تحديد التوليفة الأمثل من ضبط البارامترات الفائقة واختيار الميزات لتحسين أداء أنظمة كشف الشذوذ

المهندس علي ياسين[1] – إشراف: الدكتور كمال السلوم[2] – الدكتور وسيم رمضان[3]

الملخص

لازال اكتشاف الشذوذ محط اهتمام الكثير من الباحثين لما له من دور في حل مشكلات العالم الحقيقية مثل الكشف عن الاحتيال واكتشاف الحالات غير الطبيعية...الخ.

يتأثر أداء خوارزميات اكتشاف الشذوذ بشكل كبير بكيفية ضبط البارامترات الفائقة من جهة، واختيار الميزات الأكثر أهمية ضمن مجموعات البيانات من جهة أخرى. على الرغم من ذلك فإن الأدبيات السابقة المتعلقة باكتشاف الشذوذ لا تتطرق في معظمها لضبط البارامترات الفائقة واختيار الميزات عند المقارنة التجريبية للخوارزميات المختلفة. حيث تقارن معظم الأوراق الأداء باستخدام القيم الافتراضية للبارامترات الفائقة وجميع الميزات ضمن مجموعات البيانات.

يهدف البحث إلى تحليل أداء مصنفات تعلم الآلة المستخدمة في كشف الشذوذ عند ضبط البارامترات الفائقة الخاصة بها واختيار الميزات الأكثر أهمية، بالإضافة إلى تحديد الاستراتيجية الأفضل لضبط البارامترات الفائقة واختيار الميزات، حيث يؤدي اختيار مجموعة مختلفة من الميزات إلى اختلاف مجموعة البارامترات الفائقة والعكس صحيح.

أكدت نتائج الدراسة أن اختيار الميزات الأكثر أهمية ضمن مجموعة البيانات متبوعاً بضبط البارامترات الفائقة يساهم في تحسين أداء خوارزميات الكشف عن الشذوذ بشكل كبير، حيث تم اختزال الزمن من 50% إلى 53% وزيادة دقة الكشف من 5% إلى 6% وتحسين كل مقاييس الأداء المستخدمة.

الكلمات المفتاحية: اكتشاف الشذوذ، ضبط البارامترات الفائقة، اختيار الميزات، تعلم الآلة

[1] طالب دكتوراه - قسم هندسة البرمجيات – كلية الهندسة المعلوماتية – جامعة البعث – حمص - سوريا
[2] أستاذ دكتور - قسم هندسة البرمجيات – كلية الهندسة المعلوماتية – جامعة البعث – حمص - سوريا
[3] دكتور مشارك - قسم الاقتصاد الزراعي – كلية الزراعة – جامعة البعث – حمص - سوريا

Selecting the Optimal Combination of Hyperparameter Tuning and Feature Selection to Improve the Performance of Anomaly Detection Systems

Eng. Ali Yassin, Dr. Kamal Al-Salloum, Dr. Wassim Ramadan

Abstract

Anomaly detection is still an important research subject for many researchers. It has a crucial role in solving real-world problems such as fraud detection...etc.

The performance of anomaly detection algorithms is greatly affected by how the hyperparameters are tuned on one hand, and the selection of the most important features within datasets on the other. However, previous literature on anomaly detection mostly does not deal with hyperparameter tuning and feature selection when comparing different algorithms using experiments. Most papers compare performance using default values for hyperparameters and all features within datasets.

This research aims to analyze the performance of machine learning classifiers used in anomaly detection when tuning their hyperparameters and selecting the most important features, in addition to defining the best strategy for hyperparameters tuning and feature selection. Because selecting a different set of features causes a different set of hyperparameters and vice versa.

The results of the study confirm that selecting the most important features within the dataset followed by tuning the hyperparameters contributes significantly in improving the performance of anomaly detection algorithms, as the time was reduced by 50% to 53%, the accuracy of detection increased by 5% to 6%, and all performance measures used were improved.

Keywords: Anomaly Detection, Tuning Hyperparameters, Feature selection, Machine Learning

1- مقدمة

نتيجة ازدياد كم البيانات المتوافر في كل مكان بشكل كبير في الآونة الأخيرة (من المتوقع أن يصل إنتاج البيانات إلى 175 Zettabytes بحلول عام 2025)[1] ، أَصْبَحَ هناك حاجة ضرورية لتحليل مجموعات البيانات في العالم الحقيقي بكافة أشكالها، واستخراج أنماط البيانات الخفيّة والمعقدة، لاستخدامها في العديد من المجالات المختلفة. يتأثر تحليل البيانات في مختلف المجالات التطبيقية بالبيانات الشاذة التي يمكن أن تؤشر بشكل أو بأخر عن شيء خارج نطاق البيانات الطبيعية. وبالتالي فالتحليل السليم للبيانات للحصول على المعلومات الصحيحة والدقيقة والمعبرة عن الحالة الطبيعية يجب أن يميز بين البيانات الطبيعية وغير الطبيعية. وهذا ما يدعو للحاجة إلى الاهتمام بشكل أو بأخر بالكشف عن الشذوذ.

تَهَدُّف عملية الكشف عن الحالات الشاذة [1] إلى تحديد كل نقاط البيانات التي تسلك سلوكاً مختلفاً عن باقي نقاط المجموعة. يُمْكَنْ أن تنتج الحالات الشاذة عن خطأ في البيانات؛ ولكنها تدل أيضاً على عمليات أساسية جديدة لم تكن معروفة مسبقاً وغالباً ما تكون حرجة في مجموعة واسعة من التطبيقات. يوجد إسقاط كبير لمفهوم الشذوذ في تطبيقات العالم الحقيقي كتطبيقات اكتشاف الاحتيال المالي والاختراقات الشبكية والأمراض النادرة.

توجهت العديد من شركات القطاع المالي مؤخراً لتشغيل خدماتها عبر الانترنت كعملية الدفع الإلكتروني، وبالتالي تتزايد عمليات الاحتيال المالي من حيث الشكل والعدد حول العالم. يُعد كل من الوصول غير المصرح به والهجمات غير المنتظمة أمثلة على التهديدات التي يجب اكتشافها باستخدام أنظمة الكشف عن الاحتيال المالي. مما يؤدي إلى خسائر مالية هائلة تجعل الاحتيال المالي مشكلة كبيرة. دفع كل ذلك كبرى الشركات ومواقع التجارة الإلكترونية إلى استخدام مفهوم تعلم الآلة لبناء أنظمة قادرة على كشف تلك الحالات الشاذة (الاحتيالية). كما أصبح هذا الأمر مهماً في الجمهورية العربية السورية حيث هناك توجه كبير لاستخدام خدمات الدفع الإلكتروني بدءاً من عام 2020 مع انطلاق منظومة الدفع الإلكتروني.

يُمّكن تعريف أنظمة كشف الشذوذ على أنها أجهزة أو برامج تقوم بمراقبة البيانات المتعلقة بمجال معين لتحديد الحالات الشاذة والمختلفة عن السلوك الطبيعي باستخدام طرائق معينة مثل تعلم الآلة.

تَسْتَطِيع هذه الأنظمة والتي تعتمد على خوارزميات تعلم الآلة التعامل مع البيانات بأشكالها المختلفة كالبيانات الرقمية والنصية بنوعيها الهيكلية وغير الهيكلية بالإضافة لقدرتها على استخراج أنماط البيانات المعقدة. بالمقابل يتأثر أدائها بشكل كبير بكيفية ضبط البارامترات الفائقة (Hyperparameters Tuning) من جهة، واختيار الميزات (Features Selection) الأكثر أهمية ضمن مجموعات البيانات من جهة أخرى.

حيث تُعبر البارامترات الفائقة (Hyperparameters) عن مجموعة من البارامترات الرياضية والتي يتم ضبطها بشكل مختلف عن البارامترات العادية، حيث لن يقوم النموذج بتحديثها وفقاً لاستراتيجية التحسين ولازالت الخوارزميات حالياً تحتاج إلى تدخل يدوي لضبطها في بداية التنفيذ. يوجد عدة عوامل تحفز على ضبط هذه البارامترات الفائقة بشكل أوتوماتيكي 3]] ومنها:

- الأبعاد العالية للبارامترات الفائقة (High Dimensions of Hyperparameters): أصبحت نماذج تعلم الآلة بشكل عام أكثر تعقيداُ مع تزايد عدد البارامترات الفائقة. مما يضطر المختصون في هذا المجال إلى تقييم النماذج مع مجموعات مختلفة من قيم البارامترات الفائقة الخاصة بنماذجهم للحصول على أفضل النتائج.
- زيادة وقت التدريب: بتزايد حجم مجموعات البيانات يصبح تدريب النماذج أكثر تكلفة من الناحية الحسابية بشكل كبير، وغالباً ما يستغرق من ساعات إلى أيام على أجهزة مختصة عالية الأداء. يعد ذلك مرهقاُ بشكل خاص في سياق ضبط البارامترات الفائقة، حيث يجب تدريب نموذج جديد في كل مرة لتقييم المجموعات المختلفة من البارامترات الفائقة المرشحة.

يوجد العديد من الطرائق والأساليب المختلفة 4]] لأتمتة ضبط البارامترات الفائقة (Hyperparameters Tuning) أهمها البحث الشبكي (Grid Search) وطريقة البحث العشوائي (Random Search).

