بهینه سازی شبکه عصبی MLP با استفاده از الگوریتم ژنتیک موازی FinGrain برای تشخیص سرطان سینه

امین رضایی‌پناه[۱،*]، علی مبارکی[۲] و سعید بحرانی خادمی[۳]

اطلاعات مقاله

دریافت مقاله: ۱۳۹۷/۰۵/۲۸
پذیرش مقاله: ۱۳۹۷/۱۰/۱۲

واژگان کلیدی:
الگوریتم ژنتیک موازی، تکنیک FineGrain، شبکه عصبی MLP، تشخیص سرطان سینه، ویژگی‌های موثر.

چکیده

امروزه استفاده از سیستم‌های هوشمند در تشخیص پزشکی به تدریج در حال افزایش است. این سیستم‌ها می‌توانند به کاهش خطایی که ممکن است توسط کارشناسان کم‌تجربه اتفاق بیافتد، کمک کند. بدین منظور استفاده از سیستم‌های هوشمند مصنوعی در پیش‌بینی و تشخیص سرطان سینه که یکی از رایج‌ترین سرطان‌ها در بین زنان است، مورد توجه می‌باشد. در این تحقیق فرآیند تشخیص بیماری سرطان سینه با یک رویکرد دو مرحله‌ای انجام می‌شود. در مرحله اول دو پارامتر ویژگی‌های موثر و تعداد نودهای لایه مخفی به منظور آموزش شبکه عصبی MLP به صورت همزمان توسط یک الگوریتم ژنتیک بهینه‌سازی می‌شوند. سپس با استفاده از ویژگی‌های انتخاب شده و تعداد نودهای لایه مخفی، یک مدل طبقه‌بندی برمبنای شبکه عصبی MLP برای تشخیص بیماری سرطان سینه در مرحله دوم ایجاد می‌شود. در این مرحله از یک الگوریتم ژنتیک موازی FinGrain بر مبنای پارامترهای بهینه‌سازی شده، برای تنظیم وزن‌های شبکه عصبی MLP استفاده می‌شود. ارزیابی آزمایش‌ها نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های GAANN و CAFS روی مجموعه‌داده WBCD به نتایج بهتری رسیده است و دقت ۹۸,۷۲٪ را در حالت میانگین گزارش می‌کند.

۱-مقدمه

در دنیای امروز با عنایت به حجم وسیع و پیچیدگی داده‌های موجود، بیش‌ازپیش نیاز به ابزاری کارا، مؤثر و مطمئن به منظور کشف دانش سودمند و مورد نیاز در این داده‌ها، حس می‌شود. داده‌کاوی ابزاری است که برای حصول به چنین دانشی ما را یاری می‌کند. یکی از زمینه‌های پرکاربرد داده‌کاوی، علم پزشکی است[۱]. براساس تحقیقات گزارش شده توسط وزارت بهداشت تایوان، سرطان سینه رایج‌ترین نوع سرطان در زنان می‌باشد، درحالی که میزان مرگ‌ومیر ناشی از سرطان سینه در زنان بالای ۴۰ سال بسیار بالا است[۲ و ۳].

آزمایش آسپیراسیون سوزنی (FNA[۵]) روشی ارزان و غیرتهاجمی برای تشخیص دقیق و زودهنگام سرطان سینه است. پس از استخراج خصوصیات سیتولوژی بیمار از بافت سینه، نیاز است تا خوش‌خیم یا بدخیم بودن تومور تشخیص داده شود. در مواردی که با قاطعیت نتوان خوش‌خیم یا بدخیم بودن بیماری را تشخیص داد، استفاده از الگوریتم‌های داده‌کاوی راهنمای خوبی برای پزشک و

* پست الکترونیک نویسنده مسئول: amin.rezaeipanah@gmail.com
۱. مربی، دانشگاه رهجویان دانش برازجان، بوشهر
۲. دانشجو کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد بوشهر، ایران
۳. دانشجو کارشناسی، دانشگاه آزاد اسلامی، واحد برازجان، ایران

[5] *Fine Needle Aspiration*

متخصصین می‌باشند[۴].

تشخیص به موقع سرطان سینه (حداکثر ۵ سال پس از اولین تقسیم سلولی سرطان) احتمال زنده ماندن را از ۵۶٪ به بیش از ۸۶٪ افزایش می‌دهد. هر چند فعالیت‌های محدودی در زمینه ثبت سرطان انجام شده است، اما سرطان سینه همچنان به عنوان شایع‌ترین سرطان در بین زنان ایرانی شناخته می‌شود. از اینرو وجود یک سیستم هوشمند که دقت بالایی را در تشخیص مکان توده سرطانی ارائه دهد، بسیار ضروری خواهد بود[۵]. اخیراً با گسترش روزافزون علم، استفاده از سیستم‌های پشتیبان‌تصمیم می‌تواند کمک بسیار زیادی در سیاست‌های درمانی پزشک داشته باشد.

در تحقیقات فراوانی از الگوریتم ژنتیک و شبکه عصبی به منظور انتخاب ویژگی‌ها و طبقه‌بندی داده‌ها استفاده شده است[۶ و ۷]. از بین شبکه‌های عصبی، الگوریتم MLP[1] توجه بیشتری را به خود جلب کرده است[۸].

در ادامه این تحقیق به بررسی برخی از جدیدترین کارهای انجام شده در بخش ۲ می‌پردازیم، در بخش ۳ مدل پیشنهادی مبتنی بر طبقه‌بندی شبکه عصبی MLP و بهینه‌سازی آن با الگوریتم ژنتیک به منظور تشخیص سرطان سینه مطرح شده و عملگرهای لازم ارائه می‌شود. نتایج حاصل از ارزیابی روش پیشنهادی در بخش ۴ آورده شده و در نهایت نتیجه‌گیری در بخش ۵ ذکر شده است.

۲-پیشینه تحقیق

در این بخش تلاش بر آن داریم روش‌های مختلف ارائه شده در زمینه تشخیص سرطان سینه مورد بحث قرار گیرد. در ادامه به بررسی تعدادی از مقالات علمی که در این زمینه ارائه شده‌اند، می‌پردازیم.

به طور کلی روش‌هایی که در این حیطه ارائه شده‌اند از دو شیوه برای تشخصیص سرطان سینه استفاده می‌کنند. روش اول مبتنی بر پردازش تصویر است. در این روش با استفاده از تصاویر ماموگرافی اقدام به تشخیص سرطان می‌کنند[۹]. ماموگرافی در واقع رادیوگرافی از نسج نرم پستان‌ها است. ماموگرافی در مقایسه با دیگر روش‌های تشخیصی از جمله سونوگرافی و MRI[2] در تشخیص به موقع سرطان، دقیق‌ترین روش می‌باشد[۱۰]. هنگام انجام عمل ماموگرافی بخشی از تشعشع x ray متناسب با شرایط بافت جذب می‌شود و بخش دیگر از آن عبور می‌کند. بافت متناسب با ماهیتش بخشی از انرژی را جذب می‌کند. میزان خروج سیگنال از بافت سرطانی (دارای توده)، با بافت سالم متفاوت است. از میزان افت صورت گرفته از سیگنال ورودی به خروجی می‌توان تشخیص داد که این بافت متضمن توده می‌باشد یا خیر. روش دوم طبقه‌بندی داده‌هایی است که بوسیله نمونه‌گیری از بیماران جمع‌آوری شده است. به طور مرسوم تشخیص این بیماری به روش نمونه‌برداری از چربی‌های پستان به کمک سوزن باریک FNA انجام می‌شود، این روش بدون جراحت بوده و نوعی تست مناسب برای ارزیابیرخوش‌خیم و یا بدخیم بودن این بیماری در زنان محسوب می‌گردد. در زمینه پیش‌بینی و تشخیص سرطان سینه از طریق طبقه‌بندی داده‌ها، تحقیقات متعددی صورت گرفته است[۱۱].

