مناقشة رسالة ماجستير فى كلية الحاسبات والذكاء الأصطناعى جامعة حلوان

ناقش أ.د. علاء الدين محمد الغزالى رئيس الجمعية يوم الثلاثاء ٢٧ أغسطس ٢٠٢٤ رسالة ماجستير فى كلية الحاسبات والذكاء الأصطناعى جامعة حلوان للباحثة آية الرحمن كمال الدين رمضان المعيدة بقسم نظم المعلومات بالكلية عنوان الرساله « حول التحليل الآلي لنقاط ضعف البرمجيات بإستخدام أساليب تعلم الآلة»

On the Automated Analysis of Software Vulnerabilities using Machine Learning Approaches

الرسالة تحت أشراف:

أ. د/ أحمد بهاء (مشرفاً) أستاذ بقسم نظم المعلومات بكلية الحاسبات والذكاء الاصطناعي بجامعة حلوان ووكيل كلية الحاسبات والذكاء الاصطناعي بجامعة بنى سويف.

أ. م. د/ حنان فهمي (مشرفاً) أستاذ مساعد بقسم نظم المعلومات بكلية الحاسبات والذكاء الاصطناعي بجامعة حلوان.

د/عمرو غنيم (مشرفاً) مدرس بقسم علوم الحاسب بكلية Amr S. Ghoneim الحاسبات والذكاء الاصطناعي بجامعة حلوان.

وقد رأس جلسة الحكم

أ. د/ علاء الدين محمد الغزالى أستاذ متفرغ بقسم الحاسب الألى ونظم المعلومات بأكاديمية السادات للعلوم الإدارية. وعضوية

أ. د/ دعاء سعد الزنفلى أستاذ ورئيس قسم نظم المعلومات بكلية الحاسبات والذكاء الاصطناعي بجامعة حلوان.

أهداف الرسالة: تسعى هذه الرسالة إلى تطوير نموذج مبتكر يعتمد على تقنيات تعلم الآلة والتعلم العميق لتحليل كود البرمجيات بدقة وكفاءة، بهدف الكشف عن الثغرات الأمنية بفاعلية. تركز الرسالة على تحقيق ما يلي:

- تقديم تحليل دقيق ومفصل للكشف عن نقاط الضعف في البرمجيات.
- الكشف عن مجموعة متنوعة من نقاط الضعف البرمجية.
- تقليل عدد الثغرات الأمنية المكتشفة في البرمجيات.
- تعزيز أداء الأساليب الحالية المستخدمة في تحليل نقاط الضعف.
- خفض معدلات الإيجابيات والسلبيات الزائفة المرتفعة.
- تطبيق تقنيات متقدمة لاستخراج وتعلم خصائص كود البرمجيات بناءً على بنيته ودلالاته.
- تحسين جودة البرمجيات بشكل عام.

أدبيات الموضوع:

تقدم الرسالة مراجعة شاملة لكل الأدبيات المتعلقة بطرق التحليل المختلفة للبرمجيات لاكتشاف الثغرات الأمنية ، عملية اكتشاف الثغرات الأمنية في البرمجيات ، تقنيات تعلم الآلة والتعلم العميق لتحليل كود البرمجيات للكشف عن الثغرات الأمنية

، تضمين الكلمات السياقية وغير السياقية، تقنيات ال Transformers. تتضمن المراجعة تحديد الفجوات في الأبحاث السابقة في مجال اكتشاف الثغرات البرمجية ، استخدام تقنيات المحولات Transformers فى إنشاء تضمين الكلمات السياقية لكود البرمجيات، تطبيق النماذج الهجينة من تقنيات تعلم الآلة والتعلم العميق فى مجال البرمجيات.

منهجية البحث والإطار النظري:

تتبع الدراسة منهجية تشمل بناء نموذج مبتكر هجين لتطبيق احدث تقنيات تعلم الآلة والتعلم العميق لتحليل كود البرمجيات بدقة وكفاءة وبشكل تلقائى بالاضافة إلى دمج تقنيات المحولات Transformers لاكتشاف الثغرات الأمنية في البرمجيات.

