التزييف العميق: عندما يصبح الخداع واقعاً

Deepfakes: When deception becomes reality

Malek Algabri
dr.malekye@eiu.edu.ye
malekye@su.edu.ye

Muhannad. I.N Faqya
mohahd.2000@hotmail.com

Dr. Nasser H. Almofari3
almofaryn@gmail.com

Racha Mohamadi4
Mohamadiracha47@gmail.com

الملخص:

تشهد بيئة الأمن السيبراني تحولاً جذرياً مع ظهور ما يُعرف بـ "الهندسة الاجتماعية 2.0"، حيث لم تعد الهجمات تعتمد على التلاعب النصي فحسب، بل تجاوزته لتوظف تقنيات الذكاء الاصطناعي التوليدي، وتحديداً "التزييف العميق" (Deepfakes)، لخلق وسائط سمعية وبصرية فائقة الواقعية. تهدف هذه الورقة البحثية إلى تحليل المخاطر الناشئة عن دمج التزييف العميق في هجمات الهندسة الاجتماعية، مع التركيز على القصور الحالي في آليات الدفاع التقليدية التي تعجز عن رصد التزييف في البيئات المضغوطة وقنوات الاتصال الحية. من خلال مراجعة نقدية للأدبيات الحديثة، واقتراح إطار عمل دفاعي هجين (Hybrid Defense Framework)، تناقش الدراسة أهمية دمج التحليل التقني الآلي مع التحقق السلوكي البشري. وتخلص النتائج إلى أن الحلول التقنية وحدها غير كافية في ظل تطور الشبكات التوليدية التنافسية (GANs)، موصيةً بتبني استراتيجيات "الثقة الصفرية" والتحقق خارج النطاق كضرورة حتمية لحماية الأصول المؤسسية والمجتمعية في عصر الوسائط الاصطناعية.

الكلمات المفتاحية

التزييف العميق الهندسة الاجتماعية، الأمن السيبراني الوسائط الاصطناعية، استنساخ الصوت، الدفاع الهجين.

Abstract:

The cybersecurity landscape is undergoing a radical transformation with the emergence of "Social Engineering 2.0." Attacks no longer rely solely on text manipulation but have expanded to employ generative artificial intelligence techniques, specifically "deepfakes," to create hyper-realistic audio-visual media. This research paper aims to analyze the risks arising from the integration of deepfakes in social engineering attacks, focusing on the current shortcomings of traditional defense mechanisms that fail to detect deepfakes in compressed environments and live communication channels. Through a critical review of recent literature and the proposal of a hybrid defense framework, the study discusses the importance of combining automated technical analysis with human behavioral verification. The findings conclude that technical solutions alone are insufficient in light of the development of competitive generative networks (GANs), recommending the adoption of "zero trust" strategies and out-of-band verification as essential for protecting institutional and societal assets in the age of artificial media.

Keyword:

Deepfakes, Social Engineering, Cybersecurity, Synthetic Media, Voice Cloning, Hybrid Defense

المقدمة:

تلاشي الحدود بين الواقع والزيف في عصر "الهندسة الاجتماعية 2.0 " لطالما شُكّل العنصر البشري الحلقة الأضعف والثغرة الأكثر استعصاءً على الترقيع في منظومة الأمن السيبراني. فعلى مدار العقود الماضية، ركزت استراتيجيات الدفاع التقليدية على تحصين الشبكات وسد الثغرات البرمجية، بينما ظلت "الهندسة الاجتماعية" (Social Engineering) تعتمد في جوهرها على استغلال التحيزات المعرفية والأخطاء البشرية والثقة المفرطة لتحقيق اختراقات لا تتطلب بالضرورة مهارات برمجية معقدة. ومع ذلك، فإننا نشهد اليوم منعطفاً تاريخياً في طبيعة هذه التهديدات، حيث لم تعد محاولات الاحتيال مقتصرة على رسائل البريد الإلكتروني ركيكة الصياغة أو المكالمات الهاتفية المشبوهة، بل دخلت حقبة جديدة يمكن وصفها بـ "الهندسة الاجتماعية 2.0 "، وهي حقبة يقودها الذكاء الاصطناعي والتعلم العميق.

وقبل الخوض في تحليل هذه الظاهرة، من الضروري تعريف المفاهيم الخاصة بهذا البحث. يُقصد بـ الهندسة الاجتماعية (Social Engineering) هو فن التلاعب النفسي بالبشر لدفعهم نحو اتخاذ إجراءات أو الإفصاح عن معلومات سرية، مستغلين بذلك التحيزات المعرفية والثقة المفرطة. أما التزييف العميق (Deepfakes)، فهو مصطلح يطلق على الوسائط المتعددة التي تم صنعها أو تعديلها باستخدام تقنيات التعلم العميق (Deep Learning)، وتحديداً الشبكات التوليدية التنافسية (GANs)، لإنتاج محتوى يبدو واقعياً ولكنه غير حقيقي (Ian et al., 2014). ويندرج هذا تحت الوسائط الاصطناعية (Synthetic Media)، التي تشمل أي محتوى تم إنشاؤه أو تعديله خوارزمياً. وفي سياق بحثنا، نركز أيضاً على استنساخ الصوت (Voice Cloning)، وهي تقنية تسمح بمحاكاة نبرة وتينور صوت شخص معين بدقة عالية باستخدام عينات صوتية قصيرة، مما يشكل حجر الزاوية في الهجمات الصوتية الحديثة.

إن الطفرة الهائلة في تقنيات "الوسائط الاصطناعية" (Synthetic Media)، وتحديداً التزييف العميق (Deepfakes)، أحدثت تغييراً جذرياً في قواعد اللعبة. فبعد أن كانت قدرات التلاعب بالصوت والصورة حكراً على استوديوهات السينما الكبرى ذات الميزانيات الضخمة، أتاحت الشبكات التوليدية التنافسية (GANs) إمكانيات خطيرة لهذه الأدوات، مما مكن المهاجمين من توليد محتوى سمعي وبصري يتسم بواقعية مرعبة وتكلفة زهيدة. هذا التحول التكنولوجي، كما توضح الدراسات الحديثة، يعيد تعريف مفهوم التهديد السيبراني، حيث يتم دمج قدرات الذكاء الاصطناعي مع تكتيكات الهندسة الاجتماعية التقليدية لخلق هجمات مخصصة وشخصية للغاية يصعب على الحواس البشرية المجردة تمييزها (Federal Bureau of, 2021).