در [۱۲] الگوریتم‌های داده‌کاوی نائوبیز، رگرسیون‌لجستیک و درخت‌تصمیم را به منظور تشخیص سلول‌های سرطانی سینه مورد تجزیه و تحلیل قرار گرفتند. هدف این تحقیق یافتن کوچکترین زیر مجموعه از ویژگی‌ها است که می‌توانند طبقه‌بندی خوش‌خیم و بدخیم بودن سرطان سینه را با دقت بسیار بالایی تضمین کند. طبقه‌بندی‌ها از نظر دقت، سرعت و پیچیدگی زمانی مورد بررسی قرار گرفته‌اند که طبقه‌بندی رگرسیون‌لجستیک با دقت ۹۷,۹٪ به عنوان بهترین طبقه‌بند در مقایسه با دو طبقه‌بند دیگر اثبات شده است. در [۱۳] از الگوریتم بهینه‌سازی ازدحام ذرات به منظور تعیین پهنای‌باند و انتخاب ویژگی در تخمین چگالی کرنل مبتنی بر طبقه‌بندی، جهت تشخیص سرطان سینه استفاده شده است. انتخاب ویژگی و پهنای‌باند برای تخمین چگالی کرنل به طور قابل توجهی بر کارایی طبقه‌بندی تاثیر می‌گذارد. مدل PSO-KDE پیشنهادی دو الگوریتم PSO[3] PSO[3] و KDE[4] را به منظور تشخیص سرطان سینه به چالش می‌کشد. به منظور بررسی عملکرد این مدل از دو معیار تعداد ویژگی‌های انتخاب شده و دقت طبقه‌بندی استفاده شده است. عملکرد PSO-KDE بر روی مجموعه داده‌های سرطان سینه مورد بررسی قرار گرفته است. نتایج تجربی نشان می‌دهد که مدل PSO-KDE دارای عملکرد

[1] *Multilayer Perceptron*
[2] *Magnetic Resonance Imaging*
[3] Particle Swarm Optimization
[4] Kernel Density Estimation

متوسط بهتری نسبت به مدل GA-KDE در تشخیص سرطان سینه است.

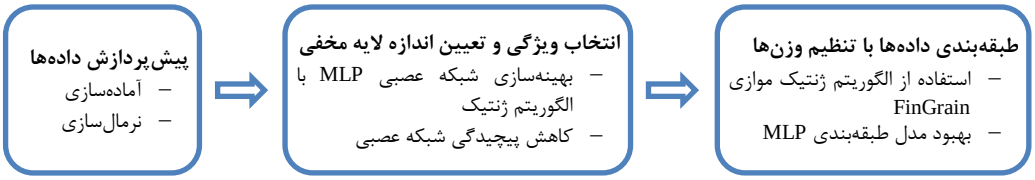
تحقیق دیگری با عنوان «مدل هوشمند پیش‌بینی سرطان سینه با استفاده از تکنیک‌های داده‌کاوی» در [۱۴] ارائه شده است. در این تحقیق برای تشخیص سرطان سینه از یک روش خصیصه انتخابی، به منظور انتخاب ویژگی‌های موثر بهره گرفته شده است، سپس با استفاده از ماشین-بردارپشتیبان[۱۵] مدل طبقه‌بندی را ایجاد می‌کند. در [۱۶] با استفاده از دو مدل شبکه عصبی BPNN[۱] و RBF[۲] به تشخیص سرطان سینه پرداخته شده است. نتایج این تحقیق با شبکه تابع‌شعاعی مقایسه شده است. نتایج مقایسه کارایی شبکه عصبی BPNN را برای تشخیص سرطان سینه نشان می‌دهد.

در [۱۷] سیستمی مبتنی بر دانش برای تشخیص سرطان سینه با استفاده از روش منطق فازی توسعه داده شده است. در این سیستم، بیماری سرطان سینه با استفاده از خوشه‌بندی، حذف‌نویز و تکنیک‌های طبقه‌بندی تشخیص داده می‌شود. از روش حداکثر انتظار برای خوشه‌بندی داده‌ها و از طبقه‌بندی درخت رگرسیون CART[۳] برای تولید قوانین فازی استفاده شده است. در [۱۸] به منظور افزایش کارایی طبقه‌بندی J48 در تشخیص سرطان سینه، ویژگی‌ها را با استفاده از خوشه‌بندی دورترین اولین[۴] گروه‌بندی کرده و سپس از هر گروه موثرترین ویژگی‌ها را انتخاب می‌کند. نتایج پیاده‌سازی دقت حدود ۹۹٪ را برای مجموعه داده‌های WBCD و WDBC نشان می‌دهد. در [۱۹] ترکیبی از روش‌های SMOTE[۵] و PSO برای تشخیص بیماران مبتلا به سرطان سینه استفاده می‌شود. در این روش برخی از مشهورترین روش‌های طبقه‌بندی مانند رگرسیون‌لجستیک، مدل درخت‌تصمیم C5 و نزدیک‌ترین همسایه ادغام می‌شوند. نتایج آزمایش‌ها نشان می‌دهد که الگوریتم ترکیبی SMOTE+PSO+C5 بهترین عملکرد را دارد.

۳-روش پیشنهادی

هدف از این تحقیق ارائه مدلی به منظور تشخیص دو نوع خوش‌خیم و بدخیم سرطان سینه بر مبنای طبقه‌بندی شبکه عصبی MLP است. برای بهبود مدل ارائه شده پارامترهای ویژگی‌های ورودی، اندازه نودهای لایه مخفی و همچنین وزن‌های شبکه را بهینه‌سازی می‌کنیم. ایده اصلی این تحقیق بهبود دقت مدل طبقه‌بندی شبکه عصبی MLP و کاهش پیچیدگی شبکه نهایی می‌باشد. در راستای تحقق این هدف از روش موازی FinGrain و بهینه‌سازی همزمان پارامترهای مختلف در شبکه عصبی MLP استفاده می‌کنیم.

روش پیشنهادی ترکیبی بوده و شامل چندین مرحله است. در ابتدا مرحله پیش‌پردازش داده‌ها انجام می‌شود. این مرحله جهت آماده‌سازی داده‌ها برای انجام آزمایش‌ها می‌باشد. سپس خروجی این مرحله در دو بخش مجزا منجر به ایجاد یک مدل تشخیصی می‌شود. شکل (۱) مراحل الگوریتم پیشنهادی را نشان می‌دهد.



شکل ۱- مراحل الگوریتم پیشنهادی

در بخش اول از یک الگوریتم ژنتیک کلاسیک برای یافتن بهترین زیرمجموعه از ویژگی‌ها بهره می‌گیریم. به این دلیل که ممکن است تعدادی از ویژگی‌های استفاده شده در مجموعه داده زائد بوده و در طبقه‌بندی داده‌ها نقشی نداشته باشند، زیرمجموعه‌ای از بهترین ویژگی‌ها انتخاب می‌شوند. ویژگی‌های انتخابی با استفاده از شبکه عصبی MLP ارزیابی می‌شوند. در این بخش علاوه بر انتخاب ویژگی‌های مطلوب، تعداد نودهای لایه مخفی در حالت بهینه جستجو می‌شود. انتخاب ویژگی‌های موثر و تعداد نودهای لایه مخفی با استفاده از بهبود الگوریتم ژنتیک

[1] Back Propagation Neural Network
[2] Radial Basis Function
[3] Classification And Regression Tree
[4] Farthest First
[5] Synthetic Minority Over-sampling Technique

برمبنای طول رشته متغیر انجام می‌شود. تکنیک طول رشته متغیر امکان تشخیص تعداد ویژگی‌های بهینه را به صورت خودکار فراهم می‌سازد.

در بخش دوم با استفاده از ویژگی‌های انتخاب شده و تعداد نودهای لایه مخفی که در بخش قبلی به صورت بهینه جستجو شدند، یک مدل طبقه‌بندی شبکه عصبی ایجاد می‌شود. در مدل طبقه‌بندی پیشنهادی از یک الگوریتم ژنتیک موازی برای یادگیری و آموزش وزن‌ها بهره گرفته شده که موازی‌سازی الگوریتم ژنتیک به روش FinGrain انجام می‌شود.

در الگوریتم ژنتیک یک عملگر ترکیب تفاضلی به منظور بهبود روند همگرایی طراحی شده است. هدف این مرحله حداقل‌سازی میزان خطای MSE[۱] در طبقه‌بندی بر روی داده‌های آموزشی می‌باشد.

