مجله دانشکده بهداشت و انستیتو تحقیقات بهداشتی
بهار ۱۴۰۳، دوره ۲۲، شماره اول صفحه ۵۴ -۴۱

مقاله پژوهشی

تشخیص بیماری دیابت با استفاده از یادگیری ماشین و الگوریتم‌های تکاملی

مهرنوش آهنگرانی[۱]، محمد جعفر تارخ[۲*]

۱- دانشجوی دوره کارشناسی ارشد، گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران

۲- استاد، گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران

*نویسنده رابط : mjtarokh@kntu.ac.ir

تاریخ دریافت: ۱۴۰۳/۳/۲۶ تاریخ پذیرش: ۱۴۰۳/۶/۲۱

چکیده

زمینه و هدف: در سال‌های اخیر، یادگیری ماشین و الگوریتم‌های تکاملی توجه پژوهشگران و متخصصین در حوزه‌های مختلف، به ویژه حوزه سلامت را به جنبه‌های کاربردی آنها در پردازش مجموعه داده‌های کلان برای ارائه بینش‌های مفید به خود جلب کرده‌اند. از طرف دیگر، تشخیص سریع و دقیق بیماری دیابت یکی از مهم‌ترین مسائل در پزشکی است و افزایش نرخ ابتلا به این بیماری برای جوامع جهانی نگرانی‌های بسیاری را به همراه داشته است. مطالعه حاضر با هدف ایجاد یک مدل تشخیصی مبتنی بر الگوریتم‌های تکاملی و یادگیری ماشین جهت تشخیص بیماری دیابت انجام شد.

روش کار: این پژوهش یک چارچوب مبتنی بر تشخیص هوشمند بیماری دیابت را ارائه می‌دهد. روش پیشنهادی شامل دو مرحله اصلی ا ست: مرحله اول شامل رویکرد طبقه‌بندی با ا ستفاده از الگوریتم‌های K-نزدیک‌ترین هم سایه و جنگل ت صادفی ا ست. مرحله دوم شامل رویکرد ترکیبی انتخاب ویژگی و طبقه‌بندی به منظور بهبود نتایج مرحله اول است که در آن از الگوریتم‌های بهینه‌ساز گرگ خاکستری، بهینه ساز نهنگ و بهینه ساز ازدحام ذرات جهت انتخاب ویژگی ا ستفاده شده ا ست. در این تحقیق از مجموعه داده دیابت هندی پیما ا ستفاده شده است. تجزیه و تحلیل مقایسه‌ای بین رویکردهای مختلف از طریق شاخص‌های ارزیابی دقت، صحت و فراخوانی و امتیاز F1 انجام شده است.

نتایج: پس از مقایسه های تطبیقی بین مدل‌های پیشنهادی، مدل جنگل تصادفی مبتنی بر بهینه‌ساز گرگ خاکستری با صحت پیش‌بینی ۸۱/۳۸٪ به عنوان مدل نهایی انتخاب و معرفی شد.

نتیجه گیری: نتایج حاصل از این پژوهش نشان می‌دهد که استفاده از الگوریتم‌های تکاملی در کنار مدل‌های یادگیری ماشینی، می‌تواند کارایی و صحت تشخیص بیماری دیابت و عوارض ناشی از آن را در بیش‌تر مواقع افزایش دهد.

واژگان کلیدی: تشخیص دیابت، یادگیری ماشین، الگوریتم‌های تکاملی، انتخاب ویژگی

مقدمه

در بدن انسان، هورمونی وجود دارد به نام انسولین که توسط لوزالمعده ترشح می‌شود و به انتقال گلوکز از خون به سلول‌هایی که بعداً برای انرژی استفاده می‌شوند، کمک می‌کند. دیابت یک بیماری مزمن است و عوارض دراز مدت و کوتاه مدت ایجاد می‌کند که عوارض کوتاه مدت شامل کم آبی بدن و کمای دیابتی و عوارض طولانی مدت شامل حمله قلبی، نابینایی، نارسایی کلیه، سکته مغزی و زخم پا است. برای حل این مشکلات، استفاده از ابزارهای هوش مصنوعی به عنوان

روش‌های جدید و نوین در تشخیص دیابت مورد توجه قرار گرفته است. ماشینی شاخه‌ای از هوش مصنوعی است که به ما کمک می‌کند تا سیستم‌هایی را توسعه دهیم که بتوانند پیش‌بینی‌هایی را بر اساس دانش قبلی انجام دهند. امروزه هوش مصنوعی یکی از حوزه های پرطرفدار برای بررسی و کاربرد در زمینه پزشکی و سلامت است. در حقیقت، ابزارها و روش‌های قدیمی و سنتی به خوبی قادر به پاسخگویی به نیاز پزشکان و کاربران این حوزه نیستند. بنابراین، ظهور فناوری یادگیری ماشین و به طور کلی فناوری هوش مصنوعی کمک بزرگی برای این حوزه بوده است. از طرف دیگر، از آنجایی که در اغلب موارد، مجموعه داده‌های در دسترس دارای ویژگی‌های اضافه و نامربوط هستند، نیاز به الگوریتم‌هایی جهت بهینه‌سازی آنها به شدت دیده می‌شود. به این منظور، الگوریتم‌های تکاملی می‌توانند کارایی خوبی داشته باشند و در کاهش ابعاد ویژگی‌ها و به طور کلی، کاهش ابعاد مسئله جهت بهبود عملکرد طبقه‌بندی بسیار مفید واقع شوند.

پیشنهادی این تحقیق برای تشخیص و پیش‌بینی دیابت، استفاده از یک رویکرد ترکیبی از الگوریتم‌های تکاملی و یادگیری ماشین است. در این روش ترکیبی، از الگوریتم‌های تکاملی بهینه‌ساز گرگ خاکستری (Gray Wolf Optimization)، بهینه‌ساز نهنگ (Whale Optimization Algorithm) و بهینه‌ساز ازدحام ذرات (Particles Swarm Optimization) برای انتخاب ویژگی‌های مفید در مجموعه داده‌های سلامت به عنوان یک فرآیند بهینه‌سازی استفاده شده است. همچنین، از الگوریتم‌های یادگیری ماشین K-نزدیک‌ترین همسایه (K-Nearest Neighbors) و جنگل تصادفی (Random Forest) به منظور طبقه‌بندی داده‌ها و تشخیص بیماری دیابت استفاده شده است. در حقیقت، در این پژوهش از مدل‌های ترکیبی جدیدی استفاده شده است که در هیچ یک از تحقیقات پیشین، این ترکیب‌ها به صورت همزمان با هم به کار گرفته نشده‌اند. روش پیشنهادی روی مجموعه داده دیابت هندی پیما اعمال شده است. هدف اصلی این تحقیق، بررسی این موضوع است که چگونه الگوریتم‌های تکاملی می‌توانند برای بهینه‌سازی مسائل مرتبط با تشخیص بیماری‌ها استفاده شوند و چگونه علم نوین یادگیری ماشین می‌تواند در پیش‌بینی بیماری‌هایی مانند دیابت به کار گرفته شود.

عمدتاً سه نوع آزمایش سنتی برای تشخیص دیابت انجام می شود که این روش‌های تشخیص دیابت زمان‌بر و پرهزینه هستند و استفاده از آن‌ها در کشورهای کم درآمد دارای برخی محدودیت ها است. روش‌های یادگیری ماشینی بر روی داده‌های موجود اجرا می‌شوند و می‌توانند دیابت را پیش‌بینی کنند. این روش‌ها زمان صرف شده برای پردازش علائم و تشخیص بیماری را کاهش می‌دهند و تقریباً هزینه‌ای برای پیش‌بینی ندارند (۱). به همین منظور تحقیقات متعددی در این زمینه انجام شده است.

به عنوان نمونه، عبدالهادی و همکارانش به منظور تشخیص دیابت، از الگوریتم‌های جنگل تصادفی، رگرسیون لجستیک، تجزیه و تحلیل تشخیص خطی و طبقه‌بندی کننده رای استفاده کردند که الگوریتم جنگل تصادفی با دقت ۸۲٪ بهترین عملکرد را نشان داد (۲). Katarya و همکارانش نیز جهت پیش‌بینی دیابت، از الگوریتم‌های جنگل تصادفی، درخت تصمیم، رگرسیون لجستیک، بیزساده، K–نزدیک ترین همسایه و ماشین بردار پشتیبان استفاده کردند که به ترتیب دارای دقت‌های ۸۴٪، ۸۲/۷٪، ۷۶٪، ۷۵٪، ۷۵٪ و ۷۴٪ هستند و الگوریتم جنگل تصادفی بهترین عملکرد را در تشخیص دیابت نشان داده است (۳).

همچنین، اشاره‌ای دقیق‌تر به مقاله‌ای می‌کنیم که توسط Saxena و همکارانش نوشته شده است. در این مقاله، ارزیابی مقایسه‌ای بین عملکرد الگوریتم‌های یادگیری ماشین به منظور تشخیص بیماری دیابت صورت گرفته است. نکته‌ای که این مقاله را از سایر مقالات این دسته متمایز کرده است، استفاده از دو مجموعه داده و همچنین انتخاب ویژگی با کمک یک الگوریتم یادگیری ماشین به نام درخت اضافی است. مجموعه داده‌هایی که در این مقاله استفاده شده‌اند، شامل مجموعه داده دیابت هندی پیما و مجموعه داده‌های پیش‌بینی خطر دیابت در مراحل اولیه است. شاید به نظر برسد انتخاب ویژگی همیشه موجب افزایش دقت می‌شود؛ اما در این تحقیق، پس از اعمال روش انتخاب ویژگی روی مجموعه داده‌های پیش‌بینی خطر

دیابت در مراحل اولیه، متوجه شدند که هرگونه کاهش بیشتر ویژگی‌ها، باعث کاهش دقت پیش‌بینی در مقایسه با مجموعه داده کامل می‌شود. این موضوع نشان داد که ویژگی‌های انتخاب شده هنگام ایجاد مجموعه داده‌های پیش‌بینی خطر دیابت در مراحل اولیه در مقایسه با مجموعه داده‌های دیابت هندی پیما برای تشخیص صحیح دیابت اهمیت بیشتری دارد و تقریبا تمامی ۱۶ مورد ویژگی آن برای تشخیص بیماری دیابت حائز اهمیت هستند (٤). محققین این پژوهش پس از مرحله پیش پردازش داده‌ها و انتخاب ویژگی، داده‌ها را به منظور تشخیص دیابت با استفاده از الگوریتم‌های یادگیری ماشین رگرسیون لجستیک، تجزیه و تحلیل تشخیص خطی، ماشین بردار پشتیبان، جنگل تصادفی، بیز ساده، نزدیک‌ترین همسایه، ماشین تقویت گرادیان، طبقه‌بندی و درخت رگرسیون و درخت اضافی طبقه‌بندی کردند. نتایج نشان دادند که برای طبقه‌بندی مجموعه داده‌های دیابت هندی پیما، الگوریتم‌های ماشین بردار پشتیبان و رگرسیون لجستیک بالاترین عملکرد را با دقت ۸٤ ٪ داشتند و برای طبقه‌بندی مجموعه داده‌های پیش‌بینی خطر دیابت در مراحل اولیه، الگوریتم‌های نزدیک‌ترین همسایه، درخت اضافی و ماشین تقویت گرادیان بالاترین عملکرد را با دقت ۹۷ ٪ داشتند (٤).

استفاده مستقیم از داده‌ها می‌تواند بر عملکرد سیستم تأثیر بگذارد. از این‌رو، انتخاب ویژگی‌های مفید می‌تواند تأثیر بیشتری بر عملکرد و نتایج یک سیستم تشخیص در پیش‌بینی داشته باشد و ممکن است در زمان صرفه‌جویی کند. به این منظور، محققان زیادی از الگوریتم‌های تکاملی جهت انتخاب ویژگی و بهینه‌سازی مدل استفاده کرده‌اند که الگوریتم‌های به کاربرده شده توسط برخی از آن‌ها در جدول ۱ قابل مشاهده است. از آنجایی استفاده از الگوریتم‌های ترکیبی در اغلب موارد نتایج را بهبود می‌دهد و امروزه بسیار مورد توجه محققان قرار گرفته است، تمرکز اصلی این مقاله بر استفاده از ترکیب الگوریتم‌های تکاملی با یادگیری ماشین به منظور تشخیص بیماری دیابت است. به این منظور، مقالات متعددی مورد مطالعه قرار گرفت که در جدول ۲ قابل مشاهده است.

به عنوان نمونه، اشاره‌ای دقیق‌تر به مقاله‌ای می‌کنیم که توسط Kumar و همکارانش، نوشته شده است. در این مقاله، محققان ابتدا از الگوریتم یادگیری ماشین بیز ساده برای طبقه‌بندی همه ویژگی‌های مجموعه داده دیابت هندی پیما و سپس از الگوریتم تکاملی ژنتیک به عنوان روشی برای انتخاب ویژگی‌های مفید این مجموعه داده استفاده کردند؛ که به وسیله آن چهار ویژگی از هشت ویژگی این مجموعه داده انتخاب شد. سپس، مجددا از الگوریتم یادگیری ماشین بیز ساده برای طبقه‌بندی ویژگی‌های انتخاب شده استفاده کردند. طبق نتایج تجربی، عملکرد الگوریتم ترکیبی پیشنهادی از نظر شاخص‌های مختلف ارزیابی عملکرد نسبت به عملکرد الگوریتم بیز ساده به صورت غیر ترکیبی، بهتر است. به عنوان نمونه، دقت عملکرد مدل تشخیص دیابت با الگوریتم بیز ساده، ۷٦,۹۵٪ است و دقت عملکرد این مدل در ترکیب با الگوریتم تکاملی ژنتیک، ۷۸,٦۹٪ است (٦ ، ۵).

روش کار

در روش پیشنهادی، Gray Wolf Optimization (GWO)، Whale Optimization Algorithm (WOA) و Particles Swarm Optimization (PSO) به عنوان روش‌های انتخاب ویژگی و K–نزدیک‌ترین همسایه و جنگل تصادفی به عنوان روش طبقه‌بندی در مجموعه داده دیابت هندی پیما استفاده شده‌اند. روش پیشنهادی شامل مراحل کلی زیر است:

۱. مرحله اول شامل پیش پردازش مجموعه داده دیابت هندی پیما است.

۲. مرحله دوم شامل طبقه‌بندی با استفاده از K–نزدیک‌ترین همسایه و جنگل تصادفی در مجموعه داده دیابت هندی پیما است.

۳. مرحله سوم شامل رویکرد ترکیبی انتخاب ویژگی و طبقه-بندی به منظور بهبود نتایج مرحله اول است که در آن از الگوریتم‌های بهینه‌ساز گرگ خاکستری، بهینه‌ساز نهنگ و بهینه-ساز ازدحام ذرات جهت انتخاب ویژگی استفاده شده است و سپس مجددا از الگوریتم‌های K-Nearest Neighbors (KNN) و Random Forest (RF) جهت طبقه‌بندی استفاده شده است.

