Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
DOI: 10.22034/jscit.2025.540111.2155
Volume 14, Number 3, Autumn 2025, pp. 61-76
Received: 08/08/2025, Revised: 09/29/2025, Accepted: 10/18/2025

Improving Zero-Day Malware Detection Based on Ensemble Learning and Gaussian Kernel Fuzzy Rough Sets

Akram Ansari[1], Javad Hamidzadeh[2*]
1- Faculty of Computer Engineering and Information Technology, Sadjad University, Mashhad, Iran.
2*- Faculty of Computer Engineering and Information Technology, Sadjad University, Mashhad, Iran.
[1]akr.ansari226@sadjad.ac.ir, [2*]J_hamidzadeh@sadjad.ac.ir

Corresponding author's address: Javad Hamidzadeh, Faculty of Computer Engineering and Information Technology, Sadjad University, Mashhad, Iran, Post Code: 9188148848.

Abstract- Zero-day malware detection, due to the dynamic and unknown nature of these threats, is one of the major challenges in cybersecurity. The presence of mixed data comprising high-dimensional numerical and categorical features further complicates this problem. Such heterogeneity not only increases the computational burden but also blurs the decision boundaries between malicious and benign classes. Numerical features capture quantitative aspects such as file size or the number of sections, whereas categorical features represent qualitative characteristics such as file type or compiler signature. The simultaneous presence of these features makes feature selection difficult, as conventional methods lack sufficient capability to model nonlinear relationships combined with uncertainty between them, leading to degraded performance in zero-day malware detection. In this study, we propose a feature selection method based on Gaussian Kernel Fuzzy Rough Sets (GKFRS), where the selected features are utilized within a stacking ensemble learning framework to weight models according to their performance. Evaluation on the standard EMBER dataset demonstrates that the proposed method achieves a precision of 96% and a high recall rate, delivering effective detection of zero-day malware. The key contributions of this research include efficient feature selection, the integration of stacking ensemble learning, and comprehensive evaluation with credible results.

Keywords- Zero-day malware, Malware detection, Fuzzy rough sets with Gaussian kernel, Hybrid features, Ensemble learning.

مجله علمی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
شناسه دیجیتال: 10.22034/jscit.2025.540111.2155
جلد ۱۴، شماره ۳، پاییز ۱۴۰۴ صفحه ۶۱-۷۶
دریافت: ۱۴۰۴/۰۵/۱۷، بازنگری: ۱۴۰۴/۰۷/۰۷، پذیرش: ۱۴۰۴/۰۷/۲۶

بهبود تشخیص بدافزار روز صفر مبتنی بر یادگیری تلفیقی و مجموعه‌های خشن فازی با کرنل گوسی

اکرم انصاری[۱]، جواد حمیدزاده*[۲]
۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه سجاد، مشهد، ایران.
*۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه سجاد، مشهد، ایران.
[1] akr.ansari226@sadjad.ac.ir, [2*]J_hamidzadeh@sadjad.ac.ir

* نشانی نویسنده مسئول: جواد حمیدزاده، مشهد، بلوار جلال آل احمد، جلال آل احمد ۶۴، دانشگاه سجاد، کد پستی: ۹۱۸۸۱۴۸۸۴۸

چکیده - تشخیص بدافزارهای روز صفر به دلیل ماهیت پویا و ناشناخته‌ی این تهدیدات، یکی از چالش‌های مهم امنیت سایبری است. وجود داده‌های ترکیبی شامل ویژگی‌های عددی و دسته‌ای با ابعاد زیاد، این مشکل را پیچیده‌تر می‌کند. این ناهمگنی در داده‌ها نه تنها بار محاسباتی را افزایش می‌دهد بلکه مرزهای تصمیم‌گیری میان کلاس‌های مخرب و سالم را مبهم می‌سازد. ویژگی‌های عددی جنبه‌های کمی مانند اندازه فایل یا تعداد بخش‌ها را دربرمی‌گیرند در حالی که ویژگی‌های دسته‌ای خصوصیات کیفی نظیر نوع فایل یا امضای کامپایلر را نشان می‌دهند. حضور همزمان این ویژگی‌ها انتخاب ویژگی را دشوار می‌کند زیرا روش‌های متداول توانایی کافی برای مدل‌سازی روابط غیرخطی همراه با عدم قطعیت بین آن‌ها را ندارند و این امر باعث کاهش عملکرد در تشخیص بدافزارهای روز صفر می‌شود. در این پژوهش، روشی بر پایه مجموعه‌های خشن فازی با کرنل گوسی برای انتخاب ویژگی‌های مؤثر ارائه شده است و ویژگی‌های منتخب در چارچوب یادگیری تلفیقی استکینگ به‌کار گرفته می‌شوند تا مدل‌ها بر اساس عملکردشان وزن‌دهی شوند. ارزیابی‌ها بر روی مجموعه‌داده استاندارد EMBER نشان می‌دهد که روش پیشنهادی به دقت ۹۶ درصد و نرخ فراخوانی بالا دست یافته و در شناسایی بدافزارهای روز صفر مؤثر عمل کرده است. مشارکت‌های اصلی پژوهش شامل انتخاب ویژگی کارآمد، به‌کارگیری یادگیری تلفیقی استکینگ و ارزیابی جامع با نتایج معتبر است.

واژه‌های کلیدی: بدافزار روز صفر، تشخیص بدافزار، مجموعه‌های خشن فازی مبتنی بر کرنل گوسی، ویژگی‌های ترکیبی، یادگیری تلفیقی.

۱- مقدمه

در عصر دیجیتال که داده‌ها از منابع گوناگون سرچشمه می‌گیرد، کاهش هدفمند ابعاد داده‌ها اهمیت حیاتی دارد. انتخاب ویژگی‌ها، به‌عنوان گامی مهم در پیش‌پردازش، با حذف دقیق ویژگی‌های غیرمرتبط یا زائد، به استخراج دانش ارزشمند از مجموعه داده‌ها کمک می‌کند [۲،۱]. این فرایند، ضمن ارتقای شفافیت تحلیل، بار محاسباتی را به‌شکل چشمگیری کاهش می‌دهد و با حفظ دقت سیستم‌های تصمیم‌گیری، توانایی مدل‌ها را در تعمیم مؤثر تقویت می‌کند. چالش شناسایی بدافزارهای روز صفر[۱]، به‌دلیل ماهیت ناشناخته و پویای این تهدیدات، در حوزه امنیت سایبری جایگاهی ویژه دارد. این نرم‌افزارهای مخرب، با بهره‌گیری از ترفندهای متنوع، از دام روش‌های سنتی تشخیص فرار می‌کنند. داده‌های مرتبط با بدافزارها، مانند مشخصه‌های استخراج‌شده از فایل‌های اجرایی، اغلب ترکیبی از ویژگی‌های عددی (مانند حجم فایل یا تعداد بخش‌ها) و دسته‌ای (مانند نوع فایل یا نشانگرهای خاص) هستند. این تنوع داده‌ها و همچنین ابعاد گسترده‌ی ویژگی‌ها، مدیریت عدم قطعیت و انتخاب ویژگی‌های کلیدی را به چالشی پیچیده تبدیل کرده است. روش‌های سنتی انتخاب ویژگی و طبقه‌بندی، اغلب در مواجهه با این چالش‌ها، به‌ویژه در داده‌های ترکیبی و تشخیص تهدیدات روز

صفر، با محدودیت‌هایی روبه‌رو هستند [۴،۳].

در فرایند انتخاب ویژگی، ماهیت ترکیبی داده‌های عددی و دسته‌ای به‌شکل معناداری بر پیچیدگی مسئله اثر می‌گذارد. هر نوع ویژگی دارای خصوصیات آماری و معیارهای شباهت متفاوتی است و همین امر باعث می‌شود روش‌های سنتی نتوانند به‌طور یکنواخت با همه ابعاد رفتار کنند [۶،۵]. علاوه بر این، در سناریوهای بدافزار روز صفر عدم قطعیت افزایش می‌یابد زیرا الگوهای مخرب ناشناخته ممکن است به‌طور همزمان در هر دو بُعد عددی و دسته‌ای ظاهر شوند [۸،۷]. یک روش کارآمد باید بتواند توزیع‌های ناهمگون داده را مدیریت کند، ماهیت اسمی و مفهوم اصلی ویژگی‌های دسته‌ای را حفظ کند و عدم قطعیت ناشی از وابستگی بین ویژگی‌ها را کاهش دهد. به عبارتی شناسایی بدافزار روز صفر به‌طور ذاتی با دو منبع اصلی عدم قطعیت مواجه است: غیرقابل پیش‌بینی بودن حملات جدید بدون نمونه قبلی در داده‌های آموزشی، ماهیت ناهمگون فضای ویژگی‌ها که شامل هر دو نوع ویژگی عددی و دسته‌ای است و وابستگی‌های غیرخطی و پیچیده ایجاد می‌کند. روش‌های متداول انتخاب ویژگی و طبقه‌بندی مانند PCA، Relief، SVM، RF، CNN و LSTM در مواجهه با این شرایط محدودیت دارند زیرا اغلب جداپذیری خطی را فرض می‌کنند، درجات عضویت را نادیده می‌گیرند، همه ویژگی‌ها را به‌طور یکسان پردازش می‌کنند و به نویز یا ویژگی‌های غیرمرتبط حساس‌تر هستند. در مقابل، نظریه مجموعه‌های خشن فازی، به‌ویژه در قالب مجموعه‌های خشن فازی با کرنل گوسی[۲] (GKFRS)، با استفاده از درجات عضویت مرزهای تصمیم‌گیری مبهم را مدل می‌کند، معیارهای شباهت را برای داده‌های ناهمگون به‌طور تطبیقی تنظیم می‌نماید و انتخاب ویژگی را با مدل‌سازی عدم قطعیت به‌طور همزمان انجام می‌دهد [۵]. این امر منجر به شناسایی ویژگی‌هایی می‌شود که بیشترین ارتباط را با کلاس‌های تصمیم دارند، همچنین افزونگی را کاهش می‌دهند و پایداری سیستم را در سناریوهای تهدید روز صفر افزایش می‌دهند.

تحقیقات اخیر نشان داده است که رویکردهای ترکیبی شامل یادگیری تلفیقی، منطق فازی و روش‌های مبتنی بر کرنل گوسی می‌توانند توانایی شناسایی بدافزارهای روز صفر را در محیط‌های پیچیده مانند اینترنت اشیا بهبود دهند [۱۰،۹،۵]. این رویکردها با بهره‌گیری از انتخاب ویژگی مقاوم و خوشه‌بندی فازی مبتنی بر کرنل، چالش‌های داده‌های پربعد، نویزی و غیرخطی را مدیریت می‌کنند و به کاهش نرخ هشدارهای کاذب کمک می‌کنند. این یافته‌ها هم راستا با اهداف پژوهش حاضر است و تأییدی بر اهمیت رویکرد پیشنهادی مبتنی بر GKFRS هستند.

این پژوهش راهکاری نوین برای شناسایی بدافزارهای روز صفر ارائه می‌دهد که بر انتخاب مؤثر ویژگی‌ها از داده‌های ترکیبی تمرکز دارد. روش پیشنهادی با بهره‌گیری از مجموعه‌های خشن فازی مبتنی بر کرنل گوسی، ویژگی‌های تأثیرگذار را با دقت بالایی شناسایی می‌کند [۱۴-۱۱]. سپس با استفاده از یادگیری تلفیقی[۳]، نمونه‌های مخرب و سالم به‌صورت کارآمدی طبقه‌بندی می‌شوند [۱۶،۱۵]. این رویکرد نه تنها دقت تشخیص را بهبود می‌بخشد بلکه با مدیریت بهتر داده‌های پیچیده، پایداری مدل را در برابر تهدیدات ناشناخته تقویت می‌کند.

ادامه‌ی مقاله به شرح زیر ساماندهی شده است:

بخش دوم به بررسی کارهای مرتبط، بخش سوم به مفاهیم پایه، بخش چهارم به روش پیشنهادی، بخش پنجم به نتایج آزمایش‌ها و بخش ششم به نتیجه‌گیری و پیشنهادهای آینده می‌پردازد.