من جهة أخرى، تعرف هندسة الميزات (Feature engineering) بأنها عملية إنشاء مجموعة ميزات باستخدام خصائص البيانات التي تعزز أداء خوارزميات تعلم الآلة. يمكن أن تكون هذه الميزات ذات أبعاد عالية (High Dimensions) ويصعب تدريبها. يعد تقليل الأبعاد (Dimensionality reduction) أحد أكثر الطرق شيوعاً لتحويل (Mapping) الميزات من فضاء ذو أبعاد عالية إلى فضاء بعدد أقل من الأبعاد والتي لها معنى [5] يوجد مجموعة من تقنيات تقليل الأبعاد وأهمها استخراج الميزات (Feature Extraction) واختيار الميزات (Feature Selection) 6]]. تنشئ تقنية استخراج الميزات مجموعة ميزات جديدة باستخدام مجموعة من الميزات الأصلية وإسقاطها إلى فضاء بأبعاد أقل. بينما يهدف اختيار الميزات إلى تحديد مجموعة فرعية من الميزات وثيقة الصلة باستخدام مقياس معياري. يوجد مجموعة من الطرائق لاختيار الميزات أهمها التصفية (Filter) والتغليف (Wrapper) وتقنيات التضمين (Embedded Methods).

يَعد كُلّ من تحديد الميزات وضبط البارامترات الفائقة مهمتين أساسيتين في تعلم الآلة. يؤثر تنفيذ أحدهما قبل الأخر على أداء النماذج، حيث أن لاختيار ميزات النموذج تأثير كبير على ضبط البارامترات الفائقة والعكس صحيح. سوف نعتمد في هذا البحث على استخدام تقنيات التضمين لتحديد الميزات الأكثر أهمية وطريقة البحث العشوائي لضبط البارامترات الفائقة، وتحديد الاستراتيجية الأفضل لترتيب تحقيقهما ضمن أنظمة كشف الشذوذ المعتمدة على خوارزميات تعلم الآلة.

2- أهمية وأهداف البحث

تتأثر أنظمة كشف الشذوذ المعتمدة على خوارزميات تعلم الآلة بتحديد الميزات الأكثر أهمية بالإضافة لاختيار أفضل البارامترات الفائقة. لكن بالمقابل فإن لترتيب تحقيق هذه العمليات دور هام في تحسين أداء هذه الأنظمة من حيث وقت ودقة الكشف.

يهدف البحث بشكل أساسي إلى إيجاد التوليفة الأمثل بين اختيار الميزات الأكثر أهمية وضبط البارامترات الفائقة لتحقيق أفضل أداء في أنظمة الكشف عن الشذوذ، وذلك من خلال تحقيق ما يلي:

1- إيجاد أفضل قيم البارامترات الفائقة لنماذج تعلم الآلة المقترحة في سياق التعلم الخاضع للإشراف (Supervised Learning) والتعلم غير الخاضع (Unsupervised Learning).
2- إيجاد الميزات الأكثر أهمية ضمن مجموعة البيانات.
3- تحديد تأثير ترتيب الخطوتين السابقتين على أداء نظام كشف الشذوذ بحيث يتم اختيار الترتيب الأمثل للوصول إلى أفضل دقة كشف.

3- مواد وطرائق البحث

3-1 البيانات البحثية

تَمَّ الاعتماد على مجموعة بيانات (Dataset) لمعاملات بطاقات ائتمان أوربية تمت على مدار يومين في أيلول عام 2013، وتُعِدّ من أكثر مجموعات البيانات الحقيقة المتاحة في الدراسات حتى الآن. تحتوي مجموعة البيانات على 284807 معاملة (سجل)، منها 492 معاملة احتيالية (حوالي 0.17٪)، مما يجعل مجموعة البيانات هذه غير متوازنة (Unbalanced Dataset) إلى حد كبير.

تَمَثَّل كل معاملة ب 30 ميزة (Feature) جميعها رقمية. تَمَّ تحويل القيم الأصلية لـ 28 من هذه الميزات باستخدام تحليل المكونات الرئيسية Principal Component) (Analysis (PCA) وسُميت بأسماء المتغيرات من V1 وحتى V28، ولم يتم الكشف عن معلومات حول هذه الميزات لأسباب تتعلق بالسرية. كما لا يتوفر أي معلومات عن معرف حامل البطاقة ID حيث يتم اعتبار كل معاملة مستقلة عن المعاملات الأخرى. تُعبر ميزة الوقت (Time) عن الثواني المنقضية بين كل معاملة والمعاملة الأولى، وتحوي ميزة الكمية (Amount) على مبلغ المعاملة. يُعِدّ المتغير Class متغير الاستجابة ويأخذ القيمة 1 في حالة المعاملات الاحتيالية و0 على خلاف ذلك7]] .

3-2 الحزم والمكتبات المستخدمة Used Packages and Libraries

تُمّ بناء نظام الكشف عن الشذوذ في مجموعة البيانات السابقة باستخدام لغة بايثون **Python** وبالاعتماد على مجموعة من المكتبات البرمجية [8] وأهمها:

- **Numpy:** هي اختصار لعبارة (Numerical Python Library) وتستخدم هذه الحزمة من أجل المصفوفات متعدد الأبعاد والعمليات الجبرية الخطية.
- **Pandas:** توفر هذه الحزمة أداة لتحليل ومعالجة البيانات. تم استخدامه لقراءة مجموعة البيانات وتحميلها.
- **Scikitlearn:** تستخدم هذه الحزمة من أجل الأساليب الإحصائية وتعلم الآلة.
- **Keras:** توفر هذه الحزمة واجهات برمجية متناسقة وبسيطة من أجل التخاطب مع المستخدم النهائي وليس الآلة ، وتحتوي على مجموعة من النماذج (Models) مثل الشبكات العصبونية وأشجار القرار وتوابع التنشيط ، كما تتميز بقابليتها للتوسع أي القدرة على إضافة نماذج جديدة. تكون المهمة الأساسية للمكتبة جعل التطبيق أكثر استجابة مع إمكانية إعطاء المستخدم المزيد من القدرة على التحكم به.
- **TensorFlow** : تُطبّق هذه الحزمة في العديد من المجالات كحساب المشتقات والمصفوفات الضخمة بالإضافة إلى استخدامها في توزيع العمليات الحاسوبية على وحدات المعالجة المركزية CPU وكذلك على شبكة موزعة مكونة من مجموعة أجهزة بعيدة تتضمن هذه المكتبة. يُستخدم TensorFlow بشكل أساسي في تعلم الآلة في الوقت الحالي.

3-3 مقاييس الأداء Performance Metrics

يعتبر مقياس الدقة Accuracy المقياس الأكثر استخداماً والأكثر منطقية في تقييم أداء خوارزميات التصنيف وذلك عندما تكون مجموعات البيانات متوازنة. فهو يمثل النسبة بين عدد العينات المصنفة بشكل صحيح وعدد العينات الكليّة. ولكن بالمقابل لا يمكن اعتباره مقياساً جيداً في البيانات غير المتوازنة (كما في حالة اكتشاف الشذوذ) لأنه يسبب ملائمة زائدة (Overfitting) لجهة صف الأغلبية.

يوجد مجموعة واسعة من المقاييس التي يمكن استخدامها لتقييم أداء أنظمة الكشف عن الشذوذ ونذكر منها هنا فقط ما تم استخدامه لتقييم الأداء كونها الأكثر استخداماً في تقييم خوارزميات اكتشاف الشذوذ:

1. مصفوفة الارتباك Confusion Matrix

يَتنبأ النموذج بصف عينات البيانات وينسب إلى كل عينة تسميتها (Label) المتوقعة (إيجابية أو سلبية)، لتقع كل عينة في نهاية المطاف ضمن أحد الحالات التالية:

1) الإيجابيات الحقيقية True Positives (TP) تدل على العينات الإيجابية المتوقعة بشكل صحيح.
2) السلبيات الحقيقية True Negatives (TN) تدل على العينات السلبية المتوقعة بشكل صحيح.
3) الإيجابيات الخاطئة False Positives (FP) والسلبيات الخاطئة False Negatives (FN) عندما يتعارض الصف الحقيقي مع الصف المتوقع.

يُمكن تلخيص الحالات السابقة ضمن مصفوفة 2*2 تسمى مصفوفة الارتباك $M = (TP\ FN\ FP\ TN)$.

2. مقياس F1 Score

يُعدّ F1 Score في التحليل الإحصائي مقياساً لدقة الاختبار، فهو المتوسط التوافقي للدقة (Precision) والاستذكار (Recall). تحسب قيمة F1 بالاعتماد على مصفوفة الارتباك بالشكل التالي:

$$F1\ score = \frac{2 * TP}{2 * TP + FP + FN} = 2 * \frac{precision * recall}{precision + recall}$$

تمثل القيمة 1 أفضل نتيجة، بينما 0 أسواءها.

3. Matthews Correlation Coefficient (MCC)

يُعدّ مقياس Matthews Correlation Coefficient (MCC) والذي يعتبر حالة خاصة من معامل phi حلاً للتغلب على مشكلة عدم توازن البيانات. فهو يعبر عن الارتباط بين القيم المتوقعة والحقيقية. تعتمد قيمة MCC على جميع عناصر مصفوفة الارتباك بالشكل التالي:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

تمثل القيمة 1 أفضل نتيجة، بينما 1− أسواءها، أما القيمة 0 فهي للدلالة على عشوائية النموذج. يعتبر MCC المقياس الوحيد [9][9] الذي يستخدم لمعرفة فيما إذا كان المصنف الثنائي قادر على توقع معظم الحالات الإيجابية والسلبية بشكل صحيح.

4. منحنيات الدقة والاستذكار Precision–Recall Curves (PR)

يوصى عادةً بمنحنيات الدقة والاستذكار [9] عند التعامل مع مجموعات بيانات غير متوازنة، حيث توفر منحنيات Receiver Operating Characteristic (ROC) رؤية مضللة ومنحازة لصف الأغلبية.

