مجلة جامعة دمشق للعلوم الهندسية
المجلد41 العدد (3) – 2025: 255-267

Damascus University Journal for engineering sciences
Vo 41 number (3) -2025: 255-267

أثر الهجمات العدائية على الشبكات العصبونية الممثلة بالبيان والكشف عنها

ضياء حسن هرموش*[1] هيام خدام[2] أغيد القطعان[3]

*[1]. طالب دكتوراه في قسم هندسة الحواسيب والأتمتة – كلية الهندسة الميكانيكية والكهربائية– جامعة دمشق.
DiaaHarmoush@damacusuniversity.edu.sy

[2]. دكتور، قسم هندسة الحواسيب والأتمتة – كلية الهندسة الميكانيكية والكهربائية– جامعة دمشق.
Hiyamkhaddam@Damascusuniversity.edu.sy

[3]. دكتور، قسم هندسة الحواسيب والأتمتة – كلية الهندسة الميكانيكية والكهربائية– جامعة دمشق.
AghiadAlkatan@Damascusuniversity.edu.sy

الملخص:

تعتبر الشبكات العصبونية الممثلة بالبيان GNN (Graph Neural Networks) أحد نماذج التعلم الآلي واسعة الانتشار وذلك لتميزها الكبير في عدد من التطبيقات الخاصة بنمذجة الرسوم البيانية وتحليلها.

وعلى الرغم من كفاءة هذه الشبكات العالية بمهام تصنيف العقد والتنبؤ بالارتباط وحتى تصنيف البيان ككل، إلا أن أي تغيير بسيط في طوبولوجيا البيان أو خصائص العقد سيؤثر سلباً على أداء هذه الشبكات واستقرارها وسيؤدي الى نتائج غير مرغوبة.

تم في هذا البحث دراسة بنية الشبكات العصبونية الممثلة بالبيان (GNN) وكيفية تدريبها لتصنيف العقد الخاصة بالبيان الشهير (Citation Network) المعروف بشبكة الاقتباسات، ودراسة أثر الهجمات العدائية على هذه الشبكات وكيفية الكشف عنها.

تظهر النتائج التجريبية الأثر السلبي للهجمات العدائية على أداء GNN والذي يتيح الفرصة للمهاجمين لاستغلال الثغرات الأمنية وتقييد تطبيقاتها، كما أظهرت النتائج القدرة على كشف هذه الهجمات الحاصلة على الشبكة.

الكلمات المفتاحية: الشبكات العصبونية الممثلة بالبيان، الهجمات العدائية.

تاريخ الايداع: 2023/2/25
تاريخ القبول: 2023/5/23

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

THE Impact and Detection of Adversarial Attacks on Graph Neural Networks

Diaa Hasan Harmosh*[1] Hiyam Khaddam[2] Aghiad ALkatan[3]

*[1]. PhD Student in The Department of Engineering Computers and Computing- Faculty of Mechanical and Electrical Engineering - Damascus University. DiaaHarmoush@damacusuniversity.edu.sy

[2]. Dr. Department of Engineering Computers and Computing- Faculty of Mechanical and Electrical Engineering - Damascus University. Hiyamkhaddam@Damascusuniversity.edu.sy

[3]. Dr. Department of Engineering Computers and Computing- Faculty of Mechanical and Electrical Engineering - Damascus University. AghiadAlkatan@Damascusuniversity.edu.sy

Received: 25/2/2023

Accepted: 23/5/2023

Abstract:

Graph neural networks are one of the widespread learning models due to its great advantage in a number of applications for modeling and analyzing graphs.

Despite the high efficiency of GNN in the tasks of classifying nodes, predicting edges, and even classifying the graph as a whole, any slight change in the topology of the graph or the characteristics of the nodes will negatively affect the performance and stability of these networks and will lead to undesirable results.

In this research, the structure of GNN was studied and how to train them to classify the nodes of the famous graph (Citation network), and to study the impact of hostile attacks on these networks and how to detect them.

Experimental results show the negative impact of hostile attacks on the performance of GNN, which allows attackers to exploit security vulnerabilities and restrict their applications.

The results also show the ability to detect these alopecia attacks on the network.

Keywords: Graph Neural Network, Adversarial Attacks.

المقدمة:

حظيت الشبكات العصبونية الممثلة بالبيان (GNN) بقدر كبير من الاهتمام في السنوات الأخيرة ،وذلك لأدائها المميز في العديد من مهام تحليل الرسوم البيانية ونمذجتها مثل تصنيف العقد والتنبؤ بالارتباط.

لكن العديد من الدراسات تحدثت عن ضعف هذه الشبكات أمام الهجمات العدائية إذ أن أي تعديل طفيف على العقد أو الحواف سيؤدي الى انخفاض في أداء هذه الشبكات واستقرارها.