۲-کارهای مرتبط

انتخاب ویژگی نقش کلیدی در کاهش پیچیدگی محاسباتی و افزایش کارایی مدل‌های یادگیری ماشین ایفا می‌کند. روش‌های انتخاب ویژگی به سه دسته‌ی اصلی فیلتر، پوشش‌دهنده[۴] و تعبیه‌شده[۵] تقسیم می‌شوند [۱۷]. روش‌های فیلتر ویژگی‌ها را براساس معیارهای آماری یا اطلاعاتی مستقل از الگوریتم‌های طبقه‌بندی ارزیابی می‌کنند. در مقابل، روش‌های پوشش‌دهنده با انتخاب زیرمجموعه‌ای از ویژگی‌ها و آزمایش آن‌ها در مدل طبقه‌بندی به بهینه‌سازی عملکرد کمک می‌کنند. روش‌های تعبیه شده با تلفیق انتخاب ویژگی و آموزش مدل، کارایی را به حداکثر می‌رسانند. مطالعات متعددی این حوزه را بررسی کرده‌اند؛ برای مثال، روش‌هایی مبتنی بر سنجش افزونگی ویژگی‌ها [۱۸]، استفاده از اطلاعات متقابل برای افزایش وابستگی میان ویژگی‌های منتخب و برچسب‌ها به‌صورت تدریجی [۱۱] و روش‌هایی که از معیارهای تمایزدهندگی درون و بین کلاسی بهره می‌برند [۱۹]. معیار Relief و بسط‌های آن نیز از جمله معیارهای رایج در روش‌های فیلتر هستند [۲۰]. همچنین، معیارهای مبتنی بر وابستگی در چارچوب مجموعه‌های خشن برای انتخاب ویژگی مورد توجه قرار گرفته‌اند [۱۴،۱۳،۱۱].

در سال‌های اخیر، رویکردهای مبتنی بر یادگیری عمیق و بصری‌سازی داده‌ها توجه زیادی را به خود جلب کرده است. با این‌وجود، این روش‌ها اغلب در حفظ خصوصیات معنایی و آماری داده‌ها با چالش مواجه هستند [۲۱-۲۷]. برخی پژوهش‌ها با افزودن ویژگی‌های معنایی به چارچوب‌های تشخیص، مقاومت در برابر

حملات خصمانه را افزایش دادند، چنین ویژگی‌هایی ممکن است به مدل کمک کنند تا انواع جدید و ناشناخته‌ی بدافزار را نیز شناسایی کند و حملات روز صفر را خنثی سازد [۲۸]. کاهش هشدارهای کاذب یکی از چالش‌های کلیدی در شناسایی بدافزارها به‌شمار می‌رود. این پژوهش با هدف کاستن از این هشدارها، روشی نوین برای استخراج ویژگی ارائه می‌دهد. استفاده از مجموعه داده‌های مبتنی بر گراف، امکان شناسایی مؤثر نرم‌افزارهای مخرب را فراهم کرده است و این داده‌ها برای آموزش شبکه‌های عصبی گراف به‌منظور طبقه‌بندی دقیق‌تر بدافزارها به‌کار گرفته شده‌اند [۲۹]. برخی پژوهشگران از مهندسی ویژگی برای شناسایی باتنت‌ها استفاده کردند اما به رویکردهای مبتنی بر منطق فازی کمتر توجه شده بود. برای رفع این کاستی، روشی مبتنی بر منطق فازی برای تولید ویژگی‌های نوین از مجموعه‌داده CTU-13 پیشنهاد شد که قابلیت بهبود عملکرد تشخیص را دارد [۳۰]. ماهیت پویا و متغیر بدافزارها ابهام و عدم قطعیت را در داده‌ها به همراه دارد. در این راستا، مجموعه‌های خشن[۶] و فازی[۷] ابزارهایی کارآمد برای مدل‌سازی این پیچیدگی‌ها ارائه می‌دهند. مجموعه‌های فازی با تمرکز بر عضویت جزئی نمونه‌ها در کلاس‌ها، انعطاف‌پذیری بالایی در تحلیل داده‌ها فراهم می‌کنند و مجموعه‌های خشن با بهره‌گیری از روابط عدم تمایز[۸]، تقریب‌های بالا و پایین را برای تحلیل دقیق‌تر داده‌ها معرفی می‌کنند [۱۴،۱۳،۱۱]. برای بهبود کارایی مجموعه‌های خشن فازی در مدیریت داده‌های پیچیده، مفهوم کرنل استفاده شده است [۳۱-۳۳]. مجموعه‌های خشن فازی مبتنی بر کرنل با تلفیق مفاهیم فازی و توابع کرنل، شباهت‌های غیرخطی میان نمونه‌ها را محاسبه می‌کنند و تقریب‌های دقیق‌تری از مجموعه‌های خشن فازی ارائه می‌دهند. این قابلیت، KFRS را در مدل‌سازی عدم قطعیت و وابستگی‌های پیچیده در داده‌های ترکیبی پربُعد[۹] توانمند ساخته است و انتخاب ویژگی‌های مرتبط با کلاس‌های هدف (مخرب/غیرمخرب) را با دقت بالاتری ممکن می‌سازد [۴،۳].

پژوهش‌های اخیر، استفاده از مفاهیمی مانند مجموعه‌های فازی تنک[۱۰] [۲۰] و معیار تمایزپذیری فازی [۴،۳۴] را در انتخاب ویژگی بررسی کرده‌اند که اثربخشی ابزارهای فازی را در این حوزه نشان می‌دهد. همچنین، روش‌های مبتنی بر مجموعه‌های خشن فازی با بهره‌گیری از کرنل‌ها برای تنظیم پویا و خودکار معیارهای ارزیابی ویژگی‌ها و مدیریت داده‌های ترکیبی پیشنهاد شده است [۱۲]. Jain و همکاران رویکرد یادگیری بدون نظارت را برای حملات روز صفر معرفی کرده‌اند [۹]، Karakaya چارچوب امنیت اینترنت اشیا مبتنی بر یادگیری تلفیقی و منطق فازی ارائه کرده است [۱۰]، Dong و همکاران انتخاب ویژگی مقاوم با مجموعه‌های خشن فازی را توسعه داده‌اند [۵]. Yang و Wang روش‌های انتخاب ویژگی و شبکه عصبی فازی خشن تطبیقی را مطرح کرده‌اند [۳۶،۳۵]. در حوزه کرنل‌های گوسی، مدل‌های خوشه‌بندی فازی مبتنی بر کرنل گوسی با تنظیم خودکار ارائه شد [۳۷]. این مطالعات هم راستا با اهداف این پژوهش است و اهمیت رویکرد پیشنهادی مبتنی بر GKFRS را تأیید می‌کنند.

پس از انتخاب ویژگی‌های مؤثر، طبقه‌بندی نمونه‌ها برای شناسایی بدافزارها گامی کلیدی است. یادگیری تلفیقی به‌عنوان رویکردی قدرتمند در سال‌های اخیر به دلیل توانایی در افزایش دقت و پایداری مدل‌های طبقه‌بندی، به‌ویژه در مسائل پیچیده مانند شناسایی بدافزار، توجه بسیاری را جلب کرده است [۱۶]. روش‌های تلفیقی با ترکیب نتایج چندین مدل پایه، خطاهای پیش‌بینی را کاهش می‌دهند و عملکرد کلی سیستم را بهبود می‌بخشند [۱۵].

روش‌های متنوعی از جمله بگینگ[۱۱]، بوستینگ[۱۲]، XGBoost و استکینگ[۱۳] برای پیاده‌سازی یادگیری تلفیقی وجود دارد. در این میان، استکینگ به‌دلیل سرعت بالا، دقت مناسب و قابلیت مدیریت داده‌های گمشده در مسائل طبقه‌بندی پیچیده کاربرد گسترده‌ای یافته است. پژوهش‌ها نشان داده است که ترکیب یادگیری تلفیقی با انتخاب ویژگی می‌تواند عملکرد طبقه‌بندی را در مسائل پیچیده بهبود بخشد [۱۶].

با مرور پژوهش‌های پیشین می‌توان دریافت که هر یک از آن‌ها دارای محدودیت‌هایی هستند. روش‌های مبتنی بر شبکه‌های عصبی عمیق مانند CNN و RNN اگرچه دقت بالایی نشان داده‌اند اما نیازمند داده‌های بسیار حجیم، منابع محاسباتی پرهزینه و زمان آموزش طولانی هستند و در مواجهه با داده‌های نامتوازن عملکرد بهینه ندارند. از سوی دیگر، روش‌های کلاسیک انتخاب ویژگی مانند PCA یا Relief عمدتاً بر کاهش بُعد داده متمرکز هستند و در مدیریت عدم قطعیت ضعف دارند. همچنین در مطالعات گذشته، به مسئله‌ی روز صفر و ناشناختگی تهدیدات واقعی کمتر توجه شده است و بیشتر پژوهش‌ها بر داده‌های استاندارد و ساده متمرکز بوده‌اند. بنابراین، یک شکاف آشکار در ادبیات وجود دارد؛ نبود رویکردی که بتواند همزمان ویژگی‌های کلیدی را از داده‌های پُربعد با استفاده از مدل‌سازی غیرخطی و مدیریت عدم قطعیت استخراج کند و این ویژگی‌ها را در یک روش یادگیری تلفیقی با وزن‌دهی مبتنی بر عملکرد مدل‌ها به‌کار گیرد، به‌گونه‌ای که سیستم تشخیص در برابر تغییرات پویا و ناشناخته بدافزارهای روز صفر پایدار بماند و عملکرد بالایی داشته باشد. بر همین اساس، پژوهش حاضر ترکیب نوینی از انتخاب ویژگی مبتنی بر مجموعه‌های خشن فازی با کرنل گوسی و

یادگیری تلفیقی استکینگ را ارائه می‌دهد تا هم دقت و هم کارایی محاسباتی در شناسایی بدافزارهای روز صفر بهبود یابد.

۳- مقدمات

در این بخش، مفاهیم پایه که مبنای روش پیشنهادی را تشکیل می‌دهند، به طور مختصر ارائه می‌شود.

۳-۱- بدافزار روز صفر

شناسایی بدافزارهای روز صفر یکی از دشوارترین چالش‌های امنیت سایبری است، زیرا هیچ اطلاعات قبلی درباره این تهدیدات در دسترس نیست. این کمبود اطلاعات، روش‌های سنتی تشخیص بدافزار را ناکارآمد می‌کند، چرا که این روش‌ها معمولاً به داده‌های قبلی وابسته است. بدافزارهای روز صفر با بهره‌گیری از تکنیک‌های رمزنگاری پیچیده به سرعت عمل می‌کنند و می‌توانند خسارت‌های جبران‌ناپذیری به سیستم‌ها وارد کنند. از این‌رو، شناسایی سریع این تهدیدات ترجیحاً پیش از آغاز فرایند رمزگذاری، اهمیتی حیاتی دارد. این پژوهش با ارائه راهکاری نوین تلاش می‌کند تا این چالش را به‌گونه‌ای کارآمد برطرف کند.

۳-۲- مجموعه‌های خشن و مجموعه‌های خشن فازی

مجموعه‌های خشن، ابزاری کارآمد برای تحلیل داده‌ها بر پایه‌ی روابط تمیزناپذیری هستند. فرض کنید یک سیستم اطلاعاتی $IS = (U, A)$ داریم که در آن $(U = \{x_1, \dots, x_m\})$ مجموعه‌ای متناهی و ناتهی از اشیا و $(A = \{a_1, a_2, \dots, a_n\})$ مجموعه‌ای متناهی و ناتهی از صفت‌ها برای توصیف اشیاست. برای هر زیرمجموعه‌ای از صفت‌ها $(B \subseteq A)$، رابطه‌ی تمیزناپذیری $IND(B)$ که اساس تقریب‌های پایین و بالا را شکل می‌دهد، به‌صورت رابطه (۱) تعریف می‌شود [۳۸]:

$$IND(B) = \{(x, y) \in U \times U : a(x) = a(y),\ \forall a \in B\} \quad (۱)$$

رابطه‌ی $IND(B)$ در (۱) به صورت شرطی بر همه صفات B تعریف می‌شود؛ x و y زمانی نسبت به B غیرقابل‌تمایز هستند که برای هر صفت $(a \in B)$ مقادیر آن‌ها برابر باشد. این تعریف یک رابطه‌ی هم‌ارزی است زیرا ویژگی‌های بازتابندگی، تقارن و تراگذری را دارد. در رابطه‌ی (۲)، این مفهوم با استفاده از نظریه روابط و اشتراک روابط تک‌ویژگی بازنویسی شده است. برای هر ویژگی منفرد $a \in B$، رابطه $IND(\{a\})$ تمیزناپذیری را تنها بر اساس آن ویژگی تعریف می‌کند. اگر بخواهیم تمیزناپذیری را براساس مجموعه‌ای از ویژگی‌های B تعریف کنیم، باید فقط زوج‌هایی را بپذیریم که در همه روابط تک ویژگی $(IND(\{a\}))$ صدق کنند. این دقیقاً همان اشتراک روابط است که در رابطه‌ی (۲) تعریف می‌شود:

$$IND(B) = \bigcap_{a \in B} IND(\{a\}) \quad (۲)$$

بنابراین رابطه‌ی (۲) نتیجه مستقیم بازنویسی رابطه‌ی (۱) بر اساس اشتراک روابط تک ویژگی است. این دو رابطه از نظر منطقی و ریاضی کاملاً معادل‌اند و فقط نحوه بیانشان متفاوت است؛ یکی با شرط $(\forall a \in B)$ و دیگری با عملگر اشتراک $(\cap)$ بیان شده است. از این رابطه‌ی هم‌ارزی، کلاس‌های هم‌ارزی طبق رابطه‌ی (۳) زیر حاصل می‌شوند:

$$[x]_B = y \in U \mid (x, y) \in IND(B) \quad (۳)$$

که مجموعه‌ی تمامی اشیاء مشابه x نسبت به B هستند. اجتماع همه‌ی این کلاس‌ها یک پارتیشن از U طبق رابطه‌ی (۴) می‌سازد:

$$U/IND(B) = \{[x]_B : x \in U\} \quad (۴)$$

حال برای هر زیرمجموعه‌ی دلخواه $(X \subseteq U)$، تقریب‌های پایین و بالای (B) از (X) به‌صورت رابطه (۵) و (۶) تعریف می‌شوند:

$$\underline{B}X = \bigcup \{[x]_B : [x]_B \subseteq X\} \quad (۵)$$

$$\overline{B}X = \bigcup \{[x]_B : [x]_B \cap X \neq \emptyset\} \quad (۶)$$

تقریب پایین $(\underline{B}X)$: اجتماع کلاس‌هایی که کاملاً در X قرار دارند.
تقریب بالا$(\overline{B}X)$: اجتماع کلاس‌هایی که با X اشتراک غیرتهی دارند.

