مقارنة تأثير الترميز الموضعي في أنظمة كشف التسلل المعتمدة على المحولات القابلة للتفسير

م. هبه تدمري* أ.د. إبراهيم الشامي* أ.د.عمار زقزوق*

الملخص

مع تزايد تعقيد التهديدات السيبرانية، برزت نماذج المحولات (Transformers) كحل فعّال لكشف الهجمات في الشبكات الحاسوبية، استناداً إلى نجاحها في معالجة اللغة والرؤية الحاسوبية. يُعد الترميز الموضعي مكوناً أساسياً في بنية المحولات لفهم العلاقات بين الميزات وتحسين دقة الكشف. تهدف هذه الدراسة إلى تقويم تأثير أساليب الترميز الموضعي في أداء المحولات في عملية الكشف عن الهجمات من خلال مقارنة ثلاثة نماذج: بدون ترميز (NoPos) وبترميز جيبي (SinPos) وبترميز قابل للتعلم (LearnPos)، تم اختبار النماذج على مجموعتي بيانات CICIDS2017 و UNSW-NB15 مع معالجة اختلال التوازن باستخدام focal loss والأوزان اللوغاريتمية. أظهر نموذج LearnPos تفوقاً بدقة 98.55% في مجموعة بيانات CICIDS2017 و97.64% في مجموعة بيانات UNSW-NB15 متفوقاً على SinPos و NoPos. كما بيّن تحليل التفسيرية باستخدام أوزان الانتباه وتقنية LIME (Local Interpretable Model-agnostic Explanations) أن LearnPos يركّز على سمات الهجوم، ما يعزز شفافية قراراته وفعاليته في كشف الفئات النادرة مثل Web Attack، ما يجعله مناسباً لتطبيقات الأمن السيبراني الذكي.

الكلمات المفتاحية: المحولات – التشفير الموضعي – كشف هجمات الشبكة – LIME – أوزان الانتباه – الأمن السيبراني.

م. هبه تدمري طالبة دراسات عليا دكتوراه في كلية الهندسة الميكانيكية والكهربائية –قسم هندسة التحكم الالي والحواسيب.

أ.د.ابراهيم الشامي أستاذ في كلية الهندسة الميكانيكية والكهربائية –قسم هندسة التحكم الالي والحواسيب.

أ.د. عمار زقزوق أستاذ في كلية الهندسة الميكانيكية والكهربائية –قسم هندسة التحكم الالي والحواسيب.

Comparing the Impact of Positional Encoding in Interpretable Transformer-Based on Intrusion Detection Systems

Abstract

With the increasing complexity of cyber threats, transformer models have emerged as an effective solution for network intrusion detection, leveraging their success in natural language processing and computer vision. Positional encoding is a critical component in transformer for understanding feature relationships and enhancing detection accuracy. This study evaluates the impact of positional encoding strategies on transformer performance by comparing three approaches: no positional encoding (NoPos), fixed sinusoidal encoding (SinPos), and learnable positional encoding (LearnPos). The models were tested on the CICIDS2017 and UNSW-NB15 datasets, with class imbalance addressed using focal loss and logarithmic weighting. LearnPos achieved superior performance with 98.55% accuracy on CICIDS2017 and 97.64% on UNSW-NB15, outperforming SinPos and NoPos. Interpretability analysis using attention weights and LIME revealed that LearnPos focuses on attack-related features, enhancing decision transparency and effectiveness in detecting rare attack categories, such as Bot and Web Attack, making it ideal for intelligent cybersecurity applications.

Keywords: Transformers - Positional Encoding - Network Attack Detection - LIME - Attention Weights - Cybersecurity.

1- مقدمة:

أدى التسارع الكبير في التطور التكنولوجي لمؤسسات القطاعين العام والخاص إلى تطور التهديدات السيبرانية وانتشارها، ما يشكل خطراً متزايداً على الاستقرار الاقتصادي والاجتماعي. وفقاً لتقرير المنتدى الاقتصادي العالمي لأمن المعلومات 2025، تتوقع 72٪ من المؤسسات تصاعداً كبيراً في الهجمات، بينما يرى 66٪ أن الذكاء الاصطناعي سيكون عنصراً محورياً في تحسين الدفاع السيبراني [1].

تُعدّ هجمات DDoS والبوت نت من أبرز التهديدات، إذ قد تتسبب بخسائر مالية وتعطيل للخدمات، فقد تعجز أنظمة الكشف التقليدية المعتمدة على قواعد ثابتة عن رصد الهجمات المتقدمة أو غير المعروفة [2]، لذا ظهرت الحاجة إلى حلول ذكية أكثر تكيفاً مع تطور التهديدات.

برزت نماذج المحولات كحل واعد ضمن تقنيات التعلم العميق لكشف التسلل، وذلك بفضل قدرتها على التعامل مع البيانات المعقدة واستخلاص الأنماط من تسلسل الحزم. لكن تطبيقها في المجال الشبكي يواجه تحديات، خاصة فيما يتعلق بترميز الموضع (Positional Encoding)، إذ أن بيانات الشبكة لا تتبع ترتيباً زمنياً صريحاً كما في النصوص، ما يستدعي تطوير استراتيجيات ترميز فعالة لفهم العلاقات البنيوية بين الميزات [3].

بناءً على ذلك، يقترح هذا البحث نموذجاً قائماً على المحولات يعالج كل سجل بيانات بشكل مستقل، مستفيداً من آلية الانتباه الذاتي لاكتشاف التفاعلات بين الميزات داخل كل عينة، كما يُجري مقارنة بين ثلاث استراتيجيات ترميز موضعي: بدون ترميز والترميز الجيبي والترميز القابل للتعلم، على مجموعتي بيانات واقعيتين (CICIDS2017 و UNSW-NB15)، ويُدمج ذلك مع تقنيات معالجة عدم التوازن (مثل Focal Loss والأوزان اللوغاريتمية)، بالإضافة إلى أدوات تفسير (LIME وتحليل أوزان الانتباه) لتعزيز شفافية النموذج وتقديم رؤى قابلة للتفسير لمتخصصي الأمن السيبراني.

تنقسم الورقة على النحو الآتي: يقدم القسم الثاني مشكلة البحث، ويعرض القسم الثالث والرابع هدف البحث وأهميته، أما القسم الخامس فيبين النموذج المقترح، بينما يوضّح القسم السادس الإعدادات التجريبية ومجموعات البيانات، ويناقش القسم السابع النتائج وتفسيراتها، ويختتم القسم الثامن بالاستنتاجات وآفاق العمل المستقبلي.

2- مشكلة البحث:

رغم تفوق نماذج المحولات في مجالات عدة، إلا أن تطبيقها في أنظمة كشف التسلل الشبكي ما زال يواجه تحديات، أبرزها كيفية تمثيل المعلومات الموضعية داخل سجلات الشبكة. يُعد ترميز الموضع عنصراً جوهرياً في هذه النماذج، لكن أثر أنواعه المختلفة على دقة النتائج وتفسيرها لم يُدرس بعد بشكل كافٍ، ما يستدعي تحليلاً معمقاً لتأثير هذه الاستراتيجيات في سياق الأمن السيبراني.

3- هدف البحث:

يهدف هذا البحث إلى دراسة تأثير ثلاث استراتيجيات ترميز الموضع (بدون والترميز الجيبي والترميز القابل للتعلم) على أداء نماذج المحولات في كشف الهجمات، مع التركيز على الدقة والتعامل مع الفئات النادرة، وتفسير القرارات باستخدام أدوات مثل LIME وخرائط الانتباه.