1. الدراسات المرجعية:

تمت دراسة الهجمات العدائية على نماذج التعلم العميق في كل من التعلم الآلي والمجتمعات الأمنية والعديد من أنواع النماذج المختلفة التي تتأثر سلباً بهذه الهجمات ، وأحد النماذج التي تمت دراسة أثر الهجمات عليها هي الشبكات العصبونية العميقة والتي تعتبر حساسة للغاية لهذه الهجمات [1,2].

تفترض الغالبية العظمى من الهجمات والدفاعات أن أمثلة البيانات التي يتم انشاءها بغرض الهجوم مستقلة ومستمرة وهذا لا ينطبق مع مهام الرسوم البيانية المتعلقة بتصنيف العقد وذلك لأن العقد المتشابهة بشكل كبير تعتبر نوع من أنواع الهجمات، من المهم جدا التمييز بين الهجمات العدائية والقيم المتطرفة ، فالقيم المتطرفة قد يكون سببها طبيعي في الرسوم البيانية لذلك غالبا ما يتم انشاء أمثلة معادية بشكل متعمد بهدف تضليل نماذج التعلم الآلي وغالبا ما يتم تصميمها بحيث تكون غير ملحوظة [3,4].

في 2017 قام أحد الباحثين بقياس التغييرات التي تحدث في الرسم البياني وذلك أثناء القيام بمهام العنقدة عن طريق اضافة ضجيج الى هذا البيان الذي يمثل نظام أسماء النطاقات DNS [5].

بدأ الباحثون مؤخرا بدراسة الهجمات العدائية على التعلم العميق للرسوم البيانية وذلك بدراسة الهجمات التي تحدث وقت الاختبار (Test time) أي بعد تدريب النموذج(evasion attack) على تصنيف البيان ككل وتصنيف العقد بشكل خاص في GNN، ومع ذلك فهم لم يتحدثوا عن الهجمات التي تحدث وقت التدريب (poisoning attack) كما أن العمل اقتصر على نقل الهجوم أو قابلية نقل الهجوم من خلال حذف الحواف فقط لتغيير التنبؤ الخاص بعقدة واحدة فقط [6].

في 2018 قام الباحث zugner وزملاؤه بالنظر في الهجمات وقت التدريب ووقت الاختبار على تصنيف العقد وذلك بالاعتماد على نموذج بديل ثابت وتنفيذ الهجمات على هذا النموذج وتقييم أثرها من خلال تدريب المصنف على البيان المعدل بواسطة GNN [7].

كما قام الباحث في [8] بتعديل سمات العقد على شكل متجهات ثنائية بهدف تغيير الصنف الخاص بعقدة معينة وقت الاختبار من خلال التأثير المباشر على هذه العقدة أي بإضافة أو حذف ضلع مباشرة عليها أو من خلال جيرانها أي بالتأثير لكن هذه الهجمات كانت جيدة لاستهداف العقد الفردية.

ظهرت أحد الدراسات الأشهر في الهجمات العدائية وكيفية اجرائها على الشبكات العصبونية الممثلة بالبيان

وذلك باستخدام نقل التعلم transfer learning لإنشاء هجمات أثناء التدريب(poisoning attack) بالتالي تعديل بيانات التدريب لتقليل أداء GNN بعد التدريب أي زيادة معدل الخطأ أثناء التدريب لتعميم الأداء المنخفض بعد التدريب [9].

في عام 2022 قام أحد الباحثين بهجوم العنقدة Cluster Attack على تصنيف العقد ، وذلك بمهاجمة بعض العقد في الرسم البياني الاصلي عن طريق حقن هذا البيان بعقد مزيفة الهدف منها التأثير على تلك العقد، مما أدى الى تقليل أداء GNN بشكل ملحوظ على العقد المستهدفة وغير المستهدفة بشكل عام [10].

كما قام عدد من الباحثين في عام 2022 بدراسة متانة GNN ضد الهجمات العدائية التي قد تحدث وقت التدريب

على تسمية العقد (label poisoning attack) وذلك من خلال تعديل تسمية بعض العقد قبل تدريب GNN عليها ودراسة الى أي مدى تكون هذه النماذج مقاومة أو عرضة لمثل هذه الهجمات [11] .

تركز الدراسات السابقة على دراسة أثر الهجمات العدائية على GNN دون الكشف عنها كما تقتصر تلك الهجمات على الرسوم البيانية غير الموزونة.

لذا تم في هذا البحث تحويل البيان citation network الى بيان موزون (يمثل اضافة الوزن اضافة ورقة بحثية عدد من المرات) بهدف اجراء هجمات على أوزان الأضلاع مع اضافة بعض الأضلاع بين عقدة وأخرى ودراسة أثر الهجوم وقت الاختبار Evasion attack على أداء هذه الشبكات واستقرارها فضلا عن الاستفادة من كل من خصائص التعلم العميق وبنية البيان للكشف عن الهجوم.

2. تعاريف ومصطلحات:

الهجمات العدائية(Adversarial Attacks): هي تعديلات على طوبولوجيا البيان لتقليل أداء الشبكات العصبونية الممثلة بالبيان GNN واعطاء العقد أصناف مختلفة عن ما هي عليه في الحقيقة.

الحالة الضمنية (embedding): لكل عقدة مجموعة من السمات (الخصائص)التي تعبر عن الحالة الضمنية لهذه العقدة في الحالة الابتدائية تمثل هذه السمات بواسطة متجه من الكلمات الخاصة ذات القيمة 0/1 والتي تشير الى وجود أو غياب الكلمة المقابلة من قاموس الكلمات والذي يتكون من 1433 كلمة فريدة تعبر عن صنف كل عقدة.

خلال عملية التدريب يتم تحديث حالة كل عقدة من خلال طبقات GNN وصولا الى الطبقة النهائية لتمثل الحالة الضمنية لكل عقدة الخرج المتوقع لها من قبل الشبكة.

تمرير الرسائل أو تبادل الرسائل (Message Passing): تعتمد GNNعلى تبادل الرسائل بين الجوار لتحديث حالة كل عقدة خلال عملية التعلم وصولا الى الاستقرار النهائي(الخطأ أقل ما يمكن).

3. المشكلة العلمية:

تندرج المهمة المراد التعامل معها من ضمن المهام شبه الخاضعة للإشراف، بالتالي ستعطى الشبكة جزء من العقد ذات التسمية المعروفة (train nodes) بينما باقي العقد ستكون غير معروفة التسمية (test nodes).

احدى السمات المميزة في هذه الشبكات هي قدرتها على التعميم فهي تستخدم شكل من أشكال تبادل الرسائل بين العقد اعتمادا على بنية البيان وتحديثها باستخدام الشبكات العصبونية ضمن طبقاتها (ممكن أي نوع من الشبكات: طبقة خطية أو التفافية) تسمى عملية الانتشار هذه بتمرير الرسائل أو تبادل الرسائل (Message Passing):

أثناء كل تكرار لتمرير الرسائل فيGNN يتم تحديث الحالة الضمنية (embedding) لكل عقدة في البيان اعتماداً على المعلومات المجمعة من جوار هذه العقدة. يمكن التعبير عن هذا التحديث لتمرير الرسائل:

$$h_u^{k+1} = \text{UPDATE}^K(h_u^k, \text{AGGREGATE}^K(\{h_v^k, \forall v \in N(u)\}))$$
$$= \text{UPDATE}^K(h_u^k, m_{N(v)}^k) \dots (1)$$

حيث أن:

h_u^{k+1} :الحالة الضمنية للعقدة u في التكرار k+1.

h_u^k :الحالة الضمنية السابقة للعقدة u .

h_V^k :الحالة الضمنية لجيران العقدة u في التكرار K.

AGGREGATE: دالة التجميع.

UPDATE: دالة التحديث.

$m_{N(v)}^k$: الرسالة المجمعة من جيران العقدة U.

وفقا لما سبق: في كل تكرار K في GNN تأخذ دالة التجميع (AGGREGATE) الحالة الضمنية الخاصة بجيران العقدة(u) كمدخلات لها وتقوم بتوليد الرسالة (m) التي تمثل المعلومات القادمة من جيران (u). ثم تقوم دالة التحديث (UPDATE) بدمج الرسالة (m) مع حالة (u) السابقة لنحصل على الحالة الضمنية النهائية للعقدة u .

لكل عقدة في الحالة الابتدائية مجموعة من المعلومات المميزة (سمات العقدة) والتي يطلق عليها الحالة الضمنية

الابتدائية (h) وتتغير هذه الحالة من طبقة لأخرى اعتمادا على أوزان طبقات الشبكة العصبونية وصولا الى الطبقة الأخيرة والتي تعبر عن الحالة الضمنية النهائية (تمثل الخرج النهائي). وفقا لما سبق يتم تعريف تمرير رسالة GNN الأساسية:

$$h_u^k = \sigma\left(w_{self}^k h_u^{k-1} + w_{neigh}^k \sum_{v \in N(u)} h_v^{k-1} + b^k\right)$$

....(2)

حيث ان w_{neigh}^k, w_{self}^k هي أوزان الطبقات العصبونية.

h_u^k :الحالة الضمنية للعقدة u في الطبقة k.

h_u^{k-1} :الحالة الضمنية للعقدة u في الطبقة السابقة.

h_v^{k-1} :الحالة الضمنية لجيران العقدة u في الطبقة السابقة.

σ : تابع تفعيل لاخطي.

مما سبق:

تعتمد عملية تمرير الرسائل في GNN على عمليات خطية (دالة المجموع) متبوعة بعنصر واحد غير خطي (σ) حيث يتم جمع الحالة الضمنية الواردة من الجيران برسالة واحدة ( )

m لتدمج مع الحالة الضمنية للعقدة في الطبقة السابقة باستخدام تابع تفعيل خطي، وأخيرًا يطبق تابع تفعيل لاخطي.

