مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (١٣) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (13) 2025

دراسة تأثير تقنيات اختيار الميزات على أداء خوارزمية الخلايا الجذعية المناعية DCA في تصنيف الهجمات في الشبكات الحاسوبية

د. يعرب ديوب*

د. جعفر سلمان **

سالي محمد عيسى ***

(تاريخ الإيداع ٢٠٢٥/٨/١٠ . قُبِل للنشر في ٢٠٢٥/١٠/٢٣)

□ ملخّص □

استدعى النمو الهائل في حركة مرور الشبكة traffic network والتهديدات الأمنية المرتبطة بها إلى ضرورة تطوير أنظمة لكشف وتصنيف الهجمات في الشبكات بما يقلل من تأثيرها على الأفراد والمؤسسات والمجتمعات ويضمن الأمن السيبراني.

تعتمد أنظمة التصنيف القائمة على تقنيات الذكاء الصنعي وخوارزمياته على كميات هائلة من بيانات الشبكة مما يجعل عملية اختيار مجموعة الميزات المناسبة خطوة حاسمة في تحسين كفاءة النظام ودقته وقابليته للتفسير، وذلك من خلال دورها في تقليل أبعاد مجموعة البيانات والحد من الضوضاء الموجودة فيها بالإضافة لإزالة الميزات المكررة أو غير المرتبطة بفئات التصنيف.

تدرس هذه الورقة البحثية مجموعة من تقنيات اختيار الميزات المعتمدة على التعلم الآلي Machine Learning مع خوارزمية إزالة الميزات المتكررة RFE مثل RF-RFE,SVM-RFE,LR-RFE بالإضافة للخوارزمية الجينية من خلال تطبيقها على مجموعة البيانات المعيارية UNSW-NB15 المكونة من 45ميزة متعلقة بحركة البيانات والاتصال في الشبكات الحاسوبية.

تم تحليل أداء هذه التقنيات وفعاليتها في اختيار مجموعة الميزات المثلى بحالتي 14-8 ميزة ومقارنة أداءها في تحسين دقة نماذج التصنيف من خلال اختبار نتائجها مع خوارزمية الخلايا الجذعية المناعية DCA لتحديد الأنشطة الضارة وكشف التهديدات في الشبكات.

أظهرت النتائج النهائية لهذه الدراسة أداء مميز لنموذج التصنيف مع تطبيق تقنيات اختيار الميزات بشكل ملحوظ مما يؤكد فعاليتها في تقليل الأبعاد والضوضاء وإزالة البيانات المكررة حيث حققت الخوارزمية الجينية أفضل النتائج بدقة وصلت إلى 99% باستخدام 8 ميزات فقط ومعدل إنذارات كاذبة FAR منخفض جداً وصل إلى 1.61% متفوقة بذلك على خوارزميات التعلم الآلي التي كانت خوارزمية الغابة العشوائية الأفضل بينها بدقة وصلت إلى 97.6%.

الكلمات المفتاحية: خوارزميات اختيار الميزات، الشبكات الحاسوبية، خوارزمية الخلايا الجذعية، مجموعة البيانات UNSW-NB15، إزالة الميزات المتكررة.

* أستاذ في قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس–سوريا.

** أستاذ مساعد في قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس–سوريا.

*** طالبة دكتوراه – قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس–سوريا.

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (١٣) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (13) 2025

Studying the Impact of Feature Selection Techniques on the Performance of Dendritic Cell Algorithms (DCA) in Classifying Network Attacks

Dr. Yaroub Dayoub*
Dr. Jaafar Salman **
Sally Mohammad Issa***

(Received 10/8/2025 . Accepted 23/10/2025)

□ ABSTRACT □

The exponential growth in network traffic and associated security threats necessitates the development of robust systems for detecting and classifying network attacks. This is crucial for mitigating their impact on individuals, organizations, and societies, thereby ensuring cybersecurity.

Classification systems based on artificial intelligence techniques and algorithms rely on massive amounts of network data. This makes the process of selecting an appropriate feature subset a critical step in enhancing system efficiency, accuracy, and interpretability. This is achieved by reducing the dimensionality of the dataset, mitigating noise, and eliminating redundant or irrelevant features that don't contribute to classification categories.

This research paper investigates a set of machine learning-based feature selection techniques including Redundant Feature Elimination (RFE) algorithms, such as RF-RFE, SVM-RFE, LR-RFE, and Genetic Algorithms (GA). These Techniques are applied to the benchmark dataset UNSW-NB15, which comprises 45 features related to network traffic and connection behavior in computer networks. The performance and effectiveness of these techniques in selecting the optimal feature set for 8-14 features were analyzed, and their performance in improving the accuracy of classification models was compared by testing their results with the immune Dendritic cell algorithm (DCA) for identifying malicious activities and detecting threats in networks.

The final results of this study demonstrated an outstanding performance of the classification model when feature selection techniques were applied. this confirms their effectiveness in dimensionality reduction noise mitigation and redundancy elimination. The Genetic Algorithm achieved the best results with an accuracy of 99% using only 8 features and a very low false alarm rate (FAR) of 1.61%, outperforming other machine learning-based algorithms, among which the Random Forest Algorithm with an accuracy of 97.6%.

Keywords: Features Selection Algorithms, Computational Networks, Dendritic Cell Algorithm, UNSW-NB15 Dataset, RFE.

---

*Professor, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.

**Assistant Professor, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria.

***PhD Student, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria

١. مقدمــة:

شهدت الشبكات الحاسوبية في السنوات الأخيرة تطوراً ملحوظاً، وذلك تزامناً مع الاعتماد المتزايد على تقنيات الاتصال وتبادل البيانات بكافة أشكالها عبر الأنترنت، مما جعلها بيئة خصبة للاستهداف من قبل الهجمات الإلكترونية وعمليات الاختراق. نتيجة لذلك، أصبح من الضروري تطوير أنظمة ذكية قادرة على كشف هذه الهجمات الإلكترونية وتحديد التهديدات التي تؤثر على سرية وسلامة وتوافر البيانات بدقة وكفاءة [1].

تعد خوارزميات التصنيف أحد أهم الأدوات المستخدمة في بناء أنظمة لكشف الهجمات وتحديد السلوك الشاذ في الشبكات، حيث تعتمد على تحليل سلوك حركة البيانات Traffic وتصنيفها إلى طبيعية أو ضارة، وبالرغم من التطور الكبير الذي يشهده هذا المجال فإن فعالية هذه الخوارزميات تتأثر بشكل كبير بكمية ونوعية الميزات (features) الموجودة في مجموعات البيانات المستخدمة في التدريب والاختبار، فوجود عدد كبير من الميزات غير المهمة أو المكررة قد يؤدي إلى عدة مشاكل منها التعقيد الحسابي، زيادة زمن التصنيف، وبالتالي يضعف أداء النموذج.

يبرز دور تقنيات اختيار الميزات (Feature Selection) في هذا السياق كخطوة رئيسية في تحسين أداء خوارزميات التصنيف بمختلف أنواعها ومجالاتها، وذلك من خلال العمل على تقليل أبعاد مجموعات البيانات والتركيز فقط على تحديد مجموعة فرعية من الميزات الأكثر تأثيراً في عملية الكشف والتصنيف [2].

طُوّرت العديد من التقنيات لمعالجة مشكلة تقليل الميزات غير ذات الصلة والمكررة التي تُشكّل عبئًا على عمليات التصنيف، والتي أثبتت فعاليتها في تحسين أداء مختلف الخوارزميات.

سيتم التركيز في هذا البحث على دراسة تأثير تقنيات استخلاص الميزات على أداء خوارزمية الخلايا الجذعية المناعية (Dendritic Cell Algorithm-DCA)، والتي تعد من الخوارزميات الحيوية المستوحاة من آليات عمل الخلايا الجذعية في جهاز المناعة البشري[3] ، حيث تتميز بقدرتها على تصنيف الأنشطة الشبكية إلى طبيعية أو ضارة بالاعتماد على إشارات بيئية وسلوكية مستخلصة من حركة البيانات ضمن الشبكة.

٢. هدف البحث:

يهدف هذا البحث إلى دراسة المتغيرات المدرجة في مجموعة البيانات UNSW-NB15 لاستخلاص الميزات الأكثر أهمية وذلك باستخدام مجموعة تقنيات استخلاص الميزات وخاصة SVM-RFE، Random Forest-RFE، والخوارزمية الجينية Genetic algorithm، مع تحليل أداء كل منها على أنظمة تصنيف الهجمات وتحديد التهديدات في الشبكات الحاسوبية من خلال دمجها مع خوارزمية الخلايا الجذعية المناعية في تصنيف الهجمات وذلك بغرض:

١- تقليل أبعاد مجموعة البيانات.

٢- تسريع عمليات التدريب والاختبار.

٣- تقليل التعقيد الحسابي من خلال اختيار مجموعات السمات الأكثر ملائمة لمشكلة التصنيف.

٤- تحسين أداء نموذج التصنيف.