يوفر منحني PR رؤية بصرية لقيم كل من الدقة والاستذكار وذلك من أجل قيم عتبات مختلفة بدلاً من قيمة واحدة. تُقيس منحنيات PR درجة الفصل بين الصفوف بالاعتماد على مساحة السطح تحت المنحني Area Under the Curve (AUC) ويُعتبر النموذج أكثر دقة مع ازدياد قيمة AUC حيث يزداد الفصل والتمايز بين الصفوف، تتراوح قيم AUC ضمن المجال [0,1].

5. معدل الفشل Failure Rate

قد يكون في بعض الحالات للمقاييس المستخدمة نفس القيم، نحتاج عندئذٍ إلى تفاضل بين هذه الحالات. اقترح البحث لتحقق ذلك ما يسمى معدل الفشل (Failure Rate). يعطي معدل الفشل نسبة كل من السلبيات الخاطئة (تصنيف الحالات الشاذة على أنها طبيعية) والإيجابيات الخاطئة (تصنيف الحالات الطبيعية على أنها شاذة). يجب أن تكون قيمة هذه النسبة أقل ما يمكن لتحقيق أفضل أداء لنماذج كشف الشذوذ.

$$failure\ rate = \frac{FP + FN}{TP + FP + TN + FN}$$

اعتمد البحث الحالي على كل من منحني PR وMCC لتحليل أداء مصنفات تعلم الآلة لما توفره من قدرة على إعطاء تصور واضح لأداء هذه المصنفات مع مسائل كشف الشذوذ.

4- الدراسة المرجعية

اشتملت أدبيات الدراسات على العديد من الوسائل والطرائق والخوارزميات لمعالجة الحالات الشاذة ومحاولة تصنيفها واكتشافها ومن أهمها خوارزميات تعلم الآلة. بالمقابل

تفتقر معظم هذه الأدبيات إلى استخدام آليات ضبط البارامترات الفائقة واختيار الميزات الأكثر أهمية من جهة، ومن جهة أخرى إلى تكامل هذه الآليات معاً.

تُشير الدراسات السابقة10]] لوجود حوالي 49 أسلوب تَعلم خاضع للإشراف في اكتشاف الحالات الشاذة 10]]، تتفوق فيها كل من والشبكات العصبونية (Neural Network)، والانحدار اللوجستي (Logistic Regression)، الغابات العشوائية (Random Forests)، وآلة شعاع الدعم (SVM) Support Vector Machines.

تمَّ تطبيق مجموعة من خوارزميات تعلم الآلة مثل Naïve Bayes وأقرب جار (K–Nearest Neighbor) وآلة شعاع الدعم والغابات العشوائية بهدف كشف الاحتيال المالي ضمن مجموعة بيانات غير متوازنة تحتوي على نسبة 3% فقط من المعاملات الاحتيالية11]] . أظهرت النتائج النهائية أن نموذج الغابات العشوائية لديه أفضل معدل للإيجابيات الحقيقية (True Positive Rate) مقارنةً مع النماذج الأخرى المستخدمة. لكن بالمقابل لم تذكر الدراسة آلية ضبط البارامترات الفائقة الخاصة بالنماذج المقترحة.

استخدمت خوارزمية الغابات العشوائية (Random Forest) بغرض كشف الاحتيال المالي ضمن أحد الشركات الصينية 12]]. اعتمدت الدراسة على بناء نموذجين من الغابات العشوائية يختلفان عن بعضهما بآلية اختيار الميزات (Features) ضمن العقد (Nodes). تعتمد الآلية الأولى على حساب المسافة ما بين نقاط البيانات وصفي البيانات (الطبيعية والشاذة)، بينما تعتمد الآلية الثانية على حساب قيمة شائبة جيني (Gini Impurity) لكل ميزة واختيار تلك التي تحقق أقل قيمة. أظهرت النتائج تفوق الآلية الثانية من حيث الاستذكار (Recall) والدقة (Precision)، حيث بلغت قيمتهما 95% و89% على التوالي. بالمقابل نوهت الدراسة أنها تتعامل مع مجموعة بيانات غير متوازنة ولذلك يمكن أن تكون دقة هذه النماذج مضللة بعض الشيء.

تمَ مقارنة عدد من تقنيات تعلم الآلة غير خاضعة للإشراف والأكثر شيوعاً في مجال اكتشاف الشذوذ، وهي غابة العزل (Isolation Forest) وخوارزمية K–Means وخوارزمية المعامل الخارجي المحلي (Local Outlier Factor) لتحديد أفضل تنقية في اكتشاف الاحتيال المالي ضمن مجموعة البيانات الأوربية13]] . أظهرت نتائج الدراسة تفوق خوارزمية غابة العزل بالاعتماد على منحني ROC حيث بلغت مساحة السطح تحت المنحني (AUC) 91%. أحد أهم قيود الدراسة هو الاعتماد على منحني ROC فقط لتقييم نماذج الكشف وهذا أمر مضلل بسبب طبيعة البيانات غير المتوازنة، كما اعتمدت أيضاً على القيم الافتراضية للبارامترات الفائقة.

تمَّ مقارنة ثماني خوارزميات تعلم آلة (Machine Learning) في مسائل كشف الاحتيال المالي ضمن بطاقات الائتمان14]] . أظهرت النتائج **Error! Reference source not found.** تفوق كل من خوارزميات آلة شعاع الدعم (SVM) والشبكات العصبونية والغابات العشوائية وفقاً لثلاث مقاييس أداء وهي الاستذكار (Recall) ومساحة السطح تحت منحني الدقة والاستذكار (AUCPR) والدقة (Accuracy). لكن من جهة أخرى، نوه الباحثون لضرورة إجراء ضبط البارامترات الفائقة لتحقيق نتائج أفضل لهذه الخوارزميات.

تم استخدام أداة Features Selector لتحديد الميزات الأكثر أهمية ضمن مجموعة البيانات الأوربية الخاصة بالاحتيال المالي حيث تم اختيار 27 ميزة لهذه التجربة. ليتم في المرحلة التالية تطبيق مجموعة من خوارزميات تعلم الآلة وهي الانحدار اللوجستي (Logistic Regression) والغابات العشوائية (Random Forests) ومصنف بايز (Naive Bayes) لتصنيف الحالات الاحتيالية ضمن مجموعة البيانات المستخدمة 15]]15]]. حيث أظهرت نتائج هذه الدراسة **Error! Reference source not found.**

تفوق الغابات العشوائية بنسبة استذكار (Recall) حوالي 81%. لكن بالمقابل لم تَطَرَّقَ الدراسة لضبط البارامترات الفائقة الخاصة بالنماذج المقترحة.

استخدمت خوارزميات الانحدار اللوجستي والغابات العشوائية وأشجار القرار (Decision tree) لاكتشاف العمليات الاحتيالية ضمن مجموعة البيانات الأوربية [16]. واختبرت الدراسة مدى فعالية هذه الخوارزميات باستخدام جميع ميزات مجموعة البيانات وعند اختيار مجموعات جزئية من الميزات مكونة من (5−10 ميزات) حيث أظهرت النتائج تفوق الغابات العشوائية بدقة (Accuracy) حوالي 90% عند استخدام جميع الميزات. لم تذكر الدراسة الطريقة التي تم اختيار الميزات بها من جهة، ومن جهة أخرى لا يمكن تبني هذه النتائج بشكل كبير بسبب اعتمادها على مقياس الدقة (Accuracy).

تمَّ تطبيق البحث الشبكي (Grid Search) لضبط البارامترات الفائقة الخاصة بخوارزميتي آلة شعاع الدعم (SVM) والغابات العشوائية من أجل بناء نماذج لاكتشاف الاحتيال المالي ضمن عدة مجموعات غير متوازنة [17][17] أظهرت نتائج الدراسة **!Error Reference source not found.** تفوق شعاع آلة الدعم على الغابات العشوائية من أجل جميع مجموعات البيانات المستخدمة. حيث تمَّ تقييم النتائج باستخدام مقياس(MCC) Matthews Correlation Coefficient، وبلغت قيمته في خوارزمية SVM حوالي 81% من أجل مجموعة البيانات الأوربية. إن أحد قيود هذه الدراسة أنها تتعامل فقط مع اكتشاف الاحتيال في سياق التعلم الخاضع للإشراف، بالإضافة لعدم تحديد الميزات الأكثر أهمية.

تم ملاحظة التفوق الكبير من خلال الدراسات السابقة للغابات العشوائية وآلة شعاع الدعم وغابات العزل في مجال اكتشاف الشذوذ على باقي الخوارزميات. لكن وبنفس الوقت تم ملاحظة القصر الحاصل فيما يتعلق بضبط البارامترات الفائقة لهذه الخوارزميات، مما يستدعي المزيد من البحث للحصول على أفضل البارامترات الفائقة. بالمقابل تم اعتماد أكثر

من طريقة لاختيار الميزات الأكثر أهمية ضمن الأدبيات، لكنها لازالت تعاني من الدقة من جهة ومن عدم تكاملها مع ضبط البارامترات الفائقة من جهة أخرى. استوجب كل ذلك المزيد من البحث لتحديد الآلية الأفضل لضبط البارامترات الفائقة، واختيار الميزات الأكثر أهمية وتكاملهما معاً لتحقيق أفضل دقة كشف عن الحالات الشاذة باستخدام هذه الخوارزميات.

5- الخوارزميات المقترحة للدراسة

بناء على ما تم ملاحظته من الدراسات السابقة فقد تم التركيز في هذا البحث على الخوارزميات التي أعطت أفضل أداء في اكتشاف الشذوذ. وعليه تم تفصيل كل منها.