يمكن أيضًا تعريف العديد من شبكات GNN بإيجاز باستخدام معادلات على مستوى الرسم البياني ككل. يمكن كتابة المعادلة للنموذج على النحو التالي:

$$H^k = \sigma\left(AH^{k-1}w_{neigh}^k + H^{k-1}w_{self}^k\right) \dots (3)$$

A: مصفوفة الجوار في الرسم البياني والتي ستساعد في تحديد الجوار الخاص بكل عقدة عند القيام بعملية تمرير الرسائل.

H^k: مصفوفة تتضمن الحالة الضمنية (embedding) لكل العقد في الطبقة k.

H^{k-1}: مصفوفة تتضمن الحالة (embedding) لكل العقد في الطبقة $k-1$.

لتبسيط نهج تمرير الرسائل من الشائع اضافة حلقات ذاتية الى الرسم البياني (تمثيل للحالة السابقة للعقدة) وحذف خطوة التحديث الصريحة أي أن التحديث ضمني . إن إضافة حلقات ذاتية (المصفوفة الواحدية I) لمصفوفة الجوار في المعادلة السابقة تمكن من مشاركة المعلمات بين مصفوفة الأوزان w_{self}^k المضروبة بالحالة الضمنية للعقدة u في الطبقة السابقة ومصفوفة الأوزان w_{neigh}^k المضروبة بالحالة الضمنية لجيران العقدة u في الطبقة السابقة بمصفوفة واحدة w^t كما تمكن من مشاركة كل من مصفوفة الحالة الضمنية لعقد الجوار ومصفوفة الحالة الضمنية للعقدة u في الطبقة السابقة بمصفوفة واحدة هي H^{t-1} لينتج المعادلة :

$$H^t = \sigma\left((A+I)H^{t-1}w^t\right) \dots (4)$$

تظهر الطريقة التي تعتمدها هذه الشبكات في التعلم على أنها عرضة للهجمات العدائية ويقصد بالهجوم بأنه تعديلات ذكية وغير ملحوظة على طوبولوجيا البيان (اضافة حواف أو ازالتها وحتى حذف عقد) أو خصائص العقد سيؤدي الى اضطرابات في تعلم هذه الشبكات أي عدم استقرارها بالتالي انخفاض كبير في الأداء ونتائج غير دقيقة ينجم عنها اعطاء العقد تسمية مختلفة عن تسمياتها الحقيقية.

لذلك تعتبر المشكلة الأساسية متانة هذه الشبكات واستقرارها أثناء عملية التعلم.

5. دراسة الحالة:

ظهرت الحاجة لاستخدام الشبكات العصبونية الممثلة بالبيان GNN لتحليل الرسوم البيانية ونمذجتها لأسباب عديدة منها:

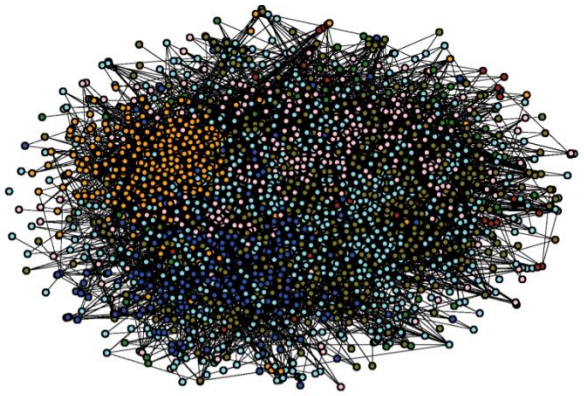
- لا يمكن للشبكات العصبونية التقليدية التعامل مع البيان كدخل فهي تكدس سمات العقد بترتيب معين فضلا عن تجاهل الترابط بين العقد (بنية البيان).
- هناك الحاجة لأكثر من مصفوفة لتمثيل البيان مثل مصفوفة الجوار ومصفوفة السمات التي تمثل خصائص العقد .
- لا يمتلك البيان شكلا ثابت ويمكن أن يكون هناك نفس المصفوفة لبيانين مختلفين.

لذلك تم استخدام شبكات GNN التي تشير الى معمارية الشبكات العصبونية التي تعمل على البيان لمهام تصنيف

العقـد الخاصـة بالبيـان المعـروف بشـبكة الاقتباسـات network (Citation) لقاعدة البيانات الشهيرة CORA dataset.

وفقا لقاعدة البيانات الشهيرة CORA فان كل عقدة في البيان تعبر عن ورقة بحثية لصنف معين من 7 أصناف موضحة في الجدول (1).

ويمثل الاقتباس بين كل عقدة وأخرى بضلع يعبر عن العلاقة التي تربط بينهما كما هو موضح في الشكل (1).



