Detection of network anomaly based on hybrid intelligence techniques

Shahbaa I. Khaleel Karam mohammed mahdi saleh
shahbaaibrkh@uomosul.edu.iq
College of Computer Sciences and Mathematics
University of Mosul

Received on: 11/09/2011 Accepted on:14/12/2011

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 practices of a security officer, 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 , Intrusion Detection Systems , Swarm Optimization, Genetic Algorithm , clustering algorithm.

كشف الشذوذ الشبكي المعتمد على التقنيات الذكائية المهجنة

شهباء إبراهيم خليل كرم محمد مهدي صالح
كلية علوم الحاسوب والرياضيات، جامعة الموصل

تاريخ استلام البحث: 2011/09/11 تاريخ قبول البحث: 2011/12/14

الملخص

تقنيات الذكاء الاصطناعي يمكن أن تجعل أنظمة كشف التطفل أسهل بكثير مما عليه اليوم وكما هو الحال دائما فأن أصعب شيء في تعلم الأنظمة المصممة بالتقنيات الذكائية هو عملية تدريبها لتعلم الأمور الصحيحة. هذا البحث يركز على عمل بيئة لأنظمة كشف التطفل وتعليمها ممارسة العمل مثل ضابط الأمن. تم في هذا البحث تصميم نظام ذكائي مهجن ومطور لكشف التطفل الشبكي, إذ قدم البحث أربعة طرائق لكشف الشذوذ الشبكي باستخدام تقنية العنقدة والاعتماد على تقنيات الذكاء الاصطناعي 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, في الطريقة الثانية *HCA* تم استخدام طريقة العنقدة المهجنة إذ تم تهجين خوارزمية *Kmeans* مع الخوارزمية الجينية. أما في الطريقة الثالثة فقد تم استخدام خوارزمية سرب الطيور *PSO*. بالاعتماد على الطريقة الثالثة أُنشأت الطريقة الرابعة وهي خوارزمية سرب الطيور المطورة *MPSO* وكانت هذه الطريقة الأفضل من بين الطرائق الأربعة المستخدمة في هذا البحث.

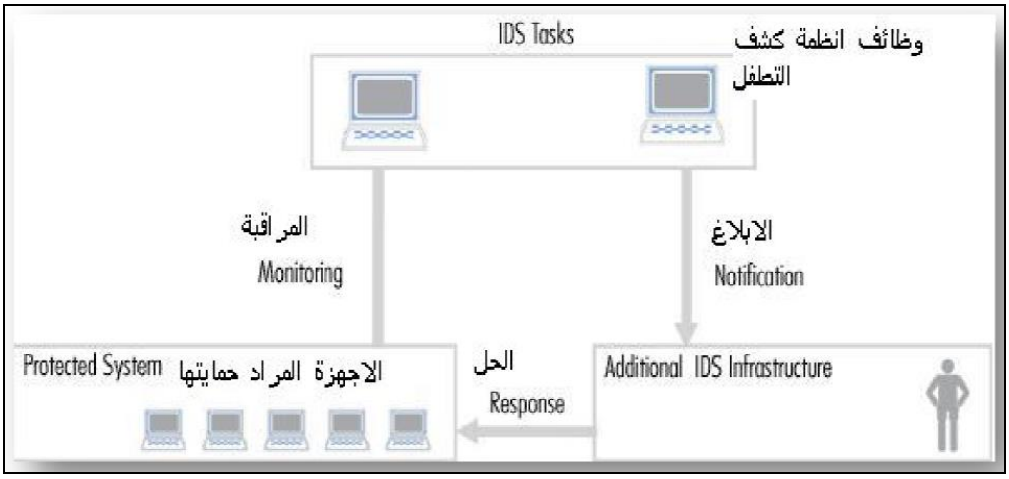
الكلمات المفتاحية: الذكاء الاصطناعي،أنظمة كشف التطفل ، أمثلية سرب ، الخوارزمية الجينية،خوارزمية العنقدة.

1. مقدمة

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

كشف التطفل يتيح رصد وتحليل نشاط المستخدم والنظام، تدقيق تكوينات النظام ونقاط الضعف، تقييم سلامة النظام وملفات البيانات، التحليل الإحصائي لنشاط النماذج *patterns* على أساس المطابقة لهجمات معروفة، تحليل النشاطات الشاذة، وتشغيل نظام المراجعة [2]. كشف التطفل له أسلوبان:

- الكشف عن الشذوذ: يشير إلى التقنيات التي تحدد وتميز السلوك العادي أو المقبول للنظام على سبيل المثال، استخدام وحدة المعالجة المركزية، وظيفة وقت التنفيذ. والتصرفات التي تحيد عن السلوك العادي المتوقع تعدها اختراقات.
- الكشف عن إساءة الاستخدام: يشير إلى التقنيات التي تميز طرائق معروفة لاختراق النظام. تتميز هذه الاختراقات بأنها نماذج *patterns* أو توقيع *signature* مخزونة في قاعدة المعرفة للنظام لذلك يقوم نظام كشف التطفل بالبحث عن النماذج والتواقيع المشابهة لها ويعدها اختراقات. النماذج أو التواقيع قد تكون جملة ثابتة أو تسلسل مجموعة من الإجراءات, وتستند استجابات النظام على الاختراقات التي تم تحديدها و الشكل رقم (1) يمثل البنية التحتية لأنظمة كشف التطفل [3].



الشكل رقم (1). البنية التحتية لأنظمة كشف التطفل

نظام كشف التطفل من ناحية مجال العمل والحماية يقسم بصورة عامه إلى قسمين:

1. نظام كشف التطفل المستند على المضيف ***Host Based Intrusion Detection***

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. يقوم كذلك بمراقبة ملفات النظام الرئيسية والتنفيذية من خلال مراقبة التواقيع في فترات منتظمة لكي يكتشف التغيرات الحاصلة عليها جراء الهجمات غير المتوقعة [4,5]

2. نظام كشف التطفل المستند على الشبكة ***Network Based Intrusion Detection***

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 ويُثبت هذا النوع من الأنظمة في جهاز الموجه *Router* أو أجهزة الشبكة الأخرى [4,5].

أما من ناحية السلوك فتقسم أنظمة كشف التطفل إلى نوعين , نظام كشف التطفل الخامل *passive* الذي يقوم بكشف التطفل و إرسال التنبيه فقط , أما النوع الثاني فهو نظام كشف التطفل الفعال *Active* الذي يقوم بكشف التطفل وأيضا 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 يتم مراقبة البيانات وتحليلها وإرسال التنبيه مباشرة في حالة اكتشاف الهجمات لكي تُؤخذ إجراءات فورية لمواجهتها ويسمى هذا النوع *On-line analysis* .أما النوع الثاني يقوم بأخذ مقطع من النظام وتقييم الوضع الأمني إذ يقوم بأجراء تحليل أكثر شمولية من النوع الأول دون أن يكون له تأثير غير مقبول على أداء الأنظمة التي يتم مراقبتها ويسمى هذا النوع *Off-line analysis* [4].

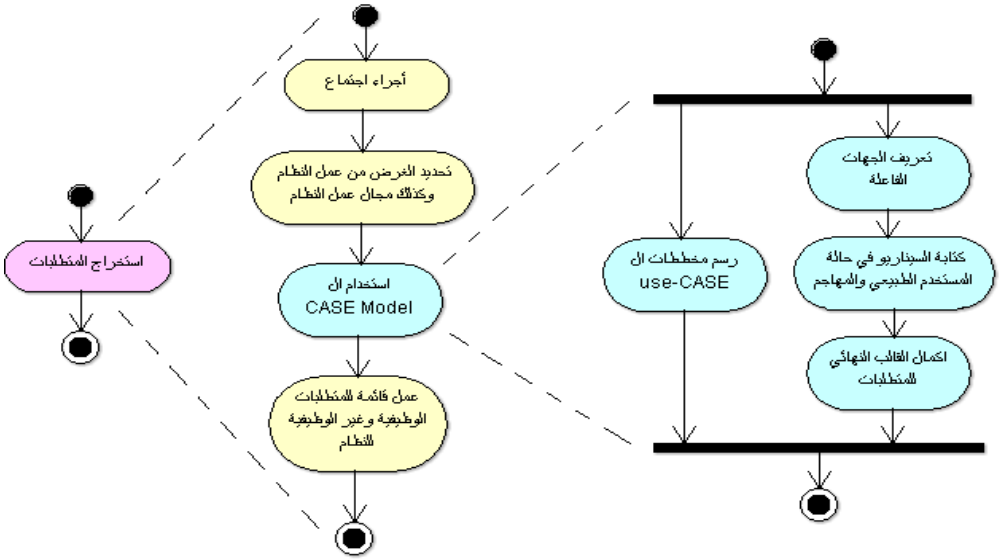
1.1 الدراسات السابقة

لقد تم دراسة مشكلة كشف التطفل وتصنيفه ضمن حقل أمنية شبكة الحاسوب ففي عام 2002. قدم الباحثان الأمريكيان *Sabhnani .M* و*Serpen .G* نظامهما لكشف التطفل بتطبيق التقنيات الذكائية على بيانات الـ *KDD99* لكشف التطفل بصيغة إساءة الاستخدام. عمد الباحثان على تطبيق أنظمتهم وبمختلف التقنيات الذكائية على هذه البيانات وكانت تقنياتهم تتضمن عدة حقول مثل الشبكات العصبية الاصطناعية والمنطق المضبب 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 كل النتائج وكانت خوارزمية *K-means* الأفضل بنتائج الكشف الخاصة بصنفي الهجوم *DOS* و*U2R* وخوارزمية *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 2005 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 *KDD99* في تدريب النظام واختباره وأظهرت النتائج أن أداء خوارزمية أشجار القرار أصبح أفضل بعد دمجها مع الخوارزمية الجينية. وأيضا في عام 2005 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 بـ *Kmeans*. في عام 2011 أقترح الباحث *Liu y.* أسلوباً جديداً 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.

2. وصف المتطلبات *Requirements specification*

تلعب المتطلبات دوراً أساسياً خلال عملية إنشاء المنتج. يتم توثيق المتطلبات في وثيقة تسمى وثيقة وصف المتطلبات. استخراج المتطلبات *Requirements elicitation* هو نشاط مهم في عملية هندسة البرمجيات يهدف إلى اكتساب وفهم المتطلبات. الشكل رقم(2) يمثل مخطط استخراج متطلبات النظام.



الشكل رقم (2). مخطط *UML activity* لاستخراج المتطلبات

1.2 المشكلة *Problem*

نظم وشبكات المعلومات تخضع 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]، الجدار الناري *Firewall* هو ليس النظام الديناميكي ذو القدرة الدفاعية المطلوبة الذي يمكن أن يتخيله المستخدم. ولكن في المقابل، يوجد أنظمة كشف التطفل *IDS* التي لقد أنظمة ديناميكية ذات قدرة دفاعية عالية, وتتعرف أنظمة كشف التطفل على الهجمات التي لا تستطيع الجدران النارية التعرف عليها [6].

2.2 الغرض *Purpose*

سيقوم النظام 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.

3.2 المجال *Scope*

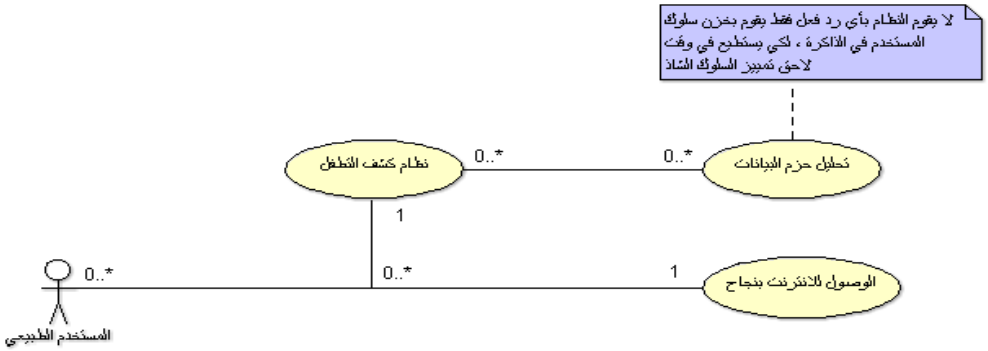
إن مجال عمل أنظمة كشف التطفل تشمل أنواع وكميات كبيره من النظم التي يتعين دعمها، إذ يمكن أن يكون النظام المحمي آلة فردية أو شبكة من الأجهزة [7].