5-1 خوارزمية الغابات العشوائية Random Forests Algorithm

تندرج خوارزمية الغابات العشوائية18]] تحت خوارزميات التعلّم الخاضع للإشراف. تتشكل الغابة (Forest) من مجموعة من أشجار القرار ويتم تدريبها باستخدام مفهوم التعبئة (Bagging). تتمثل إحدى الميزات الكبيرة للغابات العشوائية في أنه يمكن استخدامها لكل من مسائل التصنيف والانحدار، والتي تشكل غالبية أنظمة تعلّم الآلة الحالية.

باختصار: تقوم الغابات العشوائية ببناء مجموعة من أشجار القرار ودمجها معاً باستخدام تقنيات التجميع (Ensemble Technique) للحصول على تنبؤ أكثر دقة.

5-1-1 مصنفات الغابات العشوائية Random Forests Classifiers

يستخدم التصنيف في الغابات العشوائية تقنية التعبئة المعروفة أيضاً باسم "Bootstrap Aggregation" والتي تقوم باختيار مجموعة جزئية عشوائية مع الاستبدال من بيانات التدريب (مما يعني أنه يمكن اختيار نفس العينة عدة مرات، سحب مع إعادة). تُعرف الخطوة السابقة باسم Bootstrap. يتم في كل مرة اختيار عينة جديدة Bootstrap Sample ليتم تدريب النموذج عليها لنحصل في نهاية الأمر على مجموعة من النماذج المدربة، ويعتمد الناتج النهائي على تصويت الأغلبية (Majority Voting) بعد دمج نتائج جميع النماذج (بمعنى أخر يصبح الناتج الذي تختاره غالبية أشجار القرار هو الناتج النهائي لمصنف الغابة العشوائية). تُعرف الخطوة الأخيرة باسم التجميع (Aggregation).

5-1-2 البارامترات الفائقة Hyperparameters

يوجد مجموعة من البارامترات الفائقة [18] ضمن نموذج الغابات العشوائية وأهمها:

1- عدد نماذج التنبؤ Count of Estimators

تمثل قيمة هذا المتحول عدد أشجار القرار التي تبنيها الخوارزمية.

2- الحد الأعظمي لعدد الميزات Maximum Number of Features

تمثل الحد الأعظمي لعدد الميزات التي تأخذها الخوارزمية بعين الاعتبار عند تقسيم العقدة.

3- الحد الأدنى لعدد العينات Minimum Number of Samples

تمثل الحد الأدنى لعدد العينات المطلوبة لتقسيم عقدة داخلية.

4- عمق الشجرة Depth of The Tree

يحدد أقصى عمق يمكن أن تصل إليه الشجرة. في حال لم يتم تحديد قيمة ضمن هذا البارامتر فسوف يتم التوسع في الشجرة حتى تصبح جميع الأوراق نقية، أو حتى تحتوي على عينات أقل من الحد الأدنى لعدد العينات.

5-1-3 اكتشاف الشذوذ باستخدام الغابات العشوائية

تعتمد الفكرة الرئيسية في استخدام الغابات العشوائية في اكتشاف الشذوذ على دمج نتائج مجموعة من أشجار القرار المتنوعة، نظراً لأن كل شجرة تأخذ بعين الاعتبار مجموعة مختلفة من الميزات (Features). يؤدي تنوع أشجار القرار إلى اختزال فضاء أبعاد الميزات من خلال تقسيمها بين هذه الأشجار، ويحقق ذلك القدرة على التعامل مع بيانات عالية الأبعاد ولعل ذلك من أهم ما تتطلبه أنظمة كشف الشذوذ ضمن البيانات.

5-2 خوارزمية غابات العزل Isolation Forests

تشبه غابات العزل (Isolation Forests) الغابات العشوائية لحد ما حيث يتم بناء كل منهما بالاعتماد على تقنيات التجميع لأشجار القرار. تندرج غابة العزل تحت خوارزميات التعلّم غير الخاضع للإشراف

يشير مصطلح العزل (Isolation) إلى فصل نقطة بيانات (حالة/Instance) عن بقية النقاط [19][23]، ونظراً لأن الحالات الشاذة قليلة ومختلفة بالتالي فهي أكثر عرضة للعزل. يحدث في شجرة القرار انقسام لنقاط البيانات بشكل متكرر حتى يتم عزل جميع النقاط. ينتج عن هذه الانقسامات العشوائية مسارات أقصر ملحوظة للحالات الشاذة.

5-2-1 آلية عمل غابة العزل

كما ذكرنا سابقاُ أن غابات العزل ليست إلا مجموعة من أشجار القرار الثنائية، وتسمى كل شجرة في غابة العزل بشجرة العزل. تبدأ الخوارزمية بالتدريب عن طريق توليد مجموعة من أشجار العزل باستخدام مفهوم التعبئة (Bagging).

تتلخص مراحل عمل الخوارزمية [20] بالخطوات التالية:

1) يتم تحديد عينة جزئية عشوائية من بيانات التدريب باستخدام مفهوم التعبئة، وتسمى هذه العينة Bootstrap Sample.
2) يبدأ تقسيم شجرة العزل باختيار ميزة عشوائية من مجموعة الميزات، ثم يتم إجراء الانقسامات المتتالية باختيار قيمة عتبة (Threshold) عشوائية تقع بين القيم القصوى والدنيا للميزة التي تم اختيارها أولاً.
3) نختبر قيمة نقطة البيانات فإذا كانت أصغر من قيمة العتبة فإنها تصبح أبن يساري للعقدة الأب وإلا تصبح ابن يميني.
4) يتم تكرار الخطوات 2 و3 بشكل متكرر حتى يتم عزل كل نقطة بيانات تماماً أو حتى يتم الوصول إلى أقصى عمق (إذا تم تحديده).
5) يتم تكرار الخطوات المذكورة أعلاه لإنشاء أشجار ثنائية عشوائية.

نحصل في نهاية عملية تدريب الخوارزمية على مجموعة من أشجار العزل (غابة عزل). تجتاز نقطة البيانات جميع الاشجار التي تم تدريبها مسبقاً، ويتم تعيين درجة الانحراف (Anomaly Score) لكل نقطة من نقاط البيانات بناءً على طول المسار

المطلوب للوصول إلى عزل تلك النقطة. تسحب القيمة النهائية لدرجة الانحراف (الشذوذ) لنقطة ما من خلال حساب متوسط أطوال المسارات التي تم الحصول عليها من جميع أشجار العزل.

5-2-2 البارامترات الفائقة Hyperparameters

يوجد مجموعة من البارامترات الفائقة ضمن نموذج غابة العزل 23]][21] وأهمها هي:

1. **الحد الأعظمي لعدد العينات Maximum Number of Samples**

تمثل قيمة هذا البارامتر الحد الأعظمي لعدد العينات التي يمكن سحبها لتدريب كل شجرة.

2. **التلوث Contamination**

تمثل النسبة المتوقعة من القيم الشاذة في مجموعة البيانات وهي حساسة للغاية.

3. **عدد نماذج التنبؤ Count of Estimators**

يشير إلى عدد الأشجار التي سيتم بناؤها في الغابة. القيمة الافتراضية 100.

5-2-3 اكتشاف الشذوذ باستخدام غابات العزل

يتم الكشف عن الشذوذ باستخدام غابات العزل على مرحلتين: الأولى هي مرحلة التدريب لبناء شجرة عزل باستخدام عينات فرعية من مجموعة بيانات التدريب. توفر المرحلة الثانية اختبار كل نقطة بيانات باستخدام أشجار العزل الناتجة من المرحلة الأولى للحصول على قيمة درجة الشذوذ لكل حالة.

5-3 خوارزمية آلة شعاع الدعم Support Vector Machine Algorithm

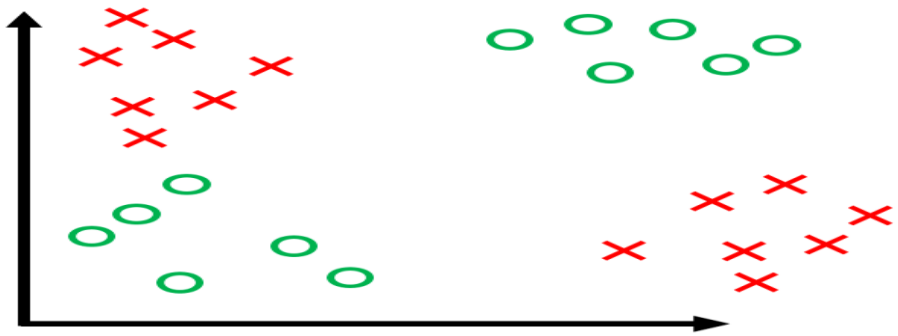
تعد خوارزمية آلة شعاع الدعم23]] (SVM) واحدة من أكثر خوارزميات تعلم الآلة شيوعاً22]] والمستخدمة على نطاق واسع. وذلك بسبب بساطة فكرة عملها وأناقتها رياضياً من جهة، وسهولة استخدامها من جهة أخرى.

تقوم فكرة الخوارزمية على الفصل بين الصفوف داخل فضاء البيانات بمستوٍ يسمى المستوى الفائق (Hyperplane). لكي تتمكن SVM من إيجاد ذلك المستوى، لا بد من أن تكون البيانات قابلة للفصل الخطي (Linearly Separable).

5-3-1 آلية عمل آلة شعاع الدعم

يتم تدريب المصنف بإيجاد قيم w و b التي تفصل مجموعة البيانات بأكبر هامش (Margin) ممكن. تسعى الخوارزمية إلى إيجاد أفضل مستوٍ فائق قادر على فصل أكبر عدد ممكن من نقاط البيانات، حيث تقوم بمعاقبة المصنف من أجل جميع نقاط البيانات التي تقع في الاتجاه الخاطئ من المستوي.