الشكل (1) شبكة الاقتباسات[14]

تم تصميم قاعدة البيانات CORA بشكل مخصص لمهام تصنيف العقد والتي تتكون من 2708عقدة تربطها 5278 وصلة ولكل عقدة 1433 سمة تمثل باقة من الكلمات التي تعبر عن صنف العقدة التي يحتويها الرسم البياني، كل عقدة تمثل ورقة بحثية والضلع يمثل الاقتباس من عقدة لأخرى يستشهد بها ضمن البيان، وهي 7 أصناف على الترتيب كما هو مبين في الجدول (1).

الجدول (1) أصناف العقد في Citation network.[14]

الصنف	عدد العقد
Case Based	298
Genetic Algorithms	418
Neural Networks	818
Probabilistic Methods	426
Reinforcement Learning	217
Rule Learning	180
Theory	351

يبين الجدول (1) عدد العقد وفقا لكل صنف في الرسم البياني.

الهدف الأساسي هو بناء نموذج GNN لتصنيف عقد هذا البيان الى الأصناف المذكورة في الجدول من خلال تقليل الخطأ على العقد المعروفة تسميتها لتعميم النتائج على باقي العقد (test nodes)، مع الأخذ بعين الاعتبار أن جميع العقد ستشارك في عملية التعلم وذلك وفقا لتمرير الرسائل في GNN.

على الرغم من قوة هذه الشبكات في مهام تصنيف العقد الا أن الطريقة التي تعتمدها تجعلها عرضة للهجمات العدائية التي يسعى المهاجم من خلالها لجعل النموذج يتجه باتجاه التلبيق الزائد (overfitting) أو التلبيق الناقص(underfitting) لمعرفة التعديلات اللازمة على البيان .

بناء على ذلك وللكشف عن الهجوم تم إضافة سمات جديدة لسمات العقد (خصائص العقد) وهي:

- مجموع أوزان كل عقدة
- عدد وصلات كل عقدة

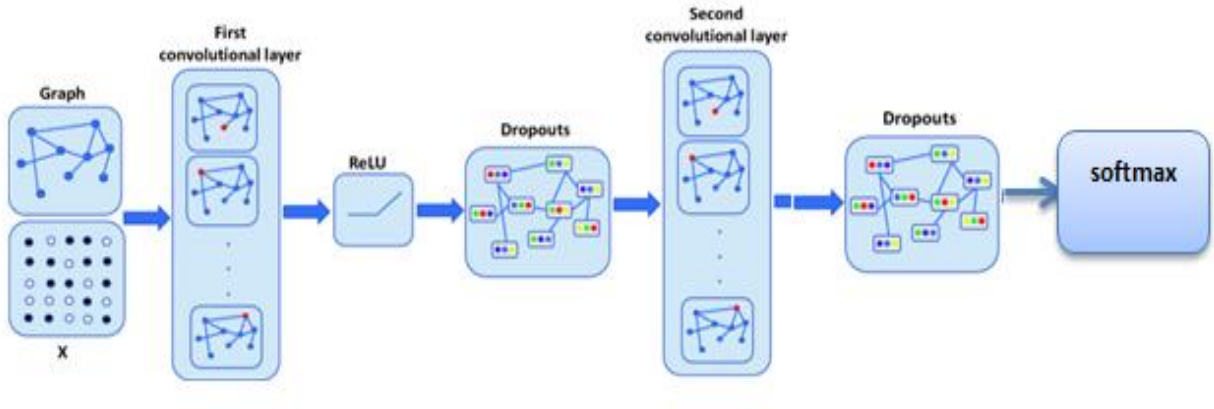
عند اضافة هذه السمات سيظهر أثر الهجوم بشكل واضح على أداء النموذج، والسبب في ذلك أن التعديلات التي سيقوم بها المهاجم ستتمثل بإضافة أو حذف الأضلاع أو حتى تغيير وزن ضلع وهذا الأمر سيؤثر بشكل كبير على عملية التعلم وظهور النتائج على الأداء بشكل واضح.

1.5. النموذج المقترح:

يبين الشكل (2) بنية النموذج حيث تمت نمذجة الشبكة من خلال طبقتين من طبقات الشبكات العصبونية الالتفافية، حيث تم استنباط الفكرة من الأداء المميز لهذه الطبقات في معالجة الصور وذلك عند استخراج الميزات الخاصة بكل بكسل من جيرانه وبشكل مشابه تم استخدام هذه الطبقات لتحديث حالة كل عقدة اعتمادا على جوارها.

- يمثل دخل الشبكة كل من مصفوفة الجوار A adjacency matrix ومصفوفة السمات (الحالة الضمنية الابتدائية للعقد) features matrix.

- طبقة GCN1 مع تابع تفعيل Relu .(rectified linear unit)
- طبقة dropout الهدف منها تقليل احتمالية حدوث overfitting .
- طبقة الخرج GCN2 مع تابع تفعيل لاخطي (softmax) لانتاج احتمال كل تصنيف (7 أصناف).



