مقاله پژوهشی

تشخیص سرطان سینه با رویکرد متوازن‌سازی مجموعه داده‌ها

زینب عباسی

چکیده: یکی از چالش‌های بزرگ در تشخیص خودکار بیماری‌ها، وجود مجموعه داده‌های نامتوازن است. عدم توازن در کلاس‌های داده، باعث شکست در تشخیص صحیح بیماری توسط سیستم‌های تشخیصی می‌شود. این پژوهش الگوریتم جدیدی برای انتخاب و متوازن سازی نمونه‌ها پیشنهاد داده که بر پایه الگوریتم ReliefF، یک الگوریتم انتخاب ویژگی، است. در الگوریتم پیشنهادی، ابتدا نمونه‌ها بر اساس شاخص مشابهت با نمونه‌های همکلاسی و کلاس مخالف، وزن‌دهی می‌شوند. پس از رتبه‌بندی نمونه‌ها بر اساس وزن آن‌ها، مجموعه داده با استفاده از روش نمونه‌برداری فزاینده‌متوازن می‌شود. الگوریتم ارائه شده توانایی کار با مجموعه داده‌های چند کلاسه و انواع داده‌های رشته‌ای و عددی و وجود مقادیر مفقود را دارد. علاوه بر این، به دلیل امکان انجام محاسبات به طور موازی برای هر نمونه، سربار محاسباتی کمتری نسبت به سایر الگوریتم‌های متوازن‌سازی دارد. این الگوریتم می‌تواند داده‌ها را به طور کامل متوازن کرده و نمونه‌های با اهمیت را تکثیر کند. الگوریتم پیشنهادی روی سه مجموعه سرطان سینه ویسکانسین (WBCD)، مجموعه تشخیصی سرطان سینه ویسکانسین (WDBCD) و مجموعه سرطان سینه SEER اجرا شده است و سپس مجموعه‌های متوازن شده با الگوریتم‌های مختلف طبقه‌بندی شدند. نتایج طبقه‌بندی نشان‌دهنده کارایی روش پیشنهادی و افزایش صحت تشخیص بیماری هستند.

کلیدواژه: **داده‌های نامتوازن، نمونه‌برداری فزاینده، oversampling، تشخیص خودکار بیماری.**

۱- مقدمه

تشخیص صحیح بیماران از افراد سالم، به ویژه برای بیماری‌های خطرناک و شایع، مهم‌ترین چالش برای سیستم‌های تشخیص پزشکی مبتنی بر کامپیوتر[1] (CAD) است. یکی از خطرناک‌ترین بیماری‌های شایع در زنان، سرطان سینه است. طبق آخرین آمارهای سازمان بین‌المللی تحقیقات سرطان[2] (IARC)، در سال ۲۰۲۲ تعداد ۲,۲۹۶,۸۴۰ مورد جدید از سرطان سینه تشخیص داده شد [۱]. شیوع سرطان سینه و میزان مرگ‌ومیر آن در جهان به سرعت در حال افزایش است [۲]. تشخیص به‌موقع این بیماری می‌تواند در درمان آن بسیار مؤثر باشد. هوش مصنوعی و ابزارهای یادگیری ماشین می‌توانند به طراحی الگوریتم‌های جدید و تشخیص زودهنگام بیماری کمک کنند. ویژگی‌های جمع‌آوری‌شده از افراد سالم و بیمار برای آموزش سیستم تشخیص استفاده می‌شود. یک سیستم CAD بر اساس دانش خود و اطلاعات داده‌شده از یک فرد، تصمیم می‌گیرد که او بیمار یا سالم است. علاوه بر این، تشخیص صحیح یک بیمار تنها نقطه پایان نیست و اطلاعات بیماران می‌تواند دیدگاه‌های جدیدی را در مورد درمان بیماری‌ها به پزشکان ارائه دهد [۳].

معمولاً در یک مجموعه داده جمع‌آوری‌شده، تعداد افراد سالم بیشتر از تعداد افراد بیمار است. بنابراین، مجموعه داده‌ها، اطلاعات کافی برای آموزش سیستم CAD را ندارد. در این نوع مجموعه داده‌ها که به آن‌ها "مجموعه داده‌های نامتوازن" گفته می‌شود، طبقه‌بند معمولاً قادر به پیش‌بینی ورودی‌های جدید نبوده و صحت طبقه‌بندی کاهش می‌یابد [۴]. در یک مجموعه داده نامتوازن، برخی کلاس‌ها نمونه‌های بیشتری نسبت به سایر کلاس‌ها دارند [۵]. حتی گاهی اوقات نسبت بین کلاس‌ها ممکن است ۱ به ۱۰۰ یا بیشتر باشد [۶]. کلاسی که نمونه‌های بیشتری دارد، کلاس اکثریت نامیده می‌شود و کلاسی که نمونه‌های کمتری دارد، کلاس اقلیت نامیده می‌شود. متأسفانه، در بسیاری از موارد در دنیای واقعی، کلاس اقلیت داده‌های مهم‌تری دارد [۷]. مجموعه داده‌های نامتوازن محدود به حوزه پزشکی نیستند. بسیاری از مجموعه داده‌های جمع‌آوری‌شده از پدیده‌های طبیعی و مشکلات دنیای واقعی نامتوازن هستند، مانند طبقه‌بندی متون [۸]، بازیابی داده‌ها، شناسایی تقلب [۹]، تعیین اعتبار مشتریان برای پرداخت وام‌ها، و مدیریت مخابرات. برخی از مثال‌های این حوزه‌ها در [۱۰] تا [۱۲] قابل مشاهده هستند.

محققان تکنیک‌های حل مشکل مجموعه داده‌های نامتوازن را به سه دسته کلی تقسیم می‌کنند [۱۳] و [۱۴]: روش‌های سطح الگوریتم، روش‌های سطح داده، و روش‌های ترکیبی.

بخشی از روش‌های سطح داده، تکنیک‌های نمونه‌گیری هستند که به دو دسته نمونه‌برداری کاهنده[3] و نمونه‌برداری فزاینده[4] تقسیم می‌شوند. روش‌های نمونه‌برداری کاهنده با حذف برخی از نمونه‌ها از کلاس اکثریت، مجموعه داده را متوازن می‌کنند. در مقابل، روش‌های نمونه‌برداری فزاینده با اضافه کردن برخی نمونه‌های جدید به کلاس اقلیت، داده‌ها را متوازن می‌کنند [۱۵].

به دلیل اهمیت موضوع تشخیص به موقع سرطان و وجود داده‌های نامتوازن در این حوزه، پژوهش حاضر الگوریتمی برای متوازن‌سازی مجموعه داده پیشنهاد می‌کند تا کارایی و صحت الگوریتم‌های تشخیصی مبتنی بر یادگیری ماشین در این حوزه افزایش یابد. این پژوهش از روش نمونه‌برداری فزاینده برای متوازن کردن نسبت دو کلاس سالم و بیمار استفاده می‌کند. وجه تمایز الگوریتم پیشنهادی در این مقاله، که به اختصار آن را OB-ReliefF، (نمونه‌برداری فزاینده مبتنی بر ReliefF[5])، می‌نامیم با سایر الگوریتم‌های نمونه‌برداری فزاینده استفاده شده برای متوازن‌سازی،

این مقاله در تاریخ ۲۶ اسفند ماه ۱۴۰۳ دریافت و در تاریخ ۲۸ خرداد ماه ۱۴۰۴ بازنگری شد.

زینب عباسی (نویسنده مسئول)، دانشکده مهندسی، مرکز آموزش عالی محلات، محلات، ایران، (email: z.abbasi@mahallat.ac.ir).

1. Computer-Aided Diagnosis
2. International Agency for Research on Cancer
3. Undersampling
4. Oversampling
5. Oversampling Based on ReliefF

اهمیت به میزان مشابهت نمونه‌های تولید شده با نمونه‌های مهم کلاس اقلیت است. نمونه‌های مهم، نمونه‌های واقع در مرز کلاس اقلیت با سایر کلاس‌ها هستند. چنین نمونه‌هایی به آموزش بهتر الگوریتم طبقه‌بند که برای تشخیص افراد بیمار از سالم استفاده می‌شود، کمک می‌کنند. علاوه بر این، همان‌گونه که در بخش آزمایش‌ها نشان داده خواهد شد، الگوریتم توانایی متوازن‌سازی کامل مجموعه داده را دارد و به ویژه، هر قدر مقدار عدم توازن مجموعه اولیه بیشتر باشد، الگوریتم OB-ReliefF عملکرد بهتری نسبت به سایر روش‌ها خواهد داشت. OB-ReliefF امکان کار با انواع مجموعه داده، شامل مجموعه داده‌های چند کلاسه، و مجموعه‌هایی با مقادیر عددی، رشته‌ای و مقادیر مفقود را دارد. هم‌چنین، چون بخش زیادی از محاسبات را می‌توان برای هر نمونه به طور مستقل انجام داد، امکان اجرای موازی و کاهش زمان محاسبات در هنگام کار با مجموعه داده‌های بزرگ وجود دارد. یکی دیگر از مزایای الگوریتم‌های نمونه‌برداری فزاینده، که OB-ReliefF نیز از آن بهره‌مند است، ممانعت از بروز مشکل بیش‌برازش[1] است. به زبان ساده، بیش‌برازش به این مفهوم است که الگوریتم طبقه‌بند فقط روی نمونه‌های که برای آموزش به آن داده شده، قدرت تشخیص خوبی دارد، اما قادر به طبقه‌بندی صحیح نمونه‌های جدید نیست. این مشکل به دلیل کمبود نمونه برای آموزش طبقه‌بند رخ می‌دهد که با استفاده از روش نمونه‌برداری فزآینده می‌توان مشکل را تا حدود زیادی حل کرد، به ویژه هنگامی که ارزش نمونه‌های تولید شده بالا باشد.

ایده الگوریتم پیشنهادی در این پژوهش الهام‌گرفته از الگوریتم ReliefF [۱۶] است. اگر چه ReliefF یک الگوریتم رتبه‌بندی و انتخاب ویژگی است، در اینجا از ایده کلی آن با تغییرات و اصلاحاتی برای ایجاد یک الگوریتم برای رتبه‌بندی نمونه‌ها و نمونه‌برداری استفاده می‌شود. ReliefF-OB برای هر نمونه کلاس اقلیت، وزنی بر اساس شباهت نمونه‌ها محاسبه می‌کند. پس از آن، برخی از بهترین نمونه‌های کلاس اقلیت که وزن بیشتری دارند، تکرار می‌شوند تا کلاس‌های مجموعه داده متوازن شده و تعداد نمونه‌های کلاس‌ها برابر شود.

کارایی الگوریتم OB-ReliefF با استفاده از مجموعه داده‌های سرطان سینه ویسکانسین[2] (WBCD)، مجموعه داده‌های تشخیصی سرطان سینه ویسکانسین[3] (WDBCD) و مجموعه سرطان سینه SEER اندازه‌گیری شد. طبق معیارهای ذکر شده در بخش آزمایش‌ها، نتایج نشان‌دهنده بهبود قابل توجهی در تشخیص بیماری هستند.

ادامه مقاله به شکل زیر سازمان‌دهی شده است: بخش ۲ کارهای مرتبط را بیان می‌کند. در بخش ۳، الگوریتم پیشنهادی شرح داده شده است. بخش ۴، مشخصات و نتایج آزمایش‌ها را بیان کرده و در مورد آن‌ها بحث می‌کند. در نهایت، در بخش ۵ جمع‌بندی و نتیجه‌گیری انجام می‌شود.