4.2 استخدام مخطط هندسة البرمجيات بمساعدة الحاسوب *Use-CASE Model*

تستخدم هذه المخططات في هندسة البرمجيات لاستخراج المتطلبات الوظيفية للنظام. وتقوم أيضاً بتعريف التفاعل بين المستخدمين *Actors* والنظام [8]. تم استخدام *ArgoUML* لتطبيق الـ*Use-CASE* على النظام المصمم في هذا البحث. *ArgoUML* هي أداة تصميم قوية سهلة الاستخدام، تدعم تصميم البرمجيات الرسومية وتطوير وتوثيق تطبيقات البرمجيات[9]، يوجد ثلاث جهات فاعلة *Actors* أساسية في النظام *IDS* وهم مسؤول النظام والمستخدم الطبيعي للنظام والمهاجم. إن نظام كشف التطفل يعمل مع سيناريوهين :

1.4.2 مستخدم طبيعي يستخدم النظام

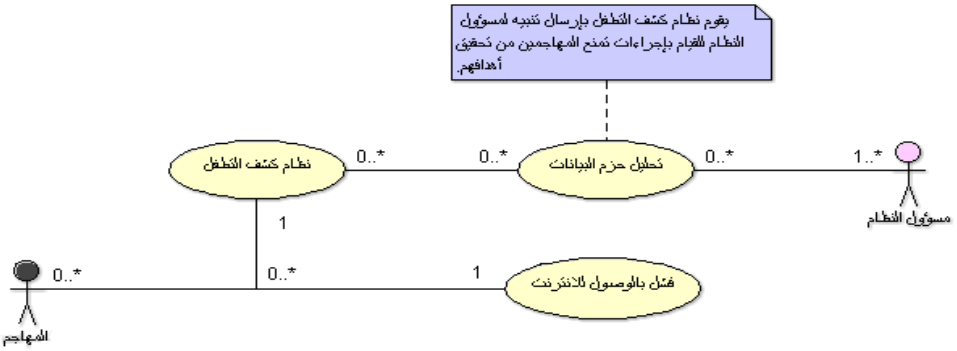
خلال الاستخدام العادي للنظام، سوف يقوم النظام برصد كل حركات المرور في الجهاز أو الشبكة المحمية. في حالة من النشاط العادي يجب أن يبقى النظام هادئاً ولا يرد على أي شيء طبيعي، ويوفر الخدمات بصورة كاملة. يقوم النظام بتحليل سلوك المستخدم ويبقيه في الذاكرة، لكي يستطيع في وقت لاحق تمييز السلوك الشاذ كما في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



الشكل رقم (3). مخطط *UML Use case* للمستخدم الطبيعي للنظام

2.4.2 شخص يحاول مهاجمة النظام

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



الشكل رقم (4). مخطط *UML Use case* لشخص يحاول مهاجمة النظام

5.2 المتطلبات الوظيفية *Functional Requirements*

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

6.2 المتطلبات الغير وظيفية *Non-Functional Requirements*

- يجب أن يكون النظام سهلاً لتدريب المستخدمين الجدد على هذا البرنامج.
- على المستخدم أن يكون قادرا على تشغيل وإيقاف النظام بسهولة.
- ضبط معلمات النظام والتحكم فيها بسهولة.
- الموثوقية *Reliability* : مسؤولو النظام لا يحتاجون 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 ثُم إخطار مسؤول النظام في أقرب وقت ممكن.
- التعزيز *Supportability* : التكوين الحيوي للنظام يسمح لمسؤولي النظام بإضافة وتغيير الأوزان والمعلمات دون التأثير على خدمات النظام، ودون تعطيل أي من الخدمات الأخرى على شبكة الاتصال.

7.2 تقييم أداء أنظمة كشف التطفل *IDSs Performance Evaluation*

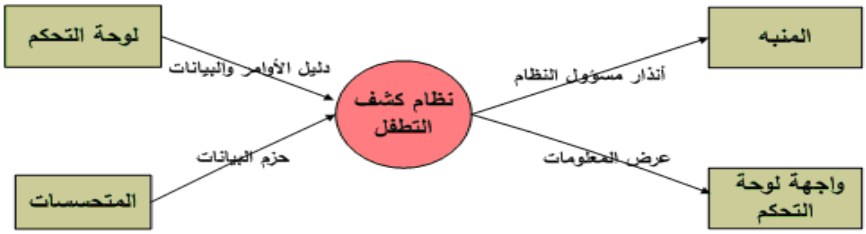
مقاييس كشف التطفل مهمة جداً لتقييم الأداء الفني لنظام كشف التطفل، فمحللو الأمنية يقومون بالإطلاع على مخرجات أنظمة كشف التطفل لكي يتعرفوا على إمكانية حدوث الهجوم ومتى يتم إصدار الإنذارات. بالإضافة إلى ذلك، فان مسؤول النظام يحتاج إلى امتلاك القدرة على المقارنة بين نقاط القوة والضعف الموجودة في أنظمة كشف التطفل الحالية واختيار الملائم منها. تعد نسبة الكشف *Detection Rate* من المعايير المهمة لقياس أداء نُظم كشف التطفل ويتم قياسها حسب المعادلة التالية[10]:

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

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

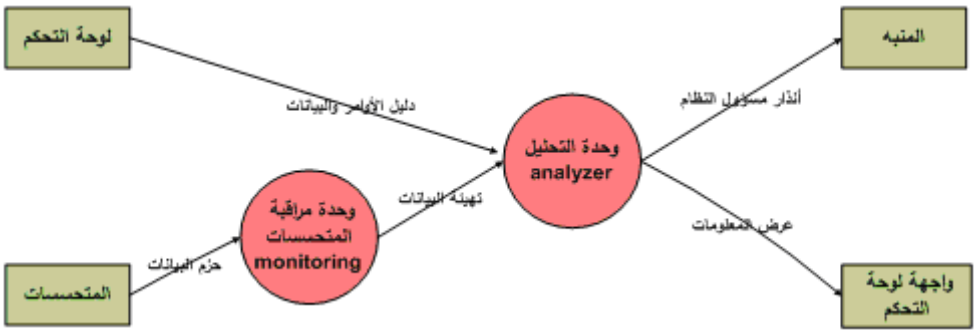
3. تحليل و تصميم النظام *Analysis and design of the system*

تضمن هذا البحث محورين أساسيين من محاور الذكاء الاصطناعي لكشف التطفل المِحور الأول يتمثل بالخوارزمية الجينية *Genetic Algorithm* والمِحور الثاني يتمثل بخوارزمية سرب الطيور *particle swarm optimization* إذ استخدمت في عملية كشف الهجمات أي كشف الشذوذ *Anomaly detection* وكذلك خوارزمية سرب الطيور المطورة *modified particle swarm optimization* التي استخدمت للغرض نفسه ولكن بنتائج أفضل وفيما يلي الشكل رقم(5) يمثل مخطط تحليل سير البيانات في نظام كشف التطفل الذي أُعتُمِدَ في هذا البحث.



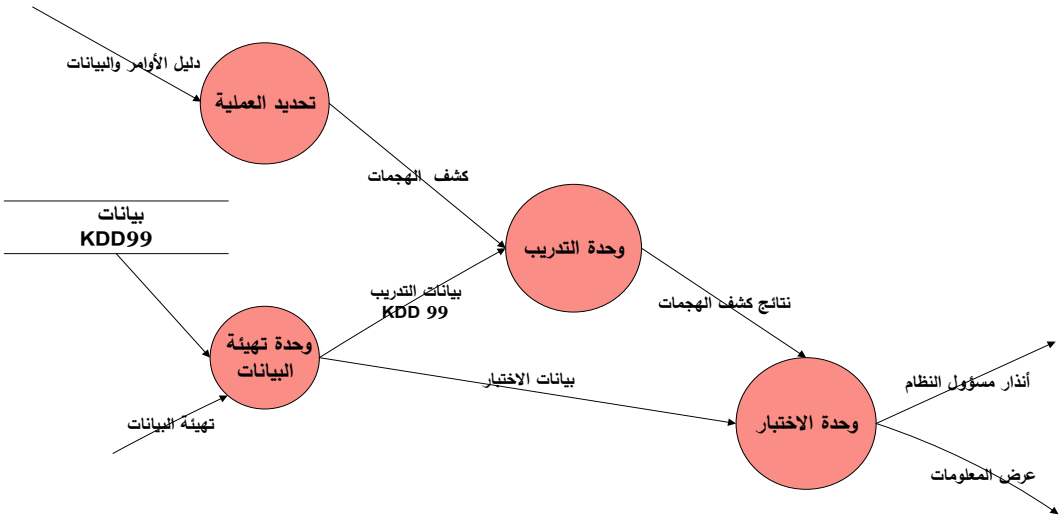
الشكل رقم (5). يمثل المرحلة رقم صفر لمخطط سير البيانات في النظام

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



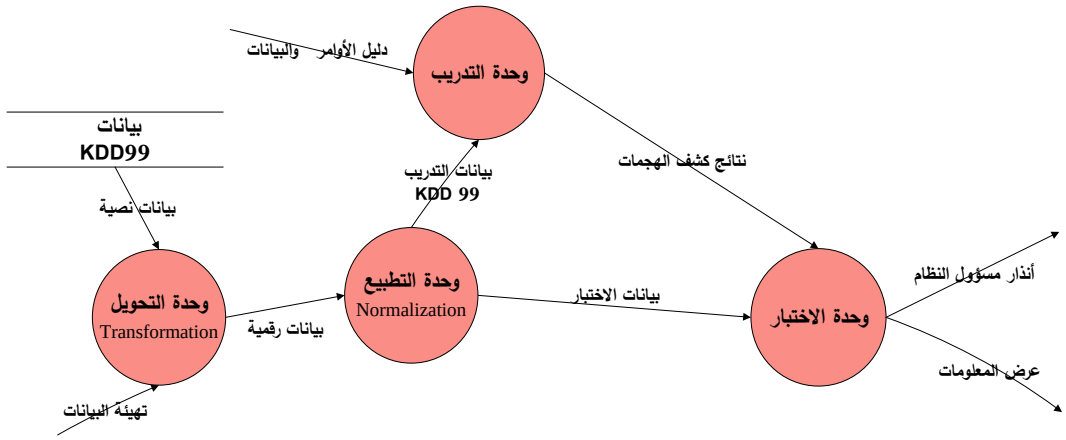
الشكل رقم (6). يمثل المرحلة رقم 1 لمخطط سير البيانات في النظام

وحدة التحليل تمثل 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 رقم (7). يمثل المرحلة رقم 2 لمخطط سير البيانات في وحدة التحليل *Analyzer*

أما البيانات المستخدمة في عملية التدريب والاختبار هي مجموعة بيانات *KDD99* وهي لقد البيانات الأكثر استخداماً في بناء نظم كشف التطفل منذ عام 1999 [14]. ولكن قبل استخدام هذه البيانات يتم إدخالها على وحدة ما قبل معالجة البيانات وهي وحدة تهيئة البيانات الخاصة بالنظام التي تقسم بدورها إلى وحدة التحويل *Transformation* لإنتاج بيانات إدخال تتوافق مع بيانات الإخراج للنظام إذ 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 هو [1.0 – 0.0] [15]. كما موضح في الشكل رقم (8).



الشكل رقم (8). يمثل المرحلة رقم 2 لمخطط سير البيانات في وحدة تهيئة البيانات

1.3 عنقدة البيانات *Data Clustering*

العنقدة هي عملية تقسيم البيانات إلى مجاميع اعتمادا على بعض المقاييس المتشابهة لهذه المجاميع ولقد عملية عنقدة البيانات عملية أساسية ومركزية في الذكاء الاصطناعي إذ يتم تعريف العنقود بواسطة مركز العنقود *cluster center* والطريقة الأكثر شيوعا لإيجاد مقاييس التشابه بين البيانات ومراكز العناقيد هي المسافة الإقليدية *Euclidean distance* التي تقاس بالمعادلة التالية :

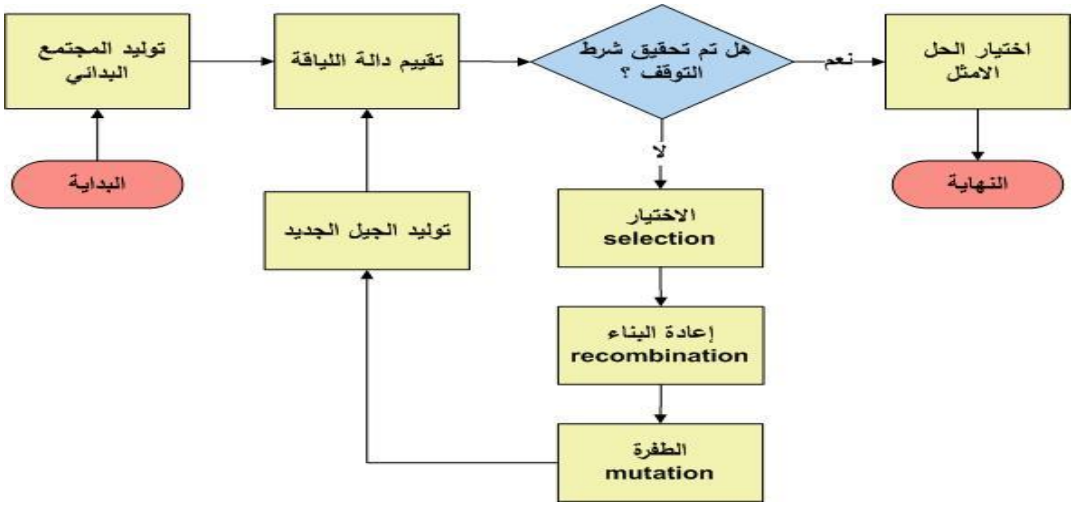
$$d(Z_u, Z_w) = \sqrt{\sum_{j=1}^{Nd} (z_{u,j} - z_{w,j})^2} = \|Z_u - Z_w\|_2 \quad \ldots(2)$$

