یک روش طبقه بندی ترکیبی برای تشخیص سرطان پستان با استفاده از الگوریتم ژنتیک و تنظیم چند مرحله ای وزنها در شبکه عصبی MLP

امین رضایی‌پناه [۱*]، سید جواد میرعابدینی [۲]، علی مبارکی [۳]

۱*: گروه مهندسی کامپیوتر، موسسه آموزش عالی رهجویان دانش برازجان، بوشهر، ایران amin.rezaeipanah@gmail.com
۲: گروه مهندسی کامپیوتر-نرم افزار، واحد تهران مرکز، دانشگاه آزاد اسلامی، تهران، ایران jvd2205@yahoo.com
۳: گروه مهندسی کامپیوتر-نرم افزار، واحد بوشهر، دانشگاه آزاد اسلامی، بوشهر، ایران am850301@gmail.com
تاریخ دریافت: ۱۳۹۷/۵/۳ تاریخ پذیرش: ۱۳۹۸/۱۲/۱۳

چکیده

امروزه با گسترش روز افزون علم، استفاده از سیستم‌های پشتیبان تصمیم می‌تواند کمک زیادی در سیاست‌های درمانی پزشک داشته باشد. بدین منظور استفاده از سیستم‌های هوشمند مصنوعی در پیش‌بینی و تشخیص سرطان پستان که یکی از رایج‌ترین سرطان‌ها در بین زنان می‌باشد، مورد توجه است. در این مقاله فرآیند تشخیص بیماری سرطان پستان با استفاده از تنظیم چند مرحله‌ای وزن‌ها در شبکه عصبی MLP در دو لایه انجام می‌شود. در لایه اول، سه طبقه‌بند وجود دارد که به طور همزمان روی داده‌های مجموعه یادگیری آموزش می‌بینند. پس از اتمام آموزش خروجی طبقه‌بندهای لایه اول جمع‌آوری شده و به همراه داده‌های مجموعه یادگیری در مجموعه‌ای جدید قرار می‌گیرند. این مجموعه به عنوان ورودی به فراطبقه‌بند لایه دوم داده می‌شود و فراطبقه‌بند نگاشت میان خروجی‌های هر یک از طبقه‌بندهای معمولی لایه اول را با کلاس‌های خروجی واقعی مدل می‌کند. ساختار سه طبقه‌بند لایه اول و همچنین فراطبقه‌بند لایه دوم یک شبکه عصبی MLP است که وزن‌ها، ویژگی‌های موثر و اندازه لایه مخفی در آن به طور همزمان با استفاده از یک الگوریتم ژنتیک ابتکاری بهینه‌سازی می‌شود. به منظور ارزیابی دقت مدل پیشنهادی از پایگاه داده ویسکانسین استفاده می‌شود که با تست FNA ایجاد شده است. نتایج آزمایش‌ها روی مجموعه داده WBCD دقت ۹۸,۷۲٪ را برای روش پیشنهادی نشان می‌دهد که نسبت به الگوریتم‌های GAANN، CAFS عملکرد بهتری ارائه داده است.

واژه های کلیدی: الگوریتم ژنتیک، ویژگی‌های موثر، فراطبقه‌بند، شبکه عصبی *MLP*، پایگاه داده ویسکانسین

۱- مقدمه

سرطان پستان به عنوان دومین عامل مرگ و میر زنان در جهان، اهمیت و جایگاه ویژه‌ای در میان بیماری‌های مربوط به بانوان دارد. بر این اساس و با توجه به ناشناخته بودن علت این بیماری تنها عامل موثر در روند درمان، تشخیص به موقع و شروع هرچه سریعتر درمان می‌باشد. با توجه به زمانگیر بودن فرآیند تشخیص توسط افراد کارشناس، استفاده از سیستم‌های کامپیوتری می‌تواند به سرعت بخشیدن و کاهش قابل توجه حجم کار عامل انسانی منجر شود. در این راستا در سال‌های اخیر کوشش‌های بسیاری در این زمینه صورت گرفته است که مهمترین آنها شامل بررسی تصاویر پزشکی و آسان‌سازی فرآیند تشخیص برای پزشک به کمک کامپیوتر و نیز تشخیص مستقیم به وسیله کامپیوتر بر اساس فاکتورهای موثر در این زمینه می‌باشد [1].

در این راستا چندین مرکز پژوهشی اقدام به جمع‌آوری اطلاعات و بررسی آنها کرده و حتی بعضی از آنها اقدام به قرار دادن این اطلاعات در اختیار عموم کرده تا از این طریق توسط محققین مختلف در سراسر دنیا روش‌های مختلف بررسی و اقدام برای بهبود این روش‌ها صورت پذیرد [2-5].

روش‌های مختلفی برای تشخیص این بیماری وجود دارد، اما تشخیص زود هنگام سرطان پستان برای ما اهمیت دارد. روش‌های تشخیصی گاهی به تنهایی و گاهاً به همراه یکدیگر به کار برده می‌شوند. معاینه بالینی توسط پزشک، BSE[1]، غربالگری با ماموگرافی، سونوگرافی، نمونه‌برداری و MRI[2] از روش‌های تشخیص زود هنگام سرطان پستان هستند[6]. تنها روش قطعی در تشخیص سرطان بیوپسی یا نمونه‌برداری می‌باشد، در این روش بخشی از تومور یا کل آن از پستان تخلیه می‌شود و با میکروسکوپ توسط متخصصین آسیب شناسی مورد بررسی قرار می‌گیرد. بر اساس شرایط بیمار و اندازه توده تحت بیهوشی عمومی یا بی‌حسی موضعی با روش جراحی نمونه‌برداری از توده انجام می‌شود[7]. محل تومور در پستان را برش داده و بخشی از تومور یا کل آن از پستان خارج شده و محل برش را بخیه می‌زنیم. تشخیص سرطان با بیوپسی دارای عوارض و دوره نقاهت حداقل ۱۰ روز را برای بیمار دارد. از اینرو نیاز به سیستم‌های هوشمند جهت تشخیص این بیماری ضروری می‌باشد.

این پژوهش بر اساس مجموعه داده‌های استاندارد UCI [8] به تشخیص سرطان پستان با استفاده از ترکیب دو روش شبکه عصبی MLP[3] و الگوریتم ژنتیک در دو لایه و تنظیم چند مرحله‌ای وزن‌ها می‌پردازد. مسئله مطرح شده در واقع به صورت یک مسئله طبقه-بندی، انتخاب ویژگی و تنظیم وزن‌های شبکه عصبی مطرح می‌باشد.

در ادامه این تحقیق به بررسی برخی از جدیدترین پژوهش‌های انجام شده در بخش ۲ می‌پردازیم، در بخش ۳ مدل پیشنهادی مبتنی بر طبقه‌بندی شبکه عصبی MLP با تظیم چند مرحله‌ای وزن‌ها به منظور تشخیص سرطان پستان مطرح شده و عملگرهای لازم ارائه می‌شود. نتایج حاصل از ارزیابی روش پیشنهادی در بخش ۴ آورده شده و در نهایت نتیجه‌گیری و پیشنهادات در بخش ۵ ذکر شده است.

