مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(2) 2022

Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (2) 2022

كشف وتصنيف الهجمات على قواعد البيانات الكبيرة باستخدام الشبكات العصبونية

د. راغب طعمه *

ريم مالك ابراهيم **

(تاريخ الإيداع 11 /8 /2021 . قُبِل للنشر في 6 /3 /2022)

□ ملخّص□

يعتبر إبقاء البيانات الضخمة أمنة من أهم التحديات وخاصة بعد تزايد حجم البيانات بشكل كبير وسريع حيث هذا التزايد يؤدي لدوث مشاكل كبيرة فيما يخص الامن والخصوصية.إذ جعلت التغيرات المستمرة والتنوع الكبير للبيانات عملية اكتشاف السلوكيات غير الطبيعية داخل البيانات وتحليلها بطريقة أمنة باستخدام تقنيات تحليل البيانات التقليدية أمر صعب ومعقد. جاءت تقنيات تعلم الالة للمساعدة في تحليل الكم الهائل من البيانات والمساعدة في اكتشاف الاختراقات ضمن النظام .في هذاالبحث تم الاعتماد على تقنية الشبكات العصبونية في تصميم مصنف قادر على كشف الاختراقات الأمنية في البيانات الضخمة وتحديد نوع هذه الاختراقات. يناقش البحث الاختيار الأمثل لهيكلية الشبكة المقترحة من حيث تحديد العدد الأمثل للطبقات الخفية والعدد الأمثل للعصبونات الموجودة فيها ،بتقييم قيم متوسط مربع الخطأ الناتج بعد كل عملية تدريب للشبكة العصبونية كما أنه تم اخضاع قاعدة البيانات الضخمة الى عملية معالجة أولية وأظهرت النتائج المنجزة في بيئة الماتلاب الأداء الأفضل للمصنف المقترح المرتكز على نموذج الشبكة العصبونية.

الكلمات المفتاحية: الاختراقات الامنية،البيانات الكبيرة، الشبكات العصبونية، المعالجة الأولية، تخفيض السمات، مصفوفة الدقة.

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

** طالبة ماجستير في قسم هندسة تكنولوجيا المعلومات – كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس– سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(2) 2022

Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (2) 2022

Detection and Classification of Attacks on Big Databases using Neural Networks

Dr.Ragheb Toemeh *
Reem Malek Ibrahem**

(Received 11/ 8/ 2021 . Accepted 6/ 3/ 2022)

□ ABSTRACT □

Keeping big data safe is one of the most important challenges, especially after the large and rapid increase in the volume of data, as this increase leads to many problems in terms of security and privacy. The constant changes and great diversity of data have made the process of detecting and analyzing abnormal behaviors within the data and analyze it in a secure way using traditional data analysis techniques difficult and complex. Machine learning techniques came to help analyze the huge amount of data and help discover breaches within the system. In this research, neural networks technology was relied on to design a classifier capable of detecting security breaches in big data and determining the type of these breaches. The research discusses the optimal choice of the proposed network architecture in terms of determining the optimal number of hidden layers and the optimal number of neurons in them, by evaluating the values of the average error squared produced after each training process for the neural network. In addition, the huge database was subjected to a preliminary processing process, and the results achieved in the Matlab environment showed the performance Best for the proposed classifier based on the neural network model.

Key Words : Security breaches, big data, neural networks, preprocessing, attribute reduction, accuracy matrix

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

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

1. المقدمة

تطورت الطرق المتبعة في عملية كشف الاختراقات الأمنية في الشبكات وقواعد البيانات الضخمة مع تطور تقنيات تعلم الألة، كما أن الزيادة السريعة في تقنيات الانترنت والأجهزة الذكية زاد عدد ونوع الهجومات على البيانات، ولمواكبة هذا التطور في أنواع الهجومات كان لابد من اجراء العديد من الأبحاث مستخدمة العديد من التقنيات التقليدية والمحسنة بهدف حماية هذه البيانات وكشف الاختراقات بدقة وسرعة.

استخدمت العديد من الأبحاث الأنظمة الخبيرة القائمة على القواعد (Rule Based Expert System) في عملية كشف الاختراقات لكن أظهرت هذه الأنظمة نتائج غير دقيقة عند استخدامها مع البيانات الضخمة لذلك تم استخدام تقنيات تعلم الالة في عملية كشف الهجومات في بيئة البيانات الضخمة، يمكن لتعلم الالة اكتشاف الارتباط بين الميزات والفئات الموجودة في بيانات التدريب عن طريق استخلاص الميزات وتقليل الابعاد ثم استخدام نتيجة هذه الطرق من اجل بناء نموذج قادر على التصنيف والتنبؤ [1].

تعد الشبكات العصبونية من أحدث التقنيات المستخدمة في مجال كشف الاختراقات الأمنية لما تبديه من دقة وسرعة وسهولة في عملية كشف الهجومات ،كما انها مناسبة من أجل عملية كشف الهجمات على البيانات الضخمة. مازالت الأبحاث جارية في مجال الشبكات العصبونية بهدف الحصول على أفضل بنية للشبكة العصبونية تساهم في عملية تصنيف الهجومات بدقة عالية. قامت العديد من الأبحاث باستخدام تقنيات الذكاء الاصطناعي في عملية كشف الاختراقات على قاعدة البيانات الكبيرة KDD-99.

قام V.N.TIWARI[1] بمقارنة أداء مصنف Naïve Bayes و Random Forest على 10% من قاعدة البيانات KDD-99 فقط وأظهرت النتائج ان خوارزميةRandom Forest أكثر دقة في التصنيف. وقام S.Yendole[2]باقتراح تقنية هجينة لكشف الاختراقات بالاعتماد على SVM في التصنيف والخوارزمية الجينية Genetic Algorithm في عملية المعالجة الأولية لقاعدة البيانات وتم تطبيقها على 10% من قاعدة البيانات KDD-99.