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 إلى العنقود الأقرب حسب دالة اللياقة الخاصة بهذه الخوارزمية[17,16]:

$$J_{K\text{-}means} = \sum_{k=1}^{K} \sum_{\forall z_p \in C_k} d^2(z_p, m_k) \quad \ldots(3)$$

2.3 الخوارزمية الجينية *Genetic Algorithm*

الخوارزمية الجينية لقد من النماذج الحسابية القائمة على مبادئ التطور والانتقاء الطبيعي .هذه الخوارزميات تقوم بنمذجة المشكلة في مجال معين باستخدام الكروموسومات وتطور هذه الكروموسومات باستخدام الاختيار *selection*، وإعادة التركيب *recombination*، وعملية الطفرة *mutation*. في تطبيقات أمن الكمبيوتر تستخدم الخوارزمية الجينية لإيجاد الحل الأمثل لمشكلة معينة. يبدأ تطوير الخوارزمية الجينية عادة مع أفراد تم اختيارها عشوائياً 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 رقم (9). يمثل 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].

3.3 خوارزمية العنقدة المهجنة *Hybrid Clustering Algorithm HCA*

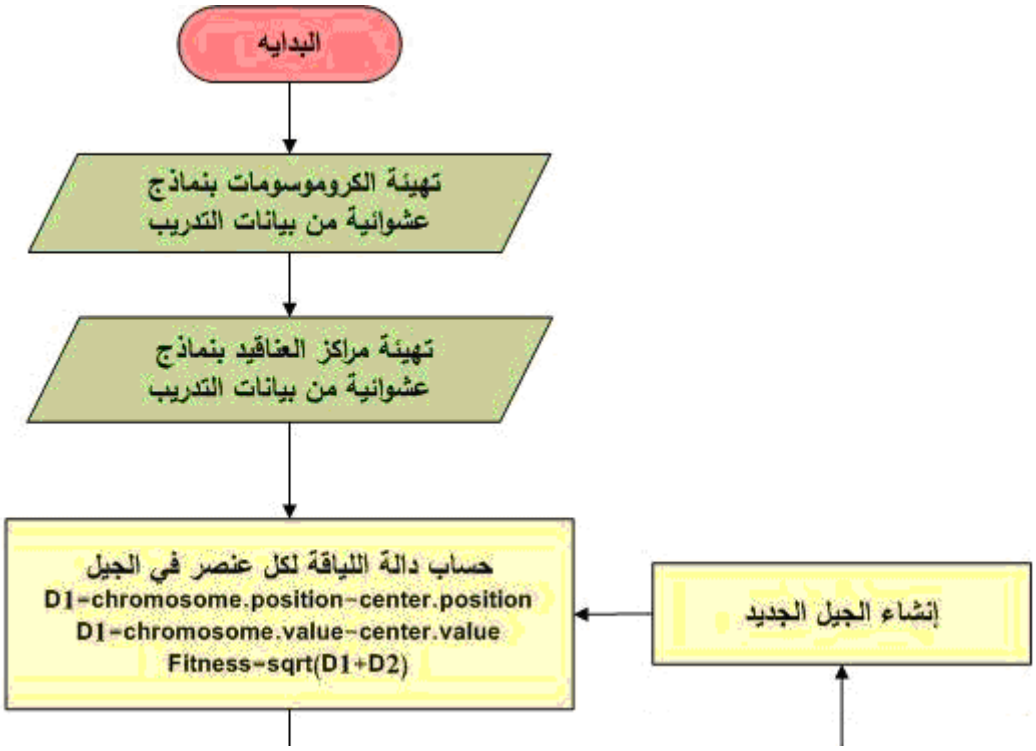
لتحسين استخدام خوارزمية *kmeans* في كشف التطفل تم دمجها مع الخوارزمية الجينية حيث تم اعتبار كل حزمة من حزم بيانات الـ*KDD99* كروموسوم وكل حزمة تحتوي على 41 صفة:

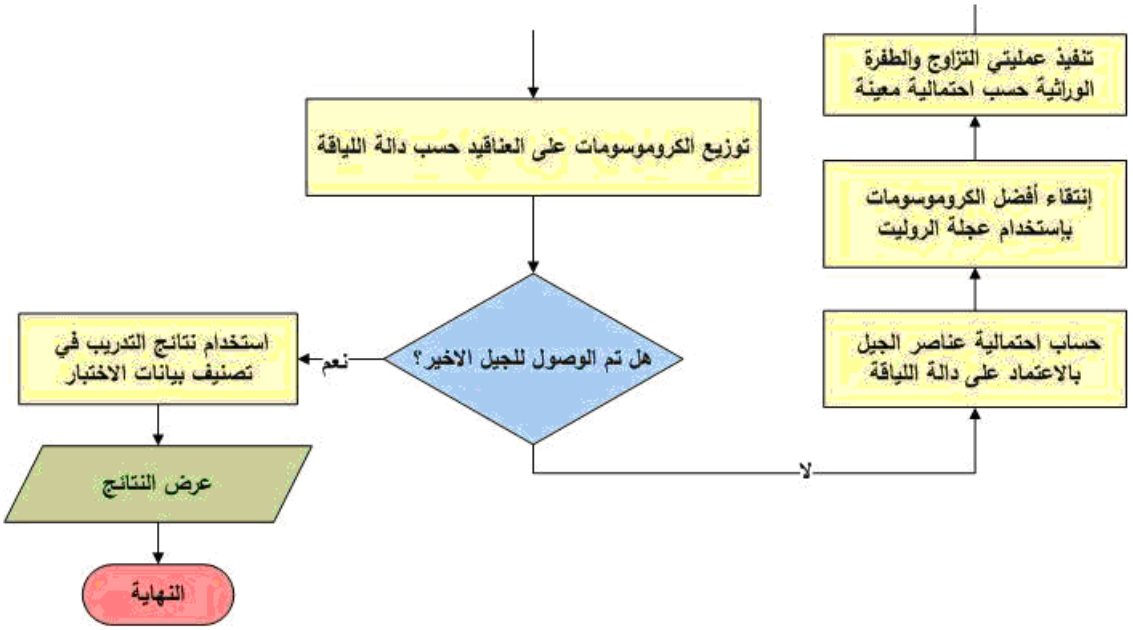
$$\text{chromosome} = \{ \text{duration} , \text{protocol type} , \text{service} , \text{flag} , \text{....etc} \}$$

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 بعجلة الروليت باعتماد المصدر [19]، باعتبار p هي احتمالية النجاح و $1\text{-}p$ هي احتمالية الفشل فأن:

$$F(x) = \begin{cases} 0, & if\ x < 0, \\ 1-p, & if\ 0 \le x < 1, \\ 1, & if\ 1 \le x \end{cases} \qquad \ldots(4)$$

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 رقم (10). يمثل المخطط الانسيابي للخوارزمية المستخدمة في نظام كشف التطفل.

