Detection of Network Anomalies Based on Hybrid Artificial Intelligence Techniques

Hassan SALEH*

(Received 22 / 7 / 2019. Accepted 2 / 9 / 2019)

□ ABSTRACT □

Artificial Intelligence could make the use of Intrusion Detection Systems a lot easier than it is today. As always, the hardest thing with learning Artificial Intelligence systems is to make them learn the right things. This research focuses on finding out how to make an Intrusion Detection Systems environment learn the preferences and work in a correct way, In this research hybrid intelligence system is designed and developed for network intrusion detection, where the research was presented four methods for network anomaly detection using clustering technology and dependence on artificial intelligence techniques, which include a Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to develop and improve the performance of intrusion detection system. The first method implemented by applying traditional clustering algorithm of KM in a way Kmeans on KDDcup99 data to detect attacks, in the way the second hybrid clustering algorithm HCA method was used where the Kmeans been hybridized with GA. In the third method PSO has been used. Depending on the third method the fourth method Modified PSO (MPSO) has been developed, This was the best method among the four methods used in this research.

Keywords: Artificial Intelligence, Clustering, Intrusion Detection, Hybrid, Genetic Algorithm

* **Master Degree - Tele-Informatics Engineering Faculty of Mechanical and Electrical Engineering, Tishreen University, Latakia, Syria. Email: eng.hassan.s.saleh@gmail.com**

كشف الشذوذ الشبكي بالاعتماد على تقنيات الذكاء الصناعي الهجين

حسان صالح*

(تاريخ الإيداع 22 / 7 / 2019. قُبِل للنشر في 2/ 9 / 2019)

□ ملخّص □

يمكن أن تجعل تقنيات الذكاء الصناعي أنظمة كشف التطفل أسهل بكثير مها عليه اليوم وكما هو الحال دائما فإن أصعب شيء في تعلم الأنظمة المصممة بتقنيات الذكاء الصنعي عملية تدريبها لتعلم الأمور الصحيحة. يركز هذا البحث على عمل بيئة لأنظمة كشف التطفل وتعليمها ممارسة العمل بصورة صحيحة. صمم في هذا البحث نظام ذكاء صنعي مهجن ومطور لكشف التطفل الشبكي، اذ قدم البحث أربعة طرائق كشف الشذوذ الشبكي باستخدام تقنية العنقدة والاعتماد على تقنيات الذكاء الصناعي التي تتضمن الخوارزمية الجينية وخوارزمية سرب الطيور لتطوير وتحسين أداء نظام كشف التطفل. نفذت الطريقة الأولى بتطبيق خوارزمية العنقدة التقليدية KM المتمثلة بطريقة Kmeans على بيانات KDDcup99 لكشف الهجمات ،واستخدمت في الطريقة الثانية HCA طريقة العنقدة المهجنة اذ تم تهجين خوارزمية Kmeans مع الخوارزمية الجينية. أما في الطريقة الثالثة فقد تم استخدام خوارزمية سرب الطيور PSO .بالاعتماد على الطريقة الثالثة أنشأت الطريقة الرابعة وهي خوارزمية سرب الطيور المطورة MPSO وكانت هذه الطريقة الأفضل من بين الطرائق الأربعة المستخدمة في هذا البحث.

الكلمات المفتاحية: ذكاء صنعي، عنقدة، كشف التطفل، هجين، خوارزمية جينية.

* **ماجستير – هندسة الاتصالات المعلوماتية – كلية الهندسة الميكانيكية والكهربائية – جامعة تشرين – اللاذقية – سورية.**

مقدمة:

أصبحت أنظمة كشف التطفل IDS في السنوات الأخيرة واحدة من المناطق الأكثر سخونة في البحوث الخاصة بأمن الحاسوب. وهي تكنولوجيا كشف مهمة تستخدم بوصفها اجراء مضادا للحفاظ على سلامه البيانات واستمرارية عمل النظام خلال عملية الاقتحام [1].

يتيح كشف التطفل رصد وتحليل نشاط المستخدم والنظام، وتدقيق تكوينات النظام ونقاط الضعف، وتقييم سلامة النظام وملفات البيانات، والتحليل الإحصائي لنشاط النماذج Patterns على أساس المطابقة لهجمات معروفة، وتحليل النشاطات الشاذة، وتشغيل نظام المراجعة [2]. لكشف التطفل أسلوبان:

1- الكشف عن الشذوذ: يشير الى التقنيات التي تحدد وتميز السلوك العادي أوالمقبول للمنظام على سبيل المثال، استخدام وحدة المعالجة المركزية، وقت التنفيذ. أوالتصرفات التي تحيد عن السلوك العادي المتوقع تعدها اختراقات.

2- الكشف عن اساءة الاستخدام: يشير الى التقنيات التي تميز طرائق معروفة لاختراق النظام. تتميز هذه الاختراقات بأنها نماذج Patterns أو توقيع Signature مخزنة في قاعدة المعرفة للنظام لذلك يقوم نظام كشف التطفل بالبحث عن النماذج والتواقيع المشابهة لها ويعدها اختراقات. قد تكون النماذج أوالتواقيع جملة ثابتة أو تسلسل مجموعة من الإجراءات، وتستند استجابات النظام على الاختراقات التي تم تحديدها ويمثل الشكل (1) البنية التحتية لأنظمة كشف التطفل [3].

الشكل (1) البنية التحتية لأنظمة كشف التطفل.

يقسم نظام كشف التطفل من ناحية مجال العمل والحماية بصورة عامة الى قسمين:

1) نظام كشف التطفل المستند على المضيف Host Based Intrusion DetectionSystem

ويرمز له HIDS يقوم هذا النوع من الأنظمة بتحليل الأحداث الموجودة في جهاز الحاسوب ويقوم بتمييز الأحداث والفعاليات الخاصة بالمستخدمين الذين يقومون بنشاطات عدائية ومضرة بنظام التشغيل. يقوم النظام بمراقبة استخدامات المضيف واقتفاء آثاره ويأخذ فعالياتهم ويعتبرها كإدخال لنظام كشف التطفل لاكتشاف نوعية فعالياتهم. ويقوم كذلك بمراقبة ملفات النظام الرئيسية والتنفيذية من خلال مراقبة التواقيع في فترات منتظمة لكي يكشف التغيرات الحاصلة عليها جراء الهجمات غير المتوقعة [4]،[5].

2) نظام كشف التطفل المستند على الشبكة Network Based Intrusion DetectionSystem

ويرمز له NIDS، لا يقوم هذا النوع من أنظمة كشف التطفل باختبار وفحص سجلات تدقيق الأثر الخاصة بالمضيف، ولكن يقوم بعمل تسجيل وتحليل لفعاليات الشبكة بتحليل البيانات الموجودة في حزم الشبكة من مختلف أجزاء الشبكة ثم يقوم بإصدار تقارير إلى وحدة الإدارة المركزية للنظام عن حالة سريان البيانات المراقبة من قبله ويثبت هذا النوع من الأنظمة في جهاز الموجه Router أواجهزة الشبكة الأخرى [4]،]5]. أما من ناحية السلوك فتقسم أنظمة كشف التطفل الى نوعين ، نظام كشف التطفل الخامل Passive الذي يقوم بكشف التطفل وارسال التنبيه فقط ، ونظام كشف التطفل الفعال Active الذي يقوم بكشف التطفل وأيضا استخدام استخدام تقنيات خاصة لمواجهتها وحماية النظام [4]،]5]. ومن ناحية تحليل البيانات تقسم أنظمة كشف التطفل إلى نوعين النوع الأول يكون نظام معتمد على الزمن الحقيقي حيث تتم مراقبة البيانات وتحليلها وارسال التنبيه مباشره في حالة اكتشاف الهجمات لكي تؤخذ اجراءات فورية لمواجهتها ويسمى هذا النوع line Analysis–on .أما النوع الثاني يقوم يأخذ مقطع من النظام وتقييم الوضع الأمني إذ يقوم بإجراء تحليل أكثر شمولية من النوع الأول دون أن يكون له تأثير غير مقبول على أداء الأنظمة التي يتم مراقبتها ويسمى هذا النوع .[4] Off0–line Analysis