مجموعه داده دیابت هندی پیما: در این تحقیق از مجموعه داده دیابت هندی پیما استفاده شده است که یکی از محبوب-ترین مجموعه داده ها در زمینه این تحقیق است. این مجموعه داده دارای ۹ ستون (هشت ویژگی با ارزش عددی و یک کلاس) و ۷۶۸ ردیف (۵۰۰ غیر دیابتی و ۲۶۸ دیابتی) است. متغیر نتیجه، طبقه‌بندی باینری مقادیر ۰ یا ۱ را می‌گیرد که در آن ۰ نشان دهنده آزمایش منفی برای دیابت و ۱ نشان دهنده یک آزمایش مثبت است (۷). در حقیقت، از افراد مورد هدف این مجموعه داده آزمایش‌هایی مطابق با استانداردهای پزشکی گرفته شده است و وضعیت سلامت آنها به لحاظ دیابتی بودن مورد بررسی قرار گرفته است. در نهایت، افرادی که دیابت داشتند با عدد یک و افراد سالم با عدد صفر در مجموعه داده مشخص و برچسب گذاری شدند.

مجموعه داده هیچ مقدار تهی و هیچ مقدار گمشده‌ای ندارد؛ اما در ویژگی های عملکرد شجره‌نامه دیابت، سن، انسولین، گلوکز، شاخص توده بدنی و فشار خون نقاط پرت وجود دارد. شایع‌ترین علت وجود نقاط پرت عبارتند از خطای انسانی، خطای اندازه گیری، خطای ناشی از آزمایش، خطای نمونه گیری و خطای پردازش داده ها. تشخیص نقاط پرت یکی از جنبه‌های مهم در طراحی مدل‌های یادگیری ماشین است؛ زیرا بر دقت مدل یادگیری ماشین تأثیر می‌گذارد (٤). بنابراین، بهتر است که داده‌ها را استاندارد کرد تا از اثرات نامطلوب موارد پرت جلوگیری شود.

پیش پردازش داده‌ها: اولین مرحله پیش پردازش داده‌ها، عادی‌سازی داده‌های خام است. عادی‌سازی فرآیندی است که در آن داده‌های خام به یک محدوده عددی خاص تغییر مقیاس می‌دهند. این کار بهبود قابلیت مقایسه و تفسیر داده‌ها را ممکن می‌سازد. انواع مختلفی از عادی‌سازی وجود دارد که از میان تکنیک‌های مختلف آن، نرمال‌سازی استاندارد و نرمال‌سازی Min-Max از تکنیک‌های پرکاربرد هستند (٤). در این تحقیق از تکنیک Min-Max استفاده شده است؛ به این معنا که مقیاس داده‌ها به بازه‌ی خاص (۱، ۱-) تبدیل شد. رابطه (۱) روش کلی اعمال این تکنیک را نشان می‌دهد که در آن X' داده نرمال شده، X داده اولیه، min(X) کمترین مقدار داده‌ها و min(X) بیشترین مقدار داده‌ها است.

$$X' = \frac{X - min(X)}{max(X) - min(X)}$$ رابطه (۱)

الگوریتم‌های تکاملی و انتخاب ویژگی: فراوانی ویژگی‌های اضافی و نامربوط در مجموعه داده‌های پزشکی مدرن، کارایی تکنیک‌های داده کاوی را کاهش می‌دهد که منجر به نتایج غیرقابل تفسیر می‌شود. با این حال، مدل‌ها با انتخاب ویژگی مناسب ممکن است نتایج قابل تفسیر و دقیقی را به همراه داشته باشند. این امر نیاز به مرحله پیش پردازش در داده کاوی را برجسته می‌کند (۸). در حقیقت، دقت طبقه‌بندی به شدت به ماهیت ویژگی‌های یک مجموعه داده بستگی دارد که ممکن است حاوی داده‌های نامربوط یا اضافی باشد. هدف اصلی انتخاب ویژگی حذف این نوع ویژگی‌ها برای افزایش دقت طبقه‌بندی است (۹). بنابراین بهینه‌سازی ویژگی‌های مجموعه داده‌های مورد استفاده محققان، به ویژه فعالان حوزه سلامت، بسیار حائز اهمیت است؛ چراکه نتایج به دست آمده توسط آن-ها، مستقیما روی سلامت و زندگی انسان‌ها تاثیر می‌گذارد. به همین منظور، در این تحقیق از الگوریتم‌های تکاملی GWO، WOA و PSO به منظور انتخاب ویژگی استفاده شده است که در ادامه به طور خلاصه به معرفی این الگوریتم‌ها پرداخته می‌شود.

الگوریتم بهینه‌سازی گرگ خاکستری: بهینه ساز گرگ خاکستری یکی از روش‌های بهینه‌سازی الهام گرفته از طبیعت است که فرآیند شکار گرگ‌های خاکستری را در طبیعت شبیه-سازی می‌کند (۱۰). در الگوریتم پیشنهادی، گرگ‌های قوی‌تر با گرگ‌های ضعیف‌تر بر اساس سطح آمادگی جسمانی جایگزین می‌شوند. در هر تکرار الگوریتم، درجه تناسب محاسبه می‌شود و در صورت بهبود، الگوریتم دوباره تکرار می‌شود؛ در غیر این صورت، الگوریتم خاتمه می‌یابد. به عبارت دیگر، الگوریتم گرگ خاکستری برگرفته از زندگی اجتماعی گرگ‌های خاکستری است. چهار نوع گرگ خاکستری شامل آلفا، بتا، دلتا و امگا برای شبیه سازی سلسله مراتب رهبری استفاده می‌شوند. همچنین سه مرحله اصلی شکار یعنی جستجوی طعمه، محاصره طعمه و حمله به طعمه برای انجام بهینه‌سازی استفاده می‌شوند. رهبران گروه، زن و مردی به نام آلفا هستند و تصمیمات آلفا به گروه ابلاغ می‌شود. کمترین درجه گرگ خاکستری، امگا است که نقش قربانی را بازی می‌کند. به نظر

می‌رسد امگا فرد مهمی در گروه نیست، اما از طرفی اگر امگا از دست برود، کل گروه با جنگ داخلی و مشکلات مواجه می‌شوند. بنابراین، امگا به حفظ کل گروه و ساختار سلسله مراتبی کمک می‌کند. اگر گرگی آلفا، بتا یا امگا نباشد به او دلتا می‌گویند. گرگ‌های دلتا باید به آلفا و بتا تسلیم شوند، اما آن‌ها بر امگا تسلط دارند. (۱۱). در مدل‌های ریاضی این الگوریتم، مناسب‌ترین راه حل آلفا نامیده می‌شود. راه‌حل‌های دوم و سوم، بتا و دلتا نام دارند. به ترتیب. بقیه راه‌حل‌های کاندید به عنوان امگا فرض می‌شوند. برای اینکه گله بتواند طعمه‌ای را شکار کند، ابتدا آن را محاصره می‌کند. به منظور مدل‌سازی ریاضی رفتار محاصره‌ای، معادلاتی در نظر گرفته شده است (۱۲). به طور کلی محققان در این الگوریتم، سلسله مراتب رهبری و مکانیسم شکار گرگ‌های خاکستری در طبیعت را برای مسائل بهینه‌سازی شبیه‌سازی می‌کنند. این روش با جمعیتی از گرگ‌های خاکستری به عنوان عوامل جستجو شروع می‌شود. سپس در هر تکرار، جامعه به صورت تصادفی به روز می‌شود.

الگوریتم بهینه‌سازی نهنگ: این الگوریتم، یک الگوریتم فراابتکاری است که از جستجوی نهنگ‌های گوژپشت تقلید می‌کند و متعلق به خانواده الگوریتم‌های مبتنی بر جمعیت تصادفی است. در اطراف نهنگ‌های گوژپشت با شنا کردن، یک دایره کوچک شده و ایجاد حباب‌های متمایز در امتداد یک مسیر دایره‌ای یا "۹" شکل، دسته‌ای از ماهی‌های کوچک را نزدیک به سطح شکار می‌کنند. (۸) به عبارت دیگر، نهنگ‌های گوژپشت با به دام انداختن طعمه در شبکه‌ای از حباب‌ها، شکار می‌کنند. آن‌ها این تور را هنگام شنا در مسیری دایره‌ای شکل ایجاد می‌کنند و محققان با استفاده از معادلات ریاضی، حرکت نهنگ را در اطراف طعمه مدل می‌کند (۱۳).

الگوریتم بهینه‌سازی ازدحام ذرات: بهینه ساز ازدحام ذرات یکی از الگوریتم‌های معروف مبتنی بر ازدحام است. این الگوریتم، رفتار اجتماعی پرندگان و ماهی‌ها را تقلید می کند. پیاده سازی بهینه ساز ازدحام ذرات ساده است و می‌تواند نقطه بهینه را به سرعت پیدا کند. این الگوریتم در ابتدا برای شبیه‌سازی فرآیند یافتن غذا توسط پرندگان مورد استفاده قرار گرفت و از تعدادی ذرات استفاده می‌کند که گروهی را تشکیل می‌دهند که در فضای جستجو در حال حرکت هستند تا بهترین راه‌حل را جستجو کنند. هر ذره در گروه، یک تابع تناسب برای محاسبه مقدار تناسب دارد (۱٤).

الگوریتم‌های یادگیری ماشین و طبقه‌بندی: امروزه الگوریتم‌های یادگیری ماشین به دلیل قدرت آن‌ها در مدیریت داده‌های حجیم و پیش‌بینی‌های کارآمد، اهمیت زیادی به دست آورده‌اند. یادگیری ماشین می‌تواند به عنوان راه حلی برای کاهش هزینه-های مربوط به مدیریت مراقبت‌های بهداشتی عمل کند. در حقیقت، برای انجام یک کار طبقه‌بندی، از ابزاری به نام یادگیری ماشین استفاده می‌شود. این ابزار توسط محقق برای تجزیه و تحلیل حجم زیادی از داده‌ها به منظور به دست آوردن یک الگوی خاص استفاده می‌شود. سپس این الگو برای به دست آوردن دانشی که می‌تواند در فرآیند تصمیم‌گیری کمک کند، تفسیر می‌شود. یادگیری ماشین در سیستم‌های تشخیص پزشکی اهمیت زیادی پیدا کرده است؛ زیرا ثابت کرده است که تا حد خوبی در تشخیص دقیق است، در کاربرد در درمان‌ها موفق است و مقرون به صرفه است (۱۵). همچنین، یادگیری ماشین به افزایش خودکارسازی و در نتیجه کاهش تلاش‌های انسان در هر زمینه کمک می‌کند (۳). به همین منظور، هدف از این تحقیق به کار گیری یادگیری ماشین جهت تشخیص و پیش‌بینی بیماری دیابت است. الگوریتم‌های یادگیری ماشینی که در این تحقیق مورد استفاده قرار گرفته‌اند، عبارتند از الگوریتم جنگل تصادفی و الگوریتم K–نزدیک‌ترین همسایه.

الگوریتم جنگل تصادفی: این الگوریتم محبوب‌ترین الگوریتم یادگیری مبتنی بر درخت است که چندین زیرمجموعه را از داده‌های مشاهده شده انتخاب می‌کند و برای هر زیر مجموعه، یک درخت تصمیم می‌سازد و آن را آموزش می‌دهد (۱٦). در واقع، جنگل تصادفی یک مدل یادگیری ماشینی گروهی است که هم برای مسائل طبقه‌بندی و هم برای رگرسیون استفاده می-شود. این الگوریتم به طور تصادفی یک زیرمجموعه را از مجموعه داده‌های آموزشی انتخاب می‌کند و درخت‌های تصمیم‌گیری مختلف را تولید می‌کند (۳). این الگوریتم از درخت‌های تصمیم‌گیری متعدد استفاده می‌کند تا به عنوان یک واحد عمل کنند. هر درخت کلاسی را که یک نمونه به آن تعلق دارد طبقه‌بندی می‌کند و کلاسی که بیش‌ترین آرا را داشته باشد،

کلاس پیش‌بینی شده نامیده می‌شود (۲). به عبارت دیگر، این الگوریتم مجموعه‌ای موازی از چندین طبقه‌بندی درخت تصمیم است و از رای اکثریت یا میانگین برای به دست آوردن نتیجه نهایی استفاده می‌کند. بنابراین مشکل برازش بیش از حد را به حداقل می‌رساند و دقت پیش بینی را افزایش می‌دهد. به همین دلیل، مدل یادگیری جنگل تصادفی معمولاً دقیق‌تر از یک مدل مبتنی بر درخت تصمیم است (۱۷).

الگوریتم k–نزدیک‌ترین همسایه: الگوریتم k–نزدیک‌ترین همسایه یکی از الگوریتم‌های محبوب در حوزه یادگیری ماشین و داده‌کاوی است که برای جستجوی همسایگان نزدیک به داده ورودی استفاده می‌شود و می‌توان آن را هم برای طبقه‌بندی و هم برای رگرسیون استفاده کرد. هنگامی که مجموعه‌ای از داده‌های ورودی و آموزشی به این الگوریتم داده می‌شود، الگوریتم با استفاده از فاصله یا شباهت بین داده ورودی و هر یک از داده‌های آموزشی، k داده‌ی نزدیک‌تر به داده ورودی را پیدا می‌کند. برای این کار، معمولا از معیار فاصله اقلیدسی استفاده می‌شود. پس از پیدا کردن k نزدیک‌ترین داده، می‌توان از آن‌ها برای پیش‌بینی داده ورودی استفاده کرد. الگوریتم k نزدیک‌ترین همسایه می‌تواند در بسیاری از کاربردهای مختلف از جمله پردازش تصویر، تشخیص چهره، دسته‌بندی متن و تحلیل شبکه‌های اجتماعی مفید باشد. بزرگ‌ترین مشکل این الگوریتم، انتخاب تعداد بهینه همسایگان است که باید در نظر گرفته شوند (۱۷). همچنین باید دانست که هرچه تعداد داده‌های مورد تحلیل بیشتر باشد و به عبارت دیگر، از مجموعه داده بزرگتری برای کاربرد مورد نظر استفاده شود، این الگوریتم عملکرد بهتری را از خود نشان خواهد داد.

معیارهای ارزیابی: میزانی که یک روش، الگوها را شناسایی می‌کند و به ما کمک می‌کند تا از داده‌ها مدلی ایجاد کنیم، اثربخشی و دقت آن را مشخص می‌کند (۳). ماتریس سردرگمی، یکی از معیارهای کمک کننده به تشخیص عملکرد یک مدل است و عملکرد طبقه‌بندی‌کننده را با تضاد کلاس‌های واقعی و کلاس‌های پیش‌بینی‌شده توصیف می‌کند. ماتریس سردرگمی یک جدول ۲×۲ است که شامل چهار نتیجه تولید شده توسط یک طبقه‌بندی کننده باینری است (٤). خروجی ماتریس سردرگمی در طبقه‌بندی داده‌های مربوط به بیماری دیابت، عبارت است از مثبت واقعی، زمانی که فرد دیابتی به عنوان دیابتی طبقه‌بندی می‌شود؛ مثبت کاذب، زمانی که یک فرد سالم به عنوان یک فرد دیابتی در نظر گرفته می‌شود؛ منفی واقعی، زمانی که فرد سالم به عنوان یک فرد سالم طبقه-بندی می‌شود و منفی کاذب، زمانی که فرد دیابتی به عنوان یک فرد سالم در نظر گرفته می‌شود (۱۸).

معیارهای مختلفی مانند میزان خطا، صحت، حساسیت (فراخوانی) و دقت از ماتریس سردرگمی مشتق می‌شوند که در ادامه به چهار مورد پرکاربرد از آن‌ها که در این تحقیق نیز مورد استفاده قرار گرفته‌اند، اشاره می‌کنیم.

صحت (Accuracy): نسبت تعداد نمونه‌های طبقه‌بندی شده واقعی به تعداد کل نمونه‌ها است و عملکرد کلی طبقه‌بندی را توصیف می‌کند. صحت پرکاربردترین معیار ارزیابی عملکرد یک طبقه‌بندی کننده یادگیری ماشین است و در واقع کسری از داده‌هایی است که به درستی پیش‌بینی شده‌اند.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$ رابطه (۲)

دقت (Precision): این شاخص، نسبت تعداد نمونه‌های مثبت واقعی پیش‌بینی شده به تعداد کل نمونه‌های پیش‌بینی شده است و نشان می‌دهد که چند درصد از پیش‌بینی‌ها درست بوده‌اند.

$$Precision = \frac{TP}{TP+FP}$$ رابطه (۳)

فراخوانی (Recall): این شاخص، نسبت تعداد نمونه‌های مثبت طبقه‌بندی شده به تعداد کل نمونه‌های مثبت است. فراخوانی به معنای اثربخشی طبقه‌بندی کننده برای شناسایی تمام نمونه‌های مثبت است.

$$Recall = \frac{TP}{TP+FN}$$ رابطه (٤)

امتیاز F1 (F1-Score) : نشان می‌دهد که چند درصد از پیش-بینی‌های مثبت درست بوده اند. امتیاز F1 یک میانگین هارمونیک وزنی از Precision و Recall است. امتیاز F1 دارای بهترین مقدار یک و بدترین مقدار صفر است.

$$F1\ Score = \frac{TP}{TP+1/2(FP+FN)}$$ رابطه (۵)

نتایج

انتخاب ویژگی یک مرحله ضروری قبل از طبقه‌بندی است که بر نتایج طبقه بندی تأثیر می‌گذارد. در این مقاله، ترکیبی از الگوریتم‌های فراابتکاری مختلف برای بهبود دقت KNN و RF در تشخیص دیابت بررسی می‌شود. به این منظور، ابتدا داده‌های مجموعه داده دیابت هندی پیما، بین ۱- و ۱ به عنوان فرایند پیش پردازش، نرمال‌سازی شدند و سپس مجموعه داده به ۷۰٪ داده های آموزشی و ۳۰٪ داده های آزمایشی تقسیم شد. در مرحله بعد، پارامترهایی مانند مقدار K در الگوریتم KNN و حداکثر تعداد تکرارها در هر بار آموزش مدل، به ترتیب روی ۱۰ و ۱۰۰ تنظیم شدند. قابل ذکر است که برای پیش پردازش داده‌ها و انجام این مدل‌سازی، از زبان برنامه نویسی پایتون و کتابخانه‌های مربوطه استفاده شده است. به این طریق که ابتدا مجموعه داده فراخوانده می-شود و سپس کدهای دستوری روی آن اجرا می‌شوند.

جدول ۳ نتایج شبیه‌سازی را بر اساس معیارهای ارزیابی دقت، صحت، حساسیت و امتیاز F1، قبل و بعد از انتخاب ویژگی نشان می‌دهند. همانطور که مشاهده می‌شود، انتخاب ویژگی سبب افزایش عملکرد مدل شده است.

از جدول ۳ و نمودار شکل ۱ (الف) می‌توان دریافت که ترکیب RF-GWO از نظر معیار ارزیابی صحت، عملکرد بهتری را در تشخیص بیماری دیابت نشان داده است. بنابراین، نتیجه می‌گیریم که از بین الگوریتم‌های تکاملی، الگوریتم گرگ خاکستری با انتخاب ٦ ویژگی از ۸ ویژگی، بیشترین تاثیر را در افزایش دقت مدل پیشنهادی این تحقیق داشته است.

نمودارهای اشکال ۱ (ب)، ۱ (ج) و ۱ (د) نتایج شبیه-سازی را به ترتیب بر اساس معیارهای ارزیابی دقت، فراخوانی و امتیاز F1 نشان می‌دهند. همانطور که مشاهده می‌شود، از نظر این سه معیار نیز ترکیب RF-GWO بهترین عملکرد را دارد. اما این معیارها نشان می‌دهند که عملکرد طبقه‌بندی کننده KNN با احتساب تمام ویژگی‌ها بهتر است. اگرچه از نظر معیار صحت، پس از ترکیب RF-GWO، ترکیب KNN-GWO عملکرد بهتری دارد.

بحث

در کل می‌توان نتیجه گرفت که انتخاب ویژگی می‌تواند بر عملکرد طبقه‌بندی کننده‌ها تاثیر ویژه‌ای بگذارد؛ اما برای رسیدن به بالاترین عملکرد، تنها انتخاب ویژگی کافی نیست و باید پارامترهای دیگر را نیز در نظر گرفت.

همانطور که ذکر شد، برای اینکه عملکرد بالایی از الگوریتم KNN بگیریم، باید آن را بر تعداد داده‌های زیاد اعمال کنیم. از آنجایی که مجموعه داده مورد هدف این تحقیق مجموعه‌ی خیلی بزرگی نیست و فقط ۸ ویژگی دارد، الگوریتم KNN با احتساب ۸ ویژگی عملکرد تقریبا بهتری را داشته است و حذف ویژگی‌ها و به عبارت دیگر کاهش داده‌ها، تاثیر مثبت زیادی روی عملکرد آن نداشته و در برخی موارد تاثیر منفی نیز داشته است.

اشکال ۲ (الف) تا ۲ (و) به ترتیب روند نزولی مقدار تابع تناسب را برای ترکیبات KNN-PSO، KNN-WOA، KNN-GWO، RF-PSO، RF-WOA و RF-GWO نشان می‌دهند. هدف در بهینه‌سازی، مینیمم کردن خطای تابع هدف است که تابع هدف در این تحقیق، خطای KNN و RF است. به همین دلیل برای رسیدن به حداقل خطای مدل پیشنهادی، نمودار تابع تناسب باید نزولی باشد و همانطور که مشاهده می‌شود، در صدمین تکرار به حداقل خطا رسیده‌ایم و الگوریتم‌ها خاتمه یافته‌اند.