اگر $\left(\underline{B}X = \overline{B}X\right)$، مجموعه‌ی (X) تعریف‌پذیر است؛ در غیر این صورت (X) یک مجموعه خشن است و مرز آن به‌صورت رابطه (۷) تعریف می‌شود:

$$BN(X) = \overline{B}X - \underline{B}X \quad (۷)$$

مجموعه‌های خشن کلاسیک به‌دلیل وابستگی به روابط هم‌ارزی، برای داده‌هایی با مقادیر نمادین مناسب‌تر هستند و در تحلیل داده‌ها با مقادیر حقیقی محدودیت دارند. برای رفع این محدودیت، مجموعه‌های خشن فازی توسط Dubois و Prade معرفی شدند که با ترکیب مفاهیم مجموعه‌های خشن و فازی، انعطاف‌پذیری بیشتری در مدل‌سازی عدم قطعیت فراهم می‌کنند. رابطه فازی (R) روی (U)، در صورتی رابطه (T)- هم‌ارزی فازی نامیده می‌شود که سه ویژگی بازتابندگی $(R(x, x) = 1)$، تقارن $\left(R(x, y) = R(y, x)\right)$، و تراگذری فازی $\left(T\left(R(x, y), R(y, z)\right) \leq R(x, z)\right)$ را نسبت به نورم مثلثی (T) داشته باشد [۳۸].

تعاریف دقیق تقریب‌های پایین و بالای فازی برای زیرمجموعه فازی ($A \subseteq U$) در روابط (۱۷) تا (۲۲) بخش ۴-۲-۲، به‌طور خاص برای کرنل گوسی ارائه شده است. این تقریب‌ها امکان تحلیل داده‌های پیچیده را در الگوریتم‌های (۱) و (۲) جهت تشخیص بدافزارهای روز صفر فراهم می‌کنند.

در چارچوب مدل مجموعه‌های خشن فازی مبتنی بر کرنل گوسی (GKFRS) که نخستین‌بار توسط Hu و همکاران در سال ۲۰۱۰ معرفی شد [۳۸]، توابع کرنل گوسی دارای ویژگی‌های بازتابندگی، تقارن و مثبت نیمه‌معین (PSD)[۱۴] هستند. برخورداری از این خواص، شرط لازم برای این است که ماتریس شباهت حاصل بتواند به‌عنوان یک رابطه T- هم‌ارزی فازی معتبر به‌کار رود. از این رو، در کاربردهای GKFRS، کرنل گوسی به دلیل حفظ خواص ریاضی، پیوستگی و مشتق‌پذیری کامل، توانایی مدل‌سازی دقیق عدم قطعیت در داده‌های ترکیبی و قابلیت استنتاج رابطه T-هم‌ارزی فازی، به‌عنوان گزینه اصلی انتخاب می‌شود و عملکرد بهتری در استخراج ویژگی‌های مؤثر فراهم می‌سازد. مطالعات اخیر نیز برتری این انتخاب را تأیید کرده‌اند و نشان داده‌اند که کرنل‌های گوسی در کاربردهای خوشه‌بندی و تشخیص ناهنجاری داده‌های پیچیده، نسبت به کرنل‌های غیر-PSD سازگاری و دقت بیشتری دارند [۳۷].

۴- روش پیشنهادی

این پژوهش راهکاری نوین برای شناسایی بدافزارهای روز صفر ارائه می‌دهد که در دو گام اصلی طراحی شده است. گام نخست، انتخاب ویژگی‌های کلیدی با استفاده از مجموعه‌های خشن فازی مبتنی بر کرنل گوسی است که ویژگی‌هایی با بیشترین ارتباط با کلاس‌های تصمیم (مخرب/غیرمخرب) و کمترین افزونگی را شناسایی می‌کند. در گام دوم، چارچوب پیشنهادی با به‌کارگیری یادگیری تلفیقی استکینگ، دقت طبقه‌بندی را به‌طور معناداری ارتقا می‌دهد. برخلاف رویکردهای تک مدلی، استکینگ از ترکیب نقاط قوت مدل‌های پایه گوناگون که هر یک تعاملات متفاوتی بین ویژگی‌ها و مرزهای تصمیم را در فضای ویژگی‌های ناهمگون شناسایی می‌کنند، بهره می‌گیرد. مدل فراگیر با یادگیری تطبیقی از داده‌های واقعی، پیش‌بینی مدل‌هایی را که در نواحی خاص عملکرد بهتری دارند وزن بیشتری می‌دهد. این بهینه‌سازی تطبیقی، بایاس و واریانس را کاهش می‌دهد و در نتیجه با کاهش خطاهای مثبت کاذب، دقت و با شناسایی بهتر الگوهای حمله جدید، نرخ فراخوانی را بهبود می‌بخشد. ترکیب این قابلیت‌ها با ویژگی‌های انتخاب شده توسط GKFRS یک تابع تصمیم‌گیری متمایز و مقاوم در برابر نویز ایجاد می‌کند و کل چارچوب را به راه‌حلی بسیار مؤثر برای شناسایی بدافزار روز صفر تبدیل می‌کند.

۴-۱- پیش‌پردازش

برای آماده‌سازی داده‌ها، از مجموعه داده EMBER با ۹۰۰،۰۰۰ نمونه آموزشی استفاده شد که شامل ۳۰۰،۰۰۰ نمونه مخرب، ۳۰۰،۰۰۰ نمونه غیرمخرب، و ۳۰۰،۰۰۰ نمونه بدون برچسب است [۳۹]. در ابتدای فرایند، تعداد ۱۰،۷۵۴ نمونه که دارای برچسب نامشخص (۱-) و ۳۰۰،۰۰۰ نمونه که فاقد برچسب بودند از مجموعه داده حذف شدند. در نتیجه، مجموعه‌داده نهایی با ۵۸۹،۲۴۶ نمونه، شامل ۳۰۰،۰۰۰ نمونه غیرمخرب و ۲۸۹،۲۴۶ نمونه مخرب، برای تحلیل آماده شد. ویژگی‌های دسته‌ای و متنی با روش کدگذاری[۱۵] مناسب به بردارهای باینری تبدیل شدند. ویژگی‌های عددی با استفاده از استانداردسازی[۱۶] نرمال‌سازی شدند تا میانگین صفر و انحراف معیار یک داشته باشند. همچنین، ویژگی‌های لیستی بر اساس فراوانی یا وجود/عدم وجود عناصر کلیدی به بردارهای عددی تبدیل شدند. این فرایند، داده‌ها را برای تحلیل‌های بعدی یکنواخت و آماده کرد.

۴-۲- انتخاب و کاهش ویژگی

در این بخش، فرایند انتخاب ویژگی‌ها با استفاده از مجموعه‌های خشن فازی مبتنی بر کرنل گوسی به‌صورت گام به گام تشریح می‌شود. فرض کنید داده‌های ورودی شامل m ویژگی عددی و دسته‌ای پیش‌پردازش‌شده (نرمال‌سازی و حذف مقادیر گمشده) است. مجموعه‌ی $B \subseteq \{1, \dots, m\}$ نمایانگر زیرمجموعه‌ی ویژگی‌های منتخب است و U مجموعه‌ی نمونه‌ها است. این روش با محاسبه‌ی روابط فازی و استفاده از تقریب‌های بالا و پایین، ویژگی‌های کلیدی را شناسایی می‌کند:

۱- استخراج بردارهای ویژگی خام: داده‌های ترکیبی (عددی/دسته‌ای) پس از پیش‌پردازش به شکل بردار استاندارد برای تحلیل آماده می‌شوند.

۲- محاسبه‌ی رابطه‌ی فازی گاوسی: برای هر ویژگی منفرد $a_j \in B$، از کرنل گوسی استفاده می‌شود تا شباهت بین دو نمونه x_i و x_k در آن ویژگی طبق رابطه‌ی (۸) محاسبه شود [۳۸]:

$$R_G^{(j)}(x_i, x_k) = \exp\left(-\frac{(x_{ij} - x_{kj})^2}{2d^2}\right) \qquad (۸)$$

در رابطه‌ی (۸)، d پارامتر پهنای باند گوسی است. این ماتریس شباهت، یک رابطه‌ی فازی بازتابنده و متقارن ایجاد می‌کند که با استفاده از نورم مثلثی کسینوسی(T_{cos}) در ادغام روابط ویژگی‌ها،

به رابطه هم‌ارزی فازی تبدیل می‌شود.

۳- تقریب‌های پایین و بالای فازی: نظریه مجموعه‌های خشن فازی، برای هر نمونه، دو نوع تقریب را تعریف می‌کند که بر اساس رابطه فازی بین نمونه‌ها و کلاس تصمیم d_i (طبق روابط (۹) و (۱۰)) محاسبه می‌شوند [۳۸]:

$$\underline{R_G^n}\mathrm{d_i(x)} = \inf_{\mathrm{y\in U}} \mathrm{h}\left(\mathrm{R_G^n(x,y)}, \mathrm{d_i(y)}\right) \qquad (۹)$$

$$\overline{R_G^n} d_i(x) = \sup_{y\in U} T\left(R_G(x,y), d_i(y)\right) \qquad (۱۰)$$

در رابطه‌ی (۹) و (۱۰)، h و T به‌ترتیب نورم‌ها/کانرم‌های فازی هستند. این دو تقریب به‌ترتیب نشان‌دهنده‌ی درجه‌ی اطمینان و میزان احتمال عضویت نمونه در کلاس هستند.

در پیاده‌سازی این پژوهش، تنها تقریب پایین برای ارزیابی ویژگی‌ها استفاده شده است، زیرا این معیار هم با رویکرد قطعیت محور ما منطبق است و هم از نظر منابع محاسباتی کم هزینه‌تر است.

۴- محاسبه‌ی ناحیه‌ی مثبت فازی و ضریب وابستگی: برای هر زیرمجموعه‌ی ویژگی(B)، ناحیه‌ی مثبت فازی به‌صورت رابطه (۱۱) تعریف می‌شود [۳۸]:

$$POS_B(D) = \bigcup_{i=1}^{I} \underline{R_G} d_i \qquad (۱۱)$$

ضریب وابستگی ویژگی‌ها به کلاس تصمیم با رابطه (۱۲) محاسبه می‌شود [۳۸]:

$$\gamma_B(D) = \frac{POS_B(D)}{|U|} \qquad (۱۲)$$

الگوریتم ۱ (وابستگی با تقریب کرنل گاوسی – DGKA[۱۷]) مقدار $\gamma_B(D)$ را برای تمام زیر مجموعه‌های ممکن B محاسبه و این زیرمجموعه‌ها را براساس مقادیر به دست آمده از $\gamma_B(D)$ رتبه‌بندی می‌کند.

۵- جستجوی حریصانه‌ی پیشرو[۱۸] (FS-GKA): الگوریتم ۲ (انتخاب ویژگی مبتنی بر تقریب‌های کرنل گاوسی) از مجموعه خالی شروع می‌کند و در هر مرحله، ویژگی‌ای را اضافه می‌کند که بیشترین افزایش در $\gamma_B(D)$ را ایجاد کند. این فرایند تا زمانی ادامه می‌یابد که بهبود بیشتری حاصل نشود یا تعداد ویژگی‌ها به حد تعیین‌شده برسد [۳۸].

۶- خروجی: زیرمجموعه‌ی بهینه B^* که بیشترین $\gamma_{B^*}(D)$ و کمترین ویژگی‌های زائد را دارد، به عنوان مجموعه‌ی ویژگی‌های منتخب خروجی داده می‌شود. این مجموعه، بیشترین قدرت تمایز بین بدافزار و نرم‌افزار سالم را فراهم می‌کند. در ادامه، جزئیات این مراحل به‌صورت کامل بیان می‌شود.