۲- کارهای مرتبط

ایجاد یک سیستم برای تشخیص سرطان سینه با استفاده از روش‌های مختلف یادگیری ماشین امکان‌پذیر است. بسیاری از مطالعات از ترکیب الگوریتم‌های مختلف طبقه‌بند استفاده می‌کنند. برخی مقالات نیز از ترکیب تکنیک‌های انتخاب ویژگی یا نمونه‌برداری با یک طبقه‌بند برای ساخت یک سیستم ACD بهره می‌گیرند. هدف تمامی این سیستم‌ها بهبود معیارهای ارزیابی و تشخیص بهتر بیماری است. با این حال، برخی از معیارهای ارزیابی استفاده شده در مقالات، مانند صحت، برای ارزیابی مجموعه داده‌های نامتوازن مناسب نیستند [۱۷] و مقالاتی که تنها از این معیارها استفاده می‌کنند برای ارجاع چندان مفید نخواهند بود. در این بخش برخی مقالات و پژوهش‌های انجام شده درباره تشخیص سرطان سینه و یا روش‌های نمونه‌برداری ارائه خواهد شد.

محققان در [۱۸] یک سیستم تشخیص سرطان سینه با استفاده از قوانین انجمنی[4] (AR) و شبکه‌های عصبی مصنوعی[5] (ANN) توسعه دادند که با استفاده از این سیستم تعدادی از ویژگی‌های موثر برای تشخیص سرطان انتخاب شدند. سپس این ویژگی‌های انتخاب شده برای طبقه‌بندی مجموعه داده WBCD با استفاده از پرسپترون چندلایه[6] (MLP)، نوعی از ANN، استفاده شدند.

در مقاله [۱۹]، ابتدا الگوریتم استخراج ویژگی PCA اجرا می‌شود. سپس، ویژگی‌ها رتبه‌بندی شده و ویژگی‌های مهم با استفاده از الگوریتم‌های نسبت سیگنال به نویز[7] (SNR) و انتخاب متوالی رو به جلو[8] جلو[8] (SFS) انتخاب می‌شوند. در نهایت، خروجی به‌دست‌آمده به طبقه‌بندهای SVM، KNN و شبکه عصبی احتمالی[9] (PNN) داده می‌شوند تا نمونه‌های خوش‌خیم و بدخیم تشخیص داده شوند. نتایج این روش و برخی روش‌های دیگر در بخش‌های بعدی مقاله ارائه می‌شود.

مؤلفان در مقاله [۲۰] مجموعه داده را با استفاده از تکنیک نمونه‌برداری فزاینده اقلیت ترکیبی[10] (SMOTE) متوازن می‌کنند. پس از آن، از سیستم شناسایی ایمنی مصنوعی[11] (AIRS) برای دسته‌بندی مجموعه داده نهایی استفاده می‌کنند.

در مقاله [۴] دو روش ترکیب نتایج طبقه‌بندهای ضعیف برای ایجاد طبقه‌بندهای قوی‌تر، یعنی Bagging و Boosting، استفاده شده است. در این مقاله بخشی از کلاس اکثریت که برابر با کلاس اقلیت است به طور مکرر انتخاب می‌شود. سپس دو مجموعه خروجی به دسته‌بندی‌کننده داده می‌شود تا خروجی نهایی به دست آید.

محققان در مقاله [۲۱] عملکرد طبقه‌بندهای مختلف را بر روی مجموعه داده WBCD بررسی کرده‌اند. طبقه‌بندهای بررسی‌شده شامل SVM، درخت تصمیم (C۴.۵)، بیز ساده (NB) و KNN هستند. نتایج نشان می‌دهند که طبقه‌بند SVM بهتر از دیگر طبقه‌بندها در این مجموعه داده عمل می‌کند.

در [۲۲] ترکیبی از چند شبکه عصبی مختلف برای تشخیص سرطان سینه استفاده شده است. آنها روش خود را بر روی WDBCD اعمال کرده‌اند. این روش با توجه به اینکه چند طبقه‌بند مختلف را اجرا می‌کند، زمانبر بوده و برای مجموعه داده‌های بزرگ مناسب نیست.

نویسندگان در [۲۳] ویژگی‌های جدیدی از مجموعه ویژگی‌های اصلی محاسبه می‌کنند. آن‌ها از الگوریتم K-means برای شناسایی تومورهای خوش‌خیم و بدخیم استفاده می‌کنند. شش ویژگی جدید از ۳۲ ویژگی پایه ایجاد می‌شود. سپس، طبقه‌بند SVM نمونه‌های سرطانی را شناسایی

1. Overfit
2. Wisconsin Breast Cancer Dataset
3. Wisconsin Diagnostic Breast Cancer Dataset
4. Association Rules
5. Artificial Neural Networks
6. Multi-Layer Perceptron
7. Signal to Noise Ratio
8. Sequential Forward Selection
9. Probabilistic Neural Network
10. Synthetic Minority Over-Sampling Technique
11. Artificial Immune Recognition System

می‌کند. آنها روش خود را بر روی مجموعه داده WDBCD آزمایش کرده و به صحت ۹۷/۳۸٪ دست یافتند. این مقاله مقادیر سایر معیارهای طبقه‌بندی را بیان نکرده است.

محققان در مقاله [۲۴] با ترکیب ویژگی‌های بافت از تصاویر اولتراسوند الاستوگرافی، یک سیستم CAD برای سرطان سینه توسعه دادند. آن‌ها از میانگین صحت طبقه‌بندی به عنوان تابع هدف در PSO استفاده می‌کنند. هم‌چنین، عملکرد سیستم تشخیص را با استفاده از چهار ویژگی مختلف بافت بررسی می‌کنند.

نویسندگان در [۲۵] از تکنیک‌های انتخاب ویژگی بهینه‌سازی مبتنی بر یاددهی-یادگیری[1] (TLBO)، بهینه‌سازی گله فیل‌ها[2] (EHO) ، و یک یک الگوریتم ترکیبی پیشنهادی از دو الگوریتم قبل، به منظوربهبود صحت طبقه‌بندی استفاده کردند. نویسندگان از چندین روش طبقه‌بندی برای ارزیابی اثربخشی رویکرد خود روی WBCD استفاده کردند.

در مقاله [۲۶] نویسندگان ۱۱ روش طبقه‌بندی مختلف را روی مجموعه داده WDBC اعمال کردند تا نتایج حاصل در معیارهای ارزیابی را با هم مقایسه کنند. نتایج نشان داد که درخت‌های تصادفی فوق‌العاده[3] (ERT) بالاترین صحت (۹۷/۳۶٪) را به دست آوردند. با این حال، این روش بهترین عملکرد را در تمامی شاخص‌های ارزیابی نداشت.

در مقاله [۲۷] تشخیص سرطان سینه با استفاده از الگوریتم‌های داده‌کاوی در سه مرحله (۱- الگوریتم SMOTE، ۲ - بهینه‌سازی بیزی، و ۳- ایجاد یک طبقه‌بند ترکیبی) انجام می‌شود.

مقاله [۲۸] یک مدل ترکیبی با استفاده از شبکه‌های عصبی مصنوعی و سیستم‌های منطق فازی و پیش پردازش با SMOTE ایجاد می‌کند. با این حال، مدل پیشنهادی نمی‌تواند روی مجموعه‌داده‌هایی حاوی مقادیر مفقود استفاده شود. این طبقه‌بندها روی چند مجموعه داده پزشکی آزمایش شده‌اند: WBCD، SPECTF، مجموعه داده بیماری قلبی، پارکینسون و مجموعه رتینوپاتی دیابتی Debrecen.

در [۲۹] محققان مجموعه از روش‌های نمونه‌برداری موجود را روی چند مجموعه داده، از جمله مجموعه داده SEER بکار بردند و سپس مجموعه‌های نسبتاً متوازن‌شده را با الگوریتم‌های طبقه‌بند مختلف آزمودند. از بین روش‌های نمونه برداری مختلف، روش SMOTE and Edited Nearest Neighbors (SMOTEENN) [۳۰] که یک روش ترکیبی از دو تکنیک نمونه‌برداری فزاینده و نمونه‌برداری کاهنده است، به بهترین نتایج در مجموعه داده‌ها دست یافته است. الگوریتم Instance Hardness Threshold (IHT) [۳۱] که یک الگوریتم نمونه‌برداری کاهنده است، در رتبه دوم از نظر کسب نتایج و کارایی، قرار دارد.

پژوهشگران در [۳۲]، با استفاده از مجموعه داده سرطان پستان SEER، کارایی مدل‌های پیش‌بینی‌کننده شامل CatBoost، Extra Tree Classifier، LightGBM و XGBoost را بررسی کردند. پیش از ایجاد مدل توسط طبقه‌بندهای ذکر شده، عملیات پیش پردازش روی داده‌ها با الگوریتم SMOTE و الگوریتم انتخاب ویژگی PCA، اجرا شد. بررسی نهایی معیارهای کارایی طبقه‌بند نشان داد که XGBoost و CatBoost در تشخیص سرطان سینه دقیق‌تر عمل می‌کنند.

در مقاله [۳۳]، از روش شبکه‌های متخاصم مولد[4] (GAN)، برای ایجاد ایجاد نمونه‌های جدید و رفع مشکل عدم توازن در مجموعه داده سرطان سینه SEER استفاده شد. GAN از معماری‌های شبکه عصبی عمیق برای تولید داده‌های مصنوعی با کیفیت بالا استفاده می‌کند. چهار طبقه‌بند Boosting، Bagging، Linear و Non-linear برای ارزیابی عملکرد مجموعه متوازن شده، استفاده شدند.

بیشتر مقالات ذکر شده در بالا بر تکنیک‌های سطح الگوریتم و مجموعه‌ای از طبقه‌بندها متمرکز هستند. در این بین، تنها مقالات [۱۸]، [۱۹]، [۲۳] تا [۲۵] از یک روش استخراج یا انتخاب ویژگی استفاده می‌کنند و مقالات [۲۰]، [۲۷] تا [۲۹]، [۳۲]، و [۳۳] الگوریتم SMOTE و یا سایر روش‌های موجود برای متوازن‌سازی نمونه‌ها را به کار می‌برند. در واقع، بسیاری از این مقالات، الگوریتم جدیدی برای تشخیص بیماری-ها و یا متوازن‌سازی داده‌ها ارائه نمی‌دهند. آن‌ها تنها از ترکیبی از روش‌های موجود برای تشخیص سرطان سینه استفاده کرده‌اند، اما الگوریتم پیشنهادی در مقاله حاضر، یک روش جدید افزایش نمونه به منظور متوازن‌سازی تعداد نمونه‌ها در مجموعه داده‌های نامتوازن است. از این الگوریتم می‌توان برای‌متوازن سازی مجموعه داده‌های متنوعی به خصوص در زمینه تشخیص بیماری استفاده نمود.

در ادامه تعدادی از پژوهش‌هایی که در زمینه متوازن‌سازی با روش‌های سطح داده انجام شده‌اند، بررسی می‌گردد. البته این روش‌ها بر روی مجموعه داده‌های پزشکی آزمایش نشده است.

محققان در [۳۴] یک روش ترکیبی در سطح داده، برای متوازن‌سازی مجموعه داده ارائه دادند. این روش نمونه‌های مرزی کلاس اکثریت را وزن دهی کرده و سپس بخشی از آن‌ها را با روش نمونه‌برداری وزن‌دار تصادفی حذف می‌کند. هم‌چنین، روش پیشنهادی برای نمونه‌گیری فزاینده در کلاس اقلیت، نمونه‌های از این کلاس را با درون‌یابی بین یک نمونه هم‌کلاسی و نزدیک‌ترین نمونه‌های کلاس اکثریت تولید می‌کند.