۲- مروری بر پژوهش‌های انجام شده

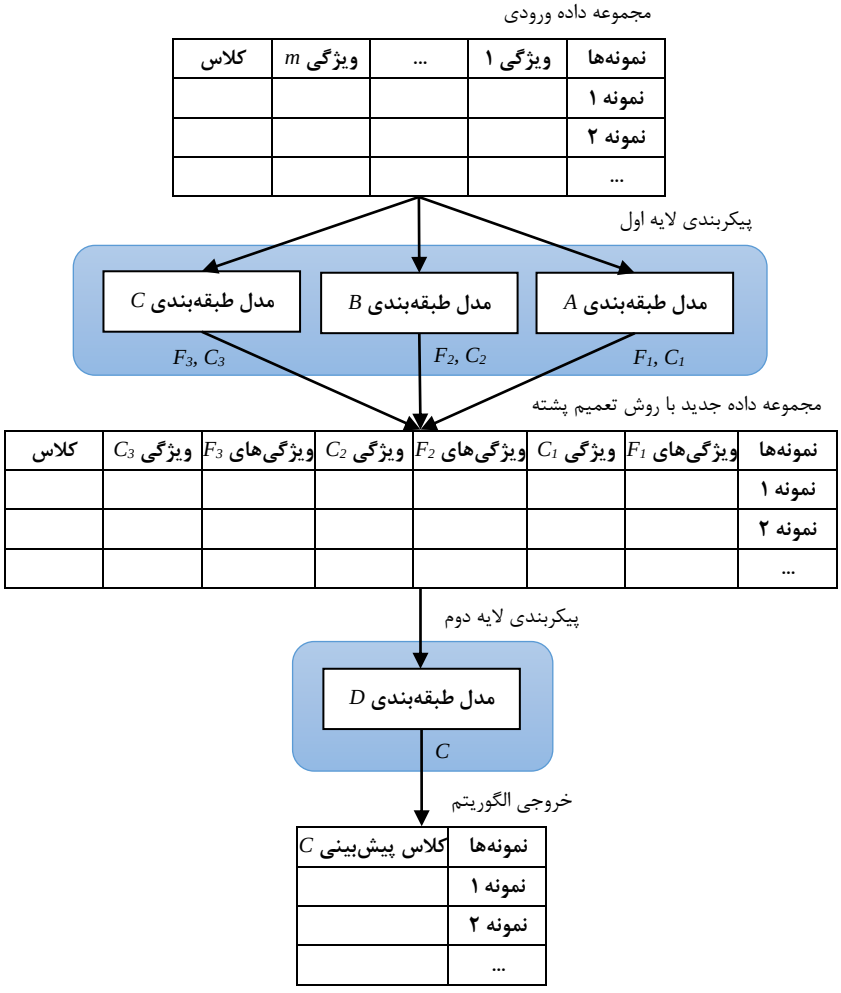
در این بخش تلاش بر آن داریم روش‌های مختلف ارائه شده در زمینه تشخیص سرطان پستان مورد بحث قرار گیرد. در ادامه به بررسی تعدادی از مقالات علمی که در این زمینه ارائه شده‌اند، می‌پردازیم.

مرزوقی و صفائی[9]، مدلی برای تشخیص سرطان پستان مبتنی بر شبکه‌های عصبی چندلایه پیشنهاد دادند[9]. این مدل توانایی طبقه‌بندی غدد سرطانی به دو گروه خوش‌خیم و بدخیم را به صورت خودکار دارد. نتایج بدست آمده حاکی از دقت ۹۸,۵٪ روی داده-های WBCD است. شیخ‌پور و همکاران[10]، انتخاب موثرترین ویژگی‌ها در تشخیص سرطان پستان را با استفاده از مدل‌های پارامتریک یادگیری ماشین معرفی کردند. انتخاب ویژگی با روش پیش‌رو و دسته‌بندی نوع تومور با انواع روش‌های پارامتریک مانند دسته‌بندی درجه دو، دسته‌بندی خطی و دسته‌بندی نزدیک‌ترین میانگین انجام گرفت.

بهمن‌یار و یثربی[11]، انتخاب ویژگی‌های موثر برای تشخیص سرطان پستان را با استفاده از یک الگوریتم شبکه عصبی با رویکرد تکاملی توسعه دادند. آنها یک الگوریتم ترکیبی جستجوی گرانشی و شبکه عصبی MLP را برای کلاس‌بندی مجموعه داده‌های ۶۸۳ بیمار مبتلا به سرطان پستان بکار بردند و به دقت حدود ۹۸٪ رسیدند. سندی و همکاران [12]، مقایسه و ارزیابی تکنیک‌های داده کاوی در جهت تشخیص بهتر سرطان پستان را ارائه دادند[12]. نتایج بدست آمده از این تحقیق نشان می‌دهد که روش بیزین یک روش قابل اطمینان در تشخیص سرطان است و با دقت ۹۷,۲٪ نسبت به دیگر روش‌ها عملکرد بهتری را گزارش می‌کند. پس از روش بیزین مشاهده شده که روش SMO و سپس به ترتیب روش‌های KNN، Baging و J48 در رتبه‌های بعد قرار می‌گیرند.

نیلاشی و همکاران[13]، سیستمی مبتنی بر دانش برای تشخیص سرطان پستان با استفاده از روش منطق فازی توسعه دادند***[13]***. در اینجا، بیماری سرطان پستان با استفاده از خوشه‌بندی، حذف‌نویز و تکنیک‌های طبقه‌بندی تشخیص داده می‌شود. از روش حداکثر انتظار[4] برای خوشه‌بندی داده‌ها و از طبقه‌بند درخت رگرسیون CART برای تولی قوانین فازی استفاده شده است. دوی و همکاران[14]، به منظور افزایش کارایی طبقه‌بندی J48 در تشخیص سرطان پستان، ویژگی‌ها را با استفاده از الگوریتم خوشه‌بندی Farthest First گروه‌بندی کرده و سپس از هر گروه موثرترین ویژگی‌ها را انتخاب می‌کنند[14]. نتایج پیاده‌سازی دقت حدود ۹۹٪ را برای مجموعه داده‌های WBCD و WDBC نشان می‌دهد. وانگ و همکاران[15]، ترکیبی از روش‌های Over Sampling جزئی (SMOTE) و بهینه‌سازی تجمعی ذره‌ای[5] (PSO) را برای تشخیص بیماران مبتلا به سرطان پستان بکار بردند[15]. در این روش برخی از مشهورترین روش‌های طبقه‌بندی مانند رگرسیون لجستیک، مدل درخت تصمیم گیری C5 و نزدیک‌ترین همسایه ادغام می‌شوند. نتایج آزمایش‌ها نشان می‌دهد که الگوریتم ترکیبی SMOTE+PSO+C5 بهترین عملکرد را دارد.

دیز و همکارانش[16]، یک رویکرد مبتنی بر داده‌کاوی برای غده‌شناسی در سرطان پستان ارائه دادند. این روش با استفاده از طبقه‌بندی چگالی پستان به شناسایی تمایزات در مجموعه داده‌ها کمک می‌کند. وایدهی و همکاران[17]، تشخیص سرطان پستان را با استفاده از طبقه‌بندی KNN ترکیبی ارائه دادند. KNN با سه ماتریس فاصله متفاوت یوکلیدان، کوزین، CITY-BLOCK و ترکیبات آن‌ها برای طبقه‌بندی مورداستفاده قرار می‌گیرد. آونان[18]، یک طبقه‌بندی نزدیکترین مجاور گنگ-سخت ترکیبی را با زیر مجموعه سازگار و گزینش موردی برای تشخیص اتوماتیک سرطان پستان مورد ارزیابی قرار می‌دهد[18]. شیخ‌پور و همکاران[19]، از الگوریتم بهینه‌سازی ذرات به منظور تعیین پهنای‌باند و انتخاب زیرمجموعه‌ای از ویژگی‌ها و همزمان از تخمین چگالی کرنل[۶] (KDE) برای طبقه‌بندی داده‌های سرطان پستان استفاده کردند[19]. مدل PSO-KDE معرفی شده دو الگوریتم PSO و KDE را با توجه به عملکرد طبقه‌بندی و تعداد ویژگی‌های انتخاب شده ارزیابی می‌کند. روش برآورد چگالی کرنل یک روش آماری بدون پارامتر است که شناسایی داده پرت در آن با مقایسه چگالی احتمال یک نمونه با نمونه‌های اطراف آن انجام می‌شود. در این مدل از تابع کرنل برای برآورد این چگالی استفاده می‌شود. در این صورت اگر برآورد چگالی کرنل برای نمونه‌ای بالا باشد، داده نرمال و در غیر اینصورت داده پرت محسوب می‌شود. علاوه بر این، الگوریتم PSO در این مدل برای کار انتخاب ویژگی‌ها استفاده شده است.



