مطالعات علوم کاربردی در مهندسی
دوره ۱۰، شماره ۴، زمستان ۱۴۰۳
صفحات ۷۳-۵۱

Online ISSN: ۲۴۷۶-۴۵۰۷
Print ISSN: ۲۵۳۸-۲۰۵۵
www.irijournals.com

تشخیص صفحات فیشینگ با استفاده از الگوریتم های داده کاوی و یادگیری ماشین

مریم حاجی اسمعیلی[۱]، حمید مظاهری طامه[۲]

[۱]استادیار دانشگاه آزاد اسلامی واحد تهران مرکزی، تهران، ایران
[۲]کارشناسی ارشد مهندسی فناوری اطلاعات گرایش تجارت الکترونیک دانشگاه آزاد اسلامی

چکیده

فیشینگ یک نوع حمله اینترنتی در سطح وب می‌باشد که هدف اساسی آن سرقت مشخصات فردی کاربران است. این نوع حمله امروزه بسیار رواج پیدا کرده و مورد توجه بسیاری از محققان قرار گرفته و به همین دلیل راهکارهای بسیاری برای تشخیص این حملات ارائه شده است که هر یک از این راهکارها دارای مزایا و معایبی هستند. در این تحقیق سعی شده است تا راهکاری بیان شود که مزایا روش‌های گذشته را دارا بوده و از معایب آن‌ها در روش پیشنهادی جلوگیری شود. در این روش پیشنهادی برای اینکه سربار کاهش پیدا کند انتخاب ویژگی‌های موثر استفاده شده است که برای این کار از راهکار ترکیبی الگوریتم شبکه عصبی-ژنتیک و الگوریتم دسته ماهی استفاده شده است. در این روش پیشنهادی از درخت تصمیم بهبود یافته استفاده شده است، به‌طوری که دارای کمترین ارتفاع می‌باشد و از این جهت نیازمند به حافظه کمتری می باشد و می تواند سربار پردازشی و حافظه ای را به شدت کاهش دهد. راهکار پیشنهادی قادر است با دقت بسیار بالایی حملات فیشینگ را شناسایی نماید و همچنین پارامترهای مورد بررسی را از حیث درجه اهمیت رتبه بندی نماید. دقت روش پیشنهادی بیان شده در این تحقیق برابر ۹۷,۰۸ درصد می‌باشد که دقت بسیار بالایی نسبت به الگوریتم‌های مشابه در این زمینه است.

واژه‌های کلیدی: حملات فیشینگ، شبکه‌های عصبی، الگوریتم دسته ماهی، الگوریتم ژنتیک، فیچرهای کلیدی

مقدمه

امروزه با پیشرفت علم و تکنولوژی‌های مختلف، مخصوصاً موارد مرتبط با سیستم‌های کامپیوتری، سطح جدیدی از زندگی را برای مردم ایجاد نموده است. با پیشرفت تکنولوژی درزمینه ارتباطات و فناوری اطلاعات، سیستم‌های ارائه‌دهنده خدمات، جایگزین خدمات مبتنی بر انسان شده‌اند. سازمان‌ها و ادارات اعم از دولتی و غیر دولتی، سعی در ارائه خدمات به مشتریان (ارباب رجوع) از راه‌دور دارند. بهترین روش برای ارائه این نوع از خدمات، ارائه سرویس‌های موردنیاز در بستر شبکه (اینترنت و یا اینترانت) می‌باشد. براین اساس تولید برنامه‌ها به‌صورت تحت‌وب بهترین راه حل برای ارائه خدمت می‌باشد (طالبی،۱۳۹۸).

بسیاری از صفحات وب از مشابهت‌های بالایی با وب‌سایت‌های اصلی برخوردارند که موجب می‌شود کاربران ناآگاه اینترنت به‌سادگی در دام این نوع فریب گرفتار شوند. لذا مساله تشخیص و شناسایی وب‌سایت‌های جعلی، از اهمیت بالایی برخوردار است و واقعاً پیچیده و پویا می‌باشد. از آنجا که تکنیک‌های داده‌کاوی روش طبیعی‌تری برای مواجهه با عوامل کیفی به جای مقادیر دقیق هستند، لذا می‌توانند در تشخیص و دسته‌بندی وب‌سایت‌های جعلی سودمند باشند. هدف اصلی ما در این مقاله تشخیص صفحات فیشینگ با استفاده از الگوریتم‌های داده‌کاوی است.

تعداد قابل توجهی از شرکت ها برای ارائه و فروش محصولات و خدمات خود از اینترنت استفاده می‌کنند و همچنین تعداد کاربران اینترنت که برای انجام کارهای روزانه خود از وب استفاده می‌کنند، در حال افزایش است. این انقلاب به سوی تجارت الکترونیک، مسیر معمولی انجام کسب و کار را تغییر داده است و همچنین این توسعه سریع باعث ایجاد مشکلاتی هم در شرکت‌ها و هم در مشتریان شده است. یکی از محصولات و خدمات اینترنت که بیش‌از هر محصول دیگر مورد استقبال قرار گرفته، برنامه‌های کاربردی تحت‌وب می‌باشند که در زمینه‌های مختلف ارتباطی، علمی، پزشکی، کسب وکار و سرویس‌های عمومی گسترش یافته‌اند. گسترش برنامه‌های کاربردی تحت‌وب باعث شده است که موضوع حفاظت از شبکه‌های کامپیوتری و داده‌های آن به یک چالش تبدیل شود. برنامه‌های کاربردی تحت‌وب به‌دلیل ساختاری که دارند، بیشتر تحت تأثیر حمله‌ها و نفوذهای غیرمجاز قرار می‌گیرند (ناصحی ،۱۳۹۸).

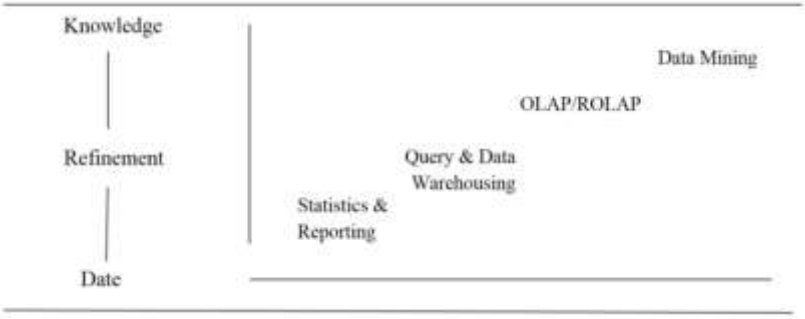
برای مثال کاربران وب برای خریداری یک محصول با انتخابات متعددی مواجه می‌شوند که باعث سردرگمی و اتلاف وقت آن‌ها می‌شود. بنابراین تدبیر استراتژی‌های بازاریابی جدید، بسیار مهم است. با گسترش استفاده از سیستم‌های پایگاهی، حجم داده‌های ذخیره‌شده در سیستم‌های تجارت الکترونیکی روزبه‌روز درحال افزایش است. از طرفی در دنیای به‌شدت رقابتی امروز، اطلاعات به‌عنوان یکی از فاکتورهای مهم پدیدار شده است.

بنابراین شرکت‌های تجارت الکترونیکی که با گذر زمان با حجم عظیمی از داده‌ها مواجه می‌شوند، نیاز به ابزاری دارند تا بتوانند این داده‌های حجیم را تحلیل کنند و اطلاعات مفیدی را از آن‌ها استخراج نمایند. درواقع امروزه نیاز به روش‌هایی است که اصطلاحاً به کشف دانش بپردازد، یعنی با کمترین دخالت کاربر به صورت خودکار الگوها و رابطه‌های منطقی را بیان نماید. داده‌کاوی یکی از مهم‌ترین این روش‌هاست که توسط الگوریتم‌های پیچیده، وابستگی‌ها و الگوهای پنهان و مفید را کشف می‌کند و اطلاعاتی را در اختیار تحلیلگران قرار می‌دهد تا براساس آن، تصمیمات مهم و حیاتی سازمان‌ها اتخاذ شود. بنابراین در سایت‌های تجارت الکترونیکی، به‌دلیل تعداد زیاد بازدیدکنندگان، وب‌کاوی امری حیاتی است. وب‌کاوی با جمع‌آوری منابع داده‌ای و پردازش آن‌ها، الگوهای استفاده کاربران وب را با به‌کارگیری تکنیک‌های داده‌کاوی استخراج می‌کند و با تحلیل این الگوهای دسترسی، پیشنهادات مناسبی را به کاربران وب ارائه می‌دهد. همچنین وب‌کاوی با شخصی‌سازی وب‌سایت‌ها باتوجه به علایق و رفتارهای کاربران می‌تواند به جذب هر چه بیشتر مشتریان و حفظ مشتریان قبلی کمک کند و همچنین کاربران می‌توانند با صرف کمترین زمان به اطلاعات موردنظر خود دست یابند که در نهایت به حجم بالایی تجارت در وب‌سایت‌های تجارت الکترونیک می‌انجامد. درنتیجه در این تحقیق ما به دنبال روشی هستیم تا براساس کاربرد داده‌کاوی وب و به کارگیری تکنیک‌های نوین داده‌کاوی، تشخیص صفحات فیشینگ سایت‌ها را انجام دهیم.

مبانی نظری

داده‌کاوی

برخلاف گذشته امروزه کمبود داده مشکل کسب و کارهای کنونی نیست، بلکه مسئله حیاتی توانایی استخراج اطلاعات مفید است. به‌دلیل افزایش ورود اطلاعات به سبب فراهم بودن بسترهای ارتباطی و نیز وجود رایانه‌های پرسرعت و لزوم پردازش آن‌ها توسط مدیریت و اتخاذ تصمیمات به موقع، استفاده از روش‌های آماری به‌عنوان ابزاری جهت استخراج اطلاعات مفید از حجم انبوهی از داده، افزایش یافته است. روش‌های آماری پدیده‌های غیر ساخت‌یافته و غیر سازماندهی شده را در قالبی ساخت‌یافته و سازماندهی شده نشان می‌دهند. شکل(۱) مراحل پالایش داده را نشان می‌دهد.



شکل۱- روند تکامل جستجو برای کسب اطلاعات

درواقع رشد سریع دسترسی داده و تکنولوژی‌های پایگاه داده نیاز به ایجاد روش‌ها و ابزارهای جدید تبدیل داده به اطلاعات و دانش مفید، به صورتی هوشمندانه و اتوماتیک را دوچندان نموده است. بنابراین داده‌کاوی به‌عنوان حوزه‌ای تحقیقاتی از اهمیت زیادی برخوردار است.

