Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
DOI: 10.22034/jscit.2026.554138.2184
Volume 15, Number 2, Summer 2026, pp. 16-30
Received: 10/20/2025, Revised: 08/05/2026, Accepted: 05/11/2026

Leukemia Classification Using a Hybrid Approach Based on Random Convolutional Kernel Transform and XGBoost

Sareh Bagheri Matak[1], Elham Askari[2*] and Sara Motamed[3]
1- Department of Computer Engineering, Ra.C., Islamic Azad University, Rasht, Iran.
2, 3- Department of Computer Engineering, FSh.C., Islamic Azad University, Fouman, Iran.
[1]sareh.bagherimatak@iau.ac.ir, [2*]askary.elham@iau.ac.ir, and [3]sara.motamed@iau.ac.ir

[*]Corresponding author's address: Elham Askari, Department of Computer Engineering, FSh.C., Islamic Azad University, Fouman, Iran

Abstract- Leukemia is one of the most common and life-threatening types of cancer worldwide. In many cases, this disease is treatable if diagnosed at an early stage. One of the most effective tools for early detection is the analysis of microarray data, which simultaneously measures the expression levels of thousands of genes. However, the high dimensionality and inherent noise of such data make the analysis process complex and time-consuming. Therefore, selecting informative genes plays a crucial role in improving accuracy and reducing the computational cost of learning models. In this study, a two-stage hybrid approach is proposed for gene selection and classification of different leukemia types. In the first stage, features are filtered using the Mutual Information criterion to select genes with the highest correlation to disease labels. In the second stage, the XGBoost model is employed for embedded ranking and stability-based selection to identify genes that consistently show the greatest importance across multiple iterations. In the final stage, classification is performed using the Random Convolutional Kernel Transform method, which enables fast and efficient learning of complex patterns among the selected genes. Experimental results on real microarray datasets demonstrate that the proposed method achieves an accuracy of 99.2%, outperforming baseline approaches and effectively identifying key genes that contribute to distinguishing between different leukemia subtypes.

Keywords- Leukemia; Random Convolutional Kernel Transform; XGBoost; Microarray; Gene.

مجله علمی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
شناسه دیجیتال: 10.22034/jscit.2026.554138.2184
جلد ۱۵، شماره ۲، تابستان ۱۴۰۵، صفحه ۱۶-۳۰
دریافت: ۱۴۰۴/۰۷/۲۸، بازنگری: ۱۴۰۵/۰۲/۱۸، پذیرش: ۱۴۰۵/۰۲/۲۱

تشخیص انواع سرطان خون با استفاده از رویکرد ترکیبی تبدیل هسته‌های پیچشی تصادفی و XG-Boost

ساره باقری[۱]، الهام عسکری*[۲]، سارا معتمد[۳]
۱- گروه کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران.
۲ و ۳- گروه کامپیوتر، واحد فومن و شفت، دانشگاه آزاد اسلامی، فومن، ایران
[1] sareh.bagherimatak@iau.ac.ir, [2*] askary.elham@iau.ac.ir, [3]sara.motamed@iau.ac.ir

* نشانی نویسنده مسئول: الهام عسکری، فومن، خیابان انتظام، دانشگاه آزاد اسلامی واحد فومن و شفت.

چکیده- سرطان خون یکی از شایع‌ترین و خطرناک‌ترین انواع سرطان در جهان است. در بسیا ری از موارد، اگر این بیماری در مراحل اولیه شناسایی شود، قابل درمان است. یکی از ابزارهای مؤثر برای تشخیص زودهنگام، تحلیل داده‌های ریزآرایه است که بیان هزاران ژن را به‌صورت هم‌زمان اندازه‌گیری می‌کند. با این حال، حجم زیاد ویژگی‌ها و وجود نویز، فرایند تحلیل را پیچیده و زمان‌بر می‌سازد. ازاین‌رو، انتخاب ژن‌های مؤثر نقش کلیدی در افزایش دقت و کاهش هزینه‌ی محاسباتی مدل‌های یادگیری دارد. در این مقاله، یک روش ترکیبی دومرحله‌ای برای انتخاب ویژگی و طبقه‌بندی انواع سرطان خون ارائه می‌شود. در گام نخست، ویژگی‌ها با استفاده از معیار اطلاعات متقابل فیلتر شده و ژن‌های دارای بیشترین ارتباط با برچسب بیماری انتخاب می‌شوند. در گام دوم، از مدل XGBoost برای رتبه‌بندی و انتخاب پایدار ویژگی‌ها استفاده می‌شود تا ژن‌هایی که در تکرارهای مختلف بیشترین اهمیت را دارند شناسایی شوند. در مرحله‌ی نهایی، طبقه‌بندی با بهره‌گیری از روش تبدیل هسته‌های پیچشی تصادفی‌انجام خواهد شد که امکان یادگیری سریع و کارآمد الگوهای پیچیده میان ژن‌های منتخب را فراهم می‌سازد. نتایج تجربی بر روی مجموعه‌داده‌های واقعی ریزآرایه نشان می‌دهد که روش پیشنهادی با دقت ۹۹,۲ درصد عملکردی برتر نسبت به روش‌های پایه داشته و توانسته است با کاهش مؤثر ابعاد داده، ژن‌های کلیدی مؤثر در تمایز انواع سرطان خون را شناسایی نماید.

واژه‌های کلیدی: سرطان خون، تبدیل هسته‌های پیچشی تصادفی، XGBoost، ریزآرایه، ژن.

۱- مقدمه

سرطان به عنوان یکی از مهم‌ترین بیماری‌های قرن حاضر و از مشکلات جوامع بشری بوده است همچنین سرطان به عنوان دومین علت مرگ و میر بعد از بیماری‌های قلبی و عروقی مطرح است. سازمان بهداشت جهانی برآورد می‌کند که سالانه بیش از ۱۰ میلیون نفر با انواع مختلف سرطان، تشخیص داده شده‌اند. انتظار می‌رود که تعداد موارد جدید ابتلا تا سال ۲۰۲۰ سالانه از ۱۰ میلیون نفر به ۱۵ میلیون نفر برسد[۱]. بدن انسان تریلیون‌ها سلول زنده می‌سازد سلول‌های طبیعی بدن رشد می‌کنند به سلول‌های جدید تقسیم می‌شوند و به طور منظم می‌میرند. سرطان یک بیماری گروهی است که توسط رشد غیر قابل کنترل و گسترش سلول‌های غیرطبیعی مشخص می‌شود. اگر گسترش این سلول‌ها کنترل نشود منجر به مرگ خواهد شد. رشد غیر قابل کنترل و حمله به بافت‌های دیگر چیزی است که یک سلول را یک سلول سرطانی می‌کند. عامل اصلی سرطان جهش‌های ژنی می‌باشد. جهش در برخی از ژن‌ها بخصوص در ژن‌هایی مثل ژن‌های بازدارنده تومور باعث عدم بیان این ژن‌ها و تغییر بیان برخی از ژن‌های دیگر می‌گردد. این تغییر در الگوی بیان ژن‌ها باعث عدم کنترل تقسیم سلولی و تکثیر بی‌رویه سلول‌ها و در نتیجه ایجاد تومور می‌گردد[۲].

کشف DNA و رشته‌های ژنتیکی مانند RNA، پروتئین و نقش آنها در کشف علل بیماری‌ها و درمان بیماری‌هایی که اساس ژنتیکی دارند تحولی عظیم در علوم زیستی ایجاد کرد. نکته‌ی اساسی در ساختار DNA که آن را متمایز و منحصر به فرد می‌کند توالی‌های آن است. رمزگشایی و یافتن توالی DNA، منجر به رمزگشایی پیغام ذخیره شده در DNA می‌شود [۳ -۵].

ریزآرایه‌ها، امکان سنجش هم‌زمان میزان بیان هزاران ژن را در شرایط مختلف فراهم می‌سازند. این ابزارها به پژوهشگران این امکان را می‌دهند تا تغییرات بیان ژن‌ها را در طول زمان یا تحت شرایط زیستی گوناگون (مانند مواجهه با یک دارو یا استرس محیطی)، رصد و تحلیل کنند. بیان ژن، فرآیندی است که طی آن اطلاعات موجود در ژن، (بخشی از DNA) به یک محصول فعال مانند پروتئین ترجمه می‌شود. این فرآیند بسته به شرایط محیطی و نیازهای سلول می‌تواند تغییر کند. اطلاعات حاصل از ریزآرایه‌ها معمولاً به صورت ماتریس‌هایی هستند که در آن‌ها سطوح بیان ژن‌ها در زمان‌ها یا شرایط مختلف ثبت شده‌اند [۴]. تحلیل این داده‌ها، به ویژه برای شناسایی الگوهای مشترک یا متفاوت در بیان ژن‌ها، می‌تواند در درک بهتر فرایندهای زیستی، کشف بیماری‌ها و طراحی داروها نقش مؤثری ایفا کند [۵].

با توجه به اینکه داده‌های ریزآرایه‌ای معمولاً در چندین بازه‌ی زمانی گردآوری می‌شوند این مسئله ذاتاً یک مسئله از نوع سری زمانی محسوب می‌شود. در این نوع داده‌ها، هدف آن است که روند تغییرات بیان ژن در گذر زمان تحلیل شده و از این طریق، رفتار ژن‌ها در پاسخ به محرک‌های مختلف شناسایی شود. چنین تحلیلی می‌تواند اطلاعات ارزشمندی درباره‌ی ژن‌های کلیدی در یک فرآیند زیستی، یا ژن‌هایی که به‌صورت هم‌زمان تنظیم می‌شوند، ارائه دهد [۵].

توسعه و پیشرفت تکنیک‌های هوش مصنوعی باعث پیشرفت‌های شگرفی در زمینه علم پزشکی شده است اما همچنان تشخیص بیماری‌های سرطانی امری دشوار است. به دست آوردن تجربه کامل در زمینه تشخیص بیماری‌های سرطانی برای یک پزشک متخصص نیاز به بازه زمانی طولانی و کار عملی بسیار دارد. وجود یک سیستم کمک‌یار پزشک که قابلیت پیش‌بینی دقیق بیماری‌های سرطانی را داشته باشد ضروری به نظر می‌رسد. شناسایی ساختار DNA و رمزگشایی آن در علم پزشکی کاربردهای فراوانی دارد. تعیین توالی DNA به علم پزشکی و مطالعات آزمایشگاهی کمک می‌کند تا عامل ایجاد سرطان خون در بدن موجودات شناخته شود.

بیماری سرطان از نظر تجربی یک بیماری غیرقابل پیش‌بینی، مخفی و با علایم نامحسوس می‌باشد. یکی از رویکردهای بررسی و پیش‌بینی در درمان و جلوگیری از پیشرفت، استفاده از سری زمانی با رویکرد عمیق می باشد. به همین دلیل این بیماری و پیش‌بین‌های قبل از اعلام خطر حاد شدن و به مرحله پیشرفته رسیدن بیماری، با شبکه عصبی پیچشی می‌تواند تشخیص داده شود. بعلاوه می‌توان بیان کرد که آیا دیتاهای بدست آمده در نهایت باعث سرطان خون می‌شوند یا خیر و این سرطان در چه بازه زمانی در چه مرحله‌ای قرار می‌گیرد. انتخاب ویژگی برای حذف ویژگی‌های اضافی و نامربوط بسیار ضروری و سبب بهبود عملکرد طبقه‌بندی می‌شود [۵].