۳-۱-پیش‌پردازش داده‌ها

اهمیت پیش‌پردازش داده‌ها به دلیل این واقعیت است که؛ «فقدان داده با کیفیت برابر با فقدان کیفیت در نتایج کاوش است» و «ورودی بد خروجی بد به دنبال دارد». به همین دلیل در اغلب مواقع اولین گام در ایجاد هر مدلی بر اساس تکنیک‌های داده‌کاوی، پیش‌پردازش داده‌ها می‌باشد. بطورکلی پیش‌پردازش داده‌ها به منظور آماده سازی داده‌ها برای پردازش و همچنین بهبود کیفیت داده‌های واقعی انجام می‌شود. در این تحقیق مرحله پیش‌پردازش شامل «آماده‌سازی» و «نرمال‌سازی» می‌باشد.

آماده‌سازی: در این مرحله فیلدهایی از پایگاه داده نظیر شناسه بیمار که نقشی در طبقه‌بندی داده‌ها ندارند را حذف می‌کنیم. همچنین اطلاعات برخی از ویژگی‌ها برای بعضی از نمونه‌ها ازدست‌رفته می‌باشند، که این نمونه‌ها از پایگاه‌داده حذف می‌شوند. مقادیر ازدست‌رفته برای نمونه‌ها با نماد «؟» گزارش شده‌اند.

نرمال‌سازی؛ هنگامی که مقادیر ویژگی‌ها از پایگاه‌داده در محدوده یا دامنه‌های متفاوتی قرار داشته باشند، آنها را در دامنه مشابهی قرار می‌دهیم. این تکنیک نرمال‌سازی نام دارد و معمولاً منجر به کسب نتایج بهتری در مدل‌های طبقه‌بندی می‌شود. در این تحقیق از روش Z-score [۱] برای نرمال‌سازی استفاده می‌کنیم. این روش بر روی هر ویژگی به صورت مجزا اعمال شده و مقادیر تمام نمونه‌ها را برای آن ویژگی نرمال می‌کند. بعد از اعمال Z-score، میانگین و انحراف معیار برای هر ویژگی به ترتیب برابر ۰ و ۱ می‌شود. بنابراین این روش داده‌ها را روی مرکزیت ۰ قرار می‌دهد. رابطه (۱) روش نرمال‌سازی Z-score را نشان می‌دهد.

$$x_{i,j}^{Z-score} = \frac{x_{i,j} - \mu_j}{\sigma_j} \qquad (۱)$$

جائیکه $x_{i,j}$ و $x_{i,j}^{Z-score}$ به ترتیب مقدار واقعی و مقدار نرمال شده j-مین ویژگی در نمونه i-ام است. μ_j و σ_j نیز به ترتیب میانگین و انحراف معیار مقادیر تمام نمونه‌ها برای j-مین ویژگی است.

۳-۲-انتخاب ویژگی‌های موثر و تعیین اندازه لایه مخفی

در این بخش از یک الگوریتم ژنتیک کلاسیک به منظور جستجو ویژگی‌های موثر در طبقه‌بندی و همچنین تعیین اندازه بهینه لایه مخفی بهره می‌گیریم.

یکی از اهداف اصلی در مسئله انتخاب ویژگی حذف ویژگی‌های نامرتبط و همچنین دارای افزونگی است. هر روش انتخاب ویژگی که به طور مؤثر قادر به حذف این ویژگی‌ها نباشد، یک روش کارا نیست و مجموعه ویژگی‌های انتخاب شده توسط آن روش، عموماً بهترین نخواهد بود. به طور کلی هر چه تعداد ویژگی‌های یک پایگاه‌داده بالا باشد، ابعاد مسئله و پیچیدگی پارامترهای طبقه‌بندی نیز بالا خواهد بود. در نتیجه این امر باعث کاهش عملکرد و دقت الگوریتم‌های طبقه‌بندی می‌شود. یکی دیگر از اهداف این مرحله، تعیین اندازه بهینه نودهای لایه مخفی است، چرا که برای کار طبقه‌بندی از شبکه عصبی MLP استفاده می‌کنیم. تعیین تعداد بهینه نودهای لایه مخفی باعث کسب نتایج بهتری می‌شود و در عین حال شبکه‌ای با پیچیدگی کمتری حاصل می‌نماید. در ادامه مراحل الگوریتم ژنتیک را جهت انتخاب ویژگی‌های موثر و همچنین تعداد بهینه نودهای لایه مخفی شرح می‌دهیم.

شکل (۲) ساختار کروموزوم پیشنهادی را نشان می‌دهد که در آن طول کروموزوم با توجه به تعداد ویژگی‌های انتخابی مشخص می‌شود.

[۱] Mean Squared Error

X	Y_1	Y_2	...	Y_{DNF}

شکل ۲-ساختار کروموزوم پیشنهادی

در اغلب الگوریتم‌های انتخاب ویژگی، تعداد ویژگی‌های موثر به عنوان ورودی الگوریتم است، در حالیکه استفاده از تکنیک طول‌رشته‌متغیر در روش پیشنهادی تعداد بهینه ویژگی‌های موثر را به صورت خودکار محاسبه می‌کند.

ساختار پیشنهادی شامل دو قسمت است، بخش ابتدایی (X) تعداد نودهای لایه مخفی را نشان می‌دهد، در حالیکه بخش دوم ($Y \in \{Y_1, Y_2, \dots, Y_{DNF}\}$) شماره ویژگی‌های انتخاب شده می‌باشد. DNF تعداد ویژگی‌های انتخاب شده است که برای هر کروموزوم متفاوت و در بازه ۱ تا NF (تعداد ویژگی‌ها) است.

در ایجاد جمعیت اولیه سعی در ایجاد یک توزیع یکنواخت در تولید کروموزوم‌ها با طول‌های متفاوت است. برای محاسبه کیفیت راه‌حل‌های تولید شده از معیار دقت طبقه‌بندی پایگاه داده با توجه ویژگی‌های انتخاب شده و اندازه نودهای لایه مخفی استفاده می‌کنیم. مدل طبقه‌بندی استفاده شده، یک شبکه عصبی MLP سه لایه (لایه ورودی-لایه مخفی-لایه خروجی) است که با توجه به مشخص بودن برچسب نمونه‌ها در مجموعه‌داده، به صورت با ناظر آموزش داده می‌شود. همچنین خروجی شبکه یک عدد باینری است که براساس تعداد کلاس‌ها در مجموعه داده تعیین شده است (۰ نشان‌دهنده «عدم بیماری» و ۱ نشان‌دهنده «بیماری»).

وزن‌های شبکه عصبی با استفاده از یک الگوریتم ژنتیک موازی محاسبه می‌شود که در بخش بعدی تشریح می‌گردد. یکی از دلایلی که وزن‌ها در اینجا بهینه‌سازی نمی‌شوند، اجتناب از ایجاد پیچیدگی محاسباتی بالا در الگوریتم می‌باشد. به‌طورکلی در روش‌های انتخاب ویژگی، نیاز به محاسبه متوالی برازندگی ویژگی‌های انتخاب شده وجود دارد، لذا استفاده از روش‌هایی که کمترین پیچیدگی را داشته باشند، در اولویت است. این امر امکان اجرای الگوریتم را روی پایگاه‌های داده بسیار بزرگ فراهم می‌سازد. در این تحقیق از روش تورنومنت (مسابقه) که یکی از متداول‌ترین روش‌های انتخاب است، استفاده می‌شود[۲۰]. برای ایجاد یک کروموزوم جدید (فرزند) از یک عملگر ترکیب با احتمال Cr استفاده می‌کنیم. در عملگر ترکیب پیشنهادی با توجه به متفاوت بودن طول والدین (تعداد ویژگی‌های انتخاب شده)، طول کروموزوم فرزند به صورت تصادفی از بین والدین انتخاب می‌شود.

