سال دوازدهم، شماره‌های ۴۳ و ۴۴، بهار و تابستان ۱۳۹۹
صفحات: ۱۵۵_۱۶۸

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

تبدیل توالی پروتئین به تصویر جهت طبقه‌بندی با شبکه عصبی کانولوشنی

رضا احسن [*] منصور ابراهیمی [**] روح الله دیانت [*]

[*] استادیار دانشکده فنی مهندسی - دانشگاه قم - قم – ایران
[**] دانشیار دانشکده علوم پایه - دانشگاه قم - قم - ایران

تاریخ دریافت: ۱۳۹۸/۱۱/۱۴ تاریخ پذیرش: ۱۳۹۹/۰۸/۱۸
نوع مقاله: پژوهشی

چکیده

از آنجا که روش‌های مخصوص طبقه‌بندی توالی یادگیری ماشین، جهت طبقه‌بندی پروتئین‌های سالم و سرطانی موفق نبودند بنابراین یافتن راهکاری برای بازنمایی این توالی‌ها جهت طبقه بندی افراد سالم و مریض با رویکردهای یادگیری عمیق ضرورت تام دارد. در این مطالعه، روش‌های مختلف بازنمایی توالی پروتئین، جهت طبقه‌بندی توالی پروتئین افراد سالم و سرطان خون، مورد بررسی قرار گرفته است. نتایج نشان داد که تبدیل حروف اسید آمینه به بردار ویژگی یک‌بعدی در طبقه بندی ۲ کلاس موفق نبود و فقط یک کلاس مریض تشخیص داده شد. با تغییر بردار ویژگی به‌صورت اعداد رنگی دقت تشخیص کلاس سالم کمی بهبود یافت. روش بازنمایی توالی پروتئینی به‌صورت یکپارچه دودویی، با ابتکار حفظ دنباله توالی در دو حالت یک‌بعدی و دوبعدی(تصویر با اعمال فیلتر گابور)، نسبت به روشهای قبلی موثرتر بود. بازنمایی توالی پروتئین به شکل تصویر دودویی با اعمال فیلتر گابور با دقت ٪۱۰۰ توالی پروتئین افراد سالم و ٪۹۸,۶ توالی پروتئین افراد دارای سرطان خون را طبقه‌بندی کرد. یافته‌های این تحقیق نشان داد که بازنمایی توالی پروتئین به شکل تصویر دودویی با اعمال فیلتر گابور، می‌تواند به‌عنوان روش موثر جدید دربازنمایی توالی‌های پروتئینی جهت طبقه‌بندی، ارایه نماید.

واژگان کلیدی: تبدیل توالی پروتئین به تصویر، فیلتر گابور، شبکه عصبی کانولوشنی، طبقه‌بندی توالی پروتئین.

۱- مقدمه

ساختار ژنومی و پروتئینی هر جاندار توسط یک سری از توالی‌های خاص تکرارای درست شده است. چهار نوکلئوتید آدنین، گوانین، تیمین، و سیتوزین ساختار اصلی توالی‌های ژنومی و بیست اسید آمینه ساختار توالی پروتئین را تشکیل می‌دهند. به‌دلیل تکراری بودن این توالی‌ها امکان استفاده از مدل‌های مختلف داده کاوی در تحلیل ژنومی فراهم شده است[۱]. در چند دهه اخیر، پیشرفت در زیست‌شناسی مولکولی و تجهیزات مورد نیاز تحقیق در این زمینه باعث افزایش سریع تعیین توالی ژنوم بسیاری از گونه‌های موجودات شد، به‌طوری‌که پروژه‌های تعیین توالی ژنوم‌ها از پروژه‌های بسیار رایج به‌حساب می‌آیند. مطالعه وابستگی این پروفایل‌های توالی پروتئینی و حالت‌های بیماری یا مراحل سلول‌ها نقش مهمی را در کاربردهای بالینی و بیولوژیکی بازی می‌کند [۲]. پروفایل‌های توالی پروتئینی می‌تواند از چندین نمونه بافت به‌دست آورده شود و در مقایسه ژن‌های بیان شده در بافت نرمال با آن ژن‌هایی که در بافت بیمار است، فرد می‌تواند به بصیرت بهتری در

نویسنده مسئول: منصور ابراهیمی mansour@future.edu

طب آسیب شناسی بیماری برسد [۳]. یکی از بیماری‌های مهم در این زمینه سرطان می‌باشد.