اقترحP.Natesan et al[3]خوارزمية محسنة لاستخلاص الميزات بالاعتماد على خوارزمية Bat Binary وتم استخدام مصنف Naïve Bayes لكشف الاختراقات تم تطبيق الخوارزمية على 10% من قاعدة البيانات KDD-99.

قام H.Wang et al[4] باقتراح خوارزمية تصنيف تفرعيه تعتمد على خوارزمية SVM بالاعتماد على التعلم المتكاملعلى 10% من قاعدة البيانات KDD-99. وقارنSH.Rezvy et al[5] آداء الشبكات العصبونية ومصنف SVM في عملية كشف الاختراقات على 151063 سجل من قاعدة البيانات NSL حيث أعطت الشبكات العصبونية الآداء الأفضل بالتصنيف بنسب دقة جيدة.

كما قام B.Ingre وA.Yadav [6] بتحليل أداء الشبكات العصبونية على 125973سجل من قاعدة البيانات NSLاذ تم اعتماد شبكة عصبونية بطبقتين خفيتين وبعدد عصبونات21، 2 لكل طبقة على الترتيب كما تم تصنيف البيانات ثنائيا الى هجوم او سجل عادي دون تحديد نوع الهجوم.

2. أهمية البحث وأهدافه

يقدم البحث مساهمة جديدة في مجال كشف الاختراقات الأمنية في البيانات من حيث تصميم مصنف معتمد على الشبكات العصبونية متعددة الطبقات ذات التغذية الامامية المدربة باستخدام خوارزمية الانتشار العكسي للخطأ وذلك بهدف اكتشاف السلوكيات غير الطبيعية في قواعد البيانات الكبيرة وتحديد نوع هذه السلوكيات، وتكمن أهمية هذا البحث في تحديد أفضل هيكلية للشبكة العصبونية المقترحة من حيث تحديدالعدد الأمثل للطبقات الخفية والعدد الأمثل للعصبونات الخفية الموجودة فيها بتقييم متوسط مربع الخطأ الناتج بعد كل عملية تدريب للشبكة العصبونية كما تبرز أهمية البحث بشكل خاص في اختبار الشبكة العصبونية المقترحة على 50% من قاعدة البيانات KDD-99 الضخمة وهو حجم جديد لم يتم اختبار الشبكة العصبونية عليه من قبل .يهدف البحث الى تقديم تقنية آمنة لتصنيف البيانات الكبيرة الى سلوكيات طبيعية وأخرى غير طبيعية مع تحديد نوع السلوك غير الطبيعي وبالتالي زيادة في أمان هذه البيانات وحمايتها من الاختراقات بالإضافة الى الحصول على تصنيف دقيق لهذه البيانات وبأقل وقت من خلال القيام بالمعالجة الأولية لقاعدة البيانات الضخمة .

3. طرائق البحث ومواده

يعتمد البحث على منظومتين أساسيتين هما قواعد البيانات والشبكات العصبونية، تم الاعتماد في هذا البحث على قاعدة البيانات الضخمة المشهورة KDD99 ، وهي عبارة عن مجموعة بيانات قياسية تم توليدها عن طريق محاكاة بيئة ضمن شبكة بيانات ضخمة، وبالتالي تضم بيانات طبيعية وغير طبيعية ، تم معالجتها معالجة مسبقة قبل استخدامها .

تم استخدام الشبكات العصبونية في عملية تصنيف البيانات، إذ تعد الشبكات العصبونية من أهم التقنيات المستخدمة في التصنيف نظراً للنتائج الدقيقة التي تحققها في عملية التصنيف وبوقت قصير.

تم تقييم التصنيف وأداء المصنف باستخدام مصفوفة الدقة لكونها من أهم الطرق المستخدمة في عملية تقييم الآداء، تم استخدام برنامج الماتلاب في عملية المعالجة الأولية و عملية التصنيف.

3-1-قاعدة البيانات Dataset (KDD-99)

قاممخبر لينكولن في معهد ماساتشوستس للتكنولوجيا في الولايات المتحدة بتوليد بيانات حركة مرور شبكة قياسية لتقييم أنظمة كشف الاختراقات الشبكية [7] .

استمرت عملية المحاكاة تسعة أسابيع تم جمعها من شبكة LAN محلية، نتج عن هذه المحاكاة كاملة حوالي خمسة ملايين سجل اتصال.في عام 1999 اعترفت KDD (Knowledge Discovery and Data mining) منظمة التنقيب عن البيانات والكشف المعرفي والتي تعد المنظمة الاحترافية الأكثر شعبية لمنقبي البيانات ووافقت على بيانات DARPA بأن تكون علامة تقليدية من أجل أنظمة كشف الاختراق وتمت تسميتها KDD99 أو KDDCup99[9][8] . تضم KDD 41 سمة بالإضافة الى سمة ال class التي تعبر عن سجل الاتصال هل هو سجل طبيعي أم سجل هجوم، كما تندرج هذه السجلات تحت أربع أنواع رئيسة من الهجومات تندرج تحتها 22 نوع هجوم فرعي .

تمتاز قاعدة البيانات المستخدمة بما يلي[10] :

1- توافريتها حيث يوجد العديد من قواعد البيانات مجهولة المصدر ويعود ذلك بمخاطر أمنية محتملة للنظام وبالتالي عدم إمكانية الوصول إليها من قبل الباحثين .