وتكمن الخطورة الحقيقية لهذا الدمج في قدرته على ضرب الركيزة الأساسية للتواصل البشري والمهني: وهي "الثقة في ما نراه ونسمعه". فلم يعد الهجوم يهدف فقط إلى سرقة البيانات، بل يمتد إلى انتحال الهوية البايومترية للأفراد.

فإن البيئات المؤسسية تواجه اليوم سيناريوهات معقدة مثل "الاحتيال عبر الرئيس التنفيذي" (CEO Fraud)، حيث يمكن لمكالمة فيديو مزيفة أو رسالة صوتية مفبركة بدقة أن تتجاوز أكثر البروتوكولات الأمنية صرامة، مستغلة السلطة الوظيفية وسياق العمل العاجل.

علاوة على ذلك، لا تتوقف التهديدات عند حدود التزييف الكامل، بل تمتد لتشمل تقنيات أكثر دهاءً ومكرًا مثل "الهجمات الصوتية الجزئية" (Partial Fake Speech)، حيث يتم التلاعب بجمل محددة ضمن سياق حقيقي، مما يجعل اكتشاف التزييف تحدياً تقنياً هائلاً. وهذا التطور ينقل المعركة من مجرد حماية "كلمات المرور" إلى حماية "بصماتنا الحيوية" وأصواتنا ووجوهنا (Shruti et al., 2019).

وبعيداً عن السياق المؤسسي، فإن التأثيرات المجتمعية لهذه التقنية لا تقل خطورة. إذ تلعب منصات التواصل الاجتماعي دور الحاضنة والمسرع لانتشار هذا النوع من التضليل، مما يؤدي إلى تآكل الثقة المجتمعية وانتشار المعلومات المضللة بسرعة تفوق قدرة آليات الكشف التقليدية على الاحتواء (معروف, 2022a).

تأسيساً على ما سبق، تهدف هذه الورقة البحثية إلى تسليط الضوء على ظاهرة "التزييف العميق كأداة للهندسة الاجتماعية"، من خلال استعراض نقدي للأدبيات الحديثة، وتحليل الآليات التقنية والنفسية التي تجعل هذه الهجمات فعالة، وصولاً إلى مناقشة استراتيجيات الدفاع المستقبلي في ظل بيئة رقمية لم يعد فيها "الرؤية" دليلاً كافياً على "التصديق" (Mika, 2019).

المشكلة وأهداف البحث:

تكمن المشكلة الجوهرية التي تعالجها هذه الدراسة في الفجوة المتسعة بين التطور المتسارع لتقنيات توليد الوسائط المزيفة (Generation) والقدرات المحدودة لتقنيات الكشف (Detection). فبينما كانت الهندسة الاجتماعية التقليدية تستغل "سذاجة" الضحية أو قلة انتباهها، فإن الهجمات الحديثة المدعومة بالتزييف العميق تستغل "الحواس الأساسية" للإنسان (البصر والسمع)، مما يجعل الضحية تشك في ذاكرتها لا في المحتوى المعروض أمامها.

وعلى الرغم من وفرة الحلول التقنية المقترحة في الأدبيات السابقة، إلا أن معظمها يعاني من قصور عملي عند التطبيق في بيئات حقيقية؛ حيث تؤدي عمليات ضغط الفيديو والصوت في تطبيقات التواصل (مثل Zoom أو WhatsApp) إلى طمس الآثار الرقمية (Artifacts) التي تعتمد عليها خوارزميات الكشف، مما يرفع معدلات الخطأ ويسمح بتمرير هجمات معقدة مثل "الاحتيال عبر الرئيس التنفيذي". وعليه، تتحدد مشكلة البحث في السؤال الرئيسي التالي: كيف يمكن بناء استراتيجية دفاعية فعالة تتجاوز الاعتماد التقني المحض، للكشف عن هجمات الهندسة الاجتماعية متعددة الوسائط في البيئات المؤسسية؟

تسعى هذه الدراسة إلى تحقيق جملة من الأهداف المترابطة التي تصب في تعزيز المنظومة الأمنية ضد التهديدات المستحدثة:

1. التحليل النقدي للتطور التاريخي: تتبع انتقال الهندسة الاجتماعية من الأساليب التقليدية إلى عصر "الذكاء الاصطناعي الهجومي"، وتصنيف أنواع التهديدات الجديدة (مثل التزييف الصوتي الجزئي والفيديو التفاعلي).
2. تقييم فاعلية أدوات الكشف الحالية: تحديد نقاط الضعف في الأنظمة الدفاعية القائمة، خاصة فيما يتعلق بقدرتها على التعامل مع الوسائط المضغوطة والهجمات الحية (Real-time attacks).
3. تصميم إطار عمل مقترح (Proposed Framework): تطوير نموذج دفاعي يدمج بين تقنيات الكشف الآلي والتحقق السلوكي البشري، لتقليل نسبة الإنذارات الخاطئة وسد الثغرات التي لا تستطيع البرمجيات وحدها تغطيتها.
4. صياغة توصيات عملية: تقديم خارطة طريق للمؤسسات وصناع القرار لتبني سياسات أمنية استباقية (مثل التحقق متعدد القنوات) لمواجهة مخاطر انتحال الشخصية الرقمية.

الدراسات السابقة:

1. الدراسة الأولى:

العنوان: الهندسة الاجتماعية المدفوعة بتقنية التزييف العميق والهجمات المولدة بالذكاء الاصطناعي: إعادة تعريف التهديدات والدفاعات في عصر الوسائط الاصطناعية (Deepfake-Driven Social Engineering and AI-Generated Attacks: Redefining Threats and Defenses in the Age of Synthetic Media).