سرطان در واقع یک بیماری ژنتیکی است که مشخصه آن جهش (تغییر در توالی ژنوم) در بخشی از DNA در یک یا چند گروه از سلول‌های طبیعی می‌باشد که منجر به تقسیم نامحدود این سلول‌ها می‌گردد [۴]. سرطان خون یا لوسمی بیماری پیش‌رونده و بدخیم اعضای خون‌ساز بدن است. این بیماری در اثر تکثیر و تکامل ناقص گویچه‌های سفید خون و پیش‌سازهای آن در خون و مغز استخوان ایجاد می‌شود. به این معنی که مغز استخوان به‌صورت غیر عادی، مقدار بسیار زیادی سلول خونی تولید می‌کند که باعث توقف در تولید سلول‌های سفید می‌شوند و توانایی فرد در مقابله با بیماری‌ها از بین می‌رود. این سلول‌ها که با سلول‌های خون نرمال متفاوت هستند بر تولید سایر انواع سلول‌های خونی که توسط مغز استخوان ساخته می‌شود مانند گویچه‌های قرمز خون که اکسیژن به بافت‌های بدن می‌رسانند و پلاکت‌های خونی که از لخته شدن جلوگیری می‌کنند، اثر می‌گذارند [۵]. پیشرفت‌های فن آوری در علم ژنتیک و تصویر برداری یک انفجاری در ایجاد حجم زیادی از نمونه‌های مولکولی و سلولی کرده است. تحلیل و بررسی این حجم زیاد از نمونه‌های مولکولی و سلولی با روشهای متعارف با توجه به ابعاد بالای داده‌های بیولوژیکی چالش برانگیز است [۶]. روشهای مدرن یادگیری ماشین، از قبیل یادگیری عمیق، نویدی برای قدرت نفوذ به ساختار مخفی بین مجموعه داده‌های بسیار بزرگ و ساخت پیش بینی‌های دقیق دارد. ارزش شبکه عصبی عمیق در این زمینه دو جنبه است. ابتدا، شیوه‌های قدیمی یادگیری ماشین نمی‌تواند مستقیما روی توالی اجرا شود، بنابراین نیازمند ویژگی‌های از پیش تعریف شده دارد که می‌تواند بر اساس دانش قبلی استخراج شود مانند حضور یا عدم حضور متغیرهای تک نوکلئوتیدی[1]؛ تعداد دفعاتی که زیر توالی[2] ظاهر شده؛ توالی‌های تکراری کوتاه نوکلئوتیدی یا آمینواسیدی[3]؛ و دنباله‌های حفظ شده که در نسلهای مختلف تغییر نکرده است.[4] شبکه‌های عصبی عمیق به‌صورت خودکار نه دستی، الگوهای مشترک از داده‌ها را از طریق یادگیری ویژگی پیدا می‌کنند. بدین معنی که بازنمایی غنی شده‌ای از داده‌های توالی ایجاد می‌کند تا بتواند وابستگی‌های غیر خطی و اثرات متقابل آنها را در محدوده توالی گسترده‌تر در مقیاس ژنومی متعدد را نشان دهد و این بازنمایی در روند دسته‌بندی، بین دسته‌ها یا کلاس‌های مختلف از توالی، بهتر تمایز قائل می‌شود [۷]. کاربرد مفید دیگر یادگیری عمیق موفقیت در پیدا کردن و وصل کردن بخشهای اطلاعاتی کدکننده ژن یعنی اگزون‌ها می‌باشد[5][۸, ۹]; یادگیری عمیق در تشخیص ویژگی‌های پروتئین‌های متصل‌شونده به DNA و RNA [۱۰]; در تشخیص ویژگی‌های علائم اپی ژنتیک که مطالعه بر روی تاثیرات محیط برای باز شدن رشته‌های DNA یا RNA پیچ خورده برای بیان شدن را دارند[۱۱]; و همچنین به منظور مطالعه کشف اثرات تغییر رشته‌های DNA [۱۲]، موفقیت آمیز عمل کرده است. یکی از بهترین و دقیق‌ترین روش‌های یادگیری عمیق در این زمینه، شبکه عصبی کانولوشنی است، در شبکه عصبی کانولوشنی استخراج ویژگی به‌صورت سلسله مرتبی انجام می‌شود [۱۳].