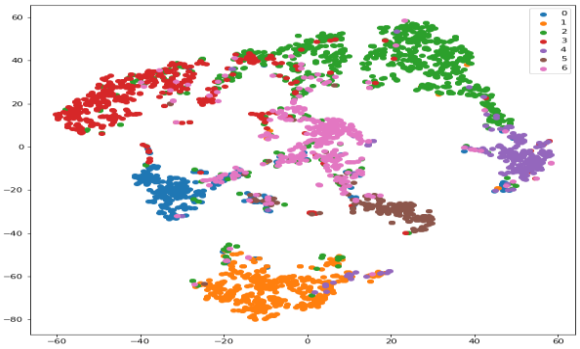
الشكل(2) بنية النموذج

يوضح الجدول (2) دقة النموذج بعد تدريبه:

الجدول (2) دقة النموذج بعد التدريب

	Precision	Recall	F1_score	Support
Case Based	0.69	0.81	0.75	298
Genetic Algorithms	0.83	0.97	0.89	418
Neural Networks	0.93	0.68	0.78	818
Probabilistic Methods	0.80	0.77	0.78	426
Reinforce_ment Learning	0.61	0.88	0.72	217
Rule Learning	0.77	0.83	0.80	180
Theory	0.66	0.68	0.67	351
ACCURACY			0.78	2708
MACRO AVG	0.76	0.80	0.77	2708
WEIGHTED AVG	0.80	0.78	0.78	2708

يبين الجدول(2) أن دقة النموذج المقترح جيدة وأن النموذج قادر على تصنيف العقد بالدقة المطلوبة بالتالي فان الشكل النهائي للتصنيف سيكون كما هو موضح في الشكل (3):



الشكل (3) الشكل النهائي للعقد بعد التصنيف

6. الهجمات العدائية (Adversarial (Attack:

لا يعتبر أي تعديل بسيط في البيان هجوما حقيقيا على البيان لأن المهاجم يسعى بشكل أساسي لإجراء تعديلات ممنهجة وذكية، الهدف منها هو تقليل أداء النموذج وزيادة معدل الخطأ بالتالي اعطاء بعض العقد تسمية مختلفة عن التسمية التي هي عليه لذلك فان المهاجم سيجري تعديلات ذكية وغير ملحوظة لتضليل GNN.

1.6 الهجوم وقت الاختبار (Evasion attack):

يعني أن النموذج المستهدف يتم تدريبه على بيان نظيف واختباره على بيان مهاجم.

من وجهة نظر المهاجم: المهاجم يمكنه تشويش بنية البيان لخداع نموذج (GNN)المدرب مسبقاً.

يسعى المهاجم لمهاجمة عقد معينة وقت الاختبار ليغير الصنف الخاص بها من خلال الهجوم المباشر على هذه العقد أو بالتأثير عليها من خلال الجيران اذ أن الهجوم يحصل على عقد معينة لتقليل أداء النموذج وتغيير الصنف الخاص بها من خلال اضافة أو حذف ضلع الى جيران هذه العقد أو حتى اجراء تعديل في الخصائص.

تم اجراء هذا الهجوم على عدد من العقد بهدف تغيير الصنف الخاص بها من خلال اجراء بعض التعديلات البسيطة على بنية البيان علما أن المشكلة الأساسية في هذه الهجمات تتمثل بقدرة المهاجم على معرفة التعديلات اللازمة لتقليل أداء النموذج وتغيير الصنف الخاص بهذه العقد بالتالي فان المهمة الأساسية للمهاجم جعل الخطأ كبيرا بما يتناسب مع اعطاء العقدة v صنف مختلف (c_{new}) عن الصنف الأصلي القديم قبل التعديل c_{old} كما توضح المعادلة :

$$l_s(A, X, W, v_0) = \max_{c \neq c_{old}} [A^2XW]_{v0c} - [\mathrm{A^2XW}]_{\mathrm{v0c_{old}}} \dots (5)$$

- يمتلك المهاجم نموذج بديل يستطيع توقع تسمية كل عقدة في البيان قبل تطبيق الهجوم على GNN.
- سيؤثر الهجوم بشكل أساسي على العقدة المراد مهاجمتها وعلى عقد الجوار وذلك لأن GNN تعتمد على تبادل الرسائل بين العقد.
- تم مهاجمة بعض العقد بشكل غير مباشر من خلال جوار هذه العقد.
- تم اجراء الهجوم باضافة أضلاع موزونة مع تغيير بعض الأوزان وملاحظة مقدار زيادة الخطأ للوصول الى التعديلات الملائمة التي التي تجعل الخطأ كبيرا بما يتناسب مع الهجوم.