4- أهمية البحث:

يُظهر هذا العمل أهمية كبيرة لأنه واحد من الدراسات القليلة التي تدرس كيفية تأثير أساليب ترميز الموضع على أداء المحولات في كشف الهجمات عبر الشبكة الحاسوبية، كما أنه يوفر رؤية تفسيرية عميقة عن سلوك النموذج وأهم الميزات اللي تؤثر فيه الشي الذي يدعم تطوير أنظمة أذكى وأكثر موثوقية في بيئات أمنية حساسة.

5- الدراسات المرجعية:

5-1 التقدم في كشف التسلل باستخدام المحولات:

تُستخدم نماذج المحولات بشكل متزايد في أنظمة كشف التسلل على الشبكات بفضل قدرتها على التقاط العلاقات المعقدة بين الميزات في البيانات عالية الأبعاد، كما أبرزت استطلاعات حديثة في التعلم الآلي [2].

اقترحت الدراسة [4] نموذج RTIDS باستخدام ترميز موضعي ثابت لتحسين أداء كشف التسلل، إذ حقق النموذج درجة F1 بنسبة 99.17% على مجموعة بيانات CICIDS2017 و 98.48% على CIC-DDoS2019، ورغم فعاليته ضد الهجمات الشائعة، فقد واجه صعوبات في اكتشاف الهجمات النادرة، وهي مشكلة شائعة في نماذج المحولات.

قدمت الدراسة [5] إطار عمل (Flow Transformer)، وهو إطار يعتمد على المحولات لكشف التسلل القائم على التدفق. يعتمد النموذج على بيانات التدفق لتحليل حركة الشبكة، مستفيداً من آليات الانتباه لالتقاط العلاقات المعقدة والسلوكيات طويلة المدى. أظهر النموذج أداءً متميزاً على مجموعات بيانات مثل NSL-KDD و CSE-CIC-IDS2018، مع تقليل التعقيد الحسابي وتحسين سرعة التدريب والاستدلال مقارنة بالنماذج التقليدية.

أما الدراسة [6] فقد اقترحت نموذج IDS-INT الهجين الذي يدمج المحولات مع CNN-LSTMs مستخدماً التعلم المنقول وتقنية SMOTE لمعالجة عدم توازن الفئات. حقق النموذج دقة 99.21% على مجموعتي بيانات UNSW-NB15 و CICIDS2017، مع تحسين الكفاءة الحسابية لتقليل استهلاك الموارد.

قدمت الدراسة [7] نموذج محولات محسّن باستخدام ترميز موضعي معدل لتعزيز تعلم التبعيات الزمنية، وقد حقق النموذج دقة تقارب 98% على CICIDS2017 ضد هجمات مثل DDoS، مع سرعة تدريب محسنة تجعله مناسباً لسيناريوهات الهجمات النادرة رغم أنهم لم يتناولوا الدراسة التفسيرية.

5-2 دور ترميز الموضع في أداء المحولات:

يُعدّ الترميز الموضعي حجر الأساس في قدرة المحولات على فهم الترتيب داخل البيانات، إلا أن بيانات الشبكة تختلف عن النصوص بغياب ترتيب زمني واضح بين الميزات ما يجعل فعالية الترميز تحدياً رئيساً. استخدمت الدراسة [4] الترميز الثابت بنجاح نسبي، في حين اقترحت الدراسة [8] أسلوباً أكثر دقة من خلال ترميز موضعي على مستوى الميزات(FPE) ، ما أدى إلى تحسين اكتشاف الهجمات النادرة. من جهة أخرى، طوّر نموذج في [9] متعدد المقاييس مع ترميز موضعي، لكنه واجه صعوبات في التصنيف متعدد الفئات، لا سيما التمييز بين الأنماط النادرة والشائعة.

5-3 أدوات التفسيرية وتحليل الانتباه:

تبرز التفسيرية كعامل حاسم في قبول نماذج كشف التسلل في البيئات الواقعية. اعتمد [10] على LIME و SHAP لشرح قرارات نموذج يعتمد على المحولات، وأظهرت النتائج أن SHAP يوفر تفسيرات عامة، بينما يقدم LIME رؤى محلية سريعة. بالمثل، استخدم [11] هذه الأدوات ضمن بيئات IoMT، ما عزز فهم تأثير الميزات، بينما قدم [12] دراسة تحليلية عامة للتفسيرية، لم تتناول أوزان الانتباه بعمق، ما يبرز الحاجة لاستكشاف تكامل أدوات التفسير مع آلية الانتباه الذاتي.

رغم التقدّم الكبير، لا تزال هناك نقطتان بحثيتان أساسيتان هما:

1- الحاجة إلى دراسة مقارنة منهجية بين استراتيجيات ترميز الموضع في بيانات الشبكة [13].
2- قلة استخدام الانتباه الذاتي كأداة تفسيرية إلى جانبLIME.

من هنا، يركّز هذا البحث على تحليل تأثير ترميز الموضع على أداء المحولات ودمج أدوات التفسير (LIME)والانتباه لتعزيز شفافية النماذج، مع معالجة مشكلة اختلال التوازن باستخدام Focal Lossوالأوزان اللوغاريتمية.

6- الإطار النظري:

6-1 المحولات:

نستخدم في هذا البحث المحولات القياسية (Encoder)، إذ إنه كافٍ لتصنيف سجلات حركة الشبكة والمكوّنة من:

- آلية الانتباه الذاتي: تُقيّم أهمية كل ميزة نسبةً إلى الأخرى، وذلك باستخدام الاستعلامات (Q) والمفاتيح (K) والقيم (V)، وتُحسب أوزان الانتباه وفق المعادلة (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \qquad (1)$$

- الطبقة الأمامية (Feed-Forward): طبقة شبكية مكونة من وحدتين كثيفتين، مزودة بتطبيع الطبقة والاتصالات المتبقية.

رغم تفوق المحولات في معالجة اللغة الطبيعية ورؤية الحاسوب، إلا أنها تواجه تحدياً في التعامل مع بيانات الشبكة غير المتسلسلة لعدم امتلاكها إدراكاً متأصلاً للترتيب على عكس RNNs و LSTM. لمعالجة ذلك، يُستخدم ترميز الموضع (Positional Encoding: PE) لإضافة معلومات الترتيب، وهو أمر جوهري لضمان التعلم الفعّال في بيانات الشبكة التي تفتقر إلى تسلسل طبيعي [14].

6-2 دور ترميز الموضع:

لا تأخذ المحولات ترتيب الرموز (الميزات) بالحسبان، لذا يُستخدم ترميز الموضع لإضافة معلومات الترتيب إلى التضمينات، ما يمكّن النموذج من فهم السياق [15].