فیشینگ

فیشینگ نوعی حمله مهندسی اجتماعی و روشی برای جمع‌آوری اطلاعات شخصی افراد با استفاده از ایمیل‌ها و وب‌سایت‌های فریبنده است. فیشینگ یکی از رایج‌ترین حملات سایبری که هکرها از ارتباطات نوشتاری (به عنوان مثال ایمیل یا پیام فوری) برای سرقت اطلاعات شخص به عنوان یک منبع معتبر استفاده می‌کنند. در واقع هدف این است که گیرنده ایمیل فریب بخورد و با تصور اینکه پیام موردنظر چیزی است که او می‌خواهد، روی لینک کلیک کرده و یا پیوست را بارگیری کند. این حملات از روزهای اولیه اینترنت وجود داشته و مجرمان سایبری اولین حملات فیشینگ را در اواسط دهه ۱۹۹۰ با استفاده از سرویس آمریکا آنلاین[1] برای سرقت رمزهای عبور و اطلاعات کارت اعتباری مورد استفاده قرار دادند. انواع حملات فیشینگ عبارتند از: فیشینگ ایمیل[2]،اچ تی تی پی اس فیشینگ[3]، اسپیر فیشینگ[4] و ...

[1] America Online(AOL)
[2] Email phishing
[3] HTTPS phishing
[4] Spear phishing

ضرورت شناسایی حملات فیشینگ

با توجه به روند رو به رشد اینترنت، تعداد رایانه های متصل به اینترنت روز به روز در حال افزایش است و اکثر شرکت ها نیز در حال انتقال اعمال اصلی خود از دنیای واقعی به دنیای سایبری هستند، اگر چه این موضوع باعث افزایش فروش آن‌ها می‌شود اما می‌تواند آسیب‌پذیری های زیادی مانند حملات سایبری برای این شرکت‌ها را درپی داشته باشد به ویژه با ساختار ناشناس اینترنت. یکی از حملات جذاب (برای مخربان) حملات فیشینگ است که از ناآگاهی کاربران جهت دستیابی به اطلاعات شخصی و مهم، اقدام به فریب کاربران می‌نمایند، لذا برای به حداقل رساندن آسیب ناشی از این گونه حملات، باید در اسرع وقت شناسایی شود. در این پروژه تمرکز ما بر روی روش های مبتنی بر یادگیری ماشین می‌باشد.

هوش مصنوعی[۵]

هوش مصنوعی که گاهی اوقات هوش ماشین نیز نامیده می‌شود شبیه‌سازی فرآیندهای هوش طبیعی[۶] توسط ماشین‌ها، به‌ویژه سیستم‌های رایانه‌ای است. به عبارت دیگر، هوش مصنوعی به سامانه‌هایی گفته می‌شود که می‌توانند واکنش‌هایی مشابه رفتارهای هوشمند انسانی، از جمله، درک شرایط پیچیده، شبیه‌سازی فرآیندهای تفکری و شیوه‌های استدلالی انسانی و پاسخ موفق به آن‌ها را یادگیری نموده و توانایی کسب دانش و استدلال برای حل مسائل را داشته باشند. هوش مصنوعی، موضوعی بسیار گسترده‌ای است که شاخه‌های متعددی دارد. شاخه‌های هوش مصنوعی عبارتند از: یادگیری ماشین[۷]،شبکه‌های عصبی[۸] سیستم‌های خبره[۹] ، پردازش زبان طبیعی [۱۰]، تشخیص گفتار[۱۱] ، بینایی ماشین[۱۲] ، رباتیک [۱۳] و منطق فازی [۱۴].

یادگیری ماشین

یادگیری ماشین بخشی از هوش مصنوعی است، که به مطالعه الگورتیم‌های رایانه‌ای می‌پردازد و می‌تواند به طور خودکار ازطریق تجربه و با استفاده از داده‌ها بهبود یابند. طراحی این شاخه از هوش مصنوعی بر اساس این ایده است که سیستم‌ها می‌توانند از داده ها بیاموزند، الگوها را تشخیص داده و با حداقل دخالت انسان تصمیم‌گیری کنند.

[5] Artificial Intelligence
[6] Natural Intelligence
[۷] machine learning
[۸] Neural network
[۹]Expert Systems
[۱۰] Natural language processing
[۱۱] speech recognition
[۱۲] Machine vision
[۱۳] robotic
[۱۴] Fuzzy logic

یادگیری ماشینی شاخه‌ای از هوش مصنوعی[۱۵] و علوم کامپیوتر[۱۶] است که بر استفاده از داده‌ها و الگوریتم‌ها برای تقلید از روشی که انسان‌ها یاد می‌گیرند تمرکز دارد و به تدریج دقت آن را بهبود می‌بخشد. الگوریتم‌های یادگیری ماشین مدلی را بر اساس داده‌های نمونه می‌سازند که به داده‌های آموزشی معروف است تا پیش‌بینی‌ها یا تصمیم‌گیری‌ها را بدون برنامه‌ریزی صریح انجام دهند. الگوریتم های یادگیری ماشین در کاربردهای متنوعی مانند پزشکی، فیلترکردن ایمیل، تشخیص گفتار و بینایی کامپیوتری استفاده می‌شوند. انواع الگوریتم‌های یادگیری ماشین عبارتنداز:

درخت تصمیم[۱۷]، لاجستیک رگرسیون[۱۸]، ماشین بردار پشتیبان[۱۹]، دسته‌بند بیز[۲۰]، نزدیک‌ترین همسایه[۲۱] و..... برخی از الگوریتم های استفاده شده در این پژوهش است.

یادگیری عمیق[۲۲]

یادگیری عمیق بخشی از خانواده یادگیری ماشین است که بر روش‌های مبتنی بر الگوریتم‌های شبکه عصبی مصنوعی[۲۳] تمرکز دارد و شبیه‌سازی مغز انسان، هدف این الگوریتم‌ها می‌باشد. به طور خلاصه در یادگیری عمیق شبکه‌های عصبی مصنوعی و الگوریتم‌های مشابه مغز بشر از مجموعه‌های عظیم‌داده، مهارت‌های مورد نظر را فرا می‌گیرند. همانطور که ما از طریق تجربه چیزهای جدید یاد می‌گیریم، مهارت الگوریتم یادگیری عمیق نیز با هر بار تکرار یک فعالیت، نسبت به دفعات قبلی بهبود می‌یابد. دلیل استفاده از عبارت یادگیری عمیق این است که شبکه‌های عصبی لایه‌های مختلف یا عمیقی دارند که یادگیری را ممکن می سازد.

پیشینه پژوهش

(مهمت کارکمز و همکاران،۲۰۲۱)،دسته‌بندی لینک‌های فیشینگ[۲۴] با استفاده از شبکه‌های عصبی عمیق بر روی دیتاست لینک‌های پرخطر

با توجه به روند رو به رشد اینترنت، تعداد رایانه‌های متصل به اینترنت روز به روز در حال افزایش است. تقریباً همه شرکت‌ها در حال انتقال اعمال اصلی خود از دنیای واقعی به دنیای سایبری هستند، اگر چه آن موضوع باعث افزایش فروش آن‌ها می‌شود اما در مقابل می‌تواند آسیب‌پذیری های زیادی مانند حملات سایبری برای این شرکت‌ها را درپی داشته باشد، به ویژه با ساختار ناشناس اینترنت. فیشینگ یکی از انواع محبوب حملات سایبری است که از ناآگاهی کاربر برای آسیب رساندن به آن‌ها سوء استفاده می شود. در برخی از پژوهش‌ها از سیستم‌های تشخیص مبتنی برقانون به عنوان یک مکانیسم پیشگیری استاتیک و سیستم‌های مبتنی بر یادگیری ماشین به عنوان یک مکانیسم پیشگیری پویا کمک می‌گیرد. در مقاله یاد شده نویسندگان یک سیستم تشخیص فیشینگ مبتنی بر شبکه‌های عصبی عمیق[۲۵] را با تجزیه و تحلیل نشانی اینترنتی وب‌سایت های مشکوک پیاده‌سازی کرده‌اند. اگر چه در تمام تحقیقاتی قبلی که توسط محققان دیگر انجام شده مجموعه داده‌های مورد استفاده از منابع مختلفی جمع آوری شده بودند که وب‌سایت های مجاز و فیشینگ در آن‌ها مشخص است، اما نویسندگان مقاله یاد شده ابتدا یک مجموعه داده

15 artificial intelligence
16 computer science
17 Decision tree
18 Logistic Regression
19 Support vector machine
20 Naive Bayes Classifiers
21 KNN
۲۲ Deep learning
۲۳ Artificial neural network
24 Phishing
25 Deep Neural Network (DNN)

پرخطر ایجاد می‌کنند که فقط شامل وب‌سایت های مشکوک است که به سایت فیشتانک[۲۶] گزارش شده‌اند. تحقیقات تجربی نویسندگان مقاله یاد شده نشان می‌دهد که سیستم پیشنهادی توسط نویسندگان مقاله، کارآیی بسیار خوبی از نظر دقت و زمان اجرا ارائه می‌دهد.

(او.اسماعیل و همکاران ،۲۰۱۰)،کاهش خطر حملات فیشینگ با استفاده از ابزار های هوش مصنوعی

در مقاله یاد شده یک پژوهش جامع به همراه تجزیه و تحلیل در مورد فیشینگ، ویشینگ[۲۷]و اسمیشینگ[۲۸] برای بهره‌برداری از دانش در پیاده‌سازی یک ابزار هوشمند برای شناسایی و محافظت در مقابل موارد ذکر شده ارائه شده است. این یک مشکل جدید مهندسی اجتماعی است که زندگی روزمره ما را آسیب‌پذیر و دشوار می‌کند. تحقیق یاد شده به ویژه بر روی فیشینگ از طریق ایمیل متمرکز است زیرا، به طور مستقیم پیامدهای جدی‌تری در مقایسه با سایر روش‌ها در رابطه با معاملات مالی دارد. شایان ذکر است که امن نمودن حجم بسیار زیاد معاملات آنلاین بسیار چالش‌برانگیز است زیرا روزانه چندین روش برای نقض حریم خصوصی افراد به منظور سرقت اعتبار آن‌ها ابداع می‌شود. هزینه این نوع حملات بیش از میلیون ها دلار به صورت سالانه است. ابزارهای زیادی برای حل این مشکل ارائه شده است. اما متأسفانه، این معضل هنوز پا برجاست. در مقاله نام برده شده توسط نویسندگان یک روش برای توسعه یک ابزار هوشمند برای مقابله با این خطر پیشنهاد شده است.

(یزدان احمد السریرا وهمکاران،۲۰۲۰)،الگوریتم های فرایادگیری و درختان تصمیم اضافی برای شناسایی صفحات فیشینگ

فیشینگ نوعی حمله مهندسی-وب اجتماعی[۲۹] در فضای مجازی است که در آن مجرمان اطلاعات ارزشمند کاربران ناآگاه را به سرقت می‌برند. اقدامات متقابل موجود در قالب نرم‌افزار ضد فیشینگ و روش های محاسباتی برای شناسایی فعالیت های فیشینگ اثبات شده است. با این حال، هکرها روش های جدیدی را برای خنثی سازی این اقدامات متقابل به کار می گیرند. با توجه به ماهیت تکامل یافته حملات فیشینگ، نیاز به اقدامات متقابل جدید و کارآمد بسیار مهم احساس می شود زیرا تأثیر حملات فیشینگ اغلب فاجعه بار است. طرح‌های هوش مصنوعی سنگ بنای روش‌های مقابله ای مدرنی است که برای کاهش حملات فیشینگ استفاده می‌شود. اقدامات متقابل یا روش‌های مقابله‌ای فیشینگ مبتنی بر هوش مصنوعی دارای نواقصی به ویژه بالا بودن میزان هشدار کاذب و عدم توانایی تفسیر نحوه عملکرد بیشتر روش‌های فیشینگ است. در این مقاله نویسندگان چهار مدل فرا آموز[۳۰]:آدابوست اکسترا تری[۳۱]، بگینگ اکسترا تری[۳۲]، روتیشن فارست اکستراتری[۳۳]، لجیت بوست اکستراتری[۳۴]، بر مبنای الگوریتم طبقه‌بندی اکستراتری توسعه داده‌اند. مدل‌ها توسط نویسندگان این مقاله روی مجموعه داده‌های وب‌سایت های فیشینگ آموزش داده شده و عملکرد آن‌ها مورد ارزیابی قرار گرفته است. دقت تشخیص مدل‌ها بالای ۹۷ درصد با نرخ مثبت کاذب ناچیز ۰،۰۲۸بوده است. علاوه بر این مدل‌های پیشنهاد شده توسط نویسندگان مقاله نام برده شده در تشخیص حملات فیشینگ در مدل های مبتنی برای یادگیری ماشینی برتری دارند. از این رو نویسندگان مقاله یاد شده مدل‌های فراآموز را برای تشخیص حملات قیشینگ اتخاذ نموده‌اند.

[26] PhishTank
[27] Vishing
[28] Smishing
[29] social web-engineering
[30] meta-learner
[31] AdaBoost-Extra Tree (ABET)
[32] Bagging - Extra tree (BET)
[33] Rotation Forest - Extra Tree (RoFBET)
[34] LogitBoost-Extra Tree (LBET)

روش پیشنهادی

ویژگی‌های مختلفی وجود دارد که از طریق آن‌ها می‌توان پی برد که یک وب‌سایت فیشینگ می‌باشد و برای کشف این موضوع، می‌بایست این ویژگی‌ها را در نظر گرفت. در این تحقیق پس از بررسی مقالات و مشاوره در این زمینه با متخصصان، ۱۸ مورد از این ویژگی‌ها انتخاب شدند که لیست آن‌ها در جدول ۱ آورده شده‌اند.

جدول ۱- ویژگی‌های مورد بررسی در روش پیشنهادی

شماره	ویژگی	توضیحات
۱	آدرس IP[35]	استفاده از آدرس IP در قسمت میزبان یک Url یکی از مهمترین مشخصات برای کشف حمله می‌باشد زیرا بدین معنا است که فردی با این آدرس IP در حال سرقت اطلاعات شخصی می‌باشد بنابراین ما این ویژگی را به عنوان یکی از مهمترین ویژگی در تشخیص حمله در نظر گرفتیم.
۲	Url طولانی	Url طولانی به عنوان یکی از مهمترین ویژگی‌های حمله می‌باشد زیرا با این کار افراد مخربی که قصد حمله دارند می‌توانند قسمت‌های مخرب Url را در میان Url طولانی طوری بگنجانند که کاربران براحتی قادر به تشخیص آن نباشد.
۳	Urlهای دارای سمبول "@"	Url دارای "@" باشد باعث می‌شود که مرورگر به محتوای قبل از این سمبول توجه نداشته و کاربر را به لینکی که بعد از آن درج شده است هدایت می‌کند یعنی حمله کننده می‌تواند در ابتدا از آدرس یک سایت کاملاً معتبر استفاده نماید و در ادامه از این سمبول استفاده و بعد از آن آدرسی که برای حمله می‌باشد را وارد نماید. براین اساس کاربر که به قسمت ابتدایی Url توجه می‌کند متوجه این حمله نمی‌شود.
۴	ابتدا و انتهای Url	حمله کنندگان برای اینکه آدرس را به عنوان یک آدرس کاملاً معتبر نشان دهند سعی می‌کنند تا آدرسی مشابه با آدرس‌های معتبر ایجاد نمایند و این کار را با استفاده از افزودن عبارت‌های ابتدایی و انتهایی انجام می‌دهند. بنابراین کاربران احساس تفاوتی نمی‌کنند و این لینک‌ها را به عنوان لینک‌های معتبر شناسایی می‌کنند.
۵	زیر دامین Url	یکی از مهمترین حملات از طریق افزودن زیر دامین‌هایی به Url می‌باشد که از این طریق افراد احساس می‌کنند که این سایت‌ها معتبر می‌باشند. بیشترین حمله‌ای که از این طریق در سال‌های اخیر صورت گرفته است جعل سایت‌های پرداخت آنلاین است.
۶	سوء استفاده از پروتکل امن HTTPs	استفاده از پروتکل امن HTTPs باعث می‌شود که اطلاعات مهم از طریقی تونلی امن ارسال شود و از این طریق کاربران احساسی آسوده در زمینه این نوع از وب سایت‌ها دارند حال آنکه بسیاری از حمله کنندگان ممکن است از نوعی HTTPs استفاده کنند که به‌صورت تقلبی است و فقط در جهت کسب اعتماد کاربران می‌باشد و در این حالت بسیاری از کاربران به این سایت‌ها اعتماد می‌نمایند. بعد از بررسی‌هایی که ما انجام دادیم متوجه شدیم که می‌توان این پروتکل‌های HTTPs را با استفاده از سایت‌های بررسی کننده‌ای همچون "GeoTrust" و یا "GoDaddy" استفاده نمود و از صحت پروتکل‌های HTTPs اطمینان کسب نمود.

[35] IP Address

شماره	ویژگی	توضیحات
۷	Url درخواستی	یک وب‌سایت معمولاً از یکسری متن و اشیایی همچون عکس و ویدئوها تشکیل می‌شود. بصورت معمول این اشیا از جایی کنار صفحات وب در همان دامین بارگذاری می‌شود. در صورتی که اشیا از دامینی غیر از دامین اصلی وبسایت بارگذاری شود، احتمال حمله بودن این وبسایت بسیار زیاد می‌باشد زیرا ممکن است وبسایت جعلی ایجاد شده باشد.
۸	Url لنگر[۳۶]	دقیقا مشابه مورد قبل یعنی "Url درخواستی" می‌باشد با این تفاوت که در این حالت لینک‌های موجود در وبسایت که به دامینی اشاره می‌کند متفاوت از دامینی باشد که در آدرس بار Url تایپ می‌شود. این ویژگی نیز مانند "Url درخواستی" شناسایی می‌شود.
۹	مدیریت کننده‌ای به شکل سرور[۳۷]	زمانی که کاربر اطلاعاتی را ارسال می‌نماید این اطلاعات به سروری برای پردازش این اطلاعات انتقال داده می‌شود به طور کامل معمول و نرمال اطلاعات به سروری در همان دامین ارسال می‌شود. خمله کنندگان معمولا این سرورها را خالی می‌گذارند و یا آن را به دامین دیگری انتقال می‌دهند.
۱۰	Url غیر طبیعی	اگر اطلاعاتی که به عنوان اطلاعات شناسایی وبسایت در پایگاه داده WHOIS (**http://who.is** یا **http://whois.net**) موجود می‌باشد غیر طبیعی باشد باعث می‌شود که وبسایت به عنوان یک وبسایت حمله کننده شناسایی شود.
۱۱	تغییر مسیر صفحه وبسایت[۳۸]	این کار بصورت کاملا معمول توسط حمله کنندگان انجام می‌شود که قصد دارند تا لینک واقعی را از دید کاربران پنهان نگه دارند که با استفاده از لینکی درخواست ارسال اطلاعات کاربران را می‌کنند و به محض تایید کاربران آن‌ها را به لینک مخربی هدایت می‌کنند.
۱۲	پنجره‌های Pop-up	در سایت‌های معتبر بسیار کم پیش می‌آید که از کاربران درخواست کنند تا اطلاعاتی همچون اطلاعات ورود را با استفاده از پنجره‌های pop-up ارسال نمایند زیرا این کار امن نیست و این کار از طرف سایت‌های معتبر کاملا غیر معمول می‌باشد.
۱۳	پنهان کردن لینک‌های مشکوک	حمله کنندگان سعی می‌کنند تا لینک‌های جعلی را در نوار وضعیت نشان دهند و یا اینکه کل نوار وضعیت را پنهان می‌سازند.
۱۴	رکورد DNS	اگر رکورد DNS پیدا نشود و یا خالی باشد وبسایت به عنوان "phishy" در نظر گرفته می‌شود و در غیر این صورت به عنوان یک سایت معتبر شناخته می‌شود.
۱۵	ترافیک وب‌سایت	ترافیک وبسایت‌های معتبر معمولا بالاست زیرا این وبسایت‌ها بصورت مرتب توسط کاربران بازدید می‌شوند و حال آنکه وب‌سایت‌های غیرمعتبر دارای ترافیک بسیار کمی می‌باشند و یا اصلاً در رنکینگ در نظر گرفته نمی‌شوند زیرا ترافیک آن‌ها زیر حداقل ترافیک مجاز است بنابراین از این طریق نیز می‌توان یک وبسایت معتبر را شناسایی نمود.
۱۶	عمر دامین[۳۹]	یک وبسایت به عنوان یک وبسایت معتبر در نظر گرفته می‌شود اگر عمر یا سن دامین حداقل ۲ سال باشد.
۱۷	غیر فعال کردن	معمولا حمله کنندگان از اسکریپت‌های جاوا استفاده می‌کنند تا کلیک راست ماوس را در

[36] URL of Anchor
[37] Server Form Handler(SFH)
[38] Redirect Page
[39] Age of Domain

شماره	ویژگی	توضیحات
	کلیک راست ماوس	وبسایت از کار بیاندازند بنابراین کاربران نمی‌توانند محتوای صفحه را مشاهده نمایند و یا آن را ذخیره نمایند.
۱۸	شماره پورت	شماره پورت در صورتی که در Url وجود داشته باشد مورد بررسی قرار می‌گیرد که در لیست پورت‌های مجاز و شناخته شده HTTP باشند پورت‌هایی همچون ۸۰، ۸۰۸۰، ۲۱، ۴۴۳، ۷۰ و ۱۰۸۰. اگر شماره پورت در لیست مجاز وجود نداشته باشد به عنوان یک وبسایت غیر معتبر و "phishy" مشخص می‌شود.

تجزیه و تحلیل یافته ها

بررسی روش پیشنهادی

در این بخش روش پیشنهادی را با چند مورد از الگوریتم‌های تصمیم‌گیری همچون ID3 مورد مقایسه قرار می‌گیرد. در این بررسی داده‌های مربوط به اطلاعات ۱۱۰۵۵ نمونه که از سایت UCI [۱۳] جمع آوری شده است مورد بررسی قرار گرفته است. روشی که برای انتخاب داده‌های آزمون و تست مورد استفاده قرار گرفته است، روش cross fold K validation است. در اینجا مقدار k برابر ۱۰ قرار داده شده است که این مقدار یک مقدار متعارف در زمینه داده کاوری می‌باشد.