در این عملگر، ۷۵ درصد ژن‌های کروموزوم فرزند به صورت تصادفی از بین ژن‌های والدین انتخاب می‌شود و مابقی به تصادف از لیست ویژگی‌های مشاهده نشده انتخاب می‌شوند. اندازه نودهای لایه مخفی نیز به صورت میانگین صحیح مقادیر دو والد برای کروموزوم فرزند محاسبه می‌شود. در عملگر جهش پیشنهادی، تمامی ویژگی‌ها به احتمال Mr به صورت تصادفی از لیست ویژگی‌های مشاهده نشده انتخاب می‌شوند. اندازه نودهای لایه مخفی نیز به احتمال Mr×۲ به صورت تصادفی با یکی از مقادیر ۱+، ۰ و ۱- جمع می‌شود.

در نهایت فرزند تولید شده با استفاده از شبکه عصبی MLP مورد ارزیابی قرار می‌گیرد. برای ایجاد جمعیت نسل بعد، تعداد NP فرزند تولید شده و از میان فرزندان و جمعیت نسل قبل (در مجموع NP×۲ کروموزوم)، تعداد NP کروموزوم با بیشترین برازندگی به نسل بعد منتقل می‌شوند. فرایند بهینه‌سازی الگوریتم ژنتیک تا پایان حداکثر نسل تکرار می‌شود و در پایان ویژگی‌های انتخاب شده به همراه اندازه نودهای لایه مخفی از بهترین کروموزوم جمعیت به بخش طبقه‌بندی منتقل می‌شوند.

۳-۳-ایجاد مدل طبقه‌بندی با تنظیم وزن‌های شبکه عصبی MLP

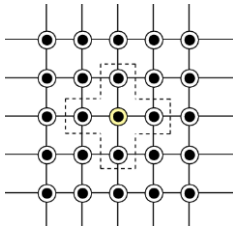
بعد از مرحله انتخاب ویژگی و تعیین اندازه بهینه برای تعداد نودهای لایه مخفی، از یک الگوریتم ژنتیک موازی مبتنی بر FinGrain برای تنظیم وزن‌های شبکه عصبی MLP استفاده می‌شود. در واقع هدف این مرحله افزایش دقت طبقه‌بندی داده‌های سرطان سینه با استفاده از فرایند یادگیری وزن‌ها می‌باشد.

از قابلیت‌های مهم الگوریتم ژنتیک امکان اجرای آن بطور موازی و نیز جستجوی فضاهایی است که بسیار پیچیده یا بزرگ است. از طرفی هزینه محاسباتی بالای این روش را می‌توان با اجرای این الگوریتم به طور موازی بر روی چند کامپیوتر یا چند پردازنده کاهش داد. کار صورت گرفته در این تحقیق تلاش دارد بجای استفاده از یک جمعیت اولیه و انجام مراحل الگوریتم ژنتیک بر روی آن، جمعیت‌های متفاوتی ایجاد نماید. هر جمعیت مرحله انتخاب متفاوتی دارد و از این طریق تبادل کروموزم‌ها بین جمعیت‌ها انجام

می‌شود. این استراتژی مانع از گسترش تاثیرات منفی یک روش نامناسب در سایر جمعیت‌ها می‌شود. از مزایای دیگر این روش امکان موازی‌سازی مراحل مربوط به هر جمعیت، مستقل از سایرین است. بطوریکه در این تحقیق تولید جمعیت اولیه و محاسبه برازندگی آنها به صورت موازی انجام می‌شود.

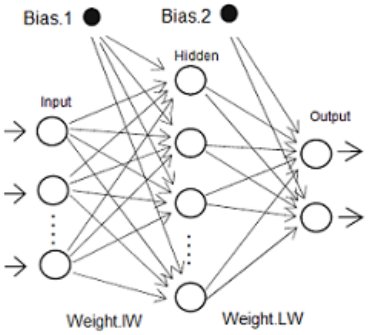
در این تحقیق از مدل FinGrain به منظور موازی‌سازی الگوریتم ژنتیک در بهینه‌سازی وزن‌های شبکه عصبی بهره گرفته شده است. در این مدل افراد بر روی یک ساختار گرد بسیار بزرگ، که انتهایش به ابتدایش متصل است، (یک شبکه دو بعدی) قرار می‌گیرند. برازندگی به طور همزمان برای همه افراد محاسبه می‌شود و عملگر انتخاب برای تعیین والدین در ترکیب و جهش به صورت محلی در یک همسایگی کوچک صورت می‌گیرد. شکل (۳) ساختار مدل FinGrain در الگوریتم ژنتیک را نشان می‌دهد.

در مدل FinGrain تعداد جمعیت‌های مستقل زیاد، ولی اندازه زیر جمعیت‌ها که به صورت گرید کنار هم قرار دارند، کوچک است. بااین‌حال اگر اندازه گرید را $G \times G$ و تعداد کروموزوم‌های هر جمعیت را NP فرض کنیم، تعداد کل جمعیت در الگوریتم ژنتیک برابر $G \times G \times NP$ است.



شکل ۳-ساختار یک الگوریتم ژنتیک موازی FinGrain

ساختار کروموزوم پیشنهادی برداری به طول N است که هر عنصر آن معرف یک وزن در بازه [۱+، ۱-] در شبکه عصبی می‌باشد. تعداد وزن‌ها با توجه به تعداد ورودی‌ها (ویژگی‌ها)، تعداد نودهای لایه مخفی و همچنین اندازه خروجی تعیین می‌شود. در ساختار کروموزوم پیشنهادی، بایاس‌ها در لایه‌های شبکه نیز به‌عنوان یک وزن در نظر گرفته می‌شوند. مطابق شکل (۴) و با توجه به در نظر گرفتن یک شبکه سه لایه (ورودی- مخفی-خروجی) تعداد وزن‌ها از مجموع دو بخش وزنی IW و LW و همچنین دو بایاس $Bias.1$ و $Bias.2$ توسط تابع شمارش وزن Γ در هر قسمت محاسبه می‌شود. در ایجاد جمعیت اولیه، برای هر کروموزوم چندین وزن تصادفی ایجاد می‌شود، سپس وزن‌ها به شبکه عصبی اعمال می‌شوند و مقدار برازندگی هر کروموزوم با توجه به خطا MSE محاسبه می‌گردد.



ساختار نمایش کروموزوم‌ها

W_1	W_2	...	W_N

$$N = \Gamma(IW) + \Gamma(LW) + \Gamma(Bias.1) + \Gamma(Bais.2)$$

شکل ۴-ساختار کروموزوم پیشنهادی در الگوریتم FinGrain

عملگر انتخاب تورنومنت به منظور انتخاب والدین مطابق با خصوصیت الگوریتم ژنتیک FinGrain کروموزوم‌هایی از بین جمعیت‌های همسایه در گرید انتخاب می‌کند. جمعیت‌های همسایه مطابق شکل (۳) برمبنای همسایه‌ها در چهار جهت بالا، پایین، چپ و راست در نظر گرفته می‌شوند. به نوعی این اشتراک‌گذاری کروموزوم‌ها به صورت محلی است که در بهبود روند همگرایی و جلوگیری از افتادن در بهینه‌های محلی موثر است. در این بخش از یک عملگر ترکیب تفاضل تکاملی (DE[1]) جهت ایجاد فرزند استفاده می‌شود که با توجه به تفاوت میان ژن‌های کروموزوم‌های والد و همچنین بهترین کروموزم جمعیت (همسایگان) اقدام به تولید فرزند می‌کند. فرزند جدید با توجه به اختلاف وزن‌های دو عضو جمعیت X_i^1 و X_i^2 و همچنین بهترین عضو جمعیت X_i^0 ایجاد می‌شود. رابطه (۲) نحوه محاسبه X_i^{new} که معرف i-مین وزن جدید است را نشان می‌دهد.