٣. الدراسات المرجعية:

هنالك العديد من المقالات والأوراق البحثية التي ناقشت عملية استخلاص الميزات وتأثيرها المباشر على نتائج التصنيف في العديد من التطبيقات الطبية والخدمية والمالية مع دراسة لأهم التقنيات المستخدمة لهذا الغرض وتحليل أدائها، نستعرض منها:

١. قدم الباحثان Marwa Hassanو Naima Kaabouch دراسة حديثة في عام 2024 والتي تستهدف اختبار تأثير تقنيات اختيار الميزات المختلفة على أداء نماذج التعلم الآلي في مجال الكشف عن الاكتئاب باستخدام بيانات تخطيط كهربائية الدماغ (EEG). قارنت الدراسة بين سته طرق لاختيار الميزات وهي:

الشبكة المرنةElastic Net ، والمعلومات المتبادلة Mutual Information (MI)، ومربع كاي Chi-Square ، واختيار الميزات الأمامي مع الانحدار التدرجي العشوائي (FFS-SGD)، وإزالة السمات المتكررة القائمة على آلة متجه الدعم (SVM-RFE)، والحد الأدنى للتكرار والأهمية القصوى (mRMR). أظهرت النتائج تفوق ملحوظ لتقنية SVM-RFE حيث حققت أعلى مستوى للدقة ومقياس F1 في تصنيف حالات الاكتئاب بناءً على الإشارات الكهربائية للدماغ [4].

٢. اقترح الباحث "Omar Almomani" ورقة بحثية [5] قدم من خلالها نموذج لاختيار الميزات في أنظمة كشف التسلل في الشبكات NIDS بالاعتماد على مجموعة من خوارزميات ذكاء الأسراب والخوارزميات التطويرية وهي: خوارزمية أسراب الطيور PSO وخوارزمية الذئب الرمادي GWO، خوارزمية اليراع FFA ، بالإضافة للخوارزمية الجينيةGA، وذلك لتحديد الميزات الأكثر تأثيراً في دقة التصنيف وبالتالي تحسين أداء النموذج المبني على خوارزمية آلة المتجهات الداعمة SVM ومصنف شجرة J48. أظهرت نتائج التصنيف أن الخوارزمية الجينية حققت قيم جيدة من حيث معدل الإيجابيات الحقيقية المرتفع TPR ومعدل السلبيات الكاذبة المنخفض FNR مع انخفاض عدد الميزات المستخدمة، بينما أظهرت خوارزمية PSO نتائج جيدة من حيث الدقة ومعدل السلبيات الحقيقية TNR.

٣. في عام 2023 تم اقتراح نموذج لاختيار الميزات باستخدام مزيجاً من تقنيني التصفية وهما: تقنية ربح المعلومات Information Gain (IG)، وتقنية الغابة العشوائية Random forest بغرض تقليل مساحة البحث عن الميزات مع تطبيق تقنية الإزالة التكرارية للميزات RFE لتقليل الميزات الزائدة[6] . تم استخدام مجموعة الميزات المستخلصة من قبل شبكة بيرسبتون العصبونية متعددة الطبقات MLP لكشف التسلل، حيث أظهرت النتائج تحسناً في دقة التصنيف من 82.25% إلى 84.24% مع تقليل عدد الميزات من 42إلى 23ميزة من مجموعة البيانات UNSW-NB15.

٤. قدم الباحث Pralhad Chapagain وآخرون، دراسة اقترحوا من خلالها نموذج هجين لكشف التسلل يعتمد على تقنيات تقليل الأبعاد وخوارزميات التجميع، حيث تم استخدام تحليل المكونات الرئيسية (PCA) لاستخلاص الميزات وتقليل الأبعاد، مما ساعد في تحسين أداء خوارزمية التجميع المحسنة K-Means المستخدمة لاحقاً في عملية التصنيف. أجريت التجارب على مجموعة البيانات NSL-KDD. أظهرت النتائج النهائية تحسناً ملحوظاً في قيم معدلات الكشف وتقليل معدل الإنذارات الكاذبة وبالتالي أداء نموذج التصنيف مقارنة بالخوارزميات التقليدية المطبقة على كافة الميزات ضمن مجموعة البيانات [7] .

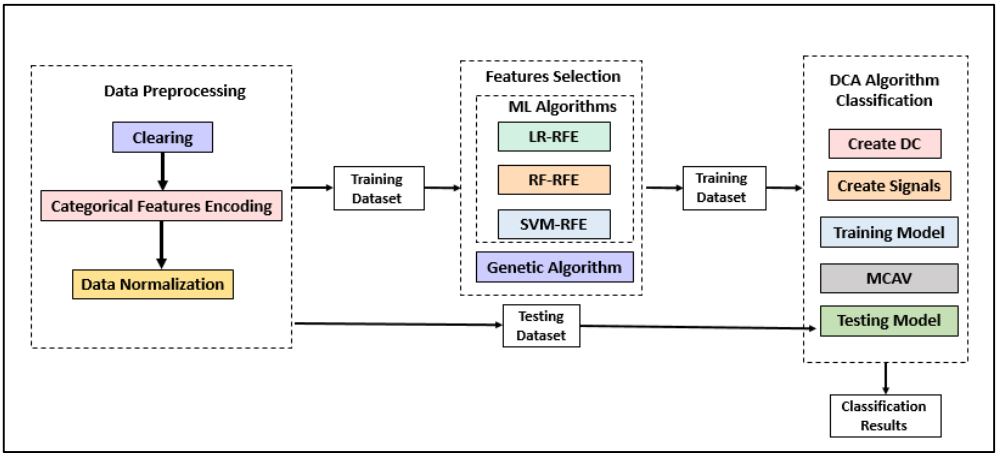
تُظهر الدراسات السابقة بوضوح مدى أهمية تقنيات استخلاص الميزات وتأثيرها الكبير في تحسين أداء أنظمة كشف التسلل، سواء من حيث رفع قيم دقة التصنيف، تقليل الأبعاد، وتقليل معدل الإنذارات الكاذبة مما يجعلها مرحلة أساسية تؤثر بشكل مباشر على نتائج خوارزميات التعلم الآلي والتعلم العميق الحديثة. مع ذلك، وعلى الرغم من التقدم الملحوظ في هذا المجال لا يزال هناك فجوات واضحة من خلال:

- ☒ التركيز على نوع واحد من الطرق لاستخلاص الميزات في أغلب الأبحاث.
- ☒ عدم دراسة تأثير الخوارزمية مع عدد مختلف من الميزات الفرعية المختارة مع قبل نفس الخوارزمية.

- ☒ الافتقار إلى مقارنات شاملة بين تقنيات استخلاص الميزات المدرجة تحت نفس النوع (التغليف Wrapper، التصفية Filtering، التضمينEmbedding).
- ☒ تتم المقارنة بين التقنيات بناءً على بعض معايير التصنيف دون غيرها.

٤. منهجية البحث:

يوضح الشكل (1) المراحل العملية والخطوات المتبعة في البحث بدءاً من معالجة مجموعة البيانات واختيار الميزات المناسبة وفق ثلاث خوارزميات وصولاً إلى تطبيق خوارزمية التصنيف للحصول على نتائج التقييم النهائية.



الشكل (1): الخطوات الرئيسية العملية للبحث

٤,١ توصيف مجموعة البيانات UNSW-NB15:

تعد مجموعة البيانات هذه إحدى أهم مجموعات البيانات الحديثة المستخدمة في تقييم أداء أنظمة كشف الهجمات وقد نُشرت لأول مرة عام 2015 من قبل مركز الأمن السيبراني الأسترالي (Cyber Range Lab at UNSW Canberra). تضم هذه المجموعة **2,540,044** سجلاً تم تجميعها باستخدام بيئة محاكاة واقعية لحركة مرور البيانات في الشبكة، كما تم تصنيف الهجمات فيها إلى تسعة أنواع رئيسية بناءً على 45 ميزة تغطي جوانب متعددة من حركة الشبكة، بما في ذلك معلومات التدفق، الخصائص الأساسية، محتوى الحزم Packets، الوقت، الأغراض العامة، والاتصال (لاحظ الجدول (2))[8].

تنقسم مجموعة البيانات إلى مجموعتي تدريب واختبار كالتالي:

جدول (1): توزيع السجلات في مجموعة البيانات UNSW-NB15

Dataset	Total Sample Size	Normal	Attacks
Training Set	175,341	56,000	119,341
Testing Set	82,332	37,000	45,332

جدول (2): الميزات المدرجة ضمن مجموعة البيانات UNSW-NB15

Feature No	Feature Name	Feature No	Feature Name	Feature No	Feature Name
1	id	16	dloss	31	response_body_len
2	dur	17	sinpkt	32	ct_srv_src
3	proto	18	dinpkt	33	ct_state_ttl
4	service	19	sjit	34	ct_dst_ltm
5	state	20	djit	35	ct_src_dport_ltm
6	spkts	21	swin	36	ct_dst_sport_ltm
7	dpkts	22	stcpb	37	ct_dst_src_ltm
8	sbytes	23	dtcpb	38	is_ftp_login
9	dbytes	24	dwin	39	ct_ftp_cmd
10	rate	25	tcprtt	40	ct_flw_http_mthd
11	sttl	26	synack	41	ct_src_ltm
12	dttl	27	ackdat	42	ct_srv_dst
13	sload	28	smean	43	is_sm_ips_ports
14	dload	29	dmean	44	attack_cat
15	sloss	30	trans_depth	45	label

٤,٢ مرحلة اختيار الميزات Feature Selection:

تعد عملية اختيار الميزات من أهم المراحل في أنظمة المعالجة الذكية للبيانات، فغالباً ما تحوي مجموعة البيانات على ميزات قد لا تكون مهمة أو فائضة عن الحاجة فحسب، بل قد تؤثر سلباً على دقة النتائج، وبالتالي فإنّ اختيار السمات المناسبة يمكن أن ينعكس إيجاباً على أداء نماذج التعلم.

يشير مفهوم اختيار الميزات إلى آلية اختيار مجموعة فرعية من البيانات تكون الأكثر أهمية وارتباطاً بمجال التطبيق مع إزالة الميزات غير ذات الصلة والمكررة، وذلك بهدف التعامل مع مشكلة الأبعاد المرتفعة وبالتالي تحسين أداء نماذج التصنيف من خلال زيادة سرعة التدريب وتقليل التعقيد الحسابي[9] .