۴-۲-۱- مدل مجموعه‌ی خشن فازی مبتنی بر کرنل گوسی

مجموعه‌های خشن فازی مبتنی بر کرنل گوسی (GKFRS) ابزاری توانمند برای انتخاب ویژگی‌های کلیدی در تشخیص بدافزارهای روز صفر فراهم می‌کنند. در این روش، کیفیت هر زیرمجموعه از ویژگی‌ها بر اساس توانایی آن در تمایز بین نمونه‌های مخرب و سالم ارزیابی می‌شود و در نهایت، مؤثرترین ویژگی‌ها استخراج می‌شوند.

کرنل گوسی که در رابطه‌ی (۸) بخش ۴-۲ معرفی شده است، به دلیل توانایی در مدل‌سازی روابط غیرخطی و سازگاری با داده‌های ترکیبی، برای ایجاد روابط هم‌ارزی فازی نوع T- نورم به‌کار می‌رود. این کرنل داده‌ها را به فضایی با ابعاد بالاتر نگاشت می‌کند و بر اساس آن، میزان شباهت بین هر دو نمونه را با دقت بالایی محاسبه می‌نماید. رابطه‌ی فازی R_G^n حاصل از کرنل گوسی، با برخورداری از ویژگی‌های بازتابندگی، تقارن و تراگذری نسبت به نورم مثلثی کسینوسی (T_{cos})، یک رابطه‌ی هم‌ارزی فازی محسوب می‌شود.

نورم مثلثی کسینوسی T_{cos}، طبق رابطه‌ی (۱۳)، به‌عنوان یک عملگر ادغام، میزان ترکیب دو درجه شباهت یا عضویت را محاسبه می‌کند [۳۸]:

$$T_{cos}(a,b) = \max\{ ab - \sqrt{(1-a^2)(1-b^2)}, 0\} \qquad (۱۳)$$

در رابطه‌ی (۱۳)، a و b بین ۰ و ۱ هستند. این نورم ضمن حفظ خواص بازتابندگی و تقارن، خروجی را در بازه $[۰,۱]$ تضمین می‌کند.

استلزام باقیمانده نورم مثلثی کسینوسی که به‌صورت $h_{T_{cos}}$ نمایش داده می‌شود، جهت محاسبه‌ی تقریب‌های فازی به کار می‌رود و به صورت رابطه‌ی (۱۴) تعریف می‌شود [۳۸]:

$$h_{T_{cos}}(a,b) = \{ ۱ \quad ,a \le b \qquad (۱۴)$$

$$ab + \sqrt{\{(1-a\hat{}2)(1-b\hat{}2)\}}, a > b\}$$

این تابع، درجه تحقق یا میزان برآورده‌سازی شرط فازی $a \Rightarrow b$ را محاسبه می‌کند. به‌عبارت دیگر، بیانگر آن است که تا چه حد مقدار aبا مقدار b در چارچوب منطق فازی سازگار است. در این تعریف، اگر $a \le b$ باشد، شرط فازی به‌طور کامل برقرار بوده و مقدار تابع برابر ۱ است؛ اما اگر $a > b$ باشد، میزان تحقق شرط بر اساس ترکیب ضربی دو مقدار و مؤلفه کسینوسی حاصل از آن‌ها محاسبه می‌شود که این مؤلفه هندسی ارتباط بین درجات عضویت را در فضای فازی بازنمایی می‌کند. این رویکرد موجب حفظ مقادیر در

بازه [۰،۱] و سازگاری با خواص بازتابندگی و تقارن در روابط فازی می‌شود. در این پژوهش، روابط (۱۳) و (۱۴) نقش پل ارتباطی بین میزان شباهت نمونه‌ها و درجه اطمینان عضویت آن‌ها در کلاس تصمیم را ایفا می‌کنند. این روابط در محاسبه‌ی تقریب‌های پایین و بالا در مدل مجموعه‌های خشن فازی مبتنی بر کرنل گوسی به‌کار می‌روند و جزئیات استفاده از آن‌ها در بخش ۴-۲-۲ ارائه شده است.

۴-۲-۲- تقریب نواحی تصمیم با کرنل گاوسی

در این بخش، فرآیند تقریب نواحی تصمیم با بهره‌گیری از کرنل گوسی تشریح می‌شود. هدف اصلی آن است که مشخص شود هر نمونه تا چه حد به دسته‌های موجود (مانند «بدافزار» یا «سالم») شباهت دارد، حتی در شرایطی که مرز بین کلاس‌ها واضح نیست. ابتدا شباهت بین نمونه‌ها برای هر ویژگی (j) با استفاده از تابع کرنل گوسی و طبق رابطه (۸) بخش ۴-۲ محاسبه می‌شود. این کرنل مانند یک فیلتر نرم عمل می‌کند و فاصله‌ها را به مقادیر بین صفر و یک تبدیل می‌نماید؛ هرچه دو نمونه شبیه‌تر باشند این مقدار بزرگ‌تر خواهد بود. نتیجه این مرحله، مجموعه‌ای از روابط فازی $\{R_G^{(1)}, R_G^{(2)}, \dots, R_G^{(n)}\}$ است که در آن (n) تعداد ویژگی‌هاست. پس از محاسبه روابط فازی برای هر ویژگی، لازم است این روابط برای تصمیم‌گیری تجمیع شوند. این روابط با استفاده از حاصل‌ضرب جبری، طبق رابطه (۱۵)، تجمیع می‌شوند و رابطه فازی نهایی (R_G^n) را تشکیل می‌دهند [۳۸].

$$R_G^n(x_i, x_k) = \exp\left(-\frac{\left||x_i - x_k|\right|^2}{2d^2}\right) = \prod_{s=1}^{n} R_G^{(s)}(x_i, x_k) \quad (۱۵)$$

ویژگی تصمیم (D) نمونه‌ها را به کلاس‌های $\{d_1, d_2, \dots, d_I\}$ تقسیم می‌کند. تابع عضویت هر نمونه در یک کلاس خاص (d_i) به‌صورت رابطه (۱۶) تعریف می‌شود:

$$\forall x \in U: di(x) = \{1 \quad x \in di \quad (۱۶)$$

$$\{\text{در غیر این صورت} \quad 0$$

برای تعیین حدود نواحی تصمیم، از دانه‌های فازی[۱۹] ایجادشده بر اساس رابطه‌ی فازی (R_G^n) استفاده می‌کنیم. این دانه‌ها، گروه‌هایی از نمونه‌های مشابه هستند که بر اساس شباهت فازی بین آن‌ها تشکیل شده‌اند. حال، اگر کلاس iام را در نظر بگیریم، درجه‌ی عضویت هر نمونه به این کلاس به‌صورت زیر محاسبه می‌شود [۳۸]:

(۱) تقریب پایین: این تقریب نمونه‌هایی را مشخص می‌کند که با قطعیت بالا متعلق به یک کلاس هستند، و در نتیجه به‌شدت احتمال هشدارهای کاذب را کاهش می‌دهد. با استفاده از تابع استلزام باقیمانده $h_{T_{cos}}$ و مطابق رابطه (۹) داریم:

اگر $d_i(y) = 1$ (یعنی $y \in d_i$)، آنگاه $R_G^n(x,y) \le d_i(y)$ و خواهیم داشت:

$$h\left(R_G^n(x,y), d_i(y)\right) = 1 \quad (۱۷)$$

- اگر $d_i(y) = 0$ (یعنی $y \notin d_i$)، آنگاه:

$$h\left(R_G^n(x,y), d_i(y)\right) = \sqrt{1 - \left(R_G^n(x,y)\right)^2} \quad (۱۸)$$

در نتیجه، تقریب پایین فازی به شکل رابطه (۱۸) ساده می‌شود:

$$\underline{R_G^n} d_i(x) = \inf_{y \notin d_i} \sqrt{1 - \left(R_G^n(x,y)\right)^2} \quad (۱۹)$$

رابطه‌ی (۱۹) کمترین شباهت فازی بین (x) و نمونه‌های کلاس‌های دیگر را نشان می‌دهد، و هرچه این مقدار بزرگ‌تر باشد اطمینان ما از تعلق نمونه به کلاس d_i بیشتر است.

(۲) تقریب بالا: این تقریب مجموعه‌ای از نمونه‌ها را مشخص می‌کند که احتمال تعلق آن‌ها به یک کلاس وجود دارد، حتی اگر این احتمال قطعی نباشد. طبق رابطه (۱۰) و با استفاده از نورم مثلثی کسینوسی T_{cos}، داریم:

اگر $d_i(y) = 1$ (یعنی $y \in d_i$):

$$T\left(R_G(x,y), d_i(y)\right) = R_G(x,y) \quad (۲۰)$$

- اگر $d_i(y) = 0$:

$$T\left(R_G(x,y), d_i(y)\right) = 0 \quad (۲۱)$$

در نتیجه، تقریب بالای فازی به شکل رابطه (۲۲) ساده می‌شود. اگر $(x \in d_i)$ باشد، این مقدار برابر ۱ است؛ در غیر این صورت، بیشینه شباهت (x) به نمونه‌های کلاس (d_i) را نشان می‌دهد.

$$\overline{R_G^n} d_i(x) = \sup_{y \in d_i} R_G(x,y) \quad (۲۲)$$

به‌طور خلاصه در این فرآیند، ابتدا میزان شباهت هر نمونه با بقیه نمونه‌ها بر اساس کرنل گوسی محاسبه می‌شود و از آن دانه‌های فازی (گروه‌های نمونه‌های مشابه) ساخته می‌شوند. سپس تقریب پایین، نمونه‌هایی را مشخص می‌کند که با اطمینان بالا به یک کلاس تعلق دارند (تمرکز بر قطعیت و کاهش هشدار کاذب) و تقریب بالا، نمونه‌هایی را مشخص می‌کند که احتمال تعلق آن‌ها به یک کلاس

وجود دارد (تمرکز بر پوشش کامل). به این ترتیب، هر یک از این دو تقریب مرزهای قطعی و احتمالی کلاس‌ها را تعیین می‌کنند. با این حال، در این پژوهش برای مراحل بعدی تنها از تقریب پایین استفاده شده است تا دقت و پایداری مدل افزایش یابد.

۴-۲-۳- ارزیابی و کاهش ویژگی با مجموعه‌های خشن مبتنی بر کرنل گاوسی

نظریه‌ی مجموعه‌های خشن فازی ابزارقدرتمندی برای سنجش توانایی مجموعه ویژگی‌های شرطی C در تفکیک دقیق نمونه‌ها بر اساس برچسب‌های تصمیم D فراهم می‌آورد. این سنجش با محاسبه‌ی وابستگی فازی $\gamma_B(D)$ بین یک زیرمجموعه ویژگی $B \subseteq C$ و کلاس‌های تصمیم انجام می‌شود. در رویکرد پیشنهادی، ابتدا رابطه‌ی شباهت فازی (R_G^n) میان نمونه‌ها محاسبه می‌شود تا درجه‌ی نزدیکی یا شباهت هر جفت نمونه را با در نظر گرفتن ویژگی‌های B، کمی‌سازی کند و ساختار ارتباطی داده‌ها مدل‌سازی شود. این رابطه، بنیان تشکیل گرانول‌های اطلاعاتی فازی است که مبنای محاسبه‌ی تقریب‌ها می‌باشد. سپس، با استفاده رابطه‌ی شباهت فازی (R_G^n)، تقریب پایین فازی $\underline{R_G^n} d_i(x)$ طبق رابطه (۱۹) برای هر کلاس d_i محاسبه می‌شود. این مقدار نشان‌دهنده‌ی میزان اطمینان تعلق یک نمونه x به کلاس d_i است و بر اساس کمترین شباهت آن با نمونه‌های سایر کلاس‌ها تعیین می‌شود. از ترکیب این مقادیر برای تمام نمونه‌ها، ناحیه‌ی مثبت فازی $POS_B(D)$ طبق رابطه (۱۱) حاصل می‌شود که شامل نمونه‌هایی است که با توجه به ویژگی‌های B، بدون ابهام به یک کلاس مشخص تعلق دارند. نسبت اندازه‌ی این ناحیه به کل مجموعه داده، تابع وابستگی فازی $\gamma_B(D)$ را طبق رابطه (۱۲) تشکیل می‌دهد که معیار اصلی توان تمایز ویژگی‌های B محسوب می‌شود هدف نهایی در فرآیند انتخاب ویژگی، یافتن کوچک‌ترین زیرمجموعه‌ی B است که ضمن بیشینه‌سازی $\gamma_B(D)$، ویژگی‌های زائد یا کم اثر را حذف کند. در این پژوهش، الگوریتم (۱) این مراحل را برای ارزیابی اهمیت ویژگی‌ها و محاسبه‌ی وابستگی فازی ($\gamma_B(D)$) به‌صورت خودکار اجرا می‌کند. ویژگی‌های نامرتبط، به‌دلیل دریافت مقادیر وابستگی پایین از مجموعه داده حذف می‌شوند. با این حال، الگوریتم (۱) به‌تنهایی قادر به شناسایی ویژگی‌های تکراری نیست؛ زیرا ممکن است دو ویژگی با وابستگی بالا، اطلاعات مشابه ارائه دهند. برای رفع این مشکل، الگوریتم (۲) با رویکرد جست‌وجوی حریصانه‌ی پیشرو به‌کار گرفته می‌شود تا زیرمجموعه‌ای بهینه از ویژگی‌ها انتخاب شود که افزونگی را حذف کند و قطعیت طبقه‌بندی را تا حد امکان افزایش دهد. ساختار دومرحله‌ای الگوریتم (۱) و (۲)، ضمن اطمینان از حداکثر تمایزپذیری، با حذف هم‌زمان ویژگی‌های نامرتبط و تکراری، مجموعه‌ی ویژگی بهینه‌ای را تولید می‌کند که در ادامه توسط یادگیری تلفیقی برای طبقه‌بندی نمونه‌ها مورد استفاده قرار می‌گیرد.