يتضمن البحث استخدام المحلل اللغوي Parser لإنشاء شجرة بناء الجملة المجردةAbstract Syntax Tree . حيث يقوم المحلل بإنشاء AST لكل دالة كود موجودة في مجموعة البيانات. يتم تسلسل ASTs هذه، مما يعني تحويل بنية الشجرة إلى تنسيق خطي. ثم يتضمن الترميز Tokenizationتحويل الرموز النصية إلى رموز عددية صحيحة باستخدام نموذج BERT . تتضمن هذه العملية تحويل الرموز النصية المحددة من AST متسلسلة إلى تسلسل من القيم الصحيحة لكل دالة. هذه الخطوة ضرورية لإنشاء ناقل التضمين لكل AST، والذي يعمل كمدخل لنموذج التعلم العميق الهجين. ومن ثم يستقبل نموذج DistilBERT تسلسل التمثيلات الرقمية للرموز المحددة كمدخلات وينتج متجه التضمين السياقي المقابل.

يتم ضبط محول DistilBert المدرب مسبقًا بدقة لتضمين الكلمات. بعد ذلك، يكتشف نموذج التعلم العميق الهجين وظائف التعليمات البرمجية المعرضة للخطر. تم بناء النموذج الهجين على شبكتين عصبيتين عميقتين DNN. النموذج الأول هو الشبكة العصبية التلافيفية CNN، والتي تستخدم لاستخراج خصائص كود البرمجيات. النموذج الثاني هو Bidirectional-LSTM (BiLSTM)، والذي تم استخدامه للحفاظ على الترتيب التسلسلي للبيانات لأنه يمكنه التعامل مع تسلسلات الرموز الطويلة. وأخيرًا، ترتبط الطبقة الكثيفة، والتي يشار إليها أيضًا بطبقة متصلة بالكامل FC، بطبقة BiLSTM لاستخلاص النتائج النهائية لاكتشاف الثغرات الأمنية.

تحليل ومناقشة النتائج:

قام الباحث بتقييم فعالية النموذج المقترح DB-CBIL من خلال تجارب كثيرة أجريت على مجموعة بيانات مرجعية لضمان البرمجيات SARDو هى Benchmark Dataset المستخدمة فى جميع التقنيات الرائدة في هذا المجال و تُوضح النتائج التجريبية النهائية أن النموذج المقترح DB-CBIL يتفوق على أداء التقنيات الرائدة لنفس مجموعة البيانات بنسبة تتراوح بين ٢,٤١٪-٨,٩٥٪، ٤,٠٪-١٦,٢٨٪، ١,٨٥٪-١٢,٧٤٪، و١٨٪ على التوالي للدقة، واسترجاع البيانات، وتقييم F1، ومعدل السلبيات الزائفة.

حقق نموذج DB-CBIL المقترح أعلى دقة بنسبة ٩٩,٨١٪، متجاوزًا بذلك التقنيات الرائدة الأخرى. كما يتمتع DB-CBIL بدرجة F1 تبلغ ٩٩,٧٥٪. يكتشف DB-CBIL بكفاءة سبع ثغرات أمنية مختلفة، بما في ذلك مشكلات تجاوز سعة المخزن المؤقت [CWE-121، CWE-124، CWE-127، CWE-122]، وحقن أوامر نظام التشغيل [CWE-78]، وأخطاء التحويل [CWE-195]، وسلاسل التنسيق غير المنضبطة [CWE-134]. أخيرًا، حقق النموذج المقترح DB-CBIL أدنى معدل للسلبيات الزائفة، مما يشكل إنجازًا ملحوظًا بالنظر إلى الطبيعة القائمة على التكلفة لأجهزة الكشف عن نقاط ضعف البرمجيات.