نهمین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات
9th National Conference on Advanced Topics in Computer and Information Technology
۲۰ آذر ماه ۱۴۰۴

بررسی فایل‌های آلوده به هرزنامه در صندوق ورودی پست الکترونیکی با استفاده از الگوریتم فراابتکاری ملخ صحرا

پریسا نقاشیان[۱]، علیرضا صفائی[۲]
(۱) موسسه آموزش عالی پاسارگاد، شیراز، ایران
Parisa.Naghashian@gmail.com
(۲)گروه کامپیوتر، واحد ماهشهر، دانشگاه آزاد اسلامی، ماهشهر، ایران
alisafasafaei@yahoo.com

چکیده : امروزه فایل‌های آلوده به هرزنامه تبدیل به یک معضل برای کاربران و شرکت‌های حوزه فناوری اطلاعات و سازمان‌ها از لحاظ اجتماعی اقتصادی و زمانی تبدیل شده است و سالانه خسارات بسیار زیادی را به بار می‌آورد و حجم زیادی از ترافیک و پهنای باند اینترنت و منابع دیتاسنترها را به هدر داده و باعث ازدحام کار مسیریاب‌های شبکه می‌شود. از این رو با توجه به اهمیت تشخیص فایل‌های آلوده به هرزنامه از عادی و توانایی الگوریتم‌های داده‌کاوی و هوش مصنوعی در این تحقیق بر آن شدیم تا با تعیین مجموعه‌ای از تنظیمات موثر شبکه عصبی مصنوعی بوسیله الگوریتم فراابتکاری ملخ صحرا به تشخیص فایل‌های آلوده به هرزنامه از فایل‌های‌های عادی بر روی مجموعه داده Spambase و در محیط شبیه‌سازی متلب بپردازیم. نتایج بدست آمده نشان داده است که روش پیشنهادی با انتخاب ۲۳ ویژگی توانسته است با دقت ۹۹.۲۵ درصد برای داده‌های آموزشی و ۹۸.۲۳ درصد برای داده‌های آزمایشی به تشخیص فایل‌های آلوده به هرزنامه بپردازد و دقت را به میزان ۸.۲ درصد بهبود دهد. بنابراین نتایج بدست آمده از مقایسات صورت گرفته با سایر الگوریتم‌های فراابتکاری و کارهای پیشین حاکی از آن است که روش پیشنهادی در این تحقیق از عملکرد خوب و بالایی جهت تشخیص فایل‌های آلوده به هرزنامه برخوردار است.

کلمات کلیدی: هرزنامه- پست الکترونیکی- داده کاوی- الگوریتم فراابتکاری ملخ صحرا

۱ – مقدمه

ارتباطات همواره نقش مهمی در زندگی انسان‌ها داشته است از همین رو در ادوار گذشته تلاش‌های زیادی برای ایجاد یک بستر ارتباطی مطمئن و امن با کارایی بالا صورت گرفته است که این تلاش‌ها در سال‌های اخیر با ورود سیستم‌های کامپیوتری و دیجیتالی به عرصه زندگی انسان‌ها سرعت بیشتری پیدا کرده است به‌طوری که در حال حاضر روزانه میلیون‌ها نامه الکترونیکی در سراسر جهان رد و بدل می‌شود. استفاده از نامه الکترونیکی بعنوان یکی از اولین سرویس‌های اینترنت امروزه دارای جایگاه مستحکمی در ارتباطات است که مانند تلفن به صورت یک شیوه ارتباطی بسیار آسان در همه جا قابل دسترس می‌باشد. در این راستا یکی از چالش‌های مهم فضای اینترنتی و وب، فایل‌های آلوده به هرزنامه‌ها می‌باشند که با رشد فناوری اطلاعات و ارتباطات در فضای وب این چالش هر روز بیشتر محسوس می‌شود[۱و۲]. فایل‌های آلوده به هرزنامه برای اهدافی نظیر تبلیغات، مزاحمت، فریب کاربر، سرقت و برای سایر اهداف اقتصادی، اجتماعی و سیاسی در وب، برای کاربران ارسال می‌گردند. ارسال فایل‌های آلوده به هرزنامه اغلب به شکل گروهی و توسط نرم‌افزارهای ارسال کننده به تعداد زیادی پست الکترونیک ارسال می‌شوند[۳]. امروزه فایل‌های آلوده به هرزنامه تبدیل به یک معضل برای کاربران و شرکت‌های حوزه فناوری اطلاعات و سازمان‌ها از لحاظ اجتماعی، اقتصادی و زمانی تبدیل شده