ماذا لو كانت البيانات غير قابلة للفصل خطياً؟ حيث إننا نرى في **Error! Reference source not found.** مثالاً على هذه الحالة. تعجز خوارزمية SVM التقليدية عن إيجاد مستوٍ فائق يفصل هذه البيانات مما يؤدي بدوره إلى عدم القدرة على تصنيف البيانات بشكل صحيح.



الشكل 1 بيانات غير قابلة للفصل خطياً

يتم حل المشكلة السابقة [22].**Error! Reference source not found** باستخدام تعديل على الخوارزمية يسمى "نوى SVM" (Kernel SVM) وهي عبارة عن تعميم لخوارزمية آلة شعاع الدعم تسمح لها بتصنيف البيانات التي لا يمكن فصلها خطياً.

5-3-2البارامترات الفائقة Hyperparameters

1. التنظيم Regularization:

تحدد قيمة هذه البارامتر مدى السماح لخوارزمية SVM في تخطي تصنيف كل نقطة تدريب. يطلق على هذه البارامتر في لغة بايثون C.

متحول Gamma:

مدى تأثير نقاط بيانات التدريب في إيجاد المستوي الفائق الأمثل.

5-3-3 اكتشاف الشذوذ باستخدام خوارزمية SVM

جعل كل ما سبق من SVM خوارزمية فعالة في مسألة اكتشاف الشذوذ نظراً لإمكانية استخدام نواة مناسبة في حال كانت البيانات غير قابلة للفصل خطياً من جهة، وقدرتها على التعامل مع البيانات ذات الأبعاد العالية (وهي سمة عامة لأغلب مسائل كشف الشذوذ) من جهة أخرى. فهي تحاول إيجاد المستوي الفائق الأفضل بناءً على عامل تعظيم الهوامش الفاصلة بين صفي البيانات.

8- النتائج والمناقشة Results and Discussion

يستعرض في هذا القسم بدايةً وصفاً تجريبياً لتصميم التجارب التي أجريناها لبناء أنظمة كشف الشذوذ (اكتشاف الاحتيال المالي) باستخدام خوارزميات تعلم الآلة المقترحة وأساليب ضبط البارامترات الفائقة واختيار الميزات. يتبع التوصيف استعراض النتائج وتقييم الأداء.

8-1 ضبط البارامترات الفائقة Hyperparameters Tuning

تَعني عملية ضبط (أو تحسين) البارامترات الفائقة إيجاد مجموعة قيم البارامترات التي تحقق أفضل أداء لنموذج تعلم الآلة. يركز هذا البحث على استخدام طريقة البحث العشوائي لضبط البارامترات الفائقة، حيث أن البحث العشوائي [23][23] أكثر كفاءة في تحسين البارامترات من البحث الشبكي من الناحية التجريبية والنظرية.

يعمل البحث العشوائي بشكل أفضل عندما يكون فضاء البحث عالي الأبعاد ويحتوي على عدد كبير من التركيبات المختلفة للبارامترات الفائقة. حيث تبحث تقنية البحث العشوائي ضمن مجموعة عشوائية من هذه التركيبات لاختيار أفضل القيم للبارامترات الفائقة

الخاصة بالنموذج المقترح، وبالتالي فإن الوقت المستغرق للعثور على المجموعة الصحيحة يكون أقل مع عدد أقل من التكرارات.

8-2 اختيار الميزات Features Selection

يعتمد البحث على الغابات العشوائية لاختيار الميزات، حيث تتمتع بقدرتها على قياس الأهمية النسبية (Relative Importance) لكل ميزة في التنبؤ 24]]. يتم قياس أهمية الميزة (Feature Importance) من خلال النظر بعدد العقد النقية في نهاية جميع الأشجار التي تستخدم هذه الميزة (يحدث الانقسام عندها باتجاه واحد). بمعنى أخر تكون العقد الشائبة (Impurity Nodes) في بداية الشجرة، بينما تحدث الملاحظات (Observations) التي تسبب نقصاً في شوائب العقد في نهاية الشجرة. بتقليم الأشجار أسفل العقد النقية يمكننا إنشاء مجموعة فرعية من أهم الميزات. يمكن من خلال النظر إلى أهمية الميزة تحديد الميزات التي من المحتمل حذفها لأنها لا تساهم بشكل كافٍ (أو أحياناً لا تساهم على الإطلاق) في عملية التنبؤ.

يبين الجدول*1* ترتيب ميزات مجموعة البيانات المقترحة باستخدام خوارزمية الغابات العشوائية ترتيب ميزات مجموعة البيانات المستخدمة وفقاً لدرجة أهمية كل منها باستخدام خوارزمية الغابات العشوائية. يتم اختيار الميزات الأكثر أهمية التي تحقق أفضل أداء للنموذج من خلال تصفية جميع الميزات التي لها درجة أهمية منخفضة، أي لها تأثير ضعيف في التصنيف.

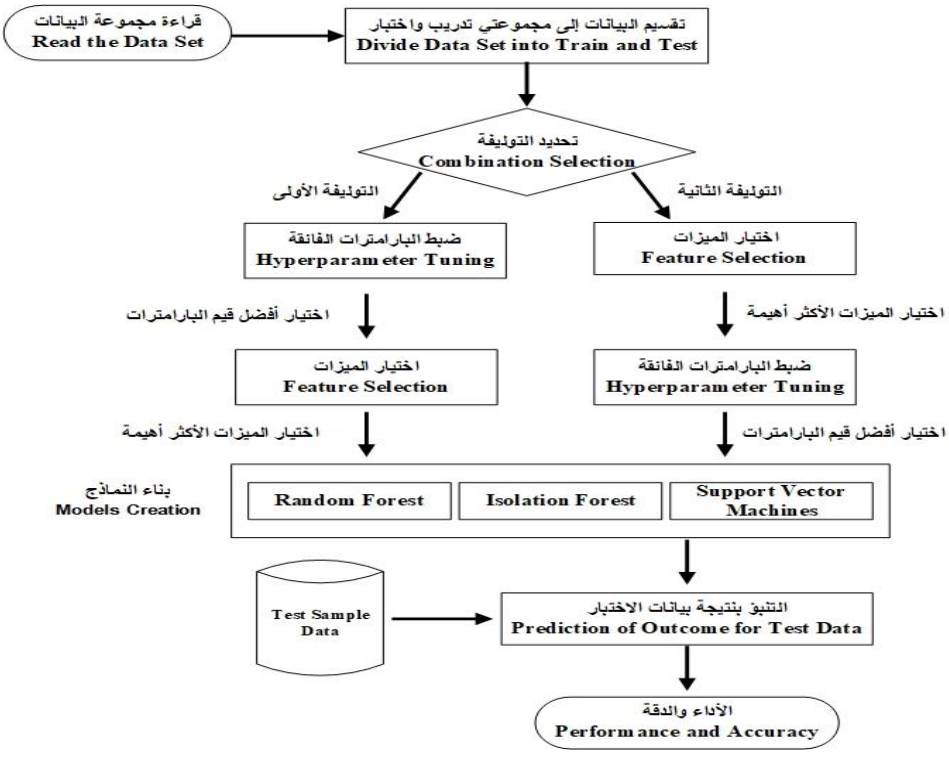
الجدول1 ترتيب ميزات مجموعة البيانات المقترحة باستخدام خوارزمية الغابات العشوائية

Feature	Score	Feature	Score	Feature	Score
V14	0.184	**V9**	0.022	**V15**	0.009
V4	0.113	**V21**	0.020	**V1**	0.007
V12	0.109	**V19**	0.017	**V13**	0.007
V10	0.101	**V27**	0.012	**V6**	0.007
V17	0.088	**V5**	0.012	**V22**	0.006
V3	0.058	**V18**	0.012	**V23**	0.006
V11	0.046	**Amount**	0.012	**V25**	0.006

V16	0.044	**V8**	0.011	**V24**	0.005
V2	0.036	**V26**	0.009	**V28**	0.005
V7	0.023	**V20**	0.009	**Time**	0.005

3-8 بناء النموذج Model Creation

يوضح المخطط التالي الاستراتيجية المقدمة ضمن هذا البحث لاستخدام الأساليب المقترحة لبناء أنظمة كشف الشذوذ (كشف الاحتيال المالي كمثال تطبيقي على النموذج).



الشكل 2 النظام المقترح لاكتشاف الشذوذ

1- اختيار التوليفة

استخدمت الخوارزميات المقترحة لبناء هذه الأنظمة وفق سيناريوهين يختلفان بترتيب التوليفة بين ضبط البارامترات الفائقة واختيار الميزات الأكثر أهمية، وذلك لتحديد التوليفة التي تحقق أفضل أداء لأنظمة كشف الشذوذ من حيث الدقة والوقت.

يكون ترتيب التوليفة من أجل **السيناريو الأول** بالشكل التالي: 1) ضبط البارامترات الفائقة للخوارزميات المقترحة، 2) اختيار الميزات الأكثر أهمية التي تحقق أفضل أداء لهذه الخوارزميات، باستخدام البارامترات الفائقة المحددة من الخطوة السابقة. بينما يكون ترتيبهما من أجل **السيناريو الثاني**: 1) اختيار الميزات الأكثر أهمية، 2) ضبط البارامترات الفائقة للخوارزميات، بالاعتماد على الميزات المحددة من الخطوة السابقة.

2- بناء النموذج

3- اختبار النموذج

يجب التنويه أنه تمَّ تقسيم مجموعة بيانات الدراسة إلى مجموعات تدريب واختبار (Training and Test Sets)، حيث تحتوي مجموعة التدريب على 70% من أجمالي عدد العينات.