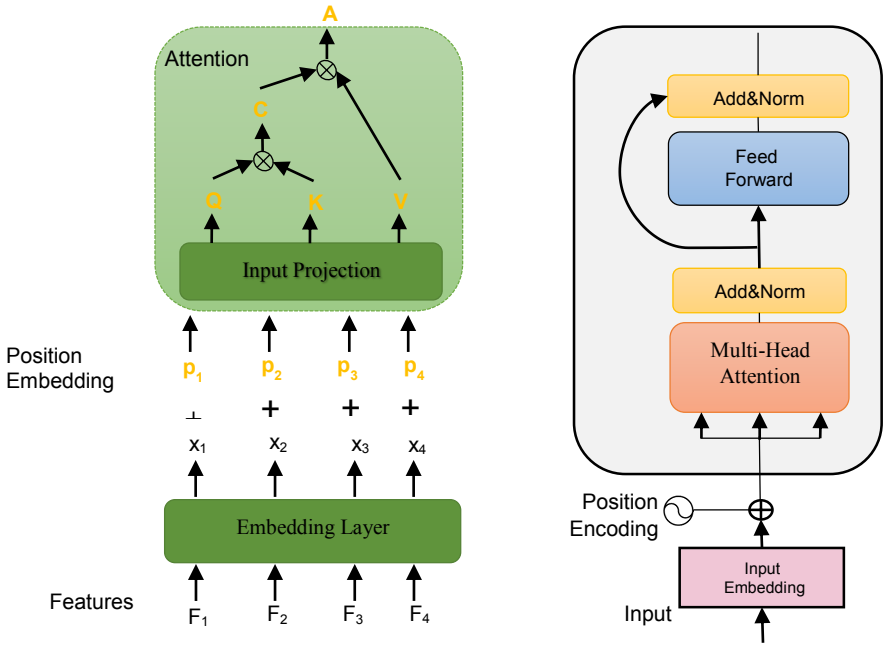
يركز بحثنا على الترميز الموضعي المطلق (Absolute Positional Encoding: APE)، إذ يُضاف متجه لكل رمز يمثل موضعه [16]، وهنا سندرس نوعين من APE هما:

- الترميز الجيبي: يُولد متجهات غير قابلة للتدريب باستخدام دوال جيبية وتجيبية، كما في المعادلة (2)، ما يساعد على تعميم التسلسلات الطويلة [13].

$$p_{pos,2i} = \sin\left(\frac{pos}{1000^{2i/d}}\right), p_{pos,2i+1} = \cos\left(\frac{pos}{1000^{2i/d}}\right), \quad (2)$$

- الترميز القابل للتعلم: يُخصص متجه موضعي قابل للتحديث لكل موقع، ويتعلم النموذج من خلاله أنماطاً موضعية خاصة بالمهمة [16].

يوضح الشكل (1) بنية المحول، إذ يتم دمج تضمينات الميزات (F) مع تضمينات الموضع (P) للحصول على الإدخال النهائي (x) المُرسل إلى وحدة التشفير [16].



الشكل (1): بنية المحول مع ترميز موضعي مطلق [14].

7- مواد البحث وطرائقه:

في بحثنا هذا، سنعمل على تقويم تأثير ثلاث استراتيجيات لترميز الموضع على أداء المحولات في كشف الهجمات باستخدام مجموعتي بيانات CICIDS2017 و UNSW-NB15.

7-1 مجموعة البيانات:

CICIDS2017: تمثل بيئة واقعية مع هجمات متنوعة مثل DDoS و Web و Bot، وتتميز بتوزيع نسبي بين الهجمات النادرة والشائعة [17].

UNSW-NB15: طُورت ببيئة هجومية متنوعة تشمل فئات مثل Analysis و Worms، وتُستخدم لتقويم التعميم [20].

اعتماد هاتين المجموعتين يُمكّن من اختبار النموذج على بيانات متنوعة من حيث التوزيع والخصائص البنيوية، ما يساعد في قياس قدرة النموذج على التكيف مع أنماط هجومية مختلفة.

تم دمج الفئات النادرة وفقاً لأساليب سابقة في معالجة بيانات التسلل، كما يبين الجدولين (1) و (2) [17]. حيث تم تطبيع جميع الميزات لضمان التوافق مع متطلبات النموذج، وتقسيم البيانات إلى 70% للتدريب و15% للتحقق و15% للاختبار باستخدام العينة الطبقية للحفاظ على نسب الفئات.

الجدول (1): توزيع الفئات في CICIDS2017.

Class	Sample Count
BENIGN	2,096,484
DoS	193,748
DDoS	128,016
Port Scan	90,819
Brute Force	9,152
Web Attack	2,143
Bot	1,953
Infiltration	36
Heartbleed	11

الجدول (2): توزيع الفئات بعد المعالجة المسبقة.

Class	Sample Count
BENIGN	2,096,484
DoS	193,748
DDoS	128,016
Port Scan	90,819
Brute Force	9,152
Web Attack	2,143
Bot	1,953
Other	47

تمت معالجة مجموعة البيانات (UNSW-NB15) بالطريقة نفسها وتقسيم البيانات بالنسب ذاتها، كما حفظنا هذه التقسيمات للمجموعتين لتدريب النماذج عليها.

يوضح الجدول (3) توزع الفئات فيها.

الجدول (3): توزيع العينات الأصلية في UNSW-NB15.

Class	Sample Count
Normal	2218760
Generic	215481
Exploits	44525
Fuzzers	24246
Dos	16353
reconnaissance	13987
Analysis	2677
Backdoor	2329
Shellcode	1511
Worms	174

7-2 معالجة عدم توازن الفئات:

للتخفيف من مشكلة عدم توازن الفئات خاصة بالنسبة للهجمات النادرة مثل البوتات وهجمات الويب، تم تطبيق دالة خسارة التركيز (Focal Loss) التي تركز على الأمثلة صعبة التصنيف من خلال تقليل وزن الأمثلة السهلة [18]، حيث (γ=0.2: معلم التركيز، α=0.25: الوزن للفئة المهيمنة)، كما اعتمدنا على تطبيق أوزان الفئات اللوغاريتمية وفق المعادلة (3)، حيث C هو إجمالي عدد العينات، و f_i هو تكرار الفئة i.

$$w_i = \log(\frac{C}{f_i}) \qquad (3)$$

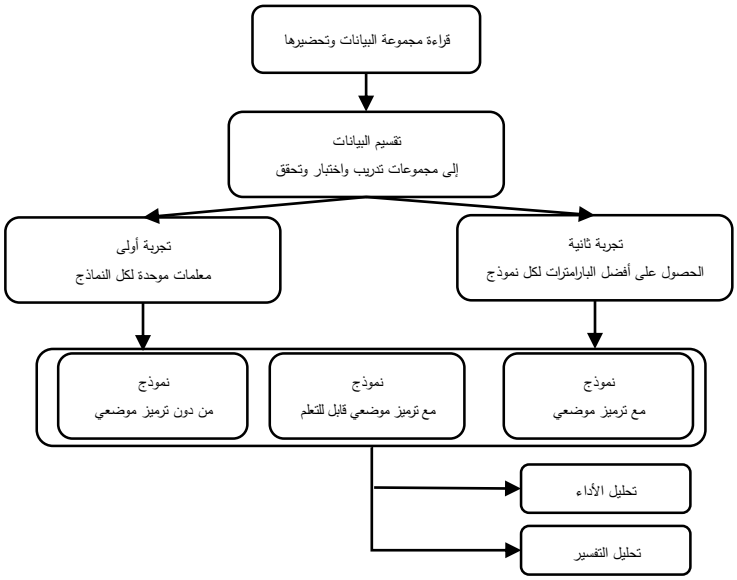
7-3 استخدام خوارزمية LIME في تفسير قرارات نموذج كشف الهجمات:

استخدامنا في بحثنا خوارزمية LIME لتفسير قرارات نموذج المحولات، كونها لا تعتمد على بنية النموذج الأصلي، بل تُنشئ نموذجاً خطياً مبسطاً حول كل حالة اختبار لتحديد الميزات الأكثر تأثيراً في التصنيف تعتمد على اختيار الميزات لتقدير أهميتها، ما يوضح سبب تصنيف تدفق معين كتهديد أو نشاط طبيعي، إذ أثبتت فعاليتها في تفسير نماذج معقدة كـ Random Forest و Decision Trees، حتى في مجالات حساسة مثل كشف خطاب الكراهية [21]، وقياساً على ذلك سنستخدمها هنا لتوضيح أثر الميزات الشبكية (مثل عدد الحزم أو البروتوكول) على قرار النموذج، ما يُعزز موثوقيته، ويُضيف طابعاً تفسيرياً مهماً في السياقات الأمنية.

4-7 الإعداد التجريبي:

كما يوضح الشكل (2)، تم اعتماد مسارين لتقييم نماذج المحولات:

1. إعداد موحد لجميع النماذج (بدون ترميز وبترميز جيبي وبترميز قابل للتعلم) باستخدام المعلمات الفائقة نفسها،والتي تم عرضها في الجدول (4).
2. إعداد محسن باستخدام أداة Keras Tuner لتخصيص المعلمات الفائقة (مثل عدد الطبقات والرؤوس و d_model و dff) لكل نموذج لتحقيق أفضل أداء، موضحة في الجدول (5).



الشكل (2): سير العمل المقترح لتقييم نماذج المحولات باستخدام إعدادات موحدة ومحسّنة.

على خلاف دراسات اعتمدت على تقنيات اختيار الميزات، فقد تم الاحتفاظ بجميع الميزات الأصلية دون تعديل، وذلك بهدف تقويم تأثير بنية النموذج والمعلمات فقط بالاعتماد على قدرة المحول على تمثيل التفاعلات المعقدة بين الميزات داخلياً [22].

الجدول (4): المعلمات الفائقة المشتركة عبر النماذج.

Parameter	Value
Number of layers	2
Number of heads	4
Hidden dimension (d_model)	64

Feed-forward size (dff)	128
Dropout rate	0.1
Learning rate (Adam)	1e-4

الجدول (5): أفضل المعلمات الفائقة لكل نموذج.

	CICIDS2017				UNSW-NB15			
Model	Layers	Heads	d_model	dff	Layers	Heads	d_model	dff
NoPos	2	8	128	256	3	8	32	128
LearnPos	3	8	32	128	3	4	64	64
SinPos	3	8	32	128	2	8	64	256

8- النتائج والمناقشة:

8-1 مقارنة الأداء العام:

تُظهر نتائج التقويم على مجموعتي البيانات CICIDS2017 و UNSW-NB15 تفوقاً واضحاً لنموذج الترميز الموضعي القابل للتعلم (LearnPos) على النموذجين الآخرين: بدون ترميز موضعي (NoPos) ومع الترميز الجيبي (SinPos)، وذلك سواءً باستخدام الإعدادات الموحدة أو المعلمات المُحسّنة.

CICIDS2017: إذ سجّل LearnPos دقة تصنيف بلغت 98.55% في الإعدادات الموحدة و 98.16% في الإعدادات المُحسّنة، مع تحقيقه لأعلى F1-score في الفئات النادرة مثل:

— Web Attack: بـ 0.8000 مقابل 0.2441 في NoPos.

— Bot: بـ 0.3480 مقابل 0.2233 في NoPos.

كما بلغ معدل ROC-AUC 0.9996، ما يعكس قدرة ممتازة على التمييز بين الفئات المختلفة.

يعرض الجدول (6) مقارنة تفصيلية لأداء النماذج حسب الفئة باستخدام كل من الإعدادات الموحدة والمحسّنة، إذ يتفوق LearnPos بشكل ملحوظ في الفئات الشائعة (مثل BENIGN و DoS) والنادرة (مثل Bot و Other).

الجدول (6). أداء النماذج حسب الفئة (F1-score) باستخدام المعلمات الموحدة والمحسّنة مجموعة بيانات CICIDS2017.

Class	SinPos (Unified)	SinPos (Optimized)	LearPos (Unified)	LearnPos (Optimized)	NoPos (Unified)	NoPos (Optimized)
BENIGN	0.9839	0.9862	0.9912	0.9889	0.9657	0.9774
Bot	0.2174	0.3125	0.1430	0.3480	0.0942	0.2233
Brute Force	0.9429	0.9553	0.9719	0.9764	0.7169	0.9515
DDoS	0.9940	0.9829	0.9993	0.9932	0.9666	0.9744
DoS	0.9721	0.9846	0.9922	0.9871	0.9124	0.9253
Other	0.0000	0.5714	0.3500	0.6000	0.0000	0.0000
Port Scan	0.8440	0.8537	0.9542	0.8537	0.8023	0.8369
WebAttack	0.2824	0.3105	0.8000	0.5717	0.2441	0.3035

UNSW-NB15: حقق LearnPos أداءً متفوقاً على مستوى: Macro F1: 0.5824 مقارنة بـ SinPos (0.5699) و NoPos (0.4674)، كما أظهر أداءً جيداً في الفئات الصعبة مثل: (Shellcode: F1 = 0.7970) و (Fuzzers: F1 = 0.5815). وقد قدم توازناً عاماً في التصنيف بين الفئات، في حين أن NoPos رغم دقته الإجمالية المرتفعة، إلا أنه تراجع بشكل كبير في الفئات قليلة التمثيل.

يوضح الجدول (7) الأداء التفصيلي لكل نموذج حسب الفئة، ما يؤكد تفوق LearnPos خاصة في الفئات التي تتطلب قدرة أعلى على التمييز، إذ يعكس تفوق LearnPos مرونة معمارية تجعله قادراً على التكيف مع بنى بيانات متنوعة، سواءً أكانت غير متوازنة (CICIDS2017) أو متوازنة نسبياً (UNSW-NB15)، ويُبرز هذا أن تصميم النموذج يلعب دوراً أكبر من مجرد تحسين المعلمات الفائقة.

الجدول (7): أداء النماذج حسب الفئة (F1-score) باستخدام المعلمات الموحدة والمحسّنة مجموعة بيانات UNSW-NB15.

Class	SinPos (Unified)	SinPos (Optimized)	LearPos (Unified)	LearnPos (Optimized)	NoPos (Unified)	NoPos (Optimized)
analysis	0.1721	0.1848	0.1582	0.1692	0.1366	0.1438
backdoor	0.1232	0.1448	0.1384	0.1757	0.0110	0.0056
dos	0.4515	0.4535	0.4502	0.4552	0.03687	0.4374

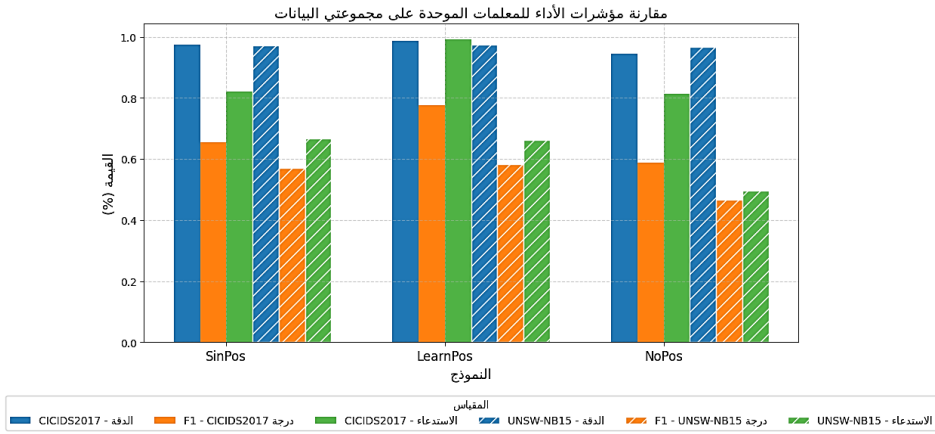
exploits	0.6380	0.6556	0.6537	0.6586	0.5808	0.5874
fuzzers	0.5755	0.5704	0.5766	0.5815	0.4472	0.4999
generic	0.9878	0.9874	0.9877	0.9886	0.9741	0.9791
normal	0.9932	0.9932	0.9932	0.9932	0.9916	0.9908
reconnaissance	0.7502	0.8122	0.803	0.8056	0.5912	0.6122
shellcode	0.7334	0.7621	0.7569	0.7970	0.4552	0.4730
worms	0.2737	0.209	0.3056	0.3226	0.1167	0.2414

رغم ذلك، لوحظ تراجع أداء LearnPos في فئة Web Attack من $F1 = 0.8000$ إلى 0.5717 بعد استخدام المعلمات المحسّنة. يُرجّح أن يكون ذلك بسبب تأثير دالة الخسارة Focal Loss، والتي تعيد توزيع الانتباه لصالح الفئات النادرة، ما يقلل من التركيز على الفئات متوسطة التكرار كما وثّقت بعض الدراسات [18–19–20].

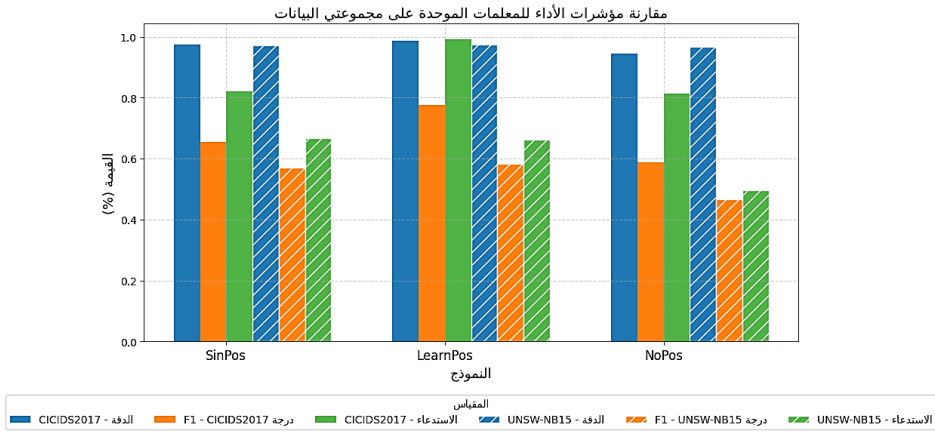
يعرض الشكل (3) مقارنة بين النماذج الثلاثة من حيث الدقة ودرجة F1 والاسترجاع باستخدام الإعدادات الموحدة، فيما يوضح الشكل (4) المقاييس نفسها مع أفضل المعلمات. أما الشكلان (5) و (6) فيبينان أخطاء التصنيف (FP و FN) لفئات حرجة مثلWeb Attack وFuzzers، إذ يظهر LearnPos بأقل عدد من الأخطاء.

8–2 تحليل التعقيد الحسابي:

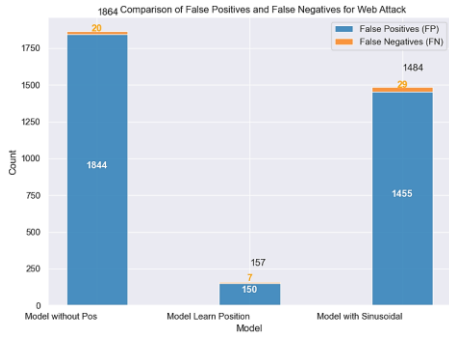
يعرض الجدول (8) مقارنة شاملة للتعقيد الحسابي بين النماذج الثلاثة باستخدام الإعدادات الموحدة للتدريب على مجموعتي البيانات CICIDS2017 و UNSW-NB15، وتُبيّن النتائج أن إدراج الترميز الموضعي — وخاصة الترميز القابل للتعلم (LearnPos) — يؤدي إلى زيادة ملحوظة في التكلفة الحسابية.



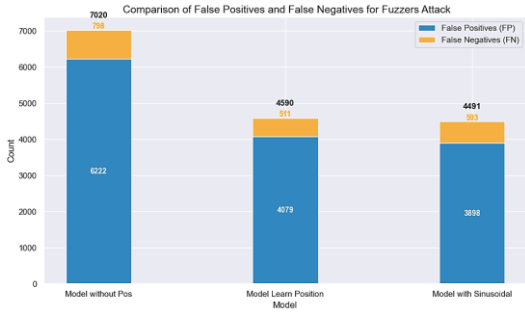
الشكل (3): مقارنة الدقة و F1 والاسترجاع تحت إعدادات موحدة.



الشكل (4): مقارنة الدقة و F1 والاسترجاع تحت إعدادات المحسنة.



الشكل (5): مقارنة أخطاء التصنيف (FP و FN) لفئة هجمات الويب عبر النماذج.



الشكل (6): مقارنة أخطاء التصنيف (FP و FN) لفئة هجمات fuzzier عبر النماذج.

الجدول (8): تحليل التعقيد الحسابي.

UNSBW			CICIDS2017			
SinPos	LearnPos	NoPos	SinPos	LearnPos	NoPos	المقياس
67,592	72,072	67,592	67,592	72,072	67,592	عدد البارامترات
5,655	6,769	4,283	6,615	10,124	4,745	الزمن الكلي للتدريب (بالثواني)
94.25	112.82	71.38	110.25	168.73	79.08	الزمن الكلي للتدريب (بالدقائق)
282.75	338.45	214.15	315.00	326.58	237.25	متوسط زمن العصور (epoch) بالثواني

يرجع ذلك إلى أن LearnPos يتطلب تحديثاً مستمراً لمتجهات الموضع أثناء كل حقبة تدريب بخلاف الترميز الجيبي (SinPos) أو غياب الترميز (NoPos)، اللذين يعتمدان على قيم ثابتة لا تحتاج إلى تعديل خلال عملية التعلم.

رغم هذه الزيادة في الزمن الكلي للتدريب وعدد المعاملات (Parameters)، فإن التحسينات الكبيرة في الأداء، خصوصاً في تصنيف الفئات النادرة، تجعل من LearnPos خياراً مُبرراً — خاصة في سياقات أمنية حرجة تتطلب دقة وموثوقية عاليتين في الكشف عن الهجمات.

8-3 تحليل أوزان الانتباه

لفهم آليات اتخاذ القرار في نماذج المحولات المستخدمة في كشف التسلل الشبكي، فقد تم تحليل أوزان الانتباه عبر خوارزمية مخصصة (خوارزمية 1) تقوم بحساب متوسط الأوزان عبر جميع الرؤوس والطبقات، ثم تُحدد أهم الميزات التي يُركز عليها النموذج خلال تصنيفه لمختلف أنواع الهجمات، وقد تم ذلك ضمن إطار الذكاء الاصطناعي التفسيري (XAI)، باستخدام خرائط حرارية ومخططات شريطية لتصوّر توزيع الانتباه وتأثير الترميز الموضعي عليه.

تشير الدراسات السابقة [14] إلى أن آليات الانتباه في نماذج المحولات قادرة على تحسين تمثيل البيانات من خلال التركيز على عناصر ذات دلالة مثل أنماط المنافذ أو خصائص تدفق الحزم، وهذا ما يُسهم في اكتشاف الأنماط المعقدة في حركة مرور الشبكة.

في هذا السياق، أجري مع الحفاظ على المعلمات (الموحدة) نفسها لتدريب النماذج، ما يضمن مقارنة عادلة لتركيز الانتباه ومدى تفسيره لسلوك النموذج.

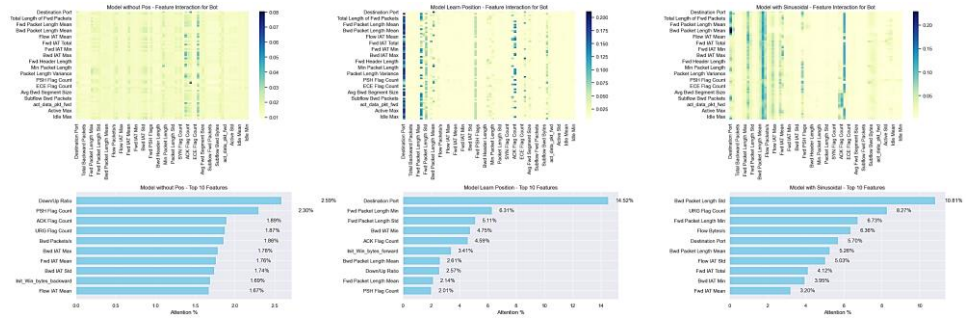
```
Input:
    - Trained transformer model M
    - Input sample X (shape: [batch_size, sequence_length, features])
    - Feature names list F = [f₁, f₂, ..., f□]
Output:
    - Attention matrix A (feature-to-feature interaction)
    - Feature importance vector I (as percentage)
Steps:
1. Forward X through the model M with attention tracking enabled:
       → Obtain attention_weights from M(X, return_attention=True)
2. Stack all attention heads and layers:
       → W ← stack(attention_weights)
         (shape: [num_layers, num_heads, seq_len, seq_len])
3. Average over all layers and heads:
       → W_mean ← mean(W, axis=(0, 1))
         (shape: [seq_len, seq_len])
4. Construct attention matrix A using W_mean:
       → A ← DataFrame(W_mean, rows=F, columns=F)

5. Compute total attention received per feature:
       → total ← sum(A, axis=0)
6. Normalize into percentage importance:
       → I ← 100 × total / sum(total)
7. Return:
       → A, I
```

خوارزمية 1: استخراج مصفوفة الانتباه ونسب التأثير لكل ميزة.

8-3-1 التحليل على مجموعة بيانات CICIDS2017:

- **هجوم Bot:**

يُظهر الشكل (7) أن نموذج LearnPos ركّز بنسبة كبيرة على ميزة منفذ الوجهة (14.52%)، وهي ميزة حاسمة في الكشف عن الأنماط المتكررة للاتصال المرتبط بشبكات البوت. كما أظهر انتباهاً موجهاً نحو الحد الأدنى لطول الحزمة الأمامية (6.31%) وعدد أعلام ACK (4.59%)، ما يعكس فهماً منظماً لسلوك الهجوم.

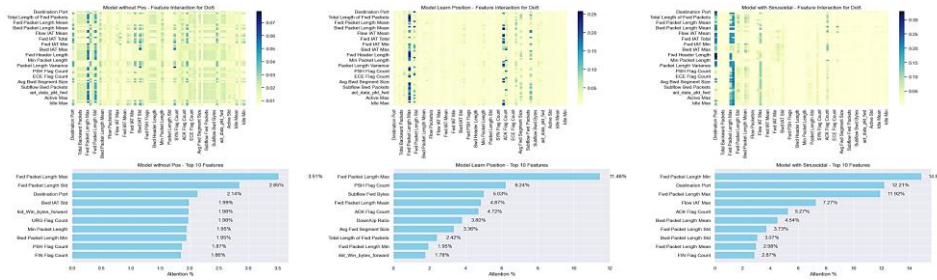


الشكل (7): خرائط حرارية الانتباه لهجوم Bot.

في المقابل، أظهر نموذج NoPos توزيع انتباه مشتت مركزاً بنسبة 2.59% فقط على نسبة التحميل/الرفع (Down/Up Ratio)، وهي ميزة ضعيفة الدلالة ما أدى إلى انخفاض أداء الكشف. أما نموذج SinPos فركّز على ميزات مثل الانحراف المعياري لطول الحزم العكسية (10.81%) وعدد أعلام URG (8.27%)، لكن افتقاره إلى التركيز على المنافذ تسبب في انخفاض F1-Score بنسبة 8% مقارنة بـ LearnPos.

- **هجوم DoS:**

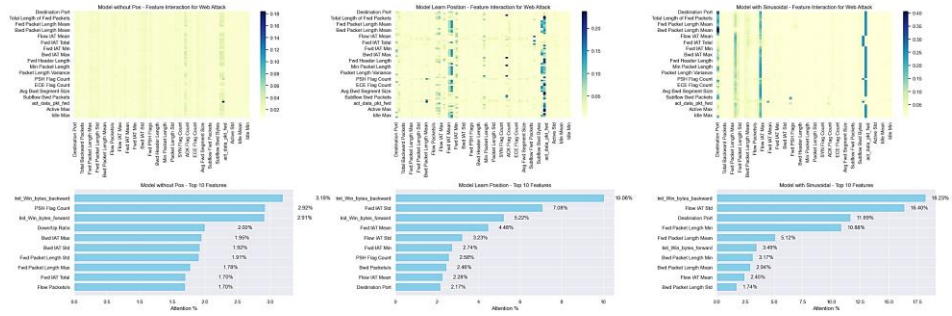
أظهر LearnPos فهماً تفسيرياً واضحاً من خلال تركيزه على ميزات مثل الحد الأقصى لطول الحزمة (11.46%) وعدد أعلام PSH (6.24%)، مما يعكس قدرته على التقاط أنماط هجمات الفيضان وساعده في خفض معدلات السلبيات الكاذبة. في المقابل ركّز SinPos على Flow IAT Max (7.27%) ومنفذ الوجهة (12.21%)، لكنه لم يُظهر تركيزاً مستقراً. أما NoPos، فلم تتجاوز أقوى ميزاته نسبة 3.51%، ما يشير إلى تشتت واضح في الانتباه، كما هو موضح في الشكل (8).



الشكل (8): خرائط حرارية الانتباه لهجمات DoS.

- **هجوم الويب (Web Attack):**

تفوق LearnPos تفسيرياً عبر تركيزه على ميزة التباين في وقت الوصول الأمامي (Fwd IAT Std) بنسبة (7.08%)، والتي تساعد في رصد الاضطرابات الزمنية المرتبطة بسلوك هجمات الويب، رغم أن SinPos أعطى وزناً مرتفعاً لميزة Init_Win_bytes_backward (18.23%)، إلا أن هذه الميزة غير مرتبطة سلوكياً بشكل مباشر بهجمات الويب، ما يُفسر أداءه المتواضع. أما NoPos فافتقر إلى التنظيم وركز على ميزات غير تفسيرية مثل PSH Flag Count و Down/Up Ratio وهذا ما يعرضه الشكل(9).



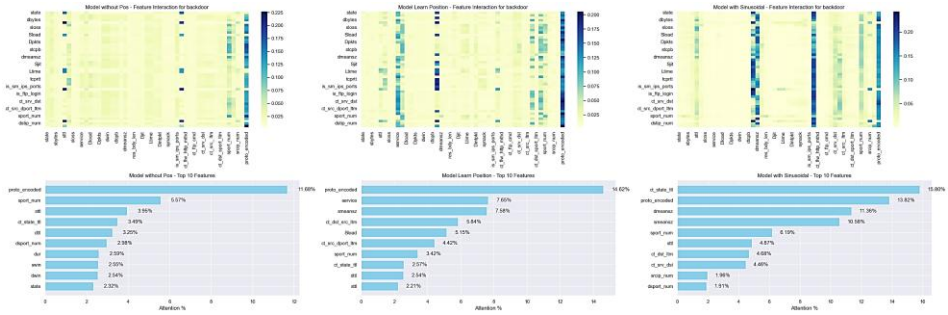
الشكل (9): خرائط حرارية الانتباه لهجمات الويب.

8-3-2 التحليل على مجموعة بيانات UNSW-NB15:

- **هجمات (backdoor):**

أظهر نموذج LearnPos تفوقاً في كشف هجمات Backdoor، إذ ركّز على ميزات تفسيرية مثل proto_encoded (ترميز البروتوكول مثل TCP/UDP بنسبة 14.62%) و service (نوع الخدمة كـ HTTP/SSH)، ما يعكس قدرته على التقاط أنماط الاتصال الخفي. في المقابل، ركّز SinPos على ct_state_ttl (عدد حالات TTL في التدفق الزمني بنسبة تركيز 15.80%)، ما يبرز فعاليته في تمثيل الخصائص الزمنية لكنه افتقر إلى التركيز على ميزات السياق مثل المنافذ، ما أدى إلى انخفاض F1-score (0.1232) مقارنة بـ LearnPos (0.1360). أما NoPos فأظهر توزيع انتباه مشتت (2.59% على نسبة التحميل/الرفع)، ما تسبب في أداء ضعيف (F1-score: 0.0110).

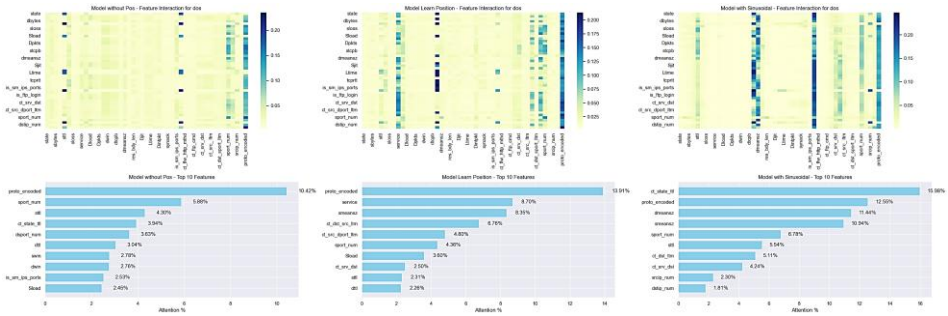
يوضح الشكل (10) توزيع الانتباه المتوازن لـ LearnPos، ما يبرر تفوقه في كشف الأنماط الخبيثة.



الشكل (10): خرائط حرارية الانتباه لهجمات backdoor.

- **هجمات DoS:**

كرر LearnPos نمط التركيز على proto_encoded و service، بينما برز SinPos مجدداً من خلال ct_state_ttl، بينما فشل النموذج NoPos في تشكيل تركيز تفسيري واضح، ما قلل من فعاليته انظر الشكل (11).



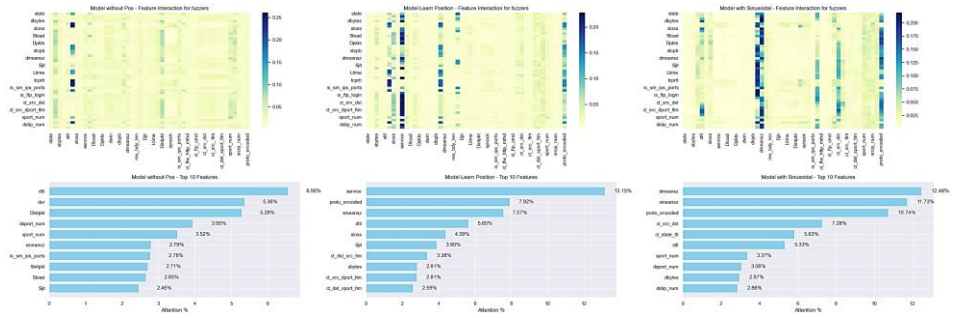
الشكل (11): خرائط حرارية الانتباه لهجمات dos.

- **هجوم Fuzzers:**

أظهر LearnPos مرونة تفسيرية قوية من خلال تركيزه على service و proto_encoded، ما مكنه من التعامل مع أنماط البيانات العشوائية ومن تمييز البروتوكولات والخدمات المستهدفة بالبيانات العشوائية. بينما ركّز نموذج SinPos على smeansz (متوسط حجم حزم المصدر) و dmeansz (متوسط حجم حزم الوجهة)، ما يعكس حساسيته لتغيرات حجم

الحزم، لكنه كان أقل اتساقاً بسبب ضعف تركيزه على ميزات السياق. أما NoPos فأظهر توزيع انتباه مشتت ما يفسر أداؤه الضعيف.

يبين الشكل (12) تفوق LearnPos في التقاط الأنماط العشوائية لهجمات Fuzzers بفضل فهمه للسياق الوظيفي للبروتوكولات والخدمات.



الشكل (12): خرائط حرارية الانتباه لهجمات Fuzzers.

يُظهر هذا التحليل أن استخدام الترميز الموضعي القابل للتعلم (LearnPos) لا يحسّن مؤشرات الأداء العددي فقط بل يُعزز تفسير قرارات النموذج بشكل واضح. ويوفر LearnPos تنظيماً بنيوياً أعلى لتركيز الانتباه عبر فئات هجومية متنوعة ومجموعات بيانات مختلفة.

في المقابل، تعاني النماذج الأخرى (NoPos و SinPos) من ضعف في التنظيم والانتباه التفسيري، ما يبرز أهمية LearnPos في تطوير نماذج تفسيرية قابلة للتعميم في سياقات أمن الشبكات.

4-8 تفسير النموذج عبر LIME:

يُعد السؤال "ما الذي جعل النموذج يتخذ هذا القرار؟" حاسماً للتحقق من صحة نماذج التصنيف في كشف التسلل على الشبكات. في هذا السياق، تمت مقارنة حالتين متباينتين لاختبار قدرات النماذج التفسيرية هما:

<u>الحالة الأولى</u>: تصنيف واضح لعينة Web Attack (من CICIDS2017).

تم اختيار عينة مصنفة كهجوم Web Attack من مجموعة بيانات (CICIDS2017) لتحليل آلية اتخاذ القرار، وقد صنفت النماذج العينة بثقة عالية (تجاوزت 90%)، وأظهرت LIME كما يوضح الشكل (13) أن السمات المؤثرة تضمنت مؤشرات زمنية مثل Flow IAT Mean و Flow

Packets/s، والتي تعكس سلوكاً هجومياً نمطياً. لوحظ تطابقاً واضحاً بين نتائج LIME وتحليل أوزان الانتباه، ما يشير إلى تفسير واضح للقرار واستناد النموذج إلى أنماط قابلة للفهم.

الشكل (13): مقارنة تفسيرات LIME وأوزان الانتباه لهجوم Web Attack عبر النماذج الثلاثة.

هذا التوافق بين نتائج LIME والتحليل الداخلي للنموذج يعزز موثوقية التفسير، ويؤكد أن النموذج لم يتخذ قراره بشكل عشوائي، بل استند إلى إشارات واضحة وقابلة للفهم ضمن البيانات.

<u>الحالة الثانية:</u> تصنيف صحيح منخفض الثقة Backdoor:

كشفت نتائج تفسير النموذج باستخدام LIME عن غياب سمات تفسيرية مباشرة، حيث أظهرت جميع السمات المفعّلة تأثيراً ضعيفاً جداً واقتصرت على "منطقة التأثير الصفري"، مما يعكس افتقار النموذج إلى إشارات واضحة تدعم قراره. وعلى الرغم من أن العينة صُنّفت بشكل صحيح، إلا أن الثقة بقيت منخفضة نسبياً. يُظهر الشكل (14) هذا التوجه حيث يتوزع الانتباه عبر سمات متعددة منخفضة الأهمية دون نمط واضح أو تركيز مهيمن، لا سيما في نموذجي Learned Pos و Sin Pos، مما يُشير إلى اعتماد النماذج على إشارات خفية وغير مستقرة يصعب تفسيرها.

الشكل (14): مقارنة تفسيرات LIME وأوزان الانتباه لهجوم back door عبر النماذج الثلاثة.

تُظهر هذه النتائج أن قابلية تفسير قرارات النموذج تختلف باختلاف طبيعة الهجوم، إذ تكون الهجمات النمطية أو الزمنية (مثل Web Attack) أسهل في التفسير، بينما تتطلب الهجمات المتخفية (مثل Backdoor) أدوات تفسير أعمق نظراً لتعقيد أنماطها.

9- الاستنتاجات والتوصيات:

أظهرت الدراسة أهمية ترميز الموضع في تعزيز أداء نماذج المحولات ضمن أنظمة كشف التسلل، وخاصة في كشف الهجمات المعقدة أو النادرة. تفوق نموذج LearnPos الذي يستخدم ترميزاً موضعياً قابلاً للتعلم، من حيث الدقة و F1-score على نموذجي NoPos و SinPos في كل من CICIDS2017 و UNSW-NB15. كما ساهم استخدام Focal Loss وأوزان الفئات في تحسين التعامل مع عدم توازن البيانات، ما انعكس في أداء مميز على الفئات النادرة مثل Bot و Web Attack.

عززت أدوات التفسير (LIME وأوزان الانتباه) شفافية النموذج مظهرةً اعتماده على سمات واضحة (كمنفذ الوجهة و Flow IAT)، ما يدل على أن LearnPos يجمع بين الفعالية التنبؤية وقابلية التفسير.

انطلاقاً من النتائج الجيدة التي تم الحصول عليها في هذا البحث، يمكننا اقتراح مجموعة النقاط التي قد تسهم في تحسين أداء أنظمة كشف التسلل المعتمدة على المحولات:

- تحسين بنية المحول (مثل زيادة العمق أو الرؤوس أو استخدام آليات انتباه مخصصة).
- تطوير ترميزات هجينة أو موجهة بالمجال.
- اعتماد أساليب ديناميكية لمعالجة عدم التوازن (مثل: SMOTE أو GAN).

تؤكد الدراسة أن استخدام ترميز موضعي قابل للتعلم لا يحسّن فقط أداء النموذج، بل يعزز أيضًا شفافيته وملاءمته للتطبيق العملي في الأمن السيبراني.

10- المراجع:

1. World Economic Forum. (2025). Global Cybersecurity Outlook 2025. Retrieved from: https://www.weforum.org/publications/global-cybersecurity-outlook-2025/
2. Momand, A., Jan, S. U., & Ramzan, N. (2023). A systematic and comprehensive survey of recent advances in intrusion detection systems using machine learning: Deep learning, datasets, and attack taxonomy. Journal of sensors, 2023(1), 6048087.
3. Corea, P. M., Liu, Y., Wang, J., Niu, S., & Song, H. (2024, July). Explainable AI for comparative analysis of intrusion detection models. In 2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom) (pp. 585-590). IEEE.
4. Wu, Z., Zhang, H., Wang, P., & Sun, Z. (2022). RTIDS: A robust transformer-based approach for intrusion detection system. IEEE Access, 10, 64375-64387.
5. Manocchio, L. D., Layeghy, S., Lo, W. W., Kulatilleke, G. K., Sarhan, M., & Portmann, M. (2024). Flow transformer: A transformer framework for flow-based network intrusion detection systems. Expert Systems with Applications, 241, 122564.
6. Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C. W. (2024). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. Digital Communications and Networks, 10(1), 190-204.
7. Liu, Y., & Wu, L. (2023). Intrusion detection model based on improved transformer. Applied Sciences, 13(10), 6251.

8. Çavşi Zaim, H., & Yolaçan, E. N. (2025). FPE–Transformer: A Feature Positional Encoding–Based Transformer Model for Attack Detection. Applied Sciences, 15(3), 1252.
9. Xi, C., Wang, H., & Wang, X. (2024). A novel multi–scale network intrusion detection model with transformer. Scientific Reports, 14(1), 23239.
10. Gaspar, D., Silva, P., & Silva, C. (2024). Explainable ai for intrusion detection systems: Lime and shap applicability on multi–layer perceptron. IEEE Access.
11. Kalakoti, R., Sharma, M., & Bhattacharya, S. (2024). Explainable transformer–based intrusion detection for IoMT networks. Journal of Medical Systems, 48(5), Article 20. https://doi.org/10.1007/s10916–024–02054–5.
12. Patil, S., Varadarajan, V., Mazhar, S. M., Sahibzada, A., Ahmed, N., Sinha, O., ... & Kotecha, K. (2022). Explainable artificial intelligence for intrusion detection system. Electronics, 11(19), 3079.
13. Dufter, P., Schmitt, M., & Schütze, H. (2022). Position information in transformers: An overview. Computational Linguistics, 48(3), 733–763.
14. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. (2017). Attention is all you need. Advances in neural information processing systems, 30.
15. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self–attention with relative position representations. arXiv preprint arXiv:1803.02155.

16. Zhao, Liang, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. (2023). Length extrapolation of transformers: A survey from the perspective of positional encoding. arXiv preprint arXiv:2312.17044.

17. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp, 1(2018), 108-116.

18. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

19. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of big data, 6(1), 1-54.

Moustafa, N., & Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In 2015 military communications and information systems conference (MilCIS) (pp. 1-6). IEEE.

20. م. ربيع محي الدين الكردي، د. كمال سلوم.، د. وسيم رمضان (2024). تفسير نموذج مُدرب لاكتشاف خطاب الكراهية في التغريدات العربية. مجلة جامعة حمص، سلسلة العلوم الهندسية الميكانيكية والكهربائية والمعلوماتية، (1)46.

21. م. علي ياسين، د. كمال سلوم.، د. وسيم رمضان (2022). تحديد عتبة التصنيف المثلى ديناميكياً في أنظمة الكشف المبكر عن الشذوذ القائمة على التعلم العميق. سلسلة العلوم الهندسية الميكانيكية والكهربائية والمعلوماتية،(4)44.