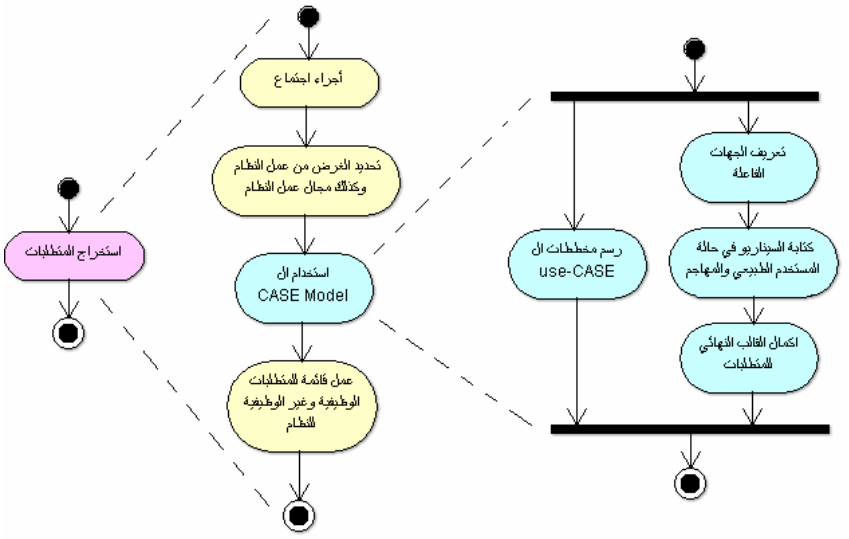
1. الدراسات السابقة:

دُرست مشكلة كشف التطفل وتصنيفه ضمن حقل أمان شبكات الحاسوب ففي عام 2012 قدم الباحثان الأمريكيانSerpen و Sabhnani نظامهما لكشف التطفل بتطبيق التقنيات الذكية على بيانات ال— KDD99 صف التطفل بصيغة اساءة الاستخدام. عمد الباحثان على تطبيق أنظمتهم وبمختلف التقنيات الذكائية على هذه البيانات وكانت تقنياتهم تتضمن عدة حقول مثل الشبكات العصبية الاصطناعية والمنطق الضبابي والنماذج الإحصائية والنماذج الاحتمالية وأشجار القرار، وقد تم تطبيق تسعة خوارزميات من مختلف الحقول المذكورة وبناء نموذج متعدد التصنيف. وقد جاءت نتائج التطبيق متفاوتة بين الخوارزميات مثلا نتائج خوارزمية MLP كانت نسبة كشفها لصنف الهجوم Probe أفضل من كل النتائج وكانت خوارزمية K–means الأفضل بنتائج الكشف الخاصة بصنفي الهجوم U2R و DOSوخوارزمية Gaussian كانت نسب كشفها لصنف الهجوم R2L الأفضل. وقد تم الاستنتاج بأن لكل خوارزمية القدرة على كشف صنف معين من أصناف الهجوم أكثر من غيرها.

في عام 2015 استخدم الباحثون .Kien A , .Annie S ,.Bing C ,.Gary S الخوارزمية الجينية لاختيار مجموعة جزئية من الخصائص الداخلة إلى المصنف الذي بني باستخدام خوارزمية أشجار القرار C4.5 لزيادة نسبة كشف التطفل وتقليل نسبة الإنذارات الكاذبة وقد تم استخدام مجموعة بيانات KDD99في تدريب النظام واختباره وأظهرت النتائج أن أداء خوارزمية أشجار القرار أصبح أفضل بعد دمجها مع الخوارزمية الجينية. وأيضا في عام 2015 استخدم الباحث .Omran M خوارزمية سرب الطيور Particle Swarm Optimization PSO المعتمدة على خوارزميات العنقدة التقليدية في تصنيف الصور بدون معلم وأظهرت النتائج أن أداء الخوارزمية الجديدة أفضل بكثير من أداء خوارزميات العنقدة التقليدية المتمثلة ب Kmeans. في عام 2016 أقترح الباحث.y Liu أسلوبا جديدا لكشف الشذوذ الشبكي عن طريق ربط شبكة دالة القاعدة الشعاعية Radial Basis Function RBF مع خوارزمية سرب الطيور Particle Swarm Optimization PSO، وقد تم استخدام مجموعة بيانات KDD99 في النظام وأظهرت النتائج أن أداء شبكة دالة القاعدة الشعاعية أصبح أفضل بعد دمجها مع خوارزمية سرب الطيور.

1. وصف المتطلبات

تلعب المتطلبات دورا أساسيا خلال عملية انشاء المنتج. يتم توثيق المتطلبات في وثيقة تسمى وثيقة وصف المتطلبات. يعد استخراج المتطلبات Requirements elicitation نشاطاً مهماً في عملة هندسة البرمجيات يهدف إلى اكتساب وفهم المتطلبات. يمثل الشكل (2) مخطط استخراج متطلبات النظام.



الشكل (2) مخطط Uml Activity لاستخراج المتطلبات

أهمية البحث وأهدافه:

تخضع نظم وشبكات المعلومات دائماً للهجمات الإلكترونية، ومحاولات اختراق أمن المعلومات تتزايد كل يوم، جنباً إلى جنب مع توافر أدوات تقييم الاختراقات التي تتوفر على نطاق واسع على شبكة الانترنت مجانا [2]، لا يعد الجدار الناري Firewall النظام الديناميكي ذوالقدرة الدفاعية المطلوبة الذي يمكن أن يتخيله المستخدم. ولكن في المقابل توجد أنظمة كشف التطفل IDS التي تعد أنظمة ديناميكية ذات قدرة دفاعية عالية، وتتعرف أنظمة كشف التطفل على الهجمات التي لا تستطيع الجدران النارية التعرف عليها [6].

الهدف من البحث

سيقوم البحث باستخدام تقنيات الذكاء الاصطناعي لتحليل البيانات واكتشاف المتطفل إذا حاول الهجوم، إذ يقوم النظام بتوليد تنبيه يتم عن طريقه إخطار مسؤول النظام لاتخاذ إجراءات معينة إذا لزم الأمر، لاتخاذ اجراءات معينة اذا لزم الأمر.

مجال البحث

يشمل مجال عمل أنظمة كشف التطفل أنواع وكميات كثيرة من النظم التي يتعين دعمها، إذ يمكن أن يكون النظام المحمي آلة فردية أو شبكة من الأجهزة [7].

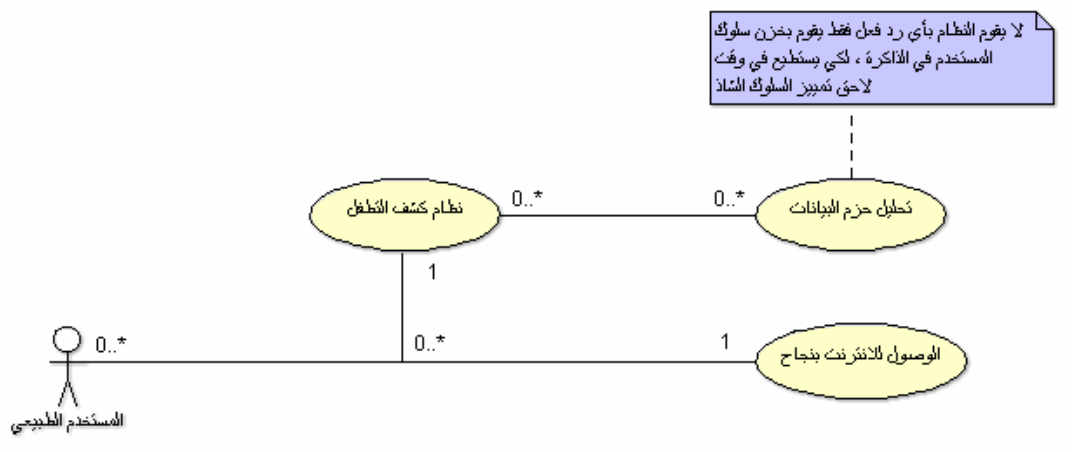
استخدام مخطط هندسة البرمجيات بمساعدة الحاسوب Use-Case Model