داده‌هایی که در اینجا مورد استفاده قرار گرفته است در ۰آورده شده است و برنامه بر اساس این پارامترها عملکرد خود را نشان می‌دهد. این پارامترها، پارامترهای تاثیر گذار در نتیجه می‌باشند که با توجه به مشاوره با افراد متخصص در این زمینه جمع آوری شده است.

در برنامه‌ای که در اینجا پیاده سازی شده است از زبان برنامه نویسی C# کمک گرفته شده است و پیاده سازی در محیط برنامه نویسی Visual Studio 2019 صورت گرفته است. در حین کار از کتابخانه‌هایی نیز کمک گرفته شده است که عبارتند از:

- Weka: نرم افزار Weka در دانشگاه Waikato واقع در نیوزلند توسعه یافته است و اسم آن از عبارت " Waikato Environment for knowledge Analysis" استخراج گشته است. همچنین Weka، نام پرنده‌ای با طبیعت جستجوگر است که پرواز نمی‌کند و در نیوزلند، یافت می‌شود. این سیستم به زبان جاوا نوشته شده و بر اساس لیسانس عمومی و فراگیر [۱۹] GNU انتشار یافته است. Weka تقریباً روی هر پلت فرمی اجرا می‌شود و نیز تحت سیستم عامل‌های لینوکس، ویندوز، و مکینتاش، و حتی روی یک منشی دیجیتالی شخصی [۲۰]، آزمایش شده است. این نرم افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش پردازش، پس از پردازش [۲۱] و ارزیابی نتایج طرح‌های یادگیری روی همه مجموعه‌های داده موجود، قابل اعمال است. نرم افزار Weka، پیاده سازی الگوریتم‌های مختلف یادگیری را فراهم می‌کند و به آسانی می‌توان آنها را به مجموعه‌های داده خود اعمال کرد. ما در این پیاده سازی از این نرم افزار به عنوان کتابخانه‌ای در کنار پروژه خود استفاده می‌کنیم.
- ZedGraph: این کتابخانه برای رسم نمودار می‌باشد که با استفاده از این کتابخانه، نمودارهای مورد نیاز برنامه ایجاد شده اند.

سیستم مورد استفاده در اینجا دارای سیستم عامل Windows 10، دارای ۱۶ گیگ RAM و Corei7 می‌باشد.

فایل ورودی برنامه با فرمت ARFF می‌باشد که ساختاری مناسب و استاندارد برای تحلیل می‌باشد.

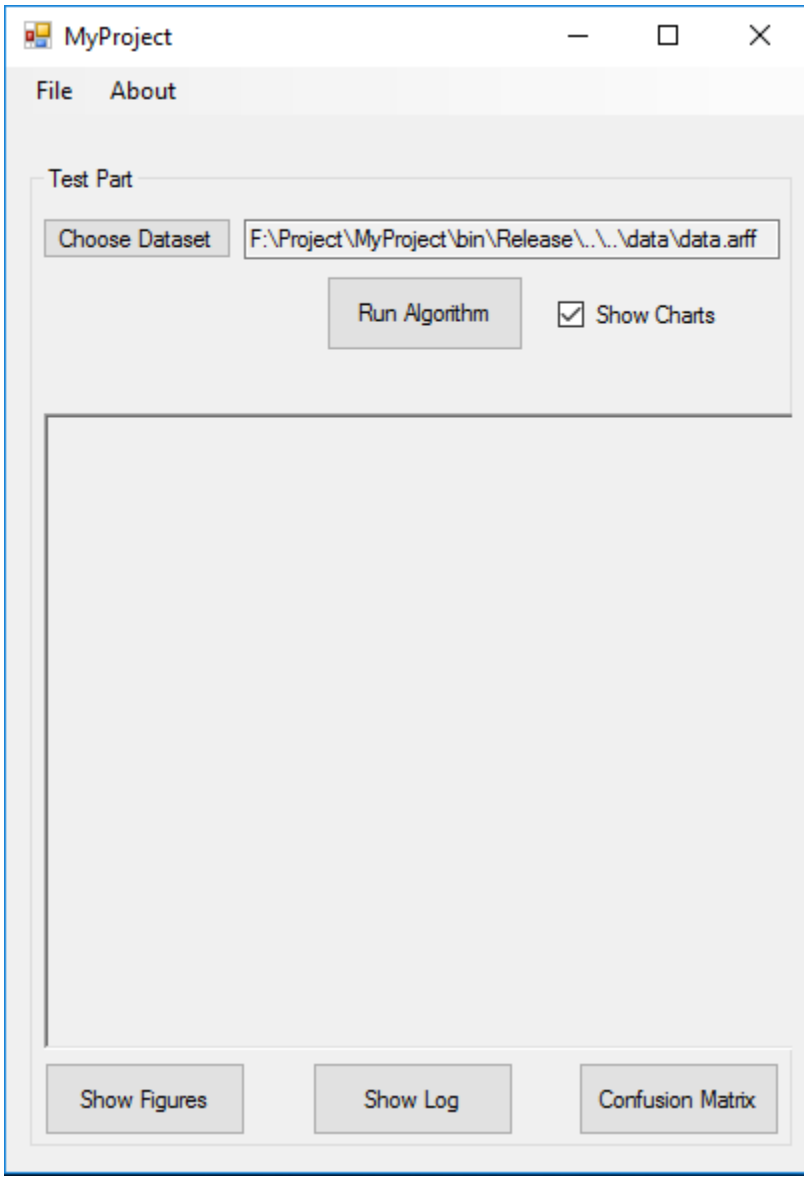
دوره ۱۰، شماره ۴، زمستان ۱۴۰۳، صفحات ۷۳-۵۱

```
@relation phishing

@attribute having_IP_Address  { -1,1 }
@attribute URL_Length   { 1,0,-1 }
@attribute Shortining_Service { 1,-1 }
@attribute having_At_Symbol   { 1,-1 }
@attribute double_slash_redirecting { -1,1 }
@attribute Prefix_Suffix  { -1,1 }
@attribute having_Sub_Domain  { -1,0,1 }
@attribute SSLfinal_State  { -1,1,0 }
@attribute Domain_registeration_length { -1,1 }
@attribute Favicon { 1,-1 }
@attribute port { 1,-1 }
@attribute HTTPS_token { -1,1 }
@attribute Request_URL  { 1,-1 }
@attribute URL_of_Anchor { -1,0,1 }
@attribute Links_in_tags { 1,-1,0 }
@attribute SFH  { -1,1,0 }
@attribute Submitting_to_email { -1,1 }
@attribute Abnormal_URL { -1,1 }
@attribute Redirect  { 0,1 }
@attribute on_mouseover  { 1,-1 }
@attribute RightClick  { 1,-1 }
@attribute popUpWidnow  { 1,-1 }
@attribute Iframe { 1,-1 }
@attribute age_of_domain  { -1,1 }
@attribute DNSRecord   { -1,1 }
@attribute web_traffic  { -1,0,1 }
@attribute Page_Rank { -1,1 }
@attribute Google_Index { 1,-1 }
@attribute Links_pointing_to_page { 1,0,-1 }
@attribute Statistical_report { -1,1 }
@attribute Result  { -1,1 }


@data
-1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-1,-1
1,1,1,1,1,-1,0,1,-1,1,1,-1,1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1,-1
```

شکل ۲- فایل Arff ورودی برنامه



شکل ۳- برنامه پیاده سازی شده از الگوریتم پیشنهادی

بعد از اجرای برنامه پیش پردازش و انتقال اطلاعات روی فایل ورودی انجام می‌گیرد و در صورت بروز مشکل فایل arff موجود حذف و فایل arff ویرایش شده در عمل انتقال و پیش پردازش جایگزین فایل اولیه می‌شود و برنامه روی این فایل کار پردازش را آغاز می‌کند.

در این پیاده سازی برای بررسی الگوریتم‌های دیگر و الگوریتم پیشنهادی از پارامترهای بررسی خطا بسیار معتبر و استاندارد استفاده شده است که عبارتند است از:

- خطای میانگین مربعات (MSE[40]): این معیار میانگین خطای مربعات را با استفاده از رابطه ۴ بدست می‌آورد و با استفاده از این معیار می‌توان به طور دقیق پیش بینی را سنجید و میزان خطای پیش بینی را محاسبه نمود.

[40] Mean squared error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(r_i - p_i)^2 \qquad ۱$$

- معیار صحیح مثبت[41](TP): در صورتی که جوابی که از پیش بینی بدست می‌آید p باشد و مقدار واقعی نیز p باشد آنگاه ۱ واحد به TP اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای تست استفاده می‌شود این مقدار جمع آوری می‌شود.

- معیار صحیح منفی[42](TN): در صورتی که جوابی که از پیش بینی بدست می‌آید n باشد و مقدار واقعی نیز n باشد آنگاه ۱ واحد به TN اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای تست استفاده می‌شود این مقدار جمع آوری می‌شود.تفاوت این قسمت با TP در این می‌باشد که در TP ما جواب مثبت و در اینجا جواب‌های منفی را شمارش می‌کنیم.

- معیار غلط مثبت[43](FP): در صورتی که جوابی که از پیش بینی بدست می‌آید p باشد و مقدار واقعی n باشد آنگاه ۱ واحد به FP اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای تست استفاده می‌شود این مقدار جمع آوری می‌شود. یعنی بطور کلی جوابی که از پیش بینی بدست می‌آید متفاوت از مقدار واقعی باشد.

- معیار غلط منفی[44](FN): در صورتی که جوابی که از پیش بینی بدست می‌آید n باشد و مقدار واقعی p باشد آنگاه ۱ واحد به Fn اضافه می‌شود بنابراین در مجموعه داده‌هایی که برای تست استفاده می‌شود این مقدار جمع آوری می‌شود. یعنی بطور کلی جوابی که از پیش بینی بدست می‌آید متفاوت از مقدار واقعی باشد. تفاوت این قسمت با قسمت قبل در این می‌باشد که در قسمت قبل ما جواب‌های منفی که به اشتباه مثبت در نظر گرفته می‌شدند را در نظر گرفتیم و در اینجا جواب‌های مثبتی که به اشتباه منفی در نظر گرفته شده‌اند در نظر می‌گیریم.