هنالك مجموعة واسعة من تقنيات استخلاص الميزات المستخدمة بكثرة في هذا المجال المعتمدة على طبيعة البيانات والتطبيق المطلوب والتي سنقوم بتحليل عدد منها وندرس أداءها عند تطبيقها مع خوارزمية الخلايا الجذعية المناعية في كشف التهديدات وتصنيف الهجمات باستخدام مجموعة البيانات UNWS-NB15 ، وهي:

٤,٢,١ خوارزمية الإزالة التكرارية للميزات RFE مع خوارزميات التعلم الآلي ML:

ظهرت هذه الخوارزمية مع بدايات تطوير تقنيات تعمل على تحسين أداء النماذج الإحصائية والتعلمية الآلية من خلال معالجة مشكلة التعقيد العالي للنماذج والناتج عن إدخال عدد كبير من السمات غير المهمة وبالتالي تقليل دقة وأداء النماذج عند تطبيقها على البيانات الجديدة، مما خلق بدوره حاجة ضرورية لاستخدام تقنيات منهجية تعمل على اختيار الميزات الأكثر أهمية وتأثيراً في نموذج التعلم (الآلي والعميق) وحذف الميزات الأقل أهمية تدريجياً والتي كانت خوارزمية RFE إحداها[10] .

الحذف التكراري هو طريقة اختيار الميزات التي تضم الخصائص الرئيسية لمجموعة البيانات وذلك من خلال الخطوات التالية:

١. اختيار مصنف(نموذج) تعلم آلي وتدريبه على كامل الميزات ضمن مجموعة البيانات المحددة.

٢. حساب أوزان السمات feature weights والتي بدورها تعكس أهمية كل ميزة Feature Ranking (FR) باستخدام خوارزمية التعلم الآلي المُختارة .

٣. ترتيب جميع الميزات تصاعدياً أو تنازلياً وفقاً لأوزانها.

٤. إزالة واستبعاد الميزات ذات الأهمية المنخفضة، ومن ثم يتم إعادة تدريب المصنف على الميزات المتبقية حتى الوصول إلى العدد المطلوب من الميزات.

بعد حذف الميزات الأضعف سنقوم بتطبيق خوارزمية RFE مع مصنفات التعلم الآلي (الانحدار اللوجستي Linear Regression، آلة دعم المتجهات Support Vector Machine(SVM) ، والغابة العشوائية Random Forest (RF)) و ودراسة آلية كل منها في اختيار الميزات الأكثر أهمية من مجموعة البيانات لدينا:

❖ **تطبيق خوارزمية RFE مع مصنف الانحدار اللوجستي (Logistic Regression-RFE):**

هو أحد تقنيات التعلم الآلي الخاضع للإشراف Supervised Learning والمستخدمة لنمذجة علاقة خطية بين مجموعة من المتغيرات المستقلة (المدخلات) ومتغير تابع واحد (الخرج).

بالرغم من استخدامها الأساسي في مجال التنبؤ، إلا أنّها نستخدم بشكل فعال في استخلاص الميزات وذلك وفق الخطوات التالية:

١- تدريب نموذج الانحدار الخطي على جميع الميزات ضمن مجموعة البيانات بحيث يتم حساب وزن كل ميزة وفق العلاقة(1) [11]:

$$W= (X^{T}\cdot X)^{-1}X^{T}.y$$ العلاقة (1)

حيث: X: مصفوفة المدخلات والتي تمثل قيم الميزات لجميع العينات ضمن مجموعة التدريب.

X^{T}: المصفوفة المنقولة لـ X.

y : متجه النتائج الحقيقية والذي يمثل القيم الفعلية للمتغير الهدف.

٢- ترتيب الميزات حسب أهمية أوزانها (أكبر وزن بالقيمة المطلقة).

٣- استبعاد الميزة الأقل أهمية في كل تكرار.

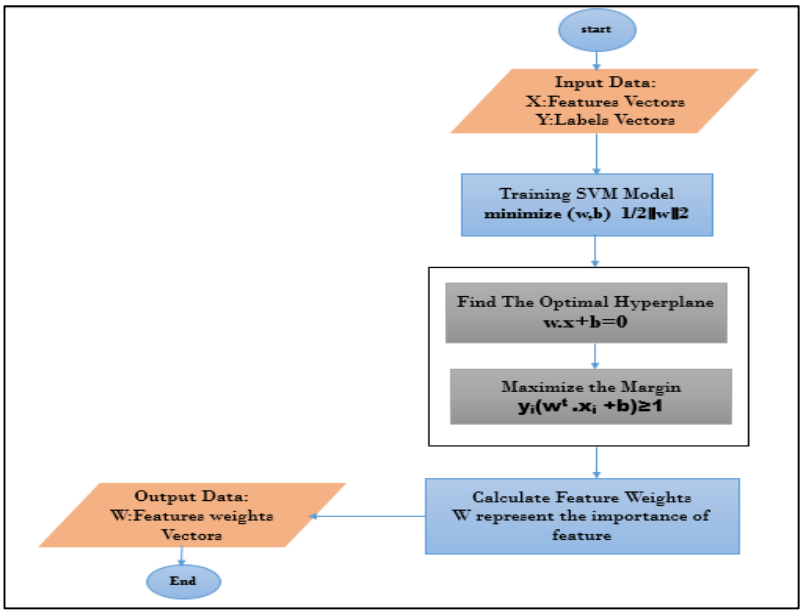
٤- تكرار العملية حتى الوصول إلى عدد الميزات المطلوب.

❖ **تطبيق خوارزمية RFE مع مصنف آلة دعم المتجهات SVM (SVM-RFE):**

تعتبر خوارزمية SVM من الأدوات المهمة للتعلم الآلي الخاضع للإشراف Supervised Learning والتي تنتمي إلى خوارزميات التصنيف Regression، حيث أصبحت في الآونة الأخيرة من أكثر تقنيات التصنيف استخداماً بفضل قدرتها العالية على التعامل مع مجموعات البيانات ذات الأبعاد العالية بكفاءة ودقة.

تكمن الفكرة الأساسية وراء خوارزمية SVM في إيجاد المستوى الفائق Hyperplane الذي يفصل بين فئتين (صنفين) عن طريق تعظيم الهامش بينهما، والذي يمثل المسافة بين المستوى الفائق وأقرب نقاط بيانات (والتي تُسمى المتجهات الداعمة Support Vectors) على كلا الجانبين.

هذا بدوره يضمن التحسين من تعميم النموذج وتقليل فرص الخطأ إلى أدنى حد له وفق الخطوات الموضحة في الشكل التالي:



الشكل (2): مراحل عمل المصنف SVM

١. يُمثل المستوى الفائق Hyperplane العلاقة التالية [12]:

$$W.X + b = 0$$ العلاقة (2)

حيث: W: وزن كل ميزة، X:متجه الميزات، b : مقدار الإزاحة.

٢. هدف SVM هو إيجاد الوزن w والحد b بحيث يتم تعظيم المسافة بين الفاصل وأقرب النقاط من كل فئة:

$$y_i\,(w_t\,.x_i + b) \geq 1$$
$$\text{minimize } (w, b)\ \ 1/2\|w\|2$$

حيث: x_i : متجه الخصائص، y_i :التصنيف الصحيح.

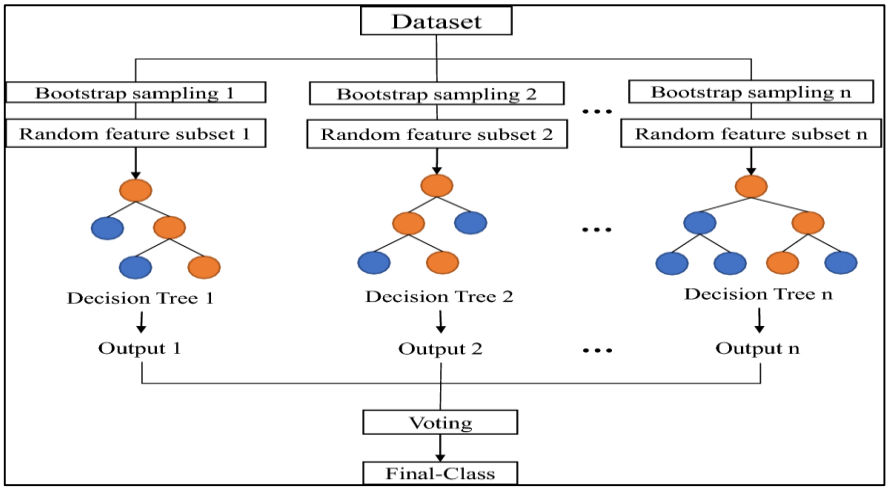
٣. بعد تدريب SVM على جميع الميزات والحصول على أوزان الميزات w من نموذج التدريب، والتي بدورها تمثل أهمية كل ميزة، يبدأ عمل خوارزمية RFE في ترتيب الميزات حسب أهميتها بحيث تكون الميزات ذات قيم الأوزان الأقل (الأقرب للصفر) ذات الأهمية الأقل.

٤. في نهاية الخوارزمية يتم الحصول على مجموعة الميزات المختارة النهائية، والتي يتم تطبيق خوارزمية التصنيف المناعية عليها.