شکل ۱ : فلوچارت روش پیشنهادی

احمد و همکاران[20]، یک مدل تشخیص سرطان پستان مبتنی بر الگوریتم ژنتیک و شبکه عصبی ANN ارائه دادند[20]. در این روش از الگوریتم ژنتیک به طور همزمان به منظور انتخاب ویژگی و بهینه‌سازی پارامترهای شبکه عصبی ANN استفاده شده است. نتایج دقت بهترین ۹۹,۲۴٪ و میانگین ۹۸,۲۹٪ را در مجموعه داده WBCD نشان می‌دهد. کبیر و همکاران[21] مدل CAFS را بر مبنای شبکه عصبی ANN ارائه دادند[21]. CAFS به طور خودکار به تعیین تعداد نودهای لایه مخفی در طول فرآیند انتخاب ویژگی

می‌پردازد. بعلاوه CAFS شامل فرآیند گروه‌بندی ویژگی [24-22] است که ویژگی‌ها را براساس میزان همبستگی قبل از یادگیری و انتخاب ویژگی، به دو گروه تقسیم می‌کند و CAFS به گروه ویژگی که دارای همبستگی کمتر است تأکید دارد.

۳- روش پیشنهادی

در مدل طبقه‌بندی پیشنهادی از ترکیب الگوریتم ژنتیک و شبکه عصبی MLP استفاده می‌شود. شبکه عصبی MLP به عنوان ساختار مدل طبقه‌بندی و الگوریتم ژنتیک به منظور بهینه‌سازی ویژگی‌های انتخابی، وزن‌ها و اندازه لایه مخفی در شبکه عصبی MLP بکار گرفته می‌شود. شکل ۱ فلوچارت روش پیشنهادی را نشان می‌دهد.

فرآیند تشخیص بیماری سرطان پستان در دو لایه انجام می‌شود. در لایه اول ۳ طبقه بند مستقل وجود دارد که هر کدام به منظور بهینه‌سازی ویژگی‌های انتخابی، وزن‌ها و اندازه لایه مخفی در شبکه عصبی MLP از یک الگوریتم ژنتیک مجزا استفاده می‌کنند. طبقه بندهای لایه اول به صورت همزمان روی داده‌های مجموعه یادگیری آموزش می‌بینند. خروجی هر یک از طبقه‌بندها زیرمجموعه ویژگی‌های انتخابی (F) و کلاس خروجی نمونه‌ها (C) است که به عنوان ورودی لایه دوم در قالب یک مجموعه داده جدید در نظر گرفته می‌شود. در لایه دوم یک فراطبقه‌بند (طبقه‌بند ترکیب) وجود دارد که نگاشت میان خروجی‌های هر یک از طبقه‌بندهای معمولی لایه اول را با کلاس‌های خروجی واقعی یاد می‌گیرد.

۳-۱- ساختار لایه اول

تکنیک‌های مختلف یادگیری پس انتشار خطا[۷] (BP)، تفاوت‌های چشمگیری در دقت طبقه‌بندی شبکه عصبی MLP ایجاد می‌کنند[20]. این تفاوت در دقت‌ها می‌تواند در یک مجموعه داده یکسان حاصل شود. بنابراین هنگام استفاده از شبکه عصبی، مهم است روش‌های مختلف BP بررسی شوند. از اینرو در این تحقیق بجای استفاده از یک مدل طبقه‌بندی MLP، از سه مدل با BPهای مختلف استفاده می‌کنیم. مطابق شکل ۱، در لایه اول مجموعه داده ورودی به هر سه طبقه‌بند به صورت مستقل و همزمان اعمال می‌شود. جدول ۱ مشخصات سه طبقه‌بند A، B و C مورد بررسی را با تکنیک‌های مختلف یادگیری BP در شبکه عصبی MLP نشان می‌دهد.

جدول ۱ : مشخصات طبقه‌بندهای لایه اول

طبقه‌بند	تکنیک استفاده شده	معادل لاتین	نماد	توضیح
مدل طبقه‌بندی A	پس انتشار بازگشتی	Resilient Backpropagation	RB	علامت مشتق برای بروزرسانی وزن‌ها استفاده میشود و اندازه مشتق اثری بر بروزرسانی وزن‌ها ندارد.
مدل طبقه‌بندی B	لونبرگ-مارکارد	Levenberg-Marquardt	LM	همانند روش‌های شبه نیوتن این روش نیز سعی در کاهش محاسبات با استفاده از عدم محاسبه ماتریس هیسین دارد.
مدل طبقه‌بندی C	گرادیان نزولی	Gradient Descent	GD	برای پیدا کردن وزن‌ها از قوانین مشتق استفاده میکند. مشتق نشان دهنده شیب خط مماس بر یک نقطه از یک تابع است.

هر یک از مدل‌های لایه اول به صورت همزمان فرایند طبقه‌بندی را انجام می‌دهند. فرایند طبقه‌بندی در این مدل‌ها بر مبنای الگوریتم ژنتیک و شبکه عصبی MLP با توجه به جدول ۱ انجام می‌شود. الگوریتم ژنتیک در هر طبقه‌بندی وظیفه بهینه‌سازی ویژگی‌های انتخابی، تعداد نودهای لایه مخفی و همچنین وزن‌های شبکه را برعهده دارد. خروجی هر یک از طبقه‌بندها زیرمجموعه ویژگی‌های انتخابی (F) و کلاس خروجی نمونه‌ها (C) است که به عنوان ورودی به فراطبقه‌بند لایه دوم ارسال می‌شود تا پیشگویی‌های انجام شده توسط این مدل‌ها را برای کسب نتایج بهتر ترکیب کند.

۳-۲-ساختار لایه دوم

ورودی لایه دوم زیرمجموعه ویژگی‌های انتخابی و کلاس خروجی نمونه‌ها به ازای هر مدل طبقه‌بندی لایه اول است. ویژگی‌های انتخابی برای i-مین طبقه‌بند لایه اول به صورت بردار $F_{i,j}$ با مقادیر باینری به طول m در نظر گرفته می‌شود جائیکه m تعداد کل ویژگی‌ها در مجموعه داده را نشان می‌دهد ($j = 1,2, \dots, m$). بنابراین،

$$F_{i,j} = \begin{cases} 0 & if\ not\ selected\ feature \\ 1 & if\ selected\ feature \end{cases} \quad (۱)$$

به طور مشابه کلاس خروجی نمونه‌ها برای i-مین طبقه‌بند لایه اول به صورت بردار $C_{i,k}$ با مقادیر کلاس پیش‌بینی به طول n در نظر گرفته می‌شود جائیکه n تعداد کل نمونه‌ها در مجموعه داده را نشان می‌دهد ($k = 1,2, \dots, n$).

در این تحقیق خروجی‌های این سه طبقه‌بند در یک مجموعه داده جهت پیش‌بینی نهایی ترکیب می‌شوند. مجموعه داده جدید در ابعاد $n \times m'$ می‌باشد که m' تعداد ویژگی‌های درنظر گرفته شده را نشان می‌دهد. ویژگی‌های مجموعه داده جدید از اجتماع ویژگی-های انتخابی و همچنین کلاس نمونه خروجی از هر سه طبقه‌بند می‌باشد. رابطه زیر تعداد ویژگی‌های انتخابی برای مجموعه داده جدید را نشان می‌دهد.

$$m' = |combin[union(F_1, F_2, F_3), C_1, C_2, C_3]| \quad (۲)$$

در این رابطه $combin$ و $union$ به ترتیب بیانگر ترکیب و اجتماع هستند. F_i و C_i به ترتیب ویژگی‌های انتخابی و کلاس نمونه خروجی برای i-مین مدل طبقه‌بندی را نشان می‌دهد.