بعد از اینکه داده‌ها به شکل ورودی برنامه ایجاد شد، الگوریتم پیشنهادی روی داده‌های ورودی اجرا می‌شود. بعد از اجرا، خروجی‌های شکل های زیر در برنامه ایجاد می‌شود که مربوط به هر الگوریتم می‌باشد.

```
----------My Algorithm--------
Precision: 0.974476032371861
Recall: 0.958758677011025
ROC: 0.993551047134636
Correct Prediction Percent = 97.0601537765717%
InCorrect Prediction Percent = 2.93984622342831%
MeanAbsoluteError(MAE) = 0.0509674908151279
MeanSquaredError(MSE) = 0.151770260792834
Kappa = 0.940332073775373
RelativeAbsoluteError(REA) = 10.3274155143527
Correct Prediction Number = 10730
InCorrect Prediction Number = 325
Error Rate = 0.0293984622342831
```

شکل۴- خروجی مربوط به الگوریتم پیشنهادی

[41] True Positive
[42] True Negative
[43] False Positive
[44] False Negative

```
----------J48--------
Precision: 0.969224370971096
Recall: 0.951612903225806
ROC: 0.989686783353228
Correct Prediction Percent = 96.5174129353234%
InCorrect Prediction Percent = 3.48258706467662%
MeanAbsoluteError(MAE) = 0.0533480158349569
MeanSquaredError(MSE) = 0.16797636530023
Kappa = 0.929301674609354
RelativeAbsoluteError(REA) = 10.8097753603823
Correct Prediction Number = 21340
InCorrect Prediction Number = 770
Error Rate = 0.0348258706467662
```

شکل ۵- خروجی مربوط به الگوریتم J48 مقاله [۶]

```
----------RandomTree--------
Precision: 0.963278597658145
Recall: 0.946168504151354
ROC: 0.988058245634793
Correct Prediction Percent = 96.0168852706166%
InCorrect Prediction Percent = 3.98311472938339%
MeanAbsoluteError(MAE) = 0.0580926993239088
MeanSquaredError(MSE) = 0.177979001810422
Kappa = 0.919144085039975
RelativeAbsoluteError(REA) = 11.7711787391014
Correct Prediction Number = 31844
InCorrect Prediction Number = 1321
Error Rate = 0.0398311472938339
```

شکل ۶- خروجی مربوط به الگوریتم RandomTree مقاله [۵]

```
--------ID3----------
Precision: 0.956167652246084
Recall: 0.940843201306656
ROC: 0.971427972210024
Correct Prediction Percent = 95.4681139755767%
InCorrect Prediction Percent = 4.53188602442334%
MeanAbsoluteError(MAE) = 0.0590150242667896
MeanSquaredError(MSE) = 0.197997206010423
Kappa = 0.908020558384315
RelativeAbsoluteError(REA) = 11.9580671413371
Correct Prediction Number = 42216
InCorrect Prediction Number = 2004
Error Rate = 0.0453188602442334
```

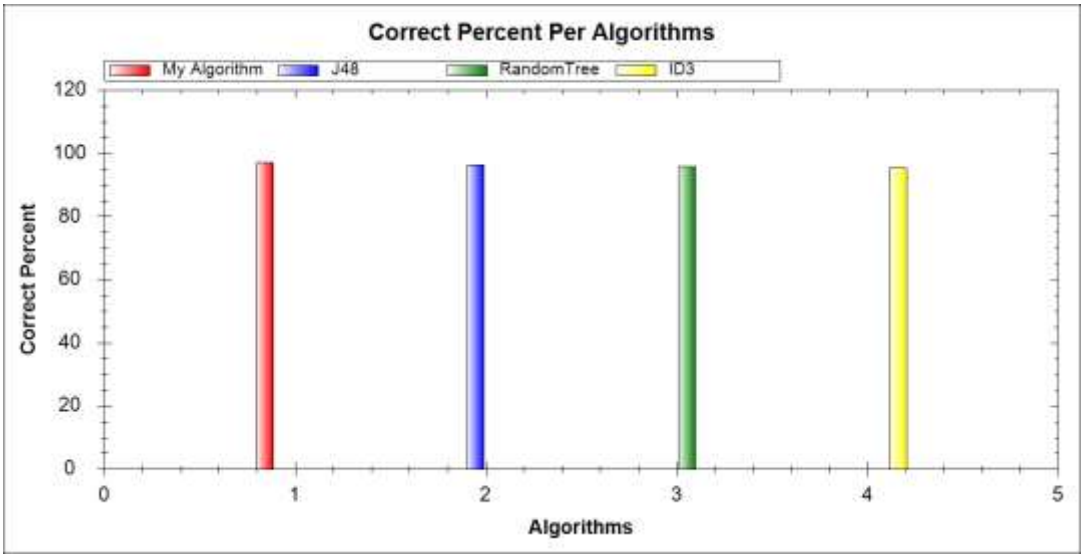
شکل ۷ - خروجی مربوط به الگوریتم ID3

با توجه به نتایج دریافتی می‌توان به وضوح مشاهده نمود که الگوریتم پیشنهادی با ٪۹۷,۰۶۰ دارای بالاترین دقت صحت و با ٪۲,۹۳۹ دارای کمترین میزان اشتباه می‌باشد . کاملا مشخص است که الگوریتم پیشنهادی از باقی بهتر و از خود الگوریتم ID3 نیز بسیار بهتر عمل می‌کند.

جدول ۲- نسبت درصد صحت پیش‌بینی‌ها و خطای پیش‌بینی‌ها

عنوان الگوریتم	درصد صحت پیش بینی ها	درصد خطای پیش بینی ها
الگوریتم پیشنهادی	٪۹۷,۰۶۰	٪۲,۹۳۹
الگوریتم J48 مقاله [۶]	٪۹۶,۵۱۷	٪۳,۴۸۲
الگوریتم RandomTree مقاله [۵]	٪۹۶,۰۱۶	٪۳,۹۸۳
الگوریتم ID3	٪۹۵,۴۶۸	٪۴,۵۳۱

در ادامه نمودارهای خطای بدست آمده از برنامه مورد بررسی قرار گرفته است. اولین نمودار، نمودار مربوط به درصد پیش‌بینی صحیح در میان داده‌های آزمون می‌باشد که می‌توان در شکل مشاهده نمود.



شکل ۸- درصد پیش بینی صحیح در میان داده‌های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم‌های مورد بررسی

همانطور که می‌توان از این نمودار دریافت روش پیشنهادی ما دارای دقت پیش بینی صحیح بیشتری نسبت به دیگر الگوریتم‌های مورد بررسی یعنی J48 مقاله [۱۰]، RandomTree مقاله [۱۲] و ID3 می‌باشد. این بدین دلیل می‌باشد که در روش پیشنهادی، برای پیش بینی تنها مواردی از آزمون در نظر گرفته می شود که تاثیر بیشتری را در نتیجه خروجی داشتند و در

نتیجه داده‌هایی که در خروجی تاثیر نداشتند استفاده نشدند و بدین شکل درصد تشخیص و زمان تحلیل بسیار کاهش یافته است و حال انکه الگوریتم‌های دیگر به دلیل استفاده از تمامی پارامترها دارای دقت کمتری هستند زیرا ممکن است بعضی از پارامترها دارای مقادیر دوری باشند که ممکن است هیچ تاثیر در نتیجه خروجی نداشته باشند ولی چون در الگوریتم‌های دیگر در ساخت مدل برای پیش بینی این پارامترها مورد استفاده قرار گرفته‌اند باعث تشخیص اشتباه و کاهش دقت می‌شوند. در روش پیشنهادی چون این پارامترها بی فایده وجود نداشتند در نتیجه دقت تشخیص در روش پیشنهادی افزایش یافته و از دیگر الگوریتم‌ها بهتر عمل نموده است.

در این تحقیق از راهکار ترکیبی الگوریتم شبکه عصبی مبتنی-ژنتیک و الگوریتم دسته ماهی برای تشخیص استفاده و یا عدم استفاده از پارامترها استفاده شده است.

جدول ۳- مقدار وزن‌های آنتروپی اختصاص داده شده به پارامترها

======Attribute Ranking Rate By Entropy======

popUpWidnow: 7.51195245533909E-09

Favicon: 8.14027878193801E-08

Iframe: 1.88255286160529E-05

Submitting_to_email: 0.000350057782657315

RightClick: 0.000449194000491072

Redirect: 0.000562792236886457

port: 0.00166104013057561

HTTPS_token: 0.00180196739611123

double_slash_redirecting: 0.00195155805708498

on_mouseover: 0.00238564859795433

having_At_Symbol: 0.00330141628292856

Links_pointing_to_page: 0.00359451402383656

Abnormal_URL: 0.00443039062770717

DNSRecord: 0.00460745174129965

URL_Length: 0.00495580259562961

Shortining_Service: 0.00605712955482196

having_IP_Address: 0.00687585209110491

Statistical_report: 0.00781172761286022

Page_Rank: 0.00971487747072772

age_of_domain: 0.0106928496701869

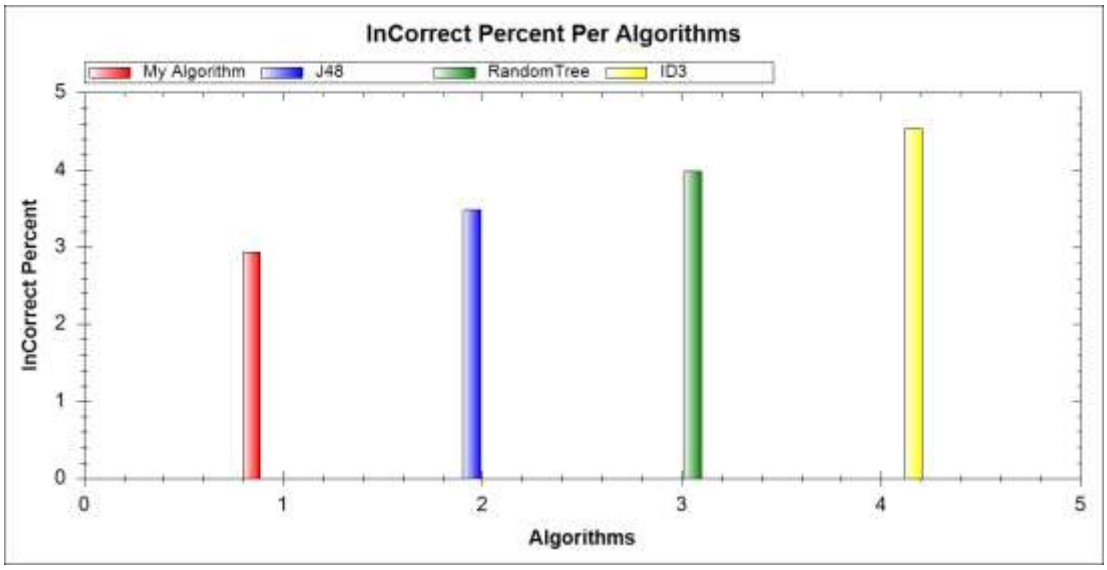
Google_Index: 0.0204870435744989

Links_in_tags: 0.0302886584872585

SFH: 0.0376771079004795
Domain_registeration_length: 0.040151249853957
Request_URL: 0.0477652449490465
having_Sub_Domain: 0.0694324755030815
web_traffic: 0.0779390619642303
Prefix_Suffix: 0.218731868560864
URL_of_Anchor: 0.316476751906907
SSLfinal_State: 0.375733072988761

می توان مشاهده نمود که با توجه به این جدول پارامتر popUpWidnow یعنی جنسیت پارامتری است که تاثیری در خروجی ندارد زیرا مقدار آن $7.5 * 10^{-9}$ می‌باشد که مقدار بسیار پایینی است و در اینجا URL_of_Anch و SSLfinal_State دارای بیشترین تاثیر در تشخیص فیشینگ می‌باشند بنابراین می‌بایست به صورت تریتبی که در جدول ۲ - آمده است به پارامترهای تاثیرگذار توجه کرد و می‌توان پارامترهایی را که دارای بیشترین تاثیر می‌باشند را انتخاب و پارامترهایی که دارای تاثیر نمی‌باشند و یا دارای تاثیر بسیار کمی می‌باشند را حذف نمود.