❖ **تطبيق خوارزمية RFE مع مصنف الغابة العشوائية (Random forest-RFE):**

تعتبر خوارزمية الغابة العشوائية من أشهر خوارزميات التعلم الآلي والتي تندرج تحت مفهوم تقنيات التعلم الجماعي Ensemble Learning المستخدم في مجال التصنيف والانحدار وفق التالي [13]:

تعتمد الخوارزمية على بناء عدة أشجار قرار decision trees(الغابة) وذلك من خلال عملية الاختيار العشوائي لعينات مختلفة من بيانات التدريب تدعى عينات التمهيد (Bootstrap Sample)، حيث يتم استخدام مجموعة بيانات فرعية مختلفة لكل شجرة، كما نلاحظ في الشكل (3)



الشكل (3): آلية عمل خوارزمية Random forest في التصنيف

عند كل عقدة (Node) يتم فيها انقسم البيانات باستخدام ميزة معينة، يتم حساب مقدار التحسين في جودة البيانات بعد هذا الانقسام وذلك من خلال حساب قيمة ربح المعلومات Information Gain المعتمدة على قيمة الانتروبيا Entropy والتي تعتبر مقياس لعدم اليقين أو العشوائية في مجموعة البيانات وتحسب وفق العلاقة (3) :[14]

$$\text{Entropy (A)} = \sum_{i=1}^{n} -\text{Pi log2 Pi}$$ العلاقة(3)

n: عدد الفئات.

P_i: الاحتمال المتكرر للفئة i (فئة التصنيف)

بناءً على قيمة Entropy لكل سمة يتم حساب التحسين الناتج عن كل منها وفق العلاقة (4) للربح بحيث كلما زادت قيمته، زادت أهمية هذه الميزة، وفي نهاية كل شجرة فرعية يتم جمع كل قيم

$$\text{Gain(S,A)=Entropy(S)} - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \text{Entropy(Si)}$$ العلاقة (4)

التحسين التي ساهمت فيها كل ميزة عبر جميع العقد المستخدمة فيها.

حيث أنّ:

A: السمات، **S**: مجموعة قيم السمات، **N**: عدد أجزاء السمة A.

|Si|: عدد البيانات في الجزء i، **|S|**: عدد البيانات في S.

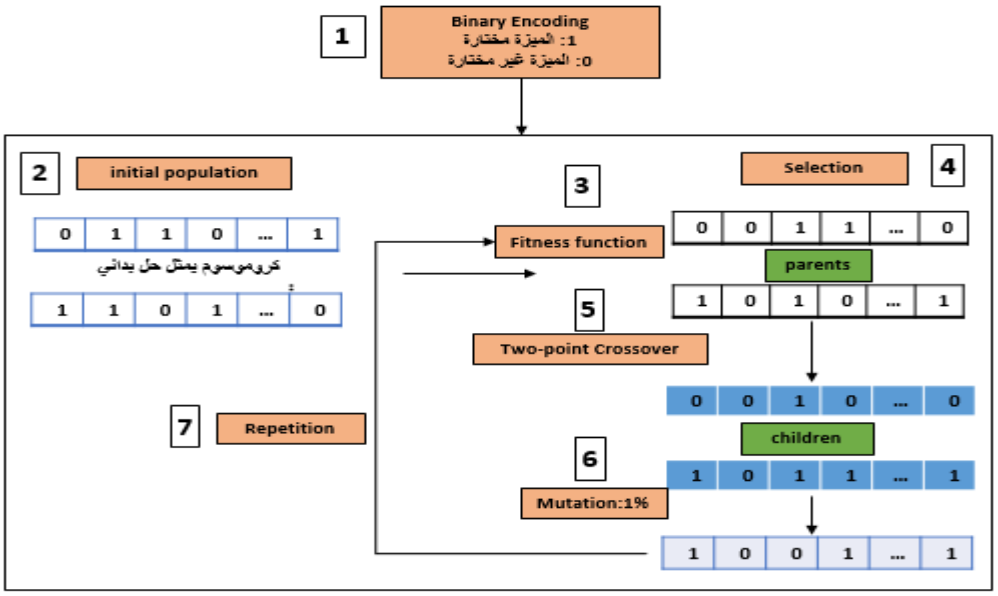
في نهاية تشكيل الغابة المكونة من مجموعة أشجار فرعية، يتم جمع مقدار التحسين لكل ميزة والتي ساهمت به في جميع أشجار الغابة يتم تحديد أهميتها وترتيبها وفقه لذلك، ليتم استخدامها من قبل خوارزمية RFE.

٤,٢,٢ الخوارزمية الجينية Genetic Algorithm:

هي إحدى خوارزميات البحث الأمثل (Optimization Algorithms) المستوحاة من نظرية التطور الطبيعي لداروين، والتي تستند في عملها إلى مبادئ الانتقاء الطبيعي والبقاء للأصلح. تستخدم هذه الخوارزمية بكفاءة في حل المسائل المعقدة التي يصعب فيها استخدام الطرق التقليدية، خاصة عند وجود عدد كبير من الاحتمالات كما في اختيار الميزات في التعلم الآلي [15].

تقوم فكرة الخوارزمية الجينية على محاكاة عملية تطور الكائنات الحية، حيث يتم تمثيل كل حل محتمل على أنه "فرد" في المجتمع Population، وتتطور هذه الأفراد على مدى عدة أجيال عبر عمليات بيولوجية مثل الطفرات (Mutation) بهدف تحسين الحلول تدريجياً وصولاً إلى أفضل حل ممكن.

يوضح الشكل (4) مراحل اختيار الميزات الأمثل باستخدام الخوارزمية الجينية:



الشكل (4): آلية عمل الخوارزمية الجينية في اختيار الميزات

٤,٣ تصنيف الهجمات في الشبكات الحاسوبية بالاعتماد على خوارزمية الخلايا الجذعية المناعية DCA:

يعد أمن الشبكات من القضايا الحيوية في العصر الرقمي الحديث، حيث تتعرض الأنظمة والشبكات الحاسوبية بشكل مستمر لهجمات تهدف إلى سرقة البيانات أو تعطيل الخدمات.

بسبب ضرورة تأمين الحماية لهذه الأنظمة وتطويرها، برزت الحاجة إلى الاعتماد على تقنيات ذكية قادرة على كشف الهجمات وتصنيفها بدقة وفعالية.

ومن بين الأساليب الحديثة في هذا المجال، ظهرت أنظمة المناعة الاصطناعية وخوارزمياتها الحيوية والتي تحاكي في كل منها آلية مناعية دفاعية في الجسم وبالأخص خوارزمية الخلايا الجذعية المناعية DCA المستوحاة من آلية عمل جهاز المناعة البشري بوصفه نهج بيولوجي قادر على تحليل إشارات حركة البيانات ضمن الشبكة وتحديد النشاط الضار [16] .

تدمج خوارزمية DCA عدة إشارات سلوكية وبيئية لتقديم نتيجة التصنيف النهائية والتي بناءً عليها يتم تمييز النشاط الطبيعي من المشبوه، وهي:

١. إشارات آمنة (Safe Signals): تشير إلى أن السلوك المراقب طبيعي ولا تدعو للقلق، وتكون الخلية بحالة تدعى غير الناضجة (Immature State).

٢. إشارات الخطر (Danger Signals): تنتج عن سلوك غير اعتيادي قد يكون مؤشر لهجوم أو نشاط مشبوه، وتكون الخلية بحالة تدعى شبه الناضجة (Semi-mature State).

٣. إشارات ممرضة (PAMP-Pathogen Associated Molecular Patterns): هي إشارات قوية جداً توحي بوجود تهديد أكيد، مثل توقيع Signature معروف لهجوم سابق.

يتضمن الشكل (1) توضيح لمراحل عمل خوارزمية الخلايا الجذعية المناعية من توليد الخلايا والإشارات المناعية وصولاً للتصنيف النهائي وتقييم النموذج.

٤,٤ القسم العملي:

تم تصميم ثمانية سيناريوهات مختلفة لتقييم أداء خوارزمية DCA تحت ظروف مختلفة من حيث تقنيات اختيار الميزات من مجموعة البيانات UNSW-NB15 وعدد الميزات المستخدمة (14 أو 8 ميزات)، بالإضافة للسيناريو الرئيسي الذي يعتمد على جميع الميزات (45 ميزة) كالتالي:

٤,٤,١ اختيار الميزات باستخدام خوارزمية LR-RFE:

في هذا السيناريو يتم استخدام خوارزمية الانحدار اللوجستي لاختيار (8-14) من ميزات مجموعة البيانات والتي سيتم استخدامها لاحقاً كمدخلات لخوارزمية التصنيف DCA، وذلك بهدف دراسة تأثير الميزات المختارة وفق هذه الخوارزمية وعددها على نتائج التصنيف. كما هو موضح في الجدول التالي:

الجدول (3): الميزات المختارة عند تطبيق خوارزمية LR-REF

الميزات المختارة من أجل 14ميزة	مجموعة A	dttl	swin	dtcpb	dload	dbytes	dpkts	spkts	ct_srv_dst
	الأهمية	16.989	11.152 –	-7.211	-6.572	−4.163	3.868	3.007	−2.859
	مجموعة B	sbytes	dmean	state	ct_state_ttl	ct_dst_sport_ltm	ct_dst_src_ltm		
	الأهمية	−2.677	2.52	2.375	1.866	1.806	1.799		
الميزات المختارة من أجل 8 ميزات		dttl	swin	dload	spkts	sbytes	state	adtcpb	dmen
أهمية الميزة		12.087	10.973 –	−9.372	4.867	-3.831	2.49	−2.289	1.727

- نظراً لأن الانحدار اللوجستي هو نموذج خطي، فإن الأوزان التي يحددها تعكس الأهمية الخطية للميزات في التنبؤ بالمتغير الهدف، حيث تشير القيمة المطلقة للوزن إلى قوة تأثير الميزة على احتمالية أن يكون المتغير الهدف هجوم (1)، فكلما زادت القيمة المطلقة للوزن تزداد أهمية الميزة.
- قد تكون إشارة الوزن:

- موجبة (+): يعني أن زيادة قيمة هذه الميزة (مع ثبات الميزات الأخرى) يزيد من احتمالية أن يكون المتغير الهدف (1) وبالتالي زيادة احتمالية الهجوم.