الگوریتم ۱: وابستگی با تقریب کرنل گاوسی (DGKA).

ورودی‌ها:

U: مجموعه داده‌ها (کل نمونه‌ها).

C: مجموعه ویژگی‌ها.

D: برچسب کلاس نمونه‌ها (ویژگی تصمیم).

d: پارامتر کرنل گاوسی که میزان دانه‌بندی (Granularity) یا سطح تقریب را کنترل می‌کند.

خروجی:

$\gamma_B(D)$: وابستگی فازی برای زیرمجموعه ویژگی‌ها $B \subseteq C$.

مرحله‌ی ۱: مقداردهی اولیه

$\mathrm{POS}_B(D) = \emptyset$ (مجموعه ناحیه مثبت اولیه).

هدف این است که این مجموعه را با نمونه‌هایی پر کنیم که با قطعیت به کلاس خود تعلق دارند.

مرحله‌ی ۲: محاسبه تقریب پایین فازی

برای هر نمونه $x \in U$:

درجه اطمینان $R_G^n(x,y)$ بین نمونه x و همه نمونه‌های دیگر y با استفاده از رابطه گوسی محاسبه می‌شود. از این مقادیر، حداقل شباهت نمونه x با نمونه‌های کلاس‌های دیگر یعنی($y \notin d_i$) گرفته می‌شود:

$$\underline{R_G^n} d_i(x) = \inf_{y \notin d_i} \sqrt{1 - \left(R_G^n(x,y)\right)^2}$$

اگر مقدار بالا صفر باشد (یعنی x هیچ شباهتی به کلاس‌های دیگر ندارد)، x به ناحیه مثبت $POS_B(D)$ اضافه می‌شود.

مرحله‌ی ۳: محاسبه وابستگی فازی

در این مرحله، با استفاده از رابطه‌ی (۱۱)، میزان وابستگی فازی $\gamma_B(D)$ محاسبه می‌شود که به‌صورت نسبت اندازه‌ی ناحیه‌ی مثبت فازی $POS_B(D)$ به اندازه‌ی کل مجموعه داده |U| تعریف می‌شود:

$$\gamma_B(D) = \frac{|\mathrm{POS}_B(D)|}{|U|}$$

این معیار نشان می‌دهد که زیرمجموعه‌ی ویژگی‌های B تا چه میزان قادر به تفکیک بدون ابهام نمونه‌ها بر اساس برچسب‌های تصمیم D است. هرچه مقدار $\gamma_B(D)$ بزرگ‌تر باشد، قدرت تمایز ویژگی‌های B بیشتر بوده و در نتیجه برای فرآیند انتخاب ویژگی مناسب‌تر خواهند بود.

خروجی: به عنوان معیاری ارزیابی می‌کند که زیرمجموعه ویژگی B تا چه حد توانسته است نمونه‌ها را قطعاً طبقه‌بندی کند. این مقدار هرچه بیشتر باشد، ویژگی‌های انتخاب شده بهتر هستند.

جست‌وجوی جامع برای یافتن زیرمجموعه‌ی بهینه از میان 2^{n-1} زیرمجموعه‌ی ممکن (که (n) تعداد کل ویژگی‌هاست)، به‌دلیل پیچیدگی محاسباتی بسیار بالا، در عمل غیرقابل اجراست. در این پژوهش، الگوریتم (۲) با رویکرد جست‌وجوی حریصانه، این مشکل را برطرف کرده است و با بهره‌گیری از وابستگی فازی ($\gamma_B(D)$) محاسبه شده از الگوریتم (۱)، زیرمجموعه‌ای بهینه از ویژگی‌ها را انتخاب می‌کند [۳۸]. در ادامه، روند اجرای الگوریتم (۲) برای یافتن این زیرمجموعه‌ی بهینه تشریح شده است.

الگوریتم۲: جستجوی حریصانه پیشرو با تقریب کرنل گوسی (FS-GKA)

ورودی:
U: مجموعه کل نمونه‌ها
C: مجموعه ویژگی‌های اولیه
D: ویژگی تصمیم
d: پارامتر کرنل گاوسی
خروجی:
$B \subseteq C$: زیرمجموعه بهینه ویژگی‌ها با وابستگی بالا
۱: مقداردهی اولیه
زیرمجموعه منتخب ویژگی‌ها $B = \emptyset$
وابستگی فعلی $\gamma_B(D) = 0$
۲: حلقه‌ی جستجوی پیشرو
تا زمانی که ویژگی باقیمانده‌ای در $C \setminus B$ وجود دارد که افزایشی در وابستگی ایجاد کند:
برای هر ویژگی $a \in C \setminus B$:
افزایش وابستگی را با افزودن a به B محاسبه کن:
$$\Delta\gamma(a) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D)$$
(در این مرحله، مقادیر γ مطابق مرحله‌ی ۳ الگوریتم ۱ محاسبه می‌شود.)
ویژگی a^* را پیدا کن که بیشترین $\Delta\gamma(a)$ را دارد.
اگر $\Delta\gamma(a^*) > 0$:
$$B = B \cup \{a^*\}$$
مقدار $\gamma_B(D)$ را به‌روزرسانی کن.
در غیر این‌صورت، حلقه را خاتمه بده (افزایشی در وابستگی حاصل نمی‌شود).
۳: خروجی نهایی
زیرمجموعه‌ی ویژگی‌های منتخب B را به‌عنوان خروجی بازگردان.

الگوریتم (۲) در هر مرحله، ویژگی a را به گونه‌ای انتخاب می‌کند که با افزودن به زیرمجموعه‌ی فعلی B، بیشترین افزایش را در وابستگی فازی نسبت به کلاس‌های تصمیم ایجاد کند. وابستگی فازی نشان‌دهنده‌ی توان یک ویژگی در ارتقای دقت تفکیک کلاس‌ها است. روند جست‌وجو تا زمانی ادامه می‌یابد که افزودن ویژگی جدید، بهبود معناداری در وابستگی فازی ایجاد نکند [۳۸]. این رویکرد، علاوه بر حذف ویژگی‌های نامرتبط، ویژگی‌های تکراری را نیز کنار می‌گذارد و در نهایت زیرمجموعه‌ای بهینه برای تشخیص دقیق بدافزارهای روز صفر فراهم می‌آورد.

۴-۳- تقسیم مجموعه داده

برای شبیه سازی شرایط بدافزار روز صفر و ارزیابی توان تعمیم مدل، مجموعه داده منتخب به دو زیرمجموعه‌ی کاملاً جدا از هم تقسیم می‌شود: کلاس‌های«دیده شده»[۲۰] و «دیده نشده»[۲۱].

تعریف:

- کلاس‌های دیده شده: Seen class = $\{X_i^{tr}, Y_i^{tr}\}$
- کلاس‌های دیده نشده: Unseen class= $\{X_i^{ts}, Y_i^{ts}\}$

در این تعریف، X_i بردار ویژگی‌های هر نمونه و Y_i برچسب متـناظر آن است.

- $Y^{tr} = \{G_i^k$, xtrat, zusy, ...$\}$
- $Y^{ts} = \{G_{K+1}^M$, installmonster, vtflooder, ...$\}$

اشتراک Y^{tr} وY^{ts} کاملاً تهی است، و برچسب‌های G_i^k وG_{K+1}^M به ترتیب نماینده‌ی نمونه‌های نرم‌افزار سالم (benign) در مجموعه‌های آموزش و آزمون هستند. کلاس‌های دیده شده برای آموزش مدل‌ها و کلاس‌های دیده نشده برای ارزیابی مدل و شبیه‌سازی سناریوی روز صفر استفاده می‌شوند.

مزیت اصلی این روش نسبت به تقسیم کلاسیک داده‌ها این است که مدل روی کلاس‌هایی ارزیابی می‌شود که در مرحله‌ی آموزش، دیده نشده است. این رویکرد در حوزه‌ی «یادگیری روز صفر»[۲۲] قرار می‌گیرد [۳۹].

در روش‌های سنتی، تقسیم تصادفی داده‌ها ممکن است نمونه‌های یک کلاس بدافزار را هم‌زمان در مجموعه‌های آموزش و آزمون قرار دهد و دقت مدل را بیش از واقعیت نشان دهد. رویکرد پیشنهادی با جداسازی کلاس‌ها به گروه‌های «دیده شده» و «دیده نشده»، سناریوی واقعی‌تری از تشخیص بدافزار روز صفر فراهم می‌کند.

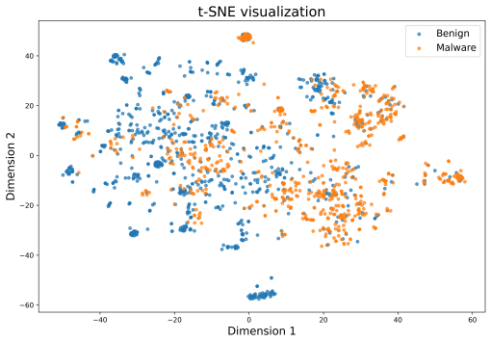
شناسایی موفق نمونه‌های کلاس‌های «دیده نشده» نشان‌دهنده‌ی توان تعمیم مدل به تهدیدات جدید است؛ امری که برای ارزیابی علمی و اعتبارسنجی عملکرد مدل ضروری است.

۴-۴- توزیع و تجسم داده‌ها با استفاده از t-SNE

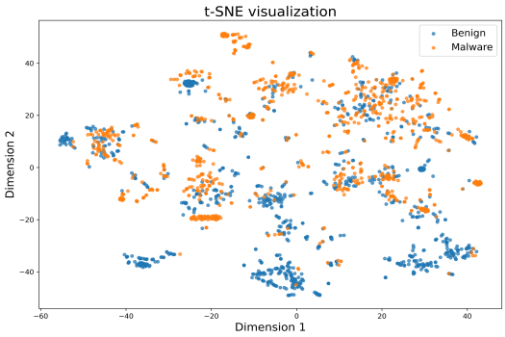
در این بخش، الگوریتم t-SNE[۲۳] برای تجسم توزیع داده‌های بدافزار و نرم‌افزارهای سالم استفاده شده است. این الگوریتم داده‌های پربُعد را به داده‌های دو یا سه‌بعدی تبدیل می‌کند و امکان تجسم خوشه‌هایی که در سطوح مختلف میان بدافزارها و نرم‌افزارهای سالم شکل می‌گیرند را فراهم می‌سازد [۳۹].

شکل (۱) داده‌های خام و شکل (۲) داده‌ها پس از استخراج ویژگی‌های بهینه با روش GKFRS را با t-SNE نمایش می‌دهد. بررسی تصاویر نشان می‌دهد که دو دسته‌ی بدافزار و نرم‌افزار سالم در برخی نقاط همپوشانی دارند؛ این موضوع نشان‌دهنده‌ی وجود شباهت‌هایی در توزیع داده‌ها است. بااین‌حال، تفاوت‌های میان بدافزار و نرم‌افزار سالم موجب تمایز چشمگیر در توزیع کلی آن‌ها شده است. به‌طورکلی، توزیع داده‌ها پس از استخراج ویژگی‌ها نشان می‌دهد که ویژگی‌های منتخب مشخصه‌های اصلی داده‌ها را به‌خوبی حفظ کرده است و رفتار توزیع آن‌ها مشابه با داده‌های اولیه

است.



شکل ۱: تجسم داده‌ها با ویژگی‌های خام.



شکل۲: تجسم داده‌ها با ویژگی‌های بهینه.