تستخدم هذه المخططات في هندسة البرمجيات لاستخراج المتطلبات الوظيفية للنظام. وتقوم أيضا بتعريف التفاعل بين المستخدمين Actors والنظام [8]. تم استخدام ArgoUML لتطبيق ال Use-Case على النظام المصمم في هذا البحث. وArgoUML هي أداة تصميم قوية سهلة الاستخدام، تدعم تصميم البرمجيات الرسومية وتطوير وتوثيق

تطبيقات البرمجيات [9] ، توجد ثلاث جهات فاعلة *Actors* أساسية في النظام IDS وهي مسؤول النظام والمستخدم الطبيعي للنظام والمهاجم. يعمل نظام كشف التطفل مع سيناريوهين :

مستخدم طبيعي يستخدم النظام

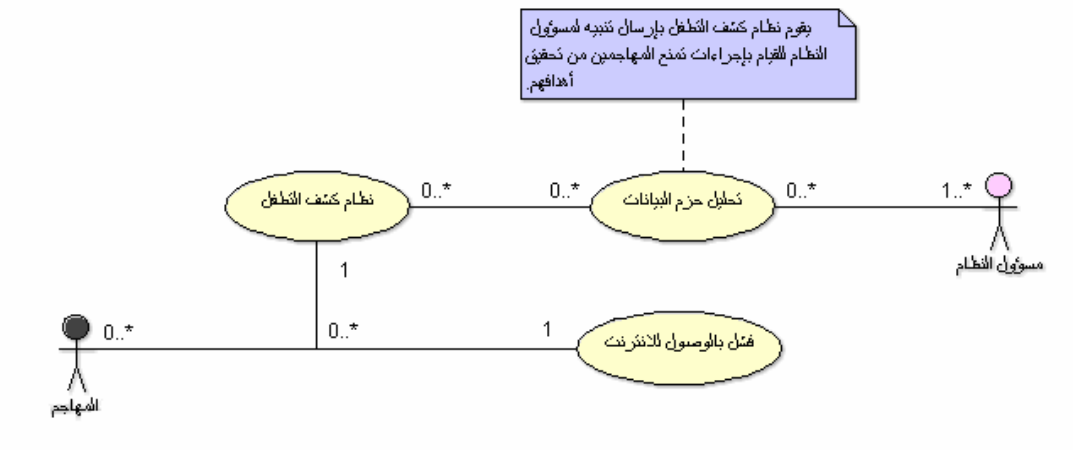
سوف يقوم النظام خلال الاستخدام العادي برصد كل حركات المرور في الجهاز أوالشبكة المحمية. يجب أن يبقى النظام هادئاً في حالة النشاط العادي ولا يرد على أي شيء طبيعي، ويوفر الخدمات بصورة كاملة. يقوم النظام بتحليل سلوك المستخدم ويبقيه في الذاكرة، لكي يستطيع في وقت لاحق تمييز السلوك الشاذ كما في الشكل (3). الجهة الفاعلة في هذه الحالة هي المستخدم الطبيعي للنظام.



الشكل (3) مخطط UML use case للمستخدم الطبيعي للنظام.

شخص يحاول مهاجمة النظام

يميز نظام كشف التطفل في مرحلة معينة هجمة معينة أو إساءة استخدام، لذلك يقوم بتوليد إنذار لإخطار مدير النظام الذي يقوم بدوره ببعض الإجراءات الوقائية كما في الشكل(4) . الجهات الفاعلة في هذه الحالة هي مسؤول النظام والمهاجم.



الشكل (4) مخطط UML use case لشخص يحاول مهاجمة النظام

المتطلبات الوظيفية

- يجب أن يكون النظام قادرا على قراءة الحزم Packets من مجموعة متنوعة من المصادر.
- على المستخدم أن يكون قادرا على تدريب النظام.
- مسؤول النظام له صلاحيات تشغيل وايقاف النظام.
- النظام قادر على تعلم الفرق بين الحزمة الطبيعية وغير الطبيعية وتصنيف الهجمات لأنواعها.
- النظام قادر على إعلام مسؤول النظام عدما يواجه سلسلة من الأحداث التي من المرجح أن تكون هجوما
- جمع الإحصاءات من الحزم وخزنها في قاعدة المعرفة ليكون النظام قادرا على التحليل والمراجعة.

المتطلبات غير الوظيفية

- بجب أن يكون النظام سهلاً لتدريب المستخدمين الجدد على هذا البرنامج.
- على المستخدم أن يكون قادراً على تشغيل وايقاف النظام بسهولة.
- ضبط محددات النظام والتحكم فيها بسهولة.
- الموثوقية Reliability: لا يحتاج مسؤولو النظام لإعادة بدء تشغيل النظام في أي لحظة إلا لأسباب الصيانة الضرورية.
- الأداء Performance: يجب أن يتم التقاط الحزم بسرعة كافية لتمكين تحليل البيانات خلال وقت قصير ومن ثم اخطار مسؤول النظام في أقرب وقت ممكن.
- الدعم Supportability: يسمح التكوين الحيوي للنظام لمسؤولي النظام بإضافة وتغيير الأوزان والمحددات دون التأثير على خدمات النظام، ودون تعطيل أي من الخدمات الأخرى على شبكة الاتصال.

تقييم أداء أنظمة كشف التطفل

إن مقاييس كشف التطفل مهمة جدا لتقييم الأداء الفني لنظام كشف التطفل، فمحللو أمان النظام يقومون بالاطلاع على مخرجات أنظمة كشف التطفل لكي يتعرفوا على إمكانية حدوث الهجوم ومتى يتم اصدار الإنذارات. بالإضافة الى ذلك، يحتاج مسؤول النظام الى امتلاك القدرة على المقارنة بين نقاط القوة والضعف الموجودة في أنظمة كشف التطفل الحالية واختيار الملائم منها. تعد نسبة الكشف Detection Rate من المعايير المهمة لقياس أداء نظم كشف التطفل ويتم قياسها حسب المعادلة التالية [1]

$$DR = \frac{TP}{TP + TN} * 100 \qquad (1)$$

حيث TP هو معيار لقياس عدد سجلات الهجوم التي يتم تصنيفها بصورة صحيحة، وTN هو معيار لقياس عدد السجلات الشرعية التي يتم تصنيفها صورة صحيحة [11,12]. وهناك أيضا الإنذارات الكاذبة الايجابية FP وتمثل النسبة لعدد سجلات الاتصال الشرعة والتي يتم تصنيفها خطأ على أنها سجلات هجوم. والإنذارات الكاذبة السلبية FN وفيها يتم تصنيف سجلات الهجوم بصورة خاطئة على أنهم سجلات اتصال شرعية[13]

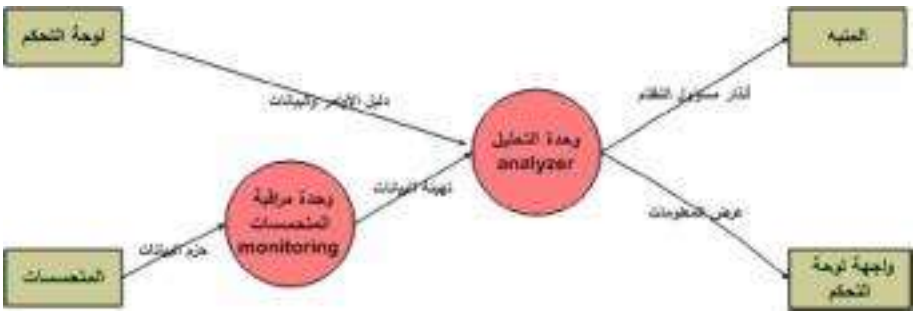
2. تحليل و تصميم النظام

تضمن هذا البحث محورين أساسيين من محاور الذكاء الاصطناعي لكشف التطفل المِحور الأول يتمثل بالخوارزمية الجينية Algorithm Genetic والمحور الثاني يتمثل بخوارزمية سرب الطيور swarm particleptimization او اذ

استخدمت في عمية كشف الهجمات أي كشف الشذوذ Anomaly detection وكذلك خوارزمية سرب الطيور المطورة modified particle swarm optimization التي استخدمت للغرض نفسه ولكن بنتائج أفضل وفي الشكل رقم (5) مخطط تحليل سير البيانات في نظام كشف التطفل الذي أعتمد في هذا البحث.