2- سمات حركة المرور المفصلة والتي تفتقر اليها قواعد بيانات أخرى مثل LBNL، والتي تساهم بشكل كبير في عملية التصنيف، وتعد سمات حركة المرور من أهم الركائز التي سنعتمد عليها في هذا البحث.

3- استخدامها الكبير من قبل الباحثين في عملية كشف الاختراق مما يتيح الفرصة للاطلاع على نتائج الأبحاث ومقارنة النتائج.

3-1-1 سمات قاعدة البيانات

تتضمن بيانات KDD عدة أنواع من السمات[11] :

1- السمات الأساسية للاتصال: توفر معلومات لأغراض تحليل حركة مرور الشبكة العامة مثل :مدة الاتصال، نوع البروتوكول،عدد البايتات المنقولة والعلم الذي يشير الى الحالة الطبيعية او حالة خطأ الاتصال .

2- سمات نفس المضيف same host : تفحص الاتصالات التي مدتها ثانيتين لنفس المضيف وتحسب الاحصائيات المتعلقة بسلوك البروتوكول، الخدمة،الخ.

3- سمات نفس الخدمة same service : تفحص الاتصالات التي مدتها ثانيتين ولها نفس الخدمة،وتسمى سمات نفس المضيف ونفس الخدمة مع بعضها السمات المعتمدة على الزمن Time-based feature .

4- السمات المعتمدة على المضيف Host-based Features : والتي تفحص مضيفين او اتصالات مدتها أكثر من ثانيتين مثلا خلال دقيقة.

5- سمات المحتوى Content Features : سمات تبحث عن السلوك المشبوه في أجزاء البيانات مثل عدد محاولات تسجيل الدخول الفاشلة.

يبين الشكل (1) سجل يعبر عن حركة غير طبيعية(هجوم):

1،tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,229,10,0.00,0.00,1.00,1.00,0.04,0.06,0.00,255,10,0.04,0.06,0.00,0.00,0.00,0.00,1.00,1.00,**Neptune**

الشكل (1) :سجل حركة غير طبيعي (هجوم)

يبين الشكل (2) سجل يعبر عن حركة طبيعية:

1،udp,http,SF,54,51,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,0.83,0.00,0.00,0.00,0.00,0.00,**normal**

الشكل (2) سجل حركة طبيعية

3-1-2 أنواع الهجومات المستخدمة في قاعدة البيانات

تتضمن قاعدة البيانات أربعة أنواع رئيسية من الهجومات يندرج تحت كل نوع من هذه الأنواع هجومات فرعية [12]:

1- هجوم حجب الخدمة Denial of service(DOS) : محاولة جعل الجهاز غير متاح لمستخدميه.

2- هجوم مستخدم الى جذر User to Root(U2R) :استغلال المهاجم للنظام الذي يبدأ بحساب طبيعي يحاول من خلاله الحصول على امتيازات مستخدم رئيسي

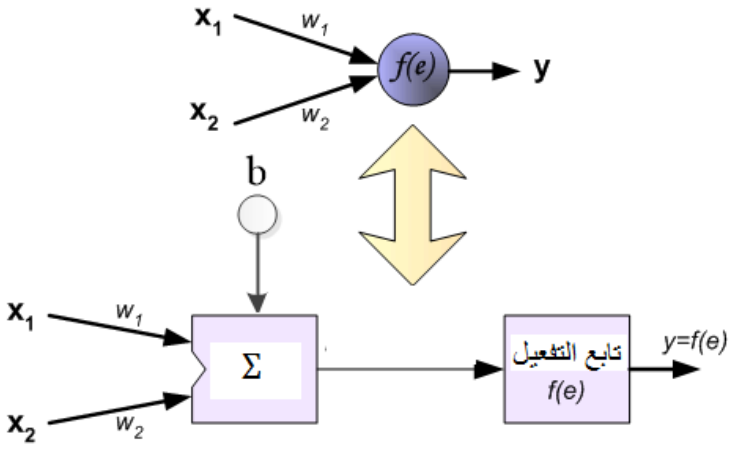
3- هجوم بعيد الى محلي Remote to Local (R2L) :يستغل المهاجم ميزات جهاز محلي من خلال ارسال حزم عبر الانترنت الى جهاز لا يملك الوصول اليه بهدف استغلال نقاط ضعف الأجهزة .

4- هجوم التحقق Prpbe :الهدف منه تعريض النظام للخطر من خلال قيام المهاجم بمسح جهاز شبكي لتحديد نقاط الضعف من اجل استغلالها فيما بعد.

3-2-الشبكات العصبونية

تعرف الشبكة العصبونية الصنعية (Artficial Neural Network) بأنها عبارة عن نظام لمعالجة البيانات بشكل يحاكي ويشابه الطريقة التي تقوم بها الشبكات العصبونية الطبيعية للإنسان حيث تتشابه الشبكة العصبونية مع الدماغ البشري في انها تكتسب المعرفة بالتدريب، وتخزن هذه المعرفة باستخدام قوى وصل داخل العصبونات تسمى الاوزان التشابكية. تتكون الشبكة العصبونية بشكل عام من طبقة الدخل،طبقة خرج،طبقة خفية والتي تتواجد بين طبقة الدخل والخرج.

يعتمد هذا البحث على استخدام شبكة عصبونية صنعية ذات تغذية امامية، ومن الأمثلة على هذا النوع من الشبكات،الشبكة العصبية ذات الانتشار الخلفي للخطأ (Error Back Propagation) يبين الشكل (3) بنية العصبون الصنعي.