است و سالانه خسارات بسیار زیادی را به بار می‌آورد و حجم زیادی از ترافیک و پهنای باند اینترنت و منابع دیتاسنترها را به هدر داده و باعث ازدحام کار مسیریاب‌های شبکه می‌شود. با وجود آنکه پست الکترونیک در توسعه فناوری اطلاعات و تجارت الکترونیک نقش مهمی را برعهده دارند اما چالشی نظیر فایل‌های آلوده به هرزنامه باعث شده است که کاربران وقت زیادی را برای مواجه با هرزنامه‌ها و مسدود کردن آنها به هدر دهند[۴]. بنابراین می‌توان گفت در عین گستردگی استفاده از پست الکترونیک طی سال‌های اخیر اما عمومیت و سادگی استفاده از پست الکترونیک باعث شده است که مورد استفاده هرزنامه‌نویسان و کلاهبرداران اینترنتی قرار گرفته شود[۵]. بنابراین باتوجه به اینکه هرزنامه به یک مشکل حیاتی تبدیل شده است نیاز به روشی جهت تشخیص فایل‌های آلوده به هرزنامه از فایل‌های عادی است بدین منظور توسط محققان طی سال‌های اخیر روش‌ها و الگوریتم‌های مختلفی مورد استفاده قرار گرفته است که یکی از این روش‌ها شناسایی و مقابله با پست‌های الکترونیکی هرزنامه با استفاده از الگوریتم‌های فراابتکاری و هوش مصنوعی می‌باشد. در ادامه به برخی از آنها بطور خلاصه بیان می‌شود. در سال ۲۰۲۲ فانی و همکاران روشی جدید برای شناسایی رایانامه‌های مخرب و فیشینگ بانکی با استفاده از الگوریتم‌های ژنتیک و جستجوی ممنوعه ارائه دادند. در مدل پیشنهادی آنها از الگوریتم‌های ترکیبی ژنتیک و جستجوی ممنوعه برای انتخاب بهترین ویژگی‌ها و از الگوریتم ماشین بردار پشتیبان برای طبقه‌بندی استفاده کردند. نتایج بدست آمده بر روی مجموعه داده SpamBase نشان داد دقت روش پیشنهادی برابر ۹۷.۲۸ درصد می‌باشد[۶]. در سال ۲۰۲۱ سلیمانیان و همکاران مدلی جهت تشخیص پست الکترونیک هرزنامه با استفاده از ترکیب الگوریتم بهینه‌سازی مغناطیسی با الگوریتم جستجوی هارمونی ارائه دادند. آنها با استفاده از الگوریتم بهینه‌سازی مغناطیسی، بهترین ویژگی‌ها را برای الگوریتم جستجوی هارمونی پیدا کرده و بوسیله الگوریتم نزدیکترین همسایه کار طبقه‌بندی را انجام دادند. نتایج بدست آمده بر روی مجموعه داده SpamBase نشان داد دقت روش پیشنهادی برابر ۹۴.۱۷ درصد می‌باشد[۷]. در سال ۲۰۲۰ سلیمانیان و همکاران مدلی جهت تشخیص پست الکترونیک هرزنامه با استفاده از الگوریتم‌های نزدیکترین همسایه و ازدحام گربه ارائه دادند. در مدل پیشنهادی آنها از الگوریتم ازدحام گربه برای انتخاب بهترین ویژگی‌ها و از الگوریتم نزدیکترین همسایه برای طبقه‌بندی استفاده کردند. نتایج بدست آمده بر روی مجموعه داده SpamBase نشان داد دقت روش پیشنهادی برابر ۹۷.۶۱ درصد می‌باشد[۸]. در سال ۲۰۱۹ Shuaib و همکاران جهت تشخیص هرزنامه‌های پست الکترونیک از الگوریتم نهنگ برای انتخاب بهترین ویژگی‌ها بر روی مجموعه داده SpamBase استفاده نمودند. آنها پس از انتخاب بهترین ویژگی‌ها از الگوریتم جنگل تصادفی جهت طبقه‌بندی استفاده کردند. نتایج بدست آمده از روش پیشنهادی نشان داد که قادر است با دقت ۹۹.۱۶ درصد در حالت انتخاب ویژگی و دقت ۹۴.۲ درصد بدون انتخاب ویژگی به تشخیص پست الکترونیک‌های هرزنامه بپردازد[۹].

۲ – پست الکترونیک

پست الکترونیک یا پست الکترونیکی یک آدرس پستی بصورت الکترونیکی در فضای مجازی می‌باشد که مسئولیت برقراری ارتباط دوسویه بین فرستنده و گیرنده پیام را دارد که بعنوان یکی از فناوری‌های برتر ارتباطی شناخته شده است. از مزایای پست الکترونیکی می‌توان به برقراری ارتباط دوسویه با سرعت بالا بین مخاطبان و شرکت‌ها اشاره نمود به نحوی که مخاطب از طریق آدرس پست الکترونیکی شرکت یا سازمان به ابراز عقاید و نظرات خود در مورد آن پیام می‌پردازد. ارسال پست الکترونیک از طریق سرویس‌های اینترنتی رایگان مانند جیمیل و یاهو به یک امر رایج جهت برقراری ارتباط بین مخاطبان تبدیل شده است به نحوی که افراد سرشناس، سیاستمداران، کارمندان و حتی مدیران مشاغل و کسب و کارها نیز برای ارسال پیام از این سرویس‌ها استفاده می‌کنند زیرا استفاده از این سرویس‌ها بسیار آسان بوده و خدمات خوبی به کاربران ارائه می‌دهد اما اطلاعات شخصی کاربران ممکن است لو رفته و در اختیار بنگاه‌های تبلیغاتی، سارقان دیتا و غیره قرار بگیرد. بنابراین می‌توان گفت در عین گستردگی استفاده از پست الکترونیک طی سال‌های اخیر، عمومیت و سادگی استفاده از پست الکترونیک سبب شده که مورد استفاده هرزنامه‌نویسان و کلاهبرداران اینترنتی قرار گرفته شود [۱۰،۱۱].

۳ – هرزنامه

هرزنامه به نامه‌های الکترونیکی ناخواسته گفته می‌شود که ممکن است با اهداف گوناگونی به صندوق پستی کاربران گوناگون فرستاده شوند، هرزنامه‌ها معمولا برای عده کثیری آزار دهنده هستند و باعث اتلاف زمان و افزایش تاخیرهای ارتباطی می‌شوند درحالی که فرستادن هرزنامه می‌تواند بدون محدودیت و با کمترین هزینه صورد

پذیرد. هرزنامه نویس‌ها[1] روزانه قادرند هزاران هرزنامه را با کمترین هزینه ارسال و حتی ممکن است این کار زمانی که باعث افزایش فروش محصولات آنها شود باعث کسب درآمدی نیز برای آنها گردد. پیشرفت سیستم‌های فیلترینگ هرزنامه نویس‌ها را مجبور می‌کند برای ادامه بقا مکانیزم‌های خود را برای دور زدن سیستم‌های فیلترینگ تغییر دهند که این امر باعث بوجود آمدن تقابل مدوام و پویایی میدان سیستم‌های فیلترینگ و هرزنامه نویس‌ها می‌شود؛ در نتیجه سیستم‌های فیلترینگ بایستی همواره به شناسایی این مکانیزم‌ها و مقابله با آنها بپردازند. پویایی ساختار هرزنامه‌ها از یک سو و قرار گرفت آنها در کنار سایر مخرب‌ها مانند ویروس‌ها، کرم‌های اینترنتی و از سوی دیگر امنیت ارتباطات اینترنتی را به طور جدی می‌توانند به مخاطره بیاندازند[۱۲]. هرزنامه‌ها بصورت انواع مختلفی در اینترنت گسترش می‌یابند که تعدادی از آنها را در ادامه بطور مختصر بیان می‌کنیم [۱۳،۱۴] :

- **هرزنامه پست الکترونیک** : هرزنامه پست الکترونیک که بنام پست الکترونیک فله‌ای ناخواسته نیز نامیده می‌شود عمل ارسال پیام‌های پست الکترونیکی بدون درخواست و به طور مکرر با محتوی تجاری در مقیاس بزرگ به دریافت کنندگان بدون استثنا قائل شدن می‌باشد. هرزنامه پست الکترونیک در اواسط ۱۹۹۰ زمانی که وارد اینترنت شد تبدیل به یک معضل گردید و روز به روز گسترش یافت به صورتی که امروزه با یک تخمین محافظه کارانه ۸۰ تا ۸۵ درصد پست الکترونیک‌ها را شامل می‌شود و در بعضی منابع از ۳۵ درصد بالاتر نیز می‌باشد.
- **هرزنامه پیام فوری** : هرزنامه پیام فوری[2] که به هرزنامه واژه‌ای مرکب از هرزنامه و پیام فوری معروف است از سیستم‌های پیام‌دهی آنی استفاده می‌کند، بسیاری از سیستم‌های پیام‌دهی آنی یک دایرکتوری کاربری شامل اطلاعات آماری را در اختیار یک ناشر برای جمع‌آوری اطلاعات، علامت‌گذاری روی سیستم و ارسال پیام‌های ناخواسته قرار می‌دهند. برای ارسال پیام‌های آنی به میلیون‌ها نفر از کاربران نیازمند یک نرم‌افزار با قابلیت ذخیره‌سازی و اسامی دریافت کنندگان است. ارسال کنندگان هرزنامه معمولا کانال‌های متداول چت اینترنتی را هدف قرار می‌دهند و با به کارگیری بوت‌های IRC که کانال‌ها را متصل و پیام‌ها را منتشر می‌کنند.
- **هرزنامه تلفن همراه** : هرزنامه تلفن همراه در سرویس پیام متنی ارسال می‌شود، ارسال هرزنامه به تلفن‌های همراه نه تنها می‌تواند موجب ناراحتی کاربران شده بلکه ممکن است هزینه‌هایی نیز برای کاربر و شبکه داشته باشد همچنین ممکن است هر پیام متنی دریافتی باعث تحریک کاربران شود. با وجود بالا بودن تعداد کاربران تلفن همراه هرزنامه هم بسیار زیاد شده است چرا که هزینه‌ای برای ارسال پیامک کمتر بوده و نصب بدافزارها از طریق پیام راحت‌تر است به نسبت به ارسال پست الکترونیک زیرا دانلود یک بدافزار باید حتما از سرور مرکزی ارائه کننده پست الکترونیک باشد ولی در پیامک چنین چیزی نیست و ممکن است کاربران مستقیما یک فایل مخرب را دانلود کنند.
- **هرزنامه وبلاگی** : هرزنامه بلاگی مخفف ارسال هرزنامه در وبلاگ‌هاست، در سال ۲۰۰۳ این نوع از هرزنامه با نوشتن مطالب تجاری در نرم‌افزارهای متحرک وبلاگی و با تکرار این توضیحات در پست‌های بلاگی ارسال می‌شود که معمولا یک لینک به وب‌سایت ارسال کننده هرزنامه است.
- **هرزنامه انجمنی:** هرزنامه انجمنی نوعی از هرزنامه است که در انجمن‌های گفتگو و پشتیبانی بسیار مشاهده شده و گاه به عنوان یک معضل و در جایی به عنوان سرگرمی تلقی می‌شود، در این نوع هرزنامه کاربران به جای پاسخ‌های مناسب از پاسخ‌های غیرمرتبط با موضوع تاپیک استفاده می‌کنند که منجر به منحرف شدن تاپیک مربوطه و به نتیجه نرسیدن آن خواهد شد.

۴ – بررسی روش‌های تشخیص هرزنامه

سرویس دهندگان خدمات پست الکترونیک برای اینکه بتوانند با ارسال هرزنامه مقابله کنند نیاز دارند تا هرزنامه را از پست‌های عادی تشخیص و تفکیک نمایند اهمیت این کار به حدی است که اگر اشتباهی در تشخیص روی دهد یا پست عادی به مقصد نمی‌رسد و یا صندوق دریافت مخاطب مملو از نامه‌های ناخواسته خواهد شد لذا قبل از مقابله با هرزنامه باید بتوان وجود آن را تشخیص داد. در ادامه انواع روش‌های تشخیص هرزنامه را بررسی می‌کنیم:

- **روش الگوی ترافیک:** در این روش الگوهای ترافیک هرزنامه‌ها مورد بحث و بررسی قرار می‌گیرد و بوسیله آزمایش یک سری از

[1] Spammer

[2] Instant messaging

ویژگی‌های پست‌های الکترونیکی و هرزنامه‌ها می‌توان بوسیله چند عنصر کلیدی مانند فرآیندهای وارد شدن پست‌های الکترونیکی، اندازه پست‌های الکترونیکی، تعداد دریافت کننده به ازاء هر پست الکترونیکی و همچنین محل دریافت پست الکترونیکی هرزنامه را از پست الکترونیکی واقعی تشخیص داد[۱۵].

- **استفاده از نتایج یادگیری مبتنی بر یادگیری ماشین:** این روش پیچیده‌تر و کاراتر از روش قبل می‌باشد که در این روش از الگوریتم‌های کشف قانون و کلاسه‌بندی بیز و درخت تصمیم‌گیری و غیره استفاده می‌شود. این روش دارای نرخ اشتباه پایینی است و اگر در کنار بقیه ضدهرزنامه‌ها قرار گیرد موثرتر خواهد بود. در این روش بر اساس یک سری از الگوریتم‌ها میزان اعتبار آدرس فرستنده مورد بررسی قرار می‌گیرد، در واقع هر آدرس پست الکترونیکی یک قالب مشخص دارد لذا از طریق قالب و روش‌های دیگر میزان اعتبار آن بررسی خواهد شد. برخی از تکنیک‌های یادگیری ماشین[3] جهت مقابله با هرزنامه شامل استفاده از روش‌هایی همچون بیزین، شبکه عصبی مصنوعی، ماشین بردار پشتیبان، درخت تصمیم‌گیری، K نزدیکترین همسایه و غیره می‌باشد.

۵ – انتخاب ویژگی

انتخاب ویژگی مرحله حساسی در بسیاری از مسائل طبقه‌بندی برای انتخاب زیرمجموعه بهینه از ویژگی‌ها است و سبب افزایش دقت یادگیری و کاهش زمان محاسباتی می‌شود که با عناوین دیگری همچون انتخاب متغیرها یا انتخاب مشخصه‌ها شناخته می‌شود و از سال ۱۹۹۶ تاکنون مورد توجه واقع شده است[۱۶]. در مجموعه داده‌ها همه ویژگی‌ها برای طبقه‌بندی مفید نیستند، ویژگی‌های وابسته و زائد هیچ اطلاعات اضافه درباره رده‌ها تهیه نمی‌کنند و اغلب تعدادی از ویژگی‌ها نوفه‌ای هستند. بنابراین روش‌های انتخاب ویژگی برای تعیین یک زیرمجموعه ویژگی بهینه استفاده می‌شوند بنابراین شناسایی ویژگی‌هایی که وابستگی زیادی با خروجی دارند مهم می‌باشند. انتخاب ویژگی، ویژگی‌هایی از مجموعه داده‌ها را که برای برای پیشگویی خروجی مؤثرتر هستند انتخاب می‌کند؛ در واقع انتخاب ویژگی فرآیند انتخاب زیرمجموعه‌ای از میان ویژگی هاست که به عنوان یک روش مهم در بازشناسی الگو و داده کاوی شناخته شده است. فرایند انتخاب ویژگی معمولا برای مسائلی مورد استفاده قرار می‌گیرد که داده‌ها از ویژگی‌های زیادی برخوردار هستند. بنابراین می‌توان گفت انتخاب ویژگی مسئله انتخاب یک زیرمجموعه از ویژگی‌ها برای توضیح هدفی کافی و ضروری است، زمانی که بردار ویژگی داده‌ها از تعداد زیادی ویژگی تشکیل شده باشد به دلیل وجود ویژگی‌های اضافی و غیرضروری دقت به دست آمده از مدل طبقه‌بندی کننده کمتر از دقتی خواهد بود که از یک زیرمجموعه مناسب برای یادگیری مدل استفاده شود. مشکل انتخاب یک زیرمجموعه از ویژگی‌ها جهت رسیدن به دقت بهینه یا دقتی نزدیک به دقت بهینه را می‌توان با استفاده از یک روش مناسب انتخاب ویژگی برطرف نمود لذا انتخاب ویژگی می‌تواند از یک سو به کاهش پیچیدگی محاسباتی منجر شود و از سوی دیگر به ایجاد شناخت بهتر از داده‌ها با کاهش ابعاد ویژگی کمک می‌کند[۱۷]. بطور کلی الگوریتم‌های انتخاب ویژگی که براساس معیارهای مختلف ارزیابی طراحی شده‌اند به دو گروه اصلی روش‌های فیلترکننده و پوشش دهنده تقسیم می‌شوند.

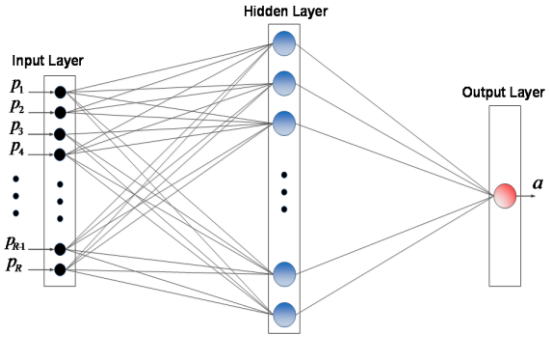
۶ –شبکه عصبی مصنوعی[4]

در سال ۱۹۴۳ شبکه‌های عصبی مصنوعی برای اولین بار توسط مک کولاک و پیتس معرفی شدند [۱۸]. شبکه عصبی مصنوعی که از سیستم عصبی زیستی الهام گرفته شده است ایده‌ای برای پردازش اطلاعات است که همانند مغز انسان به پردازش اطلاعات می‌پردازد. شبکه‌های عصبی مصنوعی که جهت مسائل طبقه‌بندی و شناسایی الگو طراحی شده‌اند، زیر شاخه‌ای از هوش مصنوعی هستند که قادر به درک ارتباطات و کشف نظم بین مجموعه‌ای از متغیرها می‌باشند. شبکه عصبی مصنوعی از اجزای سازنده‌ای همچون لایه‌ها و وزن‌ها تشکیل شده است به نحوی که رفتار این شبکه‌ها به ارتباط بین این اجزا وابسته می‌باشد. در حالت کلی می‌توان گفت شبکه‌های عصبی مصنوعی دارای سه نوع لایه هستند که وظیفه آنها دریافت داده‌ها، پردازش و تولید کمیت خروجی را به عهده دارند و عبارتند از [۱۹]:

- **لایه ورودی**[5] وظیفه‌ای شبیه به دندریت‌ها در شبکه عصبی زیستی دارد و تعداد نرون‌های آن با متغیرهایی که برای شبیه‌سازی یک مدل مورد نیاز است برابر می‌باشد.

[3] Machine Learning
[4] Artificial Neural Network (ANN)
[5] Input Layer

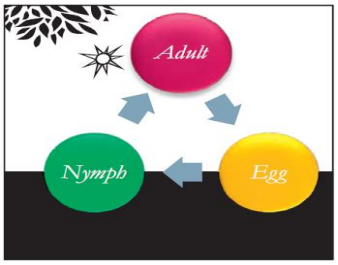
- **لایه‌های میانی (مخفی یا پنهان)**[6] معادل بدنه هسته[7] در شبکه عصبی زیستی هستند و تعداد نرون‌های این لایه‌ها بوسیله آزمون و خطا مشخص می‌گردد. با استفاده از ورودی‌ها و مقادیر وزن بین این لایه‌ها نحوه عملکرد لایه‌های پنهان تعیین می‌شود. فعال و غیرفعال بودن هر واحد پنهان بوسیله وزن‌های بین واحدهای ورودی و پنهان تعیین می‌شود.
- **لایه خروجی**[8] نتیجه نهایی شبکه را نشان می‌دهد. عملکرد واحد خروجی به فعالیت واحد پنهان و وزن ارتباط بین واحد پنهان و خروجی بستگی دارد. در شکل (۱) یک مدل ساده از شبکه‌های عصبی نمایش داده شده است.