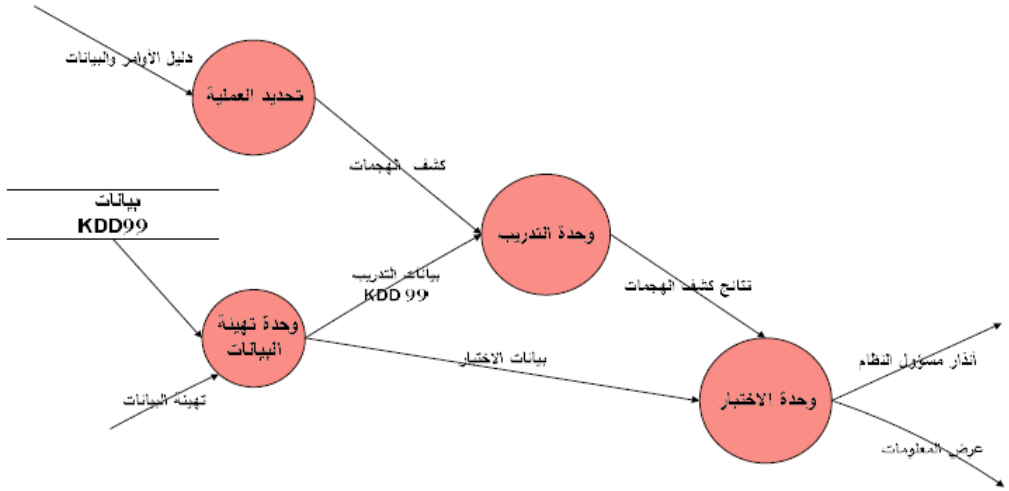
الشكل (5). المرحلة رقم صفر لمخطط سير البيانات في النظام

ثم يتم تحليل نظام كشف التطفل الى وحدتين وحدة خاصة بمراقبة المتحسسات ووحدة التحليل كما في الشكل رقم (6).



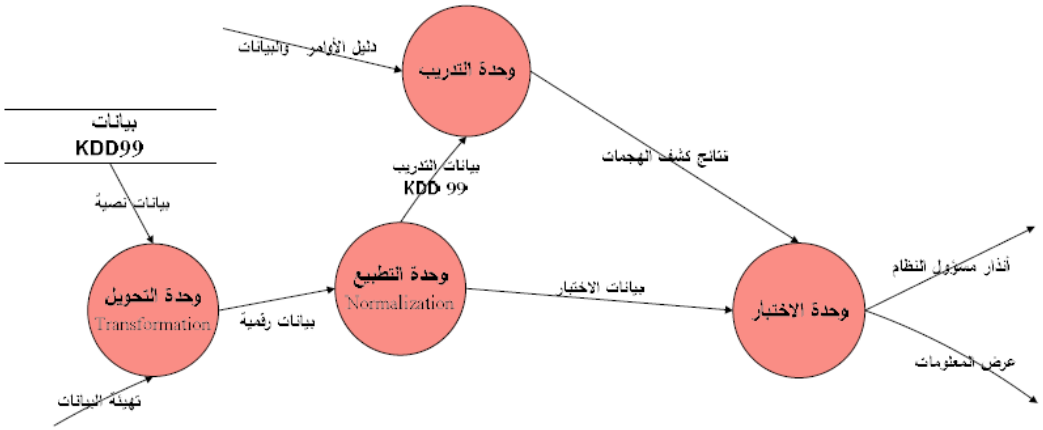
الشكل (6). المرحلة رقم 1 لمخطط سير البيانات في النظام

تمثل وحدة التحليل قلب النظام حيث تقسم إلى وحدة التدريب التي تحتوي على أسلوب كشف التطفل المعتمد في هذا البحث مثلا الخوارزمية الجينية أو خوارزمية سرب الطيور إذ يتم تدريب وتعليم النظام على التمييز بين حزم البيانات الطبيعية وغير الطبيعية ثم إرسال الحل الأفضل إلى وحدة الاختبار كما في الشكل رقم (7).



الشكل (7). المرحلة رقم 2 لمخطط سير البيانات في وحدة التحليل Analyzer

أما البيانات المستخدمة في عملية التدريب والاختبار هي مجموعة بيانات KDD99 وهي البيانات الأكثر استخداما' في بناء نظم كشف التطفل [14]. ولكن قبل استخدام هذه البيانات يتم ادخالها إلى وحدة ما قبل معالجة البيانات وهي وحدة تهيئة البيانات الخاصة بالنظام التي تقسم بدورها الى وحدة التحويل Transformation لإنتاج بيانات ادخال تتوافق مع بيانات الإخراج للنظام اذ يتم تحويل البيانات من الشكل النصي إلى الشكل الرقمي وكذلك وحدة التطبيع للبيانات normalization التي تقوم بتحسين دقة وكفاءة الخوارزميات المستخدمة بحيث يصبح المدى المستخدم للبيانات هوالمجال [0.0,0.1] [15]. كما هو موضح في الشكل (8).



الشكل (8): المرحلة رقم 2 مخطط سير البيانات في وحدة تهيئة البيانات

1. عنقدة البيانات Data Clustering

العنقدة هي عملية تقسيم البيانات الى مجاميع اعتمادا على بعض المقاييس المتشابهة لهذه المجاميع وتعد عملية عنقدة البيانات عطية أساسية ومركزية في الذكاء الصناعي اذ يتم تعرف العنقود بواسطة مركز العنقود cluster centre والطريقة الأكثر شيوعا لإيجاد مقاييس التشابه بين البيانات ومراكز العناقيد هي المسافة الإقليدية Euclidean distance التي تقاس بالمعادلة التالية :

$$d(Z_u, Z_w) = \sqrt{\sum_{j=1}^{Nd} (z_{u,j} - z_{w,j})^2} = \left\| Z_u - Z_w \right\|_2$$

(2)

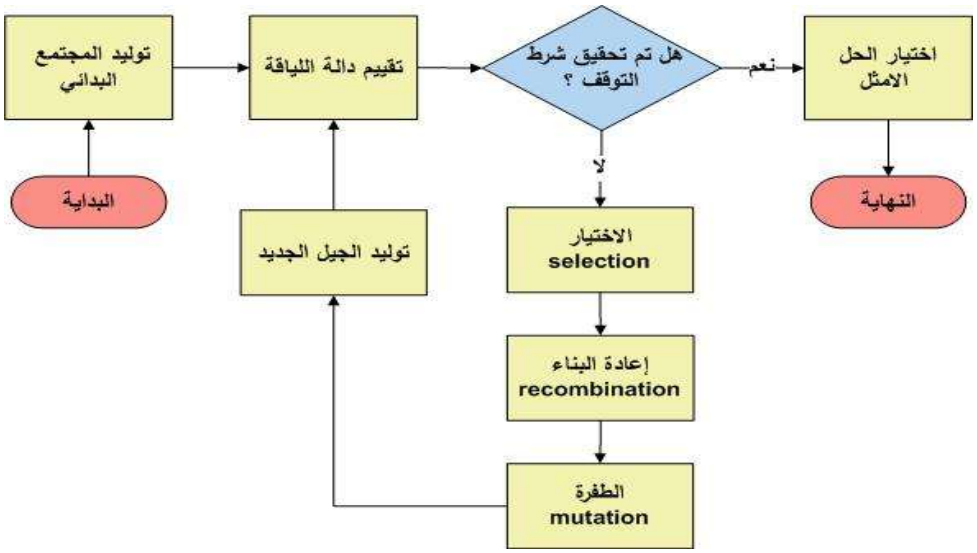
خوارزمية العنقدة الأكثر شيوعا هي خوارزمية K− −means التكرارية حيث تقوم هذه الخوارزمية بتقليل المسافة بين البيانات المتشابهة ومركز العنقود ، تبدأ هذه الخوارزمية بمراكز عناقيد عشوائية ثم تقوم بتوزيع البيانات الى العنقود الأقرب حسب دالة اللياقة الخاصة بهذه الخوارزمية [17]،[16]:

$$J_{K\text{-}means} = \sum_{k=1}^{K} \sum_{\forall z_p \in C_k} d^2(z_p, m_k)$$

(3)