۴-۵- طبقه‌بندی با استفاده از یادگیری تلفیقی استکینگ

پس از انتخاب زیرمجموعه‌ی بهینه‌ی ویژگی‌ها با چارچوب GKFRS، این ویژگی‌ها به‌عنوان ورودی مدل استکینگ استفاده می‌شوند. استکینگ یک رویکرد یادگیری تلفیقی است که با ترکیب پیش‌بینی‌های چندین مدل پایه‌ی متنوع، پیش‌بینی نهایی دقیق‌تری ارائه می‌دهد. جزئیات این فرایند در الگوریتم (۳) شرح داده شده است.

آموزش مدل استکینگ روی داده‌های آموزشی با ویژگی‌های منتخب و با استفاده از اعتبارسنجی متقابل پنج‌لایه (۵-CV fold)[۲۴] انجام می‌شود. در هر لایه، ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد برای اعتبارسنجی مدل‌های پایه به‌کار گرفته شد. برای کاهش خطر بیش برازش[۲۵] و افزایش دقت، پیش‌بینی‌های مدل‌های پایه تنها از داده‌های خارج از هر لایه[۲۶] جمع‌آوری و برای آموزش مدل فرایادگیرنده[۲۷] استفاده شد. سپس برای ارزیابی نهایی روی مجموعه‌ی دیده نشده، مدل‌های پایه دوباره با کل داده‌های مجموعه دیده شده آموزش داده شد و با مدل متا، پیش‌بینی‌ها انجام گرفت. این روش ارزیابی واقع‌گرایانه‌تری از عملکرد مدل‌های پایه روی داده‌های ناشناخته فراهم می‌کند و از انتقال خطاهای ناشی از بیش‌برازش جلوگیری می‌کند. مجموعه دیده شده از نظر تعداد نمونه تقریباً متعادل است (حدود ۴۹ درصد سالم و ۵۱ درصد بدافزار)، مجموعه داده دیده نشده نامتوازن‌تر بود (حدود ۶۰ درصد سالم و ۴۰ درصد بدافزار) و هر دو مجموعه در سطح کلاس‌ها دچار نامتوازنی بودند. برای کنترل این وضعیت، در طول آموزش از وزن‌دهی کلاس‌ها استفاده شد و هیچ‌گونه نمونه‌برداری افزایشی [۲۸] یا نمونه‌برداری کاهشی[۲۹] انجام نشد تا توزیع واقعی داده‌های عملیاتی حفظ شود.

الگوریتم۳: یادگیری تلفیقی استکینگ.

۱- ورودی:
- داده‌های آموزش (X_seen,y_seen)
- داده‌های آزمون (X_unseen ,y_unseen)
- تعداد folds برای اعتبارسنجی (k=5)
- مدل‌های پایه: LightGBM و XGBoost
- مدل متا: لجستیک رگرسیون[۳۰]

۲-پیاده‌سازی k-fold Cross Validation با k=5 بر روی داده‌های آموزش:

۱. داده‌های آموزش را به ۵ بخش تقسیم کن.

۲. برای هر فولد (i=1 تا k=5):
- چهار بخش را به‌عنوان داده آموزش مدل‌های پایه (X_base_train, y_base_train)، و یک بخش را به عنوان داده اعتبارسنجی مدل متا (X_valid,y_valid) انتخاب کن.
- مدل‌های پایه را روی داده آموزش، آموزش بده.
- روی مجموعه اعتبارسنجی، پیش‌بینی مدل‌های پایه را محاسبه کن.
- خروجی مدل‌های پایه را در اعتبارسنجی، برای آموزش مدل متا ذخیره کن.

۳. پس از ۵ تکرار، با ترکیب پیش‌بینی‌های مدل‌های پایه روی مجموعه اعتبارسنجی ، داده نهایی را برای آموزش مدل متا بساز.

۳-آموزش مدل متا (لجستیک رگرسیون):
از داده حاصل از مرحله قبلی (خروجی مدل‌های پایه روی validation) برای آموزش مدل لجستیک استفاده کن.

۴-آزمون الگوریتم روی داده‌های آزمون (X_unseen):
- مدل‌های پایه را روی تمام داده‌های آموزش، بازآموزش بده.
- با مدل‌های پایه روی داده‌های آزمون پیش‌بینی کن.
- خروجی مدل‌های پایه را به مدل متا بده، و پیش‌بینی نهایی را تولید کن.

۵-گزارش عملکرد: معیارهایی مانند دقت، فراخوانی و امتیاز F1 را محاسبه و گزارش کن.

۴-۶- پیچیدگی محاسباتی و بهینه‌سازی پیاده‌سازی

در حالت پایه، اجرای چارچوب (GKFRS) برای مجموعه داده‌های بزرگ مانند EMBER، به‌دلیل وابستگی تعاریف تقریب پایین به تمامی اشیای جهان U، نیازمند محاسبه و نگهداری ماتریس شباهت n×n برای هر ویژگی است. این ساختار باعث پیچیدگی زمانی $O(m \cdot n^2)$ و پیچیدگی فضایی $O(n^2)$ می‌شود. در داده‌های این پژوهش، این امر به معنای صدها میلیارد محاسبه و نیاز به بیش از یک ترابایت حافظه است که اجرای مستقیم الگوریتم را غیرعملی

می‌سازد. برای رفع این محدودیت و اجرای عملی روی سخت‌افزارهای متعارف، چهار راهکار کلیدی به‌کار گرفته شد:

۱. محاسبه بلوکی[۳۱]: داده‌ها به بلوک‌های کوچک‌تر تقسیم شدند تا در هر لحظه تنها یک بخش از ماتریس شباهت در RAM بارگذاری شود؛ این روش مصرف حافظه را از ۴۲/۶ گیگابایت به حدود ۱۱/۲ گیگابایت کاهش داد.

۲. ذخیره‌سازی حافظه‌نگاشتی[۳۲]: با استفاده از کتابخانه‌ی numpy.memmap، نتایج موقت بر روی دیسک ذخیره شد و فقط بخش‌های مورد نیاز به RAM خوانده شدند.

۳. شتاب‌دهی GPU[۳۳] با کتابخانه‌ی CuPy روی Tesla T4: انتقال محاسبات کرنل گوسی به GPU باعث کاهش سه برابری زمان پردازش شد (به‌طور مثال پردازش یک بلوک از ۴۵ دقیقه به ۱۲ دقیقه رسید).

۴. محدودسازی همسایه‌ها به k= ۵۰: تنها ۵۰ همسایه‌ی نزدیک هر نمونه در محاسبات استفاده شدند؛ پیچیدگی از $O(n^2)$ به $O(n \cdot k)$ کاهش یافت و حافظه مصرفی کمتر شد. در نتیجه این بهینه‌سازی زمان اجرای FS-GKA از ۱۲ ساعت و ۱۷ دقیقه و ۴۲/۶ گیگابایت مصرف RAM، به حدود ۳ ساعت و ۲۸ دقیقه و ۱۱/۲ گیگابایت مصرف RAM کاهش یافت. پس از کاهش ابعاد به ۱۰۰ ویژگی با FS-GKA، چارچوب تلفیقی استکینگ در ۱۲۲۴/۶۰ ثانیه (۲۰ دقیقه) آموزش دید و ارزیابی آن تنها ۱۵/۳۵ ثانیه طول کشید.

۵- نتایج تجربی

این بخش به بررسی نتایج آزمایش‌های انجام‌شده برای ارزیابی کارایی چارچوب پیشنهادی در تشخیص بدافزارهای روز صفر می‌پردازد. آزمایش‌ها روی زیرمجموعه‌ای از مجموعه داده EMBER انجام شده است.

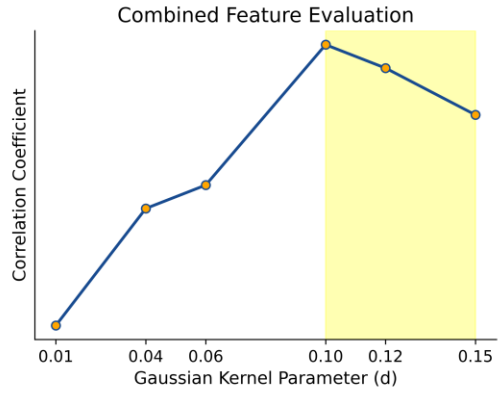
۵-۱- مجموعه داده

مجموعه داده EMBER (2018) شامل ۹۰۰،۰۰۰ نمونه آموزشی و ۲۰۰،۰۰۰ نمونه آزمایشی با ۲۳۸۱ ویژگی و ۲۷۳۱ خانواده بدافزار است. این ویژگی‌ها از تحلیل ایستای فایل‌های PE استخراج شده‌اند و شامل داده‌های عددی و دسته‌ای هستند [۴۰].

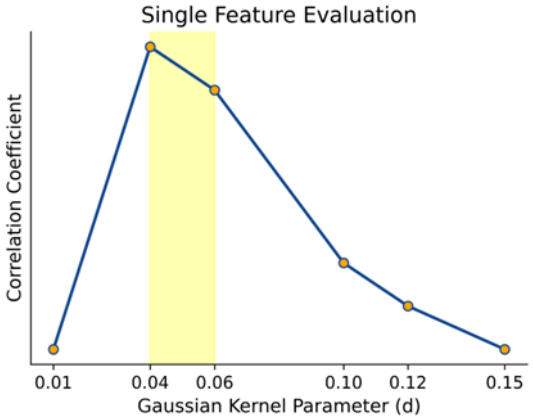
۵-۲- تنظیمات تجربی

۵-۲-۱- انتخاب پارامتر کرنل گاوسی (d)

پژوهش‌های پیشین نشان می‌دهند که پارامتر کرنل گاوسی (d) تأثیر بسزایی بر ساختار رابطه خشن فازی دارد و مقدار بهینه آن به‌صورت تجربی و متناسب با داده‌ها تعیین می‌شود [۳۸]. در این پژوهش، بازه‌های مختلفی برای (d) روی مجموعه داده EMBER بررسی شدند و مقدار بهینه با بیشینه‌سازی ضریب همبستگی میان وابستگی خشن فازی و دقت طبقه‌بندی انتخاب شد. نتایج نشان داد بازه [۰/۰۴،۰/۰۶] برای تک ویژگی‌ها و بازه [۰/۱،۰/۱۵] برای زیرمجموعه ویژگی‌ها مناسب است. مقدار نهایی d براساس بیشینه همبستگی در نواحی مشخص‌شده شکل‌های (۳) و (۴) تعیین شد.



شکل۳: تغییرات ضرایب همبستگی در مقابل پارامتر کرنل برای زیرمجموعه ویژگی‌ها.



شکل۴: تغییرات ضرایب همبستگی در مقابل پارامتر کرنل برای تک‌ویژگی‌ها.

۵-۲-۲- انتخاب ویژگی

الگوریتم GKFRS روی مجموعه داده اجرا شد تا از میان ویژگی‌ها، ۱۰۰ ویژگی بهینه برای طبقه‌بندی استخراج شود. هدف این فرایند، کاهش ابعاد داده‌ها، حذف نویز و افزایش تمایزپذیری بین کلاس‌ها بود.

۵-۲-۳- یادگیری تلفیقی استکینگ

برای ترکیب توانایی‌های مدل‌های پایه و ارتقای عملکرد نهایی، از یادگیری تلفیقی استکینگ با مدل‌های XGBoost و LightGBM همراه با اعتبارسنجی متقابل پنج‌لایه استفاده شد. مدل لجستیک رگرسیون نیز به‌عنوان فرایادگیرنده به‌کار رفت. پیکربندی مدل‌های پایه در جدول (۱) ارائه شده است.

جدول۱: پیکربندی مدل‌های پایه.

مدل	پارامترها
XGBoost	learning_rate = 0.2, max_depth = 8 subsample = 0.8, n_estimators = 2000 random_state = 42
LightGBM	n_estimators = 2000, max_depth = 8 learning_rate = 0.1, num_leaves = 31 min_child_samples = 20, subsample = 0.8 colsample_bytree = 0.8, reg_alpha = 0.1 reg_lambda = 0.1, n_jobs = -1, verbose = 1 random_state = 42

۵-۲-۴- روش ارزیابی

برای شبیه‌سازی دقیق سناریوی روز صفر، داده‌ها به‌گونه‌ای تقسیم شدند که کلاس‌های بدافزاری موجود در مجموعه آزمون، در مجموعه آموزش حضور نداشته باشند. این تقسیم‌بندی توانایی تعمیم مدل به تهدیدات ناشناخته را نشان می‌دهد. جزئیات توزیع کلاس‌ها و نمونه‌ها در جدول‌های (۲-۴) ارائه شده است.

جدول۲: کل مجموعه داده.

تعداد کل نمونه‌ها	تعداد کلاس های دیده نشده	تعداد کل کلاس های بدافزار
۵۸۹۲۴۶	۵۶۴	۲۷۳۱

جدول ۳: مجموعه دیده شده.

تعداد کل نمونه‌ها	تعداد نمونه های سالم	تعداد بدافزارها
۴۸۹۴۷۸	۲۴۰۰۰۰	۲۴۹۴۷۸

جدول۴: مجموعه دیده نشده.