الشكل (3):بنية العصبون الصنعي

يتضمن كل عصبون في الطبقة الخفية وطبقة الخرج،وحدة معالجة تقوم بتحديد خرج العصبون من خلال تطبيق عمليتين حسابيتين متتاليتين كما هو موضح بالشكل(3)تتضمن العملية الأولى،تحديد ناتج مجموع جداء متغيرات دخل الطبقة السابقة بالاوزان Wيضاف اليها قيم الانزياحb ليتم في العملية الثانية تحديد خرج العصبون Y وذلك باستخدام ناتج العملية الأولى كمتغير لتابع التفعيل f() .

3-3 الخوارزمية المقترحة:

تم في هذا البحث بناء مصنف ثنائي المراحل قادر على تمييز سجلات قاعدة البيانات الطبيعية من السجلات غير الطبيعية بالإضافة الى تحديد نوع الهجوم بدقة من خلال تحديد النوع الفرعي للهجوم والذي لم يتم التطرق له في الدراسات السابقة، اذ تم استخدام قاعدة بيانات ضخمة بحجم جديد لم يتم التطرق له في الدراسات السابقة أيضا، وقبل عملية التصنيف تم تطبيق عملية معالجة أولية لقاعدة البيانات بعدة خطوات الهدف منها زيادة دقة التصنيف وتقليل زمن التصنيف.

الخطوات المتبعة:

1- **المعالجة الأولية للبيانات Data Preprocessing :**
2- **تقسيم قاعدة البيانات الناتجة عن المرحلة السابقة الى بيانات تدريب وبيانات اختبار .**
3- **بناء المصنف باستخدام الشبكات العصبونية .**
4- **تدريب المصنف باستخدام بيانات التدريب السابقة.**
5- **اختبار المصنف باستخدام بيانات الاختبار .**
6- **تقييم آداء المصنف.**

شرح خطوات الخوارزمية :

لأهمية الخطوة الأولى في الخوارزمية المقترحة سنقوم بشرحها بالتفصيل لتبيان ماهو الجديد والخطوات التالية سيتم شرحها بشكل عملي في قسم النتائج.

1- **المعالجة الأولية للبيانات Data PreProcessing :**

تفيد المعالجة الأولية في إزالة البيانات المكررة والبيانات غير الكاملة وتحويل البيانات الى شكل موحد، تعد هذه المرحلة ضرورة أساسية قبل القيام بعملية تدريب الشبكة العصبونية الهدف منها تقليل زمن المعالجة وزمن التصنيف بالإضافة الى زيادة دقة التصنيف كون البيانات المكررة والسمات غير الهامة غالبا ماتشوش على خوارزمية التصنيف.

تضم مرحلة المعالجة الأولية :

a) **حذف السجلات المكررة :** تعتبر السجلات المكررة من قيود تصنيف سجلات قاعدة البيانات KDD،إذ أن القيام بهذه الخطوة يضمن عدم تحيز خوارزمية التعليم للسجلات المكررة و الذي يؤدي الى نتائج غير صحيحة وبالتالي تحسين دقة الكشف بالإضافة الى التخفي ضمن متطلبات مساحة التخزين كوننا نتعامل مع قاعدة بيانات ضخمة بحجم ضخم وبالتالي تقليل زمن المعالجة.

b) **الترميز الرقمي للبيانات :** يتم تحويل أنواع الهجومات الى ترميز رقمي يفيد في الحصول على نتائج التصنيف بشكل صحيح ، بالإضافة الى تحويل القيم غير الرقمية الى قيم رقمية كون الشبكة العصبونية لا تقبل إلا دخل بقيم رقمية.

c) تخفيض السمات :يمكن للسمات الإضافية أن تزيد من زمن الحساب كما يمكنها التأثير على دقة التصنيف و بالتالي عملية تخفيض السمات عملية ضرورية في مرحلة المعالجة الأولية ، لايوجد أي تصميم او تابع يعبر عن العلاقة بين السمات وبالتالي سنعتمد في هذا البحث على علاقات رياضية تساعد في حساب الترابط بين السمات وإيجاد السمات غير المفيدة ، تضم هذه المرحلة خطوتين أساسيتين هما :

1) **إيجاد السمات غير المفيدة:** من خلال حساب التباين ، بهدف الاستغناء عن السمات ذات التباينات الصفرية أي السمات التي لم تبدأ يتغير على طول قاعدة البيانات ومن أجل كل السجلات.

العلاقة الرياضية التي تعبر عن التباين:

$$\boldsymbol{Cov(x) = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{xi} - \mathrm{X}) \ldots\ldots\ldots (1)}$$

حيث:X :المتوسط الحسابي لقيم x_i ، n: عدد العينات

2) **إيجاد السمات المرتبطة** : من خلال حساب قيمة معامل الترابط لكل زوج من السمات ، تتراوح قيم معامل الارتباط بين (-1) و (1) و كلما كانت قيمته اقرب ل (1) كلما كان ارتباط زوج السمات أعلى ، وبالتالي الهدف من هذه الخطوة الاستغناء عن سمة واحدة من كل زوج من السمات ذات معامل الترابط الأعلى لان وجود احداها يعوض عن وجود الأخرى.