2. الخوارزمية الجينية Genetic Algorithm

تعد الخوارزمية الجينية من النماذج الحسابية القائمة على مبادئ التطور والانتقاء الطبيعي. تقوم هذه الخوارزميات بنمذجة المشكلة في مجال معين باستخدام الكروموسومات وتطور هذه الكروموسومات باستخدام الاختيار Selection وإعادة التركيب recombination، وعملية الطفرة mutation. في تطبيقات أمن الحواسيب تستخدم الخوارزمية الجينية لإيجاد الحل الأمثل لمشكلة معينة. يبدأ تطوير الخوارزمية الجينية عادة مع أفراد تم اختيارها عشوائيا من الكروموسومات، هذه الكروموسومات تمثل المشكلة التي يجب حلها ويتم تغييرها بشكل عشوائي خلال التطور، يطلق على مجموعة من الكروموسومات أثناء مرحلة التطوير مجتمع population. يتم استخدام دالة التقييم لحساب نسبة صلاحية كل كروموسوم خلال التقييم، تستخدم عمليتان أساسيتان (التزاوج Crossover، الطفرة Mutation)، لمحاكاة التكاثر الطبيعي والشكل رقم (9) يوضح الهيكلية العامة للخوارزمية الجينية [1,18].



الشكل (9). الهيكلية العامة للخوارزمية الجينية

أثبتت الدراسات أن هناك مجالاً واعداً في استخدام الخوارزمية الجينية لكشف التطفل. حيث تستخدم لتقييم حزم البيانات التي تمر بالشبكة وتمييز الحزم الطبيعية عن الشاذة التي تدل على احتمالية حدوث اختراق[1].

3. خوارزمية العنقدة المهجنة Hybrid Clustering Algorithm HCA

لتحسين استخدام خوارزمية kmeans في كشف التطفل تم دمجها مع الخوارزمية الجينية حيث تم اعتبار كل حزمة من حزم بيانات KDD99 كروموسوم وكل حزمة تحتوي على 41 صفة:

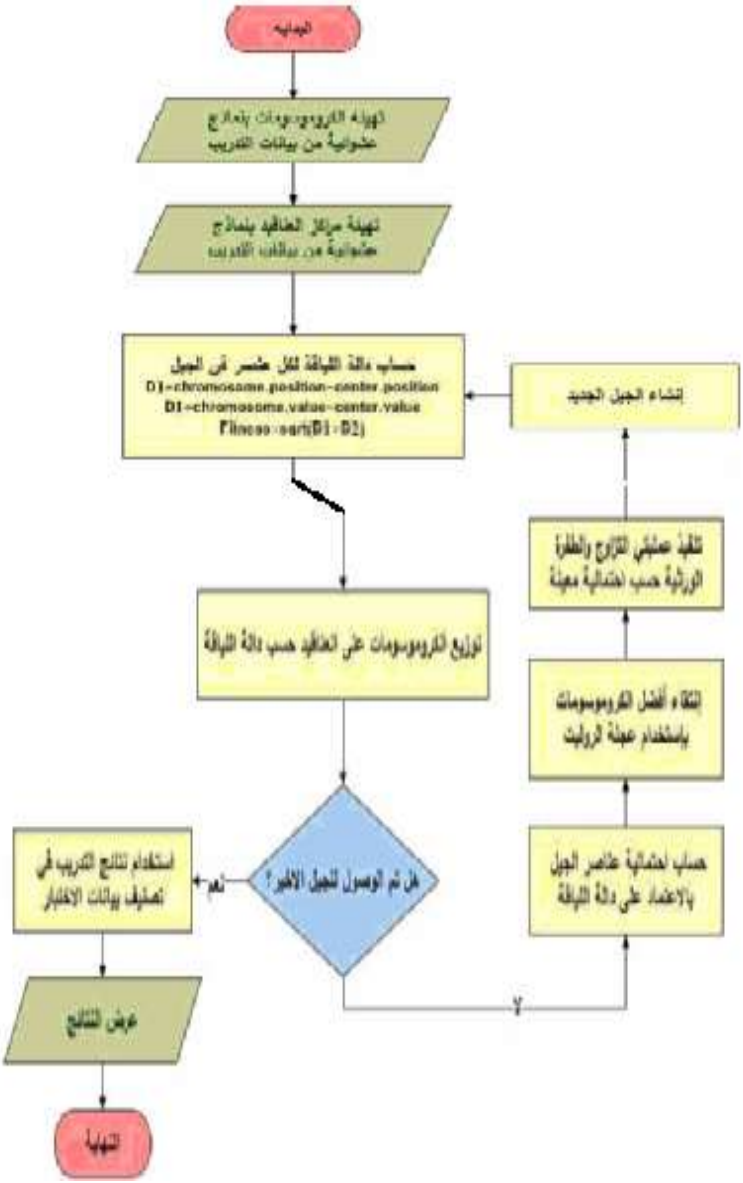
chromosome = { duration , protocol *type* , service , flag , ….etc}

حيث يتم في البداية تهيئة الجيل البدائي ومراكز العناقيد ببيانات عشوائية ثم حساب دالة اللياقة لكل عناصر الجيل البدائي وتوزيعها على مراكز العناقيد باعتماد المسافة الإقليدية حسب المعادلة رقم (2) ثم يتم اختيار أفضل الأفراد من هذا الجيل أصحاب اكبر احتمالية حسب عملية الاختيار باعتماد المصدر [19]. باعتبار p هي احتمالية النجاح و $1-p$ هي احتمالية الفشل فإن:

$$F(x)=\begin{cases}0, & if\ x<0,\\ 1-p, & if\ 0\le x<1,\\ 1, & if\ 1\le x\end{cases}$$

(4)

ويتم حساب p لكل كروموسوم عن طريق قسمة قيمة دالة اللياقة الخاصة بهذا الكروموسوم على المجموع الكلي لقيم دالة اللياقة الخاصة بالجيل . ثم يتم تطبيق عمليات التزاوج والطفرة الوراثية على الكروموسومات حسب احتمالية معينة لكي يتم تكوين الجيل التالي حيث يتم اخذ أفضل الأفراد من الجيل السابق وإهمال الأسوأ وأيضا يتم تعديل مراكز العناقيد الخاصة بكل عنقود عن طريق أخذ معدل البيانات التي تنتمي للعنقود وتستمر هذه العمليات بصورة دورية لحين تحقق شرط التوقف. والشكل رقم (10) يمثل الخوارزمية الجينية المصممة لكشف التطفل. وبعد الحصول على أفضل العناقيد يتم تصنيف البيانات على أساسها وذلك بأخذ المسافة الإقليدية بين العناقيد الناتجة من خوارزمية العنقدة المهجنة وكل نموذج من البيانات وأخذ تسلسل العنقود الذي ناتج المسافة الإقليدية بينه وبين النموذج المعين يكون أقل قيمة وهذا التسلسل يمثل صنف النموذج.



الشكل (10) المخطط الانسيابي للخوارزمية المستخدمة في نظام كشف التطفل

4. خوارزمية سرب الطيور Particle Swarm Optimization

تُعدّ الأفراد Population أساس عمل هذه الخوارزمية حيث تقوم بمحاكاة السلوك الطبيعي لأسراب الطيور في برنامج حاسوبي. يتم في البداية تهيئة الأفراد بحلول عشوائية، تسمى جسيمات Particles كل جسيم من هذه الجسيمات يرتبط بسرعة خاصة به Velocity . تطير الجسيمات في فضاء البحث بحيث يتم تعديل السرعة الخاصة بها بصورة مستمرة حسب السلوكيات الخاصة بالسرب لذلك فإن الجسيمات يكون ليها ميل لتطير نحوالحل الأفضل في فضاء البحث. كل جسيم في السرب له الخصائص التالية:

- Xi: يمثل الموقع الحالي للجسيم.
- Vl: السرعة الحالية للجسيم. ٠

- Yi: أفضل موقع اتخذه الجسيم.