7.النتائج:

- تم اختبار النموذج السابق على البيان المعدل وفقا للجدول التالي:

الجدول(3) دقة النموذج بعد الهجوم:

	Precision	Recall	F1_score	Support
Case Based	0.70	0.77	0.73	298
Genetic Algorithms	0.60	0.98	0.74	418
Neural Networks	0.93	0.60	0.73	818
Probabilistic Methods	0.81	0.73	0.77	426
Reinforcement Learning	0.65	0.86	0.74	217
Rule Learning	0.79	0.77	0.77	180
Theory	0.66	0.62	0.64	351
ACCURACY			0.73	2708
MACRO AVG	0.73	0.76	0.73	2708
WEIGHTED AVG	0.77	0.73	0.73	2708

يبين الجدول(3):

- الاعتماد على مقياس F1-Score لتقييم أداء النموذج وذلك لتعدد الفئات كما أنه تم حساب المتوسط الحسابي المرجح Macro avg وغير المرجح weighted avg لأن توزيع الفئات غير متوازن.
- الاسترجاع Recall: عدد النتائج المتعلقة بالبحث على عدد النتائج الكلية.
- الدقة Precision: عدد النتائج المتعلقة بالبحث على النتائج المسترجعة الكلية.
- انخفاض المتوسط غير المرجح Macro avg والذي يمثل درجة F1 الكلية باستخدام المتوسط الحسابي لجميع درجات F1 لكل فئة.
- انخفاض المتوسط المرجح weighted avg والذي يمثل متوسط درجات F1 لكل فئة مع مراعاة دعم كل صنف(support)
- انخفاض F1-Score لكل فئة من الفئات بشكل واضح وذلك لترابط العقد مع بعضها البعض.
- انخفاض أداء النموذج على الرغم من أن التعديلات التي تم اتخاذها على البيان لم تتجاوز 5%.

يوضح الجدول (4) أثر الهجوم على الكفاءة بإضافة السمات ومقارنة الأداء قبل وبعد الهجوم:

الجدول (4) أثر الهجوم على الكفاءة بإضافة السمات.

Accuracy	Before	After	Ratio
Accuracy	0.78	0.73	6.4%
Macro precision	0.76	0.73	3.9%
Macro recall	0.80	0.76	5%
Macro f1-score	0.77	0.73	3.8%
Weighted precision	0.80	0.77	3.7%
Weighted recall	0.78	0.73	6.4%
Weighted f1-score	0.78	0.73	6.4%

يبين الجدول (4):

- أثر الهجوم قد ظهر بشكل واضح على الرغم من أن الهجمات كانت مقتصرة على بعض الأضلاع مع تغييرات في الأوزان بنسبة 5%.
- تدهور أداء الشبكة بشكل واضح نتيجة الهجوم المطبق.

يوضح الجدول (5) أثر الهجوم على الكفاءة بدون إضافة السمات:

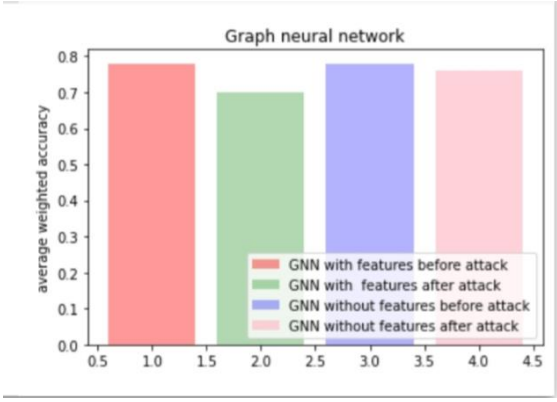
الجدول (5) أثر الهجوم على الكفاءة بدون إضافة السمات:

Accuracy	Before	After	Ratio
Accuracy	0.78	0.76	2.5%
Macro precision	0.76	0.76	1.3%
Macro recall	0.80	0.78	2.5%
Macro f1-score	0.77	0.76	1.3%
Weighted precision	0.80	0.78	2.5%
Weighted recall	0.78	0.76	2.5%
Weighted f1-score	0.78	0.76	2.5%

يبين الجدول (5):

- أثر الهجوم قد ظهر ولكن ليس كما في حالة إضافة سمات العقد المرتبطة بالبيان إلى جدول السمات
- بالتالي إضافة السمات تساعد بشكل كبير على اظهار أثر الهجوم والكشف عنه

يبين الشكل(4) أداء النموذج باضافة السمات وحذفها مع وجود نفس الهجوم على GNN:



الشكل(4) أداء النموذج باضافة السمات وحذفها قبل وبعد الهجوم

8. النتائج:

- تظهر النتائج أن اضافة السمات يظهر أثر الهجوم بشكل واضح على أداء النموذج.
- تعتبر الشبكات العصبونية الممثلة بالبيان مزيج من التعلم العميق ونظرية البيان وان أي هجوم على هذه الشبكات ستظهر آثاره على معاملات الشبكة بالتالي يمكن الاستفادة من بعض نظريات البيان للكشف عن الهجوم من خلال المطابقة بين البيان المهاجم والبيان غير المهاجم عن طريق حساب الفرق بين بعض المعاملات الممثلة: بالوسيط ، الأثر، والمحدد لكل من مصفوفة: الارتباط (Incidence matrix)، الدرجةDegree matrix)، الجوار(Adjacency matrix) قبل وبعد الهجوم .