۲- ادبیات تحقیق

سال‌های اخیر انفجاری در پیشرفت تکنیک‌هایی با تکنولوژی بالا برای دستیابی و نشان دادن جنبه‌های مختلف فعالیت ژن به وجود آمده است. اکنون با استفاده از این تکنولوژی‌های جدید، شناسایی ارتباطات جدید بین ژن‌ها را با قدرت تفکیک پذیری بالاتر نسبت به گذشته ممکن ساخته است. برای مثال خیلی زود این امکان وجود خواهد داشت که نقشه کل مجموعه کنش متقابل پروتئین برای هر ارگانیسم نیز مشخص شود. دسترسی این مجموعه داده وسیع ژنوم یک فرصت بی نظیر برای کشف ویژگی‌های سلولی جدید از منظر سیستمی می‌دهد و توانایی دانشمندان را در پیش‌بینی صحیح عملکرد ژن در حجم‌های وسیع افزایش می‌دهد [۱۴]. شماری از تکنیک‌های طبقه‌بندی یادگیری ماشین جهت طبقه‌بندی بافت به دو نوع سرطانی و نرمال استفاده شده است. بعلاوه تعداد زیاد ویژگی در مقابل تعداد کم نمونه‌های آموزشی، حل این مساله را خیلی سخت‌تر می‌کند [۱۵]. در گذشته، راه‌حل‌های زیادی جهت مساله طبقه‌بندی سرطان پیشنهاد شده است. در این روش‌ها، بیشتر از کاهش فضای مشخصه با انتخاب و یا استخراج ویژگی استفاده شده است. اگرچه این منجر به مشکلاتی با آن روش‌هایی می‌شود که اکثرا مقیاس پذیر نیستند و نمی‌توانند به انواع جدید سرطان بدون باز طراحی مشخصات جدید تعمیم داده شوند. بعلاوه این تکنیک‌ها نمی‌توانند راه‌حل موثری از نمونه‌های بافت از دیگر سرطان‌ها اتخاذ کنند [۱۶]. یادگیری