در مقاله [۳۵] یک چارچوب برای حل مشکل عدم توازن کلاس‌ها برای افزایش یادگیری از جریان‌های داده، ارائه شده است. این چارچوب شامل پردازش بخش‌های داده‌ای است که از نمونه‌های اولیه (نمونه‌های ورودی که نادرست طبقه‌بندی شده‌اند) ایجاد می‌گردد. طبقه‌بند به طور پیوسته بخش‌های داده را به روز کرده و نمونه‌های مفید را ذخیره کرده و نمونه‌های غیرمفید را حذف می‌کند.

در مقاله [۳۶]، مطالعه‌ای روی نحوه ترکیب روش‌های انتخاب ویژگی با الگوریتم نمونه‌برداری SMOTE انجام شد و مشخص شد که اجرای انتخاب ویژگی و سپس SMOTE، عملکرد بهتری ایجاد می‌کند. هم‌چنین بین روش‌های انتخاب ویژگی مقایسه شده، الگوریتم XGBoost بهترین عملکرد را در ترکیب با SMOTE ارائه می‌دهد.

همان‌گونه که ذکر شد، چند مقاله بالا روی مسئله تشخیص پزشکی کار نکرده‌اند و صرفا برای بررسی راهکارهای پیشنهادی روی مجموعه داده نامتوازن بررسی شدند، از این بین مقالات [۳۴] و [۳۵]، روش‌های ترکیبی برای اجرای همزمان نمونه‌برداری فزاینده و کاهنده ارائه داده‌اند، که چون در مجموعه داده‌های پزشکی تشخیص صحیح کلاس اقلیت موضوعیت دارد، کاهش نمونه‌های کلاس اکثریت کمک زیادی به روند تشخیص نمی‌کند. مقاله [۳۶] نیز از روش SMOTE استفاده کرده و روش جدیدی ارائه نداده است.

1. Teaching-Learning-Based Optimization
2. Elephant Herding Optimization
3. Extra Randomized Trees
4. Generative Adversarial Networks

۳- الگوریتم پیشنهادی (OB-ReliefF)

همان‌طور که قبلاً ذکر شد، OB-ReliefF از الگوریتم ReliefF الهام گرفته است. الگوریتم ReliefF یک الگوریتم انتخاب ویژگی است که به‌عنوان مرحله پیش‌پردازش در بسیاری از مسائل یادگیری ماشین استفاده می‌شود. الگوریتم ReliefF ویژگی‌ها را با مقایسه آن‌ها رتبه‌بندی می‌کند [۳۷]. این الگوریتم قادر است روی مسائل چندکلاسه، ویژگی‌های عددی و غیرعددی و داده‌های ناقص کار کند.

الگوریتم ReliefF به این صورت عمل می‌کند: ابتدا، کاربر تعداد تکرارهای حلقه اصلی را تعیین می‌کند. در هر دور از این حلقه، الگوریتم به‌طور تصادفی یک نمونه را انتخاب می‌کند. سپس، برای نمونه انتخاب‌شده، الگوریتم مجموعه Hit (B نزدیک‌ترین نمونه از همان کلاس) و مجموعه Miss (B نزدیک‌ترین نمونه از کلاس‌های دیگر) را پیدا می‌کند. سپس، یک تابع ارزیابی، شباهت بین نمونه انتخاب‌شده و نمونه‌های موجود در مجموعه‌های Hit و Miss محاسبه می‌کند. الگوریتم اصلی ReliefF از فاصله منهتن به‌عنوان تابع ارزیابی استفاده می‌کند. حالا هر ویژگی یک وزن دارد که بر اساس مقدار شباهت محاسبه می‌شود. در نهایت، ReliefF ویژگی‌ها را بر اساس وزن‌هایشان رتبه‌بندی کرده و ویژگی‌های با رتبه بالاتر را انتخاب می‌کند.

الگوریتم OB-ReliefF از ایده کلی الگوریتم ReliefF استفاده می‌کند. پیش از این، در مقاله [۳۸]، نویسنده مقاله از الگوریتم ReliefF برای تعیین و انتخاب بهترین نمونه‌ها استفاده نموده است، اما در این مقاله، الگوریتمی مشابه با روش مقاله قبلی، اما با رویکرد متوازن‌سازی نمونه‌ها و کاربرد الگوریتم پیشنهادی در تشخیص سرطان سینه استفاده می‌شود. روش کار به این شکل است که ابتدا بهترین نمونه‌های کلاس اقلیت (معمولاً کلاس افراد بیمار) که اطلاعات مفیدی برای آموزش سیستم CAD فراهم می‌کنند، شناسایی شده و سپس این نمونه‌های مهمتر با کپی کردن، تکثیر می‌شوند. ملاک اهمیت یک نمونه، میزان شباهت با نمونه‌های همکلاس یا از کلاس مخالف است که توسط تابع ارزیابی محاسبه می‌گردد. به طور خلاصه، الگوریتم OB-ReliefF به دنبال یافتن نمونه‌های مرزی از یک کلاس است، زیرا این نمونه‌ها می‌توانند به طبقه‌بند در یادگیری بهتر تفاوت‌های دو کلاس، کمک قابل توجهی کنند. هدف از ارائه این الگوریتم، ایجاد دو کلاس متوازن از تصاویر مربوط به بیماران و افراد سالم است.

با توجه به شکل ۱، OB-ReliefF به این صورت عمل می‌کند: به ازای هر کدام از نمونه‌های کلاس اقلیت، حلقه اصلی در خط ۴ مراحل زیر را انجام می‌دهد: خط 4-a یک مجموعه از نزدیک‌ترین نمونه‌های هم‌کلاس با نمونه را پیدا می‌کند. معیار مشابهت و نزدیکی دو نمونه بر اساس شاخص ژاکارد سنجیده می‌شود که در ادامه توضیح داده می‌شود. در خطوط 4-b.1 تا 4-b.3، بر اساس شاخص ژاکارد، مجموعه‌ای از نمونه‌های نزدیک به نمونه انتخابی از کلاس مخالف (یعنی کلاس اکثریت) ساخته می‌شود. شایان ذکر است که در مجموعه داده تشخیص بیماری تنها دو کلاس بیمار (اقلیت) و کلاس افراد سالم (اکثریت) وجود دارد. اما الگوریتم OB-ReliefF، توانایی کار با مجموعه داده‌هایی با چند کلاس را دارد. به همین دلیل، این خطوط در قالب یک حلقه تکرار نوشته شده است، تا مجموعه Miss set را به ازای تمام کلاس‌های مخالف بیابد. درخط 4-c، برای هر نمونه، وزنی با استفاده از فرمول داده شده محاسبه می‌شود. در این فرمول، تابع ارزیابی (δ) نشان‌دهنده میزان شباهت بین دو نمونه است. فرمول کلی وزن به این صورت است که اگر یک نمونه با دیگر نمونه‌های هم کلاس خود بسیار متفاوت باشد، وزن

1. /*X= set of training examples */
2. /*B= number of nearest neighbors to compute*/
3. /*L= number of instances in minority class*/
4. For t:=1 to L do

   a. For each instance x_t with minority class label y_t (x_t,y_t) Find B nearest instances to x_t (Hit set) by Jaccard index

   b. 1. For majority class $(c \neq y_t)$

   b. 2. Find B nearest instances to x_t from majority class (Miss set) by Jaccard index

   b.3. End for (line ۴.b.۱)

   c. Calculate the weight of each instance x_t:

$$w_t = \sum_{c \neq y_t} \frac{1}{LB} \sum_{(xj,yj)\in Miss} \delta(x_t, x_j) - \frac{1}{LB} \sum_{(xi,yi)\in Hit} \delta(x_t, x_i)$$

5. End for (line ۴)
6. Sort the weight array w_t descending.
7. For the minority class c, perform over-sampling by replicating 50% of the more weighted instances in order to the number of instances in both classes is equal.
8. End

شکل ۱: شبه کد الگوریتم OB-ReliefF.

بیشتری دارد. در خط ۶، نمونه‌ها بر اساس وزن محاسبه‌شده به ترتیب نزولی مرتب می‌شوند. در نهایت، نمونه‌های جدید با استفاده از نمونه‌های قبلی در خط ۷ ایجاد می‌شوند. در این جا، ۵۰٪ از نمونه‌هایی که وزن بیشتری دارند، تکرار می‌شوند تا تعداد نمونه‌های دو کلاس با هم برابر شوند. این عملیات به نام نمونه‌برداری فزاینده شناخته می‌شود. استفاده از روش تکثیر نمونه به این روش به سادگی اجرا می‌شود و سربار محاسباتی زیادی مانند تولید نمونه به روش الگوریتم SMOTE ندارد. انتخاب نرخ ۵۰٪ به طور تجربی و رعایت یک حد اعتدال در تکثیر نمونه‌ها است و می‌توان از هر نرخ دیگری برای تولید نمونه های جدید استفاده کرد. البته چون نرخ عدم توازن نمونه‌ها پایین است و نمونه‌های با وزن بالا، نماینده‌ی خوبی از توزیع کلاس اقلیت هستند، تکثیر با نرخ ۵۰٪ می‌تواند بدون افزودن نویز و داده‌های کم اهمیت، توازن ایجاد کند. در مورد داده های با نرخ عدم توازن بیشتر و یا داده‌های با تعداد نمونه بیشتر، می‌توان از سیاست تکثیر با کمی تفاوت استفاده کرد. برای مثال، ابتدا نمونه‌های با وزن بالا بیشتر تکثیر شوند (مثلاً ۶۰ درصد)، سپس به تدریج از درصد تکثیر (یعنی ۴۰ درصد و در دور بعد ۲۰ درصد) کم شود. یا اینکه نرخ‌های تکثیر متفاوتی را در آزمایش‌های مختلف انتخاب کرده و کارایی آن‌ها با معیارهای متفاوت بررسی گردد. البته برای بررسی میزان مفید بودن تکثیر در نرخ ۵۰ درصد نیز همانگونه که در بخش نتایج گفته خواهد شود، معیارهای کارایی غیر از صحت نیز باید استفاده شوند.

به جای فاصله منهتن در الگوریتم ReliefF اصلی، این مقاله از شاخص ژاکارد [۳۹] به عنوان تابع ارزیابی δ استفاده می‌کند. در واقع، برای اندازه‌گیری شباهت بین دو نمونه، از شاخص ژاکارد استفاده می‌شود که امکان کار روی ویژگی‌هایی با مقادیر عددی و رشته‌ای و هم‌چنین با مقادیر مفقود را دارد. اگر $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ و $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ دو نمونه با اعداد حقیقی بزرگتر از صفر باشند، شاخص ژاکارد آنها به شرح زیر محاسبه می‌شود:

$$J(\mathbf{x},\mathbf{y}) = \sum_i \min(x_i, y_i) \Big/ \sum_i \max(x_i, y_i)\,. \qquad (۱)$$

برای اجرای (۱) بر روی ویژگی‌هایی با مقادیر منفی، باید مقادیر در بازه [۰,۱] نرمال‌سازی شوند. برای دو نمونه کاملا برابر، نتیجه شاخص ژاکارد

یک است و برای دو نمونه کاملاً متفاوت، نتیجه شاخص ژاکارد صفر است. اگر $\mathbf{x}$ و $\mathbf{y}$ دو مجموعه غیر عددی باشند، شاخص ژاکارد به صورت زیر محاسبه می‌شود:

$$J(\mathbf{x},\mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}. \quad (۲)$$

اگر مقدار x_i یا y_i ناشناخته باشد، اشتراک آن‌ها تهی خواهد بود $(x_i \cap y_i = \varnothing)$. برای محاسبه تابع ارزیابی، بسته به نوع ویژگی‌ها (عددی یا غیرعددی)، از (۱) یا (۲) استفاده می‌شود. علاوه بر این، اگر ویژگی‌ای دارای مقادیر مفقود باشد، الگوریتم از (۲) استفاده می‌کند. شاخص ژاکارد هر نمونه، میانگین مقادیر شباهت بین تمام ویژگی‌های آن نمونه است. برای هر نمونه، نزدیک‌ترین همسایگانی که شاخص‌های ژاکارد بالاتری دارند، در مجموعه Hit و Miss قرار خواهند گرفت. طبق خط 4-c در شکل۱، نمونه‌هایی که شبیه به نمونه‌های دیگر از کلاس خود هستند، وزن کمتری دریافت می‌کنند. برعکس، اگر آن‌ها شبیه به نمونه‌هایی از کلاس دیگری باشند(یعنی نمونه‌های مرزی)، وزن بیشتری کسب خواهند کرد. این داده ها می‌توانند قدرت یادگیری طبقه‌بندها را افزایش داده و کارایی آن‌ها را افزایش دهند.