لنفرض أن *f* هي دالة اللياقة و *t* هوالزمن الحالي فإن أفضل موقع اتخذه الجسيم يتم تحديثه كالآتي:

$$y_i(t+1) = \begin{cases} y_i(t) & if \quad f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) & if \quad f(x_i(t+1)) < f(y_i(t)) \end{cases}$$

(5)

ومن ثم أيجاد أفضل موقع أتخذه الجسيم بالنسبة للسرب بأكمله بواسطة المتجه *y*

$$\hat{y}(t) \in \{ y_0, y_1, \ldots, y_s \} = min \{ f(y_0(t)), f(y_1(t)), \ldots, f(y_s(t)) \}$$

(6)

حيث *s* هو حجم السرب. أما سرعة الجسيمات *Vi* يتم تعديلها باستخدام المعادلة التالية:

$$v_{i,j}(t+1) = wv_{i,j}(t) + c_1 r_{1,j}(t)(y_{i,j}(t) - x_{i,j}(t)) + c_2 r_{2,j}(t)(\hat{y}(t) - x_{i,j}(t))$$

(7)

حيث *i* تمثل الجسيم و *j={1 ,.., Nd}* و حيث Nd تمثل أبعاد المسألة و *w* هو وزن القصور الذاتي قيمته بين *{0,1}* و *c1,c2* يمثلون ثوابت التسارع. c1 المسؤول عن التحكم في زيادة ونقصان البحث المحلي أما c2 فهوالمسؤول عن الحكم في زيادة ونقصان البحث الشامل. وr1,j، r2,j أرقام عشوائية بين {0,1} ويتم تحديث موقع الجسيم *i* وهو *xi* بالمعادلة التالية:

$$x_i(t+1) = x_i(t) + v_i(t+1)$$

(8)

ويمكن استخدام مقاييس كثيرة لقياس جودة خوارزميات العنقدة، مقياس الأداء الأكثر شيوعا هو مقياس الخطأ الكمي *Je*

$$Je = \frac{\sum_{k=1}^{K} \sum \forall zp \in C_k \; d(z_p, m_k)/nk}{K}$$

(9)

حيث تمثل *Ck* العنقود k و *n* هو عدد حزم البيانات التي تنتمي للعنقود k و *Zp* تمثل حزم البيانات في *خوارزمية* سرب *الطيور* كل *طير رقمي* (جسيم) يحتوي على *K* من مراكز العناقيد حيث m_{ik} تمثل مركز العنقود *k* للجسيم *i* لذلك فإن السرب يمثل مجموعة من الحلول المرشحة لعملية العنقدة. دالة اللياقة لكل جسيم يمكن حسابها بالمعادلة التالية :

$$f(x_i, Z_i) = w_1 \bar{d}_{max}(Z_i, x_i) + w_2(z_{max} - d_{min}(x_i)) + w_3 J_e$$

(10)

حيث أن z_{max} تمثل أعلى قيمة في البيانات ومصفوفة Zi تمثل البيانات الموزعة على عناقيد الجسيم *i* . أما *w1,* *w2, w3* تمثل أوزان ثابتة يحددها المستخدم.

$$\bar{d}_{max}(Z_i, x_i) = \max_{k=1,..K}\{\sum_{\forall z_p \in C_{i,k}} d(z_p, m_{i,k})/n_{i,k}\}$$

(11)

وهي أعلى قيمة لمعدل المسافة الإقليدية لمراكز العناقيد في الجسيمات والبيانات المرتبطة بها حيث أن nk هو عدد حزم البيانات التي تنتمي للعنقود k في الجسيم *i* . أما المسافة الإقليدية يتم حسابها بالمعادلة رقم (8) و d_{min} هي أقل مسافة اقليدية بين أي زوج من مراكز العناقيد:

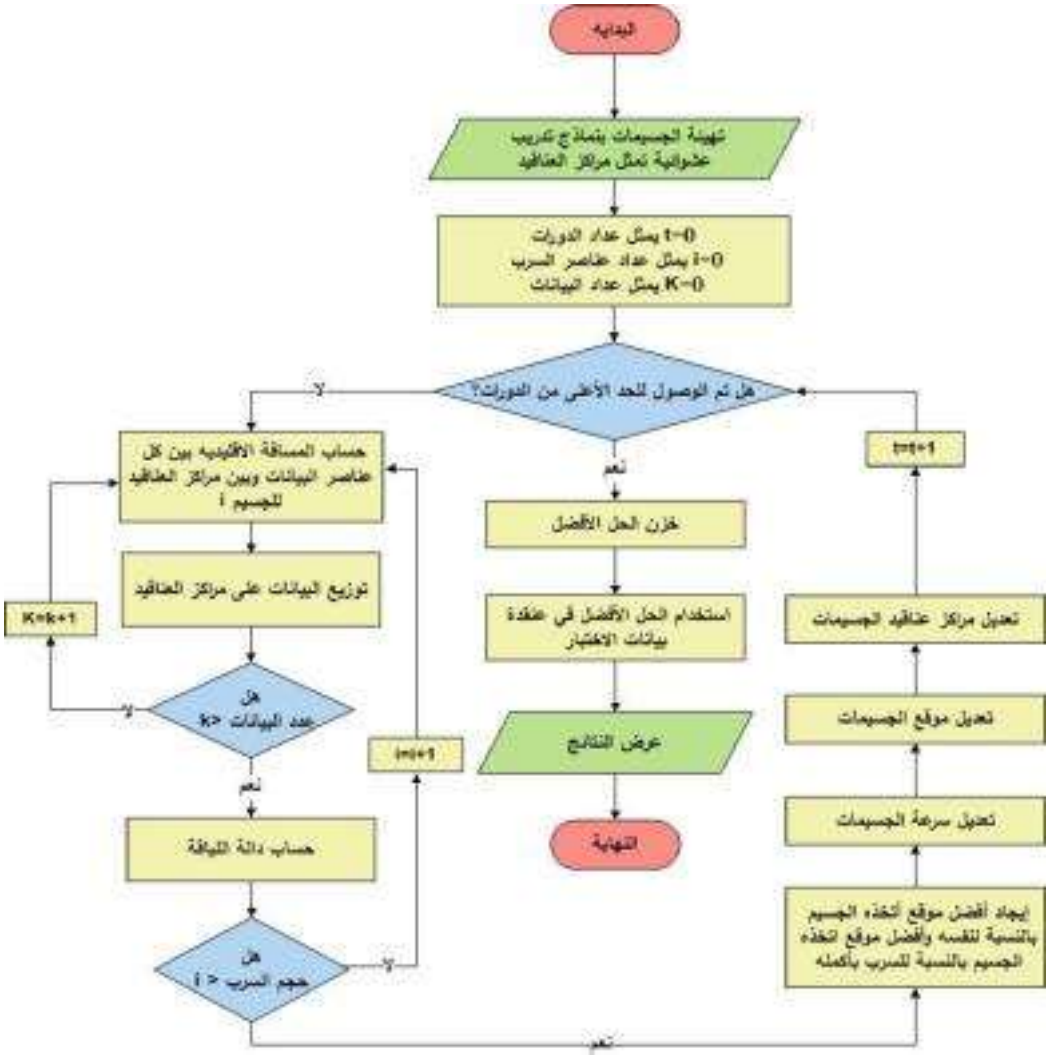
$$d_{min}(x_i) = \min_{\forall k,kk,k\neq kk}\{d(m_{i,k}, m_{i,kk})\}$$

(12)

تقوم دالة اللياقة في المعادلة رقم (10) بتقليل الخطأ الكمي *Je* عن طريق تصغير المسافة بين مراكز العناقيد والبيانات التي تنتمي اليها($d_{max}(Z_i, x_j)$ وتكبير المسافة بين مراكز العناقيد $d_{min}(x_i)$[17]].