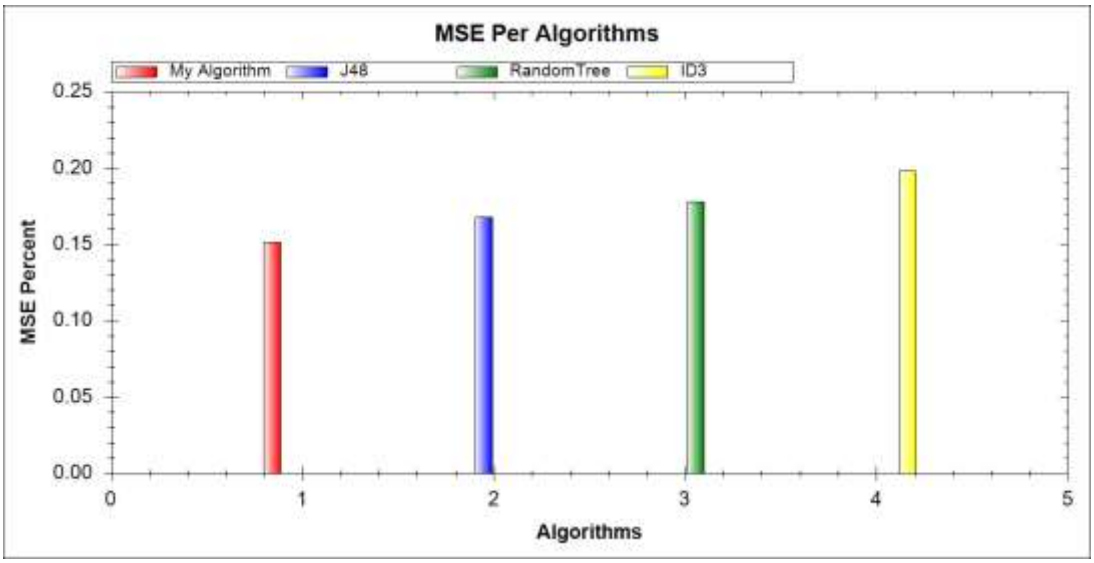
در نمودار ارائه شده در ۰می توان درصد پیش بینی غلط را مشاهده نمود. با توجه به این نمودار می‌توان درک نمود که روش پیشنهادی به همان دلیل که پیش‌تر در رابطه با درصد پیش بینی درست گفته شد از باقی روش‌ها دارای مقدار کمتری است یعنی پیش بینی اشتباه کمتری دارد بنابراین روش پیشنهادی بهتر از باقی روش‌ها عمل کرده است.



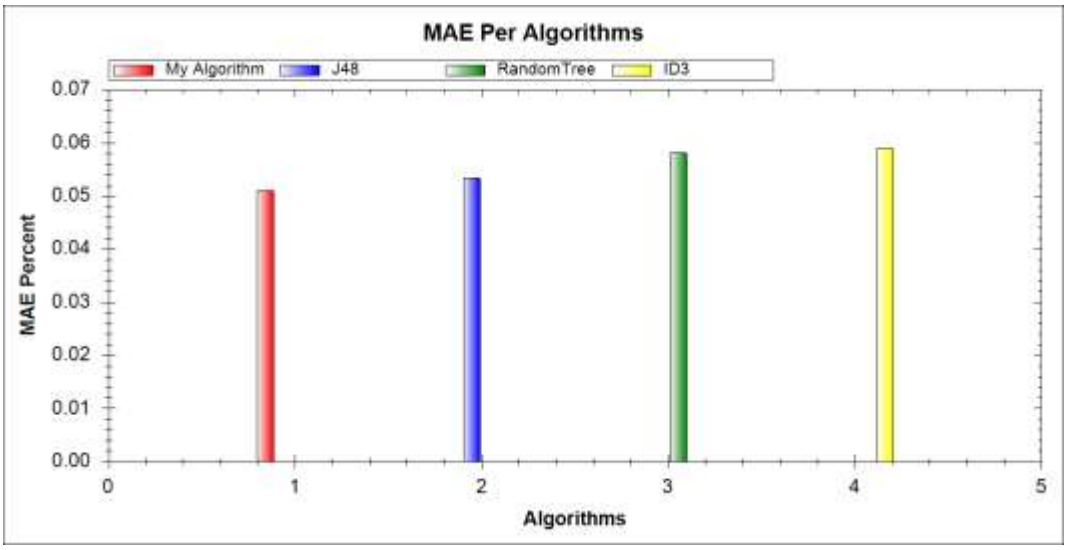
شکل ۹- درصد پیش بینی ناصحیح در میان داده‌های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم‌های مورد بررسی

با توجه به نمودار شکل می توان به این قضیه پی برد که الگوریتم پیشنهادی و الگوریتم J48 دارای خطای پیش بینی کمتری نسبت به مابقی الگوریتم‌ها هستند و روش پیشنهادی با اختلاف از الگوریتم J48 نیز بهتر عمل می‌کند و دارای خطای پیش بینی

کمتری می‌باشد و این نشان دهنده عملکرد مناسب روش پیشنهادی می‌باشد.

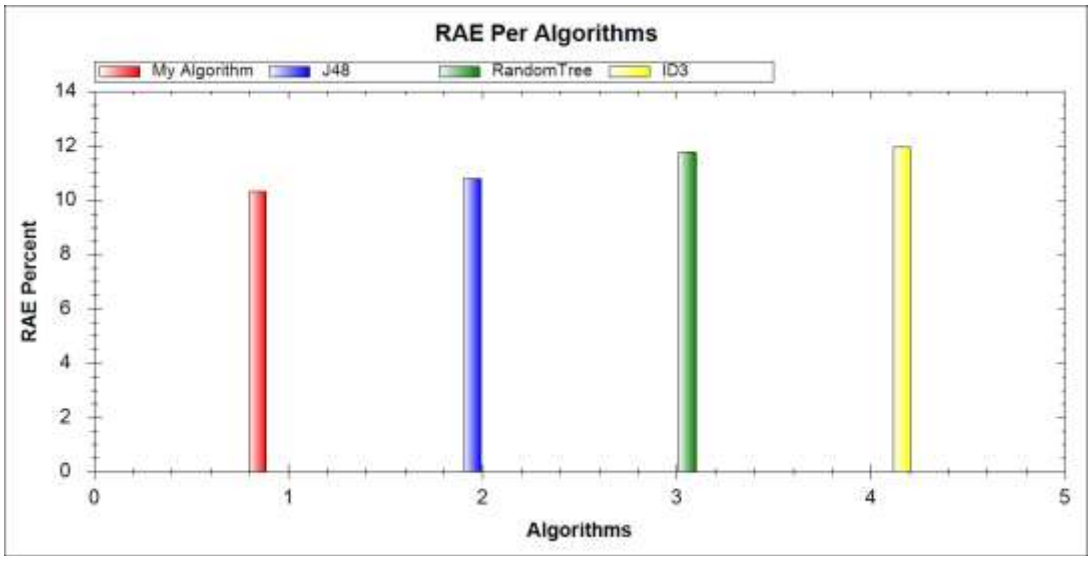


شکل ۱۰- میزان معیار MSE در میان الگوریتم پیشنهادی و دیگر الگوریتم‌های مشابه مورد بررسی



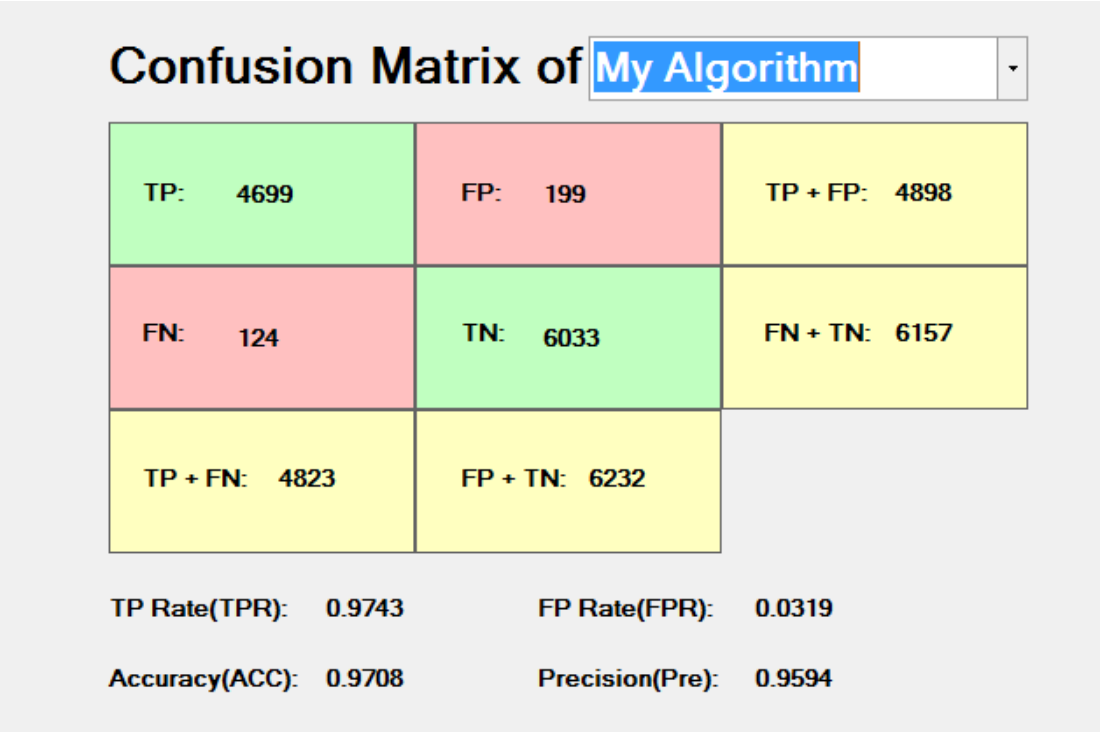
شکل ۱۱- میزان معیار MAE در میان الگوریتم پیشنهادی و دیگر الگوریتم‌های مشابه مورد بررسی