شکل ۱: یک مدل ساده از شبکه‌های عصبی مصنوعی

۷ –الگوریتم بهینه سازی ملخ صحرا

الگوریتم بهینه‌سازی ملخ در سال۲۰۱۷ معرفی شده است که یک الگوریتم فراابتکاری الهام گرفته شده از طبیعت می‌باشد که رفتار ملخها در طبیعت و حرکت گروهی ملخها به سمت منابع غذایی را تقلید کرده و آن را شبیه‌سازی می‌کند. ملخها جزو دسته حشرات هستند و به دلیل صدمه زدنشان به محصولات کشاورزی به عنوان یک آفت شناخته شده‌اند. طول عمر ملخها در شکل (۲) نشان داده شده است[۲۰].



شکل ۲: چرخه حیات ملخها

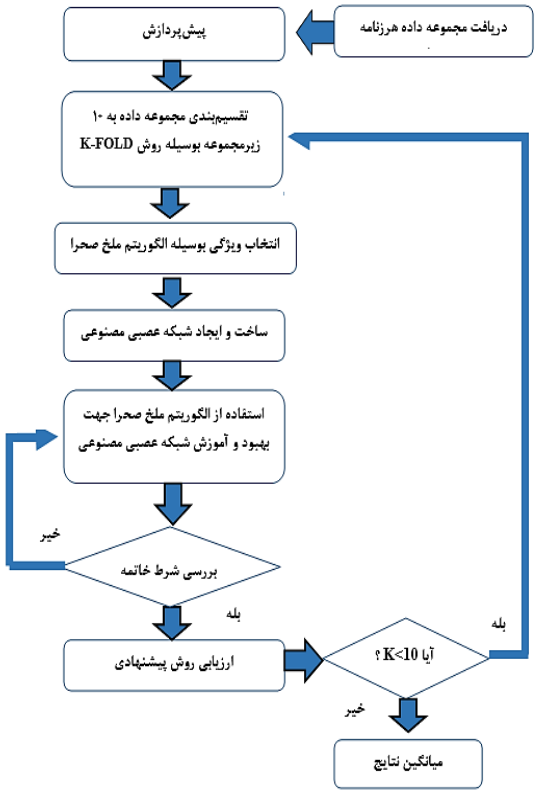
اگرچه ملخها در طبیعت معمولاً به صورت فردی دیده می‌شوند، جزو یکی از بزرگ‌ترین گروه‌ها در بین کل موجودات هستند و اندازه این گروه‌ها می‌تواند کابوسی برای کشاورزان باشد. جنبه منحصر به فرد گروه ملخها این است که رفتار زندگی گروهی آنها را می‌تواند هم در بین ملخ‌های بالغ و هم ملخ‌های نوزاد یافت. میلیون‌ها ملخ نوزاد مانند استوانه‌های در حال چرخش، پریده و حرکت می‌کنند و در مسیر حرکت، تقریباً همه گیاهان را می‌خورند. پس از این رفتار، زمانی که آنها بزرگ‌تر می‌شوند، گروهی را در هوا تشکیل می‌دهند و این همان چگونگی مهاجرت ملخها در فواصل طولانی است. مشخصه اصلی این گروه‌ها در مرحله لارو، حرکت آهسته و گام‌های کوچک ملخها می‌باشد. در مقابل حرکت طولانی مدت و ناگهانی نیز مشخصه اساسی این گروه‌ها در بین ملخ‌های بالغتر می‌باشد. جستجوی منابع غذایی مشخصه مهم زندگی گروهی در بین ملخها می‌باشد. الگوریتم‌های الهام گرفته از طبیعت، فرآیند جستجو را به دو بخش تقسیم می‌کنند اکتشاف و بهره‌برداری. در اکتشاف، عامل‌های جستجو به حرکت ناگهانی تشویق می‌گردند، در حالی که در طول بهره برداری تمایل دارند تا به صورت محلی حرکت کنند. این دو عملکرد و نیز جستجوی هدف توسط ملخها به طور غریزی انجام می‌گیرد. بنابراین، اگر روشی را برای مدل‌سازی ریاضیاتی این رفتار بیابیم می‌توانیم یک الگوریتم جدید الهام گرفته از طبیعت طراحی کنیم.

۸ –پیاده سازی روش پیشنهادی

در این تحقیق بر آن شدیم تا به ارائه روشی جهت تشخیص ایمیل‌های هرزنامه از عادی براساس الگوریتم بهینه‌سازی ملخ صحرا و شبکه عصبی مصنوعی بپردازیم. ایده اصلی روش پیشنهادی استفاده از الگوریتم شبکه عصبی مصنوعی برای طبقه‌بندی می‌باشد. فلوچارت روش پیشنهادی در شکل (۳) نشان داده شده است. همانطور که مشاهده می‌شود پس از دریافت مجموعه داده هرزنامه از مخزن مجموعه داده‌های دانشگاه کالیفرنیا[9] که شامل ۴۶۰۱ نمونه یا ایمیل می‌باشد، پیش‌پردازش بر روی داده‌های آن صورت گرفته و در ادامه پس از پیش‌پردازش‌های مورد نیاز (نرمال‌سازی) با توجه به مجموعه داده مذکور داده‌ها تقسیم‌بندی شده و شبکه عصبی مصنوعی طراحی می‌شود. سپس ویژگی‌های موثر با استفاده از الگوریتم بهینه‌سازی ملخ صحرا از بین تمامی ویژگی‌های موجود در مجموعه داده انتخاب شده و شبکه عصبی مصنوعی با استفاده

[6] Hidden Layer
[7] Nucleus
[8] Output Layer
[9] https://archive.ics.uci.edu/ml/datasets/spambase

از زیرمجموعه ویژگی‌های منتخب و اوزان و بایوس مناسبی که توسط الگوریتم بهینه‌سازی ملخ صحرا تعیین شده است مورد آموزش قرار داده می‌شود. در نهایت پس از بررسی شرط خاتمه در صورتی که به شرط موردنظر رسیده باشیم روش پیشنهادی مورد بررسی و ارزیابی قرار گرفته می‌شود.



شکل ۳: فلوچارت روش پیشنهادی

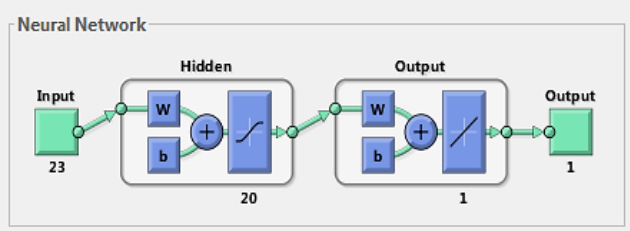
در این تحقیق جهت نرمال‌سازی داده‌ها در بازه [1,0] از تکنیک کمینه-بیشینه طبق رابطه ۱ استفاده می‌کنیم :

$$xn_1 = \frac{x_1 - min_\alpha}{max_\alpha - min_\alpha} \qquad (۱)$$

در این رابطه، xn_1 مقدار نرمال‌سازی شده، x_1 مقدار داده ورودی اولیه، min_α مقدار حداقل داده ورودی و max_α مقدار حداکثر داده ورودی می‌باشد.

پس از اعمال نرمال‌سازی، مجموعه داده موردنظر را با روش 10-FOLD جهت آموزش و ارزیابی روش پیشنهادی به دو مجموعه داده آموزشی و ارزیابی تقسیم‌بندی شدند.

در شکل (۴) معماری شبکه عصبی مورد نظر نشان داده شده است.



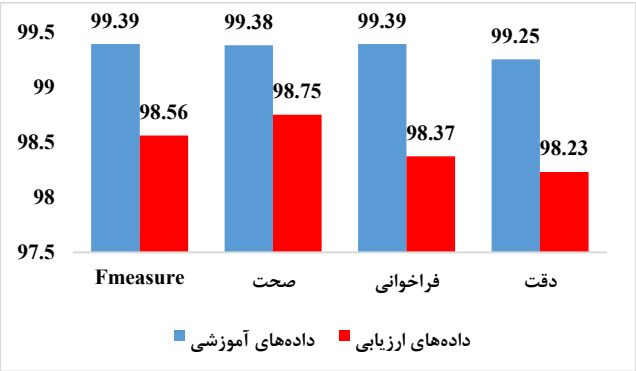
شکل ۴: معماری شبکه عصبی مصنوعی

۹ – معیارهای ارزیابی روش پیشنهادی

در این تحقیق برای سنجش و ارزیابی عملکرد روش پیشنهادی جهت تشخیص و طبقه‌بندی ایمیل‌های عادی از ایمیل‌های هرزنامه براساس ماتریس درهم‌ریختگی از معیارهای دقت، فراخوانی، صحت و F استفاده می‌شود.

۱۰- نتایج بدست آمده

در این تحقیق نتایج ارزیابی‌های به دست آمده از اجرای روش پیشنهادی با استفاده از مجموعه داده Spambase و معیارهای ارزیابی دقت، فراخوانی، صحت و F مورد بررسی قرار گرفته شده است که میانگین بدست آمده برای داده‌های آموزشی و ارزیابی با ۲۳ ویژگی در در شکل (۵) نشان داده شده است.



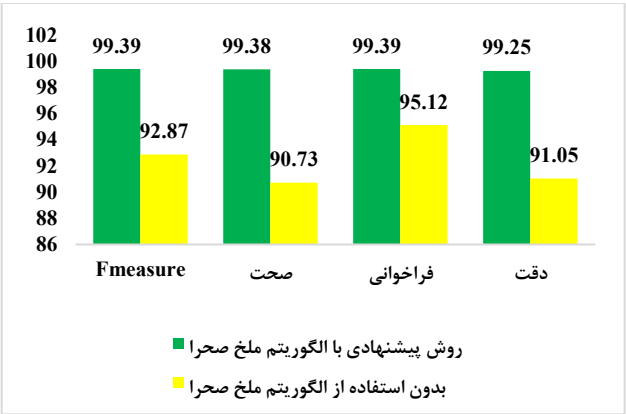
شکل ۵: نتایج بدست آمده برای داده های آموزشی

در ادامه روش پیشنهادی را بدون انتخاب ویژگی اجرا نمودیم. بنابراین پس از اجرای شبکه عصبی پیشنهادی با تمامی ۵۷ ویژگی موجود در مجموعه داده موردنظر، میانگین نتایج بدست آمده برای داده‌های آموزشی و ارزیابی در شکل (۶) نشان داده شده است.



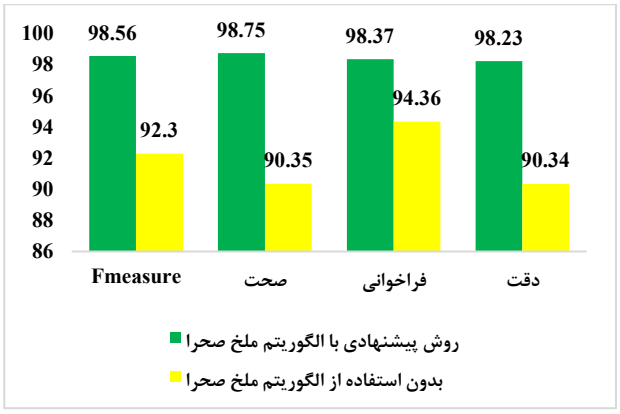
شکل ۶: نتایج بدست آمده برای داده‌های آزمایشی

در ادامه به مقایسه روش پیشنهادی با استفاده از انتخاب ویژگی و بهبود شبکه عصبی پیشنهادی بوسیله الگوریتم فراابتکاری ملخ صحرا و در حالتی که بدون انتخاب ویژگی و بهبود شبکه عصبی پیشنهادی برای داده‌های آموزشی و ارزیابی می‌باشد نشان داده شده است.



شکل ۷: مقایسه روش پیشنهادی برای داده‌های آموزشی

نتایج بدست آمده از پیاده‌سازی با استفاده از انتخاب ویژگی و بهبود شبکه عصبی پیشنهادی بوسیله الگوریتم فراابتکاری ملخ صحرا و استفاده از داده‌های آموزشی حاکی از آن است که روش پیشنهادی در این تحقیق توانسته است دقت را به میزان ۸.۲ درصد بهبود دهد همچنین معیارهای صحت، فراخوانی و FMeasure به ترتیب به میزان ۴.۲۷ درصد، ۸.۶۵ درصد و ۶.۵۲ درصد بهبود یافته‌اند.