كما ذكر سابقا يتأثر مسار كل طير رقمي (جسيم) في سرب الطيور بمسار أفضل جسيم في السرب بأكمله اذ تنجذب كل جسيمات السرب بسرعة وفي وقت واحد لأفضل جسيم في فضاء البحث ولكن هناك مشكلة وهي أن يكون أفضل جسيم في السرب بعيداً عن الحل الأمثل اذ يبدأ كل السرب بالتجمع حوله ويصلح من المستحيل على السرب اكتشاف مجالات أخرى في فضاء البحث لذلك سوف يحاصر السرب ويقع في مشكلة النهاية المحلية local optima وهذه أحدى مساوئ خوارزمية سرب الطيور التي تعتمد على أفضل جسيم في السرب بأكمله global best PSO . يوجد أسلوب آخر لخوارزمية سرب الطيور وهي أن يتأثر مسار كل طير رقمي (جسيم) بمسار أثنين من جيرانه أي الجار الأيمن والأيسر في السرب في هذا الأسلوب تزيد فرصة الاقراب من الحل الأمثل لأن البحث يكون بمجالات عديدة ويتم اكتشاف حلول عديدة ولكن مشكلة هذا الأسلوب أن التقارب يكون بطيئا جدا وتسمىlocal best PSO [17].

يمثل الشكل (11) المخطط الانسيابي لخوارزمية سرب الطيور المصممة لكشف التطفل.

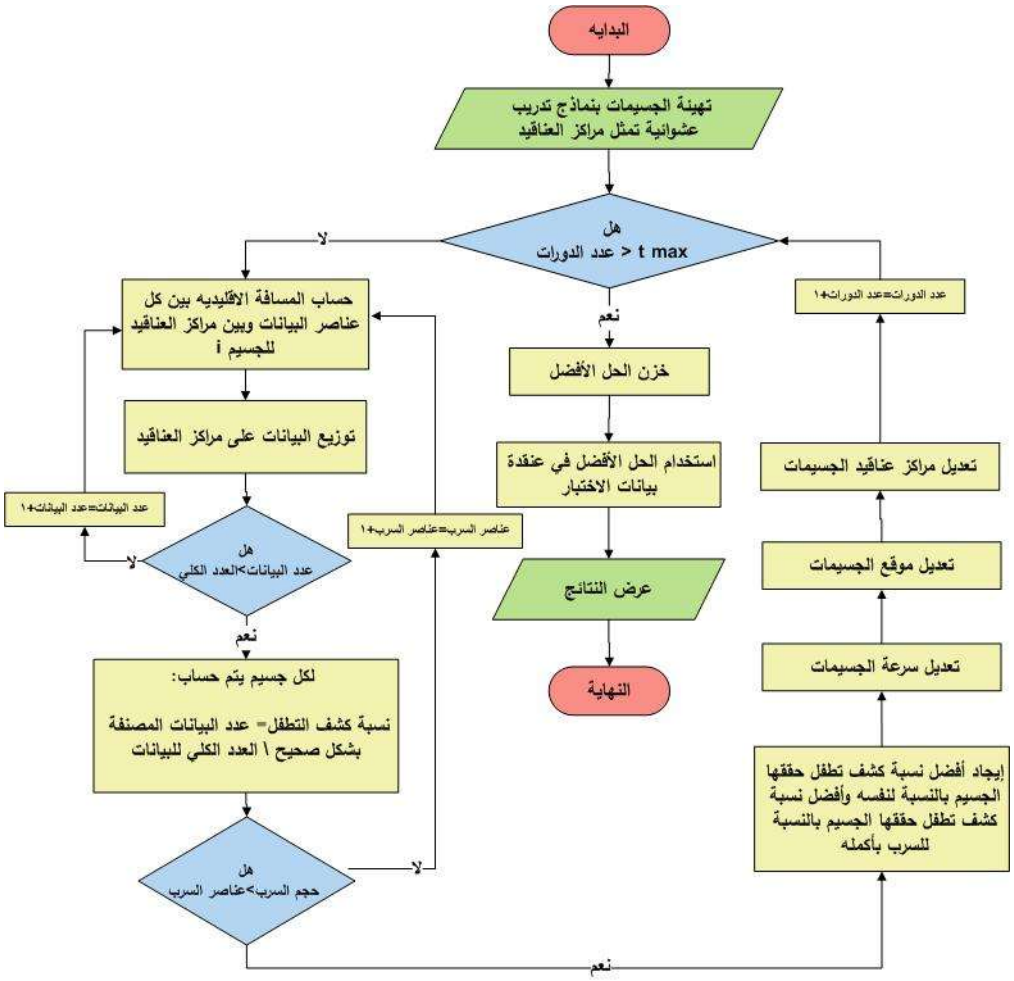


الشكل .(11) المخطط الانسيابي لخوارزمية سرب الطيور

5. خوارزمية سرب الطيور المطورة MPSO

بعد دراسة خوارزمية السرب بأنواعها — local best PSO و global best PSO تم التوصل الى خوارزمية سرب طيور جددة خاصة بنظام كشف التطفل حصرا تقضي على مساوئ — local best PSO بحيث يكون التقارب للحل الأمثل سريعا جدا بالإضافة الى التخلص من مشكلة الوقوع في النهاية المحلية local optima التي تعاني منها — PSO global best . اذ تم استبدال دالة اللياقة في الخوارزمية الأصلية المعادلة (10) بمقياس كشف التطفل detection rate الذي يتم قياسه بالمعادلة رقم (1) في كل دورة من دورات الخوارزمية يتم البحث عن الحل الأفضل في السرب اذ أن الطير الرقمي الذي يمتلك نسبة كشف تطفل أعلى من ابنقية يعد الأفضل في السرب والأقرب الى الحل الأمثل وبناء على ذلك تقوم جميع الطيور الرقمية في السرب بتغيير مواقعها وسرعاتها حسب المعادلات رقم (7,8) نسبة الى الطير الأفضل لكي يقترب كامل السرب من الحل الأفضل. ففي بداية دورات مرحلة التدريب لوحظ تباين كبير بنسب كشف التطفل لدى الطيور الرقمية وبعد الوصول للدورة التاسعة أوالعاشرة لوحظ تقارب كبير بنسب كشف التطفل لدى الطيور الرقمية حيث أصبحت جميع الطيور تقريبا تمتلك نسبة كشف تطفل 90% و 93% وفي

النهاية تم اختيار أفضل طير رقمي صاحب نسبة 93% الذي يعد هوالحل الأفضل لمشكلة كشف التطفل . يمثل الشكل (12) المخطط الانسيابي لخوارزمية سرب الطيور المطورة الخاصة بنظام كشف التطفل.



الشكل (12). المخطط الانسيابي لخوارزمية سرب الطيور المطورة.

4. التنفيذ والاختبار

تم تنفيذ النظام على 9000 عينة من بيانات *KDD99* حيث تم اختيار 5000 منها للتدريب و 4000 للاختبار كما في الجدول الآتي:

الجدول (1) محتويات ملفات كشف الهجمات المستخدمة في النظام.

محتويات ملف التدريب لكشف الهجمات

نوع الحزمة	عددها في الملف	نسبتها في الملف
Normal	2000	%40
Abnormal	3000	%60
العدد الكلي	5000	%100

محتويات ملف الاختبار لكشف

عددها في الملف	نسبتها في الملف
1500	%37.5
2500	%62.5
4000	%100

تم تدريب خوارزمية العنقدة kmeans وكذلك خوارزمية العنقدة المهجنة بالخوارزمية الجينية على5000 عينة من بيانات KDD99 واختبارها على 4000، ولوحظ بأن الخوارزمية المهجنة أعطت نتائج أفضل بكثير من خوارزمية العنقدة kmeans حيث تم الحصول على النتائج التالية:

الجدول (2). نتائج عملية تدريب واختبار خوارزميتي العنقدة والعنقدة المهجنة بالخوارزمية الجينية.