العلاقةالرياضيةالتيتعبرعنمعاملالترابط:

$$\boldsymbol{Corr(x,y) = \frac{Cov(x,y)}{\sigma x.\sigma y} \ldots\ldots\ldots\ldots (2)}$$

حيث : σx:الانحراف المعياري ل x ، σy:الانحراف المع ياري ل y ، cov(x,y) :التباين

3-4 مصفوفة الدقة

تعد مصفوفة الدقة من اهم الوسائل المستخدمة في تقييم المصنفات،إذ يتم تقييم المصنف من خلال قدرته على التصنيف الصحيح أي القدرة على تحديد الصنف الذي ينتمي اليه سجل الاتصال طبيعي أم هجوم، وعند مقارنة نتيجة التصنيف مع الواقع الفعلي نجد أربع حالات مختلفة [12]:

الإيجابيات الصحيحة(TP)True Positive : الحدث إيجابي وتم التنبؤ ان الحدث ايجابي . (صحيح)

الإيجابيات الخاطئة(FP)False Positive :الحدث سلبي وتم التنبؤ ان الحدث ايجابي . (خطأ)

السلبيات الخاطئة (FN)False Negative :الحدث إيجابي و تم التنبؤ ان الحدث سلبي . (خطأ)

السلبيات الصحيحة(TN)True Negative:الحدث سلبي وتم التنبؤ ان الحدث سلبي . (الشواذ)

من خلالها يتم حساب القيمة التالية:

الدقة Accuracy: المقياس الأكثر شيوعا لتقييم المصنف، يقيم كامل الخوارزمية يعطى بالعلاقة التالية[13]:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

4.النتائج والمناقشة

يوضح هذا الجزء من البحث نتائج عملية المعالجة الأولية لقاعدة البيانات الضخمة بالتفصيل ثم نتائج عملية التصنيف باستخدام الشبكات العصبونية،حيث تم تنفيذ العمل باستخدام البيئة البرمجية MATLAB(R2018b) كما يلي :

4-1 حذف السجلات المكررة :

نلاحظ من الجدول التالي أن نسبة تخفيض عدد السجلات كانت بمقدار 1100623*100/2333598=47.16%

جدول (1):عدد السجلات قبل عملية الحذف وبعدها

Class	عدد السجلات قبل الحذف	عدد السجلات بعد الحذف
Normal	1,199,470	627,873
DOS	891,658	320,325
Probe	140,150	90,475
R2L	100,385	60,098
U2R	1935	1852
المجموع	2,333,598	1,100,623

4-2 الترميز الرقمي لأنواع الهجومات :

من اجل مرحلة التصنيف الأولى نرمز للهجومات الرئيسية الأربعة بقيم من 1 الى 4 والحركة الطبيعية نرمزها بالقيمة 0 كما يلي : Normal→ 0 , U2R→ 4 , R2L→ 3 , Probe→2 , DOS→ 1

من أجل مرحلة التصنيف الثانية لتحديد النوع الفرعي من النوع الرئيسي يتم الترميز كما هو موضح بالجدول التالي:

هجوم DOS (1)		هجوم Probe(2)		هجوم R2L(3)		هجوم U2R(4)	
الهجوم الفرعي	الترميز	الهجوم الفرعي	الترميز	الهجوم الفرعي	الترميز	الهجوم الفرعي	الترميز
Smurf	11	Satan	21	Warezclient	31	Buffer_overflow	41
Neptune	12	Ipsweep	22	Guess_passwd	32	Rootkit	42
Back	13	Portsweep	23	Warezmaster	33	Loadmodule	43
Teardrop	14	Nmap	24	Imap	34	Perl	44
Pod	15			Ftp_write	35		
Land	16			Multihop	36		
				Phf	37		
				Spy	38		

جدول (2):ترميز أنواع الهجومات

4-3 تخفيض السمات :

4-3-1 حساب التباين : يتم بالاعتماد على العلاقة (1) وفي بيئة الماتلاب من خلال التعليمة البرمجية var(i) حيث i هي العمود المعبر عن السمة المختارة في كل مرة،كانت النتيجة كما هو موضح بالجدول التالي:

جدول (3): حساب التباين من أجل كل سمة

Var(1)=1.98	Var(5)= 2.235	Var(6)=4.5026	Var(7)=3.1042
Var(8)=0.0203	Var(9)=0.0013	Var(10)= 0.8620	Var(11)=0.0226
Var(12)=0.2467	Var(13)=52.8470	Var(14)=0.0024	Var(15)=4.4353
Var(16)=64.6676	Var(17)= 0.4581	Var(18)=0.0023	Var(19)=0.0046
Var(20)=0	**Var(21)=0**	Var(22)=0.0276	Var(23)=1.6522
Var(24)=7.9321	Var(25)=0.0872	Var(26)=0.0890	Var(27)=0.1732
Var(28)=0.1732	Var(29)=0.1702	Var(30)=0.0672	Var(31)=0.0643
Var(32)=8.84296	Var(33)=1.2496	Var(34)=0.1898	Var(35)=0.0487
Var(36)=0.938	Var(37)=0.0073	Var(38)=0.0746	Var(39)=0.0794
Var(40)=0.1499	Var(41)=0.1607		

مما سبق نجد أن السمتين 20 و 21 لها تباين صفري أي لاتتغير قيمتها من أجل كل السجلات في قاعدة البيانات وبالتالي يمكن الاستغناء عنهما مما يقلل من حجم قاعدة البيانات دون التأثير على دقة التصنيف،وأصبح عدد السمات بعد هذه المرحلة 39سمة.

السمة (20) هي Num_outbound_cmds : عدد الأوامر الصادرة في جلسة بروتوكول ftp .

السمة (21) هي Is_hot_login: تأخذ قيمة 1 اذا تم تسجيل الدخول الى hot list و0 فيما عدا ذلك .

كلا السمتين قيمتهما صفر على طول قاعدة البيانات ومن اجل كل السجلات وبالتالي حذفهما لا يؤثر على عملية التصنيف.

4-3-2 حساب قيمة معامل الترابط بين كل زوج من السمات : بالاعتماد على العلاقة (2) ومن خلال التعليمة البرمجية corrcoef(a,b)في بيئة الماتلاب حيث a وb هما السمتان المراد حساب قيمة معامل الترابط بينهما، فيما يلي جدول يوضح السمات الأكثر ترابطا أي السمات التي كان معامل الترابط لها أكبر من 0.9 :

جدول (4): قيم معامل الترابط من اجل السمات الأكثر ترابط

Corrcoef(25,26)=0.9664	Corrcoef(12,16)=0.9006	Corrcoef(13,16)=0.9960
Corrcoef(26,38)=0.9997	Corrcoef(25,39)=0.9008	Corrcoef(25,38)=0.9041
Corrcoef(27,40)=0.9852	Corrcoef(27,28)=0.9755	Corrcoef(26,39)=0.9210
Corrcoef(28,41)=0.9478	Corrcoef(28,40)=0.98	Corrcoef(27,41)=0.9310
Corrcoef(40,41)=0.9047	Corrcoef(38,39)=0.9450	Corrcoef(33,34)=0.9044

من الجدول السابق نجد أن السمة 16 من سمات قاعدة البيانات مرتبطة بشكل كبير مع السمتان13,12وبالتالي يمكن الاستغناء عن السمتين 12 و13 والإبقاء على السمة 16 ، ونجد نفس الأمر من أجل السمة 25التي تحل مكان السمات 26,38,39، والسمة 28 تحل مكان السمات 27,40,41 والسمة 33 تحل مكان السمة34 .

مثلا السمة 25 هي النسبة المئوية لعدد مرات إعادة الخطأ مرتبطة بشكل كبير مع السمة 26 التي تعبر عن النسبة المئوية لعدد الاتصالات لنفس الخدمة ونفس المضيف أي وجود قيمة لأحد هذه السمتان يغني عن وجود السمة الأخرى بمعرفة عدد مرات حصول فشل بالاتصال يمكننا من معرفة عدد مرات الاتصال الناجح، ونفس الامر ينطبق على كل السمات المترابطة.

إذاً : نبقي على السمات الأربعة16,25,28,33ونحذف السمات التسعة المتبقية 13,12,26,38,39,27,40,41

وبالتالي يصبح عدد السمات بعد هذه المرحلة هو 30سمة.

4-3-3 بيانات التدريب وبيانات الاختبار:

عدد السجلات الكلية بعد عملية المعالجة الأولية هو 1,100,623سجل يتم أخذ 70% من هذه السجلات من اجل بيانات التدريب و30% من أجل بيانات الاختبار،يمكن أخذ أي نسبة بشرط أن تكون بيانات التدريب أكبر من بيانات الاختبار لأنها مستخدمة في عملية تدريب المصنف الذي سنختبره لاحقا على بيانات الاختبار.

ومنه عدد السجلات في قاعدة بيانات التدريب هو770,437سجل وعدد السجلات في قاعدة بيانات الاختبار 330,186 سجل.

4-3-4نمذجة الشبكة العصبونية:

تتم نمذجة الشبكة العصبونية الصنعية في بيئة ماتلاب بإدخال بيانات التدريب، تمثل هذه البيانات مدخلات الشبكة المتمثلة بشعاع السمات الخاص بسجل الاتصال (30 سمة) والقيم المرغوبة للخرج المتمثلة بنتيجة التصنيف لسجل الاتصال.

لاختيار عدد الطبقات المناسب وعدد العصبونات المناسبة قمنا بإجراء عدة تجارب على الشبكة العصبونية ومن أجل قاعدة البيانات السابقة، تم في كل تجربة اختيار عدد طبقات محدد وعدد عصبونات

محددة ،وفي كل مرة تم حساب معامل الخطأ،يوضح الجدول التالي التجارب التي تم اجراؤها مع قيمة معامل الخطأ لكل تجربة :

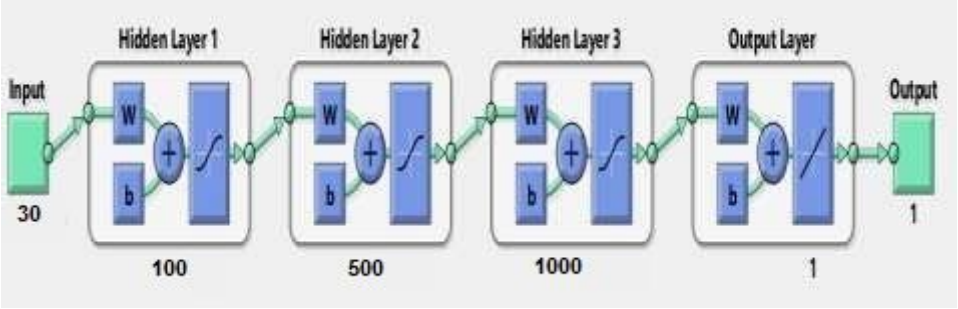
جدول (5):نتائج التجارب لاختيار بنية الشبكة العصبونية المناسبة

عدد الطبقات الخفية	عدد العصبونات في كل طبقة خفية	معامل الخطأ في مرحلة التدريب	معامل الخطأ في مرحلة الاختبار
طبقة واحدة خفية	n1=100	0.0067	0.0110
	n1=500	0.0026	0.0019
	n1=1000	0.0014	0.0015
طبقتين خفيتين	n1=100, n2=100	0.0045	0.0037
	n1=100, n2=500	0.0019	0.0026
	n1=500, n2=500	0.0018	0.0016
	n1=100, n2=1000	0.0008466	0.0010
	n1=500, n2=1000	0.0005356	0.0005
ثلاث طبقات خفية	n1=100, n2=100,n3=100	0.0009675	0.000851
	n1=100, n2=500,n3=500	0.0008987	0.000817
	n1=100, n2=100,n3=500	0.0006766	0.000494
	n1=100, n2=100,n3=1000	0.0004378	0.000484
	n1=100, n2=500,n3=1000	0.0003023	0.000482