نتیجه گیری

با رشد روز افزون داده‌های پزشکی، نقش تجزیه و تحلیل داده‌ها و استخراج اطلاعات مفید از آن‌ها در فناوری اطلاعات سلامت افزایش یافته است. از طرف دیگر، فراوانی ویژگی‌های اضافی و نامربوط در مجموعه داده‌های پزشکی، کارایی روش‌های داده کاوی موجود را کاهش می‌دهد که منجر به نتایج غیرقابل اطمینان می‌شود و نیاز به روش‌های ترکیبی را بیشتر می‌کند. به همین دلیل، انتخاب ویژگی یک گام مهم در طراحی سیستم‌های تشخیص پزشکی خودکار است. همچنین باید

دانست که روش‌های یادگیری ماشین، گاهی اوقات می‌توانند بهتر از پیش‌بینی‌های یک متخصص عمل کنند؛ به همین دلیل، می‌توانند به عنوان یک مدل مناسب برای هدایت تصمیمات پزشکان عمل کنند و هزینه‌های مربوطه را از طریق پیش‌بینی و تشخیص زودهنگام کاهش دهند. در این تحقیق، به طور خاص روی بیماری دیابت تمرکز شده است؛ چراکه امروزه به دلیل عادات غذایی ناسالم، دیابت برای اکثر مردم در سراسر جهان یک مشکل بسیار جدی شده است و افزایش سریع تعداد افراد مبتلا به دیابت، توجه همگان را به خود جلب کرده است. به طور معمول درمان دائمی برای این بیماری وجود ندارد و به همین علت، تشخیص زودهنگام بیماری دیابت اصلی‌ترین عامل برای کاهش سایر عوارض مرتبط با آن است. بنابراین، استفاده از مدل‌های خودکار ترکیبی جهت توسعه روش‌های تشخیصی ضروری است؛ چراکه آزمایش-های پزشکی سنتی می‌تواند برای بیماران پر هزینه و خسته کننده باشد و از این رو، خودکارسازی این آزمایش‌ها به منظور ایجاد جامعه‌ای سالم‌تر و کاهش خطر ابتلا به چنین بیماری‌هایی ترجیح داده می‌شود. به همین منظور، این تحقیق به ارائه یک چارچوب کلی جهت تشخیص بیماری دیابت می‌پردازد. در ابتدا، از الگوریتم‌های تکاملی گرگ خاکستری، بهینه‌ساز نهنگ و بهینه‌ساز ازدحام ذرات برای انتخاب ویژگی-های مفید و سپس از الگوریتم‌های یادگیری ماشین RF و KNN جهت طبقه‌بندی داده‌ها استفاده شد. از میان الگوریتم-های تکاملی مورد تحقیق، بهینه‌ساز گرگ خاکستری با انتخاب ٦ ویژگی هنگام اعمال الگوریتم‌های RF و KNN، بیشترین تاثیر را در افزایش صحت مدل پیشنهادی داشت. در نهایت مدل ترکیبی RF-GWO با صحت ۸۱/۳۸ ٪، بالاترین عملکرد را در تشخیص و پیش‌بینی بیماری دیابت داشت.

تشکر و قدردانی

مقاله حاضر برگرفته از پایان نامه مقطع کارشناسی ارشد مهرنوش آهنگرانی، دانشجوی رشته مهندسی فناوری اطلاعات دانشگاه خواجه نصیرالدین طوسی می باشد.

بدین وسیله از جناب آقای دکتر محمد جعفر تارخ، جهت راهنمایی های بی دریغ ایشان در طی این مسیر تحقیقاتی تشکر و قدردانی به عمل می‌آید.

جدول ۱ – الگوریتم‌های تکاملی به کار گرفته شده جهت انتخاب ویژگی توسط محققان مختلف: تشخیص بیماری دیابت با استفاده از یادگیری ماشین و الگوریتم‌های تکاملی

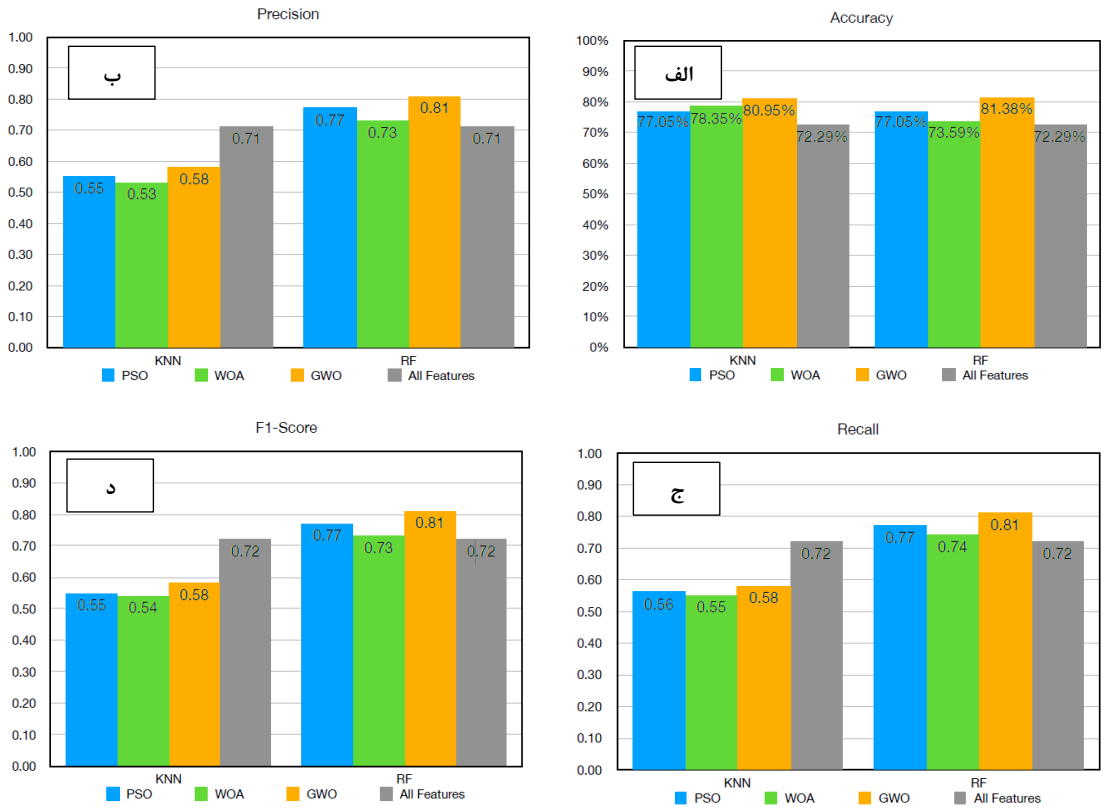
ردیف	نویسندگان	سال انتشار	الگوریتم تکاملی استفاده شده جهت انتخاب ویژگی
۱	Emary و همکاران (۱۲)	۲۰۱۵	– بهینه ساز گرگ خاکستری باینری
۲	Mafaraja و همکاران (۹)	۲۰۱۷	– بهینه ساز نهنگ
۳	Mafaraja و همکاران (۹)	۲۰۱۷	– بهینه ساز نهنگ باینری
٤	Emary و همکاران (۱۲)	۲۰۱٦	– هینه ساز شیر مورچه باینری
٥	Lu و همکاران (١٤)	۲۰۱٥	– بهینه ساز ازدحام ذرات بهبود یافته

جدول ۲ – خلاصه روش‌های ترکیبی به کار گرفته شده جهت تشخیص دیابت و یا عوارض ناشی از آن مانند رتینوپاتی دیابتی توسط محققان مختلف