در این مقاله، روشی ترکیبی برای تشخیص انواع سرطان خون بر پایه‌ی تحلیل تغییرات DNA و الگوهای بیان ژنی ارائه می‌شود. هدف این روش، افزایش دقت تشخیص از طریق شناسایی ژن‌های مؤثر و کاهش ابعاد داده‌های ریزآرایه است. در گام نخست، تمامی داده‌های ژنتیکی پیش‌پردازش شده و ویژگی‌ها با استفاده از معیار اطلاعات متقابل[۱] ارزیابی می‌شوند تا ژن‌هایی که بیشترین ارتباط آماری با برچسب بیماری دارند، شناسایی گردند. سپس، در مرحله‌ی دوم از الگوریتم XGBoost برای رتبه‌بندی جاسازشی[۲] و انتخاب پایدار ویژگی‌ها استفاده می‌شود؛ به‌گونه‌ای که تنها ژن‌هایی که در تکرارهای متعدد بیشترین اهمیت را نشان می‌دهند به‌عنوان ورودی مدل نهایی در نظر گرفته می‌شوند. در مرحله‌ی نهایی، فرآیند طبقه‌بندی با استفاده از تبدیل هسته‌های پیچشی تصادفی[۳] انجام می‌شود. در این روش، مجموعه‌ای از کرنل‌های پیچشی به‌صورت تصادفی مقداردهی اولیه شده و با استخراج الگوهای پیچیده از داده‌های ژنی، نمایه‌های متمایزی از انواع سرطان خون تولید می‌کنند ROCKET .به‌دلیل ماهیت سبک و غیرپارامتری خود، در مجموعه‌داده‌هایی با تعداد ویژگی بسیار زیاد و تعداد نمونه محدود عملکرد قابل‌توجهی دارد. بدین‌ترتیب خروجی لازم تولید و ۵ دسته از سرطان خون موجود در نمونه را پیش‌بینی خواهد نمود. در این مطالعه، هدف اصلی استفاده از ترکیب دو الگوریتم پیشرفته XGBoostو ROCKET است که به‌طور همزمان چالش‌های انتخاب ویژگی و مدل‌سازی روابط پیچیده در داده‌های ریزآرایه را حل می‌کند. برخلاف روش‌های موجود که به‌طور جداگانه از این الگوریتم‌ها استفاده کرده‌اند، این ترکیب به‌ویژه برای داده‌های با ابعاد بالا و نمونه‌های محدود بهبود عملکرد قابل توجهی را به همراه دارد. استفاده از XGBoost برای انتخاب ویژگی‌های مؤثر و ROCKET برای استخراج الگوهای پیچیده در این مطالعه به‌عنوان یک راه‌حل جدید و نوآورانه در طبقه‌بندی داده‌های ژنتیکی معرفی شده است. ادامه مقاله به این صورت تقسیم‌بندی شده است، در بخش ۲ مروری بر ادبیات موضوع مورد بررسی قرار می‌گیرد در بخش ۳ روش

پیشنهادی تشریح می‌شود و در بخش ۴ نتایج ارزیابی بحث خواهد شد و در بخش ۵ نتیجه گیری ارایه خواهد شد.

۲- مروری بر کارهای گذشتگان

شهاب و همکاران [۴]، در سال ۲۰۲۲ در مقاله‌ای به بررسی یادگیری ماشین در کاربردهای پزشکی پرداختند. این مقاله مروری جامع بر استفاده از یادگیری ماشینی در زمینه پزشکی ارائه می‌کند که بر فناوری‌های استاندارد و چگونگی تأثیر آنها بر تشخیص پزشکی تأکید می‌کند. پنج کاربرد عمده پزشکی عمیقاً در این مقاله مورد بحث قرار گرفته‌اند که بر روی تطبیق مدل‌های یادگیری ماشین برای حل مشکلات سرطان، شیمی پزشکی، مغز، تصویربرداری پزشکی و حسگرهای پوشیدنی تمرکز دارند. در نهایت، این نظرسنجی منابع و راهنمایی‌های ارزشمندی را برای محققان، متخصصان و تصمیم‌گیرندگان فراهم می‌کند که مسیرهای تحقیق و توسعه آینده را چارچوب‌بندی می‌کند. طبق نتایج روش های یادگیری ماشین و توانایی آنها در بررسی بسیاری از داده‌ها و کشف روابط پیچیده، ارائه تفسیر و شناسایی الگوها دقیق‌تر انجام می‌شود. یادگیری ماشینی می تواند به افزایش قابلیت اطمینان، پیش‌بینی‌پذیری و دقت سیستم‌های تشخیص برای بسیاری از بیماری‌ها کمک کند.

کومار و القحطانی [۵]، به تشخیص سرطان مبتنی بر یادگیری عمیق، پیشرفت‌ها، روند و چالش‌های آن پرداختند. این مطالعه مروری جامع بر روش‌های یادگیری عمیق در تشخیص سرطان، که عمدتاً بر سرطان سینه، مغز، پوست و سرطان پروستات متمرکز است، می‌باشد. آنان مدل‌های مختلف یادگیری عمیق و مراحل بکارگیری مدل‌های یادگیری عمیق در تشخیص سرطان را توصیف می‌کنند. نتیجه این مطالعه سرنخ‌های زیادی برای توسعه سیستم‌های تشخیص سرطان عملی و دقیق برای تشخیص زودهنگام و برنامه‌ریزی درمان ارائه می‌دهد. در واقع، روش‌های یادگیری عمیق به دلیل توانایی خود در استخراج خودکار ویژگی‌های ذاتی از تصاویر بدون وابستگی به ویژگی‌های استخراج‌شده دستی، نتایج فوق‌العاده‌ای را در زمینه‌های مختلف گزارش کرده‌اند.

شری و همکاران [۶]، یک رویکرد یادگیری عمیق منحصر به فرد به نام شبکه عصبی عود کننده عمیق بهینه شده[۴] برای شناسایی بیماری لوسمی با تجزیه و تحلیل تصاویر میکروسکوپی نمونه‌های خون ارائه کردند. آنها پس از استفاده از شبکه‌های عصبی عود کننده عمیق[۵]، از الگوریتم بهینه‌سازی گوزن قرمز برای بهینه‌سازی وزن به‌دست آمده استفاده نمودند. مدل ارائه شده بر روی دو مجموعه داده نمونه خون لوسمی در دسترس، AML، ALL_IDB1 و ALL_IDB2 ارزیابی شده است و موفق به تشخیص بدخیمی لوسمی شده است.

شاهین و همکاران [۷]، نقش یادگیری ماشین را با استفاده از شبکه‌های عصبی پیچیده عمیق بر روی سه مجموعه داده منحصر به فرد برای پیش‌بینی خطر ابتلا به سرطان را بررسی نمودند و با تجزیه و تحلیل تصاویر پزشکی عمل تشخیص را بهبود بخشیدند. علاوه بر این، در این مطالعه از CNNهای پیشرفته برای تشخیص سرطان در تصاویر استفاده شده است که با استفاده از لایه‌های پیچشی تشخیص دهنده، ویژگی‌ها را با لایه‌های جمع‌کننده کاهنده پیچیدگی ترکیب می‌کند و شناسایی موثر سرطان را تضمین می‌کند.

سعید و همکاران [۸]، از فناوری CNN برای ارائه راه حلی خودکار برای شناسایی لوسمی در بیماری حاد لنفاوی استفاده نمود. شبیه‌سازی‌ها از لوسمی حاد لنفوبلاستیک IDB 1 و لنفوبلاستیک lb 2 استفاده کرده است. کارایی مدل پیشنهادی براساس بررسی کیفی با تجزیه و تحلیل لایه میانی راه‌انداز، فیلترها و لایه‌های نقشه حرارتی انجام شد. مشاهده شد که روش پیشنهادی آنها با دقت ۹۹,۶۱ درصد، قادر به تشخیص سرطان لنفاوی حاد می‌باشد. آنان برای حل مشکل بیش برازش مدل، از تکنیک های افزایش داده ها برای تولید تصاویر استفاده نموده‌اند.

بتول و همکاران [۹]، از روش‌های یادگیری عمیق برای ایجاد یک روش دقیق برای تشخیص سرطان استفاده نمودند. روش آنها یک مدل مقاوم با کمک یادگیری عمیق سبک وزن را هدف قرار داده است و از پیچش‌های قابل تفکیک عمقی برای افتراق لوسمی لنفوبلاستیک حاد و سلول‌های طبیعی در مجموعه داده ای از تصاویر WBC استفاده می‌کند. در این روش، از ویژگی‌های آموزش‌پذیر کمتری برای افزایش اثربخشی و همچنین عملکرد طبقه‌بندی لوسمی استفاده می‌شود.

اکثر روش‌های پیشنهادی، استفاده از مدل‌های کامپیوتری برروی پردازش تصاویر ژن متمرکز شده‌اند [۱۰]. با وجود اینکه همه روش‌ها پیش‌بینی‌های مربوط به بیولوژیکی و عملکردی را انجام می‌دهند اما هنوز جایی برای بهبود عملکرد وجود دارد. ویژگی‌های ساختاری و ترکیب فضایی ژن‌های خاص مثل پروتئین‌ها در نظر گرفته نمی‌شود. از سوی دیگر، مدل‌های از پیش آموزش دیده شده دارای محدودیت‌هایی از جمله پیچیدگی با پارامترهای اضافی، ناکارآمدی به عنوان تابعی از زمان پردازش و ویژگی‌های تمایز بین تصاویرمختلف هستند. برای غلبه بر این مشکلات و ایجاد یک راه حل مناسب برای طبقه بندی سرطان لوسمی، به تکنیک‌های جدید

یادگیری عمیق می‌بایست پرداخته شود [۱۱-۲۳].

مومن‌زاده و همکاران [۲۴]، به بررسی روش‌های انتخاب ویژگی مبتنی بر مدل مخفی مارکوف[۶] ، برای داده‌های ریزآرایه می‌پردازد. در این مطالعه، استفاده از HMM برای رتبه‌بندی و انتخاب ویژگی‌های موثر در طبقه‌بندی سرطان‌های مختلف مورد تاکید قرار گرفته است. این روش با کاهش ابعاد داده‌ها و حذف ویژگی‌های غیرمرتبط، دقت طبقه‌بندی را افزایش می‌دهد.

دمپستر و همکاران [۲۵]، به توسعه روش طبقه‌بندی داده‌های سری زمانی پرداخته‌اند. این روش از هسته‌های تصادفی در الگوریتم‌های زمان‌سنجی استفاده می‌کند و نشان می‌دهد که استفاده از این روش‌ها می‌تواند منجر به افزایش سرعت و دقت در پردازش داده‌های بزرگ و پیچیده مانند داده‌های ژنتیکی شود. این روش برای دسته‌بندی داده‌ها با ویژگی‌های متغیر و پیچیده بسیار موثر است.

با وجود پیشرفت‌های قابل توجه در استفاده از روش‌های یادگیری ماشین و یادگیری عمیق برای تشخیص سرطان خون، هر یک از رویکردهای موجود دارای محدودیت‌هایی هستند. روش‌های مبتنی بر یادگیری عمیق، علی‌رغم دقت مناسب، اغلب به حجم بالایی از داده‌های آموزشی نیاز دارند و در مواجهه با داده‌های ریزآرایه با تعداد نمونه محدود، مستعد بیش‌برازش هستند. علاوه بر این، بسیاری از این روش‌ها فاقد سازوکار مؤثر برای انتخاب ویژگی بوده و مستقیماً بر روی داده‌های با ابعاد بسیار بالا اعمال می‌شوند که منجر به افزایش پیچیدگی محاسباتی و کاهش پایداری مدل می‌گردد. از سوی دیگر، روش‌های کلاسیک یادگیری ماشین اگرچه از نظر محاسباتی ساده‌تر هستند، اما معمولاً توانایی کافی در مدل‌سازی روابط غیرخطی و الگوهای پیچیده بیان ژن را ندارند. همچنین، اغلب مطالعات پیشین تنها بر یک مرحله انتخاب ویژگی تکیه دارند و پایداری ژن‌های انتخاب‌شده در تکرارهای مختلف را در نظر نمی‌گیرند.