4.3 خوارزمية سرب الطيور *Particle Swarm Optimization*

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

- x_i : يمثل الموقع الحالي للجسيم.
- v_i : السرعة الحالية للجسيم.
- y_i : أفضل موقع اتخذه الجسيم.

لنفرض أن f هي دالة اللياقة و t هو الزمن الحالي فإن أفضل موقع اتخذه الجسيم يتم تحديثه كالأتي:

$$y_i(t+1)=\begin{cases} y_i(t) & if \quad f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) & if \quad f(x_i(t+1)) < f(y_i(t)) \end{cases} \qquad \ldots(5)$$

ومن ثم أيجاد أفضل موقع أتخذه الجسيم بالنسبة للسرب بأكمله بواسطة المتجه $\hat{y}$

$$\hat{y}(t) \in \{ y_0, y_1, \ldots, y_s \} = min\{ f(y_0(t)), f(y_1(t)), \ldots, f(y_s(t)) \} \qquad \ldots(6)$$

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

$$v_{i,j}(t+1) = wv_{i,j}(t) + c_1 r_{1,j}(t)(y_{i,j}(t) - x_{i,j}(t)) + c_2 r_{2,j}(t)(\hat{y}(t) - x_{i,j}(t)) \qquad \ldots(7)$$

حيث i تمثل الجسيم و $j=\{1,..,Nd\}$ حيث N_d تمثل أبعاد المسألة و w هو وزن القصور الذاتي قيمته بين $\{0,1\}$ و c_1,c_2 يمثلون ثوابت التعجيل c_1 يكون المسؤول عن التحكم في زيادة ونقصان البحث المحلي أما c_2

يكون المسؤول عن التحكم في زيادة ونقصان البحث الشامل. و $r_{1,j}, r_{2,j}$ أرقام عشوائية بين {0,1} . وأخيرا يتم تحديث موقع الجسيم i وهو xi بالمعادلة التالية:

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad \ldots(8)$$

يمكن استخدام مقاييس كثيرة لقياس جودة خوارزميات العنقدة, مقياس الأداء الأكثر شيوعا هو مقياس الخطأ الكمي : Je

$$Je = \frac{\sum_{k=1}^{K} \sum \forall zp \in C_k \; d(z_p, m_k) / nk}{K} \quad \ldots(9)$$

حيث C_k تمثل العنقود k و n_k 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 $x_i = (m_{i,1}, .., m_{i,k}, ..., m_{i,K})$ و $m_{i,k}$ تمثل مركز العنقود k للجسيم i لذلك فإن السرب يمثل مجموعة من الحلول المرشحة لعملية العنقدة. دالة اللياقة لكل جسيم يمكن حسابها بالمعادلة التالية :

$$f(x_i, Z_i) = w_1 \bar{d}_{max}(Z_i, x_i) + w_2(z_{max} - d_{min}(x_i)) + w_3 J_e \quad \ldots(10)$$

حيث أن z_{max} تمثل أعلى قيمة في البيانات ومصفوفة Z_i تمثل البيانات الموزعة على عناقيد الجسيم i . أما w_1, w_2, w_3 تمثل أوزان ثابتة يحددها المستخدم.

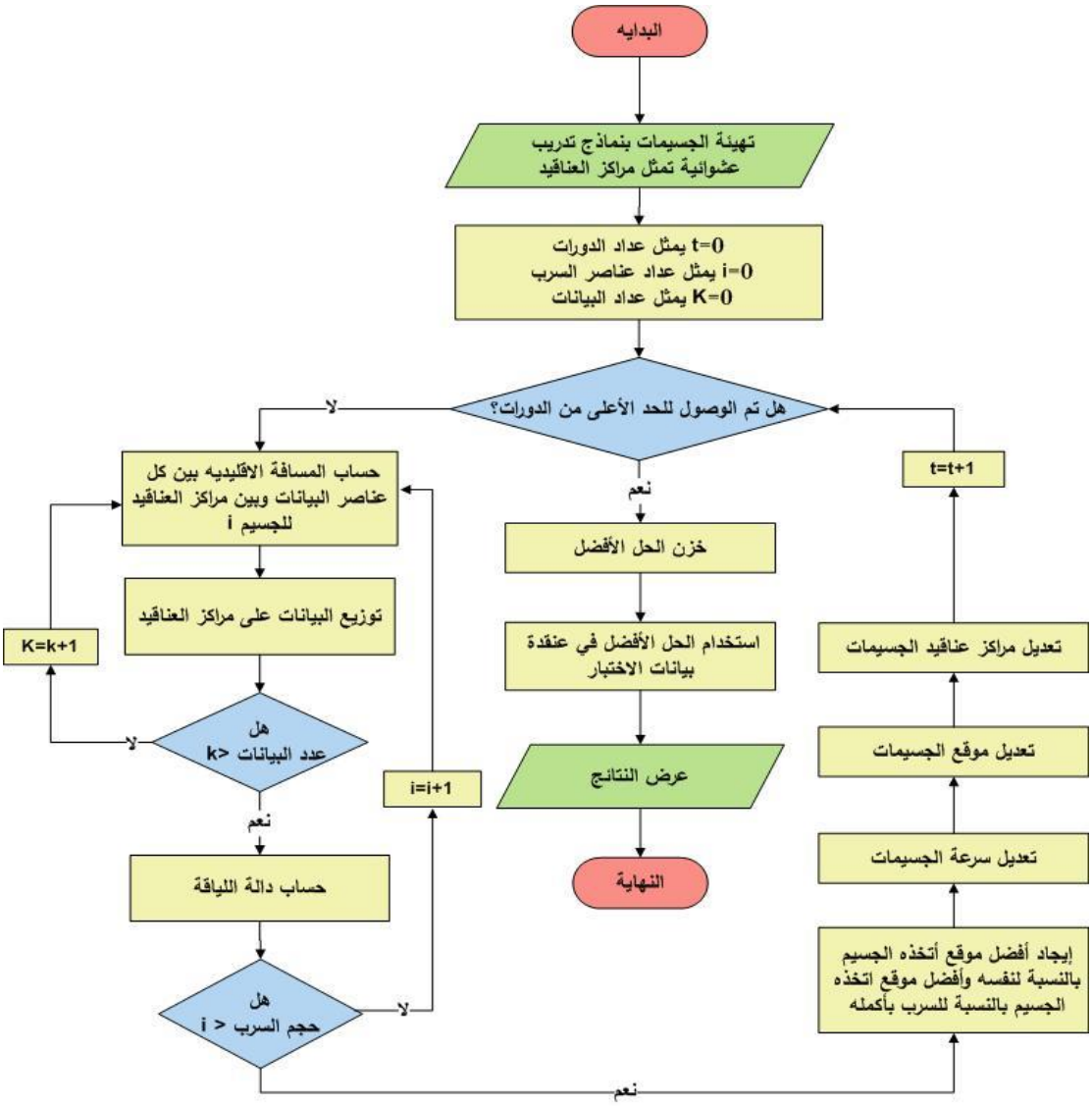
$$\bar{d}_{max}(Z_i, x_i) = \max_{k=1,...K} \{ \sum_{\forall z_p \in C_{i,k}} d(z_p, m_{i,k}) / n_{i,k} \} \quad \ldots(11)$$