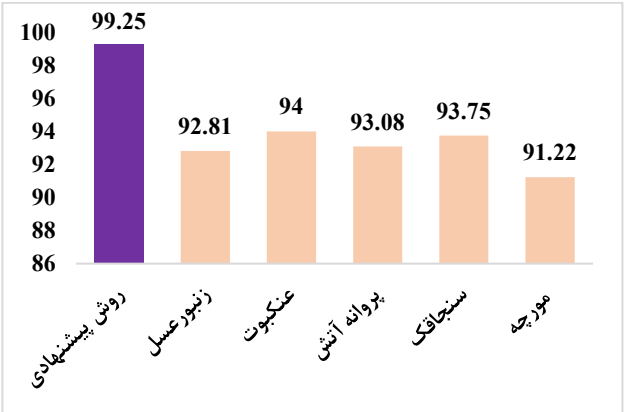


شکل ۸: مقایسه روش پیشنهادی برای داده‌های ارزیابی

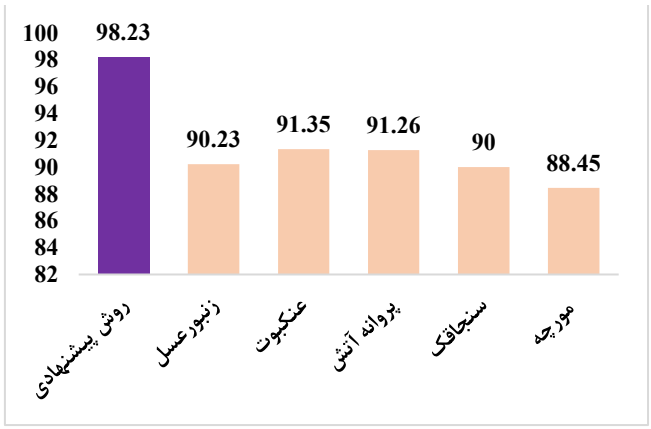
نتایج بدست آمده از پیاده‌سازی بدون استفاده از انتخاب ویژگی و بهبود شبکه عصبی پیشنهادی و با استفاده از داده‌های ارزیابی حاکی از آن است که روش پیشنهادی در این تحقیق توانسته است دقت را به میزان ۷.۸۹ درصد بهبود دهد همچنین معیارهای صحت، فراخوانی و FMeasure به ترتیب به میزان ۴.۰۱ درصد، ۸.۴ درصد و ۶.۲۶ درصد بهبود یافته‌اند.

۱۱- مقایسه دقت روش پیشنهادی با الگوریتم‌های مشابه

در ادامه نتایج بدست آمده از روش پیشنهادی را با الگوریتم‌های زنبورعسل، عنکبوت، پروانه آتش، سنجاقک و مورچه مورد بررسی قرار دادیم. برای این منظور الگوریتم‌های موردنظر ۲۳ ویژگی موثر را انتخاب کرده و شبکه عصبی مصنوعی را بهبود دادند. در شکل‌های (۹) و (۱۰) مقایسه دقت روش پیشنهادی با الگوریتم‌های ذکر شده برای داده‌های آموزشی و ارزیابی نشان داده شده است.



شکل ۹: مقایسه دقت روش پیشنهادی با سایر الگوریتم‌های فراابتکاری برای داده‌های آموزشی



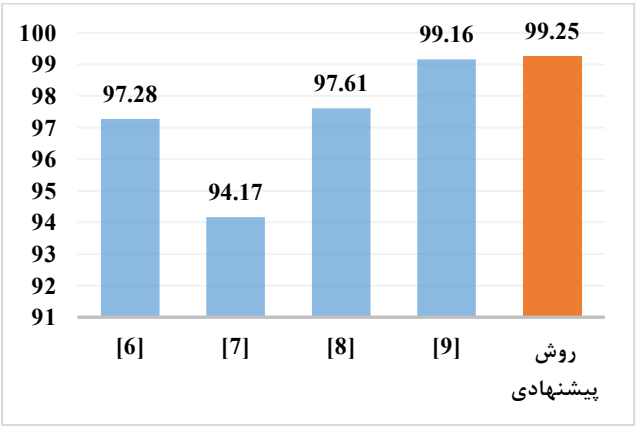
شکل ۱۰: مقایسه دقت روش پیشنهادی با سایر الگوریتم‌های فراابتکاری برای داده‌های ارزیابی

همانگونه که در شکل (۹) مشاهده می‌کنیم دقت روش پیشنهادی نسبت به الگوریتم زنبورعسل ۶.۴۴ درصد، نسبت به الگوریتم عنکبوت ۵.۲۵ درصد، نسبت به الگوریتم پروانه آتش ۶.۱۷ درصد، نسبت به الگوریتم سنجاقک ۵.۵ درصد و نسبت به الگوریتم مورچه ۸.۰۳ درصد بیشتر بوده و توانسته است با دقت بالاتری به تشخیص فایل‌های آلوده به هرزنامه از فایل‌های عادی بپردازد.

همچنین دقت روش پیشنهادی در شکل (۱۰) برای داده‌های ارزیابی نسبت به الگوریتم زنبورعسل ۸ درصد، نسبت به الگوریتم عنکبوت ۶.۸۸ درصد، نسبت به الگوریتم پروانه آتش ۶.۹۷ درصد، نسبت به الگوریتم سنجاقک ۸.۲۳ درصد و نسبت به الگوریتم مورچه ۹.۷۸ درصد بیشتر بوده و توانسته است با دقت بالاتری به تشخیص فایل‌های آلوده به هرزنامه از فایل‌های عادی بپردازد.

۱۲- مقایسه روش پیشنهادی با کارهای پیشین

در این بخش قصد داریم دقت بدست آمده از روش پیشنهادی در این تحقیق را با کارهای پیشینی که از مجموعه داده‌ای که ما در تحقیق خود از آن استفاده نمودیم، استفاده کرده‌اند مورد ارزیابی و مقایسه قرار دهیم. این مقایسه در شکل (۱۱) نشان داده شده است.



شکل ۱۱: مقایسه دقت روش پیشنهادی با کارهای پیشین

باتوجه به آنچه که نشان داده شده است دقت روش پیشنهادی در این تحقیق نسبت به دقت بدست آمده در مرجع [۶] به میزان ۱.۹۷ درصد، نسبت به مرجع [۷] به میزان ۵.۰۸ درصد، نسبت به مرجع [۸] به میزان ۱.۶۴ درصد و نسبت به مرجع [۹] به میزان ۰.۰۹ درصد بیشتر بوده و توانسته است با دقت بالاتری به تشخیص فایل‌های آلوده به هرزنامه از فایل‌های عادی بپردازد بنابراین می‌توان گفت روش پیشنهادی در این تحقیق نسبت به مقالات مورد مقایسه از عملکرد بالایی جهت تشخیص فایل‌های آلوده به هرزنامه برخوردار می‌باشد.