قدمت هذه الدراسة إطاراً نظرياً شاملاً لفهم التحول في طبيعة التهديدات. ويمكن تلخيص أبرز إسهاماتها في النقاط التالية:

1. إعادة تعريف التهديد: أوضحت كيف ينقل التزييف العميق التهديدات من "المحتوى الثابت" (نصوص وصور) إلى "الوسائط الاصطناعية الديناميكية".
2. دمج الهجمات: حللت الدراسة كيفية دمج التزييف العميق مع هجمات التصيد الاحتيالي لزيادة معدلات النجاح.
3. تجاوز الدفاعات التقليدية: أكدت أن برمجيات الحماية الحالية عاجزة عن اكتشاف المحتوى المولد بالذكاء الاصطناعي لأنه لا يحمل "توقيعاً برمجياً ضاراً" (Malware Signature) بالمعنى التقليدي.
4. الأثر النفسي: ركزت على استغلال "التحيز المعرفي" لدى الضحايا، حيث يميل البشر لتصديق ما يرونه بأعينهم (Seeing is believing).
5. سيناريوهات الهجوم: استعرضت سيناريوهات لابتزاز الأفراد وتشويه السمعة كأدوات ضغط في الهندسة الاجتماعية.
6. استراتيجيات التخفيف: اقترحت إطار عمل يعتمد على "المصادقة متعددة الوسائط" كحل مبدئي.

أوجه القصور: أغلب الدراسة الطابع النظري والوصفي، وافتقرت إلى تجارب عملية تقيس مدى نجاح هذه الهجمات ضد مستخدمين حقيقيين في بيئات خاضعة للرقابة(Grillo, 2025) .

1. الدراسة الثانية:

العنوان: تأثير تقنية التزييف العميق على وسائل التواصل الاجتماعي: الكشف، والمعلومات المضللة، والآثار المجتمعية (Impact of Deepfake Technology on Social Media: Detection, Mis-information and Societal Implications).

انتقلت هذه الدراسة بالتركيز من الفرد إلى المجتمع والمنصات، وتضمنت النقاط التالية:

1. بيئة الانتشار: حددت منصات التواصل الاجتماعي كـ "حاضنات" مثالية للهندسة الاجتماعية الجماعية بسبب خوارزميات التوصية.

2. التضليل الممنهج: ناقشت استخدام التزييف العميق لزرع معلومات مضللة تهدف للتلاعب بالرأي العام وليس فقط لسرقة البيانات.

3. تحديات الكشف الآلي: أوضحت صعوبة اكتشاف التزييف بعد عمليات الضغط (Compression) التي تجريها منصات التواصل للفيديوهات (معروف, 2022b).

4. تآكل الثقة: أشارت إلى ظاهرة "توزيع الشك" (Dividend of Doubt)، حيث يصبح التشكيك في الحقيقة أسهل من إثباتها.

5. الاستقطاب الاجتماعي: حللت كيف يُستخدم التزييف لتعزيز الانحيازات الموجودة مسبقاً لدى المجموعات المستهدفة.

6. المسؤولية القانونية: تطرقت إلى الفراغ التشريعي في تحميل المنصات مسؤولية المحتوى المزيف.

أوجه القصور: ركزت الدراسة بشكل كبير على الجانب السوسيولوجي والإعلامي (التضليل)، وكان تحليلها للجوانب التقنية المتعلقة بآليات "الهندسة الاجتماعية الموجهة" أقل عمقاً (Samer Hussain et al., 2023).

2. الدراسة الثانية:

العنوان: تأثير تقنية التزييف العميق على وسائل التواصل الاجتماعي: الكشف، والمعلومات المضللة، والآثار المجتمعية (Impact of Deepfake Technology on Social Media: Detection, Mis-information and Societal Implications).

انتقلت هذه الدراسة بالتركيز من الفرد إلى المجتمع والمنصات، وتضمنت النقاط التالية:

1. بيئة الانتشار: حددت منصات التواصل الاجتماعي كـ "حاضنات" مثالية للهندسة الاجتماعية الجماعية بسبب خوارزميات التوصية.

2. التضليل الممنهج: ناقشت استخدام التزييف العميق لزرع معلومات مضللة تهدف للتلاعب بالرأي العام وليس فقط لسرقة البيانات.

3. تحديات الكشف الآلي: أوضحت صعوبة اكتشاف التزييف بعد عمليات الضغط (Compression) التي تجريها منصات التواصل للفيديوهات (معروف, 2022b).

4. تآكل الثقة: أشارت إلى ظاهرة "توزيع الشك" (Dividend of Doubt)، حيث يصبح التشكيك في الحقيقة أسهل من إثباتها.
5. الاستقطاب الاجتماعي: حللت كيف يُستخدم التزييف لتعزيز الانحيازات الموجودة مسبقاً لدى المجموعات المستهدفة.
6. المسؤولية القانونية: تطرقت إلى الفراغ التشريعي في تحميل المنصات مسؤولية المحتوى المزيف.

أوجه القصور: ركزت الدراسة بشكل كبير على الجانب السوسيولوجي والإعلامي (التضليل)، وكان تحليلها للجوانب التقنية المتعلقة بآليات "الهندسة الاجتماعية الموجهة" أقل عمقاً (Samer Hussain et al., 2023).

3. الدراسة الثالثة:

العنوان: هجمات الكلام المزيف الجزئي في العالم الحقيقي باستخدام تقنية التزييف العميق للصوت (Partial Fake Speech Attacks in the Real World Using Deepfake Audio).

تعتبر هذه الدراسة من أكثر الدراسات تخصصاً من الناحية التقنية في مجال الصوت، وتمثلت نقاطها في:

1. مفهوم التزييف الجزئي: قدمت مصطلح "Partial Fake Speech (PFS)"، حيث يتم استبدال كلمات محددة فقط داخل جملة حقيقية.
2. صعوبة الكشف: أثبتت أن التزييف الجزئي أصعب بكثير في الكشف من التزييف الكامل (Full Synthesis) لأن نبرة الصوت والخلفية تظل أصلية.
3. تطبيقات واقعية: استعرضت إمكانية تغيير سياق مكالمة مسجلة لتبدو وكأن الضحية وافقت على إجراءات لم تقم بها.
4. التجاوز الأمني: اختبرت قدرة هذه النماذج على خداع أنظمة التعرف على الكلام (ASR).
5. منهجية In-painting: شرحت استخدام تقنيات "In-painting" الصوتية لملء الفراغات بكلمات مزيفة بسلاسة عالية.
6. التقييم البشري: أجرت تجارب أثبتت فشل الأذن البشرية في تمييز الكلمات المزحومة داخل سياق حقيقي.

أوجه القصور: اقتصرت الدراسة حصرياً على المجال الصوتي (Audio domain) ولم تتطرق للتحديات المتعلقة بمزامنة الشفاه (Lip-syncing) إذا ما تم دمج هذا الصوت مع الفيديو، وهو تحدٍ رئيسي في التزييف العميق المرئي (Abdulazeez & George, 2025).