8-4 تقييم الأداء Performance Evaluation

8-4-1 خوارزمية الغابات العشوائية

- **السيناريو الأول (scenario 1)**

تظهر النتائج المبينة في **الجدول 2** أداء نموذج الغابة العشوائية في المرحلة الأولى عند ضبط البارامترات الفائقة. بينما يُبيّن

الجدول *3* أداء النموذج في المرحلة الثانية عند اختيار الميزات الأكثر أهمية باستخدام أفضل البارامترات الفائقة المحددة من المرحلة الأولى.

الجدول 2 نتائج ضبط البارامترات الفائقة لنموذج الغابة العشوائية (السيناريو 1)

Precision	Recall	F1	MCC	AUCPR	Failure Rate	Time
0.87	0.85	0.86	0.85	0.84	0.046	31.23M

الجدول 3 نتائج اختيار الميزات لنموذج الغابة العشوائية (السيناريو 1)

#Features	Precision	Recall	F1	MCC	AUCPR	Failure Rate

1	0.27	0.74	0.39	0.44	0.53	0.364
2	0.61	0.76	0.67	0.68	0.72	0.117
3	0.82	0.83	0.83	0.83	0.82	0.055
4	0.82	0.83	0.82	0.82	0.84	0.056
5	**0.87**	**0.88**	**0.87**	**0.87**	**0.86**	**0.041**
6	0.87	0.86	0.86	0.86	0.86	0.043
7	0.86	0.85	0.86	0.85	0.86	0.046
8	0.85	0.87	0.86	0.86	0.86	0.046
9	0.86	0.88	0.87	0.87	0.86	0.043
10	0.85	0.85	0.85	0.85	0.85	0.042

1. ضبط البارامترات الفائقة

نلاحظ من **الجدول *2*** أن أفضل أداء يمكن أن يصل إليه نموذج الغابات العشوائية لاكتشاف الاحتيال المالي بالاعتماد على جميع ميزات مجموعة البيانات المقترحة يكون عند ضبط البارامترات الفائقة بالشكل التالي: عدد الأشجار100، عمق الشجرة 150، الحد الأعظمي للميزات $\sqrt{30}$ (عدد الميزات=30)، الحد الأدنى للعينات 5. حيث يحقق النموذج أداء عالي في تصنيف الحالات الاحتيالية وذلك بالنظر إلى قيم كل من AUCPR وMCC حيث تصل قيمة كل منهما إلى 84% و85% على الترتيب.

2. اختيار الميزات الأكثر أهمية

تظهر نتائج المرحلة الحالية من خلال

الجدول *3* أن نموذج الغابة العشوائية يحقق أفضل أداء له في اكتشاف العمليات الاحتيالية عند اختيار أهم **5** ميزات وهي: **(V14, V4, V10, V12, V17)** وفق ترتيب **الجدول*1***، وباستخدام البارامترات الفائقة المحددة من المرحلة السابقة. حيث تصل قيمة AUCPR إلى 86% وقيمة MCC إلى 87% وذلك عندما تكون قيمة معدل الفشل أقل ما يمكن Failure rate =0.041، وهي أفضل نتيجة تم الحصول عليها مع قيم عالية

أيضاً لباقي المقاييس، وبالتالي فأن النموذج يصنف معظم الحالات الشاذة (الاحتيالية) بشكل صحيح.

- **السيناريو الثاني (scenario 2)**

تظهر النتائج المبينة في **الجدول 4** أداء نموذج الغابة العشوائية في المرحلة الأولى عند اختيار الميزات الأكثر أهمية. بينما يُبيّن

الجدول *5* أداء النموذج في المرحلة الثانية عند ضبط البارامترات الفائقة باستخدام أهم الميزات المحددة من المرحلة الأولى.

الجدول 4 نتائج اختيار الميزات لنموذج الغابة العشوائية (السيناريو 2)

#Features	Precision	Recall	F1	MCC	AUCPR	Failure Rate
1	0.46	0.52	0.49	0.49	0.52	0.172
2	0.86	0.65	0.74	0.75	0.73	0.071
3	0.92	0.80	0.86	0.86	0.84	0.042
4	0.93	0.83	0.86	0.86	0.85	0.041
5	0.94	0.79	0.86	0.86	0.85	0.041
6	**0.96**	**0.80**	**0.87**	**0.88**	**0.85**	**0.037**
7	0.95	0.78	0.85	0.86	0.85	0.042
8	0.96	0.79	0.87	0.87	0.85	0.039
9	0.96	0.79	0.86	0.87	0.85	0.040
10	0.96	0.79	0.87	0.87	0.85	0.039

الجدول 5 نتائج ضبط البارامترات الفائقة لنموذج الغابة العشوائية (السيناريو 2)

Precision	Recall	F1	MCC	AUCPR	Failure Rate	Time
0.96	0.81	0.88	0.88	0.86	0.036	15.12M

1. اختيار الميزات الأكثر أهمية

أظهرت نتائج المرحلة الحالية المبينة في **الجدول 4** أن أفضل أداء لنموذج الغابة العشوائية باستخدام قيم البارامترات الافتراضية يكون عند اختيار أهم **6** ميزات ضمن مجموعة البيانات وفق لترتيبها في **الجدول1** وهي: **(V14, V4, V10, V12, V17,V3)**، حيث تبلغ قيمة AUCPR 85% وقيمة MCC 88% وذلك عندما تكون قيمة معدل الفشل أقل ما يمكن Failure rate=0.037، وهذا مؤشر جيد في تصنيف الحالات الشاذة (الاحتيالية).

2. ضبط البارامترات الفائقة

بالاعتماد على الميزات الأكثر أهمية التي تم اختيارها من المرحلة السابقة، أظهرت نتائج المرحلة الحالية والمبينة في

الجدول *5* تحسين بسيط في أداء نموذج الغابة العشوائية بالنسبة لقيمة AUCPR حيث تبلغ قيمتهُ 86% عند ضبط البارامترات الفائقة بالشكل التالي: عدد الأشجار :250، عمق الشجرة: 150، الحد الأعظمي للميزات: $\sqrt{6}$، الحد الأدنى للعينات:1.

يمكن مقارنة نتائج السيناريوهين السابقين من خلال **الجدول 6 نتائج نموذج الغابات العشوائية**التالي.

الجدول 6 نتائج نموذج الغابات العشوائية

Metrics / Hyperparameters	Scenario 1	Scenario 2
F1	0.87	**0.88**
MCC	0.87	**0.88**
AUCPR	**0.86**	**0.86**
Failure Rate	0.041	**0.036**
Run Time	31.23M	**15.12M**

No. Features	5	6
No. Tree	100	250
Tree Depth	150	150
Maximum Features	$RoundUp(\sqrt{30})$	$RoundUp(\sqrt{6})$

حيث نجد أن أفضل سيناريو لعمل مصنف الغابة العشوائية في اكتشاف الاحتيال المالي هو السيناريو الثاني باستخدام التوليفة "اختيار الميزات الأكثر أهمية ثم ضبط البارامترات الفائقة".

نلاحظ بدايةً تقليص الزمن المستغرق لضبط البارامترات الفائقة حوالي نصف الزمن المستهلك في السيناريو الأول (من 31د إلى 15د)، وهذا مؤشر مهم جداً في أنظمة الكشف عن الشذوذ بشكل عام. كما نجد انخفاض في عدد كل من الإيجابيات الخاطئة FP (المعاملات المشروعة التي صنفت على أنها احتيال) والسلبيات الخاطئة FN (المعاملات الاحتيالية التي صنفت على انها مشروعة) وهذا ما يظهر من خلال قيمة معدل الفشل (ينخفض من 0.041 إلى 0.036)، حيث يعد ذلك من أهم الأهداف التي يسعى إليها إي نظام كشف عن الاحتيال المالي.

8-4-2 خوارزمية غابات العزل

- **السيناريو الأول (scenario 1)**

تظهر النتائج المبينة في **الجدول 7** أداء نموذج غابة العزل في المرحلة الأولى عند ضبط البارامترات الفائقة. بينما يُبيّن **الجدول 8** أداء النموذج في المرحلة الثانية عند اختيار الميزات الأكثر أهمية باستخدام أفضل البارامترات الفائقة المحددة من المرحلة الأولى.

الجدول 7 نتائج ضبط البارامترات الفائقة لنموذج لغابة العزل (السيناريو 1)

Precision	Recall	F1	MCC	AUCPR	Failure Rate	Time
0.49	0.29	0.37	0.38	0.24	0.16	40M

الجدول 8 نتائج اختيار الميزات لنموذج غابة العزل (السيناريو 1)

#Features	Precision	Recall	F1	MCC	AUCPR	Failure Rate
1	0.75	0.41	0.53	0.55	0.55	0.116
2	0.76	0.38	0.51	0.54	0.61	0.117
3	0.84	0.47	0.60	0.63	0.68	0.098
4	0.83	0.48	0.61	0.63	0.65	0.098
5	0.83	0.51	0.64	0.65	0.66	0.094
6	0.81	0.49	0.61	0.63	0.65	0.099
7	0.80	0.51	0.62	0.64	0.64	0.098
8	**0.83**	**0.53**	**0.65**	**0.66**	**0.66**	**0.092**
9	0.81	0.49	0.61	0.63	0.62	0.099
10	0.76	0.46	0.58	0.59	0.59	0.109

1. ضبط البارامترات الفائقة

نلاحظ من **الجدول 7** أن أفضل أداء يمكن أن يصل إليه نموذج غابة العزل لاكتشاف الاحتيال المالي بالاعتماد على جميع ميزات مجموعة البيانات المقترحة يكون عند ضبط البارامترات الفائقة بالشكل التالي: عدد الأشجار:250، العدد الأعظمي للعينات: 550، التلوث: 0.001. لكن بالمقابل نلاحظ أن قيمة AUCPR تبلغ 24% وقيمة MCC 38% وهذا مؤشر على سوء تصنيف الحالات الاحتيالية.