(مرحلة التدريب)

المقاييس	KM	HCA
DR	%78.98	%97.52
FP	51	0
FN	0	124
TP	3051	2876
TN	1949	2124
FNR	%0	%4.13
FPR	%2.55	%0
TNR	%97.45	%100
TPR	%100	%95.87
Precision	0.9836	1
Accuracy	0.9899	0.9758

(مرحلة الاختبار)

KM	HCA
%73.72	%90.6
51	0
0	624
2551	1876
1449	2124
%0	%24.96
%3.4	%0
%96.6	%100
%100	%75.04
0.9804	1
0.9874	0.8651

تم أيضاً تدريب خوارزميتي سرب الطيور وسرب الطيور المطورة على 5000 عينة من بيانات KDD99 واختبارها على 4000، ولوحظ بأن خوارزمية سرب الطيور المطورة الخاصة بنظام كشف التطفل أعطت نتائج مقاربة لنتائج خوارزمية سرب الطيور الأصلية في مرحلة التدريب ولكن في مرحلة الاختبار أعطت الخوارزمية المطورة نتائج أفضل بكثير من الخوارزمية الأصلية إذ تم الحصول على النتائج التالية:

الجدول (3) نتائج عملية تدريب واختبار خوارزميتي سرب الطيور وسرب الطيور المطورة.

(مرحلة التدريب)

المقاييس	PSO	MPSO
DR	%96,96	%100
FP	152	0
FN	0	0
TP	3152	3000
TN	1848	2000
FNR	%0	%0
FPR	%7,6	%0
TNR	%92,4	%100
TPR	%100	%100
Precision	0,954	1
Accuracy	0,9705	1

(مرحلة الاختبار)

PSO	MPSO
%96,37	%100
145	0
0	0
2645	1500
1355	2500
%0	%0
%9,67	%0
%90,33	%100
%100	%100
0,948	1
0,965	1

من النتائج أعلاه نستنتج أن أفضل خوارزمية من الخوارزميات الأربعة التي تم عرض نتائجها هي خوارزمية سرب الطيور المطورة الخاصة بنظام كشف التطفل اذ أعطت أفضل النتائج وذلك بسبب اعتمادها في عملها على مقياس كشف التطفل.

الجدول (4) مقارنة بين الطرق الأربعة من ناحية نسبة كشف التطفل في مرحلة الاختبار والوقت المستغرق في عملية التدريب.

الطريقة	DR	Execution time
KM	%73,72	3.4 seconds
HCA	%90,6	20.5 seconds
PSO	%96,37	1975.5 seconds
MPSO	%100	1336.9 seconds

الاستنتاجات والتوصيات:

بعد تطبيق طرائق كشف التطفل التي تعتمد على التقنيات الذكائية باستخدام خوارزمية سرب الطيور والخوارزمية الجينية على مجموعة بيانات KDDcup99 لوحظ ما يلي:

- بالنسبة لطريقة العنقدة التقليدية المتمثلة باستخدام خوارزمية Kmeans أعطت نتائج مقبولة الى حد ما ولكنها ليست بالمستوى المطلوب، وذلك اتضح من خلال حساب قيم مقاييس أداء النظام مثل نسبة الكشف ونسبة الإنذارات الكاذبة وغيرها من المقاييس الخاصة بكشف التطفل.
- أما بالنسبة لطريقة العنقدة المهجنة بالخوارزمية الجينية HCA فقد أعطت نتائج جيدة في عملية كشف

الهجمات وامتازت هذه الطريقة بربط طريقة تقليدية متمثلة بخوارزمية Kmeans مع الخوارزمية الجينية اذ دخل مبدأ KM في حساب دالة اللياقة في الخوارزمية الجينية، اضافة لذلك بعد الحصول على أفضل العناقيد في الخوارزمية المهجنة HCA يتم تمرير جميع البيانات على هذه العناقيد وحساب المسافة الإقليدية ومن ثم كشف الهجمات.

- تم استخدام خوارزمية سرب الطيور لإجراء عملية كشف الهجمات وذلك بتنفيذ الخوارزمية على مجموعة بيانات والحصول على الحل الأمثل، وهذه الطريقة أعطت نتائج أفضل من الطريقتين أعلاه، ولكن هناك مشكلة وهي أن يكون أفضل جسيم في السرب بعيد عن الحل الأمثل حيث يبدأ كل السرب بالتجمع حوله ويصبح من المستحيل على السرب اكتشاف مجالات أخرى في فضاء البحث لذلك سوف يحاصر السرب ويقع في مشكلة النهاية المحلية optima local.
- وأخيرا تم استخدام خوارزمية سرب الطيور المطورة MPSO التي قضت على مساوئ PSO best local بحيث يكون التقارب للحل الأمثل سريع جدا بالإضافة الى التخلص من مشكلة الوقوع في النهاية المحلية local optima التي تعاني منها PSO best global . اذ تم استبدال دالة اللياقة في خوارزمية سرب الطيور بمقياس كشف التطفل rate detection الذي يعد المقياس الأهم في أنظمة كشف التطفل. وأعطت هذه الطريقة أفضل النتائج بجميع المقاييس المستخدمة حيث وصلت نسبة الكشف الى 93% ولم تؤدِ للوقوع في مشكلة النهاية المحلية

استخدام خوارزمية سرب الطيور المطورة MPSO للقضاء على مساوئ النهاية المحلية ولتحقيق تقارب للحل الأمثلي بشكل سريع مما يؤدي إلى زيادة فاعلية كشف الشذوذات الشبكية وكشف التطفل.

المراجع:

1] WEI ,Li *" Using Genetic Algorithm for Network Intrusion Detection"* Department of Computer Science and Engineering Mississippi State University, Mississippi State, MS,2015, 39762

2] ROZENBLUM .D. *"understanding Intrusion Detection System",* SANS Institute. 2017.

3] ANDERSON, R " *Security Engineering: A Guide to Building Dependable Distributed Systems",* Second Edition, 2018.

4] HERRERO, A; CORCHADO, E "*Mobile Hybrid Intrusion Detection*" Springer. 2018.

5] TOPARK ,M."*Intrusion Detection System Alert Correlation with Operating System Level Logs*" A Thesis Submitted to The Graduate School of Engineering and Sciences of Izmir Institute of Technology, 2017.

6] http://www.windowsecurity.com

7] CRAMER, L.M ; Cannady, J ; Harrell, J., "*New Methods of Intrusion Detection using Control- Loop Measurement*" , Georgia Institute of Technology Atlanta, 2016.

8] WILSON, J ; DUDLEY, P; KHAN, B., "*Requirements Specification* " , CQF- QMT, 2018.

9] ODUTOLA, K ; OGUNTIMEHIN, A ; TOLKE, L ; WULP, M. V. , "ArgoUML Quick Guide " . 2018 .

10] KANG, D. ; FULLER, D. ; HONAVAR, V., "*Learning Classifiers for Misuse and Anomaly Detection Using a Bag of System Calls Representation*" ,2015, IEEE .

11] NORSYAFAWATI, F. , NORWAWI, N. ; SEMAN, K. , "*Identifying False Alarm Rates for Intrusion Detection System with Data Mining*" , IJCSNS International Journal of Computer Science and Network Security, 2016, VOL.11 No.4.

12] HAMMERSLAND, R., "*ROC in Assessing IDS Quality*" , Norwegian Information Security Lab, Gj0vik University College. 2017.

13] CHANG, R ;LAI, L ; WANG, J. ; KOUH, J., "*Intrusion Detection by Backpropagation Neural Networks with Sample-Query and Attribute-Quer*y", International Journal of Computational Intelligence Research., 2017.Vol.3, No.1 .
14] ADETUNMBI, A.; ADEOLA, S ; DARAMOLA, O.,"*Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features"* , Proceedings of the World Congress on Engineering and Computer Science, 2016 Vol I .
15] VESCAN, A. ; HORIA, F. , "*Constraint Optimization-Based Component Selection problem*" , INFORMATICA, 2018, Volume LIII, No. 2.
16] IZENMAN A. J.,"*Modern Multivariate Statistical Techniques*", Springer ,2018.
17] ABRAHAM, A. ; GROSAN, C ; RAMOS, V., "Swarm Intelligence in Data Mining", Springer. ,2016.
18] VINU, V.; THOMAS, G ; LUMBAN, F.,"*Information Technology and Mobile Communication*" Springer , International Conference. , 2015.
19] SHONKWILER, R. W. ; MENDIVIL, F, "*Explorations in Monte Carlo Methods*",Springer ,2018.