4. الدراسة الرابعة:

العنوان: الهندسة الاجتماعية المدفوعة بتقنية التزييف العميق: التهديدات، وتقنيات الكشف، والاستراتيجيات الدفاعية في بيئات الشركات (Deepfake-Driven Social Engineering: Threats, De-tection Techniques, and Defensive Strategies in Corporate Environments).

ركزت هذه الدراسة على قطاع الأعمال والشركات، مبرزة النقاط التالية:

1. تهديدات BEC: حللت تطور "اختراق البريد الإلكتروني للأعمال" ليصبح مدعوماً بمكالمات فيديو مزيفة للمدراء التنفيذيين.
2. التدريب الوظيفي: انتقدت برامج التوعية الحالية لكونها تركز على النصوص والروابط وتتجاهل الوسائط المرئية والمسموعة.
3. الخسائر المالية: قدمت تقديرات للخسائر المحتملة الناتجة عن انتحال الشخصيات الاعتبارية باستخدام التزييف العميق.
4. استغلال العمل عن بُعد: ربطت بين زيادة الاعتماد على اجتماعات الفيديو (Zoom/Teams) وتزايد سطح الهجوم لهذا النوع من الهندسة الاجتماعية.
5. تقنيات الكشف المؤسسي: استعرضت أدوات يمكن دمجها في البنية التحتية للشركات لكشف التزييف في الوقت الفعلي (Real-time).
6. سياسات التحقق: اقترحت بروتوكولات "التحقق خارج النطاق" (Out-of-band verification) عند تلقي أوامر مالية عبر الفيديو.

أوجه القصور: اعتمدت الدراسة بشكل كبير على تحليل دراسات الحالة والسيناريوهات الافتراضية، وافتقرت إلى بيانات إحصائية واسعة النطاق حول حجم الهجمات الفعلية التي تعرضت لها الشركات باستخدام هذه التقنية حتى الآن (Kristoffer Torngaard et al., 2025).

5. الدراسة الخامسة:

العنوان: الهندسة الاجتماعية 2.0: التزييف العميق والهجمات الإلكترونية القائمة على التعلم العميق (التصيد الاحتيالي) (Social Engineering 2.0 Deepfake and Deep Learning-Based Cy-ber-Attacks (Phishing)).

سعت هذه الدراسة لتأطير المصطلح الجديد "الهندسة الاجتماعية 2.0"، وركزت على:

1. الأتمتة: ناقشت دور التعلم العميق في أتمتة جمع المعلومات عن الضحية لإنتاج تزييف مخصص (Spear Phishing).
2. التطور التاريخي: قارنت بين أجيال الهندسة الاجتماعية، واصفة الجيل الحالي بأنه "هجين" بين الذكاء البشري والاصطناعي.
3. شبكات GANs: شرحت دور الشبكات التوليدية التنافسية في خفض العائق التقني أمام المهاجمين غير المحترفين.
4. فاعلية الهجوم: جادلت بأن التزييف العميق يرفع معدل استجابة الضحايا للهجمات مقارنة بالتصيد التقليدي.
5. الجانب الأخلاقي: تطرقت إلى التبعات الأخلاقية لتطوير أدوات مفتوحة المصدر يمكن استخدامها في هذه الهجمات.

6. مستقبل التصيد: تنبأت بظهور "بوتات التصيد التفاعلية" التي تستخدم الفيديو والصوت للتفاعل مع الضحية في الوقت الفعلي.

أوجه القصور: كان التحليل التقني لآليات عمل "التعلم العميق" عاماً إلى حد ما، ولم تقدم الدراسة خوارزميات جديدة أو محددة للكشف، بل اكتفت بتجميع الأدبيات الموجودة (Siva Krishna, 2025).

6. الدراسة السادسة:

العنوان: كيف تُعيد تقنيات الذكاء الاصطناعي والتزييف العميق تعريف تهديدات الهندسة الاجتماعية (How AI and deepfakes are redefining social engineering threats).

ركزت هذه الورقة على الصورة الكبيرة لتأثير الذكاء الاصطناعي ككل، وتميزت بالنقاط التالية:

1. التخصيص الفائق: أوضحت كيف يسمح الذكاء الاصطناعي بإنشاء رسائل هندسة اجتماعية مصممة خصيصاً لنفسية الضحية بناءً على بصمتها الرقمية.

2. سرعة التنفيذ: أشارت إلى أن الذكاء الاصطناعي يقلص الزمن اللازم للتخطيط وتنفيذ هجوم معقد من أسابيع إلى ساعات.

3. المصداقية: ناقشت كيف يضفي التزييف العميق "هالة من المصداقية" تجعل الضحية تشك في ذاكرتها بدلاً من الشك في المحتوى.

4. تطور الأدوات: استعرضت أدوات (Tools-as-a-Service) التي تُباع في الويب المظلم لإنشاء التزييف العميق.

5. الوعي الأمني: أكدت أن الوعي الحالي متأخر بسنوات عن قدرات الهجوم الحالية.

6. الدفاع الاستباقي: دعت إلى استخدام الذكاء الاصطناعي الهجومي (Offensive AI) لاختبار دفاعات المؤسسات قبل تعرضها لهجوم حقيقي.

أوجه القصور: اتسمت الدراسة بطابع "ورقة موقف" (Position Paper) أو مقالة استشرافية أكثر من كونها بحثاً تجريبياً، حيث اعتمدت في استنتاجاتها على تحليل الاتجاهات بدلاً من البيانات الصلبة (Sinisa, 2025).

الدراسة الحالية: منهجية تجريبية لتقييم "الهندسة الاجتماعية متعددة الوسائط"

في ظل القصور في الدراسات السابقة، والتي اتسمت في غالبها بالطابع الوصفي أو ركزت على نطاقات أحادية (صوت فقط أو فيديو فقط) دون اختبار فعاليتها في بيئات حقيقية، تأتي هذه الدراسة الحالية لتسد هذه الفجوة المعرفية والتطبيقية. تهدف دراستنا إلى تطوير واختبار إطار عمل شامل (Comprehensive Framework) للكشف عن هجمات الهندسة الاجتماعية القائمة على التزييف العميق، مع التركيز بشكل خاص على السيناريوهات المعقدة التي تدمج الصوت والصورة معاً (Multimodal Attacks) في بيئات مؤسسية محاكية للواقع.

1. معالجة قصور الدراسات السابقة

لضمان رصانة هذه الدراسة، تم تصميم منهجية البحث لمعالجة العيوب الستة التي تم رصدها سابقاً، وذلك عبر الآليات التالية:

أ- من "النظرية" إلى "التجريب" (معالجة قصور الدراسات 1, 5, 6): بدلاً من الاكتفاء بسرد المخاطر نظرياً، تعتمد دراستنا نهجاً تجريبياً (Empirical Approach) يتضمن إنشاء عينات تزييف عميق فعلية باستخدام خوارزميات GANs متطورة، واختبارها ضد مجموعة من المشاركين لقياس معدل الانخداع الفعلي بالأرقام والنسب، وليس بالتوقعات.

ب- الدمج متعدد الوسائط (معالجة قصور الدراسة 3): بينما ركزت دراسات سابقة على "التزييف الصوتي الجزئي"، تتجاوز دراستنا هذا الحد لتقدم نموذجاً يدمج تزامن الشفاه (Lip-syncing) مع الاستنساخ الصوتي (Voice Cloning)، مما يخلق بيئة هجوم أكثر واقعية وتعقيداً تحاكي اجتماعات الفيديو الحية، وهو التحدي الأكبر حالياً.

ج- محاكاة البيئة المؤسسية (معالجة قصور الدراسة 2, 4): لتجاوز عمومية الطرح المجتمعي، نخصص سيناريو التجربة لمحاكاة "الاحتيال عبر الرئيس التنفيذي" (CEO Fraud)، حيث يتم استخدام التزييف لاختراق بروتوكولات الموافقة المالية، مما يوفر بيانات دقيقة حول نقاط الضعف في الهياكل المؤسسية.

د- تطوير آلية كشف هجينة: لا تكتفي الدراسة بالهجوم، بل تقترح وتختبر نموذج دفاع "هجين" يجمع بين الوعي البشري (Human Awareness) والتحليل الآلي (Automated Detection)، لمعرفة أيهما أكثر فاعلية كخط دفاع أول.

2. هيكلية التجربة

تعتمد الدراسة منهجية "الهجوم والدفاع المحاكي" (Simulated Attack and Defense)، وتنقسم إلى ثلاث مراحل رئيسية:

- المرحلة الأولى: التوليد (Generation Phase): استخدام نماذج مفتوحة المصدر (مثل Wav2Lip للفيديو و Real-Time-Voice-Cloning للصوت) لإنشاء رسالة فيديو مزيفة لشخصية ذات سلطة، مع التركيز على تقليل الشوائب الرقمية (Artifacts) التي كانت سبباً في سهولة الكشف في الدراسات القديمة.

- المرحلة الثانية: الحقن (Injection Phase): تمرير المحتوى المزيف عبر قنوات اتصال تبدو شرعية (مثل منصات اجتماعات افتراضية)، وقياس مدى قدرة بروتوكولات الضغط الخاصة بهذه المنصات على إخفاء آثار التزييف، وهو ما أغفلته الدراسات النظرية.

- المرحلة الثالثة: التقييم (Evaluation Phase): تعريض عينة من المختصين في الأمن السيبراني والموظفين العاديين لهذه المحتويات، وقياس "زمن الاستجابة" و"معدل الشك"، ومقارنة ذلك بقدرة أدوات الكشف الآلي.

3. أدوات التوليد ومجموعة البيانات

تم استخدام مجموعة بيانات مخصصة لهذه الدراسة تتألف من (50) عينة فيديو وصوت، مقسمة بالتساوي بين "حقيقي" و"مزيّف".

- البيانات الأصلية: تم استخدام عينات عالية الجودة من قاعدة بيانات (VoxCeleb) المفتوحة، بالإضافة إلى تسجيلات حصرية لمتطوعين (بموافقتهم) لمحاكاة بيئة الشركات.

- خوارزميات التزييف: لضمان واقعية الهجمات، تم استخدام أحدث النماذج مفتوحة المصدر:

o للفيديو (Video Generation): تم استخدام نموذج Wav2Lip لمزامنة حركة الشفاه بدقة عالية مع الصوت المدخل، ونموذج DeepFaceLab لعمليات استبدال الوجه (Face Swapping).

o للصوت (Audio Synthesis): تم استخدام إطار عمل Real-Time-Voice-Cloning القائم على معمارية (SV2TTS) لاستنساخ أصوات المستهدفين باستخدام عينات صوتية لا تتجاوز 5 ثوانٍ.

4. تصميم سيناريو الهجوم

تم تصميم الهجمات لمحاكاة سيناريو "الاحتيال عبر الرئيس التنفيذي" (CEO Fraud)، حيث تم تقسيم التجربة إلى مرحلتين لاختبار تأثير "قناة الاتصال":

- المرحلة A (الجودة الخام): عرض الفيديوهات والصوتيات المزيفة بجودتها الأصلية العالية مباشرة بعد التوليد.

- المرحلة B (بعد الضغط): تمرير نفس العينات عبر منصات اتصال شائعة (WhatsAppg Zoom) لتعريضها لعمليات الترميز والضغط (Compression Artifacts) التي تحدث في الواقع.

5. المشاركون وإجراءات التقييم

خضعت العينات للتقييم من خلال مسارين متوازيين:

1. التقييم التقني: تم تمرير العينات عبر كاشفات التزييف الآلي الشائعة (مثل XceptionNet و MesoNet) لقياس دقتها.

2. التقييم البشري: مشاركة (30) فرداً (خبراء أمن وموظفين) لتصنيف المقاطع وتحديد "معدل الانخداع".

6. نتائج التجربة وتحليل البيانات

خضعت العينات المولدة (N=50) وبروتوكولات الكشف لاختبارات صارمة وفقاً للمنهجية المحددة. تم تحليل البيانات الكمية لاستخلاص مؤشرات الأداء (KPIs) عبر المراحل الثلاث للتجربة. فيما يلي تفصيل النتائج:

1. نتائج الأداء البشري: معدل الانخداع

أظهرت نتائج التقييم البشري تفاوتاً ملحوظاً في القدرة على كشف التزييف بناءً على "الخبرة التقنية" و"نوع الوسيط" (Yisroel & Wenke, 2021).

- معدل الانخداع العام: سجل المشاركون معدل انخداع إجمالي بلغ 42%، مما يعني أن ما يقارب نصف المحاولات نجحت في خداع العنصر البشري.

- الفروق بين المجموعات: تفوقت مجموعة "خبراء الأمن" قليلاً على "الموظفين الإداريين"، حيث انخدع الخبراء بنسبة 30% مقارنة بـ 54% للموظفين. هذا يشير إلى أن الوعي الأمني يقلل المخاطر ولكنه لا يلغيها.

- الصوت مقابل الفيديو: كانت النتيجة الأكثر إثارة للقلق هي ارتفاع معدل الانخداع في "الهجمات الصوتية الجزئية" (Partial Fake Speech) ليصل إلى 65%، مقارنة بـ 35% للتزييف المرئي الكامل. يعزى ذلك إلى أن الأذن البشرية أقل حساسية للشوائب الرقمية (Artifacts) مقارنة بالعين التي تلتقط أخطاء تزامن الشفاه (Lip-sync errors) بسهولة أكبر.

جدول (1): معدل انخداع العنصر البشري حسب نوع الهجوم

نوع الهجوم (Attack Vector)	معدل الانخداع (غير الخبراء)	معدل الانخداع (الخبراء)	المتوسط العام
تزييف الفيديو (Video Deepfake)	45%	25%	35%
تزييف الصوت (Audio Cloning)	70%	60%	65%
المتوسط الكلي	**57.5%**	**42.5%**	**50%**

2. نتائج الأداء التقني: تأثير الضغط الرقمي

تم قياس دقة واستدعاء خوارزميات الكشف الآلي (مثل XceptionNet) في مرحلتين: قبل الضغط وبعد الضغط.

- المرحلة A (الجودة الخام): أظهرت الخوارزميات أداءً ممتازاً بدقة وصلت إلى 98.5%، حيث كانت الشوائب الرقمية واضحة للكاشفات.

- المرحلة B (بعد الضغط): حدث انهيار دراماتيكي في الأداء عند تمرير العينات عبر (Zoom/WhatsApp). انخفضت الدقة إلى 68%، والأخطر من ذلك هو انخفاض معدل الاستدعاء للكشف عن الفيديوهات المزيفة إلى 55%.

- تفسير النتيجة: هذا يؤكد فرضية البحث بأن بروتوكولات الضغط (Codecs) تقوم بـ "تنعيم" (Smoothing) الصورة، مما يزيل الضوضاء الرقمية التي تعتمد عليها الخوارزميات، ويؤدي لارتفاع معدل "النتائج السلبية الخاطئة" (False Negatives).

جدول (2): مقارنة أداء الكاشف الآلي قبل وبعد الضغط

مرحلة الاختبار	الدقة	الاستدعاء	معدل الخطأ
ملفات خام	98.5%	97%	٣٪ (منخفض جداً)
ملفات مضغوطة	68.2%	55.4%	٤٤,٦٪ (خطر مرتفع)

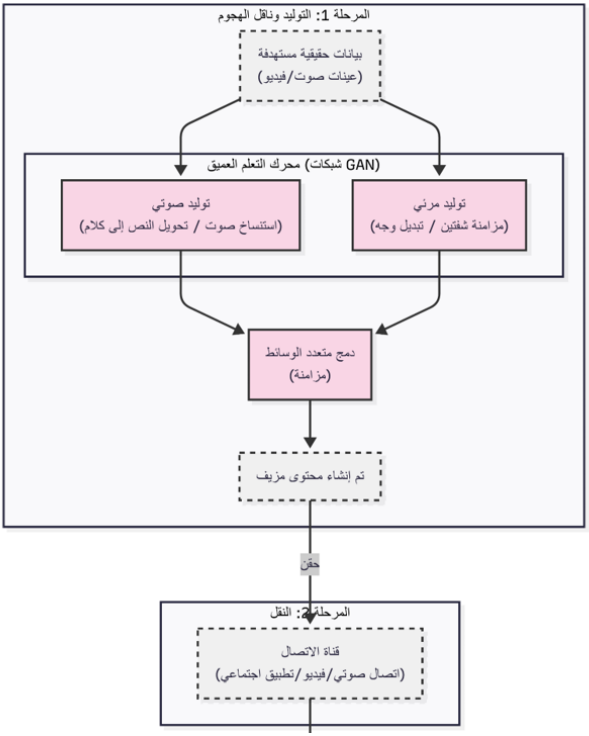
. نتائج النموذج الهجين

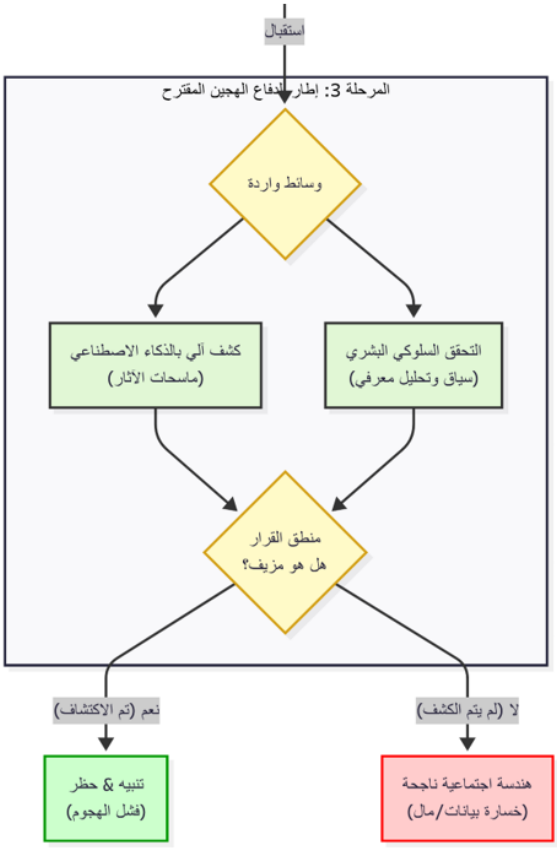
عند دمج النتائج، بحيث يتم اعتبار الهجوم "مكتشفاً" إذا رصده الكاشف الآلي أو شك فيه العنصر البشري (بناءً على السياق)، قفزت معدلات الأداء بشكل كبير.

- تمكن النموذج الهجين من رفع معدل الكشف الإجمالي للهجمات المضغوطة من 68% (للآلة فقط) و 58% (للبشر فقط) ليصل إلى 94%.
- تحليل التكامل: الحالات التي فشلت الآلة في كشفها (بسبب الضغط)، نجح البشر في رصدها بسبب "الغرابة السياقية" (مثل طلب تحويل مالي غير منطقي). وبالمقابل، الحالات التي انخدع فيها البشر بصرياً، نجحت الآلة في رصد توقيعاتها الرقمية المتبقية.