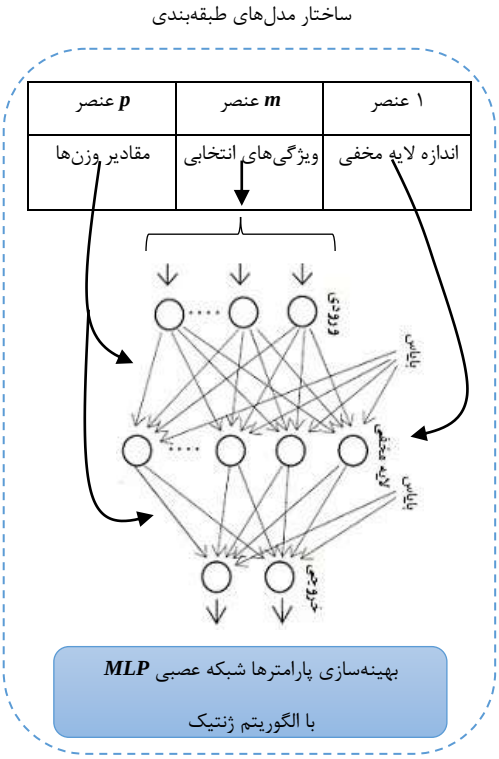
در این تحقیق جهت ترکیب خروجی‌های طبقه‌بندهای لایه اول، از یک فراطبقه‌بند با تکنیک تعمیم پشته[۸] در لایه دوم استفاده می‌کنیم. تکنیک تعمیم پشته یکی از روش‌های موجود جهت ترکیب چند طبقه‌بند و ایجاد یک طبقه‌بند گروهی می‌باشد و سعی در یادگیری خطای هر طبقه‌بند و بهبود کارایی کلی سیستم بر این اساس دارد[33]. در این تکنیک خروجی‌های هر یک از طبقه‌بندها، یک ورودی برای یادگیر سطح بالاتر در دیگر طبقه‌بندی‌ها می‌باشد. فرایند این تکنیک در شکل ۱ نشان داده شده است. تجربه نشان داده است که ترکیب پیش‌بینی‌های انجام شده توسط چند روش معمولاً پیش‌بینی‌های دقیق‌تری را نسبت به نتایجی که از یک روش بدست می‌آید، حاصل می‌کند.

تکنیک تعمیم پشته پیش‌بینی‌های انجام شده را جهت ایجاد یک طبقه‌بندی پیشگویانه بهینه نهایی در یک مجموعه داده جدید ترکیب می‌کند. از اینرو انتظار می‌رود دقت مدل ارائه شده نسبت به تکنیک رای اکثریت[36] بهتر شود، چون همزمان سه روش مختلف درک بیشتری از داده‌ها کسب کرده و پیش‌بینی‌های انجام شده توسط مدل‌ها را برای کسب نتایج بهتر ترکیب می‌کند.

مدل طبقه‌بندی استفاده شده در این لایه نیز یک شبکه عصبی MLP است که با BPهای مختلف بررسی می‌شود. فرایند طبقه-بندی در این لایه نیز مشابه لایه اول و بر مبنای الگوریتم ژنتیک به منظور بهینه‌سازی ویژگی‌های انتخابی، تعداد نودهای لایه مخفی و همچنین وزن‌های شبکه انجام می‌شود. با توجه به استفاده از تکنیک چند لایه و بهینه‌سازی چند مرحله‌ای وزن‌ها بر مبنای الگوریتم ژنتیک انتظار می‌رود دقت طبقه‌بندی نهایی بهبود یابد.

۳-۳-ساختار الگوریتم ژنتیک

به طور کلی هر چه تعداد ویژگی‌های یک پایگاه داده بالا باشد، ابعاد مسئله و تعداد پارامترهای طبقه‌بندی نیز بالا خواهد بود. در نتیجه انتخاب ویژگی‌های موثر باعث کاهش پیچیدگی و افزایش دقت الگوریتم‌های طبقه‌بندی می‌شود. در عین حال تعیین تعداد بهینه لایه مخفی باعث کسب نتایج بهتری می‌شود و در عین حال شبکه‌ای با پیچیدگی کمتری حاصل می‌نماید. همچنین فرایند یادگیری وزن‌ها به منظور افزایش دقت طبقه‌بندی داده‌های سرطان پستان ضروری می‌باشد. در مدل‌های طبقه‌بندی MLP پیشنهاد شده (در لایه اول و دوم)، از الگوریتم ژنتیک برای بهینه‌سازی همزمان سه فاکتور ویژگی‌های موثر ورودی، اندازه بهینه لایه مخفی و همچنین وزن‌های شبکه عصبی استفاده می‌کنیم. شکل ۲ ساختار بهینه‌سازی پارامترهای شبکه عصبی MLP را با الگوریتم ژنتیک نشان می‌دهد که این ساختار برای همه مدل‌های طبقه‌بندی یکسان می‌باشد.



شکل ۲ : ساختار بهینه‌سازی پارامترهای شبکه عصبی MLP را با الگوریتم ژنتیک

در اولین مرحله از الگوریتم ژنتیک مدل واقعی مساله مورد نظر به نحوه‌ای کدگذاری گردد که قابلیت استفاده در الگوریتم ژنتیک را داشته باشد. شکل ۳ ساختار نمایش کروموزوم‌ها را نشان می‌دهد.

عنصر p	عنصر m	عنصر ۱
مقادیر وزن‌ها	ویژگی‌های انتخابی	اندازه لایه مخفی

شکل ۳ : ساختار نمایش کروموزوم‌ها

در ساختار کروموزوم پیشنهادی، اندازه لایه مخفی در یک عنصر و با یک عدد صحیح بیان می‌شود. در بخش بعد ویژگی‌های انتخابی در m عنصر عنوان می‌شود. هر عنصر fi در این قسمت بیانگر ویژگی i-ام است که به صورن باینری (۰ عدم انتخاب ویژگی و ۱ انتخاب ویژگی) بیان می‌شود جائیکه $i = 1,2,\dots,m$ می‌باشد. در نهایت ساختار نمایش وزن‌ها در کروموزوم پیشنهادی برداری به طول p است که هر عنصر آن معرف یک وزن بین [-1,+1] در شبکه عصبی MLP می‌باشد. در فرایند تنظیم وزن‌ها بایاس در لایه‌های شبکه نیز به عنوان یک وزن در نظر گرفته می‌شوند. تعداد وزن‌ها با توجه به تعداد ورودی‌های (ویژگی‌ها) و تعداد نودهای لایه مخفی و همچنین اندازه خروجی تعیین می‌شود. با توجه به در نظر گرفتن یک شبکه سه لایه (ورودی-لایه مخفی-خروجی) تعداد وزن‌ها از مجموع دو بخش وزنی IW و LW و همچنین دو بایاس $B1$ و $B2$ توسط تابع شمارش وزن Γ مطابق رابطه زیر محاسبه می‌شود.

$$p = \Gamma(IW) + \Gamma(LW) + \Gamma(B1) + \Gamma(B2) \quad (۳)$$

جمعیت اولیه با توجه به ساختار کروموزوم پیشنهادی به صورت تصادفی تولید می‌شود. به منظور تعیین خودکار تعداد ویژگی‌های بهینه از تکنیک طول رشته متغیر برای نمایش کروموزوم‌ها استفاده شده است، از اینرو تعداد بیت‌های ۱ در کروموزوم‌های جمعیت متفاوت است. برای محاسبه کیفیت راه‌حل‌های تولید شده از معیار دقت طبقه‌بندی شبکه عصبی MLP با توجه به فاکتورهای ویژگی-های انتخاب شده، اندازه نودهای لایه مخفی و همچنین مقادیر وزن‌ها از هر کروموزوم استفاده می‌کنیم.