ردیف	نویسندگان	سال انتشار	روش انتخاب/استخراج ویژگی	روش طبقه‌بندی
۱	Welikala و همکاران (۲۱)	۲۰۱۵	– الگوریتم ژنتیک	– ماشین بردار پشتیبان
۲	Tamim و همکاران (۲۰)	۲۰۲۱	– الگوریتم ژنتیک	– درخت تصمیم – بیز ساده – K–نزدیک ترین همسایه – تجزیه و تحلیل تشخیص خطی
۳	Herliana و همکاران (۲۲)	۲۰۱۹	– بهینه ساز ازدحام ذرات	– شبکه عصبی
٤	Kumar و همکاران (٦)	۲۰۱۷	– بهینه ساز ازدحام ذرات	– ماشین بردار پشتیبان – بیز ساده
٥	Jayanthi و همکاران (۱۹)	۲۰۲۰	– ترکیب بهینه ساز ازدحام ذرات با شبکه عصبی کانولوشنال	– درخت تصمیم
٦	Alharbi و همکاران (۲۳)	۲۰۱۹	– الگوریتم ژنتیک	– شبکه عصبی – ماشین یادگیری افراطی
۷	Kumar و همکاران (٥)	۲۰۱۹	– الگوریتم ژنتیک	– بیز ساده
۸	Azad و همکاران [۸]	۲۰۲۲	– الگوریتم ژنتیک	– درخت تصمیم
۹	Algelal و همکاران (۱٦)	۲۰۲۱	– الگوریتم ژنتیک – تجزیه و تحلیل تشخیص خطی	– جنگل تصادفی – الگوریتم کیسه بندی – الگوریتمJRip
۱۰	Kumar و همکاران (۲٤)	۲۰۲٤	–انتخاب رو به جلو –انتخاب متوالی به عقب –الگوریتم بهینه‌سازی نهنگ –الگوریتم ژنتیک	–جنگل تصادفی –پرسپترون چند لایه – K–نزدیک ترین همسایه
۱۱	Kiran و همکاران (۲٥)	۲۰۲٤	–الگوریتم ژنتیک	–جنگل تصادفی
۱۲	Bhat و همکاران (۲٦)	۲۰۲٤	–انتخاب مبتنی بر کسب اطلاعات –انتخاب مبتنی بر همبستگی –انتخاب ویژگی متوالی	–رگرسیون لجستیک –ماشین بردار پشتیبان – بیز ساده –درخت تصمیم –جنگل تصادفی –K–نزدیک ترین همسایه
۱۳	Lohani و همکاران (۲۷)	۲۰۲٤	–الگوریتم ژنتیک	–پشتیبان ماشین بردار –جنگل تصادفی –K–نزدیک ترین همسایه –بیز ساده

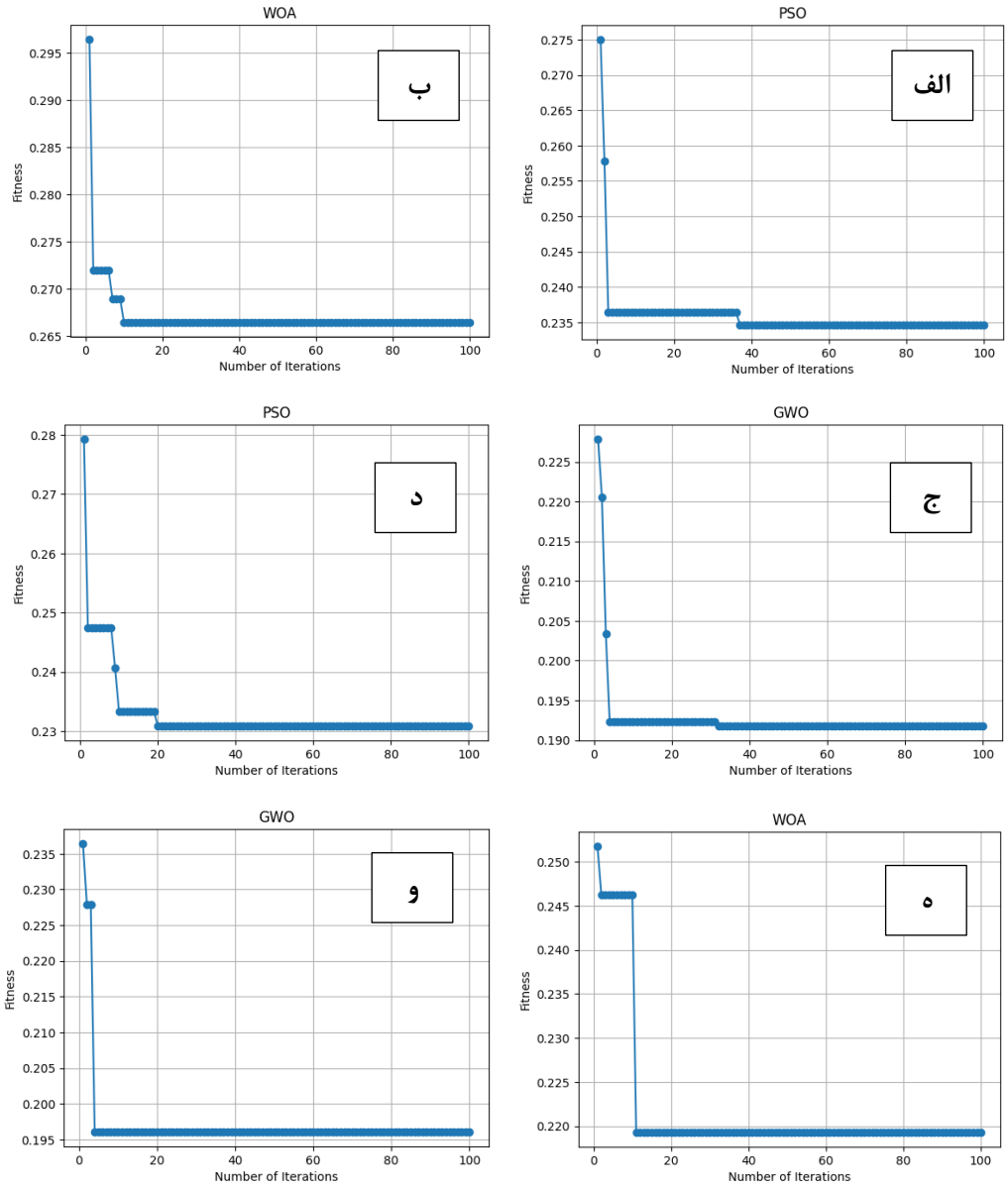
جدول ۳- نتایج شبیه‌سازی بر اساس معیارهای ارزیابی مختلف: تشخیص بیماری دیابت با استفاده از یادگیری ماشین و الگوریتم-های تکاملی

	تعداد ویژگی	صحت	دقت	فراخوانی	امتیاز F1
KNN	۸	٪۷۲/۲۹	٪۷۱	٪۷۲	٪۷۲
RF	۸	٪۷۲/۲۹	٪۷۱	٪۷۲	٪۷۲
KNN-PSO	۳	٪۷۷/۰۵	٪۵۵	٪۵۶	٪۵۵
KNN-WOA	٤	٪۷۸/۳۵	٪۵۳	٪۵۵	٪۵٤
KNN-GWO	٦	٪۸۰/۹۵	٪۵۸	٪۵۸	٪۵۸
RF-PSO	٦	٪۷۷/۰۵	٪۷۷	٪۷۷	٪۷۷
RF-WOA	٤	٪۷۳/۵۹	٪۷۳	٪۷٤	٪۷۳
RF-GWO	٦	٪۸۱/۳۸	٪۸۱	٪۸۱	٪۸۱



شکل ۱- تشخیص بیماری دیابت با استفاده از یادگیری ماشین و الگوریتم‌های تکاملی:

(الف) نمودار نتایج شبیه‌سازی بر اساس معیار ارزیابی صحت، (ب) نمودار نتایج شبیه‌سازی بر اساس معیار ارزیابی دقت، (ج) نمودار نتایج شبیه‌سازی بر اساس معیار ارزیابی فراخوانی، (د) نمودار نتایج شبیه‌سازی بر اساس معیار ارزیابی امتیاز F1