[1] Differential Evolution

$$X_i^{new} = \begin{cases} X_i^0 + F \times (X_i^1 - X_i^2) & if\ C_r > Rand(0,1) \\ X_i^0 & otherwise \end{cases} \quad (۲)$$

در این رابطه، F یک فاکتور مقیاسی در محدوده [۰، ۱] است که سرعت تکامل جمعیت را کنترل می‌کند. C_r احتمال ترکیب برای هر وزن را نشان می‌دهد. ممکن است در تکرار این فرایندها تفاوت‌های بزرگی بین مقادیر تولید شده در فضای جستجو ایجاد شود. از اینرو نیاز به تغییراتی در عملگر پیشنهادی می‌باشد. در اینجا فاکتور مقیاسی F را به صورت پویا مطابق با رابطه (۳) تغییر می‌دهیم.

$$F = \frac{C_1 \times rand(0,1)}{max(X_i^1, X_i^2)} \quad (۳)$$

در این رابطه، C_1 ثابت کوچکتر از یک می‌باشد. این روش به هر عضو جمعیت اجازه می‌دهد برای رسیدن به راه‌حل‌های بهینه در محدوده‌هایی مطابق با اندازه ویژگی نوسان کند.

در نتیجه این روش به بهبود راه‌حل‌های بهینه کمک می‌کند. عملگر جهش تعریف شده در این بخش، تغییر بیت (BC[1]) است که در صورت بهود عملگر ترکیب روی فرزند ایجاد شده و در غیر این صورت روی یکی از والدین به تصادف اعمال می‌شود. این عملگر با احتمال M_r برای هر ژن از کروموزوم سعی در یافتن بهترین همسایه وزنی دارد. همسایه‌های هر وزن در محدوده اختلافی [+۰,۱، -۰,۱] از وزن اصلی به تعداد تکرار ثابتی (در این تحقیق ۵ تکرار) جستجو می‌شوند. در هر تکرار یک مقدار تصادفی در بازه همسایگی با مقدار وزن اصلی کروموزوم جمع شده و در صورت کاهش خطا MSE، وزن بروزرسانی می‌شود.

۴-نتایج و آزمایش‌ها

نتایج حاصل از شبیه‌سازی روش پیشنهادی با عنوان «FGAMLP» در تمام آزمایش‌ها نشان داده شده است. به منظور ارزیابی و بررسی الگوریتم ارائه شده از پایگاه‌داده ویسکانسین و مجموعه‌داده WBCD استفاده می‌شود[۲۱]. همچنین از معیارهای دقت، حساسیت و ویژگی برای ارزیابی نتایج با روش اعتبارسنجی 10-fold استفاده می‌کنیم. پیچیدگی هر چه کمتر یک شبکه، باعث افزایش کیفیت در نتایج نهایی می‌شود. در تحقیق [۲۲] پیچیدگی یک شبکه (تعداد اتصالات) به صورت رابطه (۴) نشان داده شده است.

$$No.of\ connection = \alpha \times \beta + \beta \times Y + \beta + Y \quad (۴)$$

در این رابطه α، β و Y به ترتیب تعداد ویژگی انتخاب شده، اندازه نودهای لایه مخفی و اندازه لایه خروجی است. در تحقیق حاضر از این معیار نیز جهت ارزیابی روش پیشنهادی استفاده می‌شود.

مقادیر پارامترهای الگوریتم ژنتیک پیشنهادی بر مبنای روش ارائه شده در [۲۲] تنظیم شده است. برای مثال نرخ ترکیب (Cr) برابر ۰,۷، نرخ جهش (Mr) برابر ۰,۰۸ و اندازه جمعیت (NP) برابر ۱۵ می‌باشد. همچنین شبکه گرید در مدل FinGrain با ابعاد ۵×۵ فرض شده است.

روش‌های مختلف BP[2]، تفاوت‌های چشمگیری در دقت طبقه‌بندی شبکه عصبی MLP ایجاد می‌کنند. این تفاوت در دقت‌ها می‌تواند در یک مجموعه‌داده یکسان حاصل شود. بنابراین هنگام استفاده از شبکه عصبی، مهم است روش‌های مختلف BP بررسی شوند.

در جدول ۱ عملکرد الگوریتم پیشنهادی با روش‌های BP مختلف RB[3]، LM[4] و GD[5] برای بدست آوردن وزن‌ها (در بخش انتخاب ویژگی) مقایسه شده است. هر آزمایش سه بار با حداکثر تعداد نسل‌های ۵، ۱۵، ۳۰ تکرار شده است. برای هر حداکثر اندازه نسل، بهترین و متوسط عملکردها طی ۱۰ اجرا مستقل، محاسبه شده است. نتایج نشان می‌دهند که برای همه حداکثر اندازه‌های نسل، میانگین و بهترین دقت طبقه‌بندی در روش GAANN-RP بالاترین است. علاوه بر دقت، تعداد اتصالات شبکه نیز برای هر حداکثر تعداد نسل مختلف در حالت میانگین محاسبه شده است.

با توجه به نتایج حاصل شده روش FGAMLP-RP تنها نیاز به ۱۵ نسل برای رسیدن به میانگین دقت ۹۸,۷۲٪ با پیچیدگی ۱۹,۷ دارد. روش FGAMLP-GD بالاترین دقت ۹۹,۰۱٪ را در حالت میانگین ایجاد نموده اما پیچیدگی آن ۳۲,۴ اتصال است.

جدول ۲ عملکرد مدل طبقه‌بندی نهایی را به‌صورت یک ماتریس بی‌نظمی نشان می‌دهد.

[1] Bit Change
[2] Back Propagation
[3] Resilient BackPropagation
[4] Levenberg-Marquardt
[5] Gradient Descent

جدول ۱-مقایسه عملکرد الگوریتم پیشنهادی با سه یادگیری مختلف BP

تعداد نسل	FGAMLP_RP		FGAMLP _LM		FGAMLP _GD	
	دقت (٪)	تعداد اتصالات	دقت (٪)	تعداد اتصالات	دقت (٪)	تعداد اتصالات
۵	۹۸.۲۹	۲۶.۰	۹۸.۲۴	۲۷.۳	۹۸.۲۴	۲۹.۹
۱۵	۹۸.۷۲	۱۹.۷	۹۸.۵۱	۳۲.۵	۹۸.۱۱	۳۱.۲
۳۰	۹۸.۴۳	۲۴.۱	۹۸.۸۶	۲۲.۷	۹۹.۰۱	۳۲.۴

نتایج به‌صورت مجموع در ۱۰ اجرای مجزا محاسبه شده است (برای حداکثر ۱۵ نسل). در عین حال حساسیت، ویژگی و دقت در جدول ۳ ارائه شده است. حساسیت در FGAMLP-RP بهترین نتیجه را گزارش می‌دهد. از طرفی، دقت در روش FGAMLP-LM برتری نسبی نسبت به سایر روش‌ها دارد. در مورد با زمان اجرا، روش FGAMLP-RP با کمترین پردازش زمانی در هر سه حداکثر تعداد نسل مختلف، بهترین است، سپس روش FGAMLP-LM و پس از آن روش FGAMLP-GD قرار دارد. نتایج زمان اجرا در جدول ۴ نشان داده شده است.

جدول ۲-نتایج ماتریس بی‌نظمی در ۱۰ اجرا

سه نوع مختلف BP	واقعی	تعداد	پیش‌بینی	
			بدخیم	خوش‌خیم
FGAMLP_RP	بدخیم	۲۴۱	۲۳۸	۵
	خوش‌خیم	۴۴۲	۴	۴۳۹
FGAMLP _LM	بدخیم	۲۴۱	۲۳۴	۴
	خوش‌خیم	۴۴۲	۵	۴۴۰
FGAMLP _GD	بدخیم	۲۴۱	۲۳۴	۷
	خوش‌خیم	۴۴۲	۵	۴۳۷

جدول ۳- عملکرد الگوریتم پیشنهادی در معیارهای مختلف

معیار ارزیابی	FGAMLP_RP	FGAMLP _LM	FGAMLP _GD
حساسیت (٪)	۹۷.۰۱	۹۷.۲۱	۹۷.۲۱
ویژگی (٪)	۹۹.۳۱	۹۸.۹۴	۹۸.۹۴
دقت (٪)	۹۸.۲۹	۹۸.۲۴	۹۸.۲۴

نتایج این آزمایش نشان می‌دهد که دقت الگوریتم پیشنهادی در بهینه‌سازی وزن‌ها و انتخاب ویژگی مناسب است، اما دراین‌حالت تعداد اتصالات (پیچیدگی شبکه) حدود دو برابر می‌باشد. این مورد اهمیت اجرای فرایند بهینه‌سازی را هم برای انتخاب ویژگی و هم برای اندازه نودهای لایه مخفی به طور همزمان نشان می‌دهد.