بناء على ماسبق وباختيار عدة سيناريوهات لتدريب الشبكة العصبونية باستخدام تابع التدريب الممثل لخوارزمية الانحدار التدريجي للخطأ ذات معامل معدل التعلم المتغير القيمة، نجد أن هيكلية الشبكة العصبونية المختارة في البحث هي الموافقة للسيناريو الأخير والتي تمتلك القيم الأقل لمعامل الخطأ في مرحلة التدريب والاختبار كما هو مبين في الجدول (5) وبالتالي تمتلك التركيبة التالية:

- طبقة الدخل : مكونة من شعاع السمات الخاص بسجل الاتصال 30 سمة ،
- ثلاث طبقات خفية،عدد عصبونات الطبقة الخفية الأولى 100عصبون ،أما عدد عصبونات الطبقة الخفية الثانية هو500 ،و1000 عصبون في الطبقة الخفية الثالثة.
- طبقة الخرج نتيجة التصنيف لسجل الاتصال وهي بعصبون واحد.

يبين الجدول (5) ، ان الاستمرار بزيادة عدد الطبقات الخفية وزيادة عدد العصبونات فيها سيؤدي الى تحقيق قيمة أصغرية لمتوسط مربع الخطأ، ولكنه يزيد من حجم وتعقيد الشبكة،لذلك تم اعتبار ان السيناريو الأخير يحقق اختيار الهيكلية الأفضل للشبكة العصبونية المختارة في البحث، يبين الشكل (2) هيكلية الشبكة العصبونية النهائية المستخدمة في البحث.



الشكل (4):بنية الشبكة العصبونية النهائية المختارة

4-3-5 تقييم آداء المصنف

باعتماد السيناريو الأخير الذي تم التوصل اليه كمصنف للبيانات ضمن هذا البحث، تم اجراء عملية التصنيف على قاعدة البيانات الكبيرة الكلية والتي تحتوي1,100,623 سجل ، وكانت النتائج كما يلي مع اخذ بارامتر الدقة بعين الاعتبار :

جدول (6): نتائج دقة الشبكة المقترحة من اجل جميع الاصناف

	Dos	Probe	R2L	U2R	Normal
الدقة	98.54%	96.37%	92.80%	90.30%	98.5%

من الجدول السابق نجد ان الشبكة العصبونية المقترحة ومن أجل قاعدة بيانات كبيرة أظهرت نتائج جيدة من حيث الدقة بمعدلات خطأ صغيرة جدا .

يمكن مقارنة نتائج المصنف المقترح مع أنظمة أخرى تم اقتراحها في أبحاث سابقة استخدمت حجم مجموعة البيانات 10% من KDD99 لكن بالاعتماد على خوارزميات مختلفة. **لذلك تم اختبار الشبكة العصبونية على 10% فقط من قاعدة البيانات المستخدمةمع إعادة نفس الخطوات السابقة من اجل قاعدة البيانات الكلية.**

يبين الجدول(7) هذه المقارنة بين المصنف المقترح الذي استخدم خوارزمية الانتشار الخلفي للشبكات العصبونية والتقنيات المستخدمة في عدة دراسات سابقة من حيث معدلات كشف الأصناف الهجومية والصنف الطبيعي.

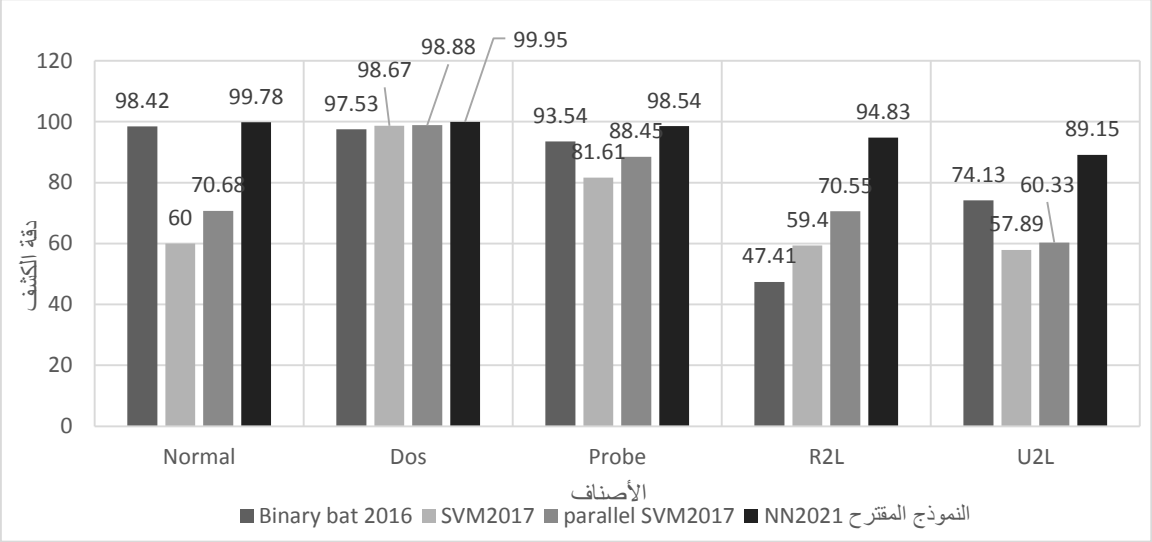
جدول (7):مقارنة المصنف المقترح مع نتائج دقة المصنفات في الدراسات السابقة