شکل ۲- تشخیص بیماری دیابت با استفاده از یادگیری ماشین و الگوریتم‌های تکاملی:
(الف) مقدار تابع تناسب مدل ترکیبی RF-PSO در ۱۰۰ تکرار، (ب) مقدار تابع تناسب مدل ترکیبی RF-WOA در ۱۰۰ تکرار، (ج) مقدار تابع تناسب مدل ترکیبی RF-GWO در ۱۰۰ تکرار، (د) مقدار تابع تناسب مدل ترکیبی KNN-PSO در ۱۰۰ تکرار، (ه) مقدار تابع تناسب مدل ترکیبی KNN-WOA در ۱۰۰ تکرار، (و) مقدار تابع تناسب مدل ترکیبی KNN-GWO در ۱۰۰ تکرار

References

1. Jaiswal V, Negi A, Pal T. A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes [Internet]. 2021 Jun 1;15(3):435-43.
2. Abdulhadi N, Al-Mousa A. Diabetes Detection Using Machine Learning Classification Methods. In: International Conference in Information Technology [Internet]. IEEE; 2021. p. 350-4.
3. Katarya R, Jain S. Comparison of Different Machine Learning Models for diabetes detection. In IEEE, 2020.
4. Saxena S, Mohapatra D, Padhee S, Sahoo GK. Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. Evolutionary Intelligence [Internet]. 2021 Nov 24;1-17.
5. Komal Kumar N, Vigneswari D, Vamsi Krishna M, Phanindra Reddy GV. An Optimized Random Forest Classifier for Diabetes Mellitus. Advances in Intelligent Systems and Computing. 2018 Sep 2;765-73.
6. Kumar SN, Dinesh D, Siddharth T, Ramkumar S, Nikhill S, Lavanya R. Selection of features using Particle Swarm Optimization for microaneurysm detection in fundus images. In: International Conference on Wireless Communications and Signal Processing [Internet]. IEEE; 2017.
7. Chang V, Bailey J, Xu Q, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications [Internet]. 2022 Mar 24;1-17.
8. Azad C, Bhushan B, Sharma R, Shankar A, Singh KK, Khamparia A. Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. Multimedia Systems [Internet]. 2021 Jun 6;1-19.
9. Mafarja M, Mirjalili S. Hybrid Whale Optimization Algorithm with Simulated Annealing for Feature Selection. Neurocomputing [Internet]. 2017 Oct 18;260:302-12.
10. Emary E, Zawbaa HM, Hassanien AE. Binary ant lion approaches for feature selection. Neurocomputing [Internet]. 2016 Nov 12;213(213):54-65.
11. Type 2 Diabetes Prediction using Gray Wolf Optimization Algorithm. Indian Journal of Forensic Medicine & Toxicology [Internet]. 2021 May 17 [cited 2024 Sep. 17];15(3):4390-5.
12. Emary E, Zawbaa HM, Hassanien AE. Binary grey wolf optimization approaches for feature selection. Neurocomputing [Internet]. 2016 Jan 8;172(8):371-81.
13. Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing [Internet]. 2018 Jan 1; 62:441-53.
14. Lu Y, Liang M, Ye Z, Cao L. Improved particle swarm optimization algorithm and its application in text feature selection. 2015 Oct 1; 35:629-36.
15. Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art. 2021 Dec; 4(1).
16. Abbas, Algelal ZM, Nabeel Salih Ali, Al-Garaawi N. Improving Classification Performance for

Diabetes with Linear Discriminant Analysis and Genetic Algorithm. 2021 Sep 1.
17. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. 2021 Mar 22; 2(3):160.
18. Haq AU, Li J, Khan J, Memon MH, Nazir S, Ahmad S, et al. Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data. Sensors [Internet]. 2020 May 6; 20(9):2649.
19. Jayanthi J, Jayasankar T, Krishnaraj N, Prakash NB, Britto ASF, Kumar KV. An Intelligent Particle Swarm Optimization with Convolutional Neural Network for Diabetic Retinopathy Classification Model. Journal of Medical Imaging and Health Informatics [Internet]. 2021 Mar 1; 11(3):803-9.
20. Tamim N, Elshrkawey M, Nassar, H. Accurate diagnosis of diabetic retinopathy and glaucoma using retinal fundus images based on hybrid features and genetic algorithm. Applied sciences (Basel, Switzerland) 2021;11(13):6178.
21. Welikala RA, Fraz MM, Dehmeshki J, Hoppe A, Tah V, Mann S, et al. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. Computerized Medical Imaging and Graphics [Internet]. 2015 Jul 1; 43:64-77.
22. [22]. Herliana A, Arifin T, Susanti S, Hikmah A. Feature Selection of Diabetic Retinopathy Disease Using Particle Swarm Optimization and Neural Network. In Repository Universitas Bina Sarana Informatika (RUBSI); 2018.
23. Alharbi A, Alghahtani M. Using Genetic Algorithm and ELM Neural Networks for Feature Extraction and Classification of Type 2-Diabetes Mellitus. Applied Artificial Intelligence [Internet]. 2019 Mar 21; 33(4):311-28.
24. Kumar P, Bhati BS, Dhanaraj RK, Iwendi C, Balusamy B, Bhati NS, et al. Feature Subset Selection using Heuristic and Metaheuristic Approaches for Diabetes Prediction on a Binary Encoded Dataset. International Journal of Modeling, Simulation, and Scientific Computing. 2024 Mar 8.
25. Kiran TSR, Sowjanya B, Srisaila A, Lakshmanarao A, Shankar GS. Machine Learning Approach for Diabetes Prediction using Genetic Algorithm based Feature selection. In 2024. p. 1-5.
26. Bhat SS, Ansari GA, Ansari MD. Performance Analysis of Machine Learning Based On Optimized Feature Selection for Type II Diabetes Mellitus. Multimedia Tools and Applications. 2024 Mar 27.
27. Lohani BP, Dagur A, Shukla D. Feature selection based hybrid machine learning classification model for diabetes mellitus type-II. In Informa; 2023. p. 96-101.

Diagnosis of Diabetes Using Machine Learning and Evolutionary Algorithms

Mehrnoosh Ahangarani[1], Mohammad Jafar Tarokh[*2]
1- MSc. Student, Department of Information Technology Engineering, Faculty of Industrial Engineering, K. N. Toosi University, Tehran, Iran
2- Professor, Department of Information Technology Engineering, Faculty of Industrial Engineering, K. N. Toosi University, Tehran, Iran
*Corresponding Author: mjtarokh@kntu.ac.ir
Received: Jun 15, 2024 **Accepted: Sep 11, 2024**

ABSTRACT
Background and Aim: In recent years, machine learning and evolutionary algorithms have drawn the attention of researchers and specialists in various fields, especially in healthcare, due to their practical applications in processing large datasets to provide valuable insights. Considering the increasing prevalence of diabetes and its rapid and accurate diagnosis being one of the most critical issues in medicine, significant concerns are faced by global communities worldwide. The present study was conducted with the aim of creating a diagnostic model based on evolutionary algorithms and machine learning to diagnose diabetes.
Materials and Methods: This research based on the Indian Pima diabetes dataset presents a framework based on intelligent diabetes diagnosis. The proposed method consists of two main stages. The first stage involves a classification approach using K-nearest neighbors and random forest algorithms. The second stage includes a combined feature selection and classification approach to enhance the results of the first stage, utilizing grey wolf optimization, whale optimization, and particle swarm optimization algorithms for feature selection. Comparative analysis among different approaches is conducted through evaluation metrics such as accuracy, precision, recall, and F1-score.
Results: After comparative comparisons among the proposed models, the random forest model based on the grey wolf optimization was selected and introduced as the final model with a prediction accuracy of 81.38%.
Conclusion: The findings of this research indicate that the use of evolutionary algorithms alongside machine learning models can often enhance the efficiency and accuracy of diabetes diagnosis and its associated complications.
Keywords: Diabetes Diagnosis, Machine Learning, Evolutionary Algorithms, Feature Selection