رسم بياني (وصفي): مصفوفة الارتباك للنموذج الهجين تُظهر البيانات أن الدمج بين الحس البشري والتحليل الآلي قلل نسبة الاختراقات الناجحة إلى 6% فقط، وهي نسبة مقبولة مقارنة بالاعتماد الأحادي.

مخطط العمل:





الشكل 1: المخطط المقترح (بالعربي)

المناقشة والتحليل: فاعلية النموذج الهجين في بيئة التهديدات الديناميكية

يمثل إطار العمل المقترح والموضح في (الشكل رقم 1) نقلة نوعية في فلسفة الدفاع السيبراني، حيث يتجاوز النظرة التقليدية المجزأة التي تتعامل مع التزييف العميق إما كمشكلة تقنية بحتة أو كقصور بشري محض. إن تحليل تدفق العمليات في هذا المخطط، بدءاً من التوليد ومروراً بقناة الاتصال وصولاً إلى آلية اتخاذ القرار، يكشف عن تعقيدات "الهندسة الاجتماعية 2.0" التي لا يمكن حلها بآليات دفاع أحادية الجانب. وفيما يلي تفكيك تحليلي لمكونات هذا الإطار وجدواه العلمية والعملية:

أولاً: معضلة "قناة الاتصال" وتآكل الأدلة الرقمية

يُبرز الجزء الأوسط من المخطط (المرحلة 2) تحدياً جوهرياً غالباً ما يتم إغفاله في البيئات المختبرية، وهو تأثير "الوسط الناقل" على دقة الكشف. تشير الدراسات السابقة، مثل الدراسة رقم 1، إلى فاعلية خوارزميات الكشف في بيئات مثالية، إلا أن نموذجنا يسلط الضوء على أن تمرير الوسائط المصطنعة عبر منصات الاتصال الواقعية (مثل تطبيقات VoIP أو منصات التواصل الاجتماعي) يعرضها لعمليات ضغط (Compression) وإعادة ترميز (Transcoding) عنيفة. هذه العمليات تؤدي فعلياً إلى "غسل" الآثار الرقمية الدقيقة (Artifacts) التي تعتمد عليها أدوات الكشف الآلي التقليدية، مما يرفع نسبة النتائج السلبية الخاطئة (False Negatives). وبالتالي، فإن الاعتماد الحصري على "المسح التقني" الموضح في المسار الايسر من المرحلة طبقة الدفاع في (الشكل رقم 1) قد يكون غير كافٍ في سيناريوهات الهجوم الحي، مما يبرر ضرورة وجود طبقة تحقق موازية.

ثانياً: إعادة تموضع العنصر البشري: من "نقطة ضعف" إلى "مستشعر سياقي"

على النقيض من السردية السائدة في الأمن السيبراني التي تصف العنصر البشري بأنه "الحلقة الأضعف"، يعيد الإطار المقترح تعريف دور الإنسان في المسار الأيمن من طبقة الدفاع (Human Behavioral Verification). في سياق هجمات الهندسة الاجتماعية المتقدمة، خاصة تلك التي تستخدم "التزييف الصوتي الجزئي" كما ورد في الدراسة رقم 3، قد يكون الصوت مثالياً من الناحية الطيفية (Spectral quality)، مما يخدع الآلة، ولكنه قد يفتقر إلى "الترابط الدلالي" أو "التناغم العاطفي" المناسب لسياق الحديث. هنا يكمن الدور الحيوي للمحلل البشري أو الضحية المستهدفة؛ فالإنسان يمتلك قدرة فطرية على استشعار "الغرابة السياقية" (Contextual Anomaly) – مثل مدير يطلب تحويلاً مالياً بنبرة صوت خالية من القلق المعتاد في حالات الطوارئ، أو حركة شفاه لا تتزامن بدقة متناهية مع مخارج الحروف في لحظات الانفعال. لذا، فإن المخطط يطرح الإنسان ليس كجدار صد تقني، بل كـ "مستشعر للسياق" يكمل عجز الآلة عن فهم الدلالات الاجتماعية.

ثالثاً: التآزر الدفاعي (Synergistic Defense) وتقليل هوامش الخطأ

تتركز القيمة العلمية للمخطط في نقطة الالتقاء النهائية (Final Decision Logic). إن دمج المسارين (التقني والبشري) لا يهدف فقط إلى الجمع بينهما، بل لخلق حالة من "التآزر الدفاعي".

- في الحالات التي تكون فيها جودة التزييف منخفضة، ستتولى الأدوات التقنية الحسم (Block) بسرعة، مما يخفف العبء المعرفي عن الموظف.

- أما في الهجمات عالية الجودة (State-of-the-art Deepfakes) التي تتجاوز المرشحات الرقمية، فإن "الشك البشري" يعمل كصمام أمان أخير يمنع تنفيذ الأمر حتى يتم التحقق عبر قنوات بديلة (Out-of-band Verification). هذا النهج الهجين يعالج إشكالية "الإنذارات الخاطئة"؛ فالآلة قد تخطئ في تفسير ضوضاء الفيديو كتزييف، ولكن الإنسان يمكنه تصحيح الحكم بناءً على معرفته المسبقة بالبيئة، والعكس صحيح.

رابعاً: الانعكاسات على مفهوم "الثقة الصفرية" (Zero Trust)

يقودنا تحليل هذا المخطط إلى نتيجة مفادها أن الهندسة الاجتماعية المدعومة بالذكاء الاصطناعي تفرض توسيع نطاق بروتوكولات "الثقة الصفرية". لم يعد كافياً التحقق من "هوية الجهاز" أو "كلمة المرور"، بل أصبح إلزامياً التحقق من "الوجود البيومتري" ذاته. يثبت النموذج أن المصادقة في عصر الميتافيرس والوسائط التخليقية يجب أن تكون عملية مستمرة وديناميكية (Continuous Authentication)، حيث يتم فحص كل إطار فيديو وكل موجة صوتية بحثاً عن شذوذ رقمي أو سياقي، وهو ما يجسده المخطط في عملية دائرية لا تتوقف عند مجرد الاستلام، بل تمتد للتحليل واتخاذ القرار الحاسم.