۱۳- نتیجه گیری

امروزه فایل‌های آلوده به هرزنامه تبدیل به یک معضل برای کاربران و شرکت‌های حوزه فناوری اطلاعات و سازمان‌ها از لحاظ اجتماعی اقتصادی و زمانی تبدیل شده است و سالانه خسارات بسیار زیادی را به بار می‌آورد و حجم زیادی از ترافیک و پهنای باند اینترنت و منابع دیتاسنترها را به هدر داده و باعث ازدحام کار مسیریاب‌های شبکه می‌شود. ز این رو با توجه به اهمیت تشخیص فایل‌های آلوده به هرزنامه از عادی و توانایی الگوریتم‌های داده‌کاوی و هوش مصنوعی در این تحقیق بر آن شدیم تا با تعیین مجموعه‌ای از تنظیمات موثر شبکه عصبی مصنوعی بوسیله الگوریتم فراابتکاری ملخ صحرا به تشخیص فایل‌های آلوده به هرزنامه از فایل‌های عادی بر روی مجموعه داده Spambase و در محیط شبیه‌سازی متلب بپردازیم. نتایج بدست آمده نشان داده است که روش پیشنهادی با انتخاب ۲۳ ویژگی توانسته است با دقت ۹۹.۲۵ درصد برای داده‌های آموزشی و ۹۸.۲۳ درصد برای داده‌های آزمایشی به تشخیص فایل‌های آلوده به هرزنامه بپردازد. بنابراین نتایج

بدست آمده از مقایسات صورت گرفته حاکی از آن است که روش پیشنهادی در این تحقیق از عملکرد خوب و بالایی جهت تشخیص فایل‌های آلوده به هرزنامه برخوردار است.

مراجع

[1] Yaseen, Q. (2021). Spam email detection using deep learning techniques. *Procedia Computer Science*, *184*, 853-858.

[2] Magdy, S., Abouelseoud, Y., & Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. *Computer Networks*, 108826.

[3] Saidani, N., Adi, K., & Allili, M. S. (2020). A Semantic-Based Classification Approach for an Enhanced Spam Detection. Computers & Security, 101716.

[4] Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). Phishing Email Detection Using Improved RCNN Model with Multilevel Vectors and Attention Mechanism. IEEE Access, 7, 56329-56340

[5] M. Ghorbanivand, F.S. Gharehchopogh, Spam Email Detection using Hybrid of Ant Colony Optimization and Adaboost, Information Technology in Engineering Design, in press, 12(2), 2019.

[6] فانی, ترابی, محمدامین, مقدم, & متینه. (۲۰۲۲). ارائه روشی جدید برای شناسایی ایمیل‌های مخرب و فیشینگ بانکی با استفاده از الگوریتم‌های ژنتیک و جستجوی ممنوع. *دوفصل نامه علمی ترویجی منادی امنیت فضای تولید و تبادل اطلاعات (افتا)*, ۹(۲), ۶۹-۸۰.

[7] سلیمانیان قره چپق, & سخی دل. (۲۰۲۱). یک مدل جدید برای تشخیص ایمیل هرزنامه با استفاده از ترکیب الگوریتم بهینه‌سازی مغناطیسی با الگوریتم جستجوی هارمونی. *دوفصل نامه علمی ترویجی منادی امنیت فضای تولید و تبادل اطلاعات (افتا)*, ۹(۱), ۵۰-۳۹.

[8] سلیمانیان قره چپق, & قربانی‌وند. (۲۰۲۰). بهبود عملکرد دسته‌بند k نزدیک‌ترین همسایه با الگوریتم بهینه‌سازی ازدحام گربه‌ها برای تشخیص ایمیل‌های هرزنامه. *فناوری اطلاعات در طراحی مهندسی*, ۱۲(۲).

[9] Shuaib, M., Adebayo, O. S., Osho, O., Idris, I., Alhassan, J. K., & Rana, N. (2019). Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification. SN Applied Sciences, 1(5), 390

[10] F.A. Hamzeh, F.S. Gharehchopogh, Feature Selection Based on Harmony Search Algorithm with Naive Bayes Algorithm to Spam Email Detection, Information Technology in Engineering Design, 11(2):31-46, 2019. (Persian)

[11] M. Ghorbanivand, F.S. Gharehchopogh, Spam Email Detection using Hybrid of Ant Colony Optimization and Adaboost, Information Technology in Engineering Design, in press, 12(2), 2019.

[12] L. Cranor and B. LaMacchia, "*Spam!*", Communications of the ACM, vol. 41, no.8, pp. 74–83, August 1998.

[13] Z. Liu, W. Lin, N. Li, and D. Lee, "Detecting and filtering instant messaging spam-a global and personalized approach," in *1st IEEE ICNP Workshop on Secure Network Protocols,2015(NPSec).* 2015, pp. 19-24.

[14] J. Bi, J. Wu, and W. Zhang, "A trust and reputation based anti-spim method"IN *INFOCOM 2018. The 27th Conference on Computer Communications. IEEE*, 2018.

[15] Liao, T., Stützle, T., de Oca, M. A. M., & Dorigo, M. (2014). A unified ant colony optimization algorithm for continuous optimization. *European Journal of Operational Research*, *234*(3), 597-609.

[16] P. Zhou, X. Hu, P. Li, X. Wu, "OFS-Density: A novel online streaming feature selection method, Pattern ecognition", Vol. 86, pp. 48-61, 2019.

[17] Z. S. Ageed *et al.*, "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Academic Journal,* vol. 1, no. 2, pp. 29-38, 2021.

[18] Nizamani, S., Memon, N., Glasdam, M., & Nguyen, D. D. (2014). Detection of fraudulent emails by employing advanced feature abundance. *Egyptian Informatics Journal*, *15*(3), 169-174.

[19] Kumar, R., Aggarwal, R. K., & Sharma, J. D. (2013). "Energy Analysis of a Building Using Artificial Neural Network: A Review". Energy And Buildings, 65, pp.352-358.

[20] Mirjalili, S. Z., Mirjalili, S., Saremi, S., Faris, H., & Aljarah, I. (2018). Grasshopper optimization algorithm for multi-objective optimization problems. Applied Intelligence, 48(4), 805-820.