- سالبة (−): يعني أن زيادة قيمة هذه الميزة (مع ثبات الميزات الأخرى) يقلل من احتمالية أن يكون المتغير الهدف (1) وبالتالي تقليل احتمالية الهجوم.

- نلاحظ من قيم sbytes, dbytes السالبة المرتفعة لعدد بايتات المصدر والوجهة أن هناك علاقة عكسية بين قيمها المرتفعة والهجمات أي بحسب النموذج أن الحركة الطبيعية غالباً تكون بحجم بايتات مرتفع مثل بث الفيديو، تحميل الملفات الكبير، تحديثات البرامج).

- تظهر قيم الأوزان إلى أن الواصفة TTL تعتبر الأعلى أهمية والتي تؤثر بشكل كبير على نتائج التصنيف، حيث أنها تعتبر ميزة حيوية توفر رؤية حول مصدر الحزمة، المسار الذي سلكته، وحتى نظام التشغيل الخاص بالمصدر أو الوجهة، بالتالي قيمها غير المتوقعة تعتبر مؤشر قوي لوجود هجوم.

- تم الاحتفاظ بالميزات الأعلى أهمية في حالة 14 ميزة واختيارها من قبل النموذج في حالة 8 ميزات.

٤,٤,٢ اختيار الميزات باستخدام خوارزمية RF-RFE:

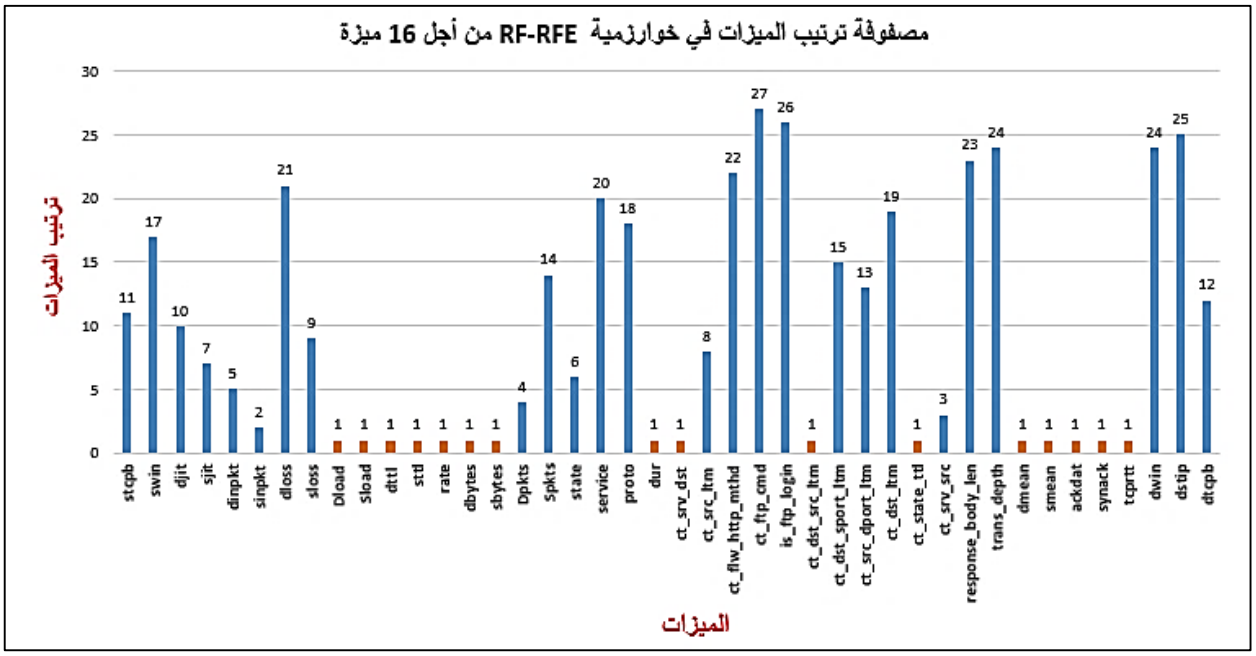
- تعمل خوارزمية RF على حساب أهمية كل ميزة بناء على مدى تقليلها للتباين (الخطأ) في كل شجرة حيث تم تحديد عدد الأشجار في الغابة بمقدار 100 شجرة مما يضمن زيادة قوة النموذج ودقته دون التكلف باستهلاك أي وقت إضافي.

الجدول (4): الميزات المختارة عند تطبيق خوارزمية RF-REF

أسماء الميزات المختارة من أجل 14ميزة	مجموعة A	'sttl'	'ct_state_ttl'	'dload'	'sbytes'	'sload'	'tcprtt'	'rate'	'dmean'
	الأهمية	0.211	0.167	0.114	0.052	0.048	0.047	0.044	0.042
	مجموعة B	'ct_srv_dst'	'dttl'	'ackdat'	'smean'	'ct_dst_src_ltm'	'dur'		
	الأهمية	0.041	0.04	0.038	0.037	0.032	0.028		
الميزات المختارة من أجل 8 ميزات		'sttl'	'ct_state_ttl'	'dload'	'rate'	sbytes	'sload'	'ct_srv_dst'	'dmeansz'
أهمية الميزة		0.293	0.154	0.125	0.109	0.09	0.084	0.072	0.069

- في نهاية عدد التكرار لخوارزمية REF تنتج لدينا مصفوفة ترتيب الميزات والتي تضم أرقام صحيحة موجبة كما هو موضح في الشكل (5)، يشير الرقم ضمنها إلى:

- الميزات المختارة الأكثر أهمية في حال كانت القيمة 1.
- ترتيب إزالة الميزة في حال كانت القيمة أكبر من 1.



الشكل (5): مصفوفة الترتيب الناتجة عن خوارزمية RF–RFE في حالة 14 ميزة

تشير المصفوفة إلى أن الميزات ذات القيم الأعلى تمت إزالتها في الجولات الأخيرة من الخوارزمية وبالتالي أهميتها العالية أعطتها الفرصة للبقاء للنهاية، على عكس الميزات ذات الترتيب المنخفض (باستثناء 1) والتي كانت الأقل أهمية ضمن مجموعة الميزات وبالتالي تم استبعادها من قبل الخوارزمية في جولاتها الأولى.

- نلاحظ عدم وجود قيم سالبة في قياس أهمية الميزة وذلك لأن RF لا تعتمد على علاقات خطية بسيطة، بل تحدد الأهمية بناءً على قدرة الميزة على فصل فئات التصنيف بشكل فعال عبر عتبات متعددة في أشجار القرار ضمن الغابة.
- الميزات ذات قيم الأهمية الأعلى وهي 'ct_state_ttl', sttl والتي تعتبر ميزات حيوية في كشف التسلل لأنها توفر معلومات حول كيفية معالجة الحزم في الشبكة والأنماط السلوكية لمصدر الاتصال وحالته، وبالتالي أي انحراف عن الأنماط المتوقعة يكون مؤشر قوي على وجود نشاط خبيث.
- نلاحظ الاحتفاظ بالميزات ذات الأهمية الأعلى عند تقليل عدد الميزات إلى 8، لكن قيم الأهمية تزداد وذلك لأن الميزات المتبقية تمتص بعضاً من القوة التنبئية للميزات التي تم إزالتها.

٤,٤,٣ اختيار الميزات باستخدام خوارزمية SVM-RFE:

- مجموعة الميزات التي يختارها نموذج SVM تمثل الميزات التي تعطي أفضل فصل بين فئات التصنيف بأكبر هامش ممكن ضمن مجموعة البيانات.
- يحدد نموذج SVM الخطي وزن لكل ميزة ضمن مجموعة البيانات لدينا، والذي يمثل أهميتها في تحديد الفاصل (hyperplane) الذي يفصل بين الفئات.
- يتم ترتيب الميزات حسب القيمة المطلقة للوزن، فكلما زاد وزنها زادت أهمية الميزة في قرار التصنيف النهائي للبيانات.

الجدول (5): الميزات المختارة عند تطبيق خوارزمية SVM-REF

الميزات المختارة من أجل 14ميزة	مجموعة A	'sloss'	dloss"	dpkts	sbytes	'sttl‘	'dttl’	ct_src_ltm'	'smean’
	الأهمية	-2.241	-2.137	1.657	1.377	1	0.929	-0.796	-0.313
	مجموعة B	tcprtt	'swin‘	proto	sinpkt'	stcpb'	dload		
	الأهمية	-0.27	0.242	0.159	-0.145	-0.137	-0.089		
الميزات المختارة من أجل 8 ميزات		'dttl'	'dtcpb'	smean	'ct_src_ltm'	'sttl'	dloss	'dload'	'sloss'
أهمية الميزة		0.844	-0.754	-0.742	0.419	-0.236	0.038	-0.019	-0.013

- إشارة الوزن قد تكون:
  - موجبة (+): تشير إلى أن زيادة قيمة الميزة تدفع التصنيف نحو فئة الهجوم (1).
  - سالبة (−): تشير إلى أن زيادة قيمة الميزة تدفع التصنيف نحو فئة الحركة الطبيعية (0).
- بحساب قيمة معامل الانحياز Bais والذي يسير إلى مقدار الإزاحة للفاصل عن نقطة الأصل عندما تكون جميع الميزات محايدة (قيمتها صفر) نجد أنه من أجل 16 ميزة حصلنا على قيمة (2.934)، أما في حال 8 ميزات كانت قيمته (6.199).

نلاحظ تغير قيم الأهمية وإشارتها لبعض الميزات بشكل واضح عند تقليل العدد إلى 8 ميزات، وهذا يشير إلى أن خوارزمية SVM قد وجدت تفاعلات جديدة بين الميزات المتبقية كون الميزات تكون مرتبطة ببعضها وبالتالي تحسن مجموعة الميزات بناءً على كيفية تأثيرها معاً على هامش الفصل بين فئات التصنيف وليس فقط على أهميتها الفردية البسيطة.

٤,٤,٤ اختيار الميزات باستخدام خوارزمية GA:

قمنا بتطبيق الخوارزمية الجينية على مجموعة البيانات لدينا بعد ضبط المعاملات وفق التالي:

- حجم السكان Population: 150، – عدد الأجيال Number Of Generations :50 جيل.
- طريقة اختيار الأفراد Tournament Selection مع قيمة Tournsize=3 .
- طريقة التزاوج: Two – Point Crossover، – معدل التزاوج: Crossover Rate: 0.8
- تطبيق مبدأ النخبة ELITISM بنسبة 0.05، – معدل الطفرة Mutation Rate:0.01.

❖ **<u>تابع اللياقة:</u>** تعتبر الأهم في عمل الخوارزمية الجينية كونها المعيار الرئيسي لاختيار أفضل الأفراد في كل جيل. اعتمدنا في بحثنا على مبدأ اختيار الميزات القائم على الارتباط CFS (Correlation-based Feature Selection) وهو عبارة عن مقياس يساعد في تقييم جودة مجموعة فرعية من الميزات التي تضمن وجود:

- ترابط خطي قوي بينها وبين المتغير الهدف Relevant (نتيجة التصنيف).

- ترابط خطي ضعيف بين بعضها البعض Non-Relevant.

وقد تم تمثيل هذا المقياس في العلاقة التالية:

حيث: K: عدد الميزات المختارة.

$$\textbf{Fitness CFS} = \sum_{i=1}^{k} r_{cf_i} \Big/ \sqrt{K + \sum_{i=1}^{k} \sum_{j=1/j\neq i}^{k} rff_{ij}}$$

العلاقة (5)

r_{cf_i} : القيمة المطلقة لمعامل الارتباط بين الميزة والقيمة الهدف.

$r_{cf_{ij}}$: القيمة المطلقة لمعامل الارتباط بين الميزات في المجموعة الفرعية.

قمنا بالاعتماد على هذا المبدأ في توليد دالة اللياقة المستخدمة في تقييم جودة الأفراد في كل جيل ضمن الخوارزمية الجينية بغرض اختيار الحلول التي لها تأثير كبير على نتيجة التصنيف وبالمقابل تكون مستقلة عن الميزات الأخرى، مما يضمن الحفاظ في كل جيل على الميزات ذات الأهمية الأعلى والدور الأعظم في تصنيف الهجمات.

❖ عند تطبيق الخوارزمية الجينية وفق البارامترات السابقة واجهتنا مشكلة وهي أن هذه الخوارزمية على خلاف الخوارزميات السابقة لا تتيح خيار تحديد عدد الميزات التي نريدها بغرض إجراء عملية المقارنة مع نتائج الخوارزميات السابقة من أجل 8 أو 14 ميزة، لحل هذه المشكلة قمنا باقتراح طريقة تعتمد على مبدأ العقوبة /المكافئة للحلول وفق التالي:

❖ **<u>طريقة العقوبة والمكافئة</u>:** تعتمد هذه الآلية المقترحة على فرض مبدأ العقوبة أو المكافئة لكل فرد ضمن المجتمع Population وذلك بناءً على مدى بعد عدد الميزات التي اختارها هذا الفرد عن العدد المستهدف على اعتبار أن كل حل هو عبارة عن عينة من الميزات المختارة من مجموعة البيانات.

يتم حساب درجة العقوبة بناءً على العلاقة التالية:

$$\textbf{Penalty(S,target\_k)=Penalty\_weight * |current\_k –target\_k\_feature|/ max\_feauture}$$

العلاقة (6)

حيث: Penalty_weight: وزن العقوبة.

current_k: عدد الميزات الحالي.

target_k_feature: العدد المستهدف الذي نريده من الميزات (*8-14*).

|current_k –target_k_feature|: الفرق المطلق بين العدد الفعلي للميزات المختارة والعدد المستهدف، كلما زاد هذا الفرق زادت قيمة العقوبة للفرد.

max_feauture: العدد الإجمالي للميزات ضمن مجموعة البيانات الأساسية (*42*).

- بعد حساب قيمة العقوبة لكل فرد ودرجة لياقته *Fitness CFS* يتم حساب قيمة اللياقة النهائية للأفراد وفق العلاقة:

$$\textbf{Final\_Fitness= Fitness\_CFS - Penalty}$$

العلاقة (7)

–بتطبيق الآلية المقترحة مع الخوارزمية الجينية حصلنا على النتائج الموضحة في الجدول (6):

الجدول (6): الميزات المختارة عند تطبيق الخوارزمية الجينية GA:

الميزات المختارة من أجل 14ميزة	مجموعة A	sttl	ct_state_ttl	state	proto	dload	rate	dmean	ct_dst_sport_ltm
	r_{cf_i}	0.6696	0.5463	0.5169	0.4358	0.3715	0.3214	0.3107	0.2614
	مجموعة B	ct_src_dport_ltm	sload	service	is_sm_ips_ports	ackdat	sinpkt		
	r_{cf_i}	0.2044	0.1844	0.1789	0.1726	0.1637	0.1631		
قيمة اللياقة Fitness	0.7444 (74.44%)								
	المميزات المختارة من أجل 8 ميزات	sttl	ct_state_ttl	state	proto	dload	rate	ct_dst_sport_ltm	service
	r_{cf_i}	0.6696	0.5463	0.5169	0.4358	0.3715	0.3214	0.2614	0.1789
	قيمة اللياقة Fitness	0.8222.22%)							

–تم اختيار الميزات التي تحقق أعلى درجة ارتباط مع الهدف بالإضافة إلى تحقيق التكامل مع باقي الميزات مما يمنع التكرار.

–نلاحظ أنه من أجل 14 ميزة تم اختيار نفس المجموعة في حالة 8 ميزات مع إضافة الميزات القريبة منها والتي تعتبر الأقرب بناءً على درجة العقوبة /المكافئة.

–إن الزيادة الملحوظة في قيمة اللياقة للحل المقترح من أجل 8 ميزات يشير إلى نجاح آلية التقييم المقترحة، حيث أن GA لم تختر الميزات الأكثر ارتباطاً فقط، بل قانت باختيار مجموعة الميزات التي تحقق توازناً مثالياً بين العلاقة بالهدف والاستقلالية المتبادلة للميزات

٠,٨٢٢٣ مرحلة التصنيف باستخدام خوارزمية الخلايا الجذعية المناعية DCA:

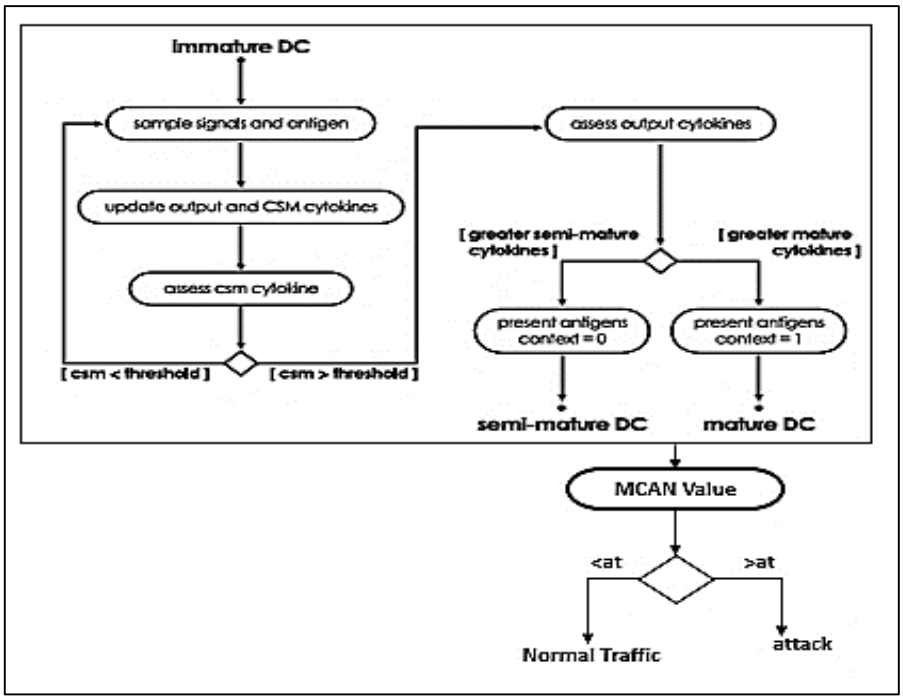
بعد تطبيق تقنيات استخلاص الميزات والحصول على مجموعة فرعية منها والتي تعتبر الأكثر أهمية وارتباطاً بتصنيف الهجمات وفق كل خوارزمية مطبقة، قمنا باستخدام مجموعة البيانات الناتجة كدخل لخوارزمية DCA المستخدمة في عملية تصنيف البيانات إلى هجوم أو حركة طبيعية وذلك بعد ضبط البارامترات التالية:

- عدد الخلايا المناعية DC_Num: 200 خلية.
- عتبة الهجرة (النضوج) Migration_Threshold :0.5%.
- مصفوفة أوزان الإشارات الناتجة في مرحلة التدريب: والتي تم حسابها بناءً على مدى أهمية قيم كل واصفة في التمييز بين السلوك الطبيعي أو الشاذ.