جدول ۴-عملکرد الگوریتم پیشنهادی در میانگین زمان اجرا

تعداد نسل	میانگین زمان اجرا (ثانیه)		
	FGAMLP_RP	FGAMLP _LM	FGAMLP _GD
۵	۴۶.۷	۸۵.۸	۱۲۴.۲
۱۵	۴۲۰.۳	۵۱۸.۴	۷۱۲.۳
۳۰	۷۰۹۸.۹	۸۴۳۸.۱	۱۲۸۴۰.۵

در آزمایش بعد کارایی تاثیر مرحله انتخاب ویژگی (FS[۱]) در الگوریتم پیشنهادی بررسی می‌شود. نتایج برای روش FGAMLP_RP محاسبه شده است (زیرا RP بهترین نتیجه را ایجاد نموده است). در این آزمایش علاوه بر اعمال FS، نتایج بدون FS نیز بررسی شده تا جایگاه بخش انتخاب ویژگی در الگوریتم پیشنهادی مشخص شود. از اینرو، ژن-های بخش زیرمجموعه انتخاب شده حذف می‌گردد و تمام راه‌حل‌ها با ویژگی‌های کامل بهینه می‌شوند. بنابراین همه ویژگی‌ها در زیرمجموعه داده‌های آموزشی و آزمایشی قابل دسترس هستند و الگوریتم ژنتیک تنها برای بهینه‌سازی تعداد نودهای لایه مخفی و الگوریتم ژنتیک موازی FinGrain برای جستجو بهترین وزن‌ها استفاده می‌شود. با توجه به جدول ۵ می‌توان مشاهده نمود، برای همه اندازه-های نسل، FS نه تنها دقت شبکه بلکه پیچیدگی آن را نیز کاهش می‌دهد. الگوریتم FGAMLP_RP با FS، مقدار بالاتر نودهای لایه پنهان را تولید می‌نماید (با تعداد نسل ۱۵ و ۳۰)، دلیل این امر اندازه کمتر ویژگی‌های انتخاب شده است. این درحالی است که تعداد اتصالات الگوریتم پیشنهادی، با FS هنوز بهتر از تعداد بدون FS است.

[۱] Features Selection

بهترین شبکه تولید شده از طریق این الگوریتم‌ها با و بدون FS در قالب ماتریس بی‌نظم در جدول ۶ مورد مقایسه قرار می‌گیرد. همچنین در این جدول، زیر مجموعه ویژگی‌های انتخاب شده قابل دسترس است.

جدول ۵-مقایسه عملکرد روش FGAMLP_RP با و بدون انتخاب ویژگی

تعداد نسل	میانگین تعداد نودهای لایه مخفی	میانگین تعداد ویژگی‌های انتخابی	میانگین اتصالات	میانگین دقت (٪)
FS با ۵	۱.۵	۶.۱	۲۶.۰	۹۸.۲۹
FS بدون ۵	۱.۷	۹.۰	۳۲.۵	۹۸.۰۶
FS با ۱۵	۱.۲	۶.۴	۱۹.۷	۹۸.۷۲
FS بدون ۱۵	۱.۱	۹.۰	۲۶.۳	۹۸.۲۳
FS با ۳۰	۱.۳	۵.۹	۲۴.۱	۹۸.۴۳
FS بدون ۳۰	۱.۲	۹.۰	۳۰.۶	۹۸.۲۰

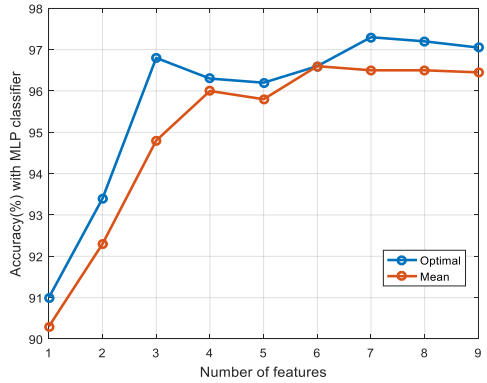
جدول ۶-ماتریس بی‌نظمی و زیرمجموعه ویژگی‌های انتخاب شده برای روش FGAMLP_RP

	واقعی	تعداد	پیش‌بینی		زیرمجموعه ویژگی‌های انتخاب شده
			خوش‌خیم	بدخیم	
با FS	خوش‌خیم	۴۸	۴۸	۱	$f_1,f_2,f_3,f_6,f_7,f_8,f_9$
	بدخیم	۲۱	۰	۲۰	
بدون FS	خوش‌خیم	۴۸	۴۷	۱	$f_1,f_2,f_3,f_4,f_5,f_6,f_7,f_8,f_9$ (همه ویژگی‌ها)
	بدخیم	۲۱	۱	۲۰	

جدول ۷-بررسی عملکرد روش FGAMLP_RP با تنظیم اندازه لایه مخفی به صورت دستی

تعداد نسل		تنظیم دستی اندازه نودهای لایه مخفی			
		۱	۲	۳	۴
۵	تعداد اتصالات	۱۴.۰	۱۹.۷	۳۲.۱	۴۲.۷
۵	دقت (٪)	۹۸.۲۰	۹۸.۸۹	۹۸.۱۸	۹۸.۰۶
۱۵	تعداد اتصالات	۱۴.۱	۲۰.۹	۳۰.۰	۳۹.۳
۱۵	دقت (٪)	۹۷.۶۸	۹۷.۹۹	۹۸.۳۳	۹۸.۷۲
۳۰	تعداد اتصالات	۱۳.۸	۲۱.۶	۲۸.۱	۳۹.۷
۳۰	دقت (٪)	۹۷.۵۴	۹۷.۹۴	۹۸.۴۳	۹۸.۰۶

در آزمایش دیگری اندازه نودهای لایه مخفی به‌صورت دستی در اندازه‌های ۱ تا ۴ بررسی شده است. نتایج این آزمایش در جدول ۷ گزارش شده است. این وضعیت نشان می‌دهد که الگوریتم ژنتیک تنها برای انتخاب ویژگی و الگوریتم ژنتیک موازی FinGrain برای بهینه‌سازی وزن‌ها مورد استفاده قرار می‌گیرد. بالاترین دقت بدست آمده از این آزمایش ۹۸,۸۹٪ است، آن هم در هنگامی که تعداد نودهای لایه مخفی و حداکثر اندازه نسل به ترتیب ۲ و ۵ تنظیم شده است.



شکل ۵-دقت با تعداد ویژگی‌های مختلف در FGAMLP_RP

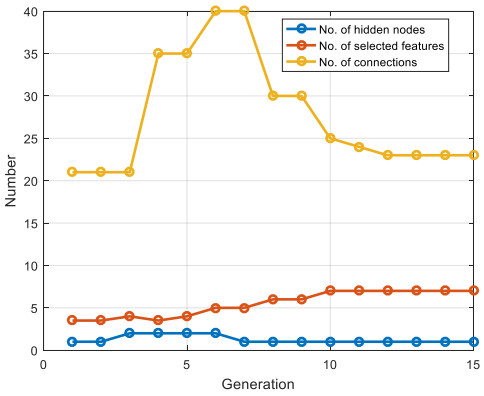
استفاده از تکنیک طول رشته متغیر در الگوریتم ژنتیک علاوه بر انتخاب ویژگی‌های موثر، تعداد بهینه این ویژگی‌ها را نیز مشخص می‌کند. شکل (۵) دقت الگوریتم پیشنهادی را با تعداد مختلف ویژگی‌ها گزارش می‌دهد.

محاسبه دقت برای تعداد ویژگی‌های مختلف در دو حالت میانگین و بهترین نشان می‌دهد که بهترین دقت طبقه‌بندی مجموعه‌داده WBCD با ۷ ویژگی ٪۹۷,۳۷ می‌باشد.

شکل (۶) اندازه‌های مختلف مدل شبکه عصبی نهایی را در روش FGAMLP_RP با حداکثر تعداد نسل ۱۵ نشان می‌دهد. در اینجا تعداد نودهای لایه مخفی، تعداد ویژگی‌های انتخاب شده و تعداد اتصالات شبکه در طول تکرار نسل گزارش شده است.