وهي أعلى قيمة لمعدل المسافة الإقليدية لمراكز العناقيد في الجسيمات والبيانات المرتبطة بها حيث أن $n_{i,k}$ 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 إقليدية بين أي زوج من مراكز العناقيد:

$$d_{min}(x_i) = \min_{\forall k, kk, k \neq kk} \{ d(m_{i,k}, m_{i,kk}) \} \quad \ldots(12)$$

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

كما ذُكر سابقا يتأثر مسار كل طير رقمي (جسيم) في سرب الطيور بمسار أفضل جسيم في السرب بأكمله إذ 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 إذ يبدأ كل السرب بالتجمع حولهُ ويصبح 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 جيرانهِ أي الجار الأيمن والأيسر في السرب في هذا الأسلوب تزيد فرصة الاقتراب من الحل الأمثل لأن البحث يكون بمجالات عديدة ويتم اكتشاف حلول عديدة ولكن مشكلة هذا الأسلوب أن التقارب يكون بطيئاً جداً وتسمى *local best PSO* [17]. الشكل رقم (11) يمثل المخطط الانسيابي لخوارزمية سرب الطيور المصممة لكشف التطفل.

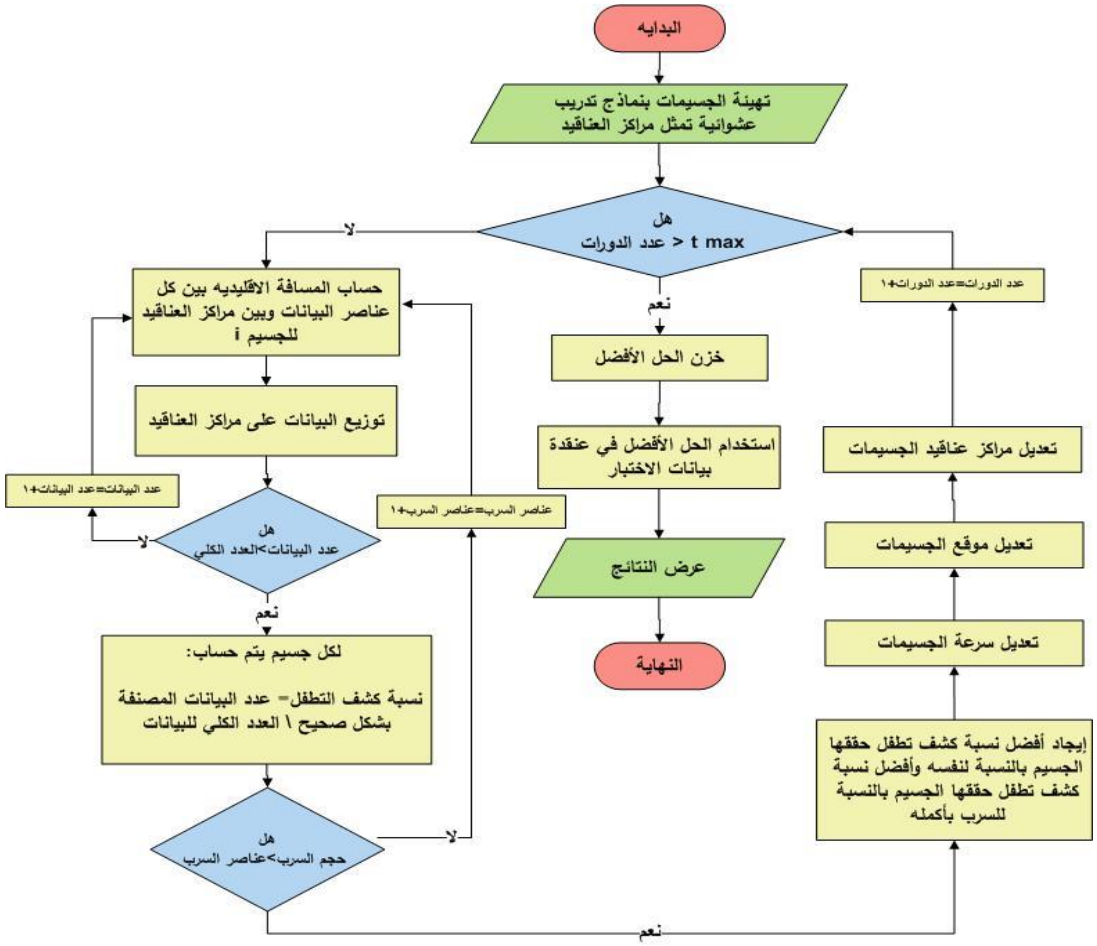


الشكل رقم (11). يمثل المخطط الانسيابي لخوارزمية سرب الطيور

5.3 خوارزمية سرب الطيور المطورة *MPSO*

بعد دراسة خوارزمية السرب بأنواعها الـ *local best PSO* و الـ*global best PSO* تم التوصل إلى خوارزمية سرب طيور جديدة خاصة بنظام كشف التطفل حصراً تقضي على مساوئ الـ *local best PSO* بحيث يكون التقارب للحل الأمثل سريعاً جداً بالإضافة إلى التخلص من مشكلة الوقوع في النهاية المحلية *local optima* التي تعاني منها الـ*global best PSO* . إذ تم استبدال دالة اللياقة في الخوارزمية الأصلية المعادلة (10) بمقياس كشف التطفل *detection rate* الذي يتم قياسهُ بالمعادلة رقم (1) في كل دورة من دورات الخوارزمية يتم البحث عن الحل الأفضل في السرب إذ أن الطير الرقمي الذي يمتلك نسبة كشف تطفل أعلى من البقية يعد الأفضل في السرب والأقرب إلى الحل الأمثل وبناء على ذلك تقوم جميع الطيور الرقمية في السرب بتغيير مواقعها وسرعاتها حسب المعادلات رقم (7,8) نسبة إلى الطير الأفضل لكي يقترب جميع السرب من الحل الأفضل. ففي بداية دورات مرحلة التدريب لوحظ تباين كبير بنسب كشف التطفل لدى الطيور الرقمية وبعد الوصول للدورة التاسعة أو العاشرة لوحظ تقارب كبير بنسب كشف التطفل لدى الطيور الرقمية حيث أصبحت جميع الطيور تقريباً تمتلك

نسبة كشف تطفل 99% و 100% وفي النهاية تم اختيار أفضل طير رقمي صاحب نسبة الـ 100% الذي يعد هو الحل الأفضل لمشكلة كشف التطفل.الشكل رقم (12) يمثل المخطط الانسيابي لخوارزمية سرب الطيور المطورة الخاصة بنظام كشف التطفل.