الجدول (7): مصفوفة الأوزان المرتبطة بحالات الخلية الثلاثة

	W_PAMP	W_SS	W_DS
CSM_Weight	0.571	0.0868	0.312
smDC_Weight	0.078	0	0.622
mDC_Weight	0.771	0.0868	0.568

يوضح الشكل (6) مراحل عمل الخلايا المناعية المدربة والناضجة في تصنيف حركة البيانات ضمن الشبكة وتمييز السلوك الطبيعي من الهجوم بناء على قيمة مستوى النضوج MCAV.



الشكل (6): خطوات تصنيف الهجمات باستخدام خوارزمية DCA

١ النتائج والمناقشة:

بتحليل أسماء ميزات مجموعة البيانات التي تم اختيارها من قبل خوارزميات اختيار الميزات الأربعة وارتباطها الفعال بفئة التصنيف وتحديد الهجمات نلاحظ أن أهم هذه الميزات هي الموضحة في الجدول التالي:

الجدول (8): الميزات الأكثر ارتباطاً وتأثيراً بنتائج التصنيف:

النوع	الواصفة	التأثير
• ميزات التحكم بالاتصال	sttl,dttl,ct_state_ttl	التلاعب بهذه القيم وأنماطها غير الاعتيادية تعتبر مؤشر قوي على الهجمات
• ميزات الحجم والمعدل	dload,rate,sbytes,dpkts	قدرة هذه الميزات على التقاط الانحرافات في الحجم والمعدل تجعلها أساسية لتحديد الأنشطة المشبوهة التي تعتمد على أنماط تدفق بيانات غير طبيعية.
• ميزات حالة الاتصال	state, proto	تكمن أهمية هذه الميزات في كشف الانحراف عن سلوك البروتوكول الطبيعي ودورة حياة الاتصال

لدراسة أداء خوارزميات اختيار الميزات المستخدمة ضمن البحث ومقارنة أداءها وتحديد الأفضل بينها في التأثير على دقة نموذج التصنيف المصمم قمنا بتوليد مصفوفة الارتباك وقياس معايير التقييم بناء على قيمها كما هو موضح في الجدول (8) الذي يتضمن قيم التقييم لخوارزمية التصنيف DCA مع الميزات الأصلية كلها بدون تطبيق أي خوارزمية لاختيار الميزات، ومقارنتها مع نتائج التصنيف لكن مع خوارزميات اختيار الميزات الأربعة:

الجدول (9): قيم معايير التقييم الناتجة عن تطبيق نموذج الكشف المقترح على مجموعة بيانات الاختبار

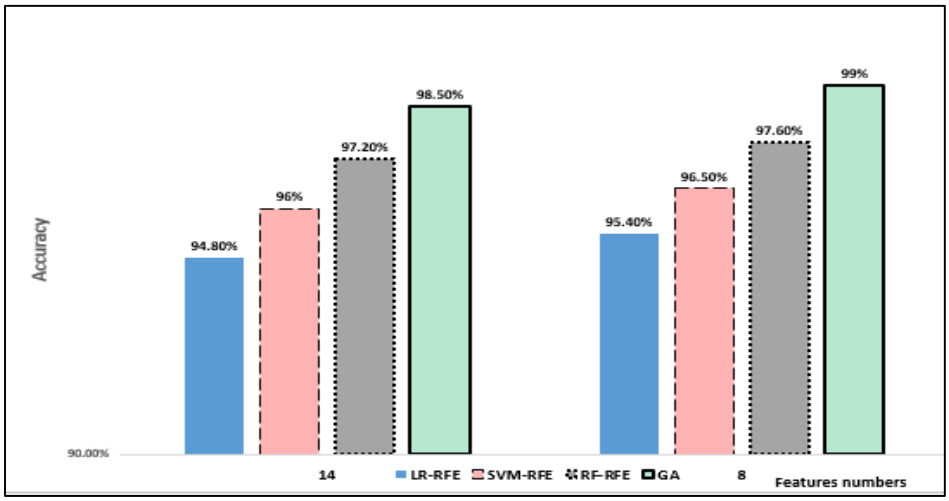
Evaluation Matrices						
Features-Selection Algorithm	Features numbers	Accuracy	precision	Recall	F1-Score	FAR
-----	42	92%	92.51%	93%	92.75%	9.22%
LR_RFE	14	94.8%	95.9%	94.54	95.25%	4.88%
	8	95.4%	96.6%	95%	95.78%	4.12%
RF_RFE	14	97.2%	96.97%	98%	97.48%	3.75%
	8	97.6%	97.20%	98.42%	97.80%	3.47%
SVM_RFE	14	96%	95.79%	95.79%	96.39%	5.22%
	8	96.5%	96.47%	97.20%	96.82%	4.35%
GA	14	98.5%	98.29%	99%	98.64%	2.11%
	8	99%	98.69%	99.5%	99.09%	1.61%

-ظهر النتائج أنّ أداء المصنف عند استخدام جميع الميزات يظهر معدل إنذارات كاذبة FAR مرتفع نسبياً مما يؤثر بشكل ملحوظ على أداء خوارزمية DCA في تصنيف حركة البيانات.

-مع تطبيق تقنيات استخلاص الميزات سواء المعتمدة على التعلم الآلي ML أو الخوارزمية الجينية نلاحظ تحسن كبير في معايير الدقة والأداء مع انخفاض ملحوظ في FAR مما يشير إلى أن إزالة الميزات الأقل ارتباطا بالتصنيف أو المكررة تؤدي إلى نماذج أكثر وضوحاً ودقة.

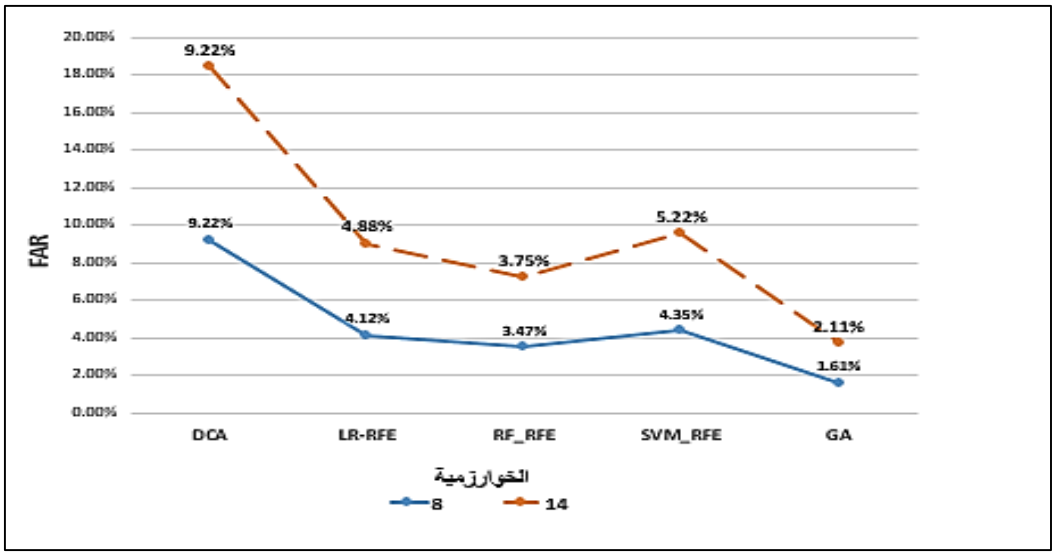
-إن تحسن قيم الأداء مع تقليل عدد الميزات من 14إلى 8 يدل إلى أن خوارزميات اختيار الميزات تعمل بشكل ناجح في تحديد مجموعة فرعية أكثر تركيزاً من الميزات الأساسية التي ترتبط بالمتغير الهدف وتقلل الضجيج ضمن مجموعة البيانات.

-نلاحظ من المخطط البياني الموضح في الشكل (7) أنّ خوارزمية RF-RFE تحقق أعلى قيم دقة وصل إلى 97.6% من أجل 8 ميزات من بين خوارزميات التعلم الآلي (LR,SVM) المطبقة مع خوارزمية RFE مع قيمة FAR منخفض للغاية وصل إلى 3.47% والسبب في ذلك يعود إلى أن خوارزمية RF تعتمد على نموذج شجري قادر على التقاط العلاقات غير الخطية والتفاعلات المعقدة بين الميزات على عكس النماذج الخطية البسيطة في SVM,LR.