در پاسخ به این چالش‌ها، روش پیشنهادی این مقاله با ارائه یک چارچوب ترکیبی دو مرحله‌ای برای انتخاب ویژگی و یک طبقه‌بند کارآمد، تلاش می‌کند محدودیت‌های موجود را برطرف نماید. استفاده از اطلاعات متقابل در مرحله نخست، کاهش سریع ابعاد داده و حذف ویژگی‌های نویزی را ممکن می‌سازد، در حالی که XGBoost با انتخاب پایدار ویژگی‌ها، ژن‌های مؤثر و قابل اعتماد را شناسایی می‌کند. در نهایت، به‌کارگیری ROCKET امکان استخراج الگوهای پیچیده و بهبود دقت طبقه‌بندی را بدون افزایش قابل توجه هزینه محاسباتی فراهم می‌آورد.

۳- روش پیشنهادی

بسیاری از تکنیک‌های تحلیلی استاندارد برای تجزیه و تحلیل داده‌های با توان بالا نامناسب یا از نظر محاسباتی غیرممکن هستند. استفاده از ژن‌های غیرمرتبط در تجزیه و تحلیل داده‌ها ابعاد مسئله را بزرگ‌تر می‌کند و هزینه محاسباتی را افزایش می‌دهد. روش پیشنهادی این پژوهش بر پایه‌ی تحلیل تغییرات DNA و شناسایی الگوهای ژنتیکی مؤثر در بروز سرطان خون طراحی شده است. با توجه به اینکه تغییرات در ساختار ژنوم، از جمله تغییر تعداد کپی DNA، نقش مهمی در ایجاد و پیشرفت سرطان خون دارد، شناسایی این تغییرات می‌تواند نقش بسزایی در تشخیص زودهنگام و دقیق بیماری ایفا کند. در این پژوهش، به‌منظور افزایش دقت مدل و کاهش ابعاد داده، از یک رویکرد ترکیبی دومرحله‌ای در انتخاب ویژگی‌ها استفاده شده است. در گام نخست، ویژگی‌های استخراج‌شده از داده‌های ژنتیکی با استفاده از معیار اطلاعات متقابل ارزیابی می‌شوند تا ژن‌هایی که بیشترین وابستگی آماری را با برچسب نوع سرطان دارند شناسایی شوند. در گام دوم، از الگوریتم XGBoostبرای رتبه‌بندی جاسازشی و انتخاب پایدار ویژگی‌ها بهره گرفته می‌شود [۲۶]. این مرحله سبب می‌شود ویژگی‌هایی که در تکرارهای مختلف بیشترین اهمیت را دارند با وزن بالاتری در مدل لحاظ شوند و ویژگی‌های غیرمؤثر حذف گردند. در نهایت طبقه‌بندی نهایی با استفاده از Rocket انجام خواهد شد[۲۵ و ۲۷]. با توجه به اینکه داده ها پویا و از نوع سری زمانی هستند و تغییر ژنها در مرور زمان اتفاق می‌افتد نیاز به استفاده از الگوریتم‌های پویا است. اصولا داده‌های سری زمانی برای طبقه‌بندی دارای پیچیدگی و محاسبات بالایی هستند و به زمان آموزش قابل توجهی نیاز دارند. برای اینکه تعداد کرنل ها در Rocket، زیاد است ویژگی‌های مختلف در نظر گرفته می‌شود و افزایش کرنل ها منجر به دقت در طبقه بندی می‌شود در حالی‌که مدلهای دیگر روی یک ویژگی تمرکز می‌کنند و بر اساس یک ویژگی (فرکانس – واریانس) خروجی ها را تشخیص می‌دهند. روش پیشنهادی این مطالعه، ترکیب XGBoost و ROCKET است که به‌طور خاص برای داده‌های ریزآرایه و تشخیص انواع لوسمی طراحی شده است. در مرحله اول، ویژگی‌ها با استفاده از معیار اطلاعات متقابل انتخاب می‌شوند تا ژن‌های مؤثرتر در ارتباط با بیماری شناسایی گردند. سپس، ازXGBoost برای رتبه‌بندی ویژگی‌ها **و** انتخاب ویژگی‌های پایدار استفاده می‌شود. در مرحله دوم، ROCKET به‌عنوان یک طبقه‌بند برای استخراج الگوهای پیچیده به کار می‌رود. این ترکیب به دلیل استفاده از الگوریتم‌های قوی در هر مرحله و توجه به ویژگی‌های خاص داده‌های ژنتیکی، قادر به حل مشکلاتی است که در روش‌های پیشین قابل شبیه‌سازی

نبوده است. در این مقاله، داده‌های ژنومیک حاصل از ریزآرایه‌ها به‌عنوان داده‌های سری‌زمانی چندمتغیره پردازش شده‌اند. هر نمونه بیمار شامل سطوح بیان هزاران ژن در چندین زمان متوالی است که پس از پیش‌پردازش (شامل کدگذاری انواع لوسمی به اعداد ۱ تا ۵، نرمال‌سازی مقادیر ژن بین ۰ و ۱، و اعتبارسنجی متقابل ۱۰بخشی) به شکل یک ماتریس سه‌بعدی N×T×G سازماندهی می‌شود N، تعداد بیماران، T تعداد نقاط زمانی (مراحل اندازه‌گیری) و G تعداد ژن‌ها (۲۲۲۸۴ ویژگی). در این نگاه، هر ژن به‌عنوان یک سیگنال سری‌زمانی مستقل در نظر گرفته شده که تغییرات آن در طول زمان می‌تواند الگوی مشخصی برای هر نوع لوسمی ایجاد کند.

برای استخراج الگوهای زمانی مؤثر از این داده‌های چند بعدی، از روش ROCKET استفاده شده که مبتنی بر اعمال هزاران کرنل پیچشی تصادفی بر روی هر سری‌زمانی ژن است. هر کرنل با پارامترهای تصادفی (طول ۷، ۹ یا ۱۱؛ وزن‌های نمونه‌گرفته از توزیع نرمال؛ بایاس یکنواخت؛ و اتساع تصادفی) ویژگی‌هایی مانند بیشینه، نسبت مقادیر مثبت و میانگین را از حاصل پیچش استخراج می‌کند. این ویژگی‌های زمانی استخراج‌شده سپس در یک بردار ویژگی یکپارچه‌شده و به یک طبقه‌بند رگرسیون ریج منتقل می‌شوند. این رویکرد به‌طور همزمان توانایی مدلسازی روابط پیچیدهی زمانی و مقیاس‌پذیری در برابر ابعاد بسیار بالای داده‌های ژنومیک را فراهم می‌آورد.

بلوک دیاگرام روش پیشنهادی در شکل ۱ نشان داده شده است. در ادامه هر یک از گام‌های روش پیشنهادی جزء به جزء شرح داده خواهد شد.

۳-۱- پیش‌پردازش داده‌ها

برروی کلیه داده‌های مجموعه داده، پیش‌پردازش لازم از جمله خارج کردن حروف اضافی و هدرها برای برخورد با داده‌ها به صورت خام انجام می‌شود. انواع لوسمی با کدهای متناظر نشان داده خواهند شد. نوع AML با کد ۱، نوع PB با کد ۲، نوع PBSC_CD34 با کد ۳، نوع Bone_Marrow با کد ۴ و نوع Bone_Marrow_CD34 با کد ۵ جایگزین شده است و ادامه روش پیشنهادی با این کدینگ‌ها انجام خواهد شد.

۳-۲- انتخاب ویژگی

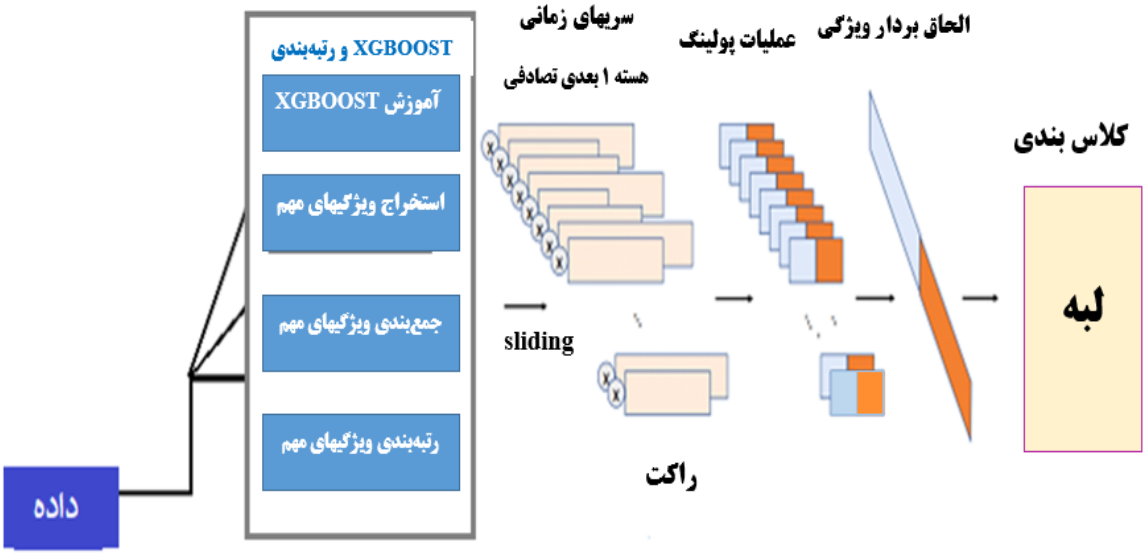
روش استفاده شده در این مقاله، روش دو مرحله‌ای فیلتر سریع با اطلاعات متقابل و XGBOOST است که در این بخش جزییات آن بیان می‌شود.

۳-۲-۱-فیلتر سریع با اطلاعات متقابل

در گام نخست، از معیار اطلاعات متقابل برای ارزیابی ارتباط آماری میان هر ژن و متغیر هدف (نوع سرطان) استفاده می‌شود. معیار MI به‌صورت زیر تعریف می‌شود:

$$\frac{p(x_i,j)}{p(x_j)p(y)}p(x_j,y)log\sum_{y\in Y}\sum_{x_j\in X}=I(X_j,Y) \qquad (۱)$$

که در آن X_j نمایانگر بیان ژن -jام و Y برچسب کلاس (نوع سرطان) است. ویژگی‌هایی که مقدار اطلاعات متقابل بالاتری دارند به‌عنوان ژن‌های مؤثرتر در تشخیص در نظر گرفته می‌شوند. بر اساس مقادیر



شکل ۱. بلوک دیاگرام روش پیشنهادی

MI، تعداد K_1 ژن برتر برای ورود به مرحله دوم انتخاب می‌شوند. این مرحله سریع بوده و باعث حذف نویز و کاهش بار محاسباتی مرحله بعد می‌شود [۲۸].