تصمیم‌گیری در مورد اینکه کدام نسل باید انتخاب شود تا الگوریتم متوقف گردد بسیار مهم است، زیرا این فرایند دقت نهایی شبکه و پیچیدگی آن را تعیین می‌کند. برای مثال، اگر تکرار قبل از نسل ۷ متوقف شود، بهترین راه‌حل دارای خطا MSE ۰/۰۱۶ می‌باشد، اما پیچیدگی با مقدار حدود ۴۰ بدترین است. در این میان، نسل ۱۲ دارای پیچیدگی حدود ۲۴ است در حالی‌که میزان خطا ۰/۰۱۶ می‌باشد. از اینرو مهم است عملکرد الگوریتم با تعداد نسل مختلف بررسی شود. با این وجود این وضعیت در این تحقیق به طور کامل بررسی نشده است.

نتایج حاصل از این تحقیق برای ارزیابی با دو روش CAFS [۲۳] و GAANN-RP [۲۲] مقایسه می‌شود. در CAFS به معرفی الگوریتمی مؤثر برای انتخاب ویژگی پرداخته شده است، بنابراین یک متدولوژی توسعه مجموعه‌داده مشابه با روش این تحقیق، جهت مقایسه انتخاب شده است.



شکل ۶-ارزیابی تعداد نودهای لایه مخفی، تعداد ویژگی‌های انتخاب شده و تعداد اتصالات شبکه

در مقایسه با الگوریتم پیشنهادی، CAFS شامل فرایند گروه‌بندی ویژگی‌ها (FG[۱]) براساس میزان همبستگی است که در آن قبل از پروسه یادگیری، انتخاب ویژگی به دو گروه تقسیم می‌گردد. در GAANN_RP یک الگوریتم انتخاب ویژگی مبتنی بر روش Wrapper طراحی شده است. در این الگوریتم علاوه بر انتخاب ویژگی‌ها، وزن‌ها و تعداد نودهای لایه مخفی شبکه را نیز با استفاده از یک الگوریتم ژنتیک سفارشی، جستجو می‌کند.

جدول ۸ عملکرد روش پیشنهادی FGAMLP_RP را در مقایسه با روش‌های GAANN-RP و CAFS نشان می‌دهد. دقت طبقه‌بندی در FGAMLP_RP بهتر از دو روش GAANN-RP و CAFS بدون گروه‌بندی ویژگی است. اما این دقت تا حدودی کمتر از CAFS با گروه‌بندی ویژگی‌ها است.

به‌طورخاص FGAMLP_RP شبکه فشرده‌تری نسبت به CAFS با گروه‌بندی ویژگی‌ها تولید می‌نماید که این مورد از نظر محاسباتی برتری قابل توجهی برای روش پیشنهای است.

جدول ۸-مقایسه مدل FGAMLP-RP با روش CAFS و GAANN_RP

روش‌ها	میانگین دقت (٪)	میانگین زمان اجرا (ثانیه)	میانگین تعداد نودهای لایه مخفی	میانگین تعداد ویژگی‌های انتخابی	میانگین اتصالات
CAFS با FG	۹۸.۷۶	۱۷	۱.۳۶	۶.۳۳	۱۴.۳۶
CAFS بدون FG	۹۶.۸۳	-	-	-	-
GAANN_RP	۹۸.۲۹	۴۲۸	۱.۴	۵.۱	۱۲.۳
FGAMLP_RP	۹۸.۷۲	۴۲۰	۱.۲	۶.۴	۱۹.۷

[۱] Features Selection

جدول ۹-مقایسه مدل FGAMLP-RP با سایر روش‌ها در مجموعه داده WBCD

نویسنده و سال انتشار	روش تحقیق	دقت طبقه‌بندی (٪)
FGAMLP_RP (حالت بهترین)	FinGrainGA+MLP	۹۹.۹۸
اوان، ۲۰۱۵ [۲۴]	Fuzzy+k-NN	۹۹.۷۱
پنگ، ۲۰۱۰ [۲۵]	CFW	۹۹.۵۰
مارکانو سدنو، ۲۰۱۱ [۲۶]	AMMLP	۹۹.۲۶
FGAMLP_RP (حالت میانگین)	FinGrainGA+MLP	۹۸.۷۲
شیخ‌پور، ۲۰۱۶ [۱۳]	PSO-KDE	۹۸.۴۵
احمد، ۲۰۱۵ [۲۲]	GA+ANN	۹۸.۲۹
کاراباتاک، ۲۰۰۹ [۲۷]	AR+NN	۹۷.۴۰
سالاما، ۲۰۱۲ [۲۸]	SMO, IBK, NB and J48	۹۷.۲۸
استوئین، ۲۰۱۳ [۲۹]	SVM+EA	۹۷.۰۷
چاورسیا، ۲۰۱۷ [۳۰]	SMO	۹۶.۱۹
نیلشی، ۲۰۱۷ [۳۱]	EM-PCA, Fuzzy	۹۳.۲

در جدول ۹ تلاش شده دقت طبقه‌بندی روش FGAMLP-RP و روش‌های مشابه پیشین که از مجموعه‌داده WBCD در آزمایش‌های خود استفاده کرده‌اند، مورد مقایسه قرار داده شود. با توجه به نتایج، میانگین دقت در روش FGAMLP-RP برای تشخیص سرطان در مقایسه با برخی از روش‌های مورد بررسی بیشتر است و در سایر موارد نیز روش پیشنهادی دقت مناسبی را ارائه می‌دهد. در نهایت، به منظور بررسی بیشتر الگوریتم پیشنهادی GAANN-RP، عملکرد آن را روی یکی دیگر از مجموعه‌داده‌های پایگاه داده ویسکانسین ارزیابی می‌کنیم. در جدول ۱۰ دقت طبقه‌بندی الگوریتم GAANN-RP و روش‌های مشابه که از مجموعه‌داده WDBC در آزمایش‌های خود استفاده می‌کنند، گزارش شده است. بر اساس نتایج مشاهده شده، روش پیشنهادی در سایر مجموعه‌داده‌ها از پایگاه‌داده ویسکانسین نیز نتایج قابل قبولی را ارائه می‌دهد، بطوریکه بعد از روش ODA + J48 و PSO-KDE در رتبه سوم قرار دارد.

جدول ۱۰-مقایسه مدل FGAMLP-RP با سایر روش‌ها در مجموعه داده WDBC

نویسنده و سال انتشار	روش تحقیق	دقت طبقه‌بندی (٪)
دیو، ۲۰۱۶ [۱۸]	ODA+J48	۹۹,۶۰
شیخ‌پور، ۲۰۱۶ [۱۳]	PSO-KDE	۹۸,۴۵
FGAMLP_RP (حالت بهترین)	FinGrainGA+MLP	۹۸,۳۸
FGAMLP_RP (حالت میانگین)	FinGrainGA+MLP	۹۷,۸۱
وانگ، ۲۰۱۸ [۳۲]	WAUCE	۹۷,۶۸
ژنگ، ۲۰۱۴ [۳۳]	K-SVM	۹۷,۳۸
پراساد، ۲۰۱۰ [۳۴]	PSO-SVM	۹۷,۳۷
پراساد، ۲۰۱۰ [۳۴]	GA-SVM	۹۷,۱۹
بامکان، ۲۰۱۴ [۳۵]	CfS+Logistic Regression	۹۶,۹۵
بامکان، ۲۰۱۴ [۳۵]	Filtered+Logistic Regression	۹۶,۶۲
پراساد، ۲۰۱۰ [۳۴]	ACO-SVM	۹۵,۹۶

۵-نتیجه گیری

در این تحقیق به منظور تشخیص سرطان سینه از یک مدل دو بخشی استفاده شده است. روش پیشنهادی، ترکیبی از الگوریتم ژنتیک و شبکه عصبی MLP می‌باشد. در بخش اول از یک الگوریتم ژنتیک کلاسیک برای جستجو تعداد ویژگی‌های بهینه و تعداد نودهای لایه مخفی در شبکه عصبی MLP استفاده شده است. با جستجو بهینه ویژگی-های موثر انتخاب شده به عنوان ورودی شبکه عصبی و تعداد نودهای لایه مخفی، یک شبکه با پیچیدگی مناسبی ایجاد می‌شود. در بخش دوم نیز از یک الگوریتم ژنتیک موازی FinGrain به منظور افزایش دقت طبقه‌بندی در داده‌های سرطان سینه با تنظیم وزن‌ها استفاده می‌شود. موازی‌سازی الگوریتم ژنتیک به منظور بهبود وزن‌های شبکه عصبی در کاهش خطا MSE است. عملگر تفاضل تکاملی و فاکتور مقیاسی پیشنهادی نقش موثری در بهبود راه‌حل‌ها و ایجاد وزن‌های بهینه داشته است.