من أجل تقليل الكلفة الحسابية تم تعريف المصفوفة التالية :

Mymat=ADJ+DEG…..(6)

تمثل هذه المصفوفة حاصل جمع كل من مصفوفة الجوار ومصفوفة الدرجة.

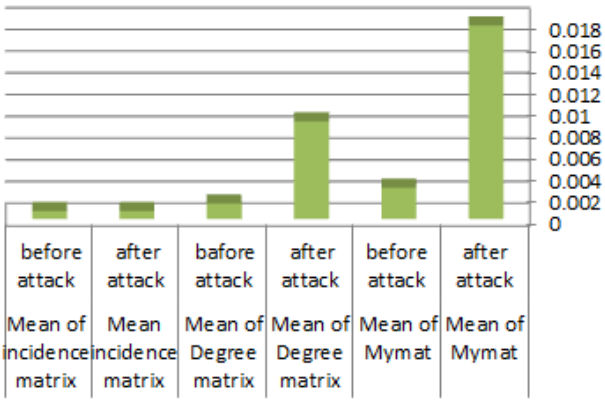
تم اجراء الحسابات التالية على هذه المصفوفة وفقا للجدول:

الجدول (6) الكشف عن الهجوم:

Accuracy	Before	After
Determinate of adjacency matrix	0	0
Mean of laplacian_ matrix	0	0
Mean of incidence matrix	0.0007385524372230429	0.0007384249247166162
Mean of Degree Matrix	0.001439468154971647	0.0089958577790900
trace Mymat	10556.0	66004.0
Det Mymat	0	0
Mean Mymat	0.002878936309943294	0.017822077590551766

من الجدول (6):

- تظهر آثار الهجوم بشكل واضح على معاملات المصفوفة Mymat.
- يوضح الشكل(5) الفرق بين المعاملات قبل وبعد الهجوم:



الشكل(5) الفرق بين المعاملات

9. الاستنتاجات والآفاق المستقبلية:

- يمكن الاستفادة من قدرة GNN بقيامها بمهام التصنيف على مستوى البيان ككل للكشف عن الهجوم في حال كان الهجوم على خصائص العقد (بالتالي أثناء التدريب).
- ان مقدار التغير الحاصل في المعاملات التي استخدمت للكشف عن الهجوم قد تساعد بمعالجة الهجوم لكن مع الأخذ بعين الاعتبار أن هذا التغيير قد يؤثر سلباً على العقد التي لم تهاجم لذلك لا بد من معرفة العقد المصابة قبل استخدام هذه المعاملات.
- لا بد من وجود معامل يساعد في حساب حجم الهجوم وذلك لتنفيذ هجمات بأحجام مختلفة.
- تتمثل معالجة الهجوم باستعادة البيان الأصلي من البيان الذي تمت مهاجمته.

التمويل: هذا البحث ممول من جامعة دمشق وفق رقم التمويل(501100020595).

References:

[1] Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In AAAI, pp. 2871–2877, 2015.

[2] Aleksandar Bojchevski and Stephan G ¨unnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In ICLR, 2018a.

[3] Aleksandar Bojchevski and Stephan G ¨unnemann. Adversarial attacks on node embeddings. arXiv preprint arXiv:1809.01093, 2018b .

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015.

[5] Yizheng Chen, Yacin Nadji, Athanasios Kountouras, Fabian Monrose, Roberto Perdisci, Manos Antonakakis, and Nikolaos Vasiloglou. Practical attacks against graph-based clustering. arXiv preprint arXiv:1708.09056, 2017.

[6] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In ICML, 2018.

[7] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Z ¨ugner, and Stephan G ¨unnemann. NetGAN: Generating graphs via random walks. In ICML, 2018.

[8] Daniel Z ¨ugner, Amir Akbarnejad, and Stephan G ¨unnemann. Adversarial attacks on neural networks for graph data. In SIGKDD, pp. 2847–2856, 2018.

[9] Zügner, D., & Günnemann, S. (2019). Adversarial attacks on graph neural networks via meta learning. arXiv preprint arXiv: 1902.08412.

[10] Wang, Z., Hao, Z., Wang, Z., Su, H., & Zhu, J. (2022). CLUSTER ATTACK: Query-based Adversarial Attacks on Graphs with Graph-Dependent Priors.

[11] Liu, G., Huang, X., & Yi, X. (2022). Adversarial Label Poisoning Attack on Graph Neural Networks via Label Propagation. In European Conference on Computer Vision (pp. 227-243). Springer, Cham.

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In ICLR, 2017.

[13] Hamilton, W. L. (2020). Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3), 1-159.

[14] Li, K., Feng, Y., Gao, Y., & Qiu, J. (2020). Hierarchical graph attention networks for semi-supervised node classification. Applied Intelligence, 50(10), 3441-3451