۲-۲-۳-رتبه‌بندی جاسازشی با XGBoost

در مرحله دوم، از الگوریتم XGBoost به‌عنوان یک مدل یادگیری جاسازشی استفاده می‌شود تا اهمیت ویژگی‌ها بر اساس سهم آن‌ها در بهبود دقت مدل تعیین گردد. XGBoost، با ترکیب مجموعه‌ای از درخت‌های تصمیم به‌صورت ترتیبی و با یادگیری گرادیانی، توانایی بالایی در مدل‌سازی روابط غیرخطی دارد. تابع هدف آن به‌صورت زیر تعریف می‌شود [۲۶]:

$$\sum_{k=1}^{K}\Omega(f_k)+\sum_{i=1}^{n}l(y_i,\hat{y_i})=\zeta^2||\lambda||w\frac{1}{2}+\gamma T \quad (۲)$$
$$=\Omega(f)$$

که در آن l تابع زیان، Ω(f) جمله‌ی منظم‌سازی برای کنترل پیچیدگی مدل، و T تعداد برگ‌های درخت است. پس از آموزش مدل، اهمیت هر ژن با استفاده از معیار Gain محاسبه می‌شود:

$$\Delta l_s\sum_{j^S}\frac{1}{|j^S|}=Gain(f_i) \quad (۳)$$

که در آن Δl_s میزان کاهش زیان در تقسیمات درختی مربوط به ویژگی f_j است. در این مقاله، مقدار K_1 (تعداد ژن‌های انتخاب‌شده در مرحله‌ی اول بر اساس معیار اطلاعات متقابل) برابر با ۲۰۰۰ در نظر گرفته شد. این مقدار از طریق تحلیل حساسیت اولیه و با هدف حفظ حداقل ٪۹۵ از اطلاعات متقابل کل تعیین گردید.

در مرحله‌ی دوم، مدل XGBoost روی ۲۰۰۰ ژن باقی‌مانده آموزش داده شد. اهمیت هر ژن با استفاده از معیار Gain محاسبه گردید. برای دستیابی به انتخاب پایدار، مدل XGBoost در ۱۰ تکرار مستقل (با دانه‌های تصادفی متفاوت) اجرا شد. ژن‌هایی که در حداقل ۸ تکرار از ۱۰ تکرار جزو ۲۰۰ ژن برتر بودند، به عنوان ویژگی‌های پایدار در نظر گرفته شدند.

این فرآیند منجر به انتخاب نهایی ۱۷۲ ژن به عنوان ورودی طبقه‌بند ROCKET شد. میانگین تعداد ژن‌های انتخابی در ۱۰ اجرای مستقل ۸±۱۷۲بود.

۳-۳- Rocket [۷]

گام نهایی نیز برای طبقه‌بندی و استفاده از نتایج مراحل قبل، استفاده از طبقه‌بند Rocket است [۲۵و۲۷]. این طبقه‌بندی برای سری‌های زمانی کاملا مناسب است و عملیات طبقه‌بندی را با دقت مناسبی انجام می‌دهد. اصولا داده های سری زمانی برای طبقه‌بندی دارای پیچیدگی و محاسبات بالایی هستند و به زمان آموزش قابل توجهی نیاز دارند اما دقت بالایی دارند. برای اینکه تعداد کرنل ها در Rocket، زیاد است ویژگی‌های مختلف در نظر گرفته می‌شود و افزایش کرنل‌ها منجر به دقت در طبقه‌بندی می‌شود در حالی‌که مدل‌های دیگر روی یک ویژگی تمرکز می‌کنند و بر اساس یک ویژگی (فرکانس – واریانس) خروجی‌ها را تشخیص می‌دهند. جزییات طبقه‌بند با این روش به تفصیل در ادامه شرح داده شده است.

اگر $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ نشان دهنده داده‌های آموزشی باشد که x_n یک سری زمانی و y_n برچسب مربوطه است. Rocket دارای سه مرحله اصلی است که عبارتند از: پیش آموزش، بهینه سازی و پس از آموزش، همانطور که در شکل ۲ نشان داده شده است.

راه‌اندازی هسته‌ها: مجموعه‌ای از D هسته‌های پیچش تصادفی $\Theta=(W_1,\ldots,W_D)$، مقداردهی اولیه می‌شود به طوری که برای هر کرنل $w\in\Theta$ مقادیر زیر تنظیم می‌شوند:

طول: طول هر هسته به طور تصادفی از {۷، ۹، ۱۱} با احتمال مساوی انتخاب می شود.

تاخیر[۸] : نمونه‌برداری از مقیاس نمایی ⌊ K=⌊2^a، که در آن a~U(,⟦log⟧_2 (|x|-1)/(|w|-1)) کاردینالیته (طول) سیگنال ورودی و |w| طول هسته است.

پدینگ[۹] : یک تصمیم تصادفی باینری با احتمال مساوی برای اعمال padding به ورودی x گرفته می شود. با padding، هسته‌ها در اولین و آخرین شاخص‌های x قرار می‌گیرند.

گام[۱۰] : مقدار گام همیشه برای همه هسته‌ها روی یک تنظیم می‌شود.

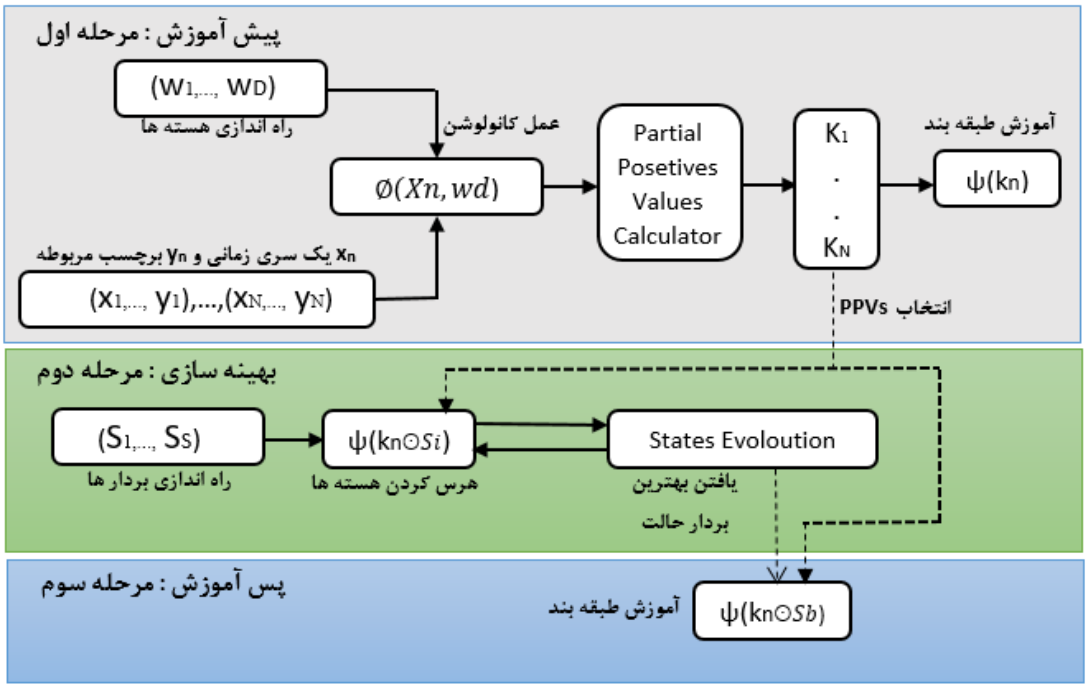
استخراج ویژگی‌ها: دو جنبه مهم Rocket از نظر دستیابی به دقت پیشرفته، استفاده از انبساط، نمونه برداری در مقیاس نمایی و استفاده از PPV است.

وزن‌ها: وزن‌ها به طور تصادفی از توزیع نرمال N(۰ ،۱) نمونه‌برداری می شوند.

بایاس: مقدار بایاس برای هر هسته به طور تصادفی از توزیع یکنواخت U (-۱ ,۱) انتخاب می‌شود.

پیچش: عمل پیچش φ(•) برای سری زمانی x و کرنل پیچش w است.

$$f=\emptyset(x,w)=(\sum_{j=1}^{|w|}x_{i+j.k}.w_j|i=1,\ldots,|x|) \quad (۴)$$



شکل ۲. بلوک دیاگرام Rocket

که در آن f بردار ویژگی استخراج شده است. از این رو، برای سری زمانی ورودی x_n، بردار ویژگی استخراج شده با استفاده از بانک هسته‌های پیچش φ با $k_n=(k_1,\ldots,k_D)$ است.

آموزش طبقه‌بندی‌کننده: طبقه‌بندی‌کننده رگرسیون ستیغی[۱۱] $\psi(k)$ با استفاده از ویژگی‌های استخراج‌شده $\{(k_1,y_1),\ldots,(k_n,y_n)\}$ مربوط به سری‌های زمانی ورودی آموزش داده می‌شود. این طبقه‌بندی‌کننده به تنظیمات فرا پارامتر گسترده نیاز ندارد، به جز پارامترهای منظم‌سازی که می‌توانند به سرعت با اعتبارسنجی متقاطع تنظیم شوند.

ضریب منظم‌سازی طبقه‌بندی‌کننده رگرسیون ستیغی بر اساس جستجوی اعتبارسنجی متقابل در مجموعه ۱۰ عددی با فاصله مساوی در محدوده مقیاس ورود به سیستم [۳، ۳-] تنظیم شد. این‌کار در واقع برای جلوگیری از بیش‌برازش موثر است. تعداد هسته‌های تصادفی D =10000 ، تعداد دوره‌های آموزشی ۵۰۰ است و میانگین نتایج ۱۰ اجرا مستقل گزارش خواهد شد. تعداد بردارهای حالت کاندید S =8 است که ۵۰٪ آنها از توزیع برنولی با احتمال ۰,۵ مقداردهی اولیه خواهد شد و بقیه به صورت ۱ مقداردهی اولیه می‌شود و سپس به هم ریخته می‌شوند [۲۹].

در اینجا تأثیر پارامترهای مختلف هسته بر دقت طبقه‌بندی بررسی می‌شود. همچنین دقت پیکربندی پیش‌فرض (یعنی با استفاده از پارامترهای مشخص‌شده در قسمت بالا) با انتخاب‌های مختلف برای تعداد هسته‌ها، طول، وزن‌ها و بایاس، اتساع، لایه‌بندی و ویژگی‌های خروجی مقایسه می‌شود. در هر مورد، تنها پارامتر داده شده تغییر می‌کند و تمام پارامترهای دیگر در مقادیر پیش فرض ثابت نگه داشته می‌شوند. نتایج میانگین نتایج بیش از ۱۰ اجرا (با استفاده از مجموعه متفاوتی از هسته‌های تصادفی در هر اجرا) هستند. در بیشتر موارد، پیکربندی‌های جایگزین نشان دهنده یک تغییر نسبتاً ظریف از پیکربندی پیش فرض است. بنابراین جای تعجب نیست که در بسیاری از موارد یک یا چند گزینه جایگزین برای پارامتر مربوطه، دقتی مشابه با پیکربندی خط پایه ایجاد کند. به عبارت دیگر، Rocket نسبتاً قوی است تا گزینه‌های مختلف را برای بسیاری از پارامترها متفاوت کند. مقادیر پارامترهای مورد استفاده در جدول ۱ ارائه شده است.

تحلیل حساسیت روی هایپرپارامترهای کلیدی شامل تعداد کرنل‌ها، طول کرنل، اتساع و پدینگ در روش ROCKET و همچنین پارامترهای تعداد درختان، عمق و نرخ یادگیری در XGBoost به‌طور سیستماتیک انجام شده است. این تحلیل از طریق جستجوی شبکه‌ای[۱۲] همراه با اعتبارسنجی متقابل ۵-بخشی روی زیرمجموعه‌ای از داده‌های آموزش انجام گرفت و عملکرد با معیار F1-Scoreوزنی ارزیابی شد. نتایج نشان داد که مدل به‌ویژه به تعداد کرنل‌های ROCKET و تنظیم وزن کلاس‌ها در XGBoost حساسیت بیشتری دارد، به‌طوری که افزایش کرنل‌ها تا ۱۰۰۰۰ و

اعمال وزن متعادل‌کننده کلاس‌ها به‌ترتیب موجب بهبود پایدار دقت و افزایش ۱/۲ درصدی فراخوانی کلاس‌های اقلیت شد. در مقابل، تغییرات در طول کرنل و پدینگ تأثیر محدودی بر عملکرد نهایی داشت.