دوره ۱۰، شماره ۴، زمستان ۱۴۰۳، صفحات ۷۳-۵۱

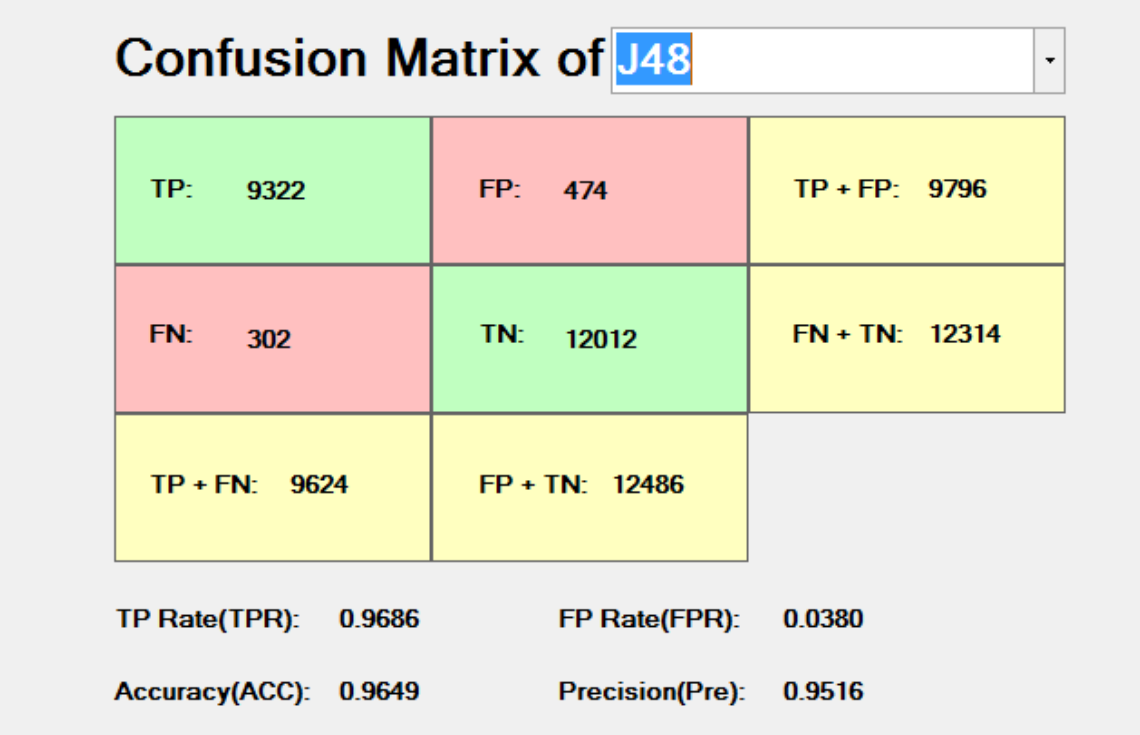


شکل ۱۲- میزان معیار RAE در میان الگوریتم پیشنهادی و دیگر الگوریتم‌های مشابه مورد بررسی

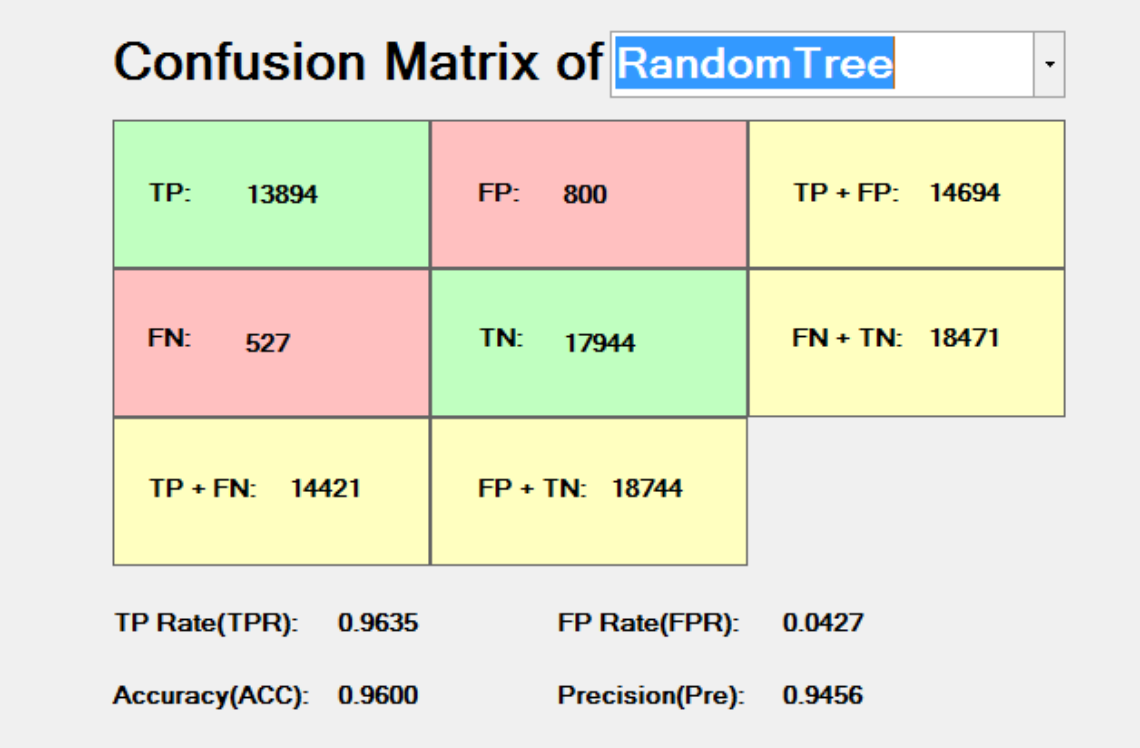
در ادامه Confusion Matrix مربوط به روش پیشنهادی و دیگر روش‌های مورد بررسی در این تحقیق نشان داده شده است.



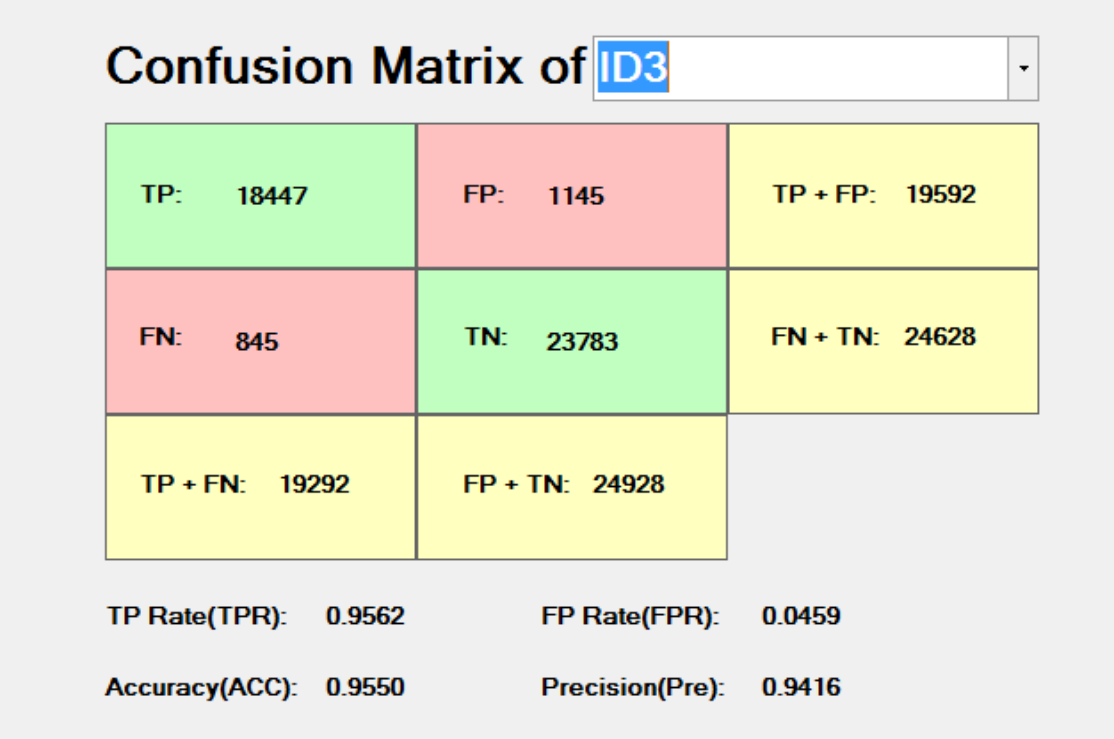
شکل ۱۳- جدول Confusion matrix روش پیشنهادی



شکل ۱۴- جدول Confusion matrix روش J48 مقاله [۱۰]



شکل ۱۵- جدول Confusion matrix روش Random tree مقاله [۱۲]



شکل ۱۶- جدول Confusion matrix روش ID3

در اینجا می‌توان مشاهده نمود که روش پیشنهادی دارای دقت می‌باشد زیرا در این جدول دارای مقدار TP و TN بیشتری از دیگر الگوریتم‌ها می‌باشد و در کنار آن نیز دارای FP و FN کمتری از دیگر الگوریتم‌های مورد بررسی در اینجا می‌باشد.
با توجه به ماتریس‌های Confusion می‌توان نرخ خطا و نرخ صحت عملکرد روش را محاسبه نمود که این روابط در رابطه ۲و۳و۴ نشان داده شده‌اند.

$$FPR = \frac{FP}{FP + TP} \quad (۲)$$

$$TPR = \frac{TP}{FP + TP} \quad (۳)$$

با توجه به جدول ۱ می توان دریافت که روش پیشنهادی دارای نرخ صحت کمتری می‌باشد ولی دارای کمترین نرخ خطا می‌باشد این بدین معنا است که روش پیشنهادی خطای بسیار کمی دارد بنابراین می‌توان گفت قطعا دارای دقت بیشتری می‌باشد.
با توجه به ماتریس‌های Confusion می‌توان دریافت که روش پیشنهادی ما دارای دقت بیشتری از دیگر روش‌ها می‌باشد زیرا دارای صحت(Accuracy) بیشتری از دیگر روش‌ها می‌باشد. صحت با استفاده از رابطه ۴ بدست می‌آید.

$$\boldsymbol{ACC = \frac{TP + TN}{TP + TN + FP + FN}} \quad (۴)$$

با توجه به این رابطه می‌توان مشاهده نمود که صحت پیش بینی‌ها در روش پیشنهادی ما از دیگر روش‌ها بیشتر می‌باشد. از طرفی دقت(Precission) به رابطه ۵ محاسبه می‌شود.

$$Precission = \frac{TP}{TP + FP} \quad (۵)$$

که می‌توان مشاهده نمود دقت در روش ما نیز از باقی روش‌ها بیشتر است.

پارامتر مهم دیگری که وجود دارد نمودار ROC می‌باشد. نمودار ROC دقت روش پیشنهادی را از جهت صحت عملکرد و کارایی روش نشان می‌دهد. در اینجا فضای زیر نمودار برای چهار روش در جدول نشان داده شده است.

جدول ۴- مقدار فضای زیر نمودار ROC برای الگوریتم پیشنهادی و روش‌های مورد بررسی

الگوریتم	فضای زیر نمودار ROC
روش پیشنهادی	۰,۹۹۳
الگوریتم J48 مقاله [۶]	۰,۹۸۹
الگوریتم RandomTree مقاله [۵]	۰,۹۸۸
الگوریتم ID3	۰,۹۴۰

با توجه به جدول می توان دریافت که روش پیشنهادی دارای دقت و صحت بیشتری می‌باشد زیرا هر چه فضای زیر نمودار ROC بزرگتر باشد نشان دهنده این قضیه می‌باشد نمودار از خط y=x بالاتر می‌باشد و در نمودار ROC هر چه بالاتر از y=x باشد نشان دهنده عملکرد بهتر الگوریتم می‌باشد. در این نمودار محور xها نشان دهنده TPR و محور yها نشان دهنده FPR می‌باشد یعنی هر چه نمودار از خط y=x بالاتر باشد دارای TPR بیشتری خواهد بود که این یعنی دارای دقت بالاتری است.