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

4. التنفيذ والاختبار *Implementation and Testing*

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

نوع الحزمة	عددها في الملف	نسبتها في الملف
normal	2000	%40
abnormal	3000	%60
العدد الكلي	5000	%100

عددها في الملف	نسبتها في الملف
1500	%37,5
2500	%62,5
4000	%100

أ)جدول محتويات ملف التدريب لكشف الهجمات ب) جدول محتويات ملف الاختبار لكشف الهجمات

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

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

(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

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

وأيضا تم تدريب خوارزميتي سرب الطيور وسرب الطيور المطورة على 5000 عينة من بيانات الـ*KDD99* واختبارها على 4000 ، ولُوحِظَ 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:

(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)

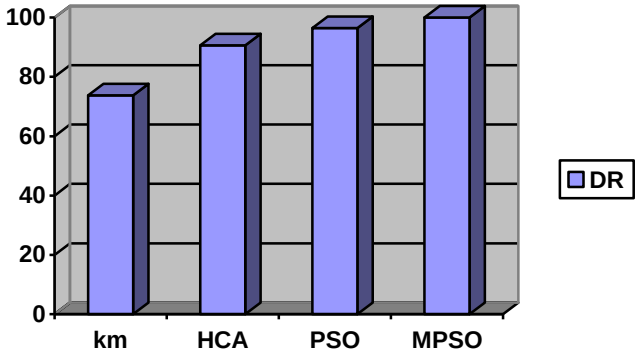
PSO	MPSO
%96,37	%100
145	0
0	0
2645	1500
1355	2500
%0	%0
%9,67	%0
%90,33	%100
%100	%100
0,948	1
0,965	1

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

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

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

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



الشكل رقم (13). يمثل مخطط المقارنة للطرق الأربعة حسب مقياس كشف التطفل

5. الاستنتاجات

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

- بالنسبة لطريقة العنقدة التقليدية المتمثلة باستخدام خوارزمية *Kmeans* أعطت نتائج مقبولة إلى 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 إذ دخل مبدأ الـ*KM* في حساب دالة اللياقة في الخوارزمية الجينية, إضافة 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 لأجراء 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 *local optima*.

- وأخيرا تم استخدام خوارزمية سرب الطيور المطورة *MPSO* التي قضت على مساوئ الـ *local best PSO* بحيث يكون التقارب للحل الأمثل سريع جداً بالإضافة إلى التخلص من مشكلة الوقوع في النهاية المحلية *local optima* التي تعاني منها الـ*global best PSO* . إذ تم استبدال دالة اللياقة في خوارزمية سرب الطيور بمقياس كشف التطفل *detection rate* الذي إذ المقياس الأهم في أنظمة كشف التطفل. وأعطت هذه الطريقة أفضل النتائج بجميع المقاييس المستخدمة حيث وصلت نسبة الكشف إلى 100% ولم تؤدِ للوقوع في مشكلة النهاية المحلية.

المصادر

[1] Wei Li " Using Genetic Algorithm for Network Intrusion Detection" Department of Computer Science and Engineering Mississippi State University, Mississippi State, MS 39762 Email: wli@cse.msstate.edu.

[2] Rozenblum D. , 2001 , "understanding Intrusion Detection System", SANS Institute.

[3] Anderson R. J. , 2008 , " Security Engineering: A Guide to Building Dependable Distributed Systems", Second Edition.

[4] Dr. Herrero A. and Prof.Dr. Corchado E. , 2011 "Mobile Hybrid Intrusion Detection" Springer.

[5] Topark M. ,2009, "Intrusion Detection System Alert Correlation with Operating System Level Logs" A Thesis Submitted to The Graduate School of Engineering and Sciences of İzmir Institute of Technology .

[6] http://www.windowsecurity.com

[7] Cramer L.M , Cannady J. and Harrell J. , 1996, "New Methods of Intrusion Detection using Control- Loop Measurement" , Georgia Institute of Technology Atlanta .

[8] Wilson J. , Dudley P. and Khan B. , 2008, "Requirements Specification " , CQF-QMT .

[9] Odutola K. , Oguntimehin A. , Tolke L. and Wulp M. V. , 2008 , "ArgoUML Quick Guide " .

[10] Kang D. , Fuller D. and Honavar V. ,2005, "Learning Classifiers for Misuse and Anomaly Detection Using a Bag of System Calls Representation" , IEEE .

[11] Norsyafawati F. , Norwawi N. and Seman K. , 2011, "Identifying False Alarm Rates for Intrusion Detection System with Data Mining" , IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.4.

[12] Hammersland R. , 2007, "ROC in Assessing IDS Quality" , Norwegian Information Security Lab, Gjøvik University College.

[13] Chang R, Lai L. , Wang J. and Kouh J. , 2007, "Intrusion Detection by Backpropagation Neural Networks with Sample-Query and Attribute-Query", International Journal of Computational Intelligence Research. Vol.3, No. 1 .

[14] Adetunmbi A., Adeola S. and Daramola O. ,2010 ,"Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features" , Proceedings of the World Congress on Engineering and Computer Science Vol I .

[15] Vescan A. and HORIA F. , 2008, "Constraint Optimization-Based Component Selection problem" , INFORMATICA,Volume LIII, No. 2.

[16] Izenman A. J. ,2008 ,"Modern Multivariate Statistical Techniques", Springer .

[17] Abraham A. , Grosan C and Ramos V. ,2006 , "Swarm Intelligence in Data Mining", Springer.

[18] Vinu V. ,Thomas G. and Lumban F. , 2011,"Information Technology and Mobile Communication" Springer , International Conference.

[19] Shonkwiler R. W. and Mendivil F. ,2009 , "Explorations in Monte Carlo Methods",Springer .