تعداد کل نمونه‌ها	تعداد نمونه های سالم	تعداد بدافزارها
۹۹۷۶۸	۶۰۰۰۰	۳۹۷۶۸

۵-۲-۵- معیارهای ارزیابی

برای ارزیابی عملکرد مدل، از معیارهای استاندارد طبقه‌بندی شامل دقت، فراخوانی و امتیاز F1 استفاده شده است. این معیارها توانایی مدل در شناسایی صحیح بدافزارها و ایجاد توازن میان نمونه‌های سالمی که به اشتباه به‌عنوان بدافزار شناسایی می‌شوند (خطای مثبت کاذب) و بدافزارهایی که شناسایی نشده است و از دید مدل پنهان می‌مانند (خطای منفی کاذب) را به‌خوبی نشان می‌دهند. این معیارها بر اساس روابط (۲۳-۲۵) محاسبه می‌شوند:

$$f_{1\,score} = \frac{2 \times precision \times recall}{precision + recall} \quad (۲۳)$$

$$recall = \frac{tp}{tp + fn} \quad (۲۴)$$

$$precision = \frac{tp}{tp + fp} \quad (۲۵)$$

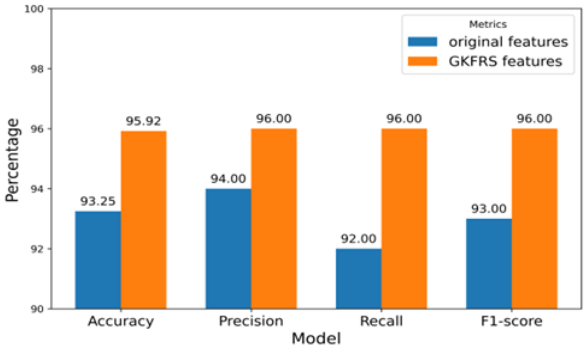
۵-۳- نتایج ارزیابی

مدل تلفیقی استکینگ مبتنی بر GKFRS در تشخیص بدافزارهای روز صفر عملکرد بسیار مطلوبی داشته است و به دقت ۹۶ درصد رسیده است. این مدل در مقایسه با سایر الگوریتم‌های پایه و پیشرفته، برتری نسبی خود را حفظ کرده است و با ایجاد تعادل مناسب بین دقت و فراخوانی، نتایج مؤثری به دست می‌آورد. نتایج حاصل در جدول (۵) ارائه شده است.

جدول۵: مقایسه با مدل‌های دیگر یادگیری ماشین.

Model	precision	recall	f1-score
Stacking Ensemble (KFRS + 5-fold)	**۹۶**	**۹۶**	**۹۶**
LightGBM	۹۵	۹۶	۹۶
XGBoost	۹۵	۹۶	۹۶
Extra tree	۹۳	۹۳	۹۲
RandomForest	۸۵/۸۹	۸۵/۸۸	۸۵/۸۹
KNN	۷۲/۷۹	۴۱/۷۹	۵۵/۷۹
AdaBoost	۸۵/۳۲	۸۵/۶۱	۸۵/۴۵
CatBoost	۹۴	۹۳	۹۳
BiLSTM	۹۳/۴۳	۰/۸۷	۰/۹۰
GNN	۹۲/۶۳	۹۳/۲۸	۹۲/۹۴
CNN-1D	۹۴/۷۴	۰/۸۲	۰/۸۸
MLP-Dense	۰/۹۳	۰/۹۲	۰/۹۳

نتایج تجربی نشان داد که به‌کارگیری روش GKFRS با کاهش هدفمند ابعاد داده‌ها، ویژگی‌های زائد و غیرمرتبط را حذف می‌کند و بار محاسباتی الگوریتم‌های پایه و مدل متا در چارچوب استکینگ را به میزان چشمگیری کاهش می‌دهد. این کاهش هزینه محاسباتی، ضمن کاهش بیش‌برازش، دقت مدل را به ۹۶ درصد رساند و یک تعادل بهینه میان پیچیدگی و عملکرد ایجاد کرد. بررسی شکل (۵) نشان می‌دهد که استفاده از روش انتخاب ویژگی GKFRS به بهبود قابل‌توجه عملکرد مدل تلفیقی استکینگ منجر شده است.



شکل۵: مقایسه بین ویژگی‌های اصلی و ویژگی‌های GKFRS.

جدول (۶) عملکرد مدل پیشنهادی را در مقایسه با سایر مطالعات مبتنی بر مجموعه داده EMBER ارائه می‌دهد. همان‌گونه که مشاهده می‌شود، بخش قابل توجهی از مطالعات پیشین فاقد مکانیزم مشخص انتخاب ویژگی بوده‌اند و کل ویژگی‌های خام مستقیماً به مدل‌های طبقه‌بندی داده شده است [۴۱،۴۰]. این امر سبب افزایش سربار محاسباتی و باقی ماندن ویژگی‌های زائد و نویزی در داده‌ها و موجب افت عملکرد تشخیص شده است. در برخی دیگر از پژوهش‌ها از الگوریتم‌هایی نظیر Gradient boostingیا Random Forest برای انتخاب ویژگی استفاده شده است [۴۳،۴۲]. اگرچه این روش‌ها تا حدی اهمیت نسبی ویژگی‌ها را تعیین می‌کنند اما به دلیل ناتوانی در مدل‌سازی روابط غیرخطی و مدیریت عدم قطعیت در داده‌های ترکیبی، بخشی از اطلاعات مؤثر از دست رفته و عملکرد نهایی محدود مانده است. مطالعه [۴۴] با بهره‌گیری از AutoGluon و یادگیری تلفیقی موازی تلاش کرده است تا فرآیند انتخاب ویژگی و ترکیب مدل‌ها را بهینه کند اما تمرکز اصلی آن بر مدل‌سازی خودکار بوده است و مسئله‌ی ناهمگنی داده‌های عددی و دسته‌ای همچنان به‌طور کامل پوشش داده نشده است. در مقابل، روش پیشنهادی با ترکیب مجموعه‌های خشن فازی همراه با کرنل گوسی توانسته است روابط غیرخطی را به‌طور همزمان مدل‌سازی کند، عدم‌قطعیت موجود در داده‌های ناهمگون را در فرآیند انتخاب ویژگی لحاظ کند و در نهایت مجموعه‌ای بهینه و فشرده از ویژگی‌های مؤثر استخراج نماید.

جدول۶: مقایسه با دیگرمدل‌های تشخیص بدافزار.

منابع	مجموعه داده	انتخاب ویژگی	مدل طبقه بند	precision	recall	f1-score	Zero_day detection
S. S. Lad and A. C. Adamuthe (2022) [40]	EMBER 2018	—	شبکه عصبی عمیق	۹۰/۱۴	۸۸/۶۶	۸۸/۸۵	خیر
Nadia M. Ali, Munera A. Jabaar (2025) [41]	EMBER 2018	—	CNN-LSTM	۹۶/۹	۹۷/۱	۰/۹۷	خیر
A. H. K, S. N S, and M. K. S D (2024) [42]	EMBER 2018	Gradient Boosting	مدل تلفیقی	۹۷	۹۷	۹۷	خیر
M. Dener and S. Gulburun (2023) [43]	EMBER 2018	Random Forest	مدل تلفیقی	—	—	۹۶/۷۷	خیر
Hoang V. Vo, Phong H. Nguyen (2023) [44]	EMBER 2018	AutoGluon	مدل تلفیقی موازی	۹۷/۸۰	۹۷/۴۹	۹۷/۶۴	خیر
مدل پیشنهادی (تقسیم تصادفی)	EMBER 2018	KFRS	مدل تلفیقی	۹۸	۹۸	۹۸	خیر
مدل پیشنهادی (روز صفر)	EMBER 2018	KFRS	مدل تلفیقی	۹۶	۹۶	۹۶	بله

ما در این ارزیابی، یک سناریوی روز صفر طراحی و آزمایش کردیم که در آن خانواده‌های بدافزار دیده نشده در مرحله آموزش حذف و فقط در آزمون استفاده شدند. این طراحی باعث شد مدل پیشنهادی در شرایطی شبیه به دنیای واقعی که بدافزارهای ناشناخته ظاهر می‌شوند، مورد آزمایش قرار گیرد. نتایج نشان داد که مدل پیشنهادی ما تنها مدلی بود که در این سناریوی واقع‌گرایانه روز صفر آزمایش شد و همچنان عملکرد مطلوبی ارائه داد. به منظور ارزیابی عملکرد مدل‌ها در شرایطی که داده‌های آموزش و آزمون از توزیع یکسانی استخراج شده‌اند، سناریوی «تقسیم تصادفی» اجرا شد. در این سناریو، داده‌ها با نسبت ۸۰ درصد برای آموزش و ۲۰ درصد برای آزمون، به‌صورت تصادفی تقسیم شدند. از آنجا که ترکیب کلاس‌ها در دو مجموعه‌ی آموزش و آزمون مشابه است، این نتایج بیانگر توانایی مدل‌ها در تعمیم‌دهی به داده‌های مشابه و حداکثر سطح عملکرد قابل دستیابی در شرایط بدون محدودیت روز صفر هستند. در جدول (۶) نتایج این سناریو نیز آمده است.

۶- نتیجه‌گیری وکارهای آینده

تشخیص بدافزارهای روز صفر به‌دلیل ویژگی‌های ناشناخته و ماهیت پویا، یکی از چالش‌های اساسی امنیت سایبری است. این پژوهش چارچوبی نوآورانه مبتنی بر ترکیب انتخاب ویژگی با مجموعه‌های خشن فازی مبتنی بر کرنل گوسی و مدل یادگیری تلفیقی استکینگ ارائه داد. هدف این چارچوب، بهبود دقت و تعمیم‌پذیری مدل در شناسایی تهدیدات جدید است.

برای ارزیابی چارچوب پیشنهادی، از زیرمجموعه‌ی گسترده از داده‌های EMBER در سناریوی واقع‌گرایانه تشخیص روز صفر استفاده شد. در این سناریو، کلاس‌های بدافزاری مجموعه آزمون از داده‌های آموزشی حذف شدند تا تعمیم‌پذیری مدل سنجیده شود. روش GKFRS با کاهش ابعاد داده‌ها و حفظ اطلاعات کلیدی، تمایز بدافزارهای روز صفر را بهبود بخشید. نتایج نشان داد که مدل استکینگ مبتنی بر ویژگی‌های منتخب GKFRS با دقت ۹۶ درصد و نرخ فراخوانی بالا، عملکردی برجسته در تشخیص بدافزارهای روز صفر ارائه داد. این دستاورد برتری رویکرد پیشنهادی نسبت به مدل‌های متداول و نقش کلیدی GKFRS در ارتقای تعمیم‌پذیری را تأیید می‌کند. این چارچوب می‌تواند راهکاری مؤثر برای بهبود سیستم‌های تشخیص بدافزار در برابر تهدیدات نوظهور باشد. برای توسعه این پژوهش، پیشنهاد می‌شود چارچوب پیشنهادی روی مجموعه داده‌های متنوع و حجیم اعتبارسنجی شود. بررسی پیکربندی‌های مختلف استکینگ و تحلیل دقیق‌تر پارامترهای GKFRS می‌تواند عملکرد مدل را بهبود بخشد. همچنین، مطالعه یادگیری افزایشی و استفاده از بازخورد عملکرد مدل برای تطبیق با تهدیدات جدید از محورهای پیشنهادی برای تحقیقات آینده است. اگرچه در این مطالعه از مجموعه داده استاندارد EMBER استفاده شده است اما باید توجه داشت که داده‌های استاندارد تمام جنبه‌های تهدیدات عملی و واقعی روز صفر را پوشش نمی‌دهند. دلیل استفاده از EMBER، اندازه بزرگ و قابلیت مقایسه با مطالعات پیشین بوده است. برای کاهش این محدودیت، ما علاوه بر تقسیم‌بندی تصادفی، یک سناریوی واقعی روز صفر طراحی کرده‌ایم که در آن خانواده‌های ناشناخته در مرحله آزمون استفاده شدند. نتایج نشان داد که مدل پیشنهادی حتی در این شرایط واقع‌گرایانه نیز عملکرد مطلوبی دارد. با این حال، در کارهای آینده قصد داریم روش پیشنهادی را روی داده‌های واقعی جمع‌آوری شده از محیط‌های عملیاتی و تهدیدات در حال ظهور آزمایش کنیم و همچنین تاب‌آوری آن را در برابر حملات خصمانه مورد ارزیابی قرار دهیم.

مراجع

[1] M. S. Raza and U. Qamar, “An incremental dependency calculation technique for feature selection using rough sets,” Inf. Sci., vol. 343–344, pp. 41–65, May 2016, doi: 10.1016/j.ins.2016.01.044.