یکی از دلایلی که وزن‌ها در بخش دوم بهینه‌سازی می-شوند، کاهش پیچیدگی زمانی الگوریتم می‌باشد. به‌طورکلی در روش‌های انتخاب ویژگی، نیاز به محاسبه متوالی برازندگی ویژگی‌های انتخاب شده وجود دارد، لذا استفاده از روش‌هایی که کمترین پیچیدگی را داشته باشند، در اولویت است. این امر امکان اجرای الگوریتم را روی پایگاه-های داده بسیار بزرگ فراهم می‌سازد و می‌تواند برای کارهای آتی انجام شود.

مراجع

[1] L. Al Shalabi, and Z. Shaaban, "Normalization as a preprocessing engine for data mining and the approach of preference matrix", Dependability of Computer Systems, 2006, DepCos-RELCOMEX'06, International Conference ,pp. 207-214.

[2] D. J. Slamon, G. Clark, S. Wong, W. Levin, A. Ullrich, and W. McGuire, "Human breast cancer", correlation of relapse and. Science, Vol. 3798106, No.177, 1987, pp. 235.

[3] A. Valachis, and C. Nilsson, "Cardiac risk in the treatment of breast cancer: assessment and management", Breast Cancer: Targets and Therapy, Vol. 7, 2015 ,p. 21.

[4] M. Nesrine, B. Mehdi, E. B. Houda, L. Soumaya, A. Mehdi, Z. Bechir, and B. Hamouda, "First site of recurrence after breast cancer adjuvant treatment in the era of multimodality therapy: which imaging for which patient during follow-up?" , Breast disease, (Preprint), 2018, pp. 1-10.

[5] National Comprehensive Cancer Network, "Breast cancer Clinical Practice Guidelines in Oncology", Journal of the National Comprehensive Cancer Network: JNCCN, Vol. 1, No. 2, 2013, p. 148.

[۶] زهرا مروج و جواد آذرخش، "شبیه سازی و طبقه بندی وقایع کیفیت توان با استفاده از شبکه عصبی"، فصلنامه مدل سازی در مهندسی، دوره ۱۳، شماره ۴۱، تابستان ۱۳۹۴، صفحه ۱۴۶-۱۳۷.

[۷] فرشاد حکیم پور، سیامک طلعت اهری و ابوالفضل رنجبر، "ارزیابی و مقایسه الگوریتم های بهینه سازی ژنتیک، شبیه سازی تبرید و فاخته ها در مکان یابی رقابتی تسهیلات (مطالعه موردی: بانکها)"، فصلنامه مدل سازی در مهندسی، دوره ۱۵، شماره ۴۸، بهار ۱۳۹۶، صفحه ۲۴۶-۲۳۱.

[۸] فاطمه کریمی زاد گوهری و اکبر شاهسوند، "مقایسه نتایج حاصل از شبکه های عصبی MLP و RBF در پیش بینی نتایج حاصل از همزمانی پدیده های انتقال جرم و انتقال حرارت"، فصلنامه مدل سازی در مهندسی، دوره ۱۱، شماره ۳۳، تابستان ۱۳۹۲، صفحه ۴۳-۲۷.

[9] Y. S. Cho, C. L. Chin, and K. C. Wang, "Based on fuzzy linear discriminant analysis for breast cancer mammography analysis", Technologies and Applications of Artificial Intelligence (TAAI), November 2011, pp. 57-61, IEEE.

[10] D. A. Schauer, and O. W. Linton, "National Council on Radiation Protection and Measurements report shows substantial medical exposure increase", Bethesda, 2009, pp. 293-29. IEEE.

[11] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008", International journal of cancer, Vol. 127, No. 12, 2010, pp. 2893-2917.

[12] S. K. Mandal, "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree", International Journal Of Engineering And Computer Science, Vol. 6, No. 2, 2017, pp. 20388-20391.

[13] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer", Applied Soft Computing, Vol. 40, 2016, pp. 113-131.

[14] R. Shen, Y. Yang, and F. Shao, "Intelligent breast cancer prediction model using data mining techniques", In Intelligent Human-Machine Systems and Cybernetics (IHMSC), Vol. 1, August 2014, pp. 384-387, IEEE.

[15] M. Ghoussaini, O. Fletcher, K. Michailidou, C. Turnbull, M. K. Schmidt, E. Dicks, and C. Baynes, "Genome-wide association analysis identifies three new breast cancer susceptibility loci", Nature genetics, Vol. 44, No. 3, 2012, p. 312.

[16] P. S. Pawar, and D. R. Patil, "Breast cancer detection using neural network models", Communication Systems and Network Technologies (CSNT), April 2013, pp. 568-572, IEEE.

[17] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method", Telematics and Informatics, Vol. 34, No. 4, 2017, pp. 133-144.

[18] R. D. H. Devi, and M. I. Devi, "Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer", International Journal of Advanced Enginneering Technology, Vol. VII/Issue II/April-June, Vol. 93, 2016, p. 98.

[19] K. J. Wang, B. Makond, K. H. Chen, and K. M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients", Applied Soft Computing, Vol. 20, 2016, pp. 15-24.

[20] D. E. Goldberg, "Genetic and evolutionary algorithms come of age", Communications of the ACM, Vol. 37, No. 3, 1994, pp. 113-120.

[21] Breast Cancer Wisconsin (Original) dataset, UCI machine language repository, 1992.

[22] F. Ahmad, N. A. M. Isa, Z. Hussain, M. K. Osman, and S. N. Sulaiman, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer", Pattern Analysis and Applications, Vol. 18, No. 4, 2015, pp. 861-870.

[23] N. M. Kabir, N. M. Islam, and K. Murase, "A new wrapper feature selection approach using neural network", Neurocomputing Vol. 73, No. 16, 2010, pp. 3273–3283.

[24] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer", Expert Systems with Applications, Vol. 42, No. 20, 2015, pp. 6844-6852.

[25] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification", Journal of Biomed Inform, Vol. 43, No. 1, 2010, pp. 15–23.

[26] A. Marcano-Ceden˜o, J. Quintanilla-Domı´nguez, and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network", Expert Systems with Applications, Vol. 38, No. 8, 2011, pp. 9573–9579.

[27] M. Karabatak, and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, Vol. 36 No. 2, 2009, pp. 3465–3469.

[28] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis", Computer Engineering & Systems (ICCES), November 2010, pp. 180-185, IEEE.

[29] R. Stoean, and C. Stoean, "Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection", Expert Syst Appl, Vol. 40, No. 7, 2013, pp. 2677–2686.

[30] V. Chaurasia, and S. Pal, "A novel approach for breast cancer detection using data mining techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, No. 1, 2017.

[31] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method", Telematics and Informatics, Vol. 34, No. 4, 2017, pp. 133-144.

[32] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis", European Journal of Operational Research, Vol. 267, No. 2, 2018, pp. 687-699.

[33] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms", Expert Systems with Applications, Vol. 41, No. 4, 2014, pp. 1476-1482.

[34] Y. Prasad, K. K. Biswas, and C. K. Jain, "SVM classifier based feature selection using GA, ACO and PSO for siRNA design", International conference in swarm intelligenceSpringer, Berlin, Heidelberg, 2010, pp. 307-314.

[35] S. M. H. Bamakan, and P. Gholami, "A novel feature selection method based on an integrated data envelopment analysis and entropy model", Procedia Computer Science, Vol. 31, pp. 632-638.