الرقم	اسم المصنف والسنة	دقة كشف كل هجوم				
		Dos	Probe	R2L	U2R	Normal
1	Binary bat 2016	97.53%	93.54%	47.41%	74.13%	98.42%
2	SVM2017	98.67%	81.62%	59.4%	57.89%	60%
3	Parallel SVM2017	98.88%	88.45%	70.55%	60.33%	70.68%
4	النموذج المقترح2021 NN	99.95%	98.54%	94.83%	89.15%	99.78%

من الجدول السابق نجد أن المصنف المقترح قام بتحسين دقة كشف الهجمات على قاعدة البيانات بمقدار 3%تقريبا من أجل الهجوم Dos وبمقدار 5% من أجل الهجزم probe بالمقارنة مع المصنف الأول و

بمقدار من 10 الى 19% بالمقارنة مع المصنفين الثاني والثالث، كما حسن المصنف الدقة من أجل الهجوم R2L بشكل كبير بنسبة تصل الى أكثر من 40% بالمقارنة مع المصنفات الثلاثة وبنسبة حوالي 20% من أجل الصنف U2R أما بالنسبة للسجلات الطبيعية كانت الدقة متقاربة بالمقارنة مع المصنف الأول وبنسبة تحسين حوالي 2% اما التحسين بالمقارنة مع الصنفين الآخرين وصلت الى 40% تقريبا، وهذا يدل على كفاءة بنية الشبكة العصبونية المستخدمة في عملية كشف الهجمات .

الشكل(5) مخطط توضيحيبينالمقارنة التي تمت في الجدول(7) بين النظام المقترح و ثلاث أنظمة أعطت أعلى دقة في التصنيف والتي استخدمت نفس مجموعة البيانات 10% من KDD–99



الشكل (5): مخطط بياني يوضح نتائج دقة الكشف من اجل المصنف المقترح ومصنفات الدراسات السابقة

5- الاستنتاجات والتوصيات:

تم في هذا البحث تصميم مصنف بالاعتماد على الشبكات العصبونية قادر على تصنيف سجلات البيانات الكبيرة وكشف الاختراقات فيها، من خلال دراستنا السابقة نجد:

- أعطت الشبكة العصبونية المقترحة افضل نتيجة من حيث دقة الكشف بالمقارنة مع المصنفات الأخرى يعود ذلك الى إجراء عدة تجارب لاختيار بنية الشبكة العصبونية المناسبة بأقل معدل خطأ
- حسنت خطوات المعالجة الأولية للبيانات دقة الكشف والتصنيف .
- الشبكة المقترحة قامت بتحديد النوع الدقيق للهجوم من خلال التعرف على الهجمات الفرعية المندرجة تحت الهجمات الرئيسية بالمقارنة مع المصنفات السابقة والتي اعتمدت على تحديد النوع الرئيسي فقط دون الفرعي للهجوم.

ومن المستحسن:

تطوير بنية الشبكة المستخدمة بحيث يتم زيادة عدد العصبونات الخفية وعدد الطبقات الخفية في الشبكة بهدف الحصول على افضل دقة، كما يمكن أن يتم تصميم المصنف باستخدام خوارزميات تصنيف أخرى غير الشبكات العصبونية، و استخدام خوارزميات مختلفة في مرحلة المعالجة الأولية تساهم في تقليل زمن التصنيف وزيادة دقة الكشف.

المراجع

[1] TIWARI, V. N. Mar-Apr 2016, *Enhanced Method for Intrusion Detection over KDD Cup 99 Dataset*. International Journal of Current Trends in Engineering & Technology, Vol. 02 No. 02.

[2] YENDOLE, S; et al. March 2017, *Identifying Intrusion Detection System using Hybrid technique with Support Vector Machine*. International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 5 , No. 3.

[3] Natesan P, et al. 2017, *Hadoop based parallel binary bat algorithm for network intrusion detection*. Int J Parallel Program, Vol. 45, No.5,1194–213.

[4] Wang H, Xiao Y, Long Y. 2017, *Research of intrusion detection algorithm based on parallel SVM*. conference on electronics information and emergency communication (ICEIEC),. Piscataway, p. 153–156.(IEEE)

[5] Rezvy,SH;et al. 2019, *Intrusion Detection and Classification with Autoencoded Deep Neural Network* . Springer Nature Switzerland AG, pp. 142–156.

[6] Ingre,B.;et al.2015, *Performance Analysis of NSL-KDD dataset using ANN* . SPACES-2015, Dept of ECE, K L UNIVERSITY.

[7] BROWN, C.; et al. July 08 2009, *Analysis of the 1999 DARPA/Lincoln Laboratory IDS Evaluation Data with NetADHICT*. In Proc. of the Second IEEE international conference 10 –on Computational intelligence for security and defense applications, Canada, 67-73.

[8]OZGUR, A.; et al. , April 14 2016, *A Review of KDD99 Dataset Usage in Intrusion Detection and Machine Learning between 2010 and 2015*. PeerJ Preprints.

[9] KDD Cup 1999 Data. https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html 13/May/2019.

[10] BHUYAN, M. H.; et al. 2015,*Towards Generating Real-life Datasets for Network Intrusion Detection*. IJ Network Security, Vol. 17, No.6, 683-701.

[11] BEKKAR, M.; et al. 2013, *Evaluation Measures for Models Assessment over Imbalanced Data Sets*.Information Engineering and Applications, Vol.3 No.10,27-39.

[12] ABDULLAH, M; ALSANEE, E; ALSEHEYMI, N. 2014, *Energy Efficient Cluster- Based Intrusion Detection System for Wireless Sensor Networks*.International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 5, No. 9, 10−15.