2. اختيار الميزات الأكثر أهمية

تظهر نتائج المرحلة الحالية من خلال **الجدول 8** تحسن كبير في أداء نموذج غابة العزل عند اختيار أهم **8** ميزات ضمن مجموعة البيانات وهي: **(V14, V4, V10, V12, V17,V3,V11,V16)** وفق ترتيبها في **الجدول1**، وباستخدام البارامترات الفائقة المحددة من المرحلة السابقة، حيث نلاحظ ازدياد قيمة AUCPR من 24% في المرحلة السابقة إلى 66% وقيمة MCC من 38% إلى 66%، كما نلاحظ انخفاض قيمة Failure Rate من 0.16 إلى 0.092.

- **السيناريو الثاني (scenario 2)**

تظهر النتائج المبينة في **الجدول *9*** أداء نموذج غابة العزل في المرحلة الأولى عند اختيار الميزات الأكثر أهمية. بينما يُبيّن **الجدول *10*** أداء النموذج في المرحلة الثانية عند ضبط البارامترات الفائقة باستخدام أهم الميزات المحددة من المرحلة الأولى.

الجدول 9 نتائج اختيار الميزات لنموذج غابة العزل (السيناريو 2)

#Features	Precision	Recall	F1	MCC	AUCPR	Failure Rate
1	0.62	0.64	0.63	0.63	0.61	0.119
2	0.58	0.58	0.58	0.58	0.61	0.113
3	0.68	0.67	0.66	0.66	0.66	0.110
4	**0.68**	**0.72**	**0.70**	**0.70**	**0.68**	**0.099**
5	0.68	0.66	0.67	0.67	0.60	0.104
6	0.70	0.65	0.67	0.67	0.62	0.101
7	0.57	0.60	0.58	0.58	0.51	0.137
8	0.67	0.69	0.68	0.68	0.65	0.103
9	0.67	0.68	0.67	0.67	0.64	0.104
10	0.55	0.64	0.59	0.59	0.60	0.142

الجدول 10 نتائج ضبط البارامترات الفائقة لنموذج لغابة العزل (السيناريو 2)

Precision	Recall	F1	MCC	AUCPR	Failure Rate	Time
0.80	0.71	0.75	0.74	0.71	0.083	18M

1. اختيار الميزات الأكثر أهمية

أظهرت نتائج المرحلة الحالية المبينة في **الجدول *9*** أن أفضل أداء لنموذج غابة العزل باستخدام قيم البارامترات الافتراضية يكون عند اختيار أهم **4** ميزات ضمن مجموعة البيانات المستخدمة وفق ترتيب **الجدول*1*** وهي: **(V14, V4, V10, V12)**. حيث تبلغ قيمة AUCPR 68% وقيمة MCC 70% وذلك عندما تكون قيمة معدل الفشل أقل ما يمكن في هذه المرحلة Failure rate=0.099، أي أن النموذج يحقق أداء مقبول في تصنيف الحالات الاحتيالية.

2. ضبط البارامترات الفائقة

بالاعتماد على الميزات الأكثر أهمية التي تم اختيارها من المرحلة السابقة، أظهرت نتائج المرحلة الحالية والمبينة في **الجدول 10** تَحسن في أداء نموذج غابة العزل عند ضبط البارامترات الفائقة بالشكل التالي: عدد الأشجار:100، العدد الأعظمي للعينات: 300، التلوث: 0.00178. حيث تصل قيمة AUCPR إلى 71% وقيمة MCC إلى 74% كما تنخفض قيمة Failure rate إلى 0.083.

يمكن مقارنة نتائج السيناريوهين السابقين من خلال **الجدول 11***الجدول 6* **نتائج نموذج الغابات العشوائية**التالي. حيث نجد أن أفضل سيناريو لعمل نموذج غابة العزل في اكتشاف الاحتيال المالي هو السيناريو الثاني باستخدام التوليفة "اختيار الميزات الأكثر أهمية ثم ضبط البارامترات الفائقة". حيث أظهرت النتائج أن غابة العزل لا يمكن أن تحقق أداء جيد في اكتشاف الحالات الاحتيالية دون اختيار الميزات الأكثر أهمية رغم ضبط البارامترات الفائقة الخاصة بها. نلاحظ أيضاً تقليص الزمن المستغرق لضبط البارامترات الفائقة أكثر من نصف الزمن المستهلك في السيناريو الأول (من 40د إلى 18د)، كما نجد انخفاض في عدد كل من الإيجابيات الخاطئة FP والسلبيات الخاطئة FN وهذا ما يظهر من خلال قيمة معدل الفشل (ينخفض من 0.092 إلى 0.083).

الجدول 11 نتائج نموذج غابات العزل

Metrics / Hyperparameters	Scenario 1	Scenario 2
F1	0.65	**0.75**
MCC	0.66	**0.74**
AUCPR	0.66	**0.71**
Failure Rate	0.092	**0.083**
Run Time	40M	**18M**
No. Features	8	4
No. Tree	250	100

Sup Sample	550	300
Contamination	0.001	0.00178

8-4-3 خوارزمية آلة شعاع الدعم

- **السيناريو الأول (scenario 1)**

تظهر النتائج المبينة في **الجدول 12** أداء نموذج آلة شعاع الدعم في المرحلة الأولى عند ضبط البارامترات الفائقة. بينما يُبيّن **الجدول 13** أداء النموذج في المرحلة الثانية عند اختيار الميزات الأكثر أهمية باستخدام أفضل البارامترات الفائقة المحددة من المرحلة الأولى.

الجدول 12 نتائج ضبط البارامترات الفائقة لنموذج آلة شعاع الدعم (السيناريو 1)

Precision	Recall	F1	MCC	AUCPR	Failure Rate	Time
0.79	0.82	0.81	0.81	0.77	0.063	15.74M

الجدول 13 نتائج اختيار الميزات لنموذج آلة شعاع الدعم (السيناريو 1)

#Features	Precision	Recall	F1	MCC	AUCPR	Failure Rate
1	0.79	0.36	0.49	0.53	0.61	0.117
2	0.74	0.40	0.52	0.55	0.67	0.117
3	0.84	0.57	0.68	0.69	0.73	0.087
4	0.85	0.62	0.71	0.72	0.76	0.078
5	0.85	0.76	0.81	0.81	0.76	0.059
6	0.84	0.76	0.80	0.80	0.75	0.061

7	**0.86**	**0.79**	**0.82**	**0.82**	**0.76**	**0.055**
8	0.84	0.80	0.82	0.82	0.76	0.055
9	0.84	0.80	0.82	0.82	0.77	0.056
10	0.82	0.82	0.82	0.82	0.77	0.057

1. ضبط البارامترات الفائقة

نلاحظ من **الجدول 12** أن أفضل أداء يمكن أن يصل إليه نموذج آلة شعاع الدعم لاكتشاف الاحتيال المالي بالاعتماد على جميع ميزات مجموعة البيانات المقترحة يكون عند ضبط البارامترات بالشكل التالي:C :3000، Gamma: 0.001، kernel: 'rbf'. حيث تبلغ قيمة كل من AUCPR وMCC 77% و81% على الترتيب، أي أن للنموذج أداء جيد في تصنيف الحالات الاحتيالية.

2. اختيار الميزات الأكثر أهمية

تظهر نتائج المرحلة الحالية من خلال **الجدول *13*** إن النموذج يحافظ على أدائه من المرحلة الأولى عند اختيار أهم **7** ميزات وهي: **(V14, V4, V10, V12, V17,V3,V11)** ضمن مجموعة البيانات وفق **الجدول*1* Error! Reference source not found.**وباستخدام البارامترات الفائقة المحددة من المرحلة السابقة. لكن من جهة أخرى نلاحظ انخفاض عدد كل من الإيجابيات الخاطئة (FP) والسلبيات الخاطئة (FN) بناء على قيمة معدل الفشل (ينخفض من 0.063 إلى 0.055) وهذا مؤشر هام في أنظمة كشف الشذوذ عموماً.

- **السيناريو الثاني (scenario 2)**

تظهر النتائج المبينة في **الجدول 14** أداء نموذج آلة شعاع الدعم في المرحلة الأولى عند اختيار الميزات الأكثر أهمية. بينما يُبيّن **الجدول 15** أداء النموذج في المرحلة الثانية عند ضبط البارامترات الفائقة باستخدام أهم الميزات المحددة من المرحلة الأولى.

الجدول 14 نتائج اختيار الميزات لنموذج آلة شعاع الدعم (السيناريو 2)

#Features	Precision	Recall	F1	MCC	AUCPR	Failure Rate
1	0.71	0.53	0.61	0.61	0.51	0.109
2	0.78	0.53	0.63	0.67	0.62	0.098
3	0.85	0.63	0.73	0.73	0.66	0.076
4	0.85	0.68	0.76	0.76	0.71	0.069
5	0.85	0.80	0.83	0.83	0.75	0.055
6	**0.85**	**0.81**	**0.83**	**0.83**	**0.77**	**0.054**
7	0.85	0.82	0.83	0.83	0.73	0.054
8	0.82	0.82	0.82	0.82	0.77	0.057
9	0.81	0.83	0.82	0.82	0.77	0.059
10	0.79	0.83	0.81	0.81	0.76	0.062

الجدول 15 نتائج ضبط البارامترات الفائقة لنموذج آلة شعاع الدعم (السيناريو 2)

Precision	Recall	F1	MCC	AUCPR	Failure Rate	Time
0.87	0.8	0.83	0.83	0.82	0.051	7.59

1. اختيار الميزات الأكثر أهمية

أظهرت نتائج المرحلة الحالية المبينة في **الجدول 14** أن أفضل أداء لنموذج آلة شعاع الدعم باستخدام قيم البارامترات الافتراضية يكون عند اختيار أهم **6** ميزات ضمن مجموعة البيانات المستخدمة وفق ترتيب **الجدول1** وهي: **(V14, V4, V10, V12, V17,V3)**. حيث تبلغ قيمة AUCPR 77% وقيمة MCC 83% وذلك عندما تكون قيمة معدل الفشل أقل ما يمكن في هذه المرحلة Failure rate=0.054، أي أن النموذج يحقق أداء جيد في تصنيف الحالات الاحتيالية.