عملگرهای ژنتیکی در روش پیشنهادی به صورت زیر است :

- عملگر انتخاب : روش تورنومنت(مسابقه)[۹]
- عملگر ترکیب : برای هر بخش از کروموزوم، عملگر ترکیب به صورت مجزا اعمال می‌شود. در این تحقیق از عملگر ترکیب تفاضل تکاملی[۱۰] (DE) استفاده می‌کنیم که توسط الانی و همکاران ارائه شده است[34]. این عملگر با توجه به تفاوت میان ژن‌های کروموزوم‌های والد و همچنین بهترین کروموزم جمعیت (همسایگان) اقدام به تولید فرزند می‌کند. بهترین کروموزوم، عضوی از جمعیت با بیشترین دقت می‌باشد که برای کمک به همگرایی بهتر الگوریتم ژنتیک استفاده می‌شود. 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 زیر نحوه محاسبه X_i^{new} را نشان می‌دهد.

$$X_i^{new} = \begin{cases} X_i^0 + F \times (X_i^1 - X_i^2) & if\ C_r > Rand(0,1) \\ X_i^0 & otherwise \end{cases} \quad (۴)$$

در این رابطه F، یک فاکتور مقیاسی در محدوده [0-1] اسن که سرعت تکامل جمعیت را کنترل می‌کند و C_r احتمال ترکیب را نشان می‌دهد.

- عملگر جهش : برای هر بخش از کروموزوم، عملگر جهش به صورت مجزا روی کروموزوم فرزند ایجاد شده از عملگر ترکیب اعمال می‌شود. در این تحقیق از عملگر جهش تغییر بیت[۱۱] (BC) استفاده می‌کنیم که توسط لی و همکاران [35] برای الگوریتم ازدحام ذرات توسعه داده شده است[35]. این عملگر با احتمال M_r برای هر ژن از کروموزوم سعی در یافتن بهترین همسایه دارد. همسایه‌ها در یک محدوده اختلافی ثابت به تعداد تکرار ثابتی (در این تحقیق تعداد ۵ تکرار) جستجو می‌شوند. در هر تکرار در صورت افزایش دقت کروموزوم فرزند، تغییرات بروزرسانی می‌شود.

۴- نتایج و آزمایش‌ها

نتایج حاصل از شبیه‌سازی روش پیشنهادی با عنوان «GAMLP» در تمام آزمایش‌ها نشان داده شده است. در این تحقیق از الگوریتم‌های GAANN، CAFS و چند روش دیگر جهت مقایسه و ارزیابی عملکرد روش GAMLP استفاده می‌کنیم. برای انجام شبیه‌سازی و تجزیه و تحلیل روش پیشنهاد شده از نرم افزار Matlab ورژن 2017a روی مجموعه داده WBCD از پایگاه داده ویسکانسین استفاده شده است. تمامی آزمایش‌ها با استفاده از یک پردازنده اینتل ۵ هسته‌ای با فرکانس ۲,۴ گیگاهرتز، حافظه ۸ گیگابایت و سیستم عامل ویندوز ۱۰ نسخه Enterprise انجام گردیده است.

مجموعه داده WBCD دارای ۶۹۹ نمونه و ۹ ویژگی (ضخامت توده، یکنواختی اندازه سلول، چسبندگی حاشیه، اندازه سلول‌های اپیتلیال، هسته خالی، کروماتین مطلوب، هستک نرمال و تقسیم میتوز) است که ۴۵۸ نمونه مربوط به نوع سرطان خوش‌خیم و ۲۴۱ نمونه مربوط به نوع سرطان بدخیم است. مقادیر ۱۶ نمونه از این مجموعه داده گمشده[۱۲] است که در ارزیابی‌های انجام شده در این تحقیق از این نمونه‌ها استفاده نمی‌شود. بنابراین تعداد نمونه‌های استفاده شده در آزمایش‌ها ۶۸۳ می‌باشد.

در این تحقیق از تکنیک اعتبارسنجی 10-Fold برای ارزیابی مدل استفاده می‌شود. در هر مرحله از اعتبارسنجی 10-Fold، مجموعه داده اصلی به دو بخش E^T (آموزش) و E^P (آزمایش) تقسیم می‌شود جائیکه ۹۰٪ نمونه‌ها برای E^T و ۱۰٪ دیگر برای E^P استفاده می‌شود. از اینرو $E = E^T \cup E^P$ و $E^T \cap E^P = \emptyset$ است، جائیکه E کل مجموعه داده را نشان می‌دهد. همچنین برای ارزیابی نتایج از معیارهای دقت[۱۳]، حساسیت[۱۴]، ویژگی[۱۵] و پیچیدگی شبکه استفاده می‌کنیم که از مهمترین شاخص‌ها برای تعیین کارایی یک الگوریتم طبقه‌بندی هستند. معیار Accuracy نسبت مجموع نمونه‌های مثبت حقیقی و منفی حقیقی را به کل نمونه‌های بررسی شده را نشان می‌دهد. معیار Sensitivity نسبت نمونه‌های مثبت حقیقی را به مجموع نمونه‌های مثبت حقیقی و منفی کاذب در نظر می‌گیرد. به طور کلی این معیار توانایی مدل طبقه‌بندی را در پیش‌بینی درست کلاس مثبت نشان می‌دهد. معیار Specificity حاصل تقسیم نمونه‌های منفی واقعی به مجموع نمونه‌های منفی واقعی و مثبت کاذب است. به طور کلی این معیار توانایی مدل طبقه‌بندی را در پیش‌بینی درست کلاس منفی نشان می‌دهد.

با توجه به استفاده از شبکه عصبی MLP نیاز است اندازه آن را مورد بررسی قرار دهیم. احمد و همکاران [20]، در تحقیق خود پیچیدگی شبکه را به صورت تعداد پیوندهای کل شبکه (تعداد اتصالات) محاسبه می‌کنند[20]. کمتر بودن مقدار پیچیدگی شبکه باعث افزایش کیفیت در نتایج نهایی می‌شود. این معیار به صورت معادله زیر محاسبه می‌شود.

$$NC = \alpha \times \beta + \beta \times Y + \beta + Y \quad (۵)$$

در این رابطه α، β و Y به ترتیب تعداد ویژگی انتخاب شده (تعداد ورودی)، اندازه لایه مخفی و اندازه لایه خروجی است.

مارکانو و همکاران [26]، نشان داده‌اند که تکنیک‌های مختلف یادگیری پس انتشار خطا (BP) منجر به تفاوت‌های قابل توجهی در صحت مدل طبقه‌بندی MLP در یک مجموعه داده می‌شود. بنابراین، مهم است که الگوریتم پیشنهادی را با روش‌های مختلف BP ارزیابی کنیم. در این تحقیق عملکرد الگوریتم پیشنهادی با BPهای مختلف پس انتشار بازگشتی[۱۶] (RB)، لونبرگ-مارکارد[۱۷] (LM) و گرادیان نزولی[۱۸] (GD) به ترتیب با نام‌های GAMLP_RB، GAMLP_LM و GAMLP_GD برای یادگیری وزن‌ها در شبکه عصبی MLP مقایسه می‌شود.

تصمیم‌گیری در مورد اینکه کدام نسل باید انتخاب شود تا الگوریتم متوقف گردد بسیار مهم است، زیرا این فرآیند دقت نهایی شبکه و پیچیدگی آن را تعیین خواهد نمود. از اینرو در این تحقیق هر آزمایش سه بار تکرار می‌شود و در هر تکرار حداکثر تعداد نسل (Gen) به صورت ۵، ۱۵ و ۳۰ تنظیم می‌شود. به منظور حصول اطمینان از نتایج گزارش شده برای هر تعداد نسل، میانگین اجرای بیش از ۱۰ مرتبه اجرای مستقل لحاظ می‌شود. سایر پارامترهای استفاده شده در شبیه‌سازی به این صورت است؛ تعداد اعضاء جمعیت(NP) برابر ۳۵، نرخ ترکیب(CR) و جهش(MR) به ترتیب برابر ۰,۸۵ , ۰,۰۵ است. در بخش طبقه‌بندی MLP حداکثر اندازه لایه مخفی ۵ در نظر گرفته شده است. حداکثر اندازه لایه مخفی با توجه به بهترین نتایج شبکه عصبی MLP به صورت سعی و خطا تنظیم شده است.

جدول ۲ عملکرد الگوریتم پیشنهادی را با سه BP مختلف را مورد مقایسه قرار می‌دهد. برای همه حداکثر تعداد نسل‌ها، میانگین دقت طبقه‌بندی در روش GAMLP-RB بالاترین است و سپس روش‌های GAMLP-LM و GAMLP-GD قرار دارد. از اینرو در سایر آزمایش‌ها و انجام مقایسه‌ها نوع گرادیان نزولی از BP توسعه داده می‌شود.

جدول ۲ : مقایسه عملکرد الگوریتم پیشنهادی با سه یادگیری مختلف BP

Gen	GAMLP_RB				GAMLP _LM				GAMLP _GD			
	Sensitivity	Specificity	Accuracy	NC	Sensitivity	Specificity	Accuracy	NC	Sensitivity	Specificity	Accuracy	NC
5	97.01	99.31	98.29	26.0	97.21	98.94	98.24	27.3	97.21	98.94	98.24	29.9
15	97.22	99.53	98.72	19.7	97.22	99.20	98.51	32.5	97.01	98.73	98.11	31.2
30	97.24	99.54	98.73	20.1	97.21	99.23	98.56	22.7	97.19	98.93	98.25	32.4

با توجه به نتایج حاصل شده روش GAMLP-RB تنها نیاز به ۱۵ تکرار برای رسیدن به میانگین دقت ۹۸,۷۲٪ با تعداد اتصالات ۱۹,۷ می‌باشد. همچنین برتری نتایج در تعداد تکرار ۳۰ نسبت به ۱۵ بسیار ناچیزی دارد، از اینرو از تعداد تکرار ۱۵ جهت گزارش نتایج در مقایسه‌ها استفاده می‌کنیم.

جدول ۳ عملکرد طبقه‌بندی نهایی در قالب یک ماتریس بی‌نظمی را نشان می‌دهد. هر یک از نتایج به صورت مجموع در ۱۰ اجرای مجزا محاسبه شده است (برای حداکثر نسل ۱۵). بعلاوه زمان اجرا روش پیشنهادی با BPهای مختلف در جدول ۴ گزارش شده است. نتایج زمان اجرا سرعت همگرایی بهتر روش GAMLP_RB را در مقایسه با دو روش GAMLP _LM و GAMLP _GD نشان می‌دهد.

در آزمایش بعد کارایی تاثیر بخش انتخاب ویژگی (FS) در الگوریتم پیشنهادی GAMLP_RB نشان داده می‌شود. در این روش علاوه بر اعمال FS، نتایج بدون FS نیز مورد بررسی قرار گرفته است تا نقش انتخاب ویژگی در الگوریتم پیشنهادی مشخص شود. با توجه به جدول ۵ می‌توان مشاهده نمود، برای همه اندازه‌های نسل، FS نه تنها دقت شبکه بلکه پیچیدگی آن را نیز کاهش می‌دهد.

دلیل این امر اندازه کمتر ویژگی‌های انتخاب شده می‌باشد، این درحالی است که براساس رابطه (۵)، تعداد اتصالات الگوریتم پیشنهادی با FS بهتر از تعداد بدون FS است.

جدول ۳ : نتایج ماتریس بی‌نظمی برای الگوریتم پیشنهادی

Three different BP	Actual	Cases	Predicted (test outcome)	
			Malignant	Benign
GAMLP_RB	Malignant	241	238	5
	Benign	442	4	439
GAMLP _LM	Malignant	241	234	4
	Benign	442	5	440
GAMLP _GD	Malignant	241	234	7
	Benign	442	5	437

جدول ۴ : میانگین زمان اجرا

Gen	Average CPU time (s)		
	GAMLP_RB	GAMLP _LM	GAMLP _GD
5	46.7	85.8	124.2
15	420.3	518.4	712.3
30	7098.9	8438.1	12840.5

جدول ۵ : مقایسه عملکرد روش GAMLP_RB با و بدون انتخاب ویژگی

Gen	Avg #hidden nodes	Avg #selected features	NC	Accuracy
5 with FS	1.5	6.1	26.0	98.29
5 without FS	1.7	9.0	32.5	98.06
15 with FS	1.2	6.4	19.7	98.72
15 without FS	1.1	9.0	26.3	98.23
30 with FS	1.3	5.9	20.1	98.73
30 without FS	1.2	9.0	30.6	98.20

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

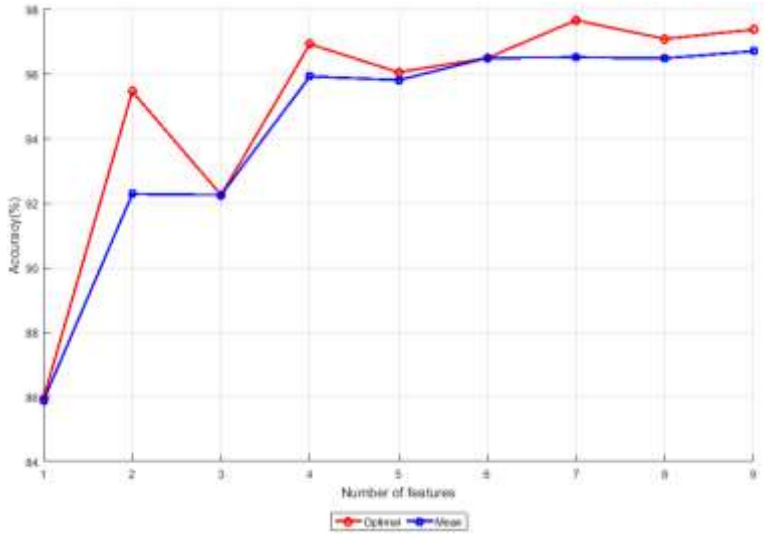
جدول ۶ : ماتریس بی‌نظمی و زیرمجموعه ویژگی‌های انتخاب شده برای روش GAMLP_RB

	Actual	Cases	Predicted (test outcome)		Selected feature subset
			Benign	Malignant	
With FS	Benign	48	48	1	f1,f2,f3,f6,f7,f8,f9
	Malignant	21	0	20	
Without FS	Benign	48	47	1	f1,f2,f3,f4,f5,f6,f7,f8,f9 (all)
	Malignant	21	1	20	

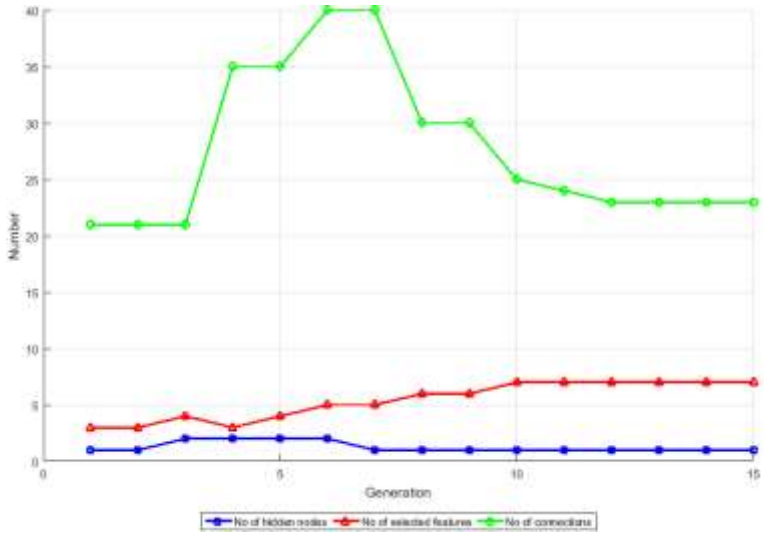
نتایج این آزمایش نشان می‌دهد که دقت الگوریتم پیشنهادی در زمان بهینه‌سازی وزن‌ها و انتخاب ویژگی خوب است اما در این حالت تعداد اتصالات (پیچیدگی شبکه) حدود دو برابر می‌باشد. این مورد اهمیت اجرای فرآیند بهینه‌سازی را هم برای انتخاب ویژگی و هم برای اندازه نودهای لایه مخفی به طور همزمان نشان می‌دهد.

تشخیص تعداد ویژگی‌های مطلوب توسط الگوریتم ژنتیک پیشنهادی بر مبنای تکنیک طول رشته متغیر انجام می‌شود. این تکنیک علاوه بر انتخاب ویژگی‌های مطلوب، تعداد بهینه این ویژگی‌ها را نیز مشخص می‌کند. در شکل ۴ نمودار دقت الگوریتم پیشنهادی در حالت یادگیری را با توجه به تعداد مختلف انتخاب ویژگی‌ها نشان داده شده است. نتایج برای روش GAMLP_RB و با تعداد نسل ۱۵ گزارش می‌شود (بهترین روش در آزمایش‌ها). محاسبه دقت برای تعداد ویژگی‌های مختلف در طول روند بهینه‌سازی و در دو حالت میانگین و بهترین نشان داده شده است.

نتایج حاصل از این آزمایش نشان می‌دهد که بهترین دقت طبقه‌بندی پایگاه داده WBCD در ۷ ویژگی و دقت ۹۷٫۸۱٪ می‌باشد. نمودار حاصل شده از روش GAMLP_RB و حداکثر تعداد نسل ۱۵ در شکل ۵، اندازه‌های مختلف در مدل شبکه عصبی نهایی را نشان می‌دهد. در اینجا تعداد نودهای لایه مخفی، تعداد ویژگی‌های انتخاب شده و پیچیدگی شبکه در طول حداکثر تکرار نسل ۱۵ گزارش شده است.



شکل ۴ : دقت طبقه‌بندی با تعداد ویژگی‌های مختلف در روش GAMLP_RB



شکل ۵ : ارزیابی تعداد نودهای لایه مخفی، تعداد ویژگی‌های انتخاب شده و پیچیدگی شبکه

نتایج حاصل از این تحقیق برای ارزیابی در ابتدا با دو روش CAFS و GAANN مقایسه می‌شود. CAFS در دو حالت با و بدون گروه‌بندی ویژگی‌ها[۱۹] (FG) بررسی شده است. جدول ۷ نتایج این مقایسه را نشان می‌دهد.

جدول ۷ : مقایسه مدل GAMLP-RB با روش‌های CAFS و GAANN

Methods	Accuracy (%)	Avg run time (s)	Avg #hidden nodes	Avg #selected features	NC
CAFS with FG	98.76	17	1.36	6.33	14.36
CAFS without FG	96.83	-	-	-	-
GAANN	98.29	428	1.4	5.1	12.3
GAMLP_RB	98.72	420	1.2	6.4	19.7

نتایج مقایسه نشان می‌دهد که دقت طبقه‌بندی در GAMLP_RB بهتر از دو روش GAANN و CAFS بدون FG است و به ترتیب بهبود ۰,۴۴٪ و ۱,۹٪ حاصل شده است. اما این دقت تا حدودی کمتر از CAFS با FG است. در هر حال GAMLP_RB شبکه فشرده‌تری نسبت به CAFS با FG تولید می‌نماید که این مورد از نظر محاسباتی برتری قابل توجهی برای روش پیشنهای است.

در جدول ۸ تلاش شده دقت طبقه‌بندی GAMLP-RB و روش‌های مشابه پیشین که از پایگاه داده WBCD در آزمایش‌های خود استفاده کرده‌اند، مورد مقایسه قرار داده شود. میانگین نتایج در GAMLP-RB برای تشخیص سرطان در مقایسه با برخی از روش‌های مورد بررسی دقت بیشتری داشته و در بقیه موارد نیز دقت مناسبی را ارائه می‌دهد.

جدول ۸ : مقایسه مدل GAMLP-RB با سایر روش‌های مشابه

First author, year	Methods	Classification accuracy (%)
Onan, 2015 [18]	Fuzzy+k-NN	99.71
Akay, 2009 [27]	SVM-CFS	99.51
Peng, 2010 [28]	CFW	99.50
Marcano-Ceden˜o, 2011 [29]	AMMLP	99.26
GAMLP_RB	GA+MLP	98.72
Ahmad, 2015 [20]	GA+ANN	98.29
Sheikhpour, 2016 [19]	PSO-KDE	98.45
Karabatak, 2009 [30]	AR+NN	97.40
Stoean, 2013 [31]	SVM+EA	97.07
Chaurasia, 2017 [32]	SMO	96.19

۵- نتیجه‌گیری و پیشنهادات آتی

پیش‌بینی و تشخیص بیماری‌های مختلف دو مولفه تاثیرگذار علوم پزشکی می‌باشند. در این تحقیق فرآیند تشخیص بیماری سرطان پستان که یکی از رایج‌ترین سرطان‌ها در بین زنان می‌باشد، با رویکردی جدید مطرح شده است. مدل پیشنهادی با تنظیم چند مرحله‌ای وزن‌ها در شبکه عصبی MLP در دو لایه و ترکیب آن با الگوریتم ژنتیک سعی در افزایش دقت طبقه‌بندی دارد. الگوریتم ژنتیک به منظور بهینه‌سازی همزمان ویژگی‌های انتخابی، تعداد نودهای لایه مخفی و همچنین وزن‌های شبکه استفاده می‌شود. در لایه اول، سه طبقه‌بند و در لایه دوم یک فراطبقه‌بند وجود دارد و برای ترکیب خروجی‌های طبقه‌بندهای لایه اول، از یک فراطبقه‌بند با تکنیک تعمیم پشته در لایه دوم استفاده می‌کنیم. در واقع در این روش ابتدا اجماعی از طبقه‌بندها ساخته می‌شود که خروجی آنها به عنوان ورودی فراطبقه‌بند لایه دوم به شمار می‌آید، سپس فراطبقه‌بند نگاشت میان خروجی‌های هر یک از طبقه‌بندهای معمولی لایه

اول را با کلاس‌های خروجی واقعی یاد می‌گیرد. نتایج آزمایش‌ها روی مجموعه داده WBCD برتری ۰,۴۴ و ۱,۹ درصد را به ترتیب در مقایسه با دو الگوریتم‌های GAANN، CAFS نشان می‌دهد.

بهبود الگوریتم ژنتیک می‌تواند به عنوان پیشنهادات آتی در ادامه کار این تحقیق انجام شود. برای مثال می‌توان از یک الگوریتم جستجو محلی برای بالا بردن دقت و کارایی کروموزوم‌های تولید شده استفاده کرد.

مراجع

[1] A. Antoniou, P. D. P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper and B. Pasini, "Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies", The American Journal of Human Genetics, vol.72, no.5, pp.1117-1130, 2003.

[2] L. C. Hartmann, D. J. Schaid, J. E. Woods, T. P. Crotty, J. L. Myers, P. G. Arnold and, M. H. Frost, "Efficacy of bilateral prophylactic mastectomy in women with a family history of breast cancer", New England Journal of Medicine, vol.340no. 2, pp.77-84, 1999.

[3], M. C. King, , J. H. Marks and J. B. Mandell, "Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2", Science, vol.302, no.5645, pp.643-646, 2003.

[4] C. DeSantis., J. Ma, L. Bryan and, A. Jemal, "Breast cancer statistics, 2013", CA: a cancer journal for clinicians, vol.64, no.1, pp.52-62, 2014.

[5] K. M. Kash, J. C. Holland, M. S. Halper and, D. G. Miller, "Psychological distress and surveillance behaviors of women with a family history of breast cancer". JNCI: Journal of the National Cancer Institute, vol.84, no.1, pp.24-30, 1992.

[6], R. L. Siegel, , K. D. Miller and, A. Jemal, "Cancer statistics, 2015", CA: a cancer journal for clinicians, vol.65, no.1, pp.5-29, 2015.

[7] E. C. Fear, X. Li, S. C. Hagness and, M. A. Stuchly, "Confocal microwave imaging for breast cancer detection: Localization of tumors in three dimensions", IEEE Transactions on Biomedical Engineering, vol.49, no.8, pp.812-822, 2002.

[8] A. Asuncion and Newman, DUCI machine learning repository, (2007).

[9] مرزوقی, فاطمه و علی اصغر صفائی (۱۳۹۵)، مدلی برای تشخیص سرطان سینه مبتنی بر شبکه های عصبی، کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات، تهران، دبیرخانه دایمی کنفرانس.

[10] شیخ پور, راضیه و مهدی آقاصرام (۱۳۹۴)، انتخاب ویژگی‌های موثر در تشخیص سرطان سینه با استفاده از مدل‌های پارامتریک یادگیری ماشین. فصلنامه علمی-پژوهشی بیماری های سینه, ۸(۲), ۱۶-۲۳.

[11] بهمن یار, حسن و بهزاد یثربی (۱۳۹۶)، انتخاب ویژگی های موثر برای تشخیص سرطان سینه با استفاده از الگوریتم شبکه عصبی مصنوعی با رویکرد تکاملی، کنفرانس ملی پژوهش های نوین در برق، کامپیوتر و مهندسی پزشکی، کازرون، دانشگاه آزاد اسلامی واحد کازرون.

[12] سندی, فاطمه؛ الهام عسکری؛ نرجس مطهری و پرستو شهابی چروده (۱۳۹۵)، مقایسه و ارزیابی تکنیک های داده کاوی در جهت تشخیص بهتر سرطان سینه، دومین کنفرانس بین المللی مدیریت و فناوری اطلاعات و ارتباطات، تهران، شرکت خدمات برتر.

[13] M. Nilashi, O. Ibrahim, H. Ahmadi and , L Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method", Telematics and Informatics, 34(4), pp.133-144, 2017.

[14], R. D. H. Devi, and, M. I. Devi, "Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer". Int J Adv Engg Tech, vol. 6,no.2 , pp. 93-98, 2016.

[15] 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, 2016, pp.15-24.

[16] J. Diz, G. Marreiros, and A. Freitas, "Applying Data Mining Techniques to Improve Breast Cancer Diagnosis". Journal of medical systems, vol.40, no.9, pp. 203-210, 2016.

[17] K. Vaidehi, and, T. S. Subashini, "Breast tissue characterization using combined K-NN classifier", Indian Journal of Science and Technology, vol.8,no.1, pp. 23-26, 2015.

[18] 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, pp.6844-6852, 2015.

[19] 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, pp.113-131, 2016.

[20] 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 Application, vol.18, no.4, pp.861-870, 2015.

[21] MM. Kabir, MM. Islam and K. Murase, "A new wrapper feature selection approach using neural network", Neurocomputing,vol. 73, pp. 3273–3283, 2010.

[22] G. L. Scott, and H. C. Longuet-Higgins, "Feature grouping by'relocalisation'of eigenvectors of the proximity matrix". in BMVC,1990, pp. 1-6.

[23] Z. Kim and R. Nevatia, "Uncertain reasoning and learning for feature grouping", Computer Vision and Image Understanding, vol.76, no.3, pp.278-288, 1999.

[24] G. Gan, and M. K. P. Ng, "Subspace clustering with automatic feature grouping" Pattern Recognition, vol.48, no.11, pp. 3703-3713, 2015.

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

[26] A. Marcano-Cedeño, J. Quintanilla-Domínguez and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network", Expert Systems with Applications, vol.38, no.8, pp. 9573-9579, 2011.

[27] MF. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis", Expert Syst Appl, vol.36, no.2, pp. 3240–3247, 2009.

[28] Y. Peng, Z. Wu, J. Jiang, "A novel feature selection approach for biomedical data classification", J Biomed Inform, vol.43, no.1, pp.15–23, 2010.

[29] A. Marcano-Ceden˜o, J. Quintanilla-Domı´nguez and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network", Expert Syst Appl, vol.38, no.8, pp.9573–9579, 2011.

[30] M. Karabatak, MC. Ince, "An expert system for detection of breast cancer based on association rules and neural network", Expert Syst Appl, vol.36, no.2, pp. 3465–3469, 2009.

[31] R. Stoean, 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, pp.2677–2686, 2013.

[32], V. Chaurasia, and, S. Pal, (2017). "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.

[33] D. H. Wolpert, "Stacked generalization", Neural networks, vol.5, no.2, 241-259, 1992.

[34] A. Al-Ani, A. Alsukker, and R. N. Khushaba, "Feature subset selection using differential evolution and a wheel based search strategy". Swarm and Evolutionary Computation, vol.9, pp.15-26, 2013.

[35] S. Lee, H. Park, and M. Jeon, "Binary particle swarm optimization with bit change mutation, IEICE transactions on fundamentals of electronics, communications and computer sciences, vol.90, no.10, pp.2253-2256, 2007.

[36] G. Tsoumakas, I. Katakis and I. Vlahavas, "Effective voting of heterogeneous classifiers". in European Conference on Machine Learning, 2004, pp. 465-476.

An Ensemble Classifier Method for Breast Cancer Detection Using Genetic Algorithm and Multistage Adjustment of Weights in the MLP Neural Network

Amin Rezaeipanah[1*], Seyed Javad Mir-Abedini [2], Ali Mobaraki [3]

1*: Department of Computer, Faculty of Computer Science, University of Rahjuyan Danesh Borazjan, Bushehr, Iran
2: Department of Computer Engineering-Software, Central Tehran Branch, Islamic Azad University, Tehran, Iran
3: Department of Computer Engineering-Software, Bushehr Branch, Islamic Azad University, Bushehr, Iran

ABSTRACT:

Today, with the increasing spread of science, the use of decision support systems can be of great help in the therapeutic policies of the Doctor. For this purpose, the use of artificial intelligence systems in predicting and diagnosing breast cancer, which is one of the most common cancers among women, is being considered. In this study, the process of diagnosis of breast cancer is done by using multistage weights in the MLP neural network in two layers. In the first layer, the three classifiers are trained simultaneously on the learning set data. Upon completion of the training, the output of the classifier of the first layer is accumulated together with the learning set data in the new sets. This set is given as an input to the second layer superconductor, and the supra-class mapping maps between the outputs of each of the ordinary classifiers of the first layer with the actual output classes. The three-layer structure of the first layer, as well as the second-layer supraclavicle, is a MLP neural network that optimizes the weights, effective properties and the size of the hidden layer simultaneously using an innovative genetic algorithm. In order to evaluate the accuracy of the proposed model, the Wisconsin database is used, which was created by the FNA test. Experiment results on the WBCD dataset the accuracy is 98.72% for the proposed method, which is relative to GAANN, CAFS algorithms provide better performance.

KEYWORDS: Genetic Algorithm, Effective Features, Meta Classifier, MLP Neural Network, Wisconsin Database.

[۱] *Bombay Stock Exchange*

[۲] *Magnetic resonance imaging*

[۳] *Multi-Layer Perceptron*

[۴] *Expectation Maximization*

[۵] *Particle Swarm Optimization*

[۶] *Kernel Density Estimation*

[۷] *Back Propagation*

[۸] *Stacked Generalization*

[۹] *Tournament*

[۱۰] *Differential Evolution*

[۱۱] *Bit Change*

[۱۲] *Missing Values*

[۱۳] *Accuracy*

[۱۴] *sensitivity*

[۱۵] *specificity*

[۱۶] *Resilient Backpropagation*

[۱۷] *Levenberg-Marquardt*

[۱۸] *Gradient Descent*

[۱۹] *Feature Grouping*