بحث و جمع بندی

مشاهده شد که روش ارائه شده در اینجا دارای عملکرد مناسبی بود و بهبود نسبتا بالایی را نسبت به الگوریتم مشابه خود یعنی ID3 نشان داده است. با استفاده از الگوریتم ارائه شده در اینجا می‌توان پی برد که چه پارامترهایی بیشترین تاثیر را در خروجی دارند و از این طریق می‌توان به پارامترهای دارای تاثیر بالاتر، توجه بیشتری نمود و از اینرو نرم افزارهایی تولید نمود که در برابر این حملات بهتر عمل کنند و با توجه به پارامترهایی که کم تاثیر می‌باشند آن‌ها را مورد بررسی قرار نداد تا اینکه سرعت بررسی افزایش یابد. به طور کلی یک سازمان می‌تواند با استفاده از این نرم افزار به نتایج بالایی دست یابد و با توجه به موفقیت روش پیشنهادی در مقابل روش‌های J48 مقاله [۱۰]، RandomTree مقاله [۱۲] و ID3، می‌توان اعتماد بالایی به روش پیشنهادی نمود و در نظر گرفت که این روش قادر است تا با سرعت بسیار بالا و همچنین دقت بالا همراه با خطای بسیار پایین پیش بینی را انجام دهد و به نتایج قابل قبولی دست یابد. در اینجا تنها از داده‌های ۱۱۰۵۵ نمونه سیستم مورد آموزش و تست قرار گرفت حال آن که اگر تعداد این نمونه‌ها بسیار بیشتر شود آنگاه نتایج بسیار بهتری نیز ایجاد خواهد شد. در این روش پیشنهادی به دلیل استفاده از راهکار انتخاب ویژگی های موثر مبتنی بر راهکار ترکیبی شبکه عصبی-ژنتیک و الگوریتم دسته ماهی و استفاده نکردن از پارامترهای بی اهمیت در نتیجه و همچنین کاهش ارتفاع درخت، درخت تصمیم قادر است تا با سرعت بالایی پیش بینی را انجام دهد و این خود می‌تواند یک نقطه عطف در نظر گرفته شود. با توجه به جداول بدست آمده از ارزیابی روش پیشنهادی در مقابل روش‌های مطرح دنیا همچون J48 مقاله [۱۰] و RandomForest مقاله [۱۲] می‌توان دریافت که روش پیشنهادی دارای TruePositive بسیار بالا و همچنین FalsePositive پایینی می‌باشد و این یعنی که این روش پیشنهادی دارای دقت بسیار

بالایی در پیش بینی‌ها می‌باشد و از طرفی با توجه به نمودار ROC که مبتنی بـر نـرخ TruePositive و نـرخ FalsePositive می‌باشد می‌توان به این نتیجه رسید که روش پیشنهادی دارای نسبت صحت بهتری از باقی روش‌ها می‌باشد یعنی رفتـه رفتـه هـر چه میزان آموزش با داده‌ها بیشتر شود نتایج بهتری نیز بدست خواهد آمد. در این بین روش J48 نیز دارای دقت مناسبی بـود کـه این از نتایج نیز قابل برداشت می‌باشد. در این روش پیشنهادی ارتفاع درخت با توجه به روش پیشنهادی به کمترین میـزان ارتفـاع تقلیل یافته که این خود باعث افزایش سرعت روش پیشنهادی می‌شود و می‌تواند به عنوان یک مزیت بسیار مناسبی در کاربردهـای مختلف مورد استفاده قرار گیرد.

پیشنهادهای آتی

- در این روش پیشنهادی از آنتروپی استفاده شده است ولی می‌توان از روش‌های دیگری نیز استفاده نمود و یا این روش را با روش‌های دیگری ادغام نمود. برای مثال در صورتی که این روش را با روشی مانند Gain که تابع ارزش می‌باشد ادغام نماییم، به احتمال زیاد دارای عملکرد بهتری می‌باشد چراکه آنتروپی نیز دارای معایبی می‌باشد ولی سرعت آن بالاست و در اینجا روشی مدنظر بود که دارای سرعت بالایی باشد ولی می‌توان با ادغام این روش و یا روش‌های جایگزین دقت این روش ارائه شده را به شدت افزایش داد.

- در اینجا مشاهده شد که درخت تصمیم‌گیری J48 نتایج نزدیکی را با الگوریتم تصمیم پیشنهادی در اینجا داشته است البته مطمئنا با افزایش تعداد داده‌های مورد آزمایش روش پیشنهادی بسیار بهتر عمل خواهد کرد و نتایج بهتری را ایجاد خواهد نمود ولی می‌توان ادغام روش پیشنهادی ناشی از عدم استفاده از پارامترهای با درجه اهمیت کمتر این درخت را ارتقا بسیاری داد و در حالتی که ممکن است از ID3 بهتر عمل کند البته نباید فراموش کرد که یکی از مشکلات درخت J48 عدم استفاده از داده‌های مختلف است بنابراین می‌توان با ایجاد ارتقایی در آن به شکلی که مانند درخت ID3 بتواند با نوع داده‌های مختلف عملکرد مناسبی داشته باشد، الگوریتمی را ایجاد نمود که دارای دقت بسیار بالاتری از روش‌های دیگر باشد.

فهرست منابع

1. -Korkmaz, M., Kocyigit, E., Sahingoz, O. K., & Diri, B. (۲۰۲۰). Deep Neural Network Based Phishing Classification on a High-Risk URL Dataset. In SoCPaR (pp. ٦٥٧-٦٤٨).
2. -Salem, O., Hossain, A., & Kamala, M. (۲۰۱۰, June). Awareness program and ai based tool to reduce risk of phishing attacks. In ۲۰۱۰ ۱۰th IEEE International Conference on Computer and Information Technology (pp. ١٤٢٣-١٤١٨). IEEE.
3. -Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (۲۰۲۰). Ai meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access, ٨, ١٤٢٥٤٢-١٤٢٥٣٢.
4. -Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. (۲۰۱۸). URLNet: Learning a URL representation with deep learning for malicious URL detection. arXiv preprint arXiv:١٨٠٢,٠٣١٦٢.
5. -Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (۲۰۰۸, November). An evaluation of machine learning-based methods for detection of phishing sites. In International Conference on Neural Information Processing (pp. ٥٤٦-٥٣٩). Springer, Berlin, Heidelberg.
6. -Khonji, M., Iraqi, Y., & Jones, A. (۲۰۱۳). Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, ١٥(٤), ٢١٢١-٢٠٩١.

7. -Cao, J., Li, Q., Ji, Y., He, Y., & Guo, D. (۲۰۱٦). Detection of forwarding-based malicious URLs in online social networks. International Journal of Parallel Programming, ٤٤(١), ١٨٠-١٦٣.

8. -Vazhayil, A., Harikrishnan, N. B., Vinayakumar, R., Soman, K. P., & Verma, A. D. R. (۲۰۱۸). PED-ML: Phishing email detection using classical machine learning techniques. In Proc. ١st AntiPhishing Shared Pilot ٤th ACM Int. Workshop Secur. Privacy Anal.(IWSPA) (pp. ٨-١). Tempe, AZ, USA.

9. -Subasi, A., Molah, E., Almkallawi, F., & Chaudhery, T. J. (۲۰۱۷, November). Intelligent phishing website detection using random forest classifier. In ۲۰۱۷ International conference on electrical and computing technologies and applications (ICECTA) (pp. ٥-١). IEEE.

10. A. Priya and E. Meenakshi, "Detection of phishing websites using c4. 5 data mining algorithm," in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1468 – 1472, IEEE, 2017

11. -Mondal, S., Maheshwari, D., Pai, N., & Biwalkar, A. (۲۰۱۹, December). A Review on Detecting Phishing URLs using Clustering Algorithms. In ۲۰۱۹ International Conference on Advances in Computing, Communication and Control (ICAC٣) (pp. ٦-١). IEEE.

12. YAZAN A. ALSARIERA, ABDULLATEEF O. BALOGUN, VICTOR E. ADEYEMO, OMAR H. TARAWNEH, HAMMED A. MOJEED, "INTELLIGENT TREE-BASED ENSEMBLE APPROACHES FOR PHISHING WEBSITE DETECTION", Journal of Engineering Science and Technology, Vol. 17, No. 1, pp. 0563 – 0582, 2022.

13. http://archive.ics.uci.edu/ml/datasets/Phishing+Websites.

۱۴ -امیری, نادری. (۲۰۲۰). استفاده از شبکه‌های خودرمزنگار موازی به منظور تشخیص صفحات جعلی اینترنتی. مجله نوآوری های فناوری اطلاعات و ارتباطات کاربردی, (۲), ۴۱-۵۰.

۱۵ -بهارلو, یاری. بهبود روش شناسایی وب سایت فیشینگ با استفاده از داده‌کاوی روی صفحات وب. دوفصلنامه فناوری اطلاعات و ارتباطات ایران, ۴۳(۴۳), ۲۷.

۱۶ -صف آرا, صباح نو. (۲۰۲۰). افزایش دقت شناسایی صفحات جعلی وب با استفاده از الگوریتم بهینه سازی کفتار و شبکه عصبی مصنوعی. فصلنامه علمی-پژوهشی فرماندهی و کنترل, ۳(۴), ۷۲-۹۱.

۱۷ -دادخواه مهدی, داورپناه جزی محمد, سعیدی مبارکه مجید. ارائه رویکردی به منظور شناسایی و پیش بینی وب سایتهای فیشینگ به وسیله الگوریتمهای کلاس بندی بر اساس مشخصه های صفحات وب.

۱۸ -کمالی زاده, سلمان, شاه‌محمدی. ارزیابی روش‌های شناسایی وب‌سایت فیشینگ. فصلنامه علمی پژوهش‌های اطلاعاتی و جنایی, ۱۱(۴۱), ۹-۳۸.

۱۹ -کتاب مدیریت کلان داده‌ها در بخش‌های خصوصی و عمومی - تألیف دکتر بابک سهرابی و حمیده ایرج - انتشارات سمت ۱۳۹۴

۲۰ -اولین کنفرانس بین المللی چشم اندازهای نو در مهندسی برق و کامپیوتر، تهران، کنفدراسیون بین المللی مخترعان جهان (IFIA)، دانشگاه جامع علمی کاربردی

۲۱ -احدزاده قناد, شیرین و فاطمه احمدی آبکناری، ۱۳۹۴، داده کاوی مشترکین شرکت مخابرات برای بهبود مدیریت ارتباط با مشتری، کنفرانس بین المللی پژوهش در مهندسی، علوم و تکنولوژی، استانبول، موسسه مدیران ایده پرداز پایتخت ویرا