2. ضبط البارامترات الفائقة

بالاعتماد على الميزات الأكثر أهمية التي تم اختيارها من المرحلة السابقة، أظهرت نتائج المرحلة الحالية والمبينة في **الجدول 15** تَحسن ملحوظ في أداء نموذج آلة شعاع الدعم عند ضبط البارامترات بالشكل التالي: C:1000، Gamma: 0.01، kernel: "rbf". حيث يحقق النموذج أداء جيد للغاية في تصنيف الحالات الاحتيالية لتصل قيمة AUCPR إلى 82% وقيمة MCC إلى 83% كما تنخفض قيمة Failure rate إلى 0.051.

يمكن مقارنة نتائج السيناريوهين السابقين من خلال **الجدول 16***الجدول 6* **نتائج نموذج الغابات العشوائية**التالي.

الجدول 16 نتائج نموذج آلة شعاع الدعم

Metrics / Hyperparameters	Scenario 1	Scenario 2
F1	0.82	**0.83**
MCC	0.82	**0.83**
AUCPR	0.76	**0.82**
Failure Rate	0.055	**0.051**
Run Time	15.74M	**7.59M**
No. Features	7	6
C	3000	1000
Gamma	0.001	0.01

حيث نجد أن أفضل سيناريو لعمل مصنف آلة شعاع الدعم في اكتشاف الاحتيال المالي هو السيناريو الثاني باستخدام التوليفة "اختيار الميزات الأكثر أهمية ثم ضبط البارامترات الفائقة". حيث أظهرت النتائج تقليص الزمن المستغرق لضبط البارامترات الفائقة حوالي نصف الزمن المستهلك في السيناريو الأول (من 15د إلى 7.5د). كما نجد انخفاض في عدد كل من الإيجابيات الخاطئة FP والسلبيات الخاطئة FN وهذا ما يظهر من خلال قيمة معدل الفشل (ينخفض من 0.055 إلى 0.051).

أكدت نتائج التجارب العملية لهذا البحث، أن اختيار الميزات الأكثر أهمية ضمن مجموعة البيانات متبوعاً بضبط البارامترات الفائقة يساهم في تحسين أداء جميع الخوارزميات المقترحة بشكل كبير. حيث أظهرت النتائج أن خوارزمية الغابات العشوائية لها أفضل أداء في كشف الحالات الشاذة MCC=88%، تليها خوارزمية آلة شعاع الدعم MCC=83%، ومن ثم خوارزمية غابة العزل MCC=74%.

كما ساهمت النتائج التي توصل إليها البحث في اختزال الزمن من 50% إلى 53%، وزيادة دقة الكشف من 5% إلى 6%، وتحسين كل مقاييس الأداء المستخدمة لجميع الخوارزميات المقترحة.

9-الخاتمة والاستنتاجات والتوصيات

تمَّ في هذا البحث دراسة أهم العوامل التي تؤثر على أداء أنظمة الكشف عن الشذوذ بشكل عام، وهي تحديد التوليفة الأمثل من ضبط البارامترات الفائقة واختيار الميزات الأكثر أهمية؛ وذلك من خلال تحليل أنظمة كشف الاحتيال المالي ضمن بطاقات الائتمان الأوربية بالاعتماد على عدة خوارزميات تعلم آلة وهي الغابات العشوائية وآلة شعاع الدعم وغابة العزل.

توصلت نتائج البحث إلى أن أفضل سياق لبناء أنظمة كشف الشذوذ هو اختيار الميزات الأكثر أهمية ضمن مجموعة البيانات ومن ثم ضبط البارامترات الفائقة الخاصة بها. حيث أظهرت التجارب العملية تقليص الزمن المستغرق لبناء هذه الأنظمة عند استخدام السياق المحدد إلى النصف أو أكثر، بالإضافة إلى زيادة في كشف الحالات الشاذة وذلك بالنظر إلى قيم AUCPR بحدود 5% لخوارزمية غابة العزل و6% لخوارزمية آلة شعاع الدعم ومن أجل جميع الخوارزميات المقترحة بشكل عام، مما يؤكد على التحسين الحاصل عند اعتماد الترتيب المذكور.

تفوقت خوارزمية الغابات العشوائية على الخوارزميات الأخرى المستخدمة من حيث دقة كشف الشذوذ وبالتالي كشف العمليات الاحتيالية، وذلك باختيار أهم 6 ميزات ضمن مجموعة البيانات وضبط البارامترات الفائقة بالشكل التالي: عدد الأشجار:250، عمق الشجرة: 150، الحد الأعظمي للميزات: $\sqrt{6}$، الحد الأدنى للعينات:1. حيث كانت أفضل

النتائج لهذه الخوارزمية 0.88 :MCC- 0.88 :F1- 0.86 :AUCPR، بينما كان الزمن المستغرق 15.12 دقيقة. بالمقابل حققت خوارزمية آلة شعاع الدعم أفضل أداء من حيث زمن التنفيذ بمقدار 7.59 دقيقة، وذلك عند اختيار أهم 6 ميزات وضبط البارامترات الفائقة بالشكل التالي: C:1000، Gamma: 0.01، kernel: "rbf". أن أفضل النتائج لهذه الخوارزمية 0.83 :MCC- 0.83 :F1- 0.82 :AUCPR. أما بالنسبة لخوارزمية غابة العزل فقد حققت أداء جيد عن اختيار أهم 4 ميزات وضبط البارامترات الفائقة بالشكل التالي: عدد الأشجار:100، العدد الأعظمي للعينات: 300، التلوث: 0.00178. أن أفضل النتائج لهذه الخوارزمية 0.74 :MCC- 0.75 :F1- 0.71 :AUCPR وبزمن تنفيذ 18 دقيقة.

وبناء على النتائج التي تم التوصل إليها يمكن أن نقدم بعض المقترحات:

- توسيع نطاق العمل ليشمل خوارزميات التعلم العميق ومجالات أخرى للكشف.
- محاولة التوسع في الدراسة لتشمل مجالات أخرى في كشف الشذوذ.
- السعي لتطبيق الدراسة على بيانات الشركات والمؤسسات المالية السورية.

المراجع

[1] How Much Data Is on the Internet? [Online] Available: seedscientific.com [Accessed 1 Jul. 2021]
[2] Zenati, H., Romain, M., Foo, C. S., Lecouat, B., & Chandrasekhar, V. (2018, November). Adversarially learned anomaly detection. In 2018 IEEE International conference on data mining (ICDM) (pp. 727-736).
[3] Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., & Talwalkar, A. (2018). Massively parallel hyperparameter tuning.
[4] Florea, A. C., & Andonie, R. (2020). Weighted random search for hyperparameter optimization. arXiv preprint arXiv:2004.01628.
[5] Van Der Maaten, L., Postma, E. and Van den Herik, J., (2009). Dimensionality reduction: a comparative. J Mach Learn Res, 10(66-71).
[6] Tang, J., Alelyani, S. and Liu, H., 2014. Feature selection for classification: A review. Data classification: Algorithms and applications, p.37.
[7] Credit card fraud detection anonymized credit card transaction labeled as fraudulent or genuine [Online] Available: https://www.kaggle.com/mlg-ulb/creditcardfraud. [Accessed 1 Jul. 2021]
[8] Python 3.10.0 documentation [Online] Available: https://docs.python.org/3/ [Accessed 1 May. 2021]
[9] Chicco, D. (2017). Ten quick tips for machine learning in computational biology. BioData mining, 10(1), 1-17.
[10] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.
[11] Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia computer science, 48(2015), 679-685.
[12] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) (pp. 1-6). IEEE.
[13] Ounacer, S., El Bour, H. A., Oubrahim, Y., Ghoumari, M. Y., & Azzouazi, M. (2018). Using Isolation Forest in anomaly detection: the case of credit card transactions. Periodicals of Engineering and Natural Sciences (PEN), 6(2), 394-400.
[14] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, 93010-93022.
[15] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.

[16] Lakshmi, S. V. S. S., & Kavilla, S. D. (2018). Machine learning for credit card fraud detection system. International Journal of Applied Engineering Research, 13(24 Pt. 1), 16819-16824.

[17] Raghavan, P., & El Gayar, N. (2019, December). Fraud detection using machine learning and deep learning. In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 334-339). IEEE.

[18] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157-175). Springer, Boston, MA.

[19] Staerman, G., Mozharovskyi, P., Clémençon, S., & d'Alché-Buc, F. (2019, October). Functional isolation forest. In Asian Conference on Machine Learning (pp. 332-347). PMLR.

[20] "Anomaly detection using Isolation Forest – A Complete Guide" [online]. https://www.analyticsvidhya.com. [Accessed 26 July 2021].

[21] "Isolation Forest Algorithm for Anomaly Detection" [online]. https://heartbeat.comet.ml. [Accessed Oct 2021]

[22] F. Cady (2017)-The Data Science Handbook,1st edition, NJ: John Wiley and Sons Ltd

[23] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

[24] Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in Society.