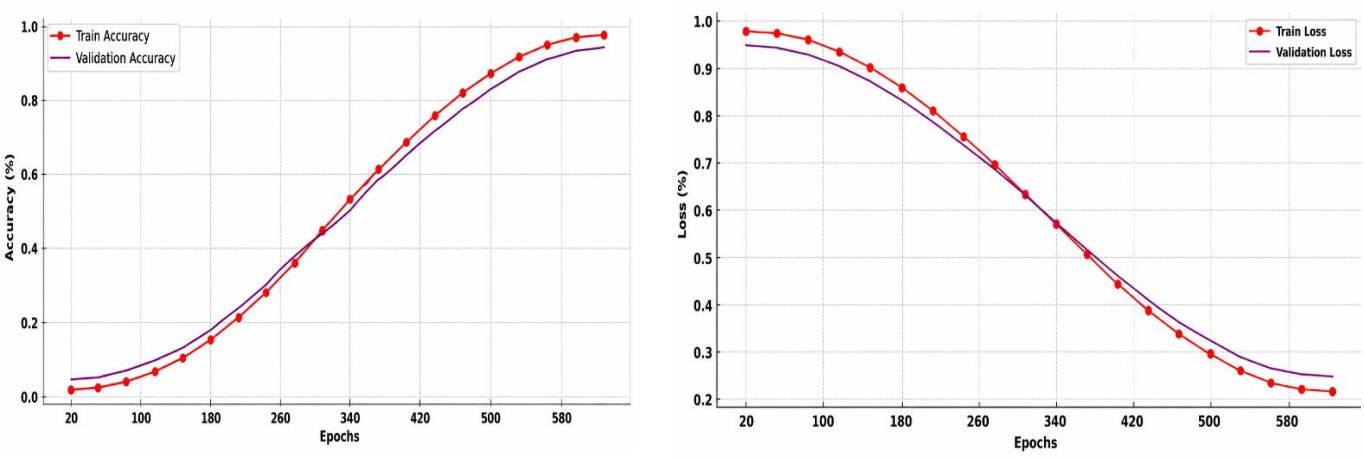
جدول۱- مقادیر پارامترهای مورد استفاده

نام پارامتر	مقدار
پارامترهای XG-BOOST	
تعداد درختان (n-estimators)	۵۰۰
حداکثر عمق درخت	۶
نرخ یادگیری	۰،۰۱
نمونه برداری از ویژگی‌ها	۰،۸
حداکثر تکرار	۱۰۰
پارامترهای ROCKET	
تعداد کرنل	۱۰۰۰۰
حداکثر تکرار	۱۰۰
طول کرنل	۷و ۹و۱۱
Padding (پدینگ)	۰،۵
Dilation (اتساع)	K=[2^a]
Stride (گام)	۱

۴- نتایج آزمایش‌ها

در این مقاله، روش پیشنهادی در محیط متلب پیاده سازی شده و تجزیه و تحلیل نتایج با معیارهای دقت، صحت، بازیابی و معیار F، انجام شده است. در کلیه آزمایش‌ها آموزش و اعتبارسنجی به روش k -Nested 10×5-fold Cross-Validatio انجام شده است. حلقه‌ی بیرونیk=10-fold، داده‌ها به ۱۰ بخش تقسیم شدند. هر بار ۱ بخش به عنوان داده‌ی تست و ۹ بخش به عنوان داده‌ی آموزش استفاده شدند. حلقه‌ی درونی k=5-Fold روی داده‌های آموزش هر تکرار، یک اعتبارسنجی متقابل ۵ بخشی برای تنظیم ابرپارامترهای مدل (تعداد کرنل‌هایROCKET، تعداد درختان و نرخ یادگیری XGBoost، و مقدار K1) انجام گرفته است. نتایج نهایی گزارش‌شده در جدول ۴، میانگین ۱۰ اجرای حلقه‌ی بیرونی هستند. این رویکرد از نشت اطلاعات جلوگیری کرده و تخمینی بدون سوگیری از عملکرد مدل ارائه می‌دهد.

برای اطمینان بیشتر از عدم وجود بیش‌برازش، شکل ۳ (الف)، که آموزش و اعتبارسنجی را نشان می‌دهد رسم شده است. همانطور که در نمودار مشاهده می‌شود، خطای مدل در هر دو مجموعه آموزش و اعتبارسنجی را نشان می‌دهد رسم شده است. همانطور که در نمودار مشاهده می‌شود، خطای مدل در هر دو مجموعه داده‌های آموزش و اعتبارسنجی با گذشت زمان به طور پیوسته کاهش می‌یابد. این نشان می‌دهد که مدل به خوبی آموزش دیده است و هیچ نشانه‌ای از بیش‌برازش مشاهده نمی‌شود، زیرا خطای داده‌های اعتبارسنجی و آموزش به طور مشابه کاهش می‌یابد.



(الف) (ب)

شکل ۳ . (الف) منحنی خطای آموزش و اعتبارسنجی مدل پیشنهادی در طول ۱۰۰ دوره آموزشی. (ب) منحنی دقت آموزش و اعتبارسنجی مدل پیشنهادی در طول ۱۰۰ دوره آموزشی

شکل ۳ (ب)، نمودار دقت مدل پیشنهادی بر روی دو مجموعه داده آموزش و اعتبارسنجی را نشان می‌دهد.

در این نمودار نیز، دقت مدل در هر دو مجموعه داده آموزش و اعتبارسنجی با گذشت زمان افزایش می‌یابد. همانطور که در نمودار مشاهده می‌شود، دقت در داده‌های اعتبارسنجی و آموزش به طور مشابه بهبود یافته و هیچ نشانه‌ای از افت دقت در مجموعه داده‌های اعتبارسنجی مشاهده نمی‌شود. این امر نیز تأکید می‌کند که مدل به درستی یاد می‌گیرد و از بیش‌برازش اجتناب می‌کند، زیرا دقت هر دو مجموعه داده تقریباً مشابه است.جدول ۲، خلاصه مراحل انتخاب ویژگی و کاهش ابعاد داده در شبیه‌سازی روش پیشنهادی را نشان می‌دهد.

جدول ۲. مراحل انتخاب ویژگی و کاهش ابعاد داده در شبیه‌سازی روش پیشنهادی

هدف	روش	تعداد ژن‌ها	مرحله
ویژگی‌های اولیه	-	۲۲،۲۸۴	داده خام
حذف نویز و کاهش ابعاد سریع	اطلاعات متقابل(MI)	K_1=۲۰۰۰	مرحله اول انتخاب
شناسایی ژن‌های مؤثر و قابل اعتماد	XGBoost + انتخاب پایدار	K_2=۱۷۲	مرحله دوم انتخاب
طبقه‌بندی نهایی	-	۱۷۲	ورودی نهایی به ROCKET

۴-۱- پایگاه داده

داده‌های استفاده‌شده از پایگاه داده‌ی عمومی Kaggle گردآوری شده‌اند. این مجموعه داده شامل اندازه‌گیری بیان ۲۲،۲۸۴ ژن در ۱۰ نقطه‌ی زمانی متوالی(T=10) برای هر بیمار است. داده‌ها از ۶۴ بیمار مبتلا به لوسمی (در مجموع ۶۴×۱۰ = ۶۴۰ نمونه) و ۶۰ فرد سالم (۶۰۰ نمونه) جمع‌آوری شده‌اند. پنج زیرگروه لوسمی مورد بررسی عبارتند از: انواع لوسمیAML (لوسمی میلوژنیک حاد)، PB (پلاسماسل)، PBSC_CD34 (سلول‌های بنیادی خونساز)، Bone_Marrow (مغز استخوان) و Bone_Marrow_CD34 (مغز استخوان با نشانگر CD34) است. لوسمی‌لنفوبلاستیک حاد (AML) شایع‌ترین نوع سرطان خون است و تقریباً ۷۵ درصد از سرطان‌های خون را تشکیل می‌دهد. ستون‌های مجموعه داده به‌ترتیب متوالی هستند و بیانگر ژن‌های موجود در مجموعه داده با تغییر شکل دقیق مورد استفاده در زمان می‌باشند [۲۸و۲۹].

ساختار داده به صورت یک ماتریس سه‌بعدی به ابعاد $N \times T \times G$ است که در آن تعداد کل افراد شامل بیمار و سالم۱۲۴= N، تعداد نقاط زمانی اندازه‌گیری ۱۰= T و تعداد ژن‌ها ۲۲۲۸۷= G می‌باشد.

مقادیر بیان ژن بین ۰ و ۱ نرمال‌سازی شده‌اند. برای هر فرد، دنباله‌ی زمانی بیان هر ژن به عنوان یک سری زمانی مستقل در نظر گرفته می‌شود. بدین ترتیب، مجموعه داده نهایی شامل ۲،۷۶۳،۲۱۶ سری زمانی است که هر کدام دارای طول T=۱۰ هستند.

جدول ۳. نمونه‌ای از داده

نمونه	ژن۱	ژن ۲	ژن ۳	...	ژن ۲۲۲۸۴	نوع لوسمی
1	0.23	0.45	0.56	…	0.12	AML

۴-۲- ساختار زمانی داده

داده‌های این مطالعه شامل اندازه‌گیری بیان ۲۲،۲۸۴ ژن در ۱۰ نوبت متوالی (T=10) برای هر فرد است .افراد شامل ۶۴ بیمار مبتلا به لوسمی و ۶۰ فرد سالم هستند. مجموعاً ۱۲۴ فرد در ۱۰ نوبت که ۱۲۴۰ نمونه داریم. فواصل زمانی دقیق بین نوبت‌های اندازه‌گیری در مجموعه داده اصلی مشخص نشده، بنابراین در این مطالعه صرفاً ترتیب زمانی نقاط حفظ شده و داده به عنوان سری زمانی گسسته با گام‌های زمانی واحد فرض شده است.

۴-۳- معیارهای ارزیابی

به‌طورکلی برای بررسی میزان موفقیت و کارایی سیستم¬های دسته‌بندی و تشخیص بیماری‌ها از ماتریس درهم ریختگی استفاده می‌شود. تحلیل ماتریس در هم ریختگی در دسته‌بندی و تشخیص بیماری‌ها منجر به ۴ حالت TP, TN, FP, FN می‌شود از نتایج ماتریس در هم ریختگی سه شاخص دقت، صحت و کارایی به دست می¬آید که برای تحلیل عملکرد سیستم‌های طبقه‌بندی استفاده می‌شود در ادامه شاخص‌ها و متغیرهای مهم در اندازه‌گیری کارایی، شرح داده می‌شود. ماتریس درهم‌ریختگی مدل پیشنهادی در جدول ۴ نشان داده شده است:

جدول ۴. ماتریس درهم‌ریختگی مدل پیشنهادی

پیش‌بینی	AML	PB	PBSC_ CD34	Bone_M arrow	Bone_M arrow_ CD34
AML	208	1	0	1	0
PB	1	138	1	0	0
PBSC_CD3 4	0	0	99	1	0
Bone_Marro w	1	0	1	108	0
Bone_Marro w_CD34	0	0	0	1	79

عملکرد مدل بر اساس معیارهای زیر مورد ارزیابی قرار می‌گیرد:

پاسخ‌های مثبت درست (True Positives | TP): نمونه‌هایی که مثبت پیش‌بینی شده و در حقیقت نیز مثبت هستند.

پاسخ‌های مثبت نادرست (False Positives | FP): نمونه‌هایی که مثبت پیش‌بینی شده اما در حقیقت منفی هستند.

پاسخ‌های منفی درست (True Negatives | TN): نمونه‌هایی که منفی پیش‌بینی شده و در حقیقت نیز منفی هستند.

پاسخ‌های منفی نادرست (False Negatives | FN): نمونه‌هایی که منفی پیش‌بینی شده اما در حقیقت مثبت هستند [۳].

معادلات صحت، دقت و بازیابی بصورت زیر می‌باشد:

$$Accuracy = \frac{TP + TN}{Total\ Examples} \quad (۵)$$

$$Precision = \frac{TP}{TP + FP} \quad (۶)$$

$$Recall = \frac{TP}{TP + FN} \quad (۷)$$

$$F - measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (۸)$$

براساس استفاده از بخش مولد مدل ژن، روش پیشنهادی با سه مجموعه داده مقایسه شده است. مجموعه داده اول DataC1، به صورت عادی بوده ، مجموعه دوم براساس باینری کردن مجموعه DataLC و مجموعه سوم براساس نرمال نمودن وزن‌ها DataC2 است. در شکل ۴، روش پیشنهادی و سایر روش‌ها از نظر معیار F-Measure با یکدیگر مورد مقایسه قرار گرفته‌اند.

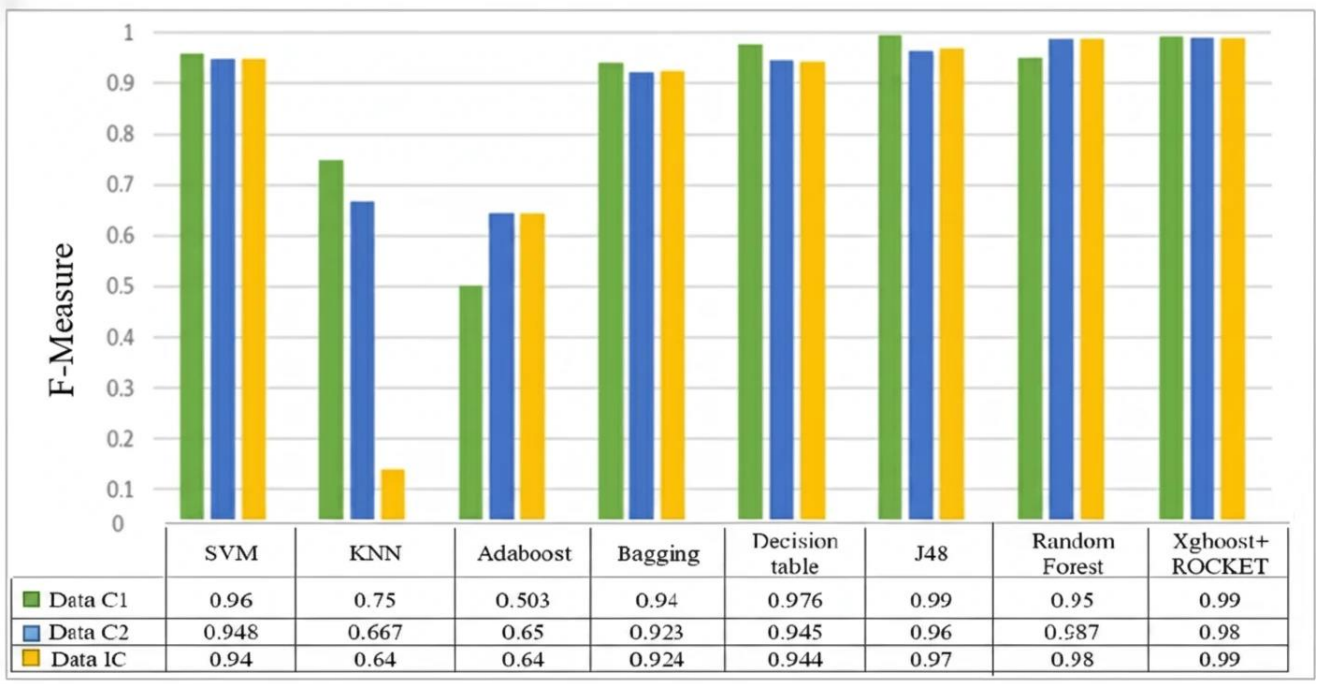
همانگونه که در شکل ۴ نشان داده شده است در معیار F-measure، روش پیشنهادی (XG-Rocket) به نسبت روش j48 و decisionTable به عنوان نزدیکترین رقبا دارای مقدار بالاتری است و دلیل این امر نیز بالا بودن هر دو مقدار Precision و Recall است و این موارد سبب افزایش این معیار شده است. روش‌های j48 و decisionTable دارای مقادیر Recall بسیار کمی بوده که این اختلاف سبب کاهش در معیار F-Measure شده است. پایین‌ترین مقدار در این معیار نیز مربوط به روش KNN است.

روش پیشنهادی در مجموعه داده C1 به ۰,۹۹۳ دست یافت در مقابل J48 مقدار ۰,۹۵ و SVM مقدار ۰,۹۶ را در خروجی نشان داده است. در مجموعه داده C2 روش پیشنهادی با مقدار ۰,۹۸۹ در برابر روش‌های جنگل تصادفی با ۰,۹۸۷ و درخت تصمیم با ۰,۹۴۵ نزدیکترین‌ها به روش پیشنهادی است و در نهایت در مجموعه داده LC روش پیشنهادی با ۰,۹۸۹ در برابر جنگل تصادفی ۰,۹۸۸ و درخت تصمیم با ۰,۹۷ نزدیکترین روش‌ها به روش پیشنهادی است. در این میان KNN با مقادیر ۰,۷۵ و ۰,۶۶۷ و ۰,۱۴ درصد در سه مجموعه داده پایین‌ترین مقادیر را در F-Measure از خود به ثبت رسانده است. این مطالعه نشان می‌دهد که روش‌های آماری قابلیت شناسایی روش‌هایی با متغییر بالا و ویژگی‌های متعدد را ندارند و باید روش‌های اکتشافی برای این منظور مانند روش پیشنهادی ارایه شود.

بطور کلی براساس نمودارهای بالا می‌توان دریافت که در روش XGBoost، از مدل درخت تصمیم سریالی استفاده می‌شود که با یادگیری گرادیانی به بهینه‌سازی عملکرد پرداخته و می‌تواند روابط غیرخطی پیچیده را به خوبی مدل‌سازی کند. الگوریتم گرادیانت بوستینگ با رتبه‌بندی ویژگی‌ها و کاهش پیچیدگی مدل به افزایش دقت و کارایی کمک می‌کند. این روش به دلیل استفاده از درخت‌های تصمیمی که به‌صورت ترتیبی ساخته می‌شوند و یادگیری از خطاهای پیشین، می‌تواند ویژگی‌های مهم را شناسایی کرده و داده‌های پیچیده را بهتر مدل‌سازی کند.

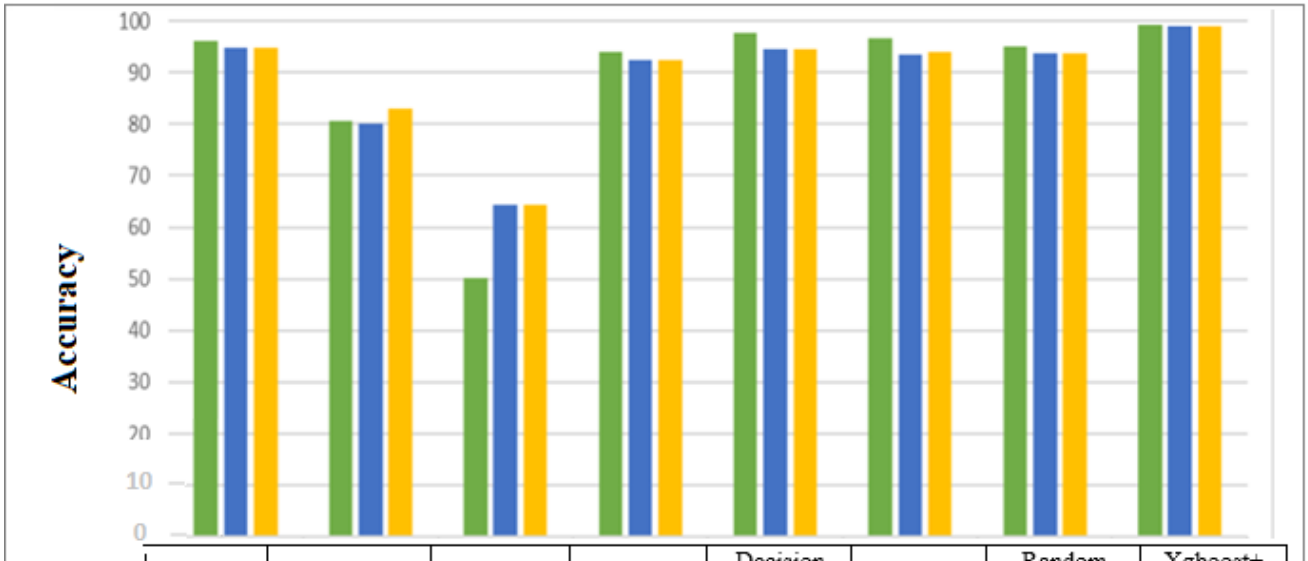
در مقایسه با سایر روش‌ها، XGBoost توانسته است با انتخاب پایدار ویژگی‌ها و رتبه‌بندی ژن‌ها، بهترین نتایج را در داده‌های با ابعاد بالا و نمونه‌های محدود ارائه دهد. این الگوریتم می‌تواند روابط پیچیده بین ویژگی‌ها و کلاس‌ها را به‌طور مؤثر در داده‌های بیولوژیکی که معمولا ویژگی‌های زیادی دارند، شبیه‌سازی کند. بنابراین، عملکرد XGBoostبه دلیل توانایی آن در کاهش احتمال بیش‌برازش و انتخاب ویژگی‌های بهینه، نسبت به بسیاری از الگوریتم‌های دیگر مانند SVM بهتر است. در مقایسه با دیگر روش‌ها، ROCKET بسیار سریع است و نیاز به تنظیمات پیچیده یا زمان‌بر ندارد. این ویژگی به‌ویژه در داده‌های با اندازه نمونه محدود بسیار حائز اهمیت است. به همین دلیل، ROCKET توانسته است عملکرد بهتری نسبت به مدل‌هایی چون CNN یا SVM در داده‌های ژنتیکی که تعداد نمونه‌ها معمولاً محدود است، ارائه دهد. Random Forest معمولاً در داده‌های با ابعاد زیاد از خود عملکرد ضعیفی نشان می‌دهد، زیرا هیچ الگوریتم داخلی برای انتخاب ویژگی‌های مؤثر ندارد و در نتیجه ممکن است عملکرد آن در داده‌هایی که ویژگی‌های زیادی دارند کاهش یابد. این مدل همچنین در مواجهه با داده‌های غیرخطی ممکن است توانایی کمی در شبیه‌سازی الگوهای پیچیده داشته باشد. روش‌های XGBoost و ROCKET به‌طور جداگانه دارای ویژگی‌های منحصر به فرد و مزایای خاص خود هستند. هر یک از این روش‌ها در شرایط خاصی عملکرد بهتری دارند، اما در ترکیب این دو روش، می‌توان از مزایای هر کدام به‌طور همزمان بهره برد.