پیچیدگی زمانی این الگوریتم همانطور که از شبه کد مشخص است، وابسته به تعداد نمونه های کلاس اقلیت (L) و تعداد همسایه هایی که باید برای هر نمونه محاسبه شود (مقدار B)، است. البته اگر مسئله چند کلاسه باشد، حلقه تکرار خطوط 4-b.1 تا 4-b.3 به ازای هر کلاس تکرار خواهد شد. اما در مجموعه داده‌های بیماری (مانند مجموعه WBCD) عموما تنها دو کلاس بیمار و سالم وجود دارد. بنابراین پیچیدگی زمانی این الگوریتم برابر با O(LB) خواهد بود. که مقدار L (تعداد نمونه‌های کلاس اقلیت) به طور معمول کمتر از حتی یک سوم مجموعه داده اصلی است و مقدار B نیز یک پارامتر تعریف شده کاربر است که مقدار پایینی دارد. برای مثال، در [۴۰]، Bمساوی ۱۰ است و در [۴۱]، یک بار B را برابر ۱ و بار دیگر برابر ۱۰ در نظر گرفته‌اند. در نرم افزار Weka نیز مقدار پیش فرض برای این پارامتر ۱۰ در نظر گرفته شده است، به این دلیل که مطابق با نظر پژوهشگران در [۳۷] انتخاب مقدار B=۱۰ برای بیشتر کاربردها، مناسب خواهد بود. شایان ذکر است که چون محاسبات به ازای هر نمونه کلاس اقلیت می‌تواند به طور مستقل انجام شود، می‌توان حلقه اصلی برنامه (خط ۴) را به صورت موازی اجرا کرد که در مجموعه داده‌های بزرگ می‌تواند زمان محاسبات را تا حد زیادی کاهش دهد. البته این نکته را باید در نظر داشت که در این مقاله هدف افزایش صحت تشخیص بیماری سرطان سینه است و زمان اجرای الگوریتم از اهمیت کمتری برخوردار است. بنابراین، سربار محاسباتی روش پیشنهادی در مقایسه با هدف افزایش کیفیت تشخیص، توجیه‌پذیر است.

٤- نتایج آزمایش‌ها

٤-١ مشخصات آزمایش

برای پیاده‌سازی الگوریتم OB-ReliefF، از نرم‌افزارهای MATLAB و Weka استفاده شده است. الگوریتم پیشنهادی روی سیستمی با پردازنده Intel core۲ Duo ۲/۲۶ GHz و ۳ گیگابایت رم، و سیستم‌عامل ۳۲ بیتی ویندوز ۷ اجرا شده است.

برای مقایسه و ارزیابی نتایج الگوریتم OB-ReliefF با دیگر روش‌ها، داده‌های به‌دست آمده با این الگوریتم به طبقه‌بندهای مختلف داده شده است. در این مقاله، از ۱۰ طبقه‌بند Weka برای ارزیابی استفاده شده است: MLP [۴۲]، جنگل تصادفی[1] (RF) [۴۳]، آدا بوست[2] [۴۴]، Bagging [۴۵]، KNN[3] (با K = ۳) [۴۶]، درخت تصادفی[4] [۴۷]، OneR، بیز ساده[5] [۴۸]، Kstar [۴۹] و C۴.۵ [۵۰]. برای آشنایی با هر یک از این روش‌های طبقه‌بندی به مقالات [۵۱] تا [۵۳] رجوع گردد. نتایج به‌دست آمده با این طبقه‌بندها از نظر معیارهای ارزیابی عملکردهای متفاوتی دارند. بنابراین، نتایج در تمامی این طبقه‌بندها مقایسه شدند، تا اثربخشی روش پیشنهادی تضمین شود.

جدول ۱: ماتریس اغتشاش.

		کلاس پیش بینی‌شده	
		منفی	مثبت
کلاس	منفی	*TN*	*FP*
واقعی	مثبت	*FN*	*TP*

علاوه بر این، از استراتژی ارزیابی متقابل ۵-بخشی[6] استفاده شده است تا از مقایسه بی‌طرفانه نتایج طبقه‌بندی اطمینان حاصل شود. برای ارزیابی دقیق‌تر عملکرد در مجموعه‌داده‌های نامتوازن، از معیارهای تعریف‌شده توسط ماتریس اغتشاش[7] استفاده می‌شود. ماتریس اغتشاش نتایج طبقه‌بندی را بر اساس اطلاعات واقعی و اطلاعات پیش‌بینی‌شده نشان می‌دهد. جدول ۱ ماتریس اغتشاش را برای مجموعه داده‌ای با دو کلاس نشان می‌دهد [۱۷].

TP: سیستم به درستی یک مورد مثبت را شناسایی می‌کند.

TN: سیستم به درستی یک مورد منفی را شناسایی می‌کند.

FN: یعنی یک مورد مثبت به اشتباه منفی شناسایی می‌شود.

FP: یعنی یک مورد منفی به اشتباه مثبت شناسایی می‌شود.

بر اساس این پارامترها، از پنج معیار ارزیابی رایج برای بررسی کارایی نتایج طبقه‌بندی استفاده شده است [۵۴]:

$$Accuracy\,(\%) = \frac{TP+TN}{TP+TN+FP+FN} \quad (۳)$$

$$Recall = Sensitivity\,(\%) = \frac{TP}{TP+FN} \quad (۴)$$

$$Specificity\,(\%) = \frac{TN}{TN+FP} \quad (۵)$$

$$G-Mean = \sqrt{Sensitivity \times Specificity} \quad (۶)$$

محدوده ROC: مساحت زیر منحنی ROC (AUC) روشی برای ارزیابی عملکرد متوسط مدل ارائه می‌دهد. زمانی که مدل کامل باشد، AUC آن برابر با یک خواهد بود.

به طور معمول، برای مقایسه عملکرد طبقه‌بندها از معیار صحت[8] استفاده می‌شود. با این حال، این معیار برای مجموعه داده‌های نامتوازن

1. Random Forest
2. AdaBoost
3. K-Nearest Neighbors
4. Random Tree
5. Naive Bayes
6. 5-fold Cross-Validation
7. Confusion Matrix
8. Accuracy

جدول ۲: مشخصات مجموعه داده‌ها.

	تعداد ویژگی	تعداد نمونه	خوش‌خیم (سالم یا زنده)	بدخیم (بیمار یا مرده)	مقادیر مفقود در داده	نرخ عدم توازن
WBCD	۱۱	۶۹۹	۴۵۸	۲۴۱	بله	۱٫۹
WDBCD	۳۲	۵۶۹	۳۵۷	۲۱۲	خیر	۱٫۶۸
SEER	۱۵	۴۰۲۴	۳۴۰۸	۶۱۶	خیر	۴٫۹۵

مناسب نیست [۱۷]. به عنوان مثال، فرض کنید ۹۵٪ از نمونه‌های یک مجموعه داده متعلق به کلاس اکثریت است و تنها ۵٪ از مجموعه داده متعلق به کلاس اقلیت است. در این صورت، حتی اگر طبقه‌بند همیشه نمونه‌ها را به عنوان کلاس اکثریت دسته‌بندی کند، صحت آن ۹۵٪ خواهد بود. البته از آنجا که صحت رایج‌ترین روش برای ارزیابی عملکرد یک طبقه‌بند است [۵]، در این مقاله از آن استفاده شده است.

حساسیت[1] و خصوصیت[2] عملکرد طبقه‌بند را در کلاس‌های مختلف نشان می‌دهند و بنابراین به تنهایی برای مجموعه داده نامتوازن مناسب نیستند. اما G-Mean چون ترکیبی از این دو معیار است، عملکرد متوازن طبقه‌بند را بین دو کلاس نشان می‌دهد [۲۴]. محدوده ROC معادلاتی بین درست مثبت‌ها و اشتباه مثبت‌ها را نشان می‌دهد و خلاصه خوبی از عملکرد یک طبقه‌بند است [۵۵].

در این مقاله، از داده‌های WBCD، WDBCD و SEER برای آزمایش قابلیت تشخیص سرطان روش پیشنهادی استفاده می‌شود. دکتر ویلیام اچ. وولبرگ داده‌های WBCD را از بیمارستان‌های دانشگاه ویسکانسین- مدیسون از سال ۱۹۸۹ تا ۱۹۹۱ جمع‌آوری کرده است [۵۶].

هم‌چنین، او و دو همکارش مجموعه داده WDBCD را در سال ۱۹۹۵ جمع‌آوری کردند [۵۷]. ویژگی‌ها از یک تصویر دیجیتالی از یک نمونه‌برداری با سوزن ظریف (FNA) از یک توده سینه محاسبه شده‌اند. این ویژگی‌ها خصوصیات هسته‌های سلولی موجود در تصاویر را توصیف می‌کنند. مجموعه داده SEER از بیماران سرطان سینه از نسخه به‌روزرسانی شده در نوامبر ۲۰۱۷ برنامه SEER موسسه ملی سرطان (NCI) به دست آمده است. این مجموعه داده‌ها شامل برخی از اطلاعات شخصی بیماران زن (مانند سن، نژاد، قومیت، مرحله سرطان و ...) مبتلا به سرطان سینه تهاجمی است که بین سال‌های ۲۰۰۰ تا ۲۰۱۷ تشخیص داده شده‌اند [۵۸]. این مجموعه داده‌ها از مخزن یادگیری ماشین UCI [۵۹] و Kaggle [۶۰] دانلود شده‌اند. مشخصات این داده‌ها در جدول ۲ نشان داده شده است.

برای مقایسه بهتر، نتایج حاصل از طبقه‌بندی داده‌های اصلی، الگوریتم SMOTE، الگوریتم نمونه‌برداری تصادفی و OB-ReliefF برای هر سه مجموعه داده ارائه می‌گردد. با این توضیح که به جز داده‌های اصلی، که مشخصات آن‌ها در جدول ۲ ارائه شد، در سایر روش‌ها، مانند تمام روش‌های پیش پردازش داده، ابتدا الگوریتم‌های نمونه‌برداری روی داده انجام شده و سپس از این داده‌ها (شامل نمونه‌های اصلی و نمونه‌های افزوده شده) هم برای آموزش و هم ارزیابی طبقه‌بندها استفاده می‌شود.

SMOTE یک الگوریتم نمونه‌برداری فزاینده معروف است. نمونه‌برداری تصادفی یک روش کاهش نمونه (نمونه‌برداری کاهنده) است که برای مقایسه کارایی روش پیشنهادی با روش‌های کاهش نمونه استفاده می‌شود. در این روش، باید یک نرخ کاهش انتخاب شود که مقدار آن ۵۰٪ در نظر گرفته شده است، یعنی نیمی از داده‌ها به‌طور تصادفی

1. Sensitivity
2. Specificity

شکل ۳: منحنی ROC برای طبقه‌بند RF برای طبقه بندی مجموعه داده‌ی WBCD متوازن‌شده با الگوریتم OB-ReliefF.