[2] J. Zhao, J. Liang, Z. Dong, D. Tang, and Z. Liu, “NEC: A nested equivalence class-based dependency calculation approach for fast feature selection using rough set theory,” Inf. Sci., vol. 536, pp. 431–453, Oct. 2020, doi: 10.1016/j.ins.2020.03.092.

[3] J. Zhang, Z. Luo, C. Li, C. Zhou, and S. Li, “Manifold regularized discriminative feature selection for multi-label learning,” Pattern Recognit., vol. 95, pp. 136–150, Nov. 2019, doi: 10.1016/j.patcog.2019.06.003.

[4] R. Huang and Z. Wu, “Multi-label feature selection via manifold regularization and dependence maximization,” Pattern Recognit., vol. 120, p. 108149, Dec. 2021, doi: 10.1016/j.patcog.2021.108149.

[5] L. Dong, R. Wang, and D. Chen, “Robust Feature Selection with Fuzzy Rough Set,” Int. J. Fuzzy Syst., July 2025, doi: 10.1007/s40815-025-02053-6.

[6] C. Wang, Y. Zhang, S. An, and T. Deng, “Adaptive feature selection based on fuzzy rough set fusion model with class variance,” Pattern Recognit., vol. 170, p. 112014, Feb. 2026, doi: 10.1016/j.patcog.2025.112014.

[7] M. Ebrahimifard, M. Abadi, and F. Deldar, “RansomHunter: zero-day Android ransomware detection using graph convolutional neural networks and code embeddings,” J. Supercomput., vol. 81, no. 15, p. 1351, Sept. 2025, doi: 10.1007/s11227-025-07777-8.

[8] M. T. Akdeniz, Z. Yeşilkaya, İ. E. Köse, İ. U. Ünal, and S. Şen, “VOLTRON: Detecting Unknown Malware Using Graph-Based Zero-Shot Learning,” 2025, arXiv. doi: 10.48550/ARXIV.2507.04275.

[9] A. Jain, R. Bagoria, and P. Arora, “An intelligent zero-day attack detection system using unsupervised machine learning for enhancing cyber security,” Knowl.-Based Syst., vol. 324, p. 113833, Aug. 2025, doi: 10.1016/j.knosys.2025.113833.

[10] A. Karakaya, “A Hybrid Approach for IoT Security: Combining Ensemble Learning with Fuzzy Logic,” Sensors, vol. 25, no. 18, p. 5668, Sept. 2025, doi: 10.3390/s25185668.

[11] Y. Xia, K. Chen, and Y. Yang, “Multi-label classification with weighted classifier selection and stacked ensemble,” Inf. Sci., vol. 557, pp. 421–442, May 2021, doi: 10.1016/j.ins.2020.06.017.

[12] I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” in Machine Learning: ECML-94, vol. 784, F. Bergadano and L. Raedt, Eds., in Lecture Notes in Computer Science, vol. 784. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 171–182. doi: 10.1007/3-540-57868-4_57.

[13] M. Moradi and J. Hamidzadeh, “Ensemble-based Top-k Recommender System Considering Incomplete Data,” J. AI Data Min., vol. 7, no. 3, July 2019, doi: 10.22044/jadm.2019.7026.1830.

[14] N. X. Vinh, S. Zhou, J. Chan, and J. Bailey, “Can high-order dependencies improve mutual information based feature selection?,” Pattern Recognit., vol. 53, pp. 46–58, May 2016, doi: 10.1016/j.patcog.2015.11.007.

[15] Y. Li, T. Li, and H. Liu, “Recent advances in feature selection and its applications,” Knowl. Inf. Syst., vol. 53, no. 3, pp. 551–577, Dec. 2017, doi: 10.1007/s10115-017-1059-8.

[16] M.-L. Zhang and Z.-H. Zhou, “A Review on Multi-Label Learning Algorithms,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, Aug. 2014, doi: 10.1109/TKDE.2013.39.

[17] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Boston, MA: Springer US, 1998. doi: 10.1007/978-1-4615-5689-3.

[18] Y. Fan, J. Liu, W. Weng, B. Chen, Y. Chen, and S. Wu, “Multi-label feature selection with constraint regression and adaptive spectral graph,” Knowl.-Based Syst., vol. 212, p. 106621, Jan. 2021, doi: 10.1016/j.knosys.2020.106621.

[19] W. Qian, J. Huang, F. Xu, W. Shu, and W. Ding, “A survey on multi-label feature selection from perspectives of label fusion,” Inf. Fusion, vol. 100, p. 101948, Dec. 2023, doi: 10.1016/j.inffus.2023.101948.

[20] N. Spolaor, E. A. Cherman, M. C. Monard, and H. D. Lee, “ReliefF for Multi-label Feature Selection,” in 2013 Brazilian Conference on Intelligent Systems, Fortaleza, Brazil: IEEE, Oct. 2013, pp. 6–11. doi: 10.1109/BRACIS.2013.10.

[21] A. M. Alnajim, S. Habib, M. Islam, R. Albelaihi, and A. Alabdulatif, “Mitigating the Risks of Malware Attacks with Deep Learning Techniques,” Electronics, vol. 12, no. 14, p. 3166, July 2023, doi: 10.3390/electronics12143166.

[22] M. Ahmed, N. Afreen, M. Ahmed, M. Sameer, and J. Ahamed, “An inception V3 approach for malware classification using machine learning and transfer learning,” Int. J. Intell. Netw., vol. 4, pp. 11–18, 2023, doi: 10.1016/j.ijin.2022.11.005.

[23] A. Ksibi, M. Zakariah, L. Almuqren, and A. S. Alluhaidan, “Efficient android malware identification with limited training data utilizing multiple convolution neural network techniques,” Eng. Appl. Artif. Intell., vol. 127, p. 107390, Jan. 2024, doi: 10.1016/j.engappai.2023.107390.

[24] H. Deng, C. Guo, G. Shen, Y. Cui, and Y. Ping, “MCTVD: A malware classification method based on three-channel visualization and deep learning,” Comput. Secur., vol. 126, p. 103084, Mar. 2023, doi: 10.1016/j.cose.2022.103084.

[25] S. Ramesh, S. Selvarayan, V. M. Raghav, and C. Arumugam, “A Multistage Malware Detection and Classification Model Using Visual Features,” in 2023 4th International Conference on Communication, Computing and Industry 6.0 (C216), Bangalore, India: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/C2I659362.2023.10431264.

[26] K. Shaukat, S. Luo, and V. Varadharajan, “A novel deep learning-based approach for malware detection,” Eng. Appl. Artif. Intell., vol. 122, p. 106030, June 2023, doi: 10.1016/j.engappai.2023.106030.

[27] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, “IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture,” Comput. Netw., vol. 171, p. 107138, Apr. 2020, doi: 10.1016/j.comnet.2020.107138.

[28] J. Chen, C. Yuan, J. Li, D. Tian, R. Ma, and X. Jia, “ELAMD: An ensemble learning framework for adversarial malware defense,” J. Inf. Secur. Appl., vol. 75, p. 103508, June 2023, doi: 10.1016/j.jisa.2023.103508.

[29] Y. Gao, H. Hasegawa, Y. Yamaguchi, and H. Shimada, “Malware Detection by Control-Flow Graph Level Representation Learning With Graph Isomorphism Network,” IEEE Access, vol. 10, pp. 111830–111841, 2022, doi: 10.1109/ACCESS.2022.3215267.

[30] C. Joshi, R. K. Ranjan, and V. Bharti, “A Fuzzy Logic based feature engineering approach for Botnet detection using ANN,” J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 9, pp. 6872–6882, Oct. 2022, doi: 10.1016/j.jksuci.2021.06.018.

[31] X. Che, D. Chen, J. Deng, and J. Mi, “Exploiting local label correlation from sample perspective for multi-label classification via three-way decision theory,” Appl. Soft Comput., vol. 149, p. 110950, Dec. 2023, doi: 10.1016/j.asoc.2023.110950.

[32] A. S. Kumar and R. Rekha, “A Dense Network Approach with Gaussian Optimizer for Cardiovascular Disease Prediction,” New Gener. Comput., vol. 41, no. 4, pp. 859–878, Nov. 2023, doi: 10.1007/s00354-023-00234-1.

[33] L. K. Singh, M. Khanna, H. Monga, R. Singh, and G. Pandey, “Nature-Inspired Algorithms-Based Optimal Features Selection Strategy for COVID-19 Detection Using Medical Images,” New Gener. Comput., vol. 42, no. 4, pp. 761–824, Nov. 2024, doi: 10.1007/s00354-024-00255-4.

[34] M. A. Zarif and J. Hamidzadeh, “Improving performance of multi-label classification using ensemble of feature selection and outlier detection,” in 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, Islamic Republic of: IEEE, Nov. 2022, pp. 073–079. doi: 10.1109/ICCKE57176.2022.9960003.

[35] Y. Yang et al., “Rational linear kernelized weighted fuzzy rough attribute selection with class separability,” Fuzzy Sets Syst., vol. 521, p. 109600, Dec. 2025, doi: 10.1016/j.fss.2025.109600.

[36] C. Wang, Y. Zhang, S. An, and W. Ding, “An Adaptive Fuzzy Rough Neural Network and Its Application in Classification,” IEEE Trans. Syst. Man Cybern. Syst., vol. 55, no. 10, pp. 7579–7590, Oct. 2025, doi: 10.1109/TSMC.2025.3599882.

[37] J. Liu, M. Wu, M. Lin, and Z. Xu, “Anisotropic-Gaussian-Kernel-Based Fuzzy Clustering Algorithm for Feature Selection,” IEEE Trans. Fuzzy Syst., vol. 33, no. 9, pp. 3061–3075, Sept. 2025, doi: 10.1109/TFUZZ.2025.3581918.

[38] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, “Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications,” Int. J. Approx. Reason., vol. 51, no. 4, pp. 453–471, Mar. 2010, doi: 10.1016/j.ijar.2010.01.004.

[39] M. Cen, X. Deng, F. Jiang, and R. Doss, “Zero-Ran Sniff: A zero-day ransomware early detection method based on zero-shot learning,” Comput. Secur., vol. 142, p. 103849, July 2024, doi: 10.1016/j.cose.2024.103849.

[40] Dept. of CSE, Rajarambapu Institute of Technology, Rajaramnagar, Sangli, Maharashtra, India, S. S. Lad., and A. C. Adamuthe, “Improved Deep Learning Model for Static PE Files Malware Detection and Classification,” Int. J. Comput. Netw. Inf. Secur., vol. 14, no. 2, pp. 14–26, Apr. 2022, doi: 10.5815/ijcnis.2022.02.02.

[41] N. M. Ali, M. A. Jabaar, and A. M. Taha, “A Dual-Branch CNN-LSTM Residual Network for Enhanced Windows Malware Detection,” J. Port Sci. Res., vol. 8, no. 4, July 2025, doi: 10.36371/port.2025.4.6.

[42] A. H. K, S. N S, and M. K. S D, “Enhancing Malware Detection: A Comparative Analysis of Ensemble Learning Approaches,” in 2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP), Bali, Indonesia: IEEE, June 2024, pp. 35–40. doi: 10.1109/TIACOMP64125.2024.00016.

[43] M. Dener and S. Gulburun, “Clustering-Aided Supervised Malware Detection with Specialized Classifiers and Early Consensus,”

Comput. Mater. Contin., vol. 75, no. 1, pp. 1235–1251, 2023, doi: 10.32604/cmc.2023.036357.

[44] H. V. Vo, P. H. Nguyen, H. T. Nguyen, D. B. Vu, and H. N. Nguyen, "Enhancing AI - Powered Malware Detection by Parallel Ensemble Learning," in 2023 RIVF International Conference on Computing and Communication Technologies (RIVF), Hanoi, Vietnam: IEEE, Dec. 2023, pp. 503–508. doi: 10.1109/RIVF60135.2023.10471855.

پاورقی‌ها:

[1] zero-day malwares
[2] Kernel-Fuzzy-Rough Sets
[3] Ensemble Learning
[4] Wrapper
[5] Embedded
[6] Rough Sets
[7] Fuzzy Sets
[8] Indiscernibility
[9] High-Dimensional
[10] Sparse Fuzzy
[11] bagging
[12] boosting
[13] Stacking Ensemble learning
[14] positive semi-definite
[15] One-Hot Encoding
[16] StandardScaler
[17] Dependency with Gaussian Kernel Approximation
[18] Forward Greedy Search – Gaussian Kernel Approximation
[19] Fuzzy granules
[20] Seen class
[21] Unseen class
[22] Zero-Shot Learning
[23] t-Distributed Stochastic Neighbor Embedding
[24] Cross-Validation
[25] Overfitting
[26] out-of-fold
[27] Meta-Learner
[28] Oversampling
[29] Undersampling
[30] Logistic Regression
[31] Block-wise computation
[32] Memory-mapped storage
[33] GPU acceleration