این ترکیب باعث می‌شود که مدل نه تنها از دقت بالاتری برخوردار باشد، بلکه از سریع‌ترین و مؤثرترین روش‌های یادگیری برای داده‌های پیچیده و ابعاد بالا استفاده کند. بنابراین، این دو الگوریتم می‌توانند نقاط ضعف یکدیگر را جبران کنند و به یک مدل جامع و بهینه دست یابند. در شکل ۵ روش‌های مختلف از نظر معیار دقت با یکدیگر مقایسه شده‌اند. همانگونه که در شکل ۵ نشان داده شده است روش پیشنهادی نسبت به کلیه روش‌های مورد مقایسه دارای دقت بالاتری بوده است. روش پیشنهادی در مجموعه‌داده C1 به ۹۹,۲۶ درصد صحت دست یافت در مقابل درخت تصمیم مقدار ۹۷,۶۱ و بردار پشنیبان تصمیم مقدار ۹۶ را در خروجی به عنوان دقت نشان داده است.



	SVM	KNN	Adaboost	Bagging	Decision table	J48	Random Forest	Xghoost+ ROCKET
Data C1	0.96	0.75	0.503	0.94	0.976	0.99	0.95	0.99
Data C2	0.948	0.667	0.65	0.923	0.945	0.96	0.987	0.98
Data IC	0.94	0.64	0.64	0.924	0.944	0.97	0.98	0.99

شکل ۴. ارزیابی روش پیشنهادی و سایر روش ها از نظر معیار F-Measure



	SVM	KNN	Adaboost	Bagging	Decision table	J48	Random Forest	Xgboost+ ROCKET
Data C1	96.2	80.52	50.27	94.1	97.61	96.54	95.3	99.26
Data C2	0.948	80.18	64.43	92.34	94.48	93.50	93.71	98.89
Data IC	0.94	82.94	64.44	92.43	94.39	93.96	93.80	98.89

شکل ۵. ارزیابی روش پیشنهادی و سایر روش ها از نظر معیار صحت

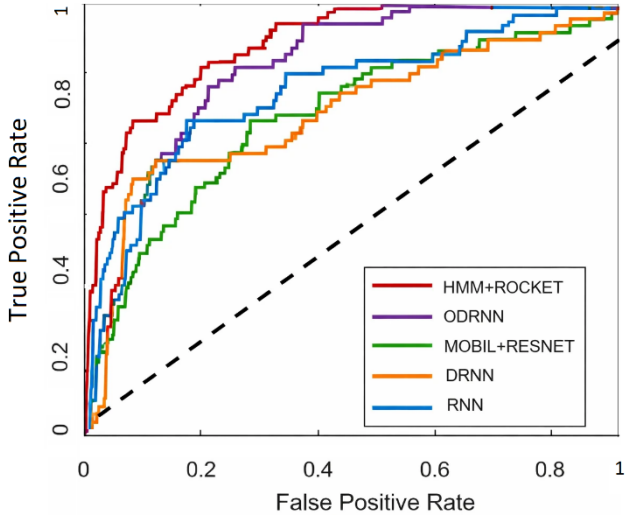
در مجموعه داده C2 روش پیشنهادی با مقدار ۹۸,۸۹ درصد دقت در برابر روش‌های بردار پشتیبان تصمیم با ۹۴,۸۰ و درخت تصمیم با ۹۴,۴۸ نزدیکترین‌ها به روش پیشنهادی است و در نهایت در مجموعه داده LC روش پیشنهادی با ۹۸,۸۹ درصد دقت در برابر بردار پشتیبان تصمیم ۹۴,۸۰ و درخت تصمیم با ۹۴,۳۹

نزدیکترین روش‌ها به روش پیشنهادی است. در این میان KNN و Adaboost در سه مجموعه داده پایین‌ترین مقادیر را در معیار دقت از خود به ثبت رسانده است. جدول ۵، کارایی روش پیشنهادی در مجموعه داده را نشان می‌دهد. همانگونه که در جدول ۴ نشان داده شده است Bone_Marrow کمترین تشخیص و PBSC_CD34 بیشترین تشخیص را در بین لوسومی‌ها داشته است.

در ادامه روش پیشنهادی با چندین روش ذکر شده مقایسه شده است و نتایج آن در جدول ۶ آورده شده است.

برای نشان دادن تفاوت‌های معتبر، نتایج به‌دست‌آمده باید با استفاده از آزمون‌های آماری تحلیل شوند. در این مقاله، مقایسه‌ها با استفاده از آزمون‌های دو طرفه (t-test) با ۹۵ درصد فاصله اطمینان انجام شده است. مقادیر احتمال (p) کمتر از ۰,۰۵ از نظر آماری معنادار در نظر گرفته شده‌اند. SD انحراف معیار است.

براساس جدول ۶، روش پیشنهادی به نسبت سایر روش‌های همسان دارای برتری مناسبی بوده است و از نزدیکترین روش خود بهبود حدود ۲ درصدی را به ثبت رسانده است. براساس این مقایسه روش CNN [۸] با مقدار دقت ۹۲,۴۲ درصد، کمترین مقدار دقت را از خود نشان داده است و روش ODRNN [۳۴] مقدار ۹۸,۹۷ درصد، در رتبه دوم دقت، پس از روش پیشنهادی با دقت ۹۹,۲۶ درصد است. در شکل ۶ نمودار ROC روش‌های مختلف رسم شده است. همانطور که مشاهده می‌شود عملکرد روش پیشنهادی (AUC=Area Under Curve) بهتر از دیگر روش‌هاست.



شکل ۶. نمودار ROC تعدادی از روش‌ها

همانطور که مشاهده می‌شود عملکرد روش پیشنهادی (AUC=Area Under Curve) بهتر از دیگر روش‌هاست.

جدول۵. بررسی کارایی روش پیشنهادی برای تشخیص انواع لوسمی

Accuracy	Precision	Recall	تشخیص صحیح (TP)	تعداد	لوسمی
99.03	99.02	99.09	208	210	AML
98.99	99.01	99.07	138	140	PB
99.17	99.11	99.21	99	100	PBSC_CD34
98.18	98.26	98.36	108	110	Bone_Marrow
99.05	99.07	99.06	79	80	Bone_Marrow_CD34

جدول ۶. مقایسه روش پیشنهادی با روش‌های همسان تشخیص سرطان

p-value	F measure (mean±SD)	Recall (mean±SD)	Precision (mean±SD)	Accuracy (mean±SD)	روش‌ها
0.02	93.32±0.011	92.21±0.012	92.56±0.02	92.41±0.012	ALNett [33]
0.04	94.65±0.021	93.45±0.011	93.65±0.00	93.50±0.020	CNN [31]
0.04	93.31±0.03	92.20±0.001	92.73±0.00	91.44±0.011	LSTM
0.02	92.85	92.05±0.012	91.92±0.01	92.28±0.04	GRU
0.02	95.45±0.030	95.62±0.04	94.95±0.01	94.61±0.011	DRNN[6]
0.01	97.98±0.002	97.85±0.013	97.23±0.013	97.98±0.003	ODRNN [6]
0.04	97.14±0.01	97.05±0.014	96.94±0.012	97.09±0.005	Mobilenet V2+Resnet 18[32]
0.05	91.14±0.05	92.05±0. 14	92.94±0.022	91.07±0.010	XG-Boost
0.05	92.98±0.022	93.85±0.014	93.33±0.031	92.98±0.014	ROCKET
0.02	99.2±0.011	98.9±0.012	99.31±0.001	99.26±0.011	Proposed Model

۵- نتیجه گیری

سرطان یکی از مهم‌ترین علت‌های مرگ و میر در جهان است و تشخیص زودهنگام آن می‌تواند باعث درمان سریع‌تر و افزایش شانس بهبودی بیمار شود. یکی از ابزارهای مؤثر در تشخیص سرطان استفاده از داده‌های ریزآرایه است که برخلاف روش‌های تصویربرداری، فاقد اشعه‌های مضر برای انسان هستند. ریزآرایه‌ها شامل هزاران ژن هستند که تحلیل آن‌ها می‌تواند بسیار پیچیده و زمان‌بر باشد. انتخاب ژن‌های مؤثر در تشخیص سرطان به‌عنوان یکی از گام‌های اساسی و چالش‌برانگیز در تحلیل داده‌های ریزآرایه مطرح است. در این مقاله، هدف اصلی تشخیص انواع لوسمی میلوژنیک و لنفوسیتیک حاد با استفاده از انتخاب ژن‌های مؤثر از داده‌های ریزآرایه بوده است. برای این منظور، یک روش دو مرحله‌ای انتخاب ویژگی پیشنهاد شد. در مرحله اول، ویژگی‌ها با استفاده از معیار اطلاعات متقابل انتخاب شدند تا ژن‌های مؤثرتر در ارتباط با انواع لوسمی شناسایی گردند. این مرحله به‌طور خودکار ویژگی‌های غیرمؤثر و نویزی را حذف کرده و تعداد ویژگی‌ها را کاهش می‌دهد. در مرحله دوم، از الگوریتم XGBoost برای رتبه‌بندی جاسازشی و انتخاب ویژگی‌های پایدار استفاده شد. سپس ویژگی‌های منتخب به طبقه‌بند ROCKET داده شدند. ROCKET با استفاده از کرنل‌های پیچشی تصادفی و استخراج الگوهای پیچیده در داده‌های ورودی، بهترین عملکرد را در داده‌های با ویژگی‌های زیاد و نمونه‌های محدود نشان داد.

روش پیشنهادی در آزمایش‌ها دقت ۹۹,۲۶ درصد را ثبت کرد و نتایج نشان داد که مدل پیشنهادی در مقایسه با روش‌های هم‌سان، دقت و کارایی بالاتری دارد و توانسته است به‌طور مؤثری انواع لوسمی را شناسایی کند. با این حال، بررسی دقیق‌تر عملکرد روش‌ها و مقایسه آن‌ها با توجه به چالش‌ها و معایب موجود، ضروری به نظر می‌رسد.

الگوریتم XGBoost به دلیل توانایی در مدل‌سازی روابط غیرخطی پیچیده و انتخاب ویژگی‌های مؤثر عملکرد خوبی از خود نشان داده است. اما نیاز به تنظیمات پیچیده دارد که در داده‌های با ابعاد بسیار زیاد ممکن است منجر به زمان محاسباتی طولانی شود. ROCKET، قادر است الگوهای پیچیده را به‌طور سریع و مؤثر از داده‌های سری زمانی استخراج کند. با این حال، عدم وجود تنظیمات دقیق و وابستگی به کرنل‌های تصادفی می‌تواند در برخی داده‌ها منجر به کاهش دقت و عملکرد ضعیف در داده‌های با کیفیت پایین یا نویز زیاد شود. در نهایت، روش پیشنهادی با ترکیب XGBoost و ROCKETتوانسته است عملکرد بهتری در شبیه‌سازی روابط پیچیده در داده‌های ریزآرایه و تشخیص انواع لوسمی نسبت به سایر روش‌ها ارائه دهد.

اگرچه روش پیشنهادی در این مطالعه عملکرد مطلوبی داشته است، اما هنوز محدودیت‌هایی در آن وجود دارد. روش پیشنهادی با استفاده از دو الگوریتم مختلف XGBoost و ROCKET نیاز به زمان محاسباتی بیشتری دارد به‌ویژه در داده‌های با ابعاد بالا و نمونه‌های زیاد. این مسئله ممکن است در داده‌های پیچیده‌تر به عنوان یک چالش مطرح شود.

برای بهبود و گسترش این تحقیق، چندین مسیر برای کارهای آینده پیشنهاد می‌شود:

۱. کاهش زمان محاسباتی: بهینه‌سازی الگوریتم‌های موجود و استفاده از مدل‌های توزیع‌شده یا پردازش موازی می‌تواند به کاهش زمان محاسباتی و بهبود سرعت مدل‌ها کمک کند.

۲. انتخاب ویژگی بهینه: استفاده از روش‌های انتخاب ویژگی خودکار و کاهش ابعاد داده‌ها می‌تواند دقت مدل را در داده‌های پیچیده‌تر و نویزی بهبود بخشد.

۳. مدل‌های مقاوم به نویز: به‌ویژه برای داده‌های ژنتیکی که ممکن است شامل نویز باشند، توسعه الگوریتم‌های مقاوم به نویز می‌تواند به بهبود دقت و کارایی کمک کند.

مراجع

[1] S. M. Melana, L. Nepomnaschy, J. Hasa, A. Djougarian, J. F. Holland, and B. G. Pogo, “Detection of Human Mammary Tumor Virus Proteins in Human Breast Cancer Cells,” Journal of Virological Methods, Vol. 163, No. 1, pp. 157–161, 2010. DOI: 10.1016/j.jviromet.2009.09.012.

[2] P. Sharma, P. Sharan, and P. Deshmukh, “Photonic Crystal Sensor for Analysis and Detection of Cancer Cells,” in International Conference on Pervasive Computing (ICPC), pp. 1–5, IEEE, Jan. 2015.

[3] M. L. Leung, A. Davis, R. Gao, A. Casasent, Y. Wang, E. Sei, et al., “Single-Cell DNA Sequencing Reveals a Late-Dissemination Model in Metastatic Colorectal Cancer,” Genome Research, Vol. 27, No. 8, pp. 1287–1299, 2017. DOI: 10.1101/gr.209973.116.

[4] M. L. Shehab, Q. Abualigah, M. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. Alsalibi, and A. H. Gandomi, “Machine Learning in Medical Applications: A Review of State-of-the-Art Methods,” Computers in Biology and Medicine, Vol. 145, Art. 105458, 2022. DOI: 10.1016/j.compbiomed.2022.105458.

[5] G. Kumar and H. Alqahtani, “Deep Learning-Based Cancer Detection—Recent Developments, Trend and Challenges,” CMES—Computer Modeling in Engineering & Sciences, Vol. 130, No. 3, 2022. DOI: 10.32604/cmes.2022.019989.

[6] K. D. Shree and S. Logeswari, “ODRNN: Optimized Deep Recurrent Neural Networks for Automatic Detection of

Cancer Nursing, Vol. 46, No. 6, pp. E384–E393, 2023. DOI: 10.1177/08445621221144594

[20] K. R. Haase, S. Sattar, S. Dhillon, H. M. Kilgour, J. Pesut, D. Howell, and J. L. Oliffe, "A Survey of Older Adults' Self-Managing Cancer," Current Oncology, Vol. 29, No. 11, pp. 8019–DOI: 8030, 2022. 10.3390/curroncol291108019.

[21] P. Dixon, R. M. Martin, and S. Harrison, "Using Mendelian Randomization to Model the Causal Effect of Cancer on Health Economic Outcomes and to Simulate the Cost-Effectiveness of Anti-Cancer Interventions," medRxiv, 2023. DOI: 10.1101/2023.05.03.23289471.

[22] H. Patadia, A. Priyadarshini, and A. Gangawane, "Integrated Proteomic, Transcriptomic, and Genomic Analysis Identifies Fibrinogen Beta and Fibrinogen Gamma as Key Modulators of Breast Cancer Progression and Metastasis," Biomedical and Biotechnology Research Journal (BBRJ), Vol. 6, No. 2, pp. 266–277, 2022. DOI: 10.2147/BBRJ.S353435.

[23] M. A. Berbís, D. S. McClintock, A. Bychkov, J. Van der Laak, L. Pantanowitz, J. Lennerz, and J. Shen, "Computational Pathology in 2030: A Delphi Study Forecasting the Role of AI in Pathology Within the Next Decade," EBioMedicine, Vol. 88, 2023. DOI: 10.1016/j.ebiom.2023.104863.

[24] M. Momenzadeh, M. Sehhati, and H. Rabbani, "A Novel Feature Selection Method for Microarray Data Classification Based on Hidden Markov Model," Journal of Biomedical Informatics, Vol. 95, Art. 103213, 2019. DOI: 10.1016/j.jbi.2019.103213.

[25] A. Dempster, D. F. Schmidt, and G. I. Webb, "MINIH: A Very Fast (Almost) Deterministic Transform for Time Series Classification," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 248–257, Aug. 2021. DOI: 10.1145/3447548.3467201.

[26] J. Wang, "An interpretable integrated machine learning framework for genomic data," Journal of Biomedical Informatics, 2025.

[27] H. Salehinejad, Y. Wang, Y. Yu, T. Jin, and S. Valaee, "S-ROCKET: Selective Random Convolution Kernels for Time Series Classification," arXiv preprint arXiv:2203.03445, 2022. DOI: 10.48550/arXiv.2203.03445.

[28] N. Papaioannou, G. Myllis, A. Tsimpiris, and V. Vrana, "The Role of Mutual Information Estimator Choice in Feature Selection: An Empirical Study on mRMR," *Information*, Vol. 16, Art. 724, 2025. DOI: 10.3390/info16090724.

[29] B. I. Grisci, B. C. Feltes, and M. Dorn, "Neuroevolution as a Tool for Microarray Gene Expression Pattern Identification in Cancer Research," Journal of Biomedical Informatics, Vol. 89, pp. 122–133, 2019. DOI: 10.1016/j.jbi.2018.11.013.

[30] S. Z. Aftabi, A. Ahmadi, and S. Farzi, "Fraud Detection in Financial Statements Using Data Mining and GAN Models," Expert Systems with Applications, Vol. 227, Art. 120144, 2023. DOI: 10.1016/j.eswa.2022.120144.

[31] M. A. Hossain, A. M. Islam, S. Islam, S. Shatabda, and A. Ahmed, "Symptom-Based Explainable Artificial Intelligence Model for Leukemia Detection," IEEE Access, Vol. 10, pp. 57283–57298, 2022. DOI: 10.1109/ACCESS.2022.3178440.

[32] P. K. Das and S. Meher, "An Efficient Deep Convolutional Neural Network-Based Detection and Classification of Acute Lymphoblastic Leukemia," Expert Systems with Applications, Vol. 183, Art. 115311, 2021. DOI: 10.1016/j.eswa.2021.115311.

[33] M. Jawahar, H. Sharen, and A. H. Gandomi, "ALNett: A Cluster Layer Deep Convolutional Neural Network for Acute Leukaemia," Signal, Image and Video Processing, pp. 1–17, 2024. DOI: 10.1007/s11760-024-03010-5.

[7] M. Shahin, F. Chen, F. A. Hosseinzadeh, and M. Maghanaki, "Deploying Deep Convolutional Neural Network to the Battle Against Cancer: Towards Flexible Healthcare Systems," Informatics in Medicine Unlocked, Vol. 47, Art. 101494, 2024. DOI:10.1016/j.imu.2024.101494.

[8] U. Saeed, K. Kumar, M. A. Khuhro, A. A. Laghari, A. A. Shaikh, and A. Rai, "DeepLeukNet—A CNN-Based Microscopy Adaptation Model for Acute Lymphoblastic Leukemia Classification," Multimedia Tools and Applications, Vol. 83, No. 7, pp. 21019–21043, 2024. DOI: 10.1007/s11042-023-16123-8.

[9] A. Batool and Y. C. Byun, "Lightweight EfficientNet-B3 Model Based on Depthwise Separable Convolutions for Enhancing Classification of Leukemia White Blood Cell Images," IEEE Access, 2023. DOI: 10.1109/ACCESS.2023.3241568.

[10] H. Xu, J. Jia, H. H. Jeong, and Z. Zhao, "Deep Learning for Detecting and Elucidating Human T-Cell Leukemia Virus Type 1 Integration in the Human Genome," Patterns, Vol. 4, No. 2, 2023. DOI: 10.1016/j.patter.2023.100742.

[11] S. A. Patel, "Functional Genomic Approaches in Acute Myeloid Leukemia: Insights Into Disease Models and the Therapeutic Potential of Reprogramming," Cancer Letters, Vol. 533, Art. DOI: 215579, 2022. DOI: 10.1016/j.canlet.2022.215579.

[12] S. A. Patel, D. Dalela, A. C. Fan, M. R. Lloyd, and T. Y. Zhang, "Niche-Directed Therapy in Acute Myeloid Leukemia: Optimization of Stem Cell Competition for Niche Occupancy," Leukemia & Lymphoma, Vol. 63, No. 1, pp. 10–18, 2022. DOI: 10.1080/10428194.2022.2029247.

[13] A. Damnernsawad, D. Bottomly, S. E. Kurtz, C. A. Eide, S. K. McWeeney, J. W. Tyner, and T. Nechiporuk, "A Genome-Wide CRISPR Screen Identifies Regulators of MAPK and mTOR Pathways That Mediate Resistance to Sorafenib in Acute Myeloid Leukemia," Haematologica, Vol. 107, No. 1, p. 77, 2022. DOI: 10.1186/s13045-022-01250-8.

[14] I. A. Basurto-Hurtado, M. Cruz-Albarran, M. Toledano-Ayala, M. A. Ibarra-Manzano, L. A. Morales-Hernandez, and C. A. Perez-Ramirez, "Diagnostic Strategies for Breast Cancer Detection: From Image Generation to Classification Using Artificial Intelligence Algorithms," Cancers, Vol. 14, No. 14, Art. 3442, 2022. DOI: 10.3390/cancers14143442.

[15] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer Statistics," CA: A Cancer Journal for Clinicians, Vol. 72, No. 1, 2022. DOI: 10.3322/caac.21708.

[16] A. Slomski, "Colonoscopy Did Not Reduce Cancer Deaths in Trial," JAMA, Vol. 328, No. 20, pp. 2003–2004, 2022. DOI: 10.1001/jama.2022.20402.

[17] M. H. Antoni, P. I. Moreno, and F. J. Penedo, "Stress Management Interventions to Facilitate Psychological and Physiological Adaptation and Optimal Health Outcomes in Cancer Patients and Survivors," Annual Review of Psychology, Vol. 74, pp. 423–455, 2023. DOI: 10.1146/annurev-psych-032420-031042.

[18] X. Zheng, X. Wang, Y. He, and H. Ge, "Systematic Analysis of Expression Profiles of HMGB Family Members for Prognostic Application in Non-Small Cell Lung Cancer," Frontiers in Molecular Biosciences, Vol. 9, Art. 844618, 2022. DOI: 10.3389/fmolb.2022.844618.

[19] M. E. Francis, P. Mohindra, and K. Mooney-Doyle, "Exploring Dyad-Based Communication During Cancer: A Pilot Study,"

Lymphoblastic Leukemia Classification," Computers in Biology and Medicine, Vol. 148, Art. 105894, 2022. DOI: 10.1016/j.compbiomed.2022.105894.

پاورقی‌ها:

[1] Mutual Information
[2] Embedded Ranking
[3] Random Convolutional Kernel Transform – ROCKET
[4] ODRNN
[5] DRNN
[6] HMM
[7] Random Convolutional Kernel Transform
[8] Dilation
[9] Padding
[10] Stride
[11] Ridge regression
[12] Grid Search