انتخاب می‌شود. مقدار ۵۰٪ به صورت تجربی و به جهت یکسان بودن با نرخ افزایش الگوریتم OB-ReliefF انتخاب شده است. زیرا در الگوریتم نمونه‌برداری تصادفی ۵۰٪ درصد از داده‌ها کاسته و در الگوریتم OB-ReliefF، ۵۰٪ به مقدار داده‌های اقلیت افزوده می‌شود. البته، همانگونه که گفته شد، در OB-ReliefF نمونه‌های افزوده شده نمونه‌های مهمی هستند که وجه تمایز بین کلاس‌ها را دقیق‌تر نشان می‌دهند. مطالعات بیشتری را می‌توان در تحقیقات آینده در مورد انتخاب نرخ کاهش یا افزایش انجام داد. مقادیر پارامترهای استفاده شده در الگوریتم SMOTE عبارتند از:

- بذر[3] مورد استفاده برای نمونه‌برداری تصادفی = ۱
- درصد نمونه‌های SMOTE که باید ایجاد شوند = ۱۰۰
- تعداد همسایگان نزدیک = ۵.

مقدار پارامتر B برای الگوریتم OB-ReliefF نیز طبق توصیه پژوهش [۱۷] برابر ۱۰ انتخاب شد.

۴-۲ نتایج پیاده‌سازی

این بخش نتایج اجرای الگوریتم OB-ReliefF و دیگر الگوریتم‌ها را روی سه مجموعه داده سرطان سینه نشان می‌دهد. جدول ۳ نتایج مربوط به WBCD را نمایش می‌دهد.

شکل ۳ منحنی ROC را برای انجام طبقه‌بند RF روی مجموعه داده WBCD متوازن‌شده با OB-ReliefF نشان می‌دهد. شایان ذکر است که این طبقه‌بند بالاترین مقادیر معیارهای کارایی را در طبقه‌بندی نمونه‌ها کسب کرده است. از ارائه سایر منحنی های ROC برای رعایت اختصار خودداری می‌گردد.

با بررسی مقادیر جدول ۳، می‌توان به نتایج زیر دست یافت:

- در اکثر موارد، الگوریتم OB-ReliefF بهترین مقادیر را در معیارهای مختلف کسب کرده است.
- الگوریتم OB-ReliefF بهترین مقدار را برای محدوده ROC در طبقه‌بند Bagging به‌دست آورده است.
- الگوریتمOB-ReliefF بالاترین مقادیر برای تمام معیارها را در

3. Seed

جدول ۳: معیارهای ارزیابی طبقه‌بندی با روش‌های مختلف برای مجموعه داده WBCD.

طبقه‌بند	الگوریتم انتخاب نمونه	صحت	حساسیت	خصوصیت	G-Mean	محدوده ROC
C۴.۵	مجموعه داده اولیه	۹۴٫۲۷	۹۴٫۳۰	۹۳٫۱	۹۳٫۷۰	۰٫۹۴۶
	SMOTE	۹۶٫۳۸	۹۶٫۴	۹۶٫۳۰	۹۶٫۳۵	۰٫۹۷۸
	Random sampling ٪۵۰	۹۵٫۹۹	۹۶٫۰۰	۹۶٫۲۰	۹۶٫۰۹	۰٫۹۷۲
	OB-ReliefF	۹۶٫۵۰	۹۶٫۵۰	۹۶٫۵۰	۹۶٫۵۰	۰٫۹۷۶
Ada Boost	مجموعه داده اولیه	۹۶٫۲۸	۹۶٫۳	۹۵٫۹۰	۹۶٫۰۹	۰٫۹۸۵
	SMOTE	۹۷٫۲۳	۹۷٫۲	۹۷٫۲۴	۹۷٫۲۲	۰٫۹۸۹
	Random sampling ٪۵۰	۹۷٫۴۲	۹۷٫۴	۹۶٫۹	۹۷٫۱۵	۰٫۹۹
	OB-ReliefF	۹۷٫۲۷	۹۷٫۳	۹۷٫۳	۹۷٫۳	۰٫۹۹۲
KNN K=۳	مجموعه داده اولیه	۹۶٫۷۱	۹۶٫۷	۹۶٫۳۰	۹۶٫۵۰	۰٫۹۸
	SMOTE	۹۷٫۸۷	۹۷٫۹	۹۷٫۸	۹۷٫۸۵	۰٫۹۸۵
	Random sampling ٪۵۰	۹۳٫۷	۹۳٫۷	۹۱٫۵	۹۲٫۵۹	۰٫۹۸۴
	OB-ReliefF	۹۷٫۲۷	۹۷٫۳۰	۹۷٫۲۰	۹۷٫۲۵	۰٫۹۸۳
Bagging	مجموعه داده اولیه	۹۵٫۴۲	۹۵٫۴۰	۹۴٫۸۰	۹۵٫۱۰	۰٫۹۸۴
	SMOTE	۹۶٫۵۹	۹۶٫۶	۹۶٫۶	۹۶٫۶	۰٫۹۸۹
	Random sampling ٪۵۰	۹۶٫۸۵	۹۶٫۸۰	۹۵٫۸۰	۹۶٫۲۹	۰٫۹۸۳
	OB-ReliefF	۹۶٫۵۱	۹۶٫۴۵	۹۶٫۵۵	۹۶٫۴۹	۰٫۹۹۴
MLP	مجموعه داده اولیه	۹۵٫۲۷	۹۵٫۳۰	۹۵٫۰۰	۹۵٫۱۵	۰٫۹۸۸
	SMOTE	۹۶٫۰۶	۹۶٫۱	۹۶٫۰۵	۹۶٫۰۷	۰٫۹۸۹
	Random sampling ٪۵۰	۹۵٫۷۰	۹۵٫۷۰	۹۵٫۲۰	۹۵٫۴۵	۰٫۹۸۱
	OB-ReliefF	۹۶٫۵۱	۹۶٫۵۰	۹۶٫۵۰	۹۶٫۵۰	۰٫۹۸۵
Random Tree	مجموعه داده اولیه	۹۳٫۴۲	۹۳٫۴۰	۹۰٫۴	۹۱٫۸۹	۰٫۹۲۳
	SMOTE	۹۶٫۱۷	۹۶٫۲	۹۶٫۱	۹۶٫۱۵	۰٫۹۶۳
	Random sampling ٪۵۰	۹۵٫۹۹	۹۶٫۰۰	۹۵٫۳۰	۹۵٫۶۴	۰٫۹۵۴
	OB-ReliefF	۹۶٫۲۹	۹۶٫۳۰	۹۶٫۳۰	۹۶٫۳۰	۰٫۹۶۵
OneR	مجموعه داده اولیه	۹۱٫۵۶	۹۱٫۶۰	۸۸٫۹	۹۰٫۲۴	۰٫۹۰۲
	SMOTE	۹۳٫۵۱	۹۳٫۵۰	۹۳٫۴۰	۹۳٫۴۵	۰٫۹۳۵
	Random sampling ٪۵۰	۹۳٫۹۸	۹۴٫۰۰	۹۱٫۲۰	۹۲٫۵۹	۰٫۹۲۶
	OB-ReliefF	۹۳٫۵۵	۹۳٫۶۳	۹۳٫۶۵	۹۳٫۶۵	۰٫۹۳۶
Naive Bayes	مجموعه داده اولیه	۹۶٫۱۴	۹۶٫۱	۹۷٫۰۰	۹۶٫۵۵	۰٫۹۸۶
	SMOTE	۹۷٫۰۲	۹۷٫۰۰	۹۶٫۹۰	۹۶٫۹۵	۰٫۹۸۵
	Random sampling ٪۵۰	۹۵٫۱۳	۹۵٫۱	۹۵٫۸۰	۹۵٫۴۵	۰٫۹۸۴
	OB-ReliefF	۹۷٫۰۵	۹۷٫۱۳	۹۷٫۱۸	۹۷٫۱۶	۰٫۹۸۷
Kstar	مجموعه داده اولیه	۸۰٫۲۶	۸۰٫۳۰	۶۷٫۴	۷۳٫۵۷	۰٫۸۲۳
	SMOTE	۸۸٫۹۴	۸۸٫۹۱	۸۹٫۰۰	۸۸٫۹۵	۰٫۹۲۸
	Random sampling ٪۵۰	۸۰٫۲۳	۸۰٫۲۰	۶۹٫۶	۷۴٫۷۱	۰٫۷۹۸
	OB-ReliefF	۹۲٫۰۳	۹۲٫۰۰	۹۲٫۰۰	۹۲٫۰۰	۰٫۹۴۸
RF	مجموعه داده اولیه	۹۶٫۷۱	۹۶٫۷	۹۶٫۳	۹۶٫۵۰	۰٫۹۸۶
	SMOTE	۹۷٫۰۲	۹۷٫۰۰	۹۷٫۰۵	۹۷٫۰۰	۰٫۹۹
	Random sampling ٪۵۰	۹۶٫۵۶	۹۶٫۶۰	۹۵٫۲	۹۵٫۸۹	۰٫۹۸۸
	OB-ReliefF	۹۸٫۰۳	۹۸٫۰۰	۹۸٫۰۲	۹۸٫۰۱	۰٫۹۹۳

طبقه‌بند RF کسب کرده است، به استثنای محدوده ROC، که مقدار محدوده ROC در طبقه‌بند RF نسبت به مقدار به‌دست آمده در طبقه‌بند Bagging اختلاف کمی دارد.

جدول ۴ معیارهای عملکرد برای WDBCD را ارائه می‌دهد. با بررسی مقادیر این جدول، می‌توان به نتایج زیر دست یافت:

- الگوریتم OB-ReliefF بهترین نتایج را در تمام طبقه‌بندها (به استثنای KNN، OneR و Naive Bayes) برای تمام معیارها کسب کرده است.
- الگوریتم OB-ReliefF بهترین مقادیر را در تمام معیارها با استفاده از RF به‌دست آورده است.

جدول ۴: معیارهای ارزیابی طبقه‌بندی با روش‌های مختلف برای مجموعه داده WDBCD.

طبقه‌بند	الگوریتم انتخاب نمونه	صحت	حساسیت	خصوصیت	G-Mean	محدوده ROC
C۴.۵	مجموعه داده اولیه	۹۳٫۸۵	۹۳٫۸	۹۳٫۵	۹۳٫۶۵	۰٫۹۳۳
	SMOTE	۹۵٫۱۳	۹۵٫۱۰	۹۵٫۲۰	۹۵٫۱۵	۰٫۹۵۴
	Random sampling ٪۵۰	۹۳٫۶۶	۹۳٫۷۰	۹۰٫۷۰	۹۲٫۱۹	۰٫۹۲۵
	OB-ReliefF	**۹۵٫۲۲**	**۹۵٫۲۴**	**۹۵٫۲۱**	**۹۵٫۲۲**	**۰٫۹۶۵**
Ada Boost	مجموعه داده اولیه	۹۵٫۴۳	۹۵٫۴۰	۹۴٫۶	۹۴٫۵	۰٫۹۸۹
	SMOTE	۹۶٫۵۴	۹۶٫۵	۹۶٫۶	۹۶٫۵۵	۰٫۹۹۲
	Random sampling ٪۵۰	۹۵٫۰۷	۹۵٫۱۰	۹۲٫۶	۹۳٫۸۴	۰٫۹۹۶
	OB-ReliefF	**۹۸٫۰۳**	**۹۸٫۰۰**	**۹۸٫۰۴**	**۹۸٫۰۲**	**۰٫۹۹۸**
KNN K=۳	مجموعه داده اولیه	۹۷٫۰۱	۹۷٫۰۰	۹۵٫۷۰	۹۶٫۳۵	۰٫۹۸۱
	SMOTE	**۹۷٫۳۱**	۹۷٫۳۰	**۹۷٫۴۰**	**۹۷٫۳۵**	۰٫۹۸۸
	Random sampling ٪۵۰	۹۵٫۷۷	۹۵٫۸۰	۹۰٫۳۰	۹۳٫۰۱	۰٫۹۷۸
	OB-ReliefF	۹۶٫۴۹	**۹۶٫۵۰**	۹۶٫۴۵	۹۶٫۴۸	**۰٫۹۹۱**
Bagging	مجموعه داده اولیه	۹۴٫۷۳	۹۴٫۷۰	۹۴٫۰۰	۹۴٫۳۵	۰٫۹۸۷
	SMOTE	۹۶٫۶۷	۹۶٫۷۰	۹۶٫۵۰	۹۶٫۶۰	۰٫۹۹۲
	Random sampling ٪۵۰	۹۳٫۳۱	۹۳٫۳۰	۸۹٫۲۰	۹۱٫۲۳	۰٫۹۸۶
	OB-ReliefF	**۹۶٫۷۷**	**۹۶٫۸**	**۹۶٫۸۲**	**۹۶٫۸۱**	**۰٫۹۹۷**
MLP	مجموعه داده اولیه	۹۶٫۳۱	۹۶٫۳۰	۹۵٫۳۰	۹۵٫۸۰	۰٫۹۸۸
	SMOTE	۹۶٫۹۳	۹۶٫۹	۹۷٫۲۰	۹۷٫۰۵	۰٫۹۹۶
	Random sampling ٪۵۰	۹۵٫۴۲	۹۸٫۰۰	۸۹٫۵	۹۳٫۶۵	۰٫۹۹۱
	OB-ReliefF	**۹۸٫۱۷**	**۹۸٫۲۰**	**۹۸٫۲۴**	**۹۸٫۲۲**	**۰٫۹۹۷**
Random Tree	مجموعه داده اولیه	۹۳٫۸۵	۹۳٫۸۰	۹۲٫۷۰	۹۳٫۲۵	۰٫۹۳۳
	SMOTE	۹۴٫۷۵	۹۴٫۸	۹۴٫۵۰	۹۴٫۶۵	۰٫۹۴۶
	Random sampling ٪۵۰	۹۴٫۷۲	۹۴٫۷۰	۹۰٫۵۰	۹۲٫۵۸	۰٫۹۲۶
	OB-ReliefF	**۹۶٫۶۳**	**۹۶٫۶۰**	**۹۶٫۶۴**	**۹۶٫۶۲**	**۰٫۹۶۶**
OneR	مجموعه داده اولیه	۸۹٫۱۰	۸۹٫۱۰	۸۶٫۴	۸۷٫۷۳	۰٫۸۷۸
	SMOTE	۸۹٫۷۶	۸۹٫۸	۸۹٫۷	**۸۹٫۷۵**	**۰٫۸۹۷**
	Random sampling ٪۵۰	**۸۹٫۷۹**	**۸۹٫۸۰**	۸۱٫۸	۸۵٫۷۱	۰٫۸۵۸
	OB-ReliefF	۸۸٫۶۲	۸۸٫۶۲	**۸۸٫۶۴**	۸۸٫۶۳	۰٫۸۸۶
Naive Bayes	مجموعه داده اولیه	۹۲٫۲۷	۹۲٫۳۰	۹۱٫۲	۹۱٫۷۵	۰٫۹۷۱
	SMOTE	**۹۴٫۳۷**	**۹۴٫۴۰**	**۹۴٫۶۰**	**۹۴٫۴۹**	۰٫۹۸۵
	Random sampling ٪۵۰	۹۱٫۵۵	۹۱٫۵۰	۹۰٫۴۰	۹۰٫۹۴	**۰٫۹۸۸**
	OB-ReliefF	۸۹٫۷۵	۸۹٫۷	۸۹٫۷۳	۸۹٫۷۱	۰٫۹۷۵
Kstar	مجموعه داده اولیه	۸۷٫۸۷	۸۷٫۹	۸۱٫۱	۸۴٫۴۳	۰٫۹۱۱
	SMOTE	۹۲٫۵۷	۹۲٫۶۰	۹۳٫۱۰	۹۲٫۸۵	۰٫۹۶۴
	Random sampling ٪۵۰	۹۰٫۱۴	۹۰٫۱۰	۷۹٫۳۰	۸۴٫۵۳	۰٫۹۳۲
	OB-ReliefF	**۹۷٫۶۱**	**۹۷٫۵۹**	**۹۷٫۶۲**	**۹۷٫۶۱**	**۰٫۹۸۴**
RF	مجموعه داده اولیه	۹۵٫۴۳	۹۵٫۴	۹۴٫۲۰	۹۴٫۸۰	۰٫۹۸۵
	SMOTE	۹۶٫۷۹	۹۶٫۸۲	۹۶٫۸۴	۹۶٫۸۳	۰٫۹۹
	Random sampling ٪۵۰	۹۵٫۰۷	۹۵٫۱	۹۰٫۰۰	۹۲٫۵۱	۰٫۹۹۴
	OB-ReliefF	**۹۸٫۳۱**	**۹۸٫۳۰**	**۹۸٫۳۴**	**۹۸٫۳۲**	**۰٫۹۹۷**

شکل ۴ منحنی ROC را برای انجام طبقه‌بند RF روی مجموعه داده WDBCD متوازن‌شده با OB-ReliefF نشان می‌دهد. شایان ذکر است که این طبقه‌بند بالاترین مقادیر معیارهای کارایی را در طبقه‌بندی نمونه‌ها کسب کرده است. از ارائه سایر منحنی‌های ROC برای رعایت اختصار خودداری می‌گردد.

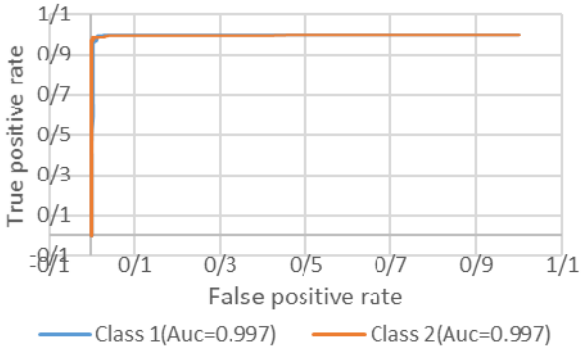
جدول ۵ معیارهای عملکرد برای SEER را ارائه می‌دهد. از این جدول نکات زیر را می‌توان استخراج کرد:

- الگوریتم OB-ReliefF بهترین نتایج را در تمام طبقه‌بندها به جز MLP ,OneR, Naive Bayes برای تمام معیارها به دست آورد.

جدول ۵: معیارهای ارزیابی طبقه‌بندی با روش‌های مختلف برای مجموعه داده SEER.

طبقه‌بند	الگوریتم انتخاب نمونه	صحت	حساسیت	خصوصیت	G-Mean	محدوده ROC
C۴.۵	مجموعه داده اولیه	۹۰٫۰۶	۹۰٫۱۰	۵۷٫۰۰	۷۱٫۶۶	۰٫۷۴۸
	SMOTE	۸۶٫۴۸	۸۶٫۵	۷۲٫۵	۷۹٫۱۹	۰٫۸۳۷
	Random sampling ٪۵۰	۹۰٫۵۶	۹۰٫۶	۵۸٫۶	۷۲٫۸۶	۰٫۸۰۴
	OB-ReliefF	**۹۳٫۷۸**	**۹۳٫۸**	**۹۳٫۸**	**۹۳٫۸**	**۰٫۹۴۷**
Ada Boost	مجموعه داده اولیه	۸۷٫۴۲	۸۷٫۴	۵۵٫۶	۶۹٫۷۱	۰٫۸۱
	SMOTE	۹۷٫۵	۹۷٫۵	۹۷٫۵	۹۷٫۵	۰٫۹۹۸
	Random sampling ٪۵۰	۹۰٫۴۱	۹۰٫۴	۶۳٫۰۰	۷۵٫۴۷	۰٫۸۵۹
	OB-ReliefF	**۹۷٫۵**	**۹۷٫۵**	**۹۷٫۵**	**۹۷٫۵**	**۰٫۹۹۸**
KNN K=۳	مجموعه داده اولیه	۸۵٫۴۱	۸۵٫۴	۳۹٫۰۰	۵۷٫۷۱	۰٫۷۰۷
	SMOTE	۸۰٫۶۵	۸۰٫۶	۶۵٫۵	۷۲٫۶۶	۰٫۸۰۴
	Random sampling ٪۵۰	۸۵٫۴۴	۸۵٫۴	۴۱٫۲	۵۹٫۳۲	۰٫۷۰۵
	OB-ReliefF	**۸۷٫۶۹**	**۸۷٫۷**	**۸۷٫۷**	**۸۷٫۷**	**۰٫۹۴۳**
Bagging	مجموعه داده اولیه	۹۰٫۰۱	۹۰٫۰۰	۵۶٫۳	۷۱٫۱۸	۰٫۸۳۷
	SMOTE	۹۵٫۹۷	۹۶٫۰	۹۶٫۰	۹۶٫۰	۰٫۹۹۶
	Random sampling ٪۵۰	۹۰٫۹۰	۹۰٫۹	۵۹٫۴	۷۳٫۴۸	۰٫۸۵۴
	OB-ReliefF	**۹۵٫۹۷**	**۹۶٫۰۰**	**۹۶٫۰۰**	**۹۶٫۰۰**	**۰٫۹۹۶**
MLP	مجموعه داده اولیه	**۸۸٫۲۵**	**۸۸٫۲**	۵۷٫۴	۷۱٫۱۵	۰٫۸۱۸
	SMOTE	۸۳٫۷۳	۸۳٫۷	۶۹٫۵	۷۶٫۲۷	۰٫۸۵۶
	Random sampling ٪۵۰	۸۸٫۱۷	۸۸٫۲	۵۸٫۱	۷۱٫۵۸	۰٫۸۳۹
	OB-ReliefF	۸۴٫۶۰	۸۴٫۶۰	**۸۴٫۶۰**	**۸۴٫۶۰**	**۰٫۹۰۱**
Random Tree	مجموعه داده اولیه	۸۳٫۳۷	۸۳٫۴	۵۵٫۲	۶۷٫۸۵	۰٫۶۹۶
	SMOTE	۸۰٫۸۴	۸۰٫۸	۶۹٫۶	۷۴٫۹۹	۰٫۷۵۵
	Random sampling ٪۵۰	۸۴٫۰۹	۸۴٫۱	۵۴٫۷	۶۷٫۸۲	۰٫۶۹۶
	OB-ReliefF	**۹۵٫۱۵**	**۹۵٫۱**	**۹۵٫۱**	**۹۵٫۱**	**۰٫۹۵۵**
OneR	مجموعه داده اولیه	۸۹٫۵۶	۸۹٫۶	۵۷٫۲	۷۱٫۵۹	۰٫۷۳۴
	SMOTE	۸۶٫۶۲	۸۶٫۶	**۶۹٫۱**	**۷۷٫۳۶**	**۰٫۷۷۹**
	Random sampling ٪۵۰	**۹۰٫۱۶**	**۹۰٫۲**	۵۸٫۷	۷۲٫۷۶	۰٫۷۴۵
	OB-ReliefF	۶۸٫۳۹	۶۸٫۴	۶۸٫۴	۶۸٫۴	۰٫۶۸۴
Naive Bayes	مجموعه داده اولیه	۸۳٫۸۰	۸۳٫۸	۵۷٫۳	۶۹٫۲۹	۰٫۸۲۲
	SMOTE	۷۹٫۶۳	۷۹٫۶	**۶۵٫۵**	**۷۲٫۲۰**	۰٫۸۲۳
	Random sampling ٪۵۰	**۸۴٫۸۹**	**۰٫۸۴۹**	۶۰٫۲	۷۱٫۴۹	**۰٫۸۳۷**
	OB-ReliefF	۶۳٫۴۳	۶۳٫۴	۶۳٫۴	۶۳٫۴	۰٫۷۲۴
Kstar	مجموعه داده اولیه	۸۵٫۹۸	۸۶٫۰۰	۴۰٫۳	۵۸٫۸۷	۰٫۷۷۷
	SMOTE	۸۵٫۴۱	۸۵٫۴	۷۴٫۳	۷۹٫۶۶	۰٫۸۸۵
	Random sampling ٪۵۰	۸۵٫۵۴	۸۵٫۵	۳۹٫۶	۵۸٫۱۹	۰٫۸۰۰
	OB-ReliefF	**۹۰٫۴۸**	**۹۰٫۵**	**۹۰٫۵**	**۹۰٫۵**	**۰٫۹۹۸**
RF	مجموعه داده اولیه	۹۰٫۰۳	۹۰٫۰۰	۵۵٫۹	۷۰٫۹۳	۰٫۸۵۰
	SMOTE	۸۸٫۱۰	۸۸٫۱۰	۷۳٫۹	۸۰٫۶۹	۰٫۹۱۶
	Random sampling ٪۵۰	۹۱٫۱۵	۹۱٫۲	۵۹٫۵	۷۳٫۶۶	۰٫۸۶۱
	OB-ReliefF	**۹۸٫۰۱**	**۹۸٫۰۰**	**۹۸٫۰۰**	**۹۸٫۰۰**	**۰٫۹۹۹**

- الگوریتم OB-ReliefF بالاترین مقادیر را در تمام معیارها با استفاده از RF کسب می‌کند.
- به استثنای الگوریتمOB-ReliefF، مقدار خصوصیت برای اکثر طبقه‌بندها مقدار پایینی است. این مورد در مجموعه‌های نامتوازن هنگامی رخ می‌دهد که طبقه‌بند به نفع کلاس اقلیت (یا اکثریت) سوگیری دارد. یعنی طبقه‌بند ممکن است بیشتر نمونه‌ها را به عنوان کلاس اقلیت (یا اکثریت) پیش‌بینی کند. در نتیجه، مقدار FP افزایش می‌یابد و خصوصیت کاهش پیدا می‌کند. در مجموعه داده‌های پزشکی که عدم تشخیص اشتباه افراد سالم به عنوان بیمار (و بالعکس) اهمیت زیادی دارد، این



شکل ۴: منحنی ROC برای طبقه‌بند RF برای طبقه بندی مجموعه داده‌ی WDBCD متوازن شده با الگوریتم OB-ReliefF.

جدول ۶: تعداد نمونه‌های انتخاب شده پس از انجام الگوریتم.

مجموعه داده	الگوریتم انتخاب نمونه	تعداد نمونه	خوش‌خیم (سالم یا زنده)	بدخیم (بیمار یا مرده)
WBCD	SMOTE	۹۴۰	۴۵۸	۴۸۲
	OB-ReliefF	۹۱۶	۴۵۸	۴۵۸
WDBCD	SMOTE	۷۸۱	۳۵۷	۴۲۴
	OB-ReliefF	۷۱۲	۳۵۶	۳۵۶
SEER	SMOTE	۴۶۴۰	۳۴۰۸	۱۲۳۲
	OB-ReliefF	۶۸۰۰	۳۴۰۰	۳۴۰۰

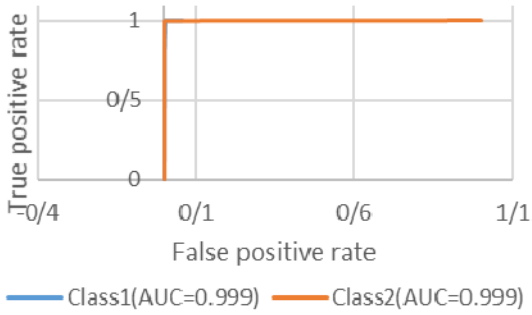
مسئله بسیار نگران کننده خواهد بود. برخی از دلایل این امر عبارتند از: ممکن است طبقه‌بند بخواهد خطای کلی را کاهش داده و توازن بین حساسیت و خصوصیت به وجود آورد. یا اینکه تکنیک نمونه‌برداری نامناسب بوده و توازن درستی بین نمونه‌های کلاس اقلیت و اکثریت ایجاد نکرده است.

شکل ۵ منحنی ROC را برای انجام طبقه‌بند RF روی مجموعه داده SEER متوازن‌شده با OB-ReliefF نشان می‌دهد. شایان ذکر است که این طبقه‌بند بالاترین مقادیر معیارهای کارایی را در طبقه‌بندی نمونه‌ها کسب کرده است. از ارائه سایر منحنی های ROC برای رعایت اختصار خودداری می‌گردد.

براساس مقادیر جداول ۳ تا ۵ نتایج زیر از اجرای الگوریتم پیشنهادی (OB-ReliefF) به‌دست آمد:

- یکی از روش‌های متوازن‌سازی مجموعه داده‌ها، استفاده از تکنیک نمونه‌برداری فزاینده است.
- هر قدر هم میزان عدم توازن مجموعه اولیه بالا باشد، OB-ReliefF قادر است کلاس‌های اقلیت و اکثریت را به‌طور کامل متوازن کند.
- به دلیل فرمول خاص تشابه نمونه‌ها، OB-ReliefF قادر به تشخیص نمونه‌های با اهمیت و تکثیر آن‌ها است.
- طبقه‌بند RF در تشخیص صحیح بیماری و معیارهای کارایی طبقه‌بند بهترین عملکرد را در تمام آزمایش‌ها داشته است.

جدول ۶ تعداد نمونه‌های انتخاب‌شده برای هر کلاس در الگوریتم SMOTE و OB-ReliefF را نشان می‌دهد. همانطور که مشاهده می‌شود، الگوریتم SMOTE نتوانسته است مجموعه داده‌ها را به طور کامل متوازن کند. علاوه بر این، با اینکه این الگوریتم در برخی موارد تعداد نمونه بیشتری ایجاد کرده است، عملکرد آن به خوبی OB-ReliefF نیست و همچنان کلاس‌ها متوازن نیستند.



شکل ۵: منحنی ROC برای طبقه‌بند RF برای طبقه‌بندی مجموعه داده‌ی SEER متوازن شده با الگوریتم OB-ReliefF.

جدول ۷: مقایسه نتایج OB-ReliefF با نتایج سایر تحقیقات.

طبقه‌بند	الگوریتم انتخاب نمونه	صحت	حساسیت	خصوصیت	G-Mean
WBCD	Asri [۲۱]	۹۷/۱۳	۹۷/۳۸	۹۶/۲۶	۹۶/۸۲
	Osareh [۱۹]	**۹۸/۸۰**	۹۵/۴۵	**۹۹/۶۳**	۹۷/۵۱
	OB-ReliefF	۹۸/۰۳	**۹۸/۰۰**	۹۸/۰۲	**۹۸/۰۱**
WDBCD	Osareh [۱۹]	۹۶/۳۳	۹۶/۸۵	۹۳/۱۱	۹۴/۹۶
	Yavuz [۲۲]	۹۶/۴۳	۹۷/۶۲	۹۵/۷۱	۹۶/۶۶
	Kadhim [۲۶]	۹۷/۳۶	۹۵/۷۴	**۹۸/۵**	۹۷/۱۱
	OB-ReliefF	**۹۸/۳۱**	**۹۸/۳۰**	۹۸/۳۴	**۹۸/۳۲**
SEER	Gurcan [۲۹]	۹۲/۸۷	-	-	-
	Sinha [۳۲]	۹۶/۷	۹۹/۷	۹۴/۰۲	۹۶/۸۲
	Gurcan [۳۳]	۹۶/۱۱	۹۴/۰۲	-	-
	OB-ReliefF	**۹۸/۰۱**	**۹۸/۰۰**	**۹۸/۰۰**	**۹۸/۰۰**

جدول ۷ مقایسه‌ای بین نتایج به‌دست‌آمده از OB-ReliefF و بهترین نتایج برخی مقالات ذکر شده در بخش کارهای مرتبط را نشان می‌دهد. همان‌طور که مشاهده می‌شود، نتایج OB-ReliefF در تمام معیارهای ارزیابی (به‌جز سه مورد) بهتر از روش‌های دیگر است. البته برخی از مقالات از معیارهایی متفاوت از این پژوهش استفاده کرده‌اند و یا تنها به معیار صحت اکتفا کرده‌اند. اما همان‌طور که پیشتر گفته شد، تنها معیار صحت برای اندازه‌گیری کارایی یک الگوریتم مناسب نیست. علاوه بر این، هرکدام از معیارهای حساسیت و خصوصیت به‌تنهایی برای ارزیابی عملکرد یک الگوریتم تشخیص کافی نیستند. زیرا هرکدام از این معیارها، عملکرد الگوریتم را در یک کلاس بررسی می‌کنند. در عمل، الگوریتمی که قادر به شناسایی صحیح نمونه‌های هر دو کلاس باشد، کارآمدتر است. بنابراین، استفاده از شاخص G-mean که ناظر به کارایی الگوریتم در هر دو کلاس است، معیار مناسبی برای ارزیابی کارایی الگوریتم در مجموعه داده نامتوازن خواهد بود .

۵- نتیجه‌گیری

یکی از خطرناک ترین بیماری‌ها در زنان سرطان سینه است. متأسفانه، به دلیل وجود مجموعه داده‌های نامتوازن جمع آوری شده در این مورد، سیستم‌های تشخیص پزشکی و الگوریتم‌های یادگیری ماشین بکار رفته در آن‌ها نمی‌توانند این بیماری را به طور دقیق تشخیص دهند. این مقاله یک روش نمونه‌برداری فزاینده به نام OB-ReliefF برای حل این مشکل ارائه می‌کند. OB-ReliefF ابتدا میزان مشابهت بین نمونه ها را

بر اساس یک شاخص محاسبه کرده و سپس نمونه‌های کلاس اقلیت را وزن‌دهی و رتبه‌بندی کرده و سپس نمونه‌های برتر در این کلاس را تکثیر می‌کند تا تعداد نمونه‌ها در کلاس‌ها متوازن شوند. این الگوریتم توانایی کار روی انواع مجموع داده‌های نامتوازن با چند کلاس یا انواع داده مختلف و حتی مقادیر مفقود را دارد. هم‌چنین، امکان متوازن‌سازی کامل کلاس‌ها و یافتن نمونه‌های باارزش که کیفیت یادگیری طبقه‌بند را افزایش می‌دهند، از دیگر ویژگی‌های این الگوریتم است. برای بررسی عملکرد الگوریتم OB-ReliefF، نتایج اجرای این الگوریتم روی سه مجموعه داده سرطان سینه با نتایج سایر الگوریتم‌ها مقایسه شد. نتایج و بررسی معیارهای کارایی حاصل از اجرای طبقه‌بندهای مختلف روی مجموعه داده‌های متوازن شده نشان می‌دهد که به‌کار بستن الگوریتم OB-ReliefF تشخیص سرطان سینه را بهبود می‌بخشد. هر چند که OB-ReliefF ممکن است در برخی موارد، برخی معیارها را بهبود نبخشد، اما در اکثریت معیارها عملکرد خوبی دارد. هم‌چنین با توجه به اینکه این الگوریتم مانند تمام الگوریتم‌های متوازن‌سازی در سطح داده، در مرحله پیش‌پردازش انجام می‌شود، سربار محاسباتی خواهد داشت. اما با توجه به اینکه کار فقط بر روی کلاس اقلیت انجام می‌شود و به علاوه امکان اجرای محاسبات به طور موازی وجود دارد، سربار محاسباتی نسبت به بسیاری از الگوریتم‌ها اندک بوده و البته با توجه به اهمیت بهبود تشخیص بیماری‌ها قابل اغماض خواهد بود. می‌توان برای کار بیشتر در آینده، الگوریتم را با روش نمونه‌برداری کاهنده ترکیب کرد و یا از OB-ReliefF برای بهبود میزان تشخیص در سایر بیماری ها استفاده کرد. هم‌چنین، می‌توان روی نرخ تکثیر و یا سایر پارامترها و معیارهای مورد استفاده در الگوریتم پیشنهادی و یا طبقه‌بندهای بکار رفته، آزمایش‌ها و تحقیقات بیشتری انجام داد.

مراجع

[1] WHO, (IARC), *International Agency for Research on Cancer, World Health Organization*, [Online]. Available: http://gco.iarc.fr/.

[2] F. Bray, *et al.*, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians,* vol. 68, no. 6, pp. 394-424. Nov 2018.

[3] L. J. Mena and J. A. Gonzalez, "Machine learning for imbalanced datasets: application in medical diagnostic," in *Proc. of the 19th Int. Florida Artificial Intelligence Research Society Conf.,* pp. 574-579, 2006.

[4] H. Parvin, B. Minaei-Bidgoli, and H. Alinejad-Rokny, "A new imbalanced learning and dictions tree method for breast cancer diagnosis," *Journal of Bionanoscience,* vol. 7, no. 6, pp. 673-678, 2013.

[5] G. Haixiang, *et al.*, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications,* vol. 73, pp. 220-239, May 2017.

[6] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino Acids,* vol. 39, no. 5, pp. 1385-1391, Nov. 2010.

[7] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data," *Knowledge-Based Systems,* vol. 94, pp. 88-104, Feb. 2016.

[8] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach," *Expert systems with Applications,* vol. 36, no. 1, pp. 690-701, Jan. 2009.

[9] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 50-59, 2004.

[10] B. W. Yap, *et al.*, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Proc. of the First Int. Conf. on Advanced Data and Information Engineering*, pp. 13-22, Kuala Lumpur, Malaysia, 16-18 Dec. 2014.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Trans. on Computer Science and Engineering,* vol. 30, pp. 25-36, 2006.

[12] N. V. Chawla, "Data mining for imbalanced datasets: An overview," In: Maimon, O., Rokach, L. (eds) *Data Mining and Knowledge Discovery Handbook,* pp. 875-886, Springer: Boston, 2005.

[13] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Applied Soft Computing,* vol. 14, pt. A, pp. 99-108, Jan. 2014.

[14] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evolutionary computation,* vol. 17, no. 3, pp. 275-306, Fall 2009.

[15] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing,* vol. 3, no. 2, pp. 224-228, Apr. 2013.

[16] E. Š. M. R. -Š. Igor Kononenko, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence,* vol. 7, no. 1, pp. 39-55, Jan. 1997.

[17] A. Luque, A. Carrasco, A. Martín, and A. D. L. Heras, "The impact of class imbalance in classification performance metrics," *Pattern Recognition,* vol. 91, pp. 216-231, Jul. 2019.

[18] 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, pt. 2, pp. 3465-3469, Mar. 2009.

[19] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *Proc. 5th Int. Symp. on Health Informatics and Bioinformatics,* pp. 114-120, Antalya, Turkey, 20-22 Apr. 2010.

[20] K. J. Wang and A. M. Adrian, "Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm," *International Journal of Computer Science and Electronics Engineering,* vol. 1, no. 3, pp. 408-412, 2013.

[21] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064-1069, 2016.

[22] E. Yavuz, C. Eyupoglu, U. Sanver, and R. Yazici, "An ensemble of neural networks for breast cancer diagnosis," in *Proc. Int. Conf. on Computer Science and Engineering*, pp. 538-543, Antalya, Turkey, 5-8 Oct. 2017.

[23] B. Zheng, S. Won 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,pt. 1, pp. 1476-1482, Mar. 2014.

[24] S. Sasikala, M. Bharathi, M. Ezhilarasi, S. Senthil, and M. R. Reddy, "Particle swarm optimization based fusion of ultrasound echographic and elastographic texture features for improved breast cancer detection," *Australasian Physical & Engineering Sciences in Medicine*, vol. 42, no. 3, pp. 677-688, Sept. 2019.

[25] M. khanna, L. k. Singh, k. Shrivastava, and R. singh, "An enhanced and efficient approach for feature selection for chronic human disease prediction: A breast cancer study," *Heliyon,* vol. 10, no. 5, Article ID: e26799, Mar. 2024.

[26] R. R. Kadhim and M. Y. Kamil, "Comparison of breast cancer classification models on Wisconsin dataset," *Int. J. Reconfigurable Embed. Syst,* vol. 11, no. 2, pp. 166-174, Jul. 2022.

[27] T. Cai, H. He, and W. Zhang, "Breast Cancer Diagnosis Using Imbalanced Learning and Ensemble Method," *Applied and Computational Mathematics,* vol. 7, no. 3, pp. 146-154, Jun. 2018.

[28] H. Ouifak and A. Idri, "On the performance and interpretability of Mamdani and Takagi-Sugeno-Kang based neuro-fuzzy systems for medical diagnosis," *Scientific African,* vol. 20, Article ID: e01610, Jul .2023.

[29] F. Gurcan and A. Soylu, "Learning from imbalanced data: Integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis," *Cancers,* vol. 16, no. 19, p. 3417, Oct.-1 2024.

[30] G. Husain, *et al.*, "SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models," *Algorithms,* vol. 18, no. 1, Article ID: 37, 16 pp., Jan. 2025.

[31] Z. Liu, H. Liu, W. Jia, D. Zhang and J. Tan, "A novel imbalanced data classification method based on weakly supervised learning for fault diagnosis.," *IEEE Trans. on Industrial Informatics*, vol. 18, no. 3, pp. 1583-1593, Mar. 2022.

[32] H. Sinha and M. Shah, "Early prediction and classification of breast cancer survival based on machine learning models," in *Proc. IEEE*

15th Annual Computing and Communication Workshop and Conf., pp. 01185-01193, Las Vegas, NV, USA, 6-8 Jan. 2025.
[33] F. Gurcan and A. Soylu, "Synthetic boosted resampling using deep generative adversarial networks: A novel approach to improve cancer prediction from imbalanced datasets," *Cancers,* vol. 16, no. 23, Article ID: 4046, Dec.-1 2024.
[34] Z. Chen, J. Duan, L. Kang, and G. Qiu, "A hybrid data-level ensemble to enable learning from highly imbalanced dataset," *Information Sciences,* vol. 554, pp. 157-176, Apr. 2021.
[35] I. Czarnowski, "Weighted ensemble with one-class classification and over-sampling and instance selection (WECOI): An approach for learning from imbalanced data streams," *Journal of Computational Science,* vol. 61, Article ID: 101614, May 2022.
[36] C.-F. Tsai, K.-C. Chen and W.-C. Lin, "Feature selection and its combination with data over-sampling for multi-class imbalanced datasets," *Applied Soft Computing,* vol. 153, Article ID: 111267, May 2024.
[37] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23-69, 2003.
[38] Z. Abbasi and M. Rahmani, "An instance selection algorithm based on ReliefF," *International Journal on Artificial Intelligence Tools,* vol. 28, no. 1, Article ID: 1950001, 2019.
[39] Z. Wang, X. Ning, and M. Blaschko, "Jaccard metric losses: Optimizing the jaccard index with soft labels," in *Proc. 37th Int. Conf. on Neural Information Processing Systems,* pp. 75259-75285, New Orleans, LA, USA ,10-16 Dec. 2023.
[40] S. Bhowmick and A. Saha, "Enhancing the performance of kNN for glass identification dataset using inverse distance weight, ReliefF ranking and SMOTE," in *Proc. 13th Int. Conf. on Material Processing and Characterization*, pp. , Hyderabad, India, 22-24 Apr. 2022, https:// https://doi.org/10.1063/5.0161083
[41] A. Moran Calderon, *Improved Distance Functions for the ReliefF Family*, MSc. Thesis, Universitat Politècnica de Catalunya, Catalunya, Spain, 2023.
[42] F. Rosenblatt, *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington DC, US, 1961.
[43] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, Oct. 2001.
[44] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences,* vol. 55, no. 1, pp. 119-139, Aug. 1997.
[45] L. Breiman, "Bagging predictors," *Machine learning ,* vol. 24, no. 2, pp. 123-140, Aug. 1996.
[46] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Information Theory ,* vol. 13, no. 1, pp. 21-27, Jan. 1967.
[47] D. Lavanya and K. U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *International Journal of Computer Applications,* vol. 26, no. 4, pp. 1-4, 2011.
[48] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, pp. 41-46, Seattle, Washington, USA, 4-6 Aug. 2001.
[49] J. G. Cleary and L. E. Trigg, "K*: An instance-based learner using an entropic distance measure," in *Proc. of the 12th Int. Conf. on Machine Learning,* pp. 108-114, Tahoe City, CA, USA, 9-12 Jul. 1995.
[50] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
[51] J. Kaur and A. Kumar, "Speech emotion recognition using CNN, k-NN, MLP and random forest," In: S. Smys, R. Palanisamy, Á.Rocha, and G. N. Beligiannis, (eds) *Computer Networks and Inventive Communication Technologies. Lecture Notes on Data Engineering and Communications Technologies*, vol 58. pp 499-509, Springer, Singapore, 2021.
[52] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," In: J. Mandal and D. Bhattacharya, (eds) *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, vol 937, 99-111, Springer, Singapore.
[53] F. Osisanwo, *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology,* vol. 48, no. 3, pp. 128-138, Jun. 2017.
[54] Q. Gu, L. Zhu and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *Proc. Int. Symp. on Intelligence Computation and Applications*, pp. 461-471, Huangshi, China, 23-25 Oct. 2009.
[55] Y. Sun, A. Wong and M. S. Kamel, "Classification of imbalanced data: A review.," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 23, no. 4, pp. 687-719, Jun. 2009.
[56] W. H. Wolberg and O. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences*, vol. 87, no. 23, pp. 9193-9196, Dec. 1990.
[57] W. Wolberg, W. Street, and O. Mangasarian, "Machine learning techniques to diagnose breast cancer from fine-needle aspirates," *Cancer Letters ,* vol. 77, no. 2-3, pp. 163-171, 1994.
[58] J. TENG, *SEER Breast Cancer Data*, IEEE Dataport, 2019.
[59] -, *UCI Machine Learning Repository*, [Online]. Available: https://archive.ics.uci.edu/ml/index.php.
[60] -, *Kaggle*, [Online]. Available: https://www.kaggle.com/datasets/sujithmandala/seer-breast-cancer-data.

زینب عباسی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر در سال ۱۳۸۵ از دانشکده دکتر شریعتی تهران و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر (گرایش نرم‌افزار) در سال ۱۳۹۰ از دانشگاه پیام نور تهران (واحد تحصیلات تکمیلی) دریافت کرد. همچنین در سال ۱۳۹۸ موفق به اخذ مدرک دکتری مهندسی کامپیوتر (گرایش نرم‌افزار) از دانشگاه اراک شد. وی در حال حاضر عضو هیأت علمی مرکز آموزش عالی محلات است. حوزه‌های پژوهشی ایشان شامل هوش مصنوعی، یادگیری ماشین و کاربردهای آن است.

.