---

[1] SNVs
[2] K-mer
[3] Motif
[4] conservation
[5] splicing

عمیق در حال حاضر با تحلیل ژن‌های بیمار برای کمک به تشخیص بیماری‌ها مورد استفاده است. این تکنیک می‌تواند سلول‌های سرطانی را تشخیص دهد که دانشمندان موفق به مشاهده آن‌ها نشده‌اند همچنین می‌تواند به محققان در درک بهتر جهش‌های عامل سرطان و توسعه درمان‌های جدید برای آن‌ها کمک کند [۱۷]. یادگیری عمیق از زیر شاخه‌های یادگیری ماشین است. این روش ویژگی را به‌صورت سلسله مراتبی از لایه‌های مختلف از طریق توابع غیر خطی استخراج می‌کند ورودی هر لایه خروجی لایه قبلی است و آموزش آن می‌تواند به‌صورت با ناظر یا بدون ناظر باشد.در واقع تک لایه مخفی در شبکه عصبی با تعدادی زیادی(عمیق) لایه جایگزین شده است [۱۸]. شبکه‌های عصبی کانولوشنی یکی از مهم‌ترین روش‌های یادگیری عمیق هستند که در آنها چندین لایه با روشی قدرتمند آموزش می‌بینند. این روش بسیار کارآمد بوده و یکی از رایج‌ترین روشها در کاربردهای مختلف بینایی کامپیوتر است [۱۹]. پروتئین‌ها می‌توانند فعال کننده یا مهار کننده بیماری باشند علاوه بر نقش خود به عنوان یک عامل تمایز، سرکوبگر تومور نیز می‌باشند. در حدود ۹۰٪ از سلول‌های سرطانی دارای فعالیت بالای ترکیبات نوکلئوپروتئینی می‌باشند که سبب می‌شود سلول‌ها رشد غیرعادی داشته باشند [۲۰]. بنابراین بررسی فعالیت توالی پروتئینی سلول‌های سرطانی، می‌تواند به عنوان ابزاری برای تشخیص و طبقه‌بندی بیماری سرطان مورد استفاده قرار گیرد. در ارتباط با این مساله و جهت تسهیل و توسعه نسخه‌های تعمیم یافته‌تر دسته کننده‌های سرطان، در این تحقیق، ما راه کلی‌تری از یادگیری مشخصه‌ها به وسیله کاربرد یادگیری مشخصه با ناظر و روش‌های یادگیری عمیق ، در واقع شبکه کانولوشنی را پیشنهاد می‌دهیم. در روش پیشنهادی از داده‌های توالی پروتئینی بیماران مبتلا به سرطان خون و انسان سالم، استفاده شده است. در این مقاله ما در رابطه با نوع جدیدی از برنامه‌های کاربردی در آنالیز دستاوردهایی از بازنمایی توالی پروتئین به تصویر را بحث می‌کنیم. هدف اصلی این مطالعه استفاده از قابلیت‌های شبکه عصبی کانولوشنی جهت طبقه‌بندی تصاویر بازنمایی شده از توالی پروتئین می‌باشد. ما ابتدا نوآوری در این تحقیق یعنی تبدیل توالی پروتئین به تصویر را ارائه می‌دهیم برای این منظور روش‌های مختلف در تبدیل توالی پروتئین به تصاویر را پیشنهاد می‌دهیم. سپس با اعمال فیلتر گایوربا زاویه و طول موج مختلف به تصویر دودویی، دقت طبقه‌بندی معماری پیشنهاد شده شبکه عصبی کانولوشنی را بررسی کرده و تعیین می‌کنیم با چه تنظیماتی می‌توانیم در بینش بیولوژیکی با تبدیل توالی پروتئین به تصویر باینری از آن استفاده نماییم. ما همچنین زمان آموزش برای رسیدن به دقت کل ۱۰۰٪ جهت طبقه‌بندی توالی‌های پروتئین سالم و سرطان خون بازنمایی شده با تصویر دودویی را با اعمال حالت‌های مختلف زوایا و طول موج فیلتر گابور و چگونگی بهترین استفاده از این تکنولوژی جدید را مورد بحث قرار می‌دهیم.

۳- شبکه عصبی کانولوشنی

شبکه عصبی کانولوشنی رده‌ای از یادگیری عمیق هستند که معمولا برای تحلیل تصاویر در یادگیری‌ماشین استفاده می‌شوند. ساختار شبکه کانولوشنی از فرایندهای زیستی قشر بینایی گربه الهام گرفته شده است. این ساختار به‌گونه‌ای است‌که تک نورون‌ها، تنها در یک ناحیه محدود به تحریک پاسخ می‌دهند که به آن ناحیه ادراکی گفته می‌شود. نواحی ادراکی نورون‌های مختلف، به‌صورت جزئی باهم هم-پوشانی دارند به گونه‌ای که کل میدان دید را پوشش می‌دهند. یک شبکه عصبی کانولوشنی از سه لایه اصلی تشکیل می‌شود که عبارتند از : لایه کانولوشنی[۶] لایه ادغام[۷]و لایه تماما متصل[۸] لایه های مختلف وظایف مختلفی را انجام می‌دهد. در هر شبکه عصبی کانولوشنی دو مرحله برