anomaly detection and the time required to detect anomalies. Despite these challenges, anomaly detection using machine learning remains a core component of modern network security and management strategies, and is constantly evolving to address emerging threats and complex network environments. Therefore, in this research, the effect of features on machine learning algorithms was studied using three feature selection algorithms: Feature Importance, correlation, chi-square, and three machine learning classifiers: decision tree, random forest, and nearest neighbor algorithm. The results showed that the features selected using the FI algorithm are the best of the rest of the algorithms, as the random forest and decision tree classifier were the best of the rest of the classifiers with an f1 score of 1.

Keywords: Anomaly detection, network data, machine learning, feature engineering, data classification.

مقدمة

في عالمنا الرقمي السريع الخطى اليوم، أصبح حماية أمن وسلامة بيانات الشبكة أمرًا بالغ الأهمية. لقد أدى الارتفاع المفاجئ في التهديدات السيبرانية والتعقيد المتزايد للأنشطة الخبيثة إلى جعل تدابير الأمن التقليدية غير كافية بمفردها. وقد دفع هذا الإدراك إلى استكشاف التقنيات المتطورة، وخاصة اكتشاف الشذوذ، والتي يمكنها تحديد المخالفات أو الانحرافات عن السلوك القياسي داخل تدفقات بيانات الشبكة.

يستخدم اكتشاف الشذوذ خوارزميات التعلم الآلي للتمييز بين السلوكيات الطبيعية والشاذة، وبالتالي تمكين استراتيجيات الدفاع الاستباقية ضد نقاط الضعف الأمنية المحتملة. يسمح هذا النهج بتحديد الأنماط الدقيقة التي قد تفلت من الأساليب التقليدية، مما يوفر إطارًا قويًا لتحليل بيانات الشبكة.

من خلال تدريب النماذج على كميات هائلة من بيانات حركة المرور على الشبكة، يمكن لهذه الخوارزميات استيعاب السلوك النموذجي لأنظمة الشبكة في ظل الظروف العادية. عندما تنحرف البيانات التي تم مواجهتها حديثًا بشكل كبير عن هذه القاعدة الراسخة، يتم وضع علامة عليها باعتبارها شذوذًا محتملاً، مما يشير إلى محاولة اقتحام أو خلل في النظام.

تتعمق هذه الدراسة في تطبيق التعلم الآلي للكشف عن الشذوذ في بيانات الشبكة. وتفحص منهجيات التعلم الآلي المختلفة، التي تشمل كل من الأساليب الخاضعة للإشراف وغير الخاضعة للإشراف، وتناقش كيف يمكن تخصيصها لتناسب بيئات الشبكة المحددة وملفات تعريف التهديد. ومن خلال هذا الاستكشاف، يمكن لمسؤولي الشبكات ومتخصصي الأمن السيبراني تعزيز قدراتهم على حماية شبكاتهم بشكل استباقي ضد التهديدات الناشئة.

ومع استمرار التقدم التكنولوجي في إعادة تشكيل المشهد الرقمي لدينا، فإن أهمية التعلم الآلي في تعزيز البنية التحتية الرقمية تنمو بشكل كبير. ومن خلال تبني هذه التقنيات المبتكرة، يمكن لأصحاب المصلحة

في جميع أنحاء طيف الأمن السيبراني تعزيز دفاعاتهم، وإنشاء نظام بيئي رقمي أكثر مرونة وقدرة على التكيف.

مشكلة البحث

في مجال شبكات الكمبيوتر، يمكن تحديد المشكلات كمشاهدات شاذة في حركة البيانات، حيث يتعارض السلوك غير المعتاد مع التوقعات الطبيعية. على سبيل المثال، قد يؤدي عطل جهاز شبكي إلى توليد حركة مرور غير متوقعة في مناطق أخرى من الشبكة. استخدام خوارزميات لاكتشاف الشذوذ في هذا السياق يوفر طريقة فعالة لتحديد هذه الحالات الشاذة.

مع ذلك، يوجد تحدي في اختيار الميزات المناسبة لتدريب النماذج، حيث يمكن أن يكون اختيار جميع الميزات متاحًا، ولكنه يعيق سرعة التدريب وتأثر على دقة النموذج الناتج. البحث عن طرق لاختيار الميزات بشكل أكثر فعالية يظل موضوعًا هامًا، حيث يمكن أن تساهم في تحسين سرعة التدريب وتحسين دقة التصنيف للنموذج.

على الرغم من التقدم في استخدام طرائق لاختيار الميزات، فإن تحدي كبير يكمن في القدرة على تحديد الميزات الأكثر فائدة التي تساهم بشكل مباشر في التصنيف. هذه الميزات ليست دائمًا واضحة، وقد تتطلب دراسة متعمقة للبيانات ورصد السلوك الشائع لتحديدها بشكل صحيح.

هدف البحث

يهدف البحث الى اكتشاف الشذوذ في البيانات الشبكية عن طريق اختيار الميزات الأكثر الأهمية لتدريب مصنفات اكتشاف الشذوذ من خلال تقييم فعالية خوارزميات اختيار الميزات على هذه المصنفات.

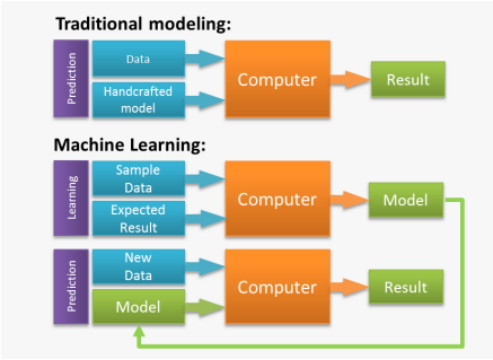
الفصل الأول الدراسة النظرية

1.1 مقدمة

يتضمن هذا الفصل شرح لأهم المفاهيم النظرية التي تم ذكرها ضمن هذا البحث، كالتعلّم الآلي والشذوذ في البيانات الشبكية وطرائق اكتشافه.

1.2 التعلم الآلي

التعلم الآلي هو مجال ضمن الذكاء الاصطناعي وعلوم الكمبيوتر، يهدف إلى استغلال البيانات والخوارزميات لتمتلك القدرة على التعلم والتطور بنفس الطريقة التي يفعلها البشر، مع التحسن المستمر في الدقة. هذا المجال يلعب دورًا بارزًا في تطوير علوم البيانات، حيث يستخدم الأساليب الإحصائية لتدريب الخوارزميات على القيام بتصنيفات أو تنبؤات في عمليات البحث عن البيانات. هذه العملية تساعد في اتخاذ قرارات أكثر فعالية داخل التطبيقات والشركات. بينما يعتمد التعلم الآلي على البيانات كمصدر رئيسي للتعلم دون الحاجة إلى برمجة محددة بالقواعد، يختلف عن النهج التقليدي في البرمجة. في البرمجة التقليدية، يتم تقديم البيانات والبرنامج كمعلمات للجهاز، والذي يقوم بإنتاج نتائج محددة. ومع ذلك، في التعلم الآلي، تكون البيانات المدخلة هي العنصر الرئيسي، ويتم تعلم البرنامج نفسه من خلال هذه البيانات لتنفيذ وظائف جديدة وتوقع نتائج لمشكلات غير معروفة[1]. يوضح الشكل 1-1 الفرق الرئيسي بين النهج التقليدي والنهج الذي يتبعه التعلم الآلي.



الشكل 1-1:الفرق بين بين النهج التقليدي ونهج التعلم الآلي.

المصدر [2]

1.3 تحديات التعلم الالي

التعلم الآلي يواجه العديد من التحديات، منها[3] :

1. عدم توفر كمية كافية من البيانات للتدريب: غالبًا ما يتطلب التعلم الآلي الكثير من البيانات لتشغيله بكفاءة. حتى للأمثلة البسيطة، قد تكون هناك حاجة إلى آلاف الأمثلة، وفي حالات معقدة مثل التعرف على الصور أو الكلام، قد يكون هناك حاجة إلى ملايين الأمثلة.
2. جودة البيانات السيئة: إذا كانت بيانات التدريب مليئة بالأخطاء، القيم المتطرفة، والضوضاء، فإن هذا يجعل من الصعب على النظام اكتشف الأنماط الأساسية، مما يجعله من الصعب على النموذج أن يعمل بشكل جيد. يشير العلماء إلى أن جزء كبير من الوقت يتم استخدامه في تنظيف البيانات. مثلاً، إذا كانت بعض الأمثلة متطرفة بشكل واضح، قد يساعد تجاهلها أو محاولة إصلاح الأخطاء يدويًا. وفي حالة عدم وجود بعض الميزات في بعض الحالات، يجب اتخاذ قرار بشأن كيفية التعامل مع هذه الحالات، سواء كان تجاهلها أو تعبئتها بالقيم المفقودة أو تدريب نماذج مختلفة بناءً على هذه القرارات.
3. وجود ميزات غير ذات صلة: النموذج يمكن أن يتعلم فقط إذا كانت بيانات التدريب تحتوي على ميزات ذات صلة كافية ولا تحتوي على الكثير من الميزات غير ذات الصلة. جزء مهم من نجاح مشروع التعلم الآلي هو اختيار مجموعة جيدة من الميزات للتدريب عليها، وهو ما يتضمن اختيار الميزات الأكثر فائدة، واستخراج الميزات من الميزات الموجودة، وخلق ميزات جديدة من خلال جمع بيانات جديدة.
4. فرط الملاءمة (Overfitting) يحدث عندما يعمل النموذج بشكل جيد على بيانات التدريب ولكنه لا يعمل بشكل جيد على البيانات الخارجية. هذا يمكن أن يحدث عندما يكون النموذج معقدًا جدًا

بالنسبة إلى كمية وضوضاء في بيانات التدريب. الحلول الممكنة تشمل جمع المزيد من بيانات التدريب أو تقليل الضوضاء في بيانات التدريب.

5. الملاءمة الناقصة (Underfitting) يحدث عندما يكون النموذج بسيطًا جدًا لتعلم البنية الأساسية للبيانات. الحلول الرئيسية لإصلاح هذه المشكلة تشمل استخدام نموذج أكثر قوة مع مزيد من الميزات أو تحسين نوعية الميزات التي يتم تقديمها للخوارزمية.

1.4 أنواع التعلم الآلي

التعلم الآلي يمكن تقسيمه إلى أربعة أنواع أساسية بناءً على طريقة التعلم [2]:

التعلم الآلي الخاضع للإشراف: يتم فيه تدريب الآلات باستخدام بيانات "معنونة"، حيث يتم تقديم مجموعة من البيانات المرتبطة بحالات معينة، ومن ثم تستخدم هذه البيانات لتدريب الآلة على التنبؤ أو التصنيف. هذا النوع من التعلم يتم التركيز عليه بشكل خاص في العديد من الدراسات.

التعلم الآلي بدون إشراف: يختلف عن التعلم الخاضع للإشراف حيث لا يتم تقديم بيانات معنونة للمعالجة. بدلاً من ذلك، يتم تدريب الآلات على التعامل مع البيانات غير المعنونة، مما يتطلب من الآلات التعلم من الهيكل الطبيعي للبيانات.

التعلم الآلي شبه الخاضع للإشراف: يقع هذا النوع بين التعلم الخاضع للإشراف والتعلم بدون إشراف، حيث يستخدم مزيجاً من البيانات المصنفة وغير المصنفة خلال عملية التدريب. يساعد هذا النوع على تعزيز قدرات الآلات على التعلم من البيانات بطريقة أكثر مرونة.

التعلم الآلي المعزز: يعتمد على مفهوم التغذية الراجعة، حيث يتم تعليم الوكيل الذكائي كيفية التعلم من تجاربه الخاصة عبر مكافأة الجوانب الإيجابية والسلبية من السلوكيات. هذا النوع لا يعتمد على بيانات مصنفة مسبقاً ويعتبر من الأنظمة العكسية حيث يتعلم الوكيل من تجاربه مباشرة.

كل نوع من هذه الأنواع يلعب دورًا مختلفًا في مجال التعلم الآلي، ويمكن اختيار النوع المناسب بناءً على طبيعة المشكلة والبيانات المتاحة.

1.5 خطوات التعلم الالي

عملية التعلم الآلي تتبع عدة خطوات أساسية لتحقيق أقصى قدر من الدقة والفعالية في النماذج المُنشأة:

1. استخراج البيانات: تتعلق هذه الخطوة بتجميع البيانات من مصادر متعددة والتي ستستخدم في عملية التعلم الآلي. الهدف منها هو تحديد البيانات الأكثر ملاءمة للاستخدام في النموذج.
2. تحليل البيانات: يتم في هذه المرحلة إجراء تحليل استكشافي للبيانات لتقديم فهم أفضل لها. تتضمن هذه الخطوة:

- فهم البيانات، الهيكل، وتوزيعاتها لاستخدامها كمدخل أو تسمية للنموذج.
- تحديد الإجراءات اللازمة لتنظيف البيانات وتحسين الميزات لتحسين القدرة التنبؤية للنموذج.

3. إعداد البيانات: تشمل هذه الخطوة تنظيف البيانات وإزالة البيانات الغير منطقية، وتقسيم البيانات إلى مجموعات تدريب، تحقق من الصحة، واختبار. بالإضافة إلى ذلك، قد تتضمن هندسة الميزات لإنشاء ميزات جديدة التي يمكن أن تحسن القدرة التنبؤية للنموذج.
4. تدريب النموذج: في هذه المرحلة، يتم استخدام البيانات المستعدة لتدريب نماذج التعلم الآلي المختلفة باستخدام خوارزميات متنوعة.

5. تقييم النموذج: يتم تقييم النموذج الأفضل الذي تم تدريبه في الخطوة السابقة باستخدام بيانات الاختبار. قبل هذه الخطوة، يجب تحديد مقياس تقييم مناسب لتحديد مدى فعالية النموذج.

كل خطوة في هذه العملية تلعب دورًا حاسمًا في ضمان أن النموذج النهائي قادر على تحقيق أقصى قدر من الدقة والفعالية في حل المشكلة التي تم تصميمها له.

1.6 الشذوذ

يعرف الشذوذ بالشكل العام بأنه أنماط في البيانات لا تتوافق مع مفهوم محدد جيداً للسلوك الطبيعي.

يلعب الشذوذ دوراً هاماً في مجالات مختلفة تتجاوز العلوم والإحصاء فقط. ففي عالم الاقتصاد والتمويل، يشير الشذوذ إلى الحالات التي تختلف فيها النتائج الفعلية عن التوقعات القائمة على نماذج راسخة. وكثيراً ما تتحدى هذه التناقضات الافتراضات الأساسية وتسلط الضوء على المجالات التي قد تكون فيها النظريات الحالية غير كافية. على سبيل المثال، تعتبر أنماط السوق المحددة التي تتناقض مع فرضية السوق الفعالة، مثل تأثيرات التقويم، شذوذات في الدراسات المالية. في شبكات الكمبيوتر والأمن السيبراني، تعمل الشذوذات كمؤشرات للتهديدات الأمنية المحتملة. تُستخدم الخوارزميات المتقدمة وتقنيات التعلم الآلي بشكل متزايد للكشف عن الأنماط أو السلوكيات غير المعتادة داخل حركة المرور على الشبكة، مما يسمح بالكشف المبكر عن الهجمات الإلكترونية والتخفيف من حدتها قبل أن تتسبب في أضرار جسيمة.

إن فهم وتحليل الشذوذ عبر مختلف المجالات أمر بالغ الأهمية لتطوير المعرفة وتحسين النماذج وتطوير حلول مبتكرة. سواء في العلوم أو الاقتصاد أو الآثار أو الأحياء أو الطيران أو الطب أو التكنولوجيا، تمثل الشذوذ فرصًا للاكتشاف والنمو، مما يشجعنا على تحدي الافتراضات واستكشاف إمكانيات جديدة.

1.6.1 أنواع الشذوذ

شذوذ نقطة البيانات: يحدث عندما ينحرف مثيل بيانات معين عن النمط العادي لمجموعة البيانات، مما يجعله شاذاً. مثلاً، استهلاك بيانات بشكل غير معتاد.

الشذوذ السياقي: يصف حالات حيث يتصرف مثيل البيانات بشكل غير طبيعي في سياق معين، مما يجعله شاذاً. مثلاً، الاتصالات على سيرفر معين خلال فترة زمنية معينة.

الشذوذ الجماعي: يحدث عندما تتصرف مجموعة من حالات البيانات المتشابهة بشكل غير طبيعي بالنسبة لمجموعة البيانات بأكملها، مما يجعلهم شاذين جماعياً. مثلاً، محاولة تسجيل الدخول باستخدام كلمة مرور خاطئة لعدة حسابات.

1.7 اكتشاف الشذوذ

يعتبر اكتشاف الشذوذ تحديًا بسبب عدة عوامل:

1. تحديد السلوك الطبيعي يمكن أن يكون صعبًا ومتغيرًا بناءً على السياق.
2. الحدود بين السلوك الطبيعي والشاذ غالبًا ما تكون غير واضحة.
3. الأعداء الخبثاء قد يتعاونوا لجعل الشذوذ يبدو طبيعيًا.
4. السلوك الطبيعي قد يتغير مع الزمن، مما يجعل تعريف السلوك الطبيعي أصعب.
5. مفهوم الشذوذ يختلف بين مجالات مختلفة، مما يجعل تطبيق تقنيات اكتشاف الشذوذ صعبًا.
6. توافر البيانات المصنفة للتدريب والتحقق من صحة النماذج هو مشكلة كبيرة.
7. البيانات غالبًا ما تحتوي على ضوضاء مشابهة للشذوذ، مما يصعب تمييزها وإزالتها.

نظراً للتحديات المذكورة أعلاه، فإن مشكلة اكتشاف الشذوذ، في شكلها الأكثر عمومية، ليس من السهل حلها.

1.7.2 نتيجة كشف الشذوذ

في مجال الكشف عن الشذوذ، يعتبر الطريقة التي يتم بها الإبلاغ عن الحالات الشاذة جزءًا أساسيًا من التقنيات المستخدمة. تنتج تقنيات اكتشاف الشذوذ عادةً بمخرجات من نوعين رئيسيين:

1. الدرجات Scores تقوم هذه التقنيات بتعيين درجة شذوذ لكل حالة في بيانات الاختبار، حيث تعتبر الحالة شاذة إذا كانت الدرجة فوق معين. النتيجة هي قائمة مرتبة بالحالات الشاذة، حيث يمكن للمحلل تحليل بعض الحالات الشاذة أو استخدام حد قطع معين لتحديد الحالات الشاذة.
2. التسميات Label تقدم هذه الأساليب تسمية (عادية أو غير طبيعية) لكل حالة اختبار. تسمح هذه التقنيات للمحلل باستخدام عتبة خاصة بالمجال لتحديد الحالات الشاذة الأكثر ارتباطًا بالمشكلة، دون السماح للمحللين بإجراء اختيار مباشر.

تستخدم تقنيات الكشف عن الشذوذ في مجالات متعددة، بما في ذلك:

– كشف الاحتيال: في قطاعات التأمين والبنوك، حيث يتم استخدامها لاكتشاف النشاطات الشاذة التي قد تشير إلى الاحتيال.

– كشف التسلل: في شبكات الكمبيوتر، حيث تساعد في التعرف على الجهود غير القانونية لاختراق شبكة أو نظام معلومات.

– الكشف عن الخطأ / الضرر: في التجارة والصناعة، حيث تساعد في تحديد الأحداث الشاذة التي قد تشير إلى أخطاء أو أضرار في العمليات التجارية.

هذه التقنيات تلعب دورًا حاسمًا في تحسين الأمان والكفاءة في العديد من المجالات، حيث تساعد في التعرف على النشاطات الشاذة والتعامل معها بشكل فعال.

1.8 تقنيات الكشف عن الشذوذ

1.8.1 الطرق الإحصائية

تقنيات الكشف عن الشذوذ تعتمد على مجموعة واسعة من الأساليب والخوارزميات، وأحد أقدمها هو الأساليب الإحصائية. هذه الأساليب تناسب بشكل أفضل البيانات ذات القيمة الحقيقية الكمية أو توزيعات البيانات الترتيبية الكمية، حيث يمكن تحويل البيانات الترتيبية إلى قيم عددية مناسبة للمعالجة الإحصائية. ومع ذلك، قد يكون تحويل البيانات المعقدة ضروريًا قبل المعالجة، مما يزيد من وقت المعالجة وتقليل قابلية التطبيق.

أحد الأساليب الإحصائية البسيطة هو استخدام المخططات الصندوقية غير الرسمية لتحديد القيم المتطرفة في مجموعات البيانات أحادية المتغير ومتعددة المتغيرات. هذه التقنية تتيح لمدقق بشري تحديد النقاط البعيدة بصرياً، مما يجعلها مفيدة في التعامل مع السمات الحقيقية القيمة والترتيبية والفئوية (بدون ترتيب)[4] .

المخططات الصندوقية ترسم الحد الأدنى، والربيع الأدنى، والربيع الأعلى، والربيع الأعلى، والنقاط القصوى القصوى. بالنسبة للبيانات أحادية المتغير، هذا مخطط بسيط من 5 نقاط. القيم المتطرفة هي النقاط التي تتجاوز القيم القصوى الدنيا والعليا لمخطط الصندوق، مثل V و Y و .Zلا تقدم مخططات الصندوق أي افتراضات حول نموذج توزيع البيانات ولكنها تعتمد على الإنسان لملاحظة القيم العظمى المرسومة في مخطط الصندوق كما هو موضح بالشكل 2-1.

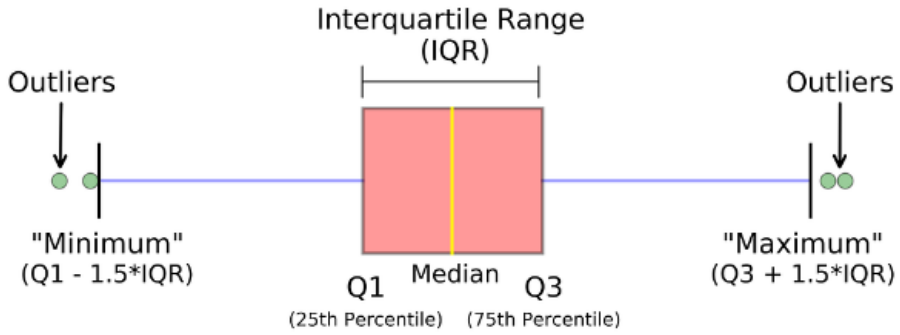
بالإضافة إلى الأساليب الإحصائية، هناك العديد من التقنيات الحديثة الأخرى التي تستخدم في الكشف عن الشذوذ، منها:

– الخوارزميات اللينة: تستخدم لتحديد الشذوذ بناءً على القواعد المحددة من قبل المستخدم أو تعلمها من البيانات.

–الخوارزميات العصبية: مثل الشبكات العصبية الاصطناعية والخوارزميات العصبية الأخرى، التي تستخدم لتعلم الأنماط وتحديد الشذوذ بناءً على البيانات المدخلة.

– خوارزميات التعلم العميق: تستخدم لتحليل البيانات المعقدة وتحديد الشذوذ بناءً على الأنماط المعقدة التي تعلمتها من البيانات.

كل هذه التقنيات تقدم طرقًا متعددة لتحديد الشذوذ في البيانات، وتعتمد على طبيعة البيانات والمجال الذي يتم تطبيقه فيه.



الشكل 1_2 كشف الشذوذ بالطرق الإحصائية

تتميز الأساليب التقليدية (الإحصائية) في كشف الشذوذ الشبكي بقدرتها على تحديد السلوك المتوقع لحركة المرور في الشبكات، وذلك عبر الاعتماد على نظرية التغيرات المفاجئة التي تطلق إنذاراً عند حدوث انحراف كبير عن النمط الطبيعي. هذه الأساليب لا تتطلب أي معرفة مسبقة حول النظام كمدخلات، مما يجعلها قادرة على التعامل مع مجموعة واسعة من البيئات الشبكية.

ومع ذلك، هناك عيوب تظهر في هذه الأساليب، بما في ذلك:

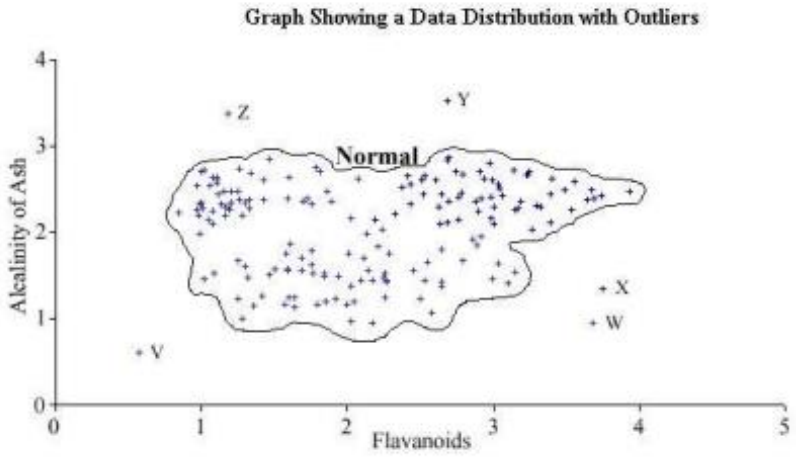
– قد تكون بعض أنواع الهجمات جزءًا طبيعيًا من مجموعة البيانات، مما يجعلها تبدو طبيعية بدلاً من شاذة.

– يتطلب تدريب النماذج وقتًا لضبط الإنذار الأول، مما يعيق الاستجابة السريعة للشذوذ.

– قد لا يكون استخدام العتبات موثوقًا في بعض الحالات الحقيقية بسبب طبيعتها الثابتة والمحدودة.

1.8.2 الطرق البارامترية

تسمح الطرق البارامترية بتقييم النموذج بسرعة كبيرة بالنسبة للحالات الجديدة وتكون مناسبة لمجموعات البيانات الكبيرة؛ ينمو النموذج فقط مع تعقيد النموذج وليس حجم البيانات. ومع ذلك، فإنها تحد من قابليتها للتطبيق من خلال فرض نموذج توزيع محدد مسبقاً ليناسب البيانات. إذا كان المستخدم يعرف أن بياناته تناسب نموذج التوزيع هذا، فستكون هذه الأساليب دقيقة للغاية ولكن العديد من مجموعات البيانات لا تناسب نموذجاً معيناً. أحد هذه الأساليب هو الحد الأدنى لتقدير الحجم الإهليلجي Minimum Volume ellipsoid Estimator [5] (MVE)، (والذي يناسب أصغر حجم إهليلجي مسموح به حول غالبية نموذج توزيع البيانات (يغطي بشكل عام 50٪ من نقاط البيانات). يمثل هذا المنطقة العادية المكتظة الموضحة في الشكل 1_3 (مع القيم المتطرفة الموضحة).



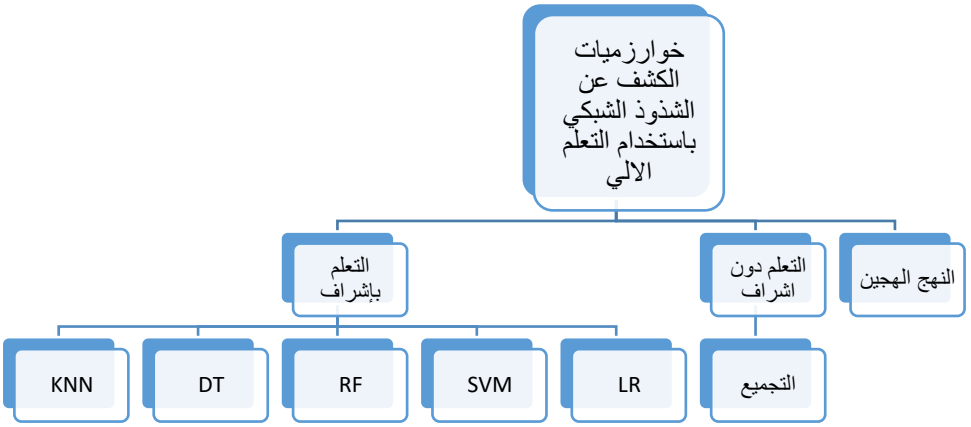
الشكل 1_3: mve مع القيم المتطرفة.

1.8.3 طرق التعلم الالي

تقسم طرق التعلم الآلي الى خوارزميات خاضعة للإشراف وغير خاضعة للإشراف.

يتم استخدام منهجيات التعلم الآلي على نطاق واسع من قبل الباحثين في مجال اكتشاف شذوذ الشبكة لعدة أسباب أبرزها قدرات نماذج التعلم الآلي التعميمية التي تساعد على فهم المعرفة التقنية حول الشذوذ التي لا تحتوي على أي أنماط محددة مسبقاً [11].

في اكتشاف الشذوذ الشبكي باستخدام التعلم الآلي تم استخدام مجموعة من الخوارزميات حيث لوحظ من خلال الدراسات السابقة استخدام الخوارزميات التالية اكثر من غيرها كما يوضح الشكل 1_4:



الشكل 1_4 خوارزميات الكشف عن الشذوذ الشبكي في أدبيات الدراسات السابقة.

- **خوارزميات تعلم الآلة الخاضعة للإشراف**

في مجال التعلم الآلي، يتم تقسيم التقنيات إلى خوارزميات تعلم الآلة الخاضعة للإشراف وغير خاضعة للإشراف. الخوارزميات الخاضعة للإشراف تستخدم بيانات مصنفة لتدريب النماذج، حيث يتم تعيين تسميات صحيحة لكل حالة بيانات، سواء كانت عادية أو شاذة. هذه التقنيات تُستخدم بشكل واسع في أنظمة الكشف عن الشذوذ، حيث يتم استخدامها لتحديد الأنماط الشاذة بناءً على البيانات المدخلة.

من الأمثلة على خوارزميات التعلم الآلي الخاضعة للإشراف، نجد:

– شجرة القرار Decision Tree تقوم بتقسيم البيانات إلى فرعين أو أكثر بناءً على قيم معينة، مما يجعلها أداة فعالة لتحليل البيانات وفهم الأنماط.

– آلة المتجهات الداعمة Support Vector Machine تعتبر من الخوارزميات القوية لتحديد الحدود بين مجموعتين مختلفتين من البيانات، وهي مفيدة خاصة في التعلم العميق والتعلم الآلي الخاضع للإشراف[6] .

– الشبكة العصبونية Neural Network تستخدم شبكات عصبونية لتكرار الأنماط المعقدة في البيانات، وهي قادرة على التعلم من البيانات غير المصنفة وكذلك البيانات المصنفة[7] .

– الغابة العشوائية Random Forest تعتبر من الخوارزميات المتقدمة التي تستخدم مجموعة من الشجرات لتحسين دقة التنبؤ والتقليل من الخطأ.

تتميز هذه الخوارزميات بقدرتها على التعلم من البيانات المصنفة وتحديد الأنماط الشاذة بناءً على تلك البيانات. تستخدم بشكل واسع في مجالات متعددة، بما في ذلك الكشف عن الاحتيال، الكشف عن التسلل، والكشف عن الأخطاء أو الأضرار في البيانات.

- **خوارزميات تعلم الآلة غير الخاضعة للإشراف**

في مجال التعلم الآلي، هناك خوارزميات تعلم الآلة غير الخاضعة للإشراف التي لا تتطلب بيانات مصنفة لتدريب النماذج. هذه الخوارزميات تعتمد على اكتشاف وتحليل بنية البيانات الإدخالية مباشرة. بعض الخوارزميات الأكثر شيوعاً في هذا المجال تشمل الخوارزميات الجينية وتجميع الوسائل K– means.

1. الخوارزمية الجينية :تستخدم في كشف التسلل لاشتقاق قواعد التصنيف من بيانات تدقيق الشبكة. Genetic algorithm GA تمتاز بقدرتها على التعلم الذاتي ومقاومتها للضوضاء، مما يجعلها فعالة في

التعامل مع البيانات غير المنسقة. تم الإبلاغ عن استخدام GA في تحسين معدلات اكتشاف الهجمات وتقليل معدلات الإيجابية الكاذبة.

2. (K-means Clustering):هي واحدة من الخوارزميات البسيطة والشائعة في التعلم الآلي غير الخاضع للإشراف. تقوم K-means بتقسيم مجموعة البيانات إلى K مجموعات بناءً على تشابه النقاط داخل كل مجموعة. يتم تحديد K، وهو عدد المجموعات، مسبقًا، وهدفها هو تجميع النقاط التي تشارك خصائص مشتركة. النقطة الوسطى لكل مجموعة هي متوسط للنقاط داخل المجموعة، وتمثل مركز الكتلة.

تتميز هذه الخوارزميات بقدرتهما على التعامل مع البيانات غير المصنفة وتحديد الأنماط الموجودة في البيانات بدون الحاجة إلى تعليمات أو بيانات مصنفة مسبقًا. تستخدم بشكل واسع في مجالات متعددة، بما في ذلك الكشف عن الشذوذ، التنقيب عن البيانات، وتحسين الأداء في الأنظمة.

1.9 الخلاصة

تم في هذا الفصل الدراسة النظرية للمفاهيم المستخدمة في هذا البحث حيث تم التطرق الى مفهوم التعلم الالي وانواعه وتحدياته والشذوذ وماهيته وانواعه وكيفية اكتشافه.

الفصل الثاني الدراسة المرجعية

2.1 مقدمة

تشتمل الأدبيات العلمية على مجموعة متنوعة من الخوارزميات والتقنيات لتحليل الشذوذ في البيانات، وهذا التنوع يعود إلى طبيعة البيانات الشبكية الخاصة التي تتميز بتنوعها وخصوصيتها. تشمل هذه الأساليب الأساسية الإحصائية، التعلم الآلي، والتعلم العميق، حيث يتم استخدام هذه التقنيات لتحديد الأنماط الشاذة في حركة البيانات الشبكية.

في هذا الفصل، سنستعرض الدراسات السابقة بطريقة ترتبط بخوارزمية الكشف عن الشذوذ المستخدمة في البيانات الشبكية، وهذا التصنيف يساعد في فهم التقنيات الحالية وكيفية اختيارها. الهدف من هذا التصنيف هو:

1. التعرف على التقنيات الحالية المستخدمة لكشف الشذوذ في البيانات الشبكية.

2. التعرف على الطرق الأكثر فعالية من خلال مقارنة مختلف طرق كشف الشذوذ الشبكي.

3. التعرف على الطرق التي تم اختيارها لاختيار الميزات المهمة التي تمثل المعرفة الحقيقية في الشبكة.

4. تحديد الأسباب التي تجعل خوارزمية معينة تفضل على أخرى في كشف الشذوذ الشبكي.

تم التركيز بشكل أساسي على استخدام طرق التعلم الآلي، حيث يتم التدريب تحت الإشراف، بالإضافة إلى التعلم الآلي غير الخاضع للإشراف، نظرًا لاعتبار مشكلة الشذوذ كمسألة تصنيف، حيث يتم تصنيف حركة المرور الشبكية كطابع طبيعي أم غير طبيعي. ومع ذلك، استكشفت العديد من الدراسات أيضًا

استخدام التعلم العميق بسبب قدرته على التعلم من البيانات غير المتوازنة بكفاءة وتقديمه لاكتشاف أنماط شذوذ جديدة.

2.2 الدراسات المرجعية

في الدراسة [8]، تم اقتراح نظام للكشف عن الشذوذ الإحصائي يمكنه العمل مع الشبكات السلكية واللاسلكية على حد سواء، مستخدماً النمذجة الإحصائية.

في دراسة أخرى [9]، تم استخدام استراتيجيات اكتشاف الشذوذ المستندة إلى المستخدم والمجموعة، حيث يتم تحديد مجموعة من القيم التي تعتبر طبيعية لكل ميزة. إذا كانت إحدى الميزات خارج النطاق الطبيعي أثناء إحدى الجلسات، سيتم اعتبارها شذوذ.

أما في دراسة [10]، تم استخدام محرك كشف الشذوذ الإحصائي Statistical Packet Anomaly Detection Engine (SPADE)، الذي يعتبر واحدًا من الأوراق الأولى التي اقترحت استخدام مفهوم درجة الشذوذ لاكتشاف عمليات مسح المنفذ.

قام الباحثون في [12]، بمقارنة مجموعة من خوارزميات الكشف عن الشذوذ في البيانات الشبكية لتحديد أفضل خوارزمية لهذا الغرض. الخوارزميات التي تمت مقارنتها تشمل التدرج العشوائي اللائق، الغابات العشوائية، الانحدار اللوجستي، آلة المتجهات الداعمة، والنموذج المتسلسل. للمساعدة في هذه العملية، تم استخدام مجموعة بيانات KDD التي تمت معالجتها مسبقًا، حيث تم تغيير ثلاث ميزات لقيم الكائن إلى تنسيق رقمي. هذه الميزات هي "نوع البروتوكول"، "الخدمة"، و"العلم". بعد تطبيق تقنية تشفير واحد ساخن، وصل عدد الميزات إلى 122 بالإضافة إلى تسمية المصنفات. تم تقييم الخوارزميات باستخدام مقاييس الأداء مثل الاستدعاء، درجة F1، Receiver Operating Characteristic (ROC)،

والدقة. النتائج أظهرت أن خوارزمية الغابات العشوائية تفوقت على المصنفات الأربعة الأخرى، حيث حققت أعلى قيمة للاسترجاع، مما يعني تحقيق الحد الأدنى من السلبيات الكاذبة.

بالإضافة إلى ذلك، تم تطوير خوارزمية شجرة القرار بناءً على نهج شجرة القرار في [13]. حيث يعتبر اختيار الميزة وقيمة التقسيم من الأمور الحاسمة لإنشاء شجرة قرار. لذلك، تم تصميم الخوارزمية لمعالجة هاتين المسألتين. يتم تحديد الميزات الأكثر صلة باستخدام اكتساب المعلومات، ويتم تحديد قيمة الانقسام بطريقة تجعل المصنف غير متحيز تجاه القيم الأكثر شيوعًا. تم إجراء التجارب على مجموعة بيانات NSL-KDD استنادًا إلى عدد الميزات، مما يوفر نظرة ثاقبة على كيفية تحسين دقة الكشف عن الشذوذ في البيانات الشبكية.

في الدراسة [14]، تم استخدام 20% من السجلات في KDDTrain كمجموعة بيانات للتدريب ، حيث تم تدريب خوارزميات الكشف عن الشذوذ مثل الشجرة العشوائية، C4.5، وNBTreeشجرة نايف بازين . بعد ذلك، تم تطبيق النماذج التي تم تعلمها على مجموعتين من الاختبارات، وهما KDDTest و KDDTest-21بالإضافة إلى الدقة الشاملة، تم التعبير عن أداء التصنيف أيضًا من حيث الحساسية والخصوصية. حيث يُظهر مقياس الحساسية مدى جودة الخوارزميات المقترحة عند اكتشاف هجمات الشبكة، بينما تمثل الخصوصية مقياس الاكتشافات الخاطئة، أو أداء المصنف في اكتشاف الهجمات العادية. تم اختيار ثلاث خوارزميات كشف تحقق أعلى درجات الدقة: الشجرة العشوائية، C4.5، و .NBTree الخوارزميات الأخرى المستخدمة في المقارنة هي Naive-Bayes، perceptron متعدد الطبقات، وآلات ناقلات الدعم، وكان الحد الأدنى لعدد المثيلات لكل ورقة 6، وتم استخدام خمسة أضعاف لتقليل الخطأ. تم استخدام NBTree مع المعلمات الافتراضية.

تم تحقيق دقة الكشف بنسبة 89.24٪ باستخدام مزيج من الشجرة العشوائية وخوارزميات NBTree. في حالة المصنفات الفردية، كان أداء الشجرة العشوائية أفضل أداء على الرغم من أن C4.5 كان يحتوي على أقل عدد من الإنذارات الكاذبة. كان أداء NBTree أسوأ بناءً على أي معايير تقييم تم استخدامها في هذا البحث.

في دراسة أخرى [15]، تم تقييم أداء أربع خوارزميات تعلم الآلي وهي RF random forest الغابة العشوائية و شجرة القرار Decision tree ونايف بايزن Naïve Bayes و التدرج العشوائي Stochastic Gradient جنباً إلى جنب مع تقنيات اختيار الميزات المختلفة مرشح ترتيب الارتباط Correlation Ranking Filter (CRF) و Feature Evaluator Gain Ratio (GRFE)أداة تقييم ميزة نسبة الربح ، حيث تم في كل تقنية اختيار 15 ميزة على حدى لتنفيذ نماذج نظام كشف التسلل Network Intrusion detection system (NIDS). بناء على التحليل والمقارنات التي تم إجراؤها، أثبتت النتائج أن أداء Random Forest أفضل من الخوارزميات الأخرى في الدقة المتوقعة accuracy وخطأ الكشف. حيث تقدم Random Forest أداءً أفضل من الخوارزميات الأخرى حيث كانت الدقة المتوقعة 99.3172٪ وخطأ الكشف 0.6828٪ والذي تم تحقيقه باستخدام الميزات المحددة بواسطة مرشح ترتيب الارتباط (CRF، بينما الدقة المتوقعة 99.7618٪ وخطأ الكشف 0.2382 بواسطة GRFE، بينما كان أقل اكتشاف (87.7104٪) أثناء استخدام خوارزمية تعلم Decision tree مع خطأ اكتشاف بنسبة 12.2896٪ بعد استخدام الميزات التي تم إنشاؤها بواسطة CRF. لكن تم تحسين الأداء بعد استخدام الميزات المحددة بواسطة GRFE. ومع ذلك، فإن أداء الخوارزمية مثل SGD أفضل أيضاً (بدقة تبلغ 96.7450٪ وخطأ كشف يبلغ 3.255٪). يتم أيضاً قياس وقت التدريب (الوقت المستغرق لتدريب النموذج) وهو 10.55 ثانية أثناء تدريب نموذج NIDS على أساس الانحدار

العشوائي بعد إجراء اختيار الميزة باستخدام CRF. ومن سلبيات هذه الدراسة عدم الحديث عن الوقت اللازم لاكتشاف الشذوذ حيث تم التطرق فقط الى الوقت اللازم لتدريب النماذج.

في الدراسة [16]، قام الباحثون بتحقيق دراسة في تقنيات التنقيب عن البيانات المختلفة التي يمكن أن تقلل من عدد السلبيات الكاذبة والإيجابيات الكاذبة في اكتشاف الشذوذ في البيانات الشبكية. تم فحص فعالية المصنف C5 مقارنةً مع المصنفات الأخرى، باستخدام مجموعة بيانات NSL KDD. أظهرت النتائج أن C5 قد حقق دقة عالية وإنذارات كاذبة منخفضة للكشف عن التسلل، حيث بلغ معدل اكتشاف حركة مرور الضارة بنسبة 99.82%. تمت مقارنة C5 مع C4.5 و آلة دعم المتجهات SVM وNaive Bayes، وتبين أن C5 كان الأفضل في هذه المقارنة، مع استخدام الذاكرة الضئيلة وسرعة جيدة ودقة ممتازة أيضًا.

في الدراسة [17]، كان الهدف الرئيسي هو اختبار تأثير المعالجة المسبقة للبيانات والاختيار من بين الميزات على مجموعة بيانات NSL-KDD المستخدمة لتصنيف التدخلات. تم استخدام نظرية نظرية المجموعة الخشنة Rough set theory (RST) وكسب المعلومات لاختيار الميزة. بعد تطبيق هذا الاختيار، استخدم المصنف J48 لتقييم الأداء. النتائج أظهرت انخفاضًا كبيرًا في وقت المعالجة مع تقليل عدد الميزات. بالإضافة إلى ذلك، كانت النتائج ذات معدل اكتشاف جيد ودقة عالية، مع الحفاظ على الإيجابيات الخاطئة في الحد الأدنى.

في الدراسة [18]، تم تطوير فكرة تقليل عدد الميزات الأصلية التي تمثل جميع سجلات الاتصال في المجموعة البيانات لغرض اكتشاف التسلل. هذه الفكرة تعتمد على حقيقة أن تعدين البيانات يعطي أنماطًا مخفية ويستكشف البيانات بطريقة مختلفة. العمل المقدم في هذا البحث يشرح كيف يمكن استخراج المعلومات ذات الصلة باستخدام PCAFC-KNN من أجل بناء معرفات قوية مع أقصى معدل للكشف

مع الحد الأدنى من الإنذارات الكاذبة. تظهر النتائج التجريبية أن مع زيادة حجم البيانات، ستنخفض كفاءة ودقة خوارزمية كشف التسلل تدريجياً.

في الدراسة [19]، قام الباحثون باستخدام مجموعة بيانات KDD لتدريب خوارزمية التعلم الآلي غير الخاضعة للرقابة تسمى غابة العزلة. .Isolation Forest تم إجراء المعالجة المسبقة للبيانات أولاً، واستخدموا أيضًا التعلم غير الخاضع للإشراف. بالنسبة لهذه الشبكة، تم استخدام غابة عزل نموذج الكشف عن الشذوذ المستند إلى حركة مرور الشبكة لاكتشاف القيم المتطرفة والهجمات المحتملة. تم تقييم النتائج باستخدام درجة الانحراف AUC، حيث وصلت إلى 98.3% عند n_estimators مساوي لل100.

في الدراسة [20]، قام الباحثون بإقتراح وتنفيذ سبعة نماذج ML هجينة تستخدم خوارزميات التعلم الخاضعة للإشراف NN وSVM في المستوى الأول، وخوارزمية K-Means غير الخاضعة للإشراف في المستوى الثاني. تم استخدام خوارزمياتPrincipal component analysis تحليل البيانات الشخصية PCA وGFR لاختيار الميزات. الهدف من هذا النهج الهجين هو تعظيم اكتشاف التسلل في شبكات الكمبيوتر، حيث يتم اكتشاف الهجمات المعروفة من خلال التعلم الخاضع للإشراف، بينما يتم اكتشاف الهجمات غير المعروفة عن طريق التعلم غير الخاضع للإشراف.

في الدراسة [21]، تم تطوير النموذج الهجين الذكي لاستكشاف أي اختراق داخل الشبكة. يندرج هذا النموذج تحت مرحلتين أساسيتين. تتضمن المرحلة الأولى استخدام الخوارزمية الجينية في اختيار الخصائص، مع اعتمادها على عملية الاستخراج والتقدير وتقليل الأبعاد من خلال التقسيم النسبي للفاصل الزمني (PKID) والتحليل الخطي فيشر (FLDA) على التوالي. في نهاية المرحلة الأولى، تم جمع مصنف Naïve Bayes (NB) وشجرة القرار (DT) باستخدام مجموعة بيانات NSL-KDD

مقسمة إلى مجموعتين للتدريب والاختبار. تعتمد المرحلة الثانية بشكل كامل على مخرجات المرحلة الأولى (الفئة المتوقعة) وإعادة تصنيفها باستخدام تصورات متعددة الطبقات باستخدام Deep Learning (DL) واستخدام خوارزمية Stochastic Gradient Descent (SGD). الهدف من هذا النهج هو تحسين الأداء من حيث الدقة في تصنيف الاختراقات ورفع معدل الاكتشاف وتقليل الإنذارات الكاذبة. توضح المقارنة بين النموذج المقترح والنماذج التقليدية تفوق النموذج المقترح، مع دقة تصنيف تبلغ 99.9325%، ومعدل اكتشاف يصل إلى 99.9738%، ودراسة إنذارات كاذبة بنسبة 0.00093%.

بالإضافة إلى التحليل، بدلاً من استخدام جميع الميزات كمدخل إلى DNN، تم تطبيق تحليل المكون الرئيسي على 41 ميزة لاستخراج 15 ميزة مختصرة ثم إدخالها في DNN. تم ضبط المعلمات الفائقة لجميع النماذج المذكورة أعلاه، وتم تشغيل جميع النماذج على بيانات التدريب NSL-KDD واختبارها لاحقاً على بيانات اختبار NSL-KDD. كما كان أداء نموذج DNN أفضل على مجموعة البيانات المخفضة.

تم اقتراح نهجاً قائماً على التعلم العميق حيث تم استخدام التعلم الذاتي [23] (STL)، وهو أسلوب قائم على التعلم العميق، على مجموعة بيانات NSL-KDD. تم مقارنة أداء هذا النهج ببعض الأعمال السابقة. تشمل المقاييس المقارنة قيم الدقة والدقة والتذكر والقياس. تم ملاحظة أن STL تُظهر أداءً أفضل للتصنيف الثنائي مقارنةً بـ SMR. ومع ذلك، فإن أداءها مشابه جداً في حالة 5 فئات و23 فئة. كما تم ملاحظة أن STL حققت معدل دقة تصنيف أكثر من 98٪ لجميع أنواع التصنيف.

يوضح الجدول 2-1 مقارنة الدراسات السابقة

جدول 2-1 مقارنة الدراسات السابقة

الدراسة	المنهجية	النتائج
[8]	اقتراح نظام للكشف عن الشذوذ الإحصائي	العمل مع الشبكات السلكية واللاسلكية على حد سواء، مستخدماً النمذجة الإحصائية.
[9]	استخدام استراتيجيات اكتشاف الشذوذ المستندة إلى المستخدم والمجموعة	يتم تحديد مجموعة من القيم التي تعتبر طبيعية لكل ميزة. إذا كانت إحدى الميزات خارج النطاق الطبيعي أثناء إحدى الجلسات، سيتم اعتبارها شذوذ.
[10]	استخدام محرك كشف الشذوذ الإحصائي Statistical Packet Anomaly Detection Engine (SPADE)	اكتشاف عمليات مسح المنفذ
[12]	مقارنة مجموعة من خوارزميات الكشف عن الشذوذ في البيانات الشبكية لتحديد أفضل خوارزمية لهذا الغرض	النتائج أظهرت أن خوارزمية الغابات العشوائية تفوقت على المصنفات الأربعة الأخرى، حيث حققت أعلى قيمة للاسترجاع، مما يعني تحقيق الحد الأدنى من السلبيات الكاذبة.
[13]	تم تطوير خوارزمية شجرة القرار بناءً على نهج شجرة القرار	تحسين دقة الكشف عن الشذوذ في البيانات الشبكية.
[14]	تم استخدام 20% من السجلات في KDDTrain كمجموعة بيانات للتدريب	تم اختيار ثلاث خوارزميات كشف تحقق أعلى درجات الدقة: الشجرة العشوائية، C4.5، و .INBTree
[15]	تم تقييم أداء أربع خوارزميات تعلم الآلي وهي RF و Decision tree و Naïve Bayes و Stochastic Gradient جنباً إلى جنب مع تقنيات اختيار الميزات المختلفة Correlation Ranking Filter (CRF) و Gain Ratio	. حيث تقدم Random Forest أداءً أفضل من الخوارزميات الأخرى حيث كانت الدقة المتوقعة 99.3172٪ وخطأ الكشف 0.6828٪ والذي تم تحقيقه باستخدام الميزات المحددة بواسطة مرشح ترتيب الارتباط

	Feature Evaluator(GRFE)،	
[16]	قام الباحثون بتحقيق دراسة في تقنيات التنقيب عن البيانات المختلفة التي يمكن أن تقلل من عدد السلبيات الكاذبة والإيجابيات الكاذبة في اكتشاف الشذوذ في البيانات الشبكية	أظهرت النتائج أن C5 قد حقق دقة عالية وإنذارات كاذبة منخفضة للكشف عن التسلل، حيث بلغ معدل اكتشاف حركة مرور الضارة بنسبة 99.82%. تمت مقارنة C5 مع C4.5 وSVM وNaive Bayes، وتبين أن C5 كان الأفضل في هذه المقارنة،
[17]	الهدف الرئيسي هو اختبار تأثير المعالجة المسبقة للبيانات والاختيار من بين الميزات على مجموعة بيانات NSL-KDD	كانت النتائج ذات معدل اكتشاف جيد ودقة عالية
[18]	تم تطوير فكرة تقليل عدد الميزات الأصلية التي تمثل جميع سجلات الاتصال في المجموعة البيانات لغرض اكتشاف التسلل	تظهر النتائج التجريبية أن مع زيادة حجم البيانات، ستنخفض كفاءة ودقة خوارزمية كشف التسلل تدريجياً.
[19]	قام الباحثون باستخدام مجموعة بيانات KDD لتدريب خوارزمية التعلم الآلي غير الخاضعة للرقابة تسمى .Isolation Forest	تم تقييم النتائج باستخدام درجة الانحراف AUC، حيث وصلت إلى 98.3% عند n_estimators مساوي لل100
[20]	قام الباحثون بإقتراح وتنفيذ سبعة نماذج ML هجينة تستخدم خوارزميات التعلم الخاضعة للإشراف NN وSVM في المستوى الأول، وخوارزمية K-Means غير الخاضعة للإشراف في المستوى الثاني	اكتشاف الهجمات المعروفة من خلال التعلم الخاضع للإشراف، بينما يتم اكتشاف الهجمات غير المعروفة عن طريق التعلم غير الخاضع للإشراف
[21]	]، تم تطوير النموذج الهجين الذكي لاستكشاف أي اختراق داخل الشبكة	توضح المقارنة بين النموذج المقترح والنماذج التقليدية تفوق النموذج المقترح، مع دقة تصنيف تبلغ 99.9325%، ومعدل اكتشاف يصل إلى 99.9738%، ودراسة إنذارات

		كاذبة بنسبة 0.00093%.
[23]	تم اقتراح نهجاً قائماً على التعلم العميق حيث تم استخدام التعلم الذاتي	تم ملاحظة أن STL حققت معدل دقة تصنيف أكثر من 98٪ لجميع أنواع التصنيف

الفصل الثالث المنهجية المقترحة

3.1 مقدمة

يكون الكشف عن شذوذ البيانات الشبكية من أهمية قصوى لمسؤولي الشبكات، حيث يكون للحالات الشاذة تأثير كبير على أداء الشبكة ومتطلبات جودة الخدمة لمستخدمي الشبكة. ومع ذلك، فإن هذه المهمة تستغرق وقتاً طويلاً ومكلفة للغاية وتتطلب غالبًا اتخاذ قرار من خبراء الشبكة والأمن. لذلك، بدأت العديد من الأعمال البحثية في هذا المجال، مما أدى إلى تنوع الطرق المستخدمة، من بينها الطرق الاحصائية والتعلم الآلي.

لا يزال تحديد المجموعات المقابلة لفئات حركات المرور الشاذة بين المجموعة الكاملة يمثل تحدياً، ويرجع ذلك في الغالب إلى طبيعة الخوارزميات المستخدمة. هذه الخوارزميات تتطلب أهم السمات المميزة في البيانات التي تكون قادرة على تشكيل نمط معين، يكون من خلاله يمكن للخوارزميات تحديد النشاطات المشبوهة وحركة المرور الشاذة بدقة أكبر وبوقت أقسر.

انطلاقًا مما سبق، ظهرت الحاجة إلى توجه جديد لحل مشكلة اكتشاف الشذوذ الشبكي من خلال اكتشاف واختيار الميزات والسمات الأفضل لتزويد خوارزميات التعلم الآلي بها لبناء نموذج أفضل قادر على اكتشاف الشذوذ بدقة أكبر.

سيتم في هذا الفصل شرح الخوارزميات المستخدمة في هذه الدراسة، التي تنقسم إلى خوارزميات التعلم الآلي وهي خوارزمية أقرب جار وخوارزمية شجرة القرار وخوارزمية الغابة العشوائية، وخوارزميات اختيار الميزات وهي مربع كاي Chi-Square وcorrelationالارتباط و أهمية الميزة Feature Importance

كما سيتم تقديم في الفصل اللاحق مقارنة بين هذه الخوارزميات وبين بعضها البعض وبين هذه الخوارزميات والنموذج المقترح لتحديد الخوارزمية ذات الفعالية الأفضل لاستخدامها في مجال اكتشاف الشذوذ في البيانات الشبكية. هذا من خلال مجموعة واسعة من التجارب التي تم إجراؤها.

3.2 خوارزميات اختيار الميزات

وهي مربع كاي و أهمية الميزة و الارتباط

3.2.1 مربع كاي Chi-Square

مربع كاي هو اختبار إحصائي يقيس مدى مقارنة النموذج بالبيانات الفعلية المرصودة. والغرض الأساسي منه هو تقييم ما إذا كان هناك فرق كبير بين الترددات المرصودة والمتوقعة في البيانات التصنيفية. يتم استخدام اختبار مربع كاي في الإحصاء لاختبار استقلالية حدثين. وبالنظر إلى بيانات متغيرين، يمكننا الحصول على العدد المرصود O والعدد المتوقع E. يقيس مربع Chi كيف ينحرف العد المتوقع E والعدد الملحوظ O. [24]

ويحسب بالطريقة المبينة في المعادلة (1).

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \ldots.(3.1)$$

حيث:

C درجة الحرية degrees of freedom ، O القيمة المرصودة ، E القيمة المتوقعة

لنفكر في سيناريو نحتاج فيه إلى تحديد العلاقة بين ميزة الفئة المستقلة (المتنبئ) وميزة الفئة التابعة (الاستجابة). في اختيار الميزة، نهدف إلى تحديد الميزات التي تعتمد بشكل كبير على الاستجابة. عندما تكون ميزتان مستقلتان، يكون العدد المرصود قريباً من العدد المتوقع، وبالتالي سيكون لدينا قيمة Chi-Square أصغر. تشير قيمة Chi-Square العالية إلى أن فرضية الاستقلال غير صحيحة.

ترتبط الخوارزمية بمصطلح يدعي درجات الحرية Degrees of freedom والذي يشير إلى الحد الأقصى لعدد القيم المستقلة منطقياً، والتي تتمتع بحرية التغيير. بكلمات بسيطة، يمكن تعريفه على أنه العدد الإجمالي للملاحظات مطروحاً منه عدد القيود المستقلة المفروضة على الملاحظات.

خطوات إجراء اختبار Chi-Square:

- تحديد الفرضية

فرضية لاغية Null Hypothesis (H0): المتغيران مستقلان.

الفرضية البديلة (H1): المتغيرين ليسا مستقلين.

- بناء جدول Contingency table

هو جدول يوضح توزيع متغير واحد في صفوف وآخر في أعمدة. يتم استخدامه لدراسة العلاقة بين متغيرين

- البحث عن القيم المتوقعة

بناءً على الفرضية الصفرية بأن المتغيرين مستقلين. يمكننا القول ما إذا كان A، B حدثان مستقلان.

- حساب إحصائية Chi-Square

تلخيص القيم المرصودة والقيم المتوقعة المحسوبة في جدول وتحديد قيمة Chi-Square.

- قبول أو رفض فرضية Null(Null Hypothesis)

يمكن تحديد قيم Chi-Square باستخدام جدول Chi-Square من أجل قبول او رفض الفرضية.

3.2.2 الارتباط correlation

يقيس الارتباط العلاقة بين متغيرين. ويقيس مدى قوة ارتباط متغيرين وفي أي اتجاه.

نقاط رئيسية حول الارتباط:

1. يقيس العلاقات الخطية بين المتغيرات
2. تتراوح قوة العلاقة من -1 إلى 1
3. الارتباط الإيجابي يعني زيادة/نقصان المتغيرات معًا

الارتباط السلبي يعني زيادة متغير واحد مع نقصان الآخر

هناك ثلاثة أنواع رئيسية من الارتباط:

1. الارتباط الإيجابي: زيادة/نقصان المتغيرات معًا
2. الارتباط السلبي: زيادة متغير واحد مع نقصان الآخر
3. عدم الارتباط: لا توجد علاقة خطية بين المتغيرات

3.2.3 Feature Importance أهمية الميزة

عندما نقوم بتدريب نموذج Random Forest على مجموعة بيانات ذات ميزات معينة، فإن كائن النموذج الذي نحصل عليه لديه القدرة على إخبارنا بأهم الميزات في التدريب؛ بمعنى آخر أي منهم له التأثير الأكبر على المتغير المستهدف.

في الأساس، في كل انقسام من الشجرة، تكون الميزة المختارة للتقسيم هي الميزة التي تزيد من تقليل نوع معين من الخطأ، مثل Gini أوImpurity هناك بعض العشوائية المخصصة لهذه العملية (ومن هنا جاء اسم Random)، حيث يتم اختيار الميزات التي تدخل في العملية ليتم اختيارها على عقدة بشكل عشوائي. على الرغم من ذلك، يجب أن يكون الهدف الرئيسي هو أنه في كل مرة يتم فيها إنشاء شجرة قرار، سواء بشكل فردي أو لتشكيل مجموعة، فإن المتغيرات أو الميزات المختارة في كل عقدة هي التي تقلل من الخطأ معين. بعد ذلك، بالنسبة لشجرة قرارات فردية، يتم ترتيب أهم المتغيرات حسب مقدار تقليلها للخطأ عند ظهورها، مع ترجيح تقليل هذا الخطأ بعدد الملاحظات على العقدة. في الغابة العشوائية، يتم إجراء ذلك لكل شجرة في الغابة، ثم يتم حساب المتوسط لمعرفة أهمية الميزة الفردية. لإعطاء حدس أفضل، تعد الميزات المحددة في الجزء العلوي من الأشجار أكثر أهمية بشكل عام من الميزات المحددة في العقد النهائية للأشجار، حيث تؤدي التقسيمات العلوية عموماً إلى مكاسب معلومات أكبر.

3.3 خوارزميات التعلم الآلي

وهي خوارزمية شجرة القرار والغابة العشوائية وخوارزمية اقرب جار

3.3.1 خوارزمية شجرة القرار

تعد خوارزمية شجرة القرار خوارزمية مستخدمة على نطاق واسع للتصنيف، والتي تستخدم قيم السمات لتقسيم مساحة القرار إلى مساحات فرعية أصغر بطريقة تكرارية. جوهرها هو نهج فرق تسد، تبدأ بعقدة جذر وتنمو تدريجياً إلى تصنيف نهائي (أو ورقة). يمكن تمثيل عمليات اتخاذ القرار بيانياً كشجرة، على الرغم من رسم هذه الشجرة رأساً على عقب [25].

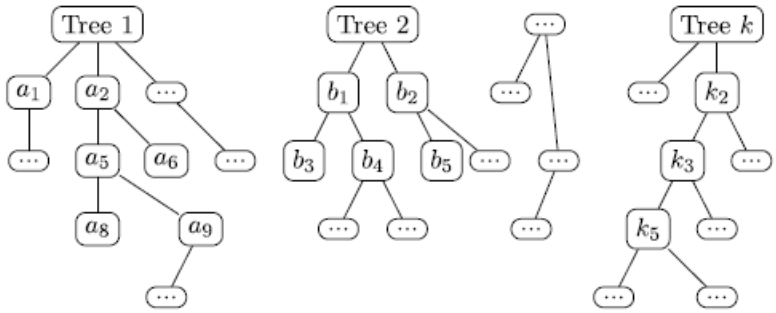
المكونات الرئيسية لنموذج شجرة القرار هي العقد والفروع و أهم خطوات بناء النموذج هي الانقسام والتوقف والتقليم [26].

- العقد:هناك ثلاثة أنواع من العقد. (أ) تمثل العقدة الجذرية، التي تسمى أيضاً عقدة القرار، خياراً سينتج عنه تقسيم جميع السجلات إلى مجموعتين فرعيتين متنافيتين أو أكثر. (ب) العقد الداخلية، التي تسمى أيضاً عقد الصدفة، تمثل أحد الخيارات الممكنة المتاحة في تلك المرحلة في هيكل الشجرة، (ج) تمثل العقد الورقية، التي تسمى أيضًا العقد النهائية، النتيجة النهائية لمجموعة من القرارات أو الأحداث.
- الفروع :تمثل الفروع نتائج الصدفة أو الأحداث التي تنبثق من العقد الجذرية والعقد الداخلية.
- Splitting الانقسام: يتم استخدام متغيرات الإدخال المتعلقة بالمتغير الهدف فقط لتقسيم العقد الأصلية إلى عقد فرعية أنقى للمتغير الهدف. يمكن استخدام كل من متغيرات الإدخال المنفصلة ومتغيرات الإدخال المستمرة.
- التوقف: تشمل المعلمات الشائعة المستخدمة في قواعد الإيقاف ما يلي: (أ) الحد الأدنى لعدد السجلات في ورقة؛ (ب) الحد الأدنى لعدد السجلات في العقدة قبل التقسيم؛ و (ج) العمق (أي عدد الخطوات) لأي ورقة من عقدة الجذر.
- Pruning التقليم: في بعض الحالات، لا تعمل قواعد الإيقاف بشكل جيد. تتمثل إحدى الطرق البديلة لبناء نموذج شجرة القرار في تنمية شجرة كبيرة أولاً، ثم تقليمها إلى الحجم الأمثل عن طريق إزالة العقد التي توفر معلومات إضافية أقل [48].

Random Forest 3.3.2

على وجه التحديد، إنها مجموعة من أشجار القرار التي تم إنشاؤها من مجموعة البيانات، هناك العديد من المتغيرات للغابة العشوائية التي تتميز بـ (1) الطريقة التي يتم بها إنشاء كل شجرة فردية، (2) الإجراء المستخدم لإنشاء مجموعات البيانات المعدلة التي يتم بناء كل شجرة على أساسها، و (3) طريقة

التنبؤات يتم تجميع كل شجرة على حدة لإنتاج تنبؤ فريد بالإجماع. الفكرة الرئيسية لمصنف الغابة العشوائية هي إنشاء أشجار قرارات متعددة بشكل عشوائي باستخدام أخذ العينات مع الاستبدال، وستستند القرارات النهائية إلى قاعدة، غالباً من حيث نظام التصويت لمجموعة أشجار القرار. تم تطوير الفكرة الأساسية للغابات ذات القرار العشوائي في عام 1995 وتم توسيعها بشكل كبير لاحقًا بواسطة Breiman باستخدام فكرة التعبئة. يمكن استخدام التعبئة لتناسب العديد من الأشجار الكبيرة باستخدام تصويت الأغلبية كقاعدة للتصنيف. خوارزمية الغابة العشوائية عبارة عن مجموعة من الأشجار المترابطة مع التعبئة. في جوهره، يعتبر مصنف الغابة العشوائية طريقة تجميعية أو تعلم شجرة قرار قائم على المجموعة بهدف تحسين الدقة مع تقليل التباين وتجنب الإفراط في التجهيز[27] . يمكن لكل شجرة أن تدلي بصوت واحد للفئة الأكثر شيوعاً لمتجه ميزة إدخال معينة. يتم تمثيل الفكرة الأساسية للغابات العشوائية بشكل تخطيطي في الشكل 3_1:



الشكل 3_1: الفكرة الأساسية للغابة العشوائية.

3.3.3 خوارزمية اقرب جار KNN

بغض النظر عما إذا كانت المشكلة هي تخمين رقم أو فصل، فإن الفكرة الكامنة وراء استراتيجية التعلم لخوارزمية K–Nearest Neighbours (KNN) هي نفسها دائماً. تجد الخوارزمية أكثر الملاحظات

تشابهاً مع تلك التي يجب أن تتنبأ بها والتي تستمد منها حدساً جيداً للإجابة المحتملة عن طريق حساب متوسط القيم المجاورة، أو عن طريق اختيار فئة الإجابة الأكثر شيوعاً فيما بينها.

يمكن لـ KNN التعامل بسهولة مع مئات التسميات [28]. عادة، يعمل KNN على معرفة جيران الملاحظة بعد استخدام مقياس المسافة مثل الإقليدية (الخيار الأكثر شيوعاً) أو مانهاتن (يعمل بشكل أفضل عندما يكون هناك العديد من الميزات الزائدة في البيانات). لا توجد قواعد مطلقة بشأن قياس المسافة الأفضل للاستخدام. حيث انها تعتمد على التجربة.

تعد خوارزمية k–الأقرب إلى الجار (kNN) خوارزمية تصويت لأنها تستخدم نقاط عينة k من حيث مقياس مسافة معين لتحديد فئة نقطة البيانات قيد الدراسة. من السهل فهم الفكرة الأساسية لهذه الخوارزمية، والخوارزمية سهلة التنفيذ نسبياً. يعتمد التصنيف على المعلومات والمسافات المحلية، وبالتالي يمكن أن تكون حدود القرار مرنة إلى حد ما للتعامل مع الأشكال المعقدة غير المنتظمة.

تُستخدم خوارزمية KNN على نطاق واسع نظراً لبساطتها، وهي فعالة بدرجة كافية في العديد من التطبيقات. تستخدمه العديد من الأنظمة عبر الإنترنت للتوصية بالمنتجات والأفلام. بالإضافة إلى ذلك، تستخدمه العديد من أنظمة تصنيف النصوص والصور.

KNN هي خوارزمية حساسة للقيم المتطرفة. قد يكون الجيران على حدود سحابة في مساحة البيانات حالات بعيدة، مما يتسبب في جعل التنبؤات غير منتظمة. لذك يجب تنظيف البيانات قبل استخدامها. يمكن أن يساعد استخدام (K–mean) أولاً في تحديد القيم المتطرفة المجمعة في مجموعات خاصة بهم.

الفصل الرابع التنفيذ العملي

4.1 مقدمة

سيتم في هذا الفصل التطرق الى أدوات الدراسة والية العمل المنجزة

4.2 أدوات الدراسة

- **Jupyter**

تطبيق Jupyter Notebook هو تطبيق خادم عميل يسمح بتحرير مستندات دفتر الملاحظات وتشغيلها عبر مستعرض ويب. يمكن تنفيذ تطبيق Jupyter Notebook على سطح مكتب محلي لا يتطلب الوصول إلى الإنترنت أو يمكن تثبيته على خادم بعيد والوصول إليه عبر الإنترنت.

بالإضافة إلى عرض / تحرير / تشغيل مستندات دفتر الملاحظات، يحتوي تطبيق Jupyter Notebook على "Dashboard" (لوحة معلومات Notebook)، و "لوحة تحكم" تعرض الملفات المحلية وتسمح بفتح مستندات دفتر الملاحظات [29].

- **Pandas**

هي أداة تحليل ومعالجة للبيانات مفتوحة المصدر سريعة وقوية ومرنة وسهلة الاستخدام، مبنية على قمة لغة برمجة Python. تم إنشاء Pandas فوق مكتبتين أساسيتين من مكتبات Python – matplotlib لتصور البيانات و NumPy للعمليات الرياضية. تعمل Pandas كغلاف فوق هذه المكتبات، مما يسمح بالوصول إلى العديد من أساليب matplotlib و NumPy برمز أقل [30].

قبل الباندا، استخدم معظم المحللين Python للتحكم في البيانات وإعدادها، ثم تحولوا إلى لغة أكثر تحديداً للمجال مثل R لبقية سير العمل. قدمت Pandas نوعين جديدين من الكائنات لتخزين البيانات

التي تجعل المهام التحليلية أسهل وتزيل الحاجة إلى تبديل الأدوات: السلسلة، التي لها هيكل يشبه القائمة، وأطر البيانات، التي لها هيكل جدولي.

- **NumPy**

(اختصار لـ Numerical Python) هي "الحزمة الأساسية للحوسبة العلمية باستخدام Python" [31]، العمليات باستخدام NumPy:

العمليات الحسابية والمنطقية على المصفوفات.تحولات فوريه وإجراءاتها الخاصة بمعالجة الشكل.

العمليات المتعلقة بالجبر الخطي: يحتوي NumPy على وظائف مضمنة للجبر الخطي وتوليد الأرقام العشوائية.

- **Matplotlib**

هي مكتبة شاملة لإنشاء تصورات visualizations ثابتة ومتحركة وتفاعلية في Python. يجعل Matplotlib الأشياء السهلة سهلة والأشياء الصعبة ممكنة.

تستخدم هذه المكتبة من أجل:

صناعة رسومات figures تفاعلية يمكنها التكبير والتحريك. تخصيص النمط والتخطيط المرئي. وتصدير إلى العديد من تنسيقات الملفات. تضمين في JupyterLab وواجهات المستخدم الرسومية [32].

- **Seaborn**

هي مكتبة لتصور البياناتvisualize مبنية على قمة matplotlib ومتكاملة بشكل وثيق مع هياكل بيانات الباندا في Python. التصور هو الجزء المركزي من Seaborn الذي يساعد في استكشاف البيانات وفهمها [33].

- **Scikit-learn**

scikit-Learn هي المكتبة الأكثر فائدة للتعلم الآلي في Python. تحتوي مكتبة sklearn على الكثير من الأدوات الفعالة للتعلم الآلي والنمذجة الإحصائية بما في ذلك التصنيف والانحدار والتكتل وتقليل الأبعاد. حيث يتم استخدام sklearn لبناء نماذج التعلم الآلي. لا يجوز استخدامها لقراءة البيانات ومعالجتها وتلخيصها. هناك مكتبات أفضل لذلك (مثل NumPy و Pandas وما إلى ذلك). تستخدم هذه المكتبة لأمور مثل: خوارزميات التعلم الخاضع للإشراف ,التحقق المتقاطع ,خوارزميات التعلم غير الخاضعة للإشراف [34].

- **مواصفات الحاسوب المستخدم**

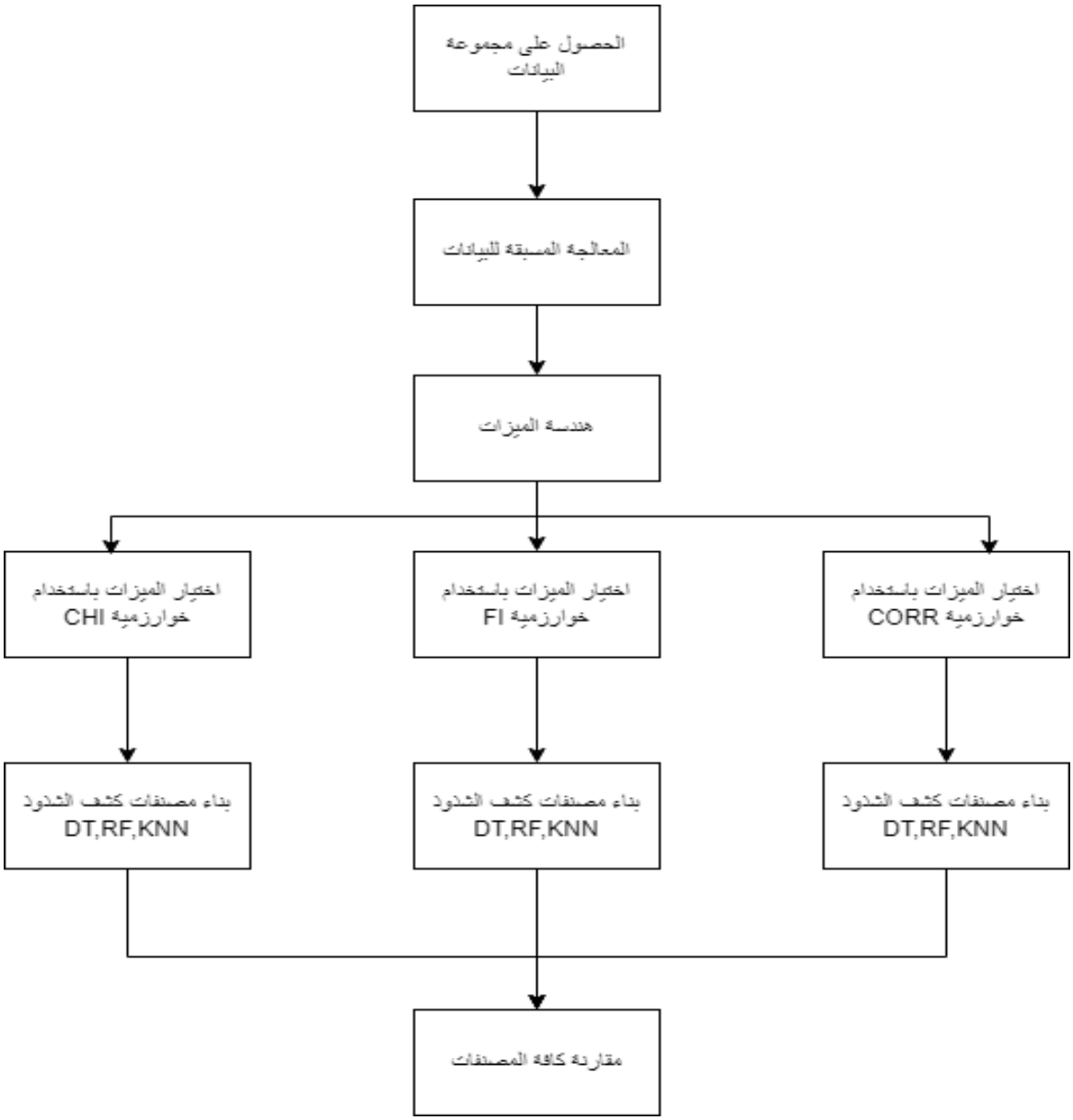
يجدر الذكر بأنه تم استخدام جهاز حاسوب محمول ذو المواصفات التالية كما يوضح الجدول4_1:

المواصفة	القيمة
تردد المعالج	GHZ 2.60
الذاكرة العشوائية RAM	Gb 16
عدد الأنوية	7
جيل الحاسوب	10

الجدول 4_1 يوضح مواصفات الحاسوب المستخدم

4.3 آلية العمل المنجزة

سيتم استخدام مجموعة البيانات الشبكية NSL KDD والقيام بالمعالجة المسبقة لها ومن ثم سيتم اختيار الميزات باستخدام ثلاث خوارزميات اختيار ميزات FI, CHI,CORR كل على حدى وسيتم بعدها استخدام الميزات المستخرجة (مجموعة البيانات الجزئية من مجموعة البيانات الكلية) لبناء ثلاث نماذج تعلم الي وهي DT,RF, KNN كما هي موضحة بالشكل 4_1 .



الشكل 4_1 مخطط العمل المقترح في البحث

الفصل الخامس النتائج والاختبارات

5.1 مجموعة البيانات المستخدمة

تم استخدام مجموعة البيانات NSLKDD في هذا البحث كونها مجموعة البيانات المعيارية التي يتم استخدامها في مجال اكتشاف الشذوذ الشبكي وهي عبارة عن مجموعة بيانات شبكية تحتوي على 41 ميزة (feature) تصف البيانات الشبكية بمافيها تسمية البيانات كبيانات طبيعية او هجمة شبكية تم الحصول على مجموعة البيانات من [35] وتتألف من مجموعة من الملفات كما هو موضح في الجدول 5_1:

جدول 5_1 ملفات مجموعة البيانات

اسم الملف	الوصف	لاحقة الملف
KDDTrain	مجموعة التدريب NSL– KDD الكاملة مع تسميات ثنائية	**ARFF**
KDDTrain	مجموعة التدريب NSL– KDD الكاملة بما في ذلك تسميات نوع الهجوم	**TXT**
KDDTrain_20Percent	مجموعة فرعية 20٪ من ملف KDDTrain.arff	**ARFF**
KDDTrain_20Percent	مجموعة فرعية 20٪ من ملف KDDTrain.txt	**TXT**
KDDTest	مجموعة اختبار NSL– KDD الكامل مع تسميات ثنائية	**ARFF**
KDDTest+	مجموعة اختبار NSL– KDD الكاملة بما في ذلك تسميات نوع الهجوم	**TXT**
KDDTest-21	مجموعة فرعية من ملف KDDTest +. arff	**ARFF**
KDDTest-21	مجموعة فرعية من ملف KDDTest + .txt	**TXT**

تتكون مجموعة بيانات التدريب من 21 هجوماً مختلفاً من أصل 37 هجوماً موجوداً في مجموعة بيانات الاختبار. أنواع الهجمات المعروفة هي تلك الموجودة في مجموعة بيانات التدريب بينما الهجمات الجديدة هي الهجمات الإضافية في مجموعة بيانات الاختبار، أي غير متوفرة في مجموعات بيانات التدريب. يتم تجميع أنواع الهجوم في أربع فئات: DoS و Probe و U2R و R2L. فيما يلي شرح للهجمات الأربعة المتواجدة في مجموعة البيانات:

1. DOS: رفض الخدمة هو فئة هجوم تستنفد موارد الضحية مما يجعلها غير قادرة على التعامل مع الطلبات المشروعة.

2. Probe: هدف هذه الجهمة هي الحصول على معلومات حول الضحية البعيدة.

3. U2R: الوصول غير المصرح به إلى امتيازات المستخدم الفائق المحلي (الجذر) هو نوع هجوم، يستخدم من خلاله المهاجم حساباً عادياً لتسجيل الدخول إلى نظام الضحية ويحاول الحصول على امتيازات الجذر / المسؤول عن طريق استغلال بعض نقاط الضعف في الضحية.

4. R2L: وصول غير مصرح به من جهاز بعيد، يتطفل المهاجم على آلة بعيدة ويحصل على وصول محلي للجهاز الضحية. على سبيل المثال تخمين كلمة المرور.

تحتوي مجموعة البيانات الأولى NSLKDDTRAIN على 125973 سجل تمثل البيانات الطبيعة نسبة 53.46% والباقي هجمات بمختلف انواعها المذكورة سابقاً، في حين تحتوي مجموعة البيانات الثانية NSLKDDTEST على 22544 سجل تمثل البيانات الطبيعية نسبة 43.08% من اجمالي البيانات والنسبة المتبقية هجمات.

يوضح الشكل 5_1 الهجمات الرئيسية في كل من مجموعة بيانات التدريب والاختبار [36].

Attacks in Dataset	Attack Type (37)
DOS	Back,Land,Neptune,Pod,Smurf, Teardrop,Mailbomb,Processtable,Udpstorm,Apache2,Worm
Probe	Satan,IPsweep,Nmap,Portsweep,Mscan,Saint
R2L	Guess_password,Ftp_write,Imap,Phf,Multihop,Warezmaster,Xlock,Xsnoop,Snmpguess,Snmpgetattack,Httptunnel,Sendmail, Named
U2R	Buffer_overflow,Loadmodule,Rootkit,Perl,Sqlattack,Xterm,Ps

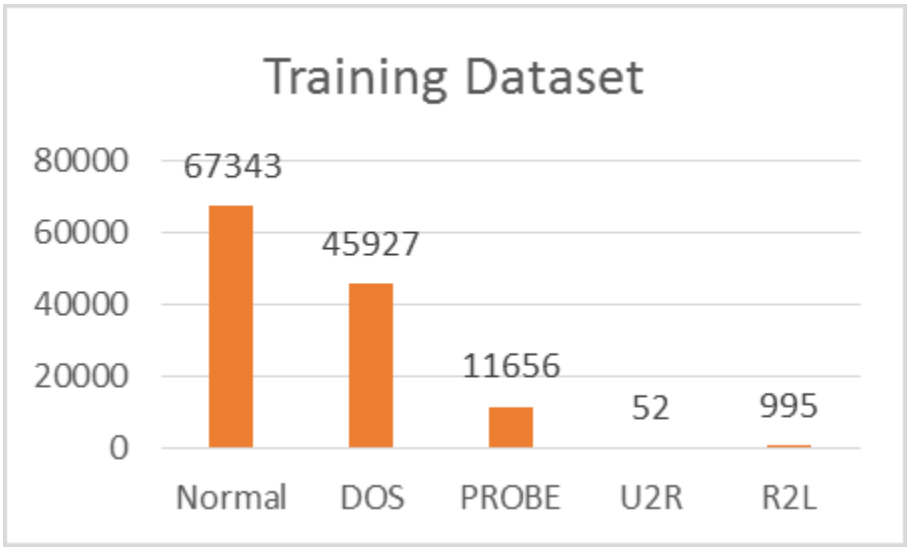
الشكل 5_1: الهجمات الرئيسية في مجموعة البيانات.

يوضح الجدول 2-5 توزع الهجمات الأساسية والهجمات الفرعية (sub calsses)، وكما تم ذكره سابقاً فإن مجموعة بيانات NSLKDDTEST تحتوي على الهجمات الفرعية نفسها المتواجدة في NSLKDDTRAIN بالإضافة الى هجمات اضافية جديدة.

جدول 5_2 توزع الهجمات الأساسية والهجمات الفرعية في مجموعتي البيانات

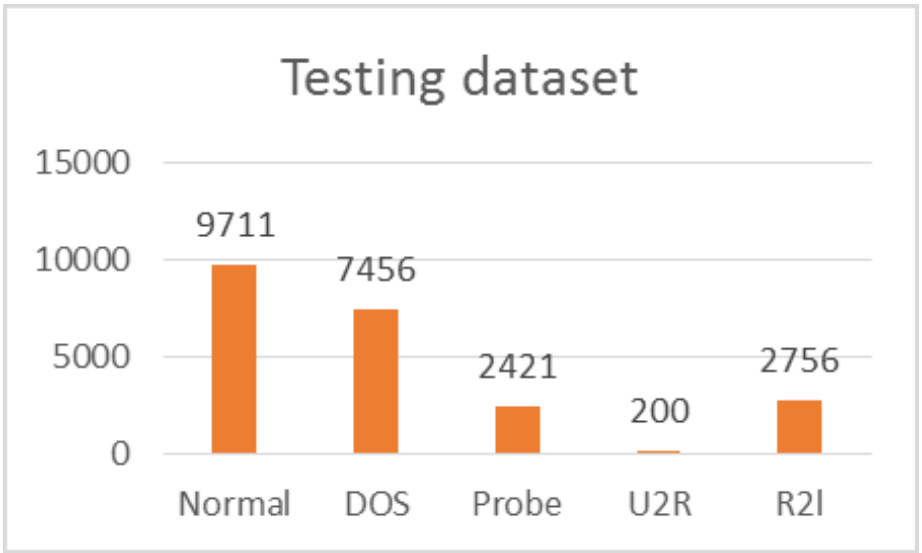
Attack in dataset	**In training dataset**	**In testing dataset**
DOS	6	**11**
Prob	4	**6**
R2L	7	**13**
U2R	4	**7**
Sum	21	**37**

في حين يوضح الشكل5_2 عدد الحالات الطبيعة وأنواع الهجمات في مجموعة بيانات التدريب من الواضح أن النصيب الأكبر من البيانات هي بيانات طبيعية وبالنسبة للهجمات كانت أقل عدد حالات هي لهجوم U2R وأكثر عدد حالات لهجوم DOS.



الشكل 5_2 : عدد الحالات الطبيعية وعدد الهجمات في مجموعة بيانات التدريب.

والشكل 5_3 يوضح عدد الحالات الطبيعة وأنواع الهجمات في مجموعة بيانات الإختبار من الواضح أن عدد حالات الهجمات أكثر من البيانات الطبيعية وزيادة عدد الحالات لهجمات U2Rو R2L عن مجموعة بيانات التدريب نظراً لوجود أنواع هجمات جديدة غير متواجدة في مجموعة بيانات التدريب.



الشكل 5_3: عدد الحالات الطبيعية وعدد الهجمات في مجموعة بيانات الاختبار.

5.2 التجارب

سيتم استخدام ثلاث خوارزميات لاختيار الميزات (Chisquare,Corr,Feature Importance) بالاضافة الى النموذج المقترح و ثلاث خوارزميات تعلم آلي (DT,RF,KNN) لتدريب النموذج حيث سيتم استخدام كل خوارزمية اختيار ميزات على حدى مع خوارزميات التعلم الالي الثلاث.

في كل مرة سيتم اختيار خوارزمية اختيار ميزات وتجريبها على خوارزميات التعلم الآلي الثلاث سيتم تقسيم مجموعة بيانات التدريب NSLKDDTRAIN الى

80% تدريب training	20% تحقق validation

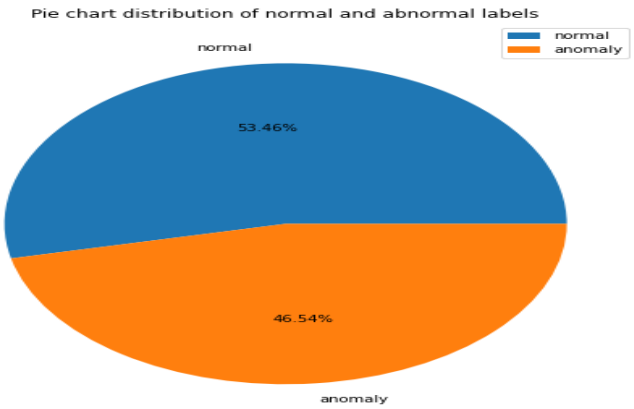
وذلك لكي نقيم قدرة النموذج على اكتشاف الهجمات التي تدرب عليها. ثم سيتم اختبارها على كامل بيانات الاختبار NSLKDDTEST لتقييم قدرة كل نموذج على اكتشاف أنواع جديدة من الهجمات من خلال مقاييس الأداء المذكورة سابقاً.

- **تحليل وتحضير البيانات**

في البداية تم العمل على تحليل البيانات وذلك من خلال القيام ببعض العمليات الاحصائية حيث يتم التعرف على عدد السجلات ومتوسط القيم الموجودة للبيانات الرقمية واكبر قيمة واصغر قيمة في الميزات ليتم اكتشاف البيانات ويكون هناك تصور واضح عن ماهية البيانات التي يتم التعامل معها.

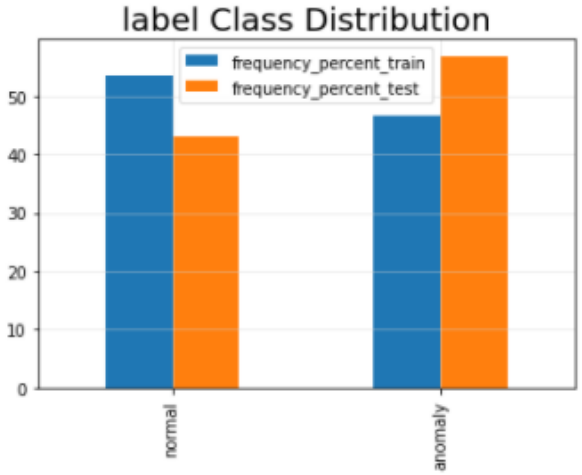
ليتم بعد ذلك احتساب مجموعة القيم المتواجدة ضمن الميزات (أي مجموعة القيم الفريدة الموجودة ضمن هذه الميزات) حيث لم تحتو الميزة num_outbound_cmds سوى على قيمة فريدة واحدة في كل السجلات وهي القيمة صفر في مجموعتي البيانات التدريب والاختبار لذلك يجب حذفها كونها تحتوي على قيمة فريدة واحدة فقط.

ثم يتم بعد ذلك تجميع الهجمات كلها ووضعها تحت مسمى غير طبيعي (anomaly) ويتم اظهار القيم الاحصائية كما هو موضح في الشكل 5_4 فكانت نسبة البيانات الطبيعية 53% بالتالي تعتبر البيانات متوازنة ولاحاجة لإجراء أي عملية لموازنة البيانات.



الشكل 5_4:توزع البيانات الطبيعة والشاذة في مجموعة بيانات التدريب

يوضح الشكل 5-5 توزع البيانات الطبيعية والشاذة في مجموعتي بيانات التدريب والاختبار بعد تحويل الهجمات الى شذوذ.



الشكل 5-5: توزع البيانات الطبيعة والشاذة في مجموعتي بيانات التدريب والإختبار

- **تطبيع البيانات**

تطبيع الميزة هو طريقة مستخدمة لتطبيع نطاق المتغيرات المستقلة أو ميزات البيانات. في معالجة البيانات، يُعرف أيضاً باسم تطبيع البيانات ويتم إجراؤه عموماً أثناء خطوة المعالجة المسبقة للبيانات.

وبعد عملية اظهار البيانات بشكل احصائي تم تقييس البيانات الرقمية باستخدام StandardScaler حيث يتم استخدام StandardScaler لتغيير حجم توزيع القيم بحيث يكون متوسط القيم الملاحظة صفر والانحراف المعياري هو 1.

وفي المرحلة الثانية تم العمل على ترميز البيانات الفئوية categorical وهي ['protocol_type', 'service', 'flag'] باستخدام LabelEncoder حيث يعدّ LabelEncoder أسلوب ترميز شائع للتعامل مع المتغيرات الفئوية. في هذه التقنية، يتم تعيين عدد صحيح فريد لكل تسمية بناءً على الترتيب الأبجدي.

5.2.1 التجربة الأولى

سيتم استخدام خوارزمية Chi-Square لاختيار الميزات وبعدها استخدام هذه الميزات لتدريب النموذج باستخدام ثلاثة خوارزميات وهي DT وRF وKNN كل على حدى مع العلم ان كافة خوارزميات التعلم الآلي التي تم استخدامها قد تم الإبقاء على بارامتراتها الافتراضية المقدمة من مكتبة sicikt learn.

- **تقييم أداء الخوارزمية:**

يتم تقييم أداء الخوارزميات بناء على المعايير التي عرجنا عليها سابقاً على كل من بيانات التدريب وبيانات الاختبار وعلى كامل مجموعة الاختبار NSLKDDTEST.

1. **معيار الدقة accuracy**

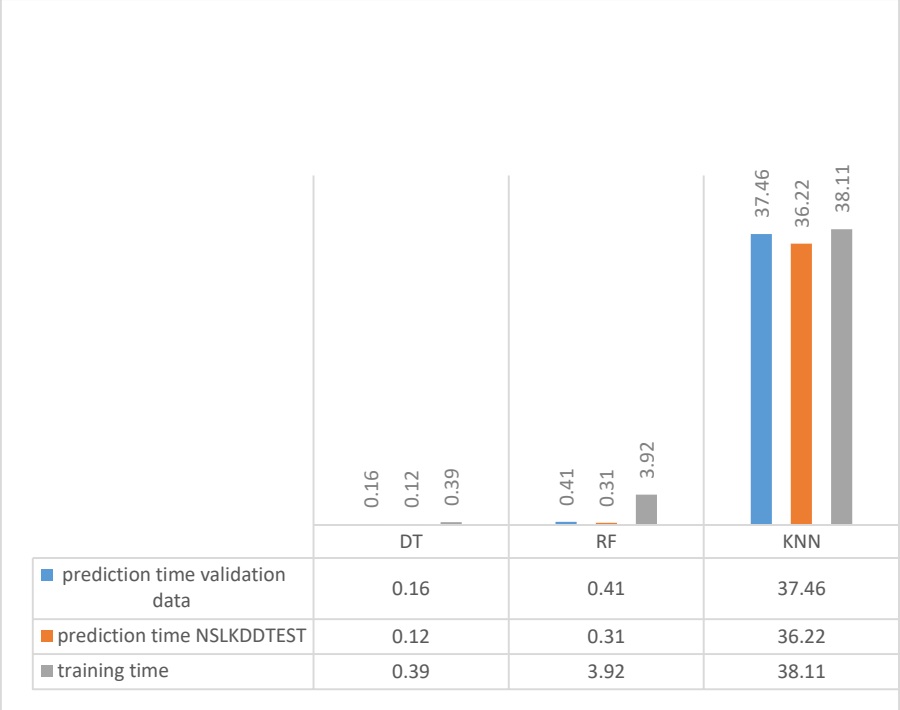
يوضح الشكل 5_6 دقة مصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية Chi-Square.

الشكل 5_6: دقة المصنفات بعد استخدام ميزات المختارة من قبل خوارزمية ChiSquare.

يظهر المخطط ان النتائج متقاربة جداً في حين يظهر تفوق بسيط لمصنف الغابة العشوائية.

1. **معيار الوقت اللازم لبناء النموذج واختباره**

يوضح الشكل 5_7 الوقت اللازم لتدريب والوقت اللازم لكشف طبيعة البيانات لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية Chi-Square.



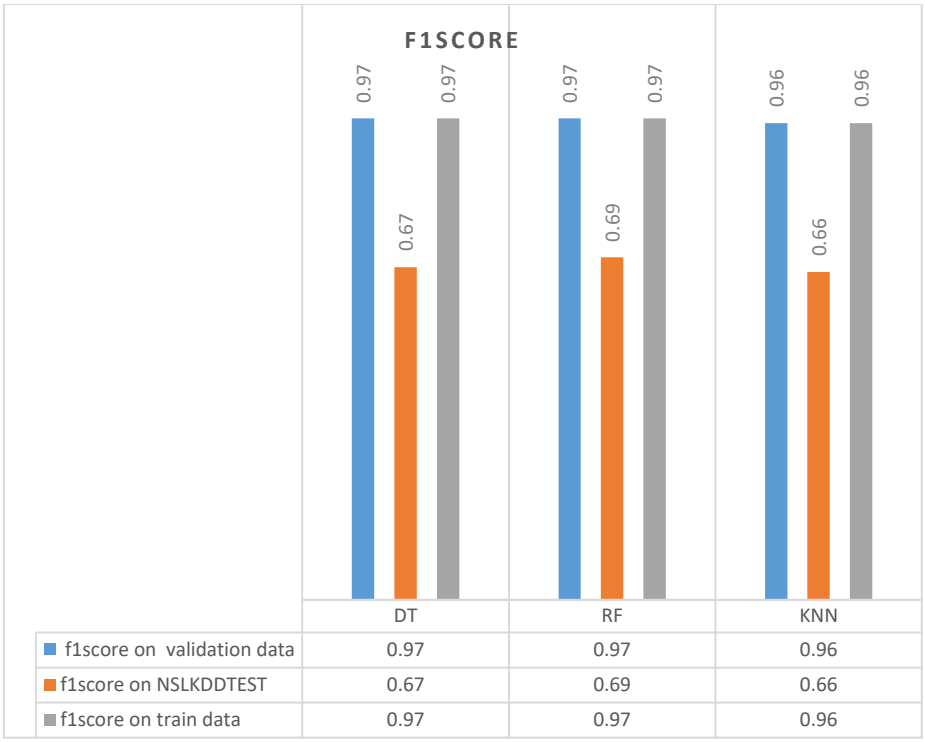
	DT	RF	KNN
prediction time validation data	0.16	0.41	37.46
prediction time NSLKDDTEST	0.12	0.31	36.22
training time	0.39	3.92	38.11

الشكل 5_7:الوقت اللازم لتدريب المصنفات واكتشاف طبيعة البيانات باستخدام ميزات Chi-Square.

من الواضح ان خوارزمية اقرب جار قد أعطت اسوء أداء في الوقت اللازم للتدريب ولاكتشاف طبيعة البيانات في حين مصنف شجرة القرار قد اعطى اقل وقت في التدريب واكتشاف طبيعة البيانات.

1. معيار f1 score

يوضح الشكل 5_8 معيار f1score لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية Chi-Square.



	DT	RF	KNN
f1score on validation data	0.97	0.97	0.96
f1score on NSLKDDTEST	0.67	0.69	0.66
f1score on train data	0.97	0.97	0.96

الشكل 5_8: f1score لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل chisquare.

يظهر من المخطط تقارب النتائج مع تفضيل لمصنف الغابة العشوائية.

5.2.2 التجربة الثانية

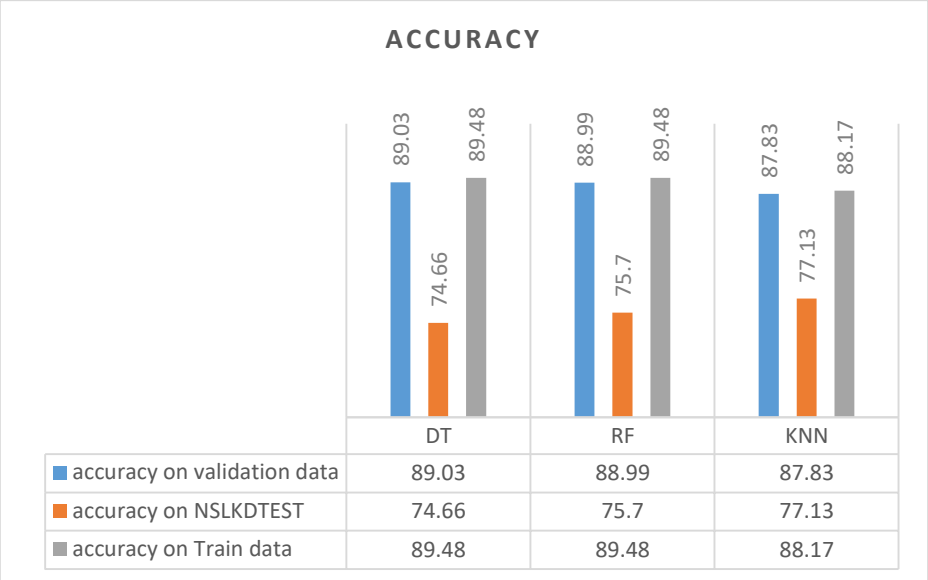
سيتم استخدام خوارزمية الارتباط لاختيار الميزات وبعدها استخدام هذه الميزات لتدريب النموذج باستخدام ثلاثة خوارزميات وهي DT و RF و KNN كل على حدى حيث تم اختيار الميزات المرتبطة بارتباط جيد جداً مع المتغير التابع (corr_y >0.7)

- **تقييم أداء الخوارزمية:**

يتم تقييم أداء الخوارزميات بناء على المعايير التي عرجنا عليها سابقاً على كل من بيانات التدريب وبيانات الاختبار وعلى كامل مجموعة الاختبار NSLKDDTEST.

2. معيار الدقة accuracy

يوضح الشكل 5_9 دقة مصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية الارتباط.



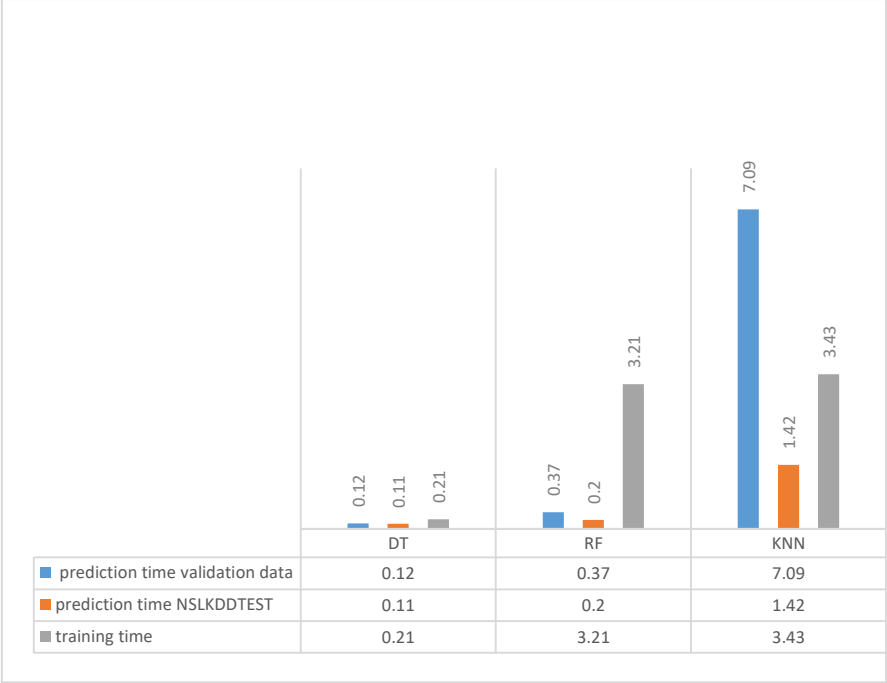
	DT	RF	KNN
accuracy on validation data	89.03	88.99	87.83
accuracy on NSLKDTEST	74.66	75.7	77.13
accuracy on Train data	89.48	89.48	88.17

الشكل 5_9: دقة المصنفات بعد استخدام ميزات المختارة من قبل خوارزمية الارتباط.

من الواضح من المخطط ان مصنف الغابة العشوائية أفضل على بيانات الاختبار في حين مصنف شجرة القرار افضل على بيانات التحقق.

2. معيار الوقت اللازم لبناء النموذج واختباره

يوضح الشكل 5_10 الوقت اللازم لتدريب والوقت اللازم لكشف طبيعة البيانات لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية الارتباط.

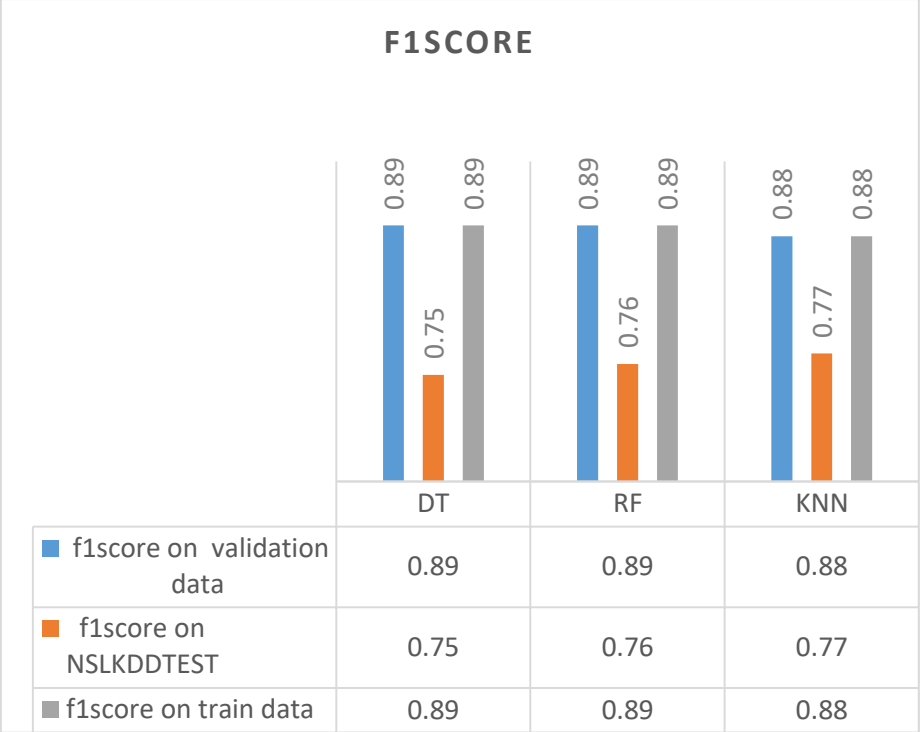


الشكل 5_10:الوقت اللازم لتدريب المصنفات واكتشاف طبيعة البيانات باستخدام ميزات الارتباط.

من الواضح من المخطط ان مصنف شجرة القرار هو الأفضل بكافة المقاييس.

2. معيار f1 score

يوضح الشكل 5_11 معيار f1score لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية الارتباط.



الشكل 5_11: f1score لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل الارتباط.

يوضح المخطط تقارب النتائج في حين كان مصنف الغابة العشوائية هو الأفضل.

5.2.3 التجربة الثالثة

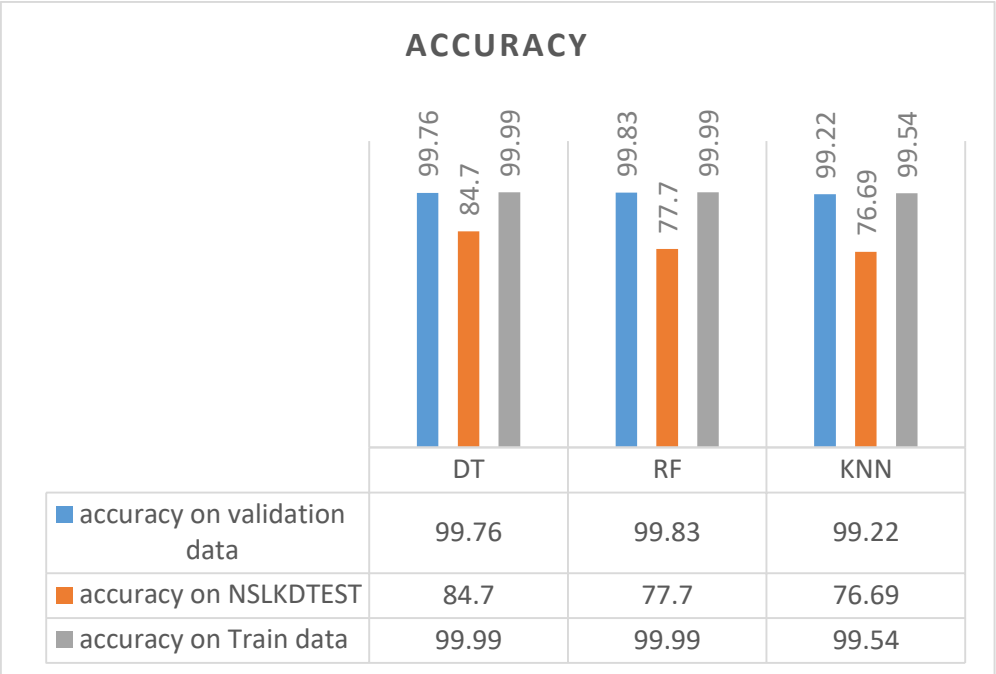
سيتم استخدام خوارزمية Feature Importance لاختيار الميزات وبعدها استخدام هذه الميزات لتدريب النموذج باستخدام ثلاثة خوارزميات وهي DT و RF و KNN كل على حدى.

- **تقييم أداء الخوارزمية:**

يتم تقييم أداء الخوارزميات بناء على المعايير التي عرجنا عليها سابقاً على كل من بيانات التدريب وبيانات الاختبار وعلى كامل مجموعة الاختبار NSLKDDTEST.

3. معيار الدقة accuracy

يوضح الشكل 5_12 دقة مصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية الارتباط.



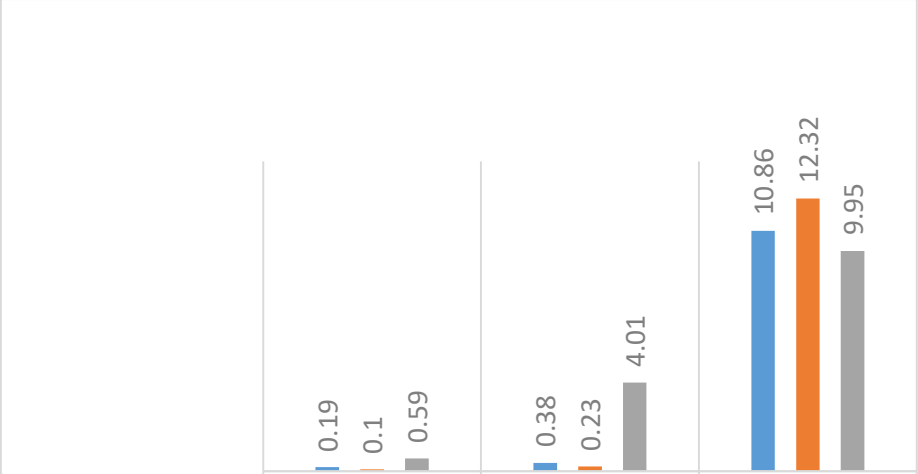
	DT	RF	KNN
accuracy on validation data	99.76	99.83	99.22
accuracy on NSLKDTEST	84.7	77.7	76.69
accuracy on Train data	99.99	99.99	99.54

الشكل 5_12: دقة المصنفات بعد استخدام ميزات المختارة من قبل خوارزمية FI.

من المخطط السابق نرى ان مصنف شجرة القرار هو الأفضل على الرغم من تقارب النتائج مع باقي المصنفات.

3. معيار الوقت اللازم لبناء النموذج واختباره

يوضح الشكل 5_13 الوقت اللازم لتدريب والوقت اللازم لكشف طبيعة البيانات لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية FI.



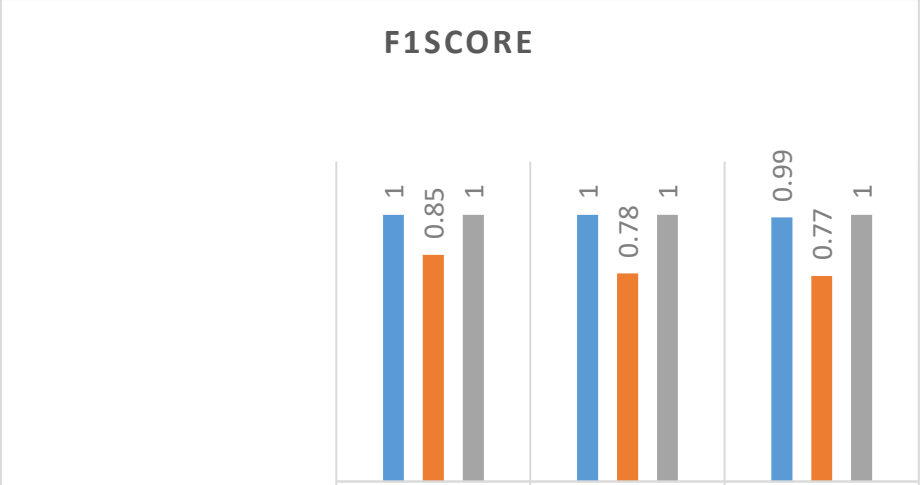
	DT	RF	KNN
prediction time validation data	0.19	0.38	10.86
prediction time NSLKDDTEST	0.1	0.23	12.32
training time	0.59	4.01	9.95

الشكل 5_13:الوقت اللازم لتدريب المصنفات واكتشاف طبيعة البيانات باستخدام ميزات FI.

من المخطط السابق نرى ان مصنف شجرة القرار هو الأفضل على الرغم من تقارب النتائج مع باقي المصنفات.

3. معيار f1 score

يوضح الشكل 5_14 معيار f1score لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل خوارزمية FI.



	DT	RF	KNN
f1score on validation data	1	1	0.99
f1score on NSLKDDTEST	0.85	0.78	0.77
f1score on train data	1	1	1

الشكل 5_14: f1score لمصنفات كشف الشذوذ بعد تدريبها على الميزات التي تم اختيارها من قبل FI.

من المخطط السابق نرى ان مصنف شجرة القرار هو الأفضل على الرغم من تقارب النتائج مع باقي المصنفات.

الفصل السادس التقييم والمقارنة

6.1 مقدمة

سيتم في هذا التطرق الى مقاييس الأداء والتقييم ومقارنة النتائج

6.2 مقاييس الاداء

6.2.1 دقة التصنيف accuracy

دقة التصنيف هي ما نعنيه عادةً عندما نستخدم مصطلح الدقة. إنها نسبة عدد التنبؤات الصحيحة إلى العدد الإجمالي لعينات الإدخال.

إنه يعمل بشكل جيد فقط إذا كان هناك عدد متساوٍ من العينات التي تنتمي إلى كل فئة. على سبيل المثال، لنضع في الاعتبار أن هناك 98٪ عينات من الفئة "أ" و 2٪ عينات من الفئة "ب" في مجموعة التدريب الخاصة بنا. بعد ذلك، يمكن أن يحصل نموذجنا بسهولة على 98٪ من دقة التدريب عن طريق التنبؤ بكل عينة تدريب تنتمي إلى الفئة أ. عندما يتم اختبار نفس النموذج على مجموعة اختبار مع 60٪ عينات من الفئة أ و 40٪ عينات من الفئة ب، فإن دقة الاختبار تنخفض إلى 60٪. دقة التصنيف رائعة، لكنها تعطي إحساساً زائفاً بتحقيق الدقة العالية. تبرز المشكلة الحقيقية، عندما تكون تكلفة التصنيف الخاطئ لعينات الصنف الثانوي عالية جداً. إذا تعاملنا مع مرض نادر ولكنه قاتل، فإن تكلفة الفشل في تشخيص مرض شخص مريض أعلى بكثير من تكلفة إرسال شخص سليم لمزيد من الفحوصات [37].

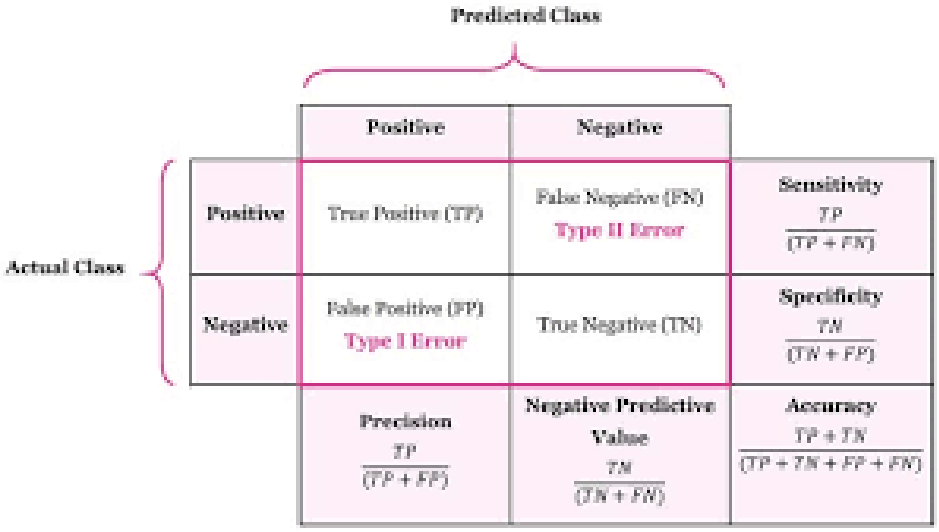
وتعطى الدقة بالمعادلة 1:

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \quad \ldots(6.1)$$

6.2.2 Confusion Matrix

مصفوفة الارتباك كما يوحي الاسم تعطي مصفوفة كإخراج وتصف الأداء الكامل للنموذج.

لنفترض أن لدينا مشكلة تصنيف ثنائي. لدينا بعض العينات التي تنتمي إلى فئتين: نعم أو لا. أيضاً، لدينا المصنف الخاص بنا والذي يتوقع فئة لعينة إدخال معينة [37] يوضح الشكل 6_1 مصفوفة الارتباك:



الشكل 6_1: مصفوفة الارتباك.

المصدر [38]

هناك 4 مصطلحات مهمة موضحة بالشكل 2-6:

- ❖ الإيجابيات الحقيقية TP: الحالات التي توقعنا فيها نعم وكان الناتج الفعلي نعم أيضاً.
- ❖ سلبيات حقيقية TN: الحالات التي توقعنا فيها لا وكان الناتج الفعلي لا.
- ❖ الإيجابيات الكاذبةFP: الحالات التي توقعنا فيها نعم وكان الناتج الفعلي لا.
- ❖ السلبيات الكاذبة FN : الحالات التي توقعنا فيها لا والفعلي كان الإخراج نعم.

يمكن حساب دقة المصفوفة بأخذ متوسط القيم الموجودة عبر "القطر الرئيسي" كما هو موضح بالمعادلة 2:

$$\text{Accuracy} = \frac{\text{TruePositive+TrueNegative}}{\text{TotalSample}} \quad \ldots(6.2)$$

6.2.3 Precision الدقة

هي عدد النتائج الإيجابية الصحيحة مقسوماً على عدد النتائج الإيجابية التي تنبأ بها المصنف كما هو موضح في المعادلة 3:

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TotalPredicted Positive}} \quad \ldots(6.3)$$

بعبارات أبسط، الدقة هي النسبة بين الإيجابيات الحقيقية وجميع الإيجابيات.

6.2.4 Recall الاسترجاع

هو عدد النتائج الإيجابية الصحيحة مقسوماً على عدد جميع العينات ذات الصلة (جميع العينات التي كان ينبغي تحديدها على أنها إيجابية) [37].

الاسترجاع هو مقياس النموذج الذي يحدد الإيجابيات الحقيقية بشكل صحيح. رياضياً موضح بالمعادلة 4:

$$\text{Recall}=\frac{\text{TruePositive}}{\text{TruePositive+FasleNegative}} \quad \ldots(6.4)$$

6.2.5 F1 Score

F1 Score هو المتوسط التوافقي بين الدقة والاسترجاع. نطاق نقاط F1 هو [0، 1]. يخبر بمدى دقة المصنف(عدد الحالات التي يصنفها بشكل صحيح)، بالإضافة إلى مدى قوته.

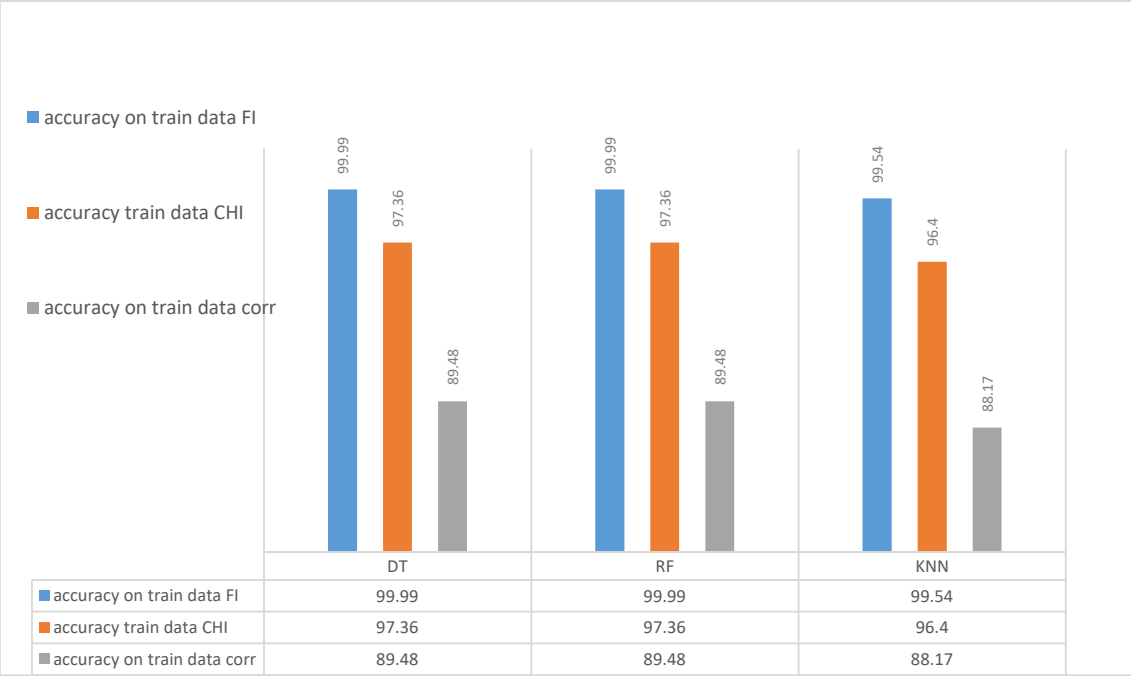
يمنح الدقة العالية ولكن الاسترجاع الأقل دقة بالغة، ولكنها تفوت بعد ذلك عدداً كبيراً من الحالات التي يصعب تصنيفها. كلما زادت درجة F1، كان أداء النموذج أفضل. تحاول F1 Score إيجاد التوازن بين الدقة والاستدعاء[37]، رياضياً موضح بالمعادلة 5.

$$F1score=\frac{2*Precision*Recall}{Precision+Recall} \quad ...(6.5)$$

6.3 المقارنة

نورد في مايلي ملخصاً لجميع نتائج التجارب:

حيث يوضح 6_2 الدقة على بيانات التدريب (80% NSLKDDTRAIN).

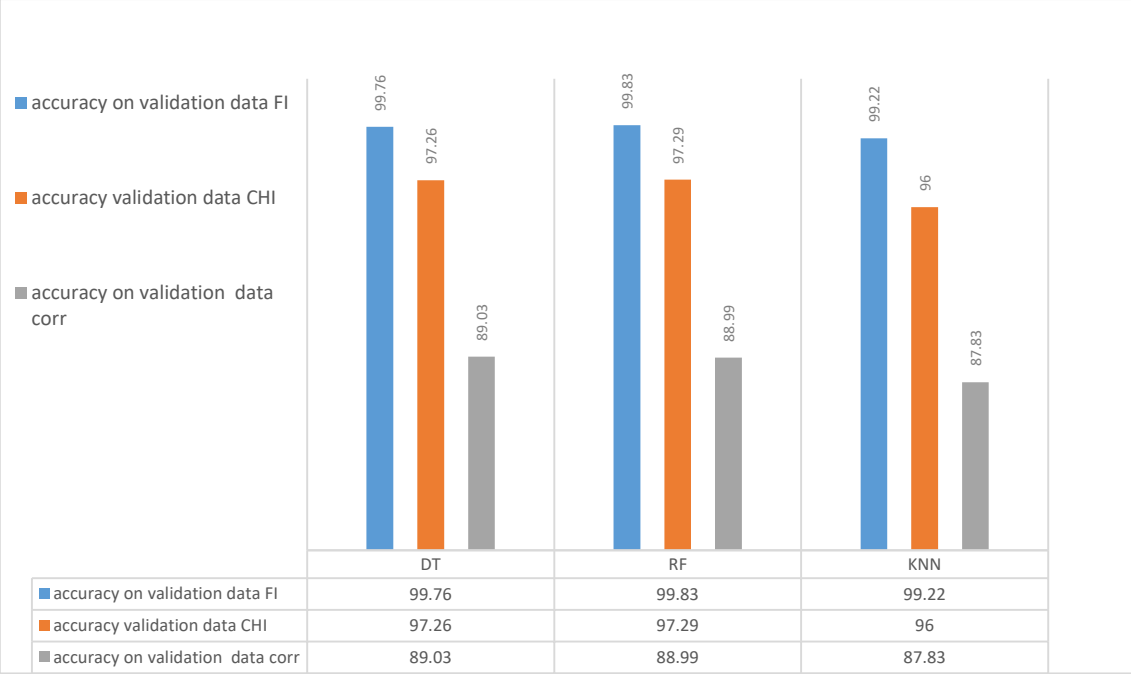


	DT	RF	KNN
accuracy on train data FI	99.99	99.99	99.54
accuracy train data CHI	97.36	97.36	96.4
accuracy on train data corr	89.48	89.48	88.17

الشكل 6_2: ملخص لنتائج دقة الخوارزميات على بيانات التدريب

من الواضح ان النتائج التي تم الحصول عليها على بيانات التدريب باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 99.99 .

حيث يوضح الشكل 6_3 الدقة على بيانات التحقق (20% NSLKDDTRAIN).

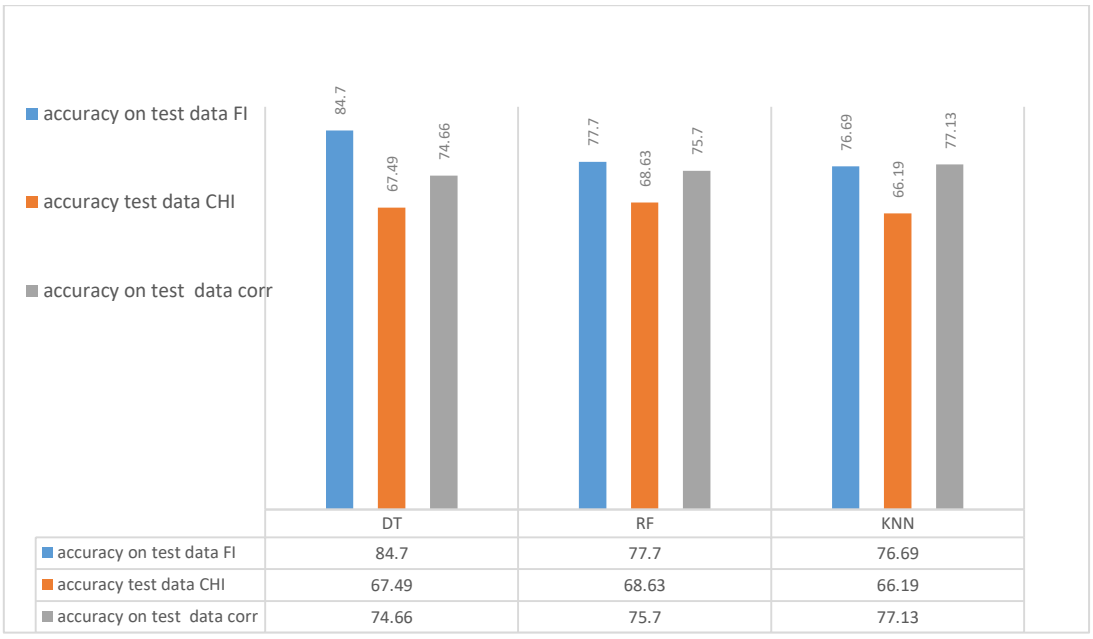


	DT	RF	KNN
accuracy on validation data FI	99.76	99.83	99.22
accuracy validation data CHI	97.26	97.29	96
accuracy on validation data corr	89.03	88.99	87.83

الشكل 6_3: ملخص لنتائج دقة الخوارزميات على بيانات التحقق

من الواضح ان النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 99.83 و 99.76 على التوالي.

يوضح الشكل 6_4: ملخص لنتائج دقة الخوارزميات على بيانات الاختبار

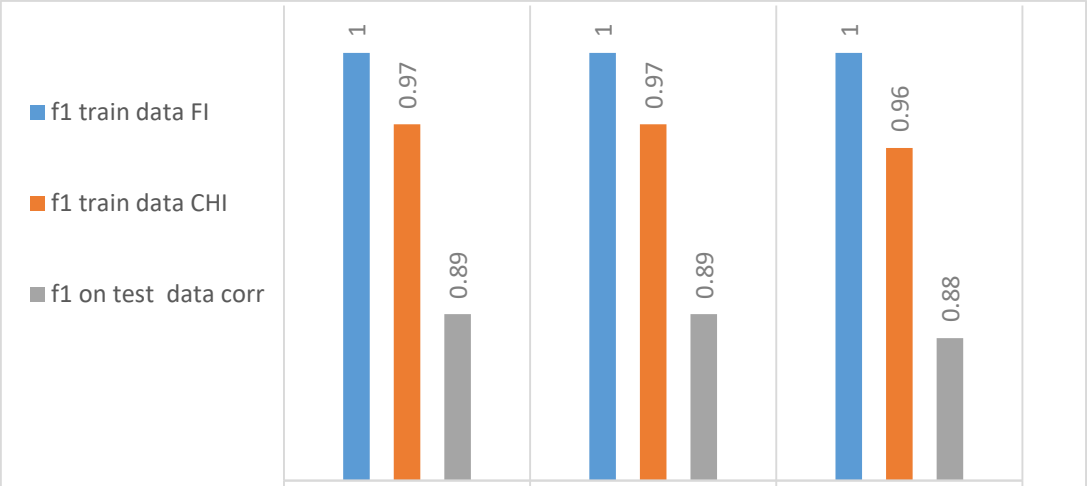


	DT	RF	KNN
accuracy on test data FI	84.7	77.7	76.69
accuracy test data CHI	67.49	68.63	66.19
accuracy on test data corr	74.66	75.7	77.13

الشكل 6_4: ملخص لنتائج دقة الخوارزميات على بيانات الاختبار

من الواضح ان النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف شجرة القرار على 84.7.

يوضح الشكل 6_5 ملخص لنتائج f1score للخوارزميات على بيانات التدريب

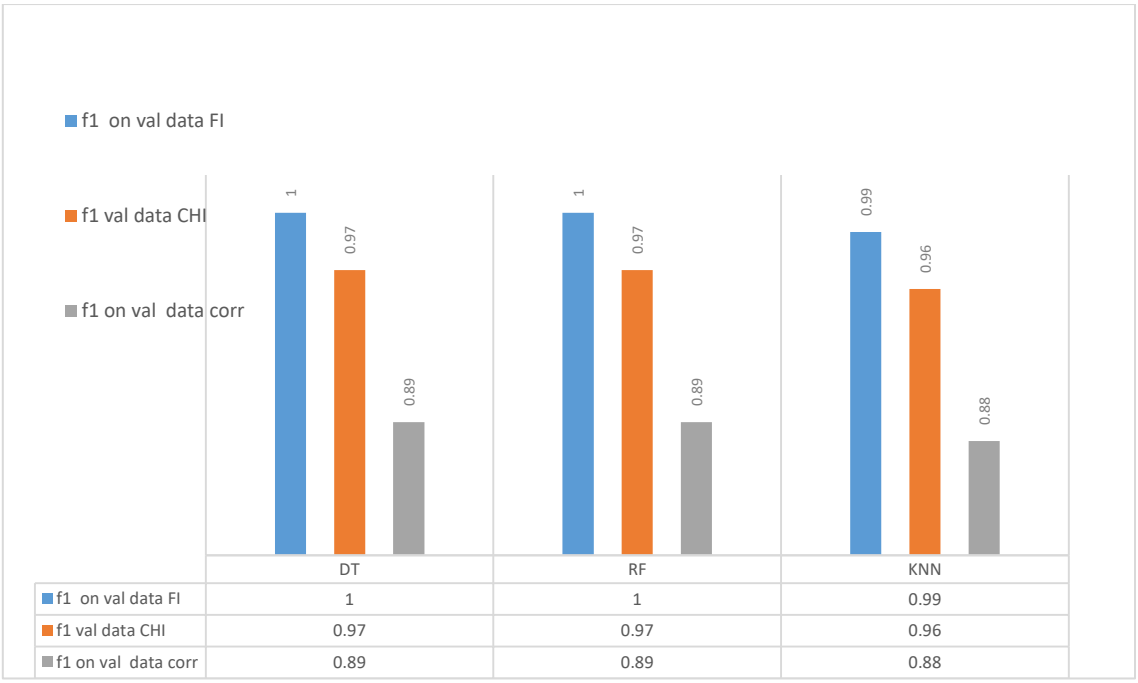


	DT	RF	KNN
f1 train data FI	1	1	1
f1 train data CHI	0.97	0.97	0.96
f1 on test data corr	0.89	0.89	0.88

الشكل 6_5: ملخص لنتائج f1score الخوارزميات على بيانات التدريب

من الواضح ان النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 1.

يوضح الشكل 6-6 ملخص لنتائج f1score الخوارزميات على بيانات التحقق

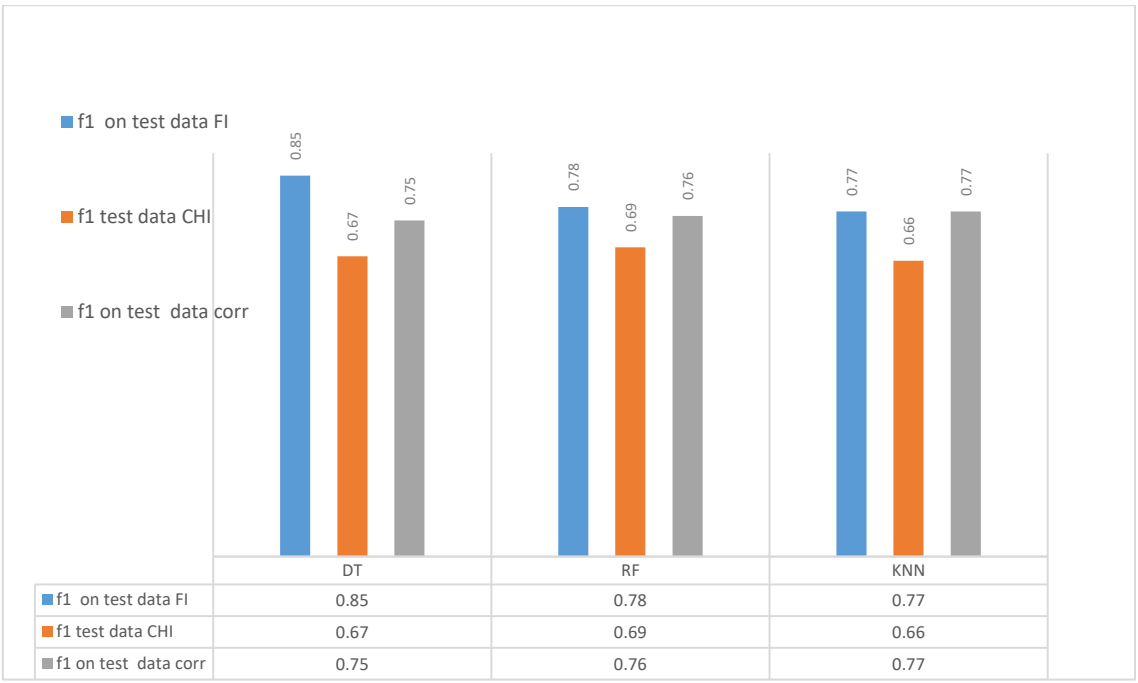


	DT	RF	KNN
f1 on val data FI	1	1	0.99
f1 val data CHI	0.97	0.97	0.96
f1 on val data corr	0.89	0.89	0.88

الشكل 6-6: ملخص لنتائج f1score الخوارزميات على بيانات التحقق

من الواضح ان النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 1 .

يوضح الشكل 6_7 ملخص لنتائج f1score الخوارزميات على بيانات الاختبار.

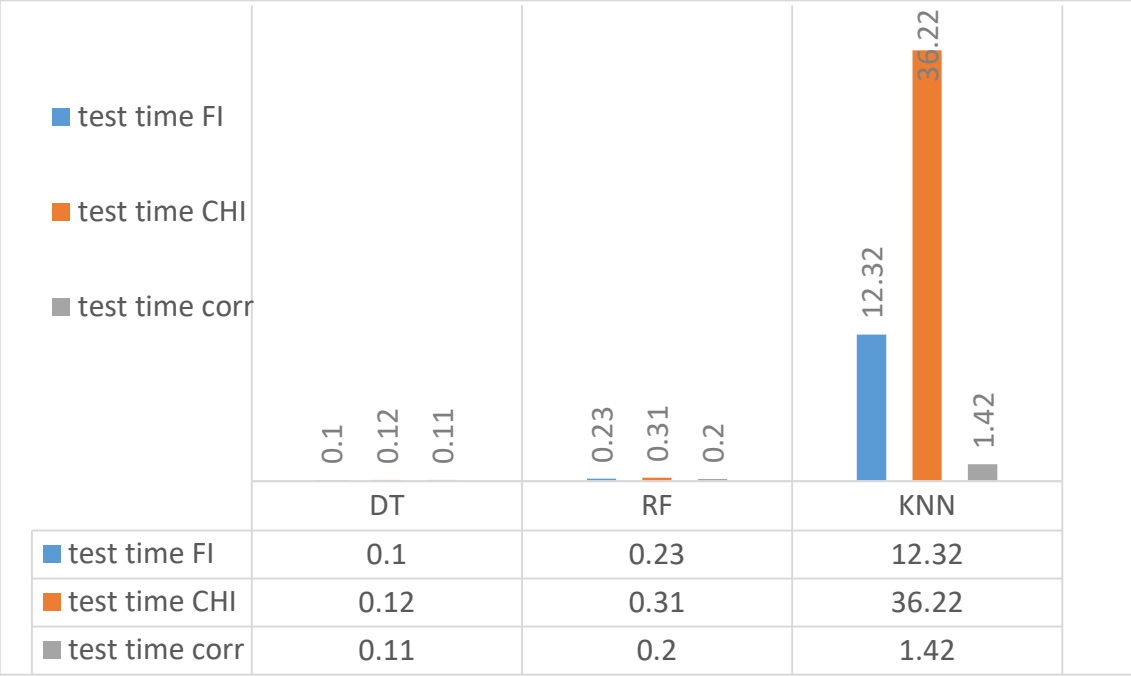


	DT	RF	KNN
f1 on test data FI	0.85	0.78	0.77
f1 test data CHI	0.67	0.69	0.66
f1 on test data corr	0.75	0.76	0.77

الشكل 6_7: ملخص لنتائج f1score الخوارزميات على بيانات الاختبار

من الواضح ان النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف شجرة القرار على 0.85.

يوضح الشكل 6_8 ملخص لنتائج وقت الكشف على بيانات الاختبار.



	DT	RF	KNN
test time FI	0.1	0.23	12.32
test time CHI	0.12	0.31	36.22
test time corr	0.11	0.2	1.42

الشكل 6_8: ملخص لنتائج وقت الكشف على بيانات الاختبار

يظهر المخطط تقارب في النتائج بين خوارزمية FI و corr مع تفوق بسيط لخوارزمية FI في النهاية.

1. النتائج التي تم الحصول عليها على بيانات التدريب باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 99.99 .
2. النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 99.83 و 99.76 على التوالي.
3. النتائج التي تم الحصول عليها باستخدام ميزات FI هي الأفضل حيث حصل مصنف شجرة القرار على 84.7 .
4. النتائج التي تم الحصول عليها بمعيار f1 score باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 1 على بيانات التدريب.
5. النتائج التي تم الحصول عليها بمعيار f1score باستخدام ميزات FI هي الأفضل حيث حصل مصنف الغابة العشوائية وشجرة القرار على 1 على بيانات التحقق.
6. النتائج التي تم الحصول عليهابمعيار f1score باستخدام ميزات FI هي الأفضل حيث حصل مصنف شجرة القرار على 0.85 على بيانات الاختبار.

الفصل السابع النتائج والافاق المستقبلية

7.1 النتائج

تم في هذا البحث :

.1

تقييم تطور طرائق كشف الشذوذ الشبكي، بدءًا من الطرق الإحصائية التقليدية إلى طرائق التعلم الآلي حيث أظهرت الطرق التقليدية عجزًا في مواجهة تطور الشبكات الحاسوبية وزيادة كمية حركة المرور وانماط الشذوذ الجديدة، مما أدى إلى ضعفها في اكتشاف أنماط الشذوذ بكفاءة وفعالية عالية.

.2

تم تحديد هذه العجز واستيفاده في طرائق التعلم الآلي، التي تتميز بدراسة الشذوذ في البيانات الشبكية بدقة وسرعة عالية. ومع ذلك، ما زالت هذه الطرق تواجه مشكلة اختيار الميزات لتدريب خوارزميات اكتشاف الشذوذ، هذه الميزات تمثل المعارف الحقيقية المهمة في الشبكة، لكن عملية اختيارها ما تزال قيد الدراسة الحالية وتحتاج إلى طرائق أفضل للحصول على أفضل الميزات لتدريب الخوارزميات عليها.

.3

أثناء هذا البحث، تم تقييم فعالية ثلاث خوارزميات اختيار الميزات هي Chi-Square و Corr و Feature Importance على أداء مصنفات التعلم الآلي.

.4

تم استخدام هذه الخوارزميات مع ثلاثة خوارزميات اكتشاف شذوذ الشبكة هي KNN و DT و RF باستخدام معايير الدقة f1–score والزمن اللازم لتدريب النموذج واكتشاف البيانات الشاذة.

.5

أثبتت النتائج تأثير الميزات على أداء مصنفات كشف الشذوذ الشبكي، مما يسلط الضوء على أهمية اختيار الميزات المناسبة لتطوير خوارزميات اكتشاف الشذوذ الشبكي الأكفأ.

.6

وقد أظهرت النتائج ان الميزات التي تم اختيارها باستخدام خوارزمية FI هي الأفضل من باقي الخوارزميات حيث كان مصنف الغابة العشوائية وشجرة القرار هما الأفضل من باقي المصنفات بدرجة f1 وصلت ل 1.

7.2 الآفاق المستقبلية

بالنسبة للآفاق المستقبلية برز التعلم العميق كأداة قوية للكشف عن الشذوذ في شبكات الكمبيوتر، حيث يوفر العديد من المزايا مقارنة بالطرق التقليدية. إن قدرته على اكتشاف أنماط الشذوذ الجديدة تجعله فعالاً بشكل خاص في هذا المجال. على عكس خوارزميات التعلم الآلي التقليدية، فإن نماذج التعلم العميق أقل عرضة لمشكلة عدم التوازن في البيانات، والتي تنتشر في سيناريوهات اكتشاف الشذوذ. تسمح هذه الخاصية لأنظمة التعلم العميق بتحديد الأنماط غير العادية حتى عندما تحدث بشكل غير متكرر أو غير متوقع.

تمكن بنية الشبكة العصبية لنماذج التعلم العميق من تعلم الأنماط المعقدة في بيانات حركة المرور على الشبكة. يمكن لهذه النماذج التقاط الاختلافات الدقيقة في السلوك التي قد تشير إلى نشاط غير طبيعي. على سبيل المثال، يمكن لنظام التعلم العميق اكتشاف ارتفاعات غير عادية في استخدام النطاق الترددي، أو أنماط الاتصال غير المتوقعة بين الأجهزة، أو محاولات تسجيل الدخول غير المعهودة من عناوين IP غير مألوفة.

تتمثل إحدى المزايا الرئيسية للتعلم العميق للكشف عن الشذوذ في قدرته على تعلم تمثيلات الميزات تلقائيًا مباشرة من بيانات الشبكة الخام. وهذا يلغي الحاجة إلى هندسة الميزات اليدوية، والتي يمكن أن تستغرق وقتًا طويلاً وعرضة للتحيز البشري. بدلاً من ذلك، يمكن لنماذج التعلم العميق استخراج الميزات ذات الصلة تلقائيًا، مما يسمح لها بالتكيف مع ظروف الشبكة المتغيرة ومناظر التهديدات المتطورة.

علاوة على ذلك، يمكن لأساليب التعلم العميق التعامل مع البيانات عالية الأبعاد بكفاءة، مما يجعلها مناسبة تمامًا لتحليل كميات كبيرة من بيانات حركة مرور الشبكة. هذه القدرة قيمة بشكل خاص في بيئات المؤسسات الحديثة حيث يمكن لحركة مرور الشبكة توليد كميات هائلة من البيانات عبر العديد من الأجهزة والتطبيقات.

ومع ذلك، من المهم ملاحظة أنه في حين يوفر التعلم العميق إمكانات كبيرة لاكتشاف الشذوذ، فإنه يأتي أيضًا مع التحديات. وتشمل هذه الحاجة إلى كميات كبيرة من بيانات التدريب المصنفة، وخطر الإفراط في التكيف مع أنماط محددة، والموارد الحسابية المطلوبة لتدريب ونشر هذه النماذج. بالإضافة إلى ذلك، يظل ضمان قابلية تفسير قرارات نماذج التعلم العميق تحديًا مستمرًا في مجال اكتشاف الشذوذ.

المراجع

[1]https://www.ibm.com/cloud/learn/machine-learning

[2]Mehra, Sidharth & Hasanuzzaman, Mohammed, "Detection of Offensive Language in Social Media Posts", 2020.

[3] Aurelien Geron, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow", 2019.

[4] Laurikkala, J., Juhola, M., and Kentala, E, 'Informal Identification of Outliers in Medical Data". In: Fifth International Workshop on Intelligent Data Analysis in Medicine, 2000.

[5] Rousseeuw, P. and Leroy, A, "Robust Regression and Outlier Detection", 3 edition,1996.

[6] Tang D. H., Cao Z, "Machine Learning-based Intrusion Detection Algorithm" Journal of Computational Information Systems;5(6) p. 1825-1831, 2009.

[7] Kou Y., Lu C. T., Sirwongwattana S., Huang Y. P., "Survey of fraud detection techniques", In Proceedings of the IEEE International conference Networking, sensing and control; 2; p. 749-754, 2004.

[8] Manikopoulos, C., Papavassiliou, S., "Network intrusion and fault detection: a statistical anomaly approach". IEEE Commun. Mag. 40(10), 76–82 (2002)

[9] S.E. Smaha, Haystack: "An intrusion detection system", in: Proceedings of the IEEE Fourth Aerospace Computer Security Applications Conference, Orlando, FL, pp. 37–44. 1988.

[10] S. Staniford, J.A. Hoagland, J.M. McAlerney, "Practica automated detection of stealthy portscans", Journal of Computer Security 10, pp. 105–136, 2002.

[11] Panda, Mrutyunjaya & Abraham, Ajith & Das, Swagatam & Patra, Manas, "Network intrusion detection system: A machine learning approach". Intelligent Decision Technologies (IDT), 2011.

[12] F. Yihunie, E. Abdelfattah and A. Regmi, "Applying Machine Learning to Anomaly-Based Intrusion Detection Systems," IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1-5, 2019.

[13] Rai, Kajal & Devi, Mandalika & Professor, Devi & Guleria, Ajay, "Decision Tree Based Algorithm for Intrusion Detection", International Journal of Advanced Networking and Applications. 07. 2828-2834, 2016.

[14] Kevric, Jasmin , Jukic, Samed ,Subasi, AbdulhamitPY, "An effective combining classifier approach using tree algorithms for network intrusion detectionJO", Neural Computing and ApplicationsSP - 1051, 2017.

[15] P. Maniriho and T. Ahmad, "Analyzing the performance of machine learning algorithms in anomaly network intrusion detection systems", In: Proc. of the 4th International Conference on Science and Technology, 2018.

[16] Khraisat A., Gondal I., Vamplew P, "An Anomaly Intrusion Detection System Using C5 Decision Tree Classifier" Trends and Applications in Knowledge Discovery and Data Mining. PAKDD Springer ,2018.

[17] M.S Irfan Ahmed, Riyad A.M, R.L Raheemaa Khan, Mohamed Jamshad K, Shamsudeen E, " Information based feature selection for intrusion detection systems ", International Journal of Scientific & Engineering Research Volume 8, Issue 7, July-2017.

[18] BENADDI, H., IBRAHIMI, K., & BENSLIMANE, A, "Improving the Intrusion Detection System for NSL-KDD Dataset based on PCA-Fuzzy Clustering-KNN", 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), 2018.

[19] Vikram, A., & Mohana. "Anomaly detection in Network Traffic Using Unsupervised Machine learning Approach". 5th International Conference on Communication and Electronics Systems (ICCES), 2020.

[20] D. Perez, M. A. Astor, D. P. Abreu and E. Scalise, "Intrusion detection in computer networks using hybrid machine learning techniques,", XLIII Latin American Computer Conference (CLEI), pp. 1-10, 2017.

[21] Mohammed Tabash, Mohamed Abd Allah, and Bella Tawfik , "Intrusion Detection Model Using Naive Bayes and Deep Learning Technique", The International Arab Journal of Information Technology, Vol. 17, No. 2, March 2020.

[22] Shisrut Rawat1, Aishwarya Srinivasan, and Vinayakumar, "Intrusion detection systems using classical machine learning techniques versus integrated unsupervised feature learning and deep neural network", 1 Self-employed Data Scientist, New Delhi, India.

[23] Kwon, Donghwoon, et al. "An Empirical Study on Network Anomaly Detection Using Convolutional Neural Networks." ICDCS. 2018.

[24] https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223.

[25] L. Rokach, O. Maimon, Decision trees, in: O. Maimon, L. Rokach (Eds.), "Data Mining and Knowledge Discovery Handbook" , Springer, Boston, MA, pp. 165–192, 2005.

[26] Song, Y. Y., & Lu, Y, "Decision tree methods: applications for classification and prediction". Shanghai archives of psychiatry, 27(2), 130–135. 2015.

[27] Introduction to Algorithms for Data Mining and Machine Learning Xin-She Yang Middlesex University School of Science and Technology London, United Kingdom.

[28] John Wiley & Sons, "Machine Learning For Dummies" Published by: New Jersey Media and software compilation 2016

[29] https://jupyter.org/

[30] https://mode.com/python-tutorial/libraries/pandas

[31]https://towardsdatascience.com/a-quick-introduction-to-the-numpy-library-6f61b7dee4db

[32] https://matplotlib.org/

[33] Seaborn: Python - Towards Data Science

[34] https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool

[35] https://www.unb.ca/cic/datasets/nsl.html.

[36] A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015.

[37] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[38] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9