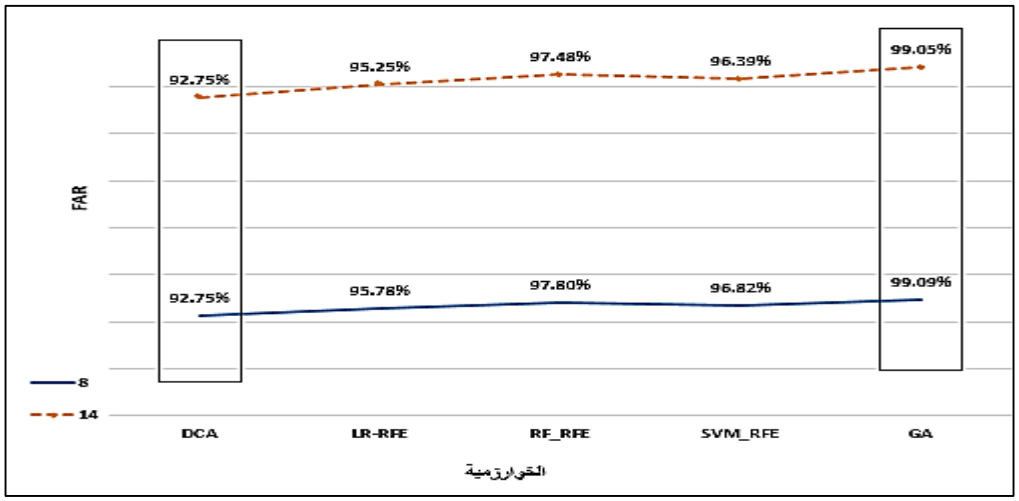
الشكل (7): مقارنة دقة نموذج التصنيف مع خوارزميات اختيار الميزات الأربعة

- نلاحظ عند تطبيق خوارزمية DCA مع خوارزمية SVM-RFE من أجل 8 ميزات انخفاض ملحوظ في قيمة معدل الإنذارات الكاذبة FAR وذلك بسبب قيم الأوزان السالبة الكبيرة للميزات مثل (sloss,dloss,smean,dtcpb) والتي تعتبر مهمة جداً لتمييز السلوك الطبيعي بدقة.
- تظهر خوارزمية التصنيف المناعية DCA أفضل أداء مع الخوارزمية الجينية GA مع أفضل قيم لمعايير التقييم وذلك يعود لاستخدام دالة لياقة فعالة في تقييم الحلول ضمن مجموعة البيانات والعثور على المجموعة المثلى التي تحقق التوازن الأعلى بين الارتباط بالهدف والاستقلالية المتبادلة بين الميزات، حيث وصلت قيم اللياقة له إلى 82.22%.



الشكل (8): معدل الإنذارات الكاذبة FAR لنتائج التصنيف مع الخوارزميات الأربعة

نلاحظ ارتفاع كبير بقيمة F-score لجميع خوارزميات اختيار الميزات مقارنة بالخط الأساسي (من 92.75% إلى قيم أعلى من 99%) كما هو موضح في الشكل (9) والذي يقدم مؤشر رئيسي لأداء النموذج في تحديد وكشف الهجمات مع ضمان التوازن بين دقة التصنيف (precision) وحساسيته (recall) مما يؤكد أنّ عملية اختيار الميزات أمر بالغ الأهمية لتحسين أداء تصنيف الهجمات حيث تعمل على التقليل من الضوضاء والتكرار في البيانات.



الشكل (9): قيم F-Score لنتائج التصنيف مع الخوارزميات الأربعة

تم مقارنة نتائج بحثنا مع نتائج مجموعة من الأوراق البحثية في مجال كشف التسلل وتصنيف الهجمات في الشبكات بالاعتماد على مجموعة من تقنيات الذكاء الصنعي والتعلم الآلي في كل من مرحلتي اختيار الميزات والتصنيف كما هو موضح في الجدول (10)، حيث يُظهر تفوقاً ملحوظاً لخوارزمية DCA في التصنيف مع أغلب تقنيات اختيار الميزات المستخدمة ضمن البحث مقارنة مع تقنيات التعلم الآلي في الدراسات الأخرى محققة الأداء الأعلى مع الخوارزمية الجينية والتي أظهرت فعاليتها قدرتها العالية في البحث عن الميزات المثلى ضمن مجموعة البيانات.

الجدول (10): مقارنة نتائج التصنيف للبحث مع نتائج أوراق بحثية في مجال كشف وتصنيف الهجمات في الشبكات

Reference	Dataset	Feature selection Algorithm	Classification Algorithm	Accuracy
[5]	UNSW-NB15	PSO	J48	89,013%
			SVM	89.152%
		GWO	J48	85.676%
			SVM	84.485%
		FFW	J48	86.037%
			SVM	85.429%
		GA	J48	86.864%
			SVM	86.397%
[6]	UNSW-NB15	IGRF-RFE	MLP	84.24%
[17]	UNSW-NB15	TS-RF	RF	83.12%
[18]	NSL-KDD	GA-SUS	XG boost and gradient boost	97.60%
Proposed Research	UNSW-NB15	LR_RFE	DCA	95.4%
		RF_RFE		97.6%
		SVM_RFE		96.5%
		GA		99%

٢ الاســـتنتاجات والتوصيـــات:

مع الاستمرار في تزايد حجم مجموعات البيانات والتعقيد الحسابي والزمني المرتبط بعملية معالجتها، أصبح من الضروري العمل على اختيار مجموعات فرعية مثلى من الميزات والتي تضم الحصول على أفضل

أداء في مهام التنبؤ والتصنيف وخاصة في مجال الأمن السيبراني، فقد تؤدي البيانات ذات الأبعاد العالية إلى انخفاض دقة النماذج الذكية وتشتت تركيزها.

قدمنا خلال هذا البحث دراسة لمجموعة من تقنيات اختيار الميزات المعتمدة على التعلم الآلي والحوسبة التطويرية، وتحليل أداءها مع مجموعة البيانات UNSW_NB15 المكونة مع كمية ضخمة من الميزات التي وصلت إلى 45 ميزة المتعلقة بحركة البيانات والاتصال في الشبكات، وذلك من خلال استخدام المجموعات المختارة من كل خوارزمية كدخل لخوارزمية الخلايا الجذعية المناعية DCA لبناء نموذج لكشف الهجمات في الشبكات.

أظهرت النتائج قدرة خوارزميات اختيار الميزات على تقليل أبعاد مجموعة البيانات بشكل فعال مما انعكس إيجاباً على أداء نموذج التصنيف ودقته، حيث تفوقت خوارزمية الغابة العشوائية RF-RFE على نظيراتها من خوارزميات التعلم الآلي (SVM-LR) بدقة وصلت إلى 97.6% مع مجموعة فرعية من 8 ميزات بينما أظهرت الخوارزمية الجينية GA نتائج ممتازة مع خوارزمية التصنيف DCA بوصفها نموذج حوسبة تطويرية بدقة تصنيف مرتفعة جداً وصلت إلى 99% وبمعدل إنذارات خاطئة منخفض جداً وصل إلى 1.6% من أجل (8)ميزات مختارة، وذلك بفضل دالة اللياقة المقترحة لتقييم المجموعات الفرعية من قبل الخوارزمية.

وبالنتيجة يقترح البحث التوصيات التالية:

- ❖ تطبيق تقنيات استخلاص الميزات مثل خوارزمية PCA التي تعمل على توليد ميزات جديدة مستخلصة من الميزات الأساسية ومقارنتها مع تقنيات استخراج الميزات الطبقة في هذا البحث.
- ❖ دراسة تأثير خوارزمية تصنيف الهجمات المطبقة في أداء تقنيات اختيار الميزات وذلك من خلال إجراء مقارنة بين نتائج التقييم لمجموعة من خوارزميات الكشف والتصنيف الذكية المعتمدة على نفس خوارزميات اختيار الميزات.
- ❖ دراسة وتحليل أداء تقنيات استخلاص الميزات المطبقة في هذا البحث مع مجموعات بيانات أخرى مثل NSL-KDD ومقارنة نتائجها مع نتائج بحثنا.

المراجع:

[1] Gupta, B. B., Perez, G. M., Agrawal, D. P., & Gupta, D. (2020). *Handbook of computer networks and cyber security.* Springer, 10, 978-3.

[2] Theng, D., & Bhoyar, K. K. (2024). *Feature selection techniques for machine learning: a survey of more than two decades of research*. Knowledge and Information Systems, 66(3), 1575-1637.

[3] Chelly, Z., & Elouedi, Z. (2016*). A survey of the dendritic cell algorithm*. Knowledge and Information Systems, 48, 505-535.

[4] Hassan, M., & Kaabouch, N. (2024). *Impact of feature selection techniques on the performance of machine learning models for depression detection using EEG dat*a. Applied Sciences, 14(22), 10532.

[5] Almomani, O. (2020*). A feature selection model for network intrusion detection system based on PSO*, GWO, FFA and GA algorithms. Symmetry, 12(6), 1046.

[6] Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., & Kwak, J. (2023*). IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset*. Journal of Big data, 10(1), 15.

[7] Chapagain, P., Timalsina, A., Bhandari, M., & Chitrakar, R. (2022, January). *Intrusion detection based on PCA with improved K-means*. In International conference on electrical and electronics engineering (pp. 13-27). Singapore: Springer Singapore.

[8] Kanimozhi, V., & Jacob, P. (2019). *UNSW-NB15 dataset feature selection and network intrusion detection using deep learning*. International Journal of Recent Technology and Engineering, 7(5), 443-446.

[9]Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017*). Feature selection: A data perspective*. ACM computing surveys (CSUR), 50(6), 1-45.

[10] Jeon, H., & Oh, S. (2020). *Hybrid-recursive feature elimination for efficient feature selection*. Applied Sciences, 10(9), 3211.

[11] Mathew, T. E. (2019). *A logistic regression with recursive feature elimination model for breast cancer diagnosis*. International Journal on Emerging Technologies, 10(3), 55-63.

[12] Azman, N. S., Samah, A. A., Lin, J. T., Majid, H. A., Shah, Z. A., Wen, N. H., & Howe, C. W. (2023). *Support vector machine–Recursive feature elimination for feature selection on multi-omics lung cancer data*. Progress In Microbes & Molecular Biology, 6(1).

[13] Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018*). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data.* BMC genetics, 19(Suppl 1), 65.

[14] Xia, S., & Yang, Y. (2023*). A model-free feature selection technique of feature screening and random forest-based recursive feature elimination*. International Journal of Intelligent Systems, 2023(1), 2400194.

[15]Sohail, A. (2023). *Genetic algorithms in the fields of artificial intelligence and data sciences.* Annals of Data Science, 10(4), 1007-1018.

[16]Dagdia, Z. C. (2019*). A scalable and distributed dendritic cell algorithm for big data classification*. Swarm and Evolutionary Computation, 50, 100432.