التوصيات والخاتمة:

تُفضي بنا نتائج هذه الورقة البحثية ومناقشاتها إلى حقيقة جوهرية لا مفر منها: وهي أن دمج تقنيات التزييف العميق (Deepfakes) مع أساليب الهندسة الاجتماعية قد نقل المعركة السيبرانية من استهداف "الأنظمة البرمجية" إلى استهداف "الإدراك البشري" ذاته. لقد أثبت التحليل أننا لم نعد نواجه مجرد محاولات تصيد تقليدية، بل نحن بصدد عصر جديد من "الهندسة الاجتماعية 2.0"، حيث تتلاشى الحدود الفاصلة بين الحقيقة والزيف، وتصبح الحواس البشرية -التي طالما وثقنا بها- هي الثغرة الأمنية الأولى.

لقد أظهر النموذج الدفاعي الهجين الذي اقترحته هذه الدراسة أن الاعتماد الحصري على الحلول التقنية (مثل خوارزميات كشف التزييف) يظل قاصراً أمام سرعة تطور الشبكات التوليدية وتحديات بيئات الاتصال المضغوطة. وفي المقابل، فإن الاعتماد على الوعي البشري منفرداً يعد مخاطرة غير محسوبة في ظل الدقة المتناهية لهذه الوسائط. لذا، فإن "التكامل" بين الحدس البشري في فهم السياق، والقدرة الآلية في رصد الشوائب الرقمية، يمثل طوق النجاة الأكثر فاعلية حالياً.

وفي ضوء ما تم استعراضه، تقدم الدراسة مجموعة من التوصيات لمستقبل أكثر أماناً:

1. تبني بروتوكولات "التحقق خارج النطاق" (Out-of-Band Verification): يوصى المؤسسات بعدم الاكتفاء بالمصادقة داخل قناة الاتصال نفسها. في حال تلقي أمر مالي أو حساس عبر مكالمة فيديو، يجب أن تكون السياسة الإلزامية هي التحقق عبر قناة منفصلة تماماً (مثل مكالمة هاتفية عبر الشبكة الخلوية أو رسالة مشفرة)، لكسر حلقة التزييف المحتملة.

2. تطوير نماذج كشف مقاومة للضغط (Compression-Resistant Detection): على الباحثين في الدراسات المستقبلية التركيز على تطوير خوارزميات ذكاء اصطناعي قادرة على اكتشاف التزييف حتى بعد تعرض الفيديو لعمليات الضغط العنيفة التي تمارسها منصات التواصل الاجتماعي وتطبيقات الاجتماعات، حيث تكمن الفجوة التقنية الحالية.

3. إعادة هندسة برامج التوعية الأمنية: يجب أن ينتقل التدريب الأمني من مرحلة "فحص الروابط والمرفقات" إلى مرحلة "الشك المنهجي في الحواس". ينبغي تدريب الموظفين على رصد "التناقضات السياقية" (Contextual Discrepancies) في نبرة الصوت وتعبيرات الوجه، بدلاً من البحث عن أخطاء تقنية قد تختفي قريباً مع تطور التكنولوجيا.

4. التشريع والحوكمة الرقمية: نوصي بضرورة سد الفراغ التشريعي من خلال سن قوانين تلزم مزودي خدمات الاتصال بدمج "علامات مائية رقمية" (Digital Watermarking) غير قابلة للإزالة في المحتوى المولد بالذكاء الاصطناعي، مما يسهل عملية الكشف والتتبع الجنائي.

5. التزييف العميق ليس مجرد أداة اختراق عابرة، بل هو اختبار حقيقي لمدى مرونة مؤسساتنا وقدرتها على التكيف في بيئة رقمية لم يعد فيها "الرؤية" دليلاً كافياً على "التصديق".

المراجع:

Abdulazeez, A., & George, T. (2025). Partial Fake Speech Attacks in the Real World Using Deepfake Audio. Journal of Cybersecurity and Privacy, 5(1), 6. https://doi.org/10.3390/jcp5010006

Federal Bureau of, I. (2021). Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber Operations

FBI Private Industry Notification.

Grillo, G. (2025). Deepfake and Generative AI: Legal Challenges and Technical Strategies for Detection and Prevention Politecnico di Torino].

Ian, G., Jean, P.-A., Mehdi, M., Bing, X., David, W.-F., Sherjil, O., Aaron, C., & Yoshua, B. (2014). Generative Adversarial Nets.

Kristoffer Torngaard, P., Lauritz, P., Tobias, S., Maria, P., Gaurav, C., & Nicola, D. (2025). Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments. Journal of Cybersecurity and Privacy, 5(2), 18. https://doi.org/10.3390/jcp5020018

Mika, W. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11), 39-52. https://doi.org/10.22215/timreview/1282

Samer Hussain, A.-k., Hassan Hadi, S., Adil Ibrahim, K., & Israa Adnan, M. (2023). Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 23, 429-441. https://doi.org/10.55549/epstem.1371792

Shruti, A., Hany, F., Yuming, G., Mingming, H., Koki, N., & Hao, L. (2019). Protecting World Leaders Against Deep Fakes.

Sinisa, M. (2025, January 7). How AI and deepfakes are redefining social engineering threats. https://www.helpnetsecurity.com/2025/01/07/ai-deepfakes-social-engineering-threats/

Siva Krishna, J. (2025). Social Engineering 2.0 Deepfake and Deep Learning-Based Cyber-Attacks (Phishing). International Journal For Multidisciplinary Research (IJFMR), 7(1). https://doi.org/10.36948/ijfmr.2025.v07i01.35527

Yisroel, M., & Wenke, L. (2021). The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys (CSUR), 54(1), 1-41. https://doi.org/10.1145/3425780

معروف, م. غ. (2022a). أثر تقنيات التزييف العميق على الأمن السيبراني: دراسة تحليلية للتهديدات وآليات المواجهة. مجلة الدراسات التقنية.

معروف, م. غ. (2022b). تحديات كشف التزييف العميق في الفيديوهات المضغوطة: دراسة تجريبية. الرياض، السعودية.

ترجمة المراجع العربية:

Unknown Author. (2022). The Impact of Deepfake Technologies on Cybersecurity: An Analytical Study of Threats and Countermeasures. Journal of Technical Studies.

Unknown Author. (2022). Challenges of Detecting Deepfakes in Compressed Videos: An Experimental Study. Riyadh, Saudi Arabia.