مجلة جامعة دمشق للعلوم الهندسية
المجلد 42 العدد 2 (2026): 77- 91

Damascus University Journal of Engineering Sceinces
Vo 42 No2. (2023): 77 - 91

تحسين متانة الشبكات العصبونية الممثلة بالبيان ضد الهجمات العدائية

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

[1*] طالب دكتوراه، مهندس في قسم هندسة الحواسيب والأتمتة –كلية الهندسة الميكانيكية والكهربائية–جامعة دمشق. ddiaahh34@Damascusuniversity.edu.sy

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

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

الملخص:

حققت الشبكات العصبونية الممثلة بالبيان GNN (Graph Neural Networks) نجاحا ملحوظا في العديد من التطبيقات الخاصة بتحليل الرسوم البيانية ونمذجتها.

ويعود سر النجاح الكبير الذي حققته GNN في العديد من التطبيقات المتعلقة بالرسوم البيانية الى مخطط تمرير الرسائل الذي تعتمده أثناء التعلم حيث تقوم بتجميع رسائل الجوار لكل عقدة في كل طبقة من طبقاتها أثناء التدريب مما يسمح للنموذج في الطبقة النهائية من معرفة البيان بشكل كامل وفقا للرسائل المجمعة من كل عقدة وجوارها.

وعلى الرغم من قوة هذا المبدأ في مهام تصنيف العقد الخاصة بالبيان الا أن اعتماد GNN على بنية البيان بشكل كبير أثناء تبادل الرسائل يجعلها عرضة للهجمات العدائية التي تؤثر سلبا على متانة هذه الشبكات واستقرارها وبالتالي انخفاض كبير في الأداء ونتائج غير دقيقة ينجم عنها اعطاء العقد تسمية مختلفة عن تسمياتها الحقيقية.

تم في هذا البحث اقتراح خوارزمية لتحسين متانة GNN واستقرارها ضد الهجمات العدائية التي تم اجراءها بنسب مختلفة على نموذج GNN المدرب مسبقا على مهام تصنيف العقد الخاصة بالبيان المعروف بشبكة الاقتباسات (Citation Network) لقاعدة البيانات الشهيرة CORA dataset.

الكلمات المفتاحية: الشبكات العصبونية الممثلة بالبيان، الهجمات العدائية.

تاريخ الإيداع: 2023/7/19
تاريخ القبول: 2023/9/19

Improve The Robustness Of Graph Neural Networks Against Adversarial Attacks

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

*[1].PhD Student, Eng in The Department of Engineering Computers and Computing- Faculty of Mechanical and Electrical Engineering - Damascus University.
ddiaahh34@Damascusuniversity.edu.sy

[2]. Dr. Department of Engineering Computers and Computing- Faculty of Mechanicl 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: 19/7/2023
Accepted: 19/9/2023

Abstract:

Graph neural networks (GNN) have achieved remarkable success in many applications graph analysis and modeling.

The secret of the great success achieved by GNN in many applicationsrelated to graphs is due to the message passing scheme that it adopts during learning, as it collects neighbor messages for each node in each of its layers during training, which allows the model in the final layer to know the graph completely.

Despite the strength of this principle in the tasks of classifying nodes for the graph, GNN's reliance on the graph structure greatly during the message passing makes it vulnerable to adversarial attacks that negatively affect the Robustness and stability of these networks and thus a significant decrease in performance and inaccurate results that result in giving nodes A different class from its real classes.

In this research, an algorithm was proposed to improve the robustness and stability of GNN against adversarial attacks that were performed in different proportions on the GNN model pre-trained on the tasks of classifying nodes of the graph known as the Citation Network of the famous CORA dataset.

Keywords: Graph Neural Network, Adversarial Attacks**.**

المقدمة:

مع زيادة الاهتمام بتقنيات التعلم الآلي القائمة على الرسم البياني كانت هناك العديد من الدراسات التي تشير الى أن هذه النماذج عرضة للهجمات العدائية، على وجه الخصوص الشبكات العصبونية الممثلة بالبيان GNN.[8]

تعتبر الشبكات العصبونية الممثلة بالبيان أحد أهم نماذج التعلم الآلي لتحليل الرسوم البياني ونمذجتها وعلى الرغم من كفاءة هذه الشبكات الا أن العديد من الدراسات تظهر الأثر السلبي للهجمات العدائية على متانة GNN والذي يتيح الفرصة للمهاجمين لاستغلال الثغرات الأمنية وتقييد تطبيقاتها لذلك كان لا بد من الحد من الأثر السلبي لهذه الهجمات وتحسين متانة GNN ضدها.[8]

1. الدراسات المرجعية:

قامت العديد من الدراسات البحثية بدراسة أثر الهجمات العدائية على متانة الشبكات العصبونية الممثلة بالبيان واستقرارها، من هذه الدراسات:

قام الباحث 2018) ugner بدراسة نوعين من الهجمات العدائية: الهجمات التي تحدث وقت التدريبTime (Training) أو ما يسمى(poisoning attack) والهجمات التي تحدث وقت الاختبار(Testing time) أو ما يسمىattack (evasion).

اقتصرت هذه الهجمات على الرسوم البيانية غير الموزونة وتمثل الهجوم باضافة أو حذف بعض الأضلاع للتأثير على بعض العقد لمهام تصنيف العقد لGNN.[2][1]

قام Liu واخرون(2022) بدراسة متانة GNN ضد الهجمات التي تحدث وقت التدريب على تسمية العقد poisoning (label attack) وذلك عن طريق تغيير تسمية بعض العقد وتدريب النموذج على البيان الذي تمت مهاجمته.[3]

كما قام Wang واخرون(2022) بدراسة هجوم يسمى (Cluster attack) على بعض العقد أثناء قيام GNN بمهام تصنيف العقد لهذا البيان، وذلك من خلال حقن البيان السابق ببعض العقد المزيفة ودراسة أداء النموذج على هذا البيان مما أدى الى انخفاض أداء النموذج بشكل ملحوظ. [6]

ومن الدراسات التي تحدثت عن تحسين متانة GNN ضد الهجمات العدائية:

قام Tang واخرون(2020) بنقل المتانة ل GNN ضد الهجمات التي تحدث وقت التدريب من خلال مفهوم الجزاء أي اجراء عقوبة على العقد المصابة من خلال تقليل الاهتمام بها أي تقليل تأثير هذه العقد على عملية التدريب عن طريق اضافة معامل يسمى معامل الأهمية.[5]

يسمح هذا المعامل بزيادة الاهتمام بالوصلات غير المصابة ويقلل من الاهتمام بالوصلات المصابة لذلك يفترض البحث وجود معرفة مسبقة بالأضلاع (الوصلات) التي سيتم اضافتها الى البيان كما ركز الباحث Zhang (2020) على تحسين متانة GNN على الرسوم البيانية المتجانسة من خلال تقدير قيمة الوزن وفقا لمعامل يسمى معامل التشابه والذي يقيس التشابه بين خصائص عقدتين بالتالي عندما يكون التشابه كبير(قيمة الوزن كبيرة) سيسمح بتمرير الرسائل، وفي حال كان التشابه قليل سيمنع الرسالة من المرور.[8]

قام الباحث Regol (2019) بتحسين متانة هذه الشبكات ضد الهجمات التي تحدث وقت التدريب من خلال نسخ الحالة الضمنية للعقد التي تمت مهاجمتها الى جوارها وتفترض هذه الخوارزمية أن البيان المراد التعامل معه متجانس. كما أن الهجمات اقتصرت على بعض العقد المعروفة من قبل الباحث لأن عملية النسخ قد تستغرق وقتا طويلا، كما أنه لا بد من معرفة هذه العقد حتى لا تتم عملية النسخ على كامل البيان.[4]

اقترح الباحث Wu (2022) نموذج دفاع ضد الهجمات العدائية يسمى(C2OG). اعتمد النموذج السابق أو ما يسمى باطار التدريب المشترك

ثنائي الرؤية للرسوم البيانية على فصل بيانات الرسم البياني الى عرضين :عرض الميزة (feature view) وعرض الهيكل (structure view).

يوفر هذان العرضان معلومات مختلفة حول العقد في الرسم البياني حيث يصف عرض الميزة الخصائص الأساسية للعقد بينما يصف عرض الهيكل العلاقة بين العقد ، وبالتالي فانه من أجل كل عقدة في البيان: يتم فصل معلوماتها في عرض الميزة وعرض الهيكل ليتم تدريب اثنين من المصنفات بشكل منفصل على تلك المعلومات وحساب درجة الثقة لجميع العقد غير معروفة التسمية لكل مصنف واضافة العقد الأكثر ثقة من كل مصنف الى مجموعة البيانات بعد عدد معين

من التكرار أو عدم وجود عقد اختبار.

في النهاية يتم الحصول على نموذجين مدربين جيدا ليتم انشاء مجموعة بيانات منهما من خلال متوسط توقعاتهم. بالتالي فان C2OG يسمح للنماذج الفرعية من تصحيح بعضها البعض بشكل متبادل وبالتالي تعزيز متانة مجموعات البيانات لينتج عن ذلك تحسين متانة GNN. [7]

استفادت الأبحاث السابقة من فكرة تجانس البيان لتحسين متانة GNN اعتمادا على معامل التشابه لإعادة تصميم مخطط تمرير الرسائل أو نسخ الحالة الضمنية للعقد المصابة الى جوارها في البيان بالتالي هناك معرفة مسبقة بالأضلاع والعقد التي يتم مهاجمتها في البيان غير الموزون.

لذلك تم في هذا البحث تطوير خوارزمية لتحسين أداء GNN بغض النظر ان كان البيان متجانس أو غير متجانس ضد الهجمات العدائية التي تحدث على البيان الموزون.

3.المشكلة العلمية:

يعتبر تصنيف العقد من أكثر مهام الشبكات العصبونية الممثلة بالبيان شيوعا ويعود سر نجاح هذه الشبكات في هذه المهام الى طريقة انتشارها اذ أنها 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 في GNN يتم تحديث الحالة الضمنية (embedding) لكل عقدة في البيان اعتماداً على المعلومات المجمعة من جوار هذه العقدة.

يمكن التعبير عن هذا التحديث لتمرير الرسائل كما يلي:

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

حيث أن:

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

h_u^k :الحالة الضمنية السابقة للعقدة u .

h_v^k :الحالة الضمنية لجيران العقدة u في الطبقة K.

AGGREGATE: دالة التجميع.

UPDATE: دالة التحديث.

N(u):مجموعة العقد لجوار العقدة u.

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

وفقا لما سبق: في كل طبقة (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 للحصول على الحالة الضمنية النهائية للعقدة u .

الحالة الضمنية (embedding):

لكل عقدة في الحالة الابتدائية مجموعة من المعلومات المميزة (سمات أو خصائص العقدة) والتي يطلق عليها الحالة الضمنية الابتدائية (h) وتتغير هذه الحالة من طبقة لأخرى اعتمادا على

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

أوزان طبقات الشبكة العصبونية وصولا الى الطبقة الأخيرة والتي تعبر عن الحالة الضمنية النهائية.
وفقا لما سبق يتم تعريف تمرير رسالة GNN الأساسية كما يلي:

$$h_u^k = \sigma(w_{self}^k h_u^{k-1} + w_{neigh}^k \sum_{v \in N(u)} h_v^{k-1} + b^k) \quad \forall v \in N(u) \ldots (2)$$

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:

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

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 $k-1$.
على الرغم من قوة هذه الشبكات في مهام تصنيف العقد الا أن اعتمادها الكبير على بنية البيان يجعلها عرضة للهجمات العدائية التي تؤثر على متانتها واستقرارها لينتج عن ذلك انخفاض كبير في الأداء ونتائج غير دقيقة ينجم عنها اعطاء بعض العقد تسمية مختلفة عن تسمياتها الحقيقية.

4.الهجمات العدائية (Adversarial Attacks):

هي تعديلات ممنهجة على طوبولوجيا البيان (اضافة أو حذف بعض الأضلاع) تؤثر سلبا على أداء GNN.
يسعى المهاجم لتقليل أداء GNN في مهام تصنيف العقد للبيان من خلال مهاجمة هذه العقد بشكل مباشر أو بالتأثير غير المباشر عليها (من خلال جوار هذه العقد) وذلك كله بالاستفادة من نهج هذه الشبكات في التعلم (مخطط تبادل الرسائل).
يتمثل الهجوم المراد تنفيذه باضافة أو حذف بعض الأضلاع مع تغيير أوزانها أو تغيير بعض الأوزان لأضلاع موجودة مسبقا، لذلك فان المهاجم سيقوم بإجراء التعديلات اللازمة على البيان الأصلي G(V,E)
حيث أن:
V: مجموعة العقد في البيان G، E: مجموعة الأضلاع في البيان G.
للحصول على $\acute{G}\ (\acute{V}, \acute{E})$
حيث أن:
$\acute{E}$: مجموعة الأضلاع في البيان المهاجم $\acute{G}$، $\acute{V}$: مجموعة العقد في البيان المهاجم $\acute{G}$.
هذه التعديلات ستؤثر سلبا على الأداء العام لنموذج GNN المدرب مسبقا على البيان الأصلي G(V,E) لينتج عن ذلك اعطاء بعض العقد تسميات مختلفة عن التسميات التي كانت عليها قبل تنفيذ الهجوم.

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

وفقا لذلك: من أجل كل عقدة v في البيان G(V,E):

تم تعريف:

v_c: $number\ of\ edges\ for\ vertex\ v.$

v_s: $sum\ of\ edges\ weights\ for\ vertex\ v.$

يستهدف المهاجم العقد التي ليس لها وصلات كثيرة أو وصلات كثيرة مع أوزان قليلة.

لذلك تم تعريف:

th1: الحد الأعظمي لعدد الأضلاع (الوصلات) المرتبطة بالعقدة.

th2: الحدالأعظمي لمجموع أوزان الأضلاع المرتبطة بالعقدة .

كما هو موضح في الشكل(1):

```
def attacker (G, th1, th2, edge_limit,
weight_limit, attack_limit, Beta)
int attack_limit_temp=0;
int edge_limit_temp=0;
int weight_limit_temp=0;
For each U in G.V
if U_c < th1
attack_limit_temp +=1
for each M ~= U in G.V
if(edge_limit_temp<edge_limit)
if E(U, M) not exist
edge_limit_temp +=1
add E(U,M)=1 # with weight 1
end
end
end
print('ratio of attack of node U')
print(100*edge_limit/U_c)
else
if U_s < th2
attack_limit_temp +=1
for each M ~= U in G.V
if(weight_limit_temp<weight_limit)
if E(U, M) exist
add E(U,M)=Beta+E(U,M)
edge_limit_temp +=Beta
end
end
end
print('ratio of attack of node U')

end
end
```

```
if (attack_limit_temp > attack_limit)
break (out of for loop)
end
end
end
```

الشكل(1) نموذج الهجوم المقترح

يبين الشكل (1):

attack limit : الحد الأعظمي للعقد المراد مهاجمتها .

edge limit: الحد الأعظمي للوصلات المراد اضافتها لعقدة معينة من البيان

Weight limit: الحد الأعظمي لمجموع الأوزان على عقدة ما في البيان.

5.خوارزمية التصحيح المقترحة:

وفقا لما سبق فان احدى السمات المميزة في هذه الشبكات هي قدرتها على التعميم فهي تستخدم شكل من أشكال تبادل الرسائل بين العقد اعتمادا على بنية البيان، بالتالي يمكن القول أن هذه الشبكات هي مزيج بين التعلم العميق ونظرية البيان لذلك تم الاستفادة من البيان الذي تمت مهاجمته لاستعادة البيان الأصلي.

على الرغم من عدم المعرفة الكاملة بتفاصيل الهجوم أو الاضطراب الذي سبب انخفاض في أداء الشبكات الا أنه من الممكن معرفة مقدار التغير الحاصل بين البيان المهاجم والبيان غير المهاجم واضافة الفرق الى مصفوفة الجوار بعد الهجوم لاستعادة استقرار النموذج كما هو موضح في المعادلات التالية:

$$CM^{t}_{[i,j]} = \begin{cases} 0 & \text{if} \quad A^{t-1}_{[i,j]} = A^{t}_{[i,j]} \\ A^{t-1}_{[i,j]} - A^{t}_{[i,j]} & \text{if} \quad A^{t-1}_{[i,j]} \neq A^{t}_{[i,j]} \end{cases} \quad ..(4)$$

$$CM^{t+1}_{[i,j]} = A^{t}_{[i,j]} + CM^{t}_{[i,j]} \ldots (5)$$

حيث أن:

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

$CM^{t}_{[i,j]}$: مصفوفة التصحيح في اللحظة tوتمثل مقدار التغير الحاصل بين كل لحظة وأخرى

$A^{t-1}_{[i,j]}$: مصفوفة التجاور للبيان في اللحظة t-1 (قبل الهجوم)

$A^{t}_{[i,j]}$: مصفوفة التجاور للبيان في اللحظة t (بعد الهجوم)

$CM^{t+1}_{[i,j]}$: مصفوفة التصحيح في اللحظة$t+1$

يتم حساب نسبة الهجوم بالاعتماد على مجموع عناصر مصفوفة التصحيح في اللحظة t على عدد الوصلات الكلي:

$$\text{Attack Ratio}=100*\frac{\sum_1^n CM^t}{E} \dots (6)$$

حيث أن:

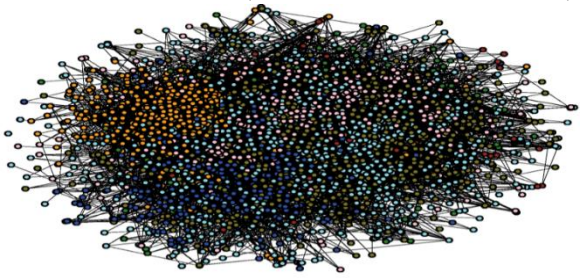
E: عدد الوصلات الكلي.

6. دراسة الحالة:

يظهر الشكل (2) البيان المعروف بشبكة الاقتباسات (Citation network) 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 (2).



الشكل (2) شبكة الاقتباسات [10]

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.[10]

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.

تتدرج 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).

تم تحويل البيان Citation network الموضح في الشكل (2) الى بيان موزون.

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

ثم القيام بتنفيذ الهجوم على النموذج ودراسة أثر هذه الهجمات على متانة واستقرار GNN ثم تطبيق خوارزمية التصحيح المقترحة لتحسين الأداء واستعادة استقرار الشبكة.

1.6.النموذج المقترح:

يبين الشكل (3) بنية النموذج حيث تمت نمذجة الشبكة من خلال طبقتين من طبقات الشبكات العصبونية الالتفافية، وذلك

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

لتشابه طريقة تحديث حالة كل عقدة اعتمادا على جوارها من استخراج المميزات الأساسية للبكسلات من جوارها في الصور.

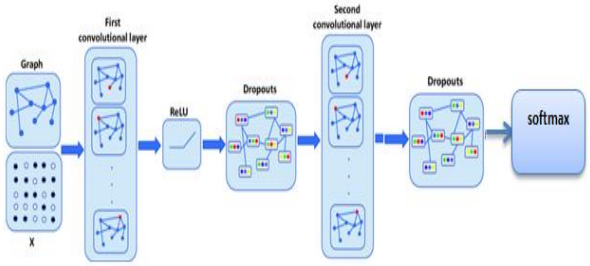
وفقا للشكل(3):

• 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) لإنتاج احتمال كل تصنيف (7 أصناف).



الشكل(3) بنية النموذج المقترح

7-النتائج:

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):

الاعتماد على مقياس F1-Score لتقييم أداء النموذج وذلك لأنه أحد المقاييس الشائعة الاستخدام في مهام التصنيف متعددة الفئات، كما أنه تم حساب المتوسط الحسابي غير المرجح Macro avg والمتوسط 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.

المتوسط غير المرجح Macro avg: والذي يمثل درجة F1 الكلية باستخدام المتوسط الحسابي لجميع درجات F1 لكل فئة.

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

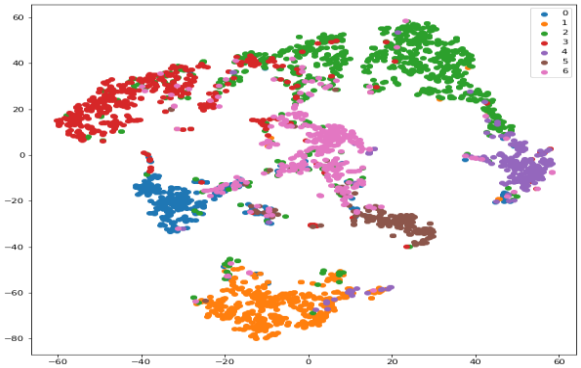
المتوسط المرجح weighted avg: والذي يمثل متوسط درجات F1 لكل فئة مع مراعاة دعم كل صنف (support).

$$F1\_Score = \frac{2*(precision*recall)}{precision+recall} \ldots(8)$$

$$MACRO\ AVG = \frac{sum\ of\ F1\_score\ \ for\ every\ class}{number\ of\ classes} \ldots(9)$$

$$WEIGHTED\ AVG= \frac{(F1-Score\ \ for\ every\ class)*(Support\ this\ class)}{number\ of\ classes} \ldots(10)$$

يبين الجدول أن دقة النموذج المقترح جيدة وأن النموذج قادر على تصنيف العقد بالدقة المطلوبة بالتالي فان الشكل النهائي للتصنيف سيكون كما هو موضح في الشكل (4):



الشكل (4) الشكل النهائي للعقد بعد التصنيف

تم اجراء هجوم بنسبة 20.5% على 4 عقد في البيان ثم اختبار نموذج GNN (المدرب مسبقا) على البيان المهاجم (الذي تم تعديله) لينتج:

الجدول(3) دقة النموذج بعد الهجوم المطبق.

	Precision	**Recall**	**F1_Score**	Support
Case Based	0.86	0.53	0.65	298
Genetic Algorithms	0.67	0.96	0.79	418
Neural Networks	0.94	0.51	0.66	818
Probabilistic Methods	0.77	0.75	0.76	426
Reinforce_ment Learning	0.71	0.71	0.71	217
Rule Learning	0.37	0.92	0.53	180
Theory	0.59	0.68	0.63	351
ACCURACY			0.68	2708
MACRO AVG	0.70	0.72	0.68	2708
WEIGHTED AVG	0.76	0.68	0.69	2708

نلاحظ من الجدول(3):

انخفاض أداء النموذج بشكل واضح.

انخفاض دقة النموذج F1-Score لكل فئة من الفئات بشكل واضح وذلك لترابط العقد مع بعضها البعض.

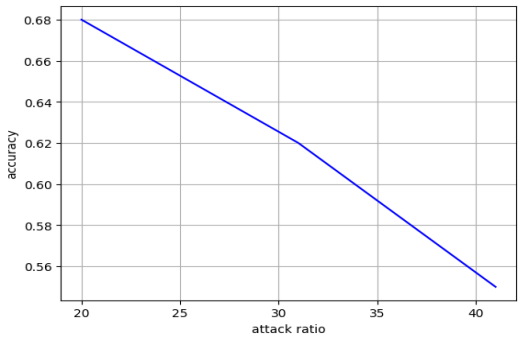
انخفاض كل من المتوسط غير المرجح والمرجح.

يبين الجدول(4) كفاءة النموذج من أجل نسب الهجوم المقترح:

الجدول (4) كفاءة النموذج وفقا لنسب الهجوم.

Attack ratio%	20	31	42
Accuracy %	0.68	0.62	0.55

يوضح الشكل (5) كفاءة النموذج وفقا لنسب الهجوم السابقة:



الشكل(5) كفاءة النموذج بعد اجراء الهجمات.

نلاحظ من الشكل(5) انخفاض كفاءة النموذج بشكل واضح وفقا لنسب الهجمات المختلفة.

يظهر الجدول(5) دقة النموذج بعد تطبيق خوارزمية التصحيح على النموذج المهاجم بنسبة الهجوم (20.5%):

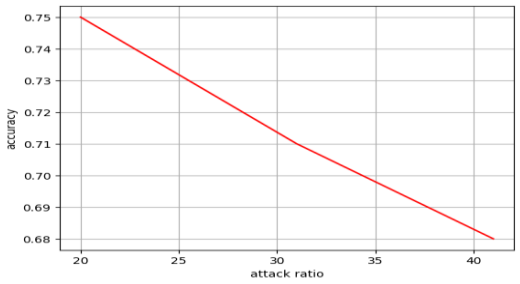
الجدول(5) دقة النموذج بعد تطبيق الخوارزمية

	Precision	Recall	F1_ Score	Support
Case Based	0.81	0.68	0.74	298
Genetic Algorithms	0.78	0.97	0.87	418
Neural Networks	0.94	0.59	0.72	818
Probabilistic Methods	0.78	0.78	0.78	426
Reinforce_ment Learning	0.76	0.80	0.78	217
Rule Learning	0.55	0.88	0.68	180
Theory	0.56	0.78	0.65	351
ACCURACY			0.75	2708
MACRO AVG	0.74	0.78	0.75	2708
WEIGHTED AVG	0.79	0.75	0.75	2708

نلاحظ من الجدول(5):

- تحسن أداء النموذج بشكل واضح.
- تحسن دقة النموذج F1-Score لكل فئة من الفئات بشكل واضح.
- تحسن كل من المتوسط غير المرجح والمرجح.

يوضح الشكل (6) كفاءة النموذج بعد تطبيق الخوارزمية على النموذج الذي تمت مهاجمته وفقا لنسب الهجوم في الجدول (6):



الشكل (6) كفاءة النموذج بعد تطبيق الخوارزمية

نلاحظ من الشكل(6) تحسن كفاءة النموذج بشكل واضح بعد تطبيق الخوارزمية.

يبين الجدول (6) معدل انخفاض كل من المتوسط المرجح (weighted avg) وغير المرجح (macro avg) على النموذج المهاجم على 4 عقد قبل تطبيق خوارزمية التصحيح وبعد تطبيقها:

الجدول (6) معدل انخفاض دقة النموذج قبل التصحيح وبعد التصحيح وفقا لـ4 عقد تمت مهاجمتها.

Ratio attack (4 nodes)	f1-score	After attack %	After correction %
3.27%	macro avg	4.7	4.1
	weighted avg	1.44	0.98
18.4%	macro avg	12.73	5.01
	weighted avg	8.92	1.97
25.43%	macro avg	15	6.16
	weighted avg	12	2.95
31.42%	macro avg	24.64	7.86
	weighted avg	19.07	4.51
49.86	macro avg	79.56	19.26
	weighted avg	73.31	13.03

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

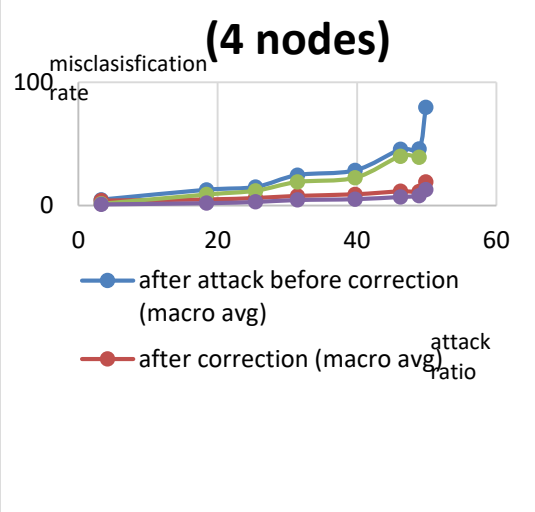
يبين الجدول (7) معدل انخفاض كل من المتوسط المرجح (weighted avg) وغير المرجح (macro avg) على النموذج المهاجم على 10 عقد قبل تطبيق خوارزمية التصحيح وبعد تطبيقها:

الجدول (7) معدل انخفاض دقة النموذج قبل التصحيح وبعد التصحيح وفقا ل10 عقد تمت مهاجمتها.

Ratio attack (10 nodes)	f1-score	After attack %	After correction %
10.81	macro avg	10.43	4.64
	weighted avg	6.61	1.62
20.81	macro avg	13.03	5.59
	weighted avg	9.7	2.57
23.79	macro avg	17.69	6.22
	weighted avg	12.78	2.98
30.01	macro avg	24.37	7.13
	weighted avg	19.73	3.94
38.14	macro avg	41.29	9.56
	weighted avg	34.88	5.87
45.44	macro avg	66.3	11.7
	weighted avg	60.3	7.87
47.96	macro avg	70.37	13.78
	weighted avg	64.38	9.29
49.88	macro avg	95.75	46.23
	weighted avg	94.81	37.4

8.مناقشة النتائج:

الشكل (7) معدل انخفاض دقة النموذج قبل تطبيق الخوارزمية وبعدها وفقا ل 4 عقد.

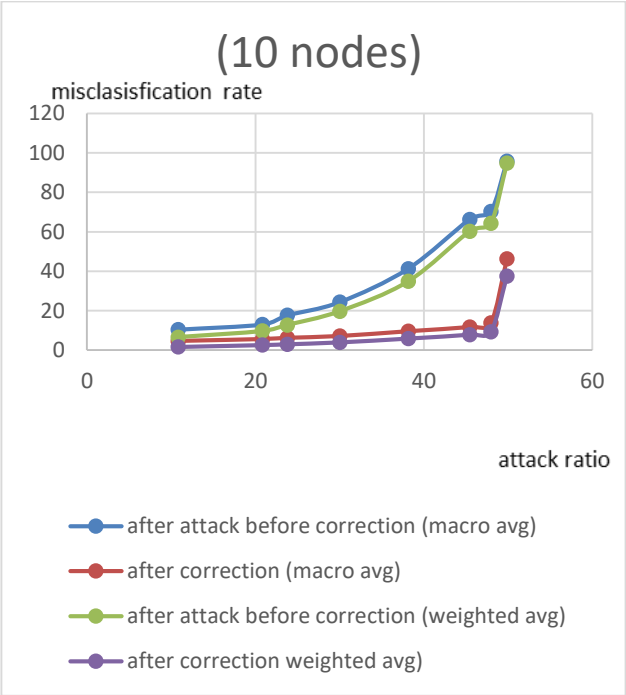


الشكل (7) معدل انخفاض دقة النموذج قبل تطبيق الخوارزمية وبعدها وفقا ل 4 عقد.

يظهر الشكل (7) أثر الهجمات السلبي على دقة النموذج اذ أن الهجوم من نسبة 39.71% يسبب انخفاض المتوسط غير المرجح بمعدل 28.49%(من النسبة % 77 قبل الهجوم). كما يظهر انخفاض المتوسط المرجح بمعدل 22.48% (من النسبة % 78 قبل الهجوم).

كما يظهر الشكل التحسن الملحوظ في دقة النموذج بعد تطبيق خوارزمية التصحيح اذ أنه من أجل نسبة الهجوم السابقة ينخفض المتوسط غير المرجح بمعدل 9.15%(من النسبة % 78 قبل الهجوم بدلا من 28.49%) والمتوسط المرجح بمعدل 5.16% (من النسبة % 78 قبل الهجوم بدلا من 22.48%). بالتالي نسبة التحسن تقريبا 19.34% للمتوسط غير المرجح و17.32% للمتوسط المرجح وفقا لنسبة الهجوم السابقة.

كما يظهر الشكل (8) معدل انخفاض دقة النموذج قبل تطبيق الخوارزمية وبعدها وفقا ل 4 عقد.



الشكل (8) معدل انخفاض دقة النموذج قبل تطبيق الخوارزمية وبعدها وفقا ل 10 عقد.

يظهر الشكل(8) أثر الهجمات السلبي على دقة النموذج اذ أن الهجوم من نسبة 30.01% يسبب انخفاض المتوسط غير المرجح بمعدل 24.37% (من النسبة 77% قبل الهجوم) كما يظهر انخفاض المتوسط المرجح بمعدل 19.73%(من النسبة 78 % قبل الهجوم).

كما يظهر الشكل التحسن الملحوظ في دقة النموذج بعد تطبيق خوارزمية التصحيح اذ أنه من أجل نسبة الهجوم السابقة ينخفض المتوسط غير المرجح بمعدل 7.13% (من النسبة 77% قبل الهجوم بدلا من 24.37%) والمتوسط المرجح بمعدل 3.94% (من النسبة 78% قبل الهجوم بدلا من 19.73%).

بالتالي نسبة التحسن تقريبا 17.13% للمتوسط غير المرجح و15.79% للمتوسط المرجح وفقا لنسبة الهجوم السابقة.

9.الاستنتاجات والآفاق المستقبلية:

تظهر الخوارزمية المقترحة امكانية تحسين متانة GNN ضد الهجمات التي تحدث على بنية البيان.

يمكن الاستفادة من بعض المعاملات الوزنية واضافتها الى خوارزمية التصحيح وملاحظة التغيرات التي قد تحدث.

تتمثل معالجة الهجوم باستعادة البيان الأصلي من البيان الذي تمت مهاجمته.

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

Computer Science and Technology, 37(5), 1161-1175.

[8] Zhang, X., & Zitnik, M. (2020). Gnnguard: Defending graph neural networks against adversarial attacks. Advances in neural information processing systems, 33, 9263-9275

[9] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI open, 1, 57-81

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

References:

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

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

[3] 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.

[4] Regol, F., Pal, S., & Coates, M. (2019, December). Node copying for protection against graph neural network topology attacks. In 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) (pp. 709-713). IEEE.

[5]Tang, X., Li, Y., Sun, Y., Yao, H., Mitra, P., & Wang, S. (2020, January). Transferring robustness for graph neural network against poisoning attacks. In Proceedings of the 13th international conference on web search and data mining (pp. 600-608).

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

[7] Wu, X. G., Wu, H. J., Zhou, X., Zhao, X., & Lu, K. (2022). Towards Defense Against Adversarial Attacks on Graph Neural Networks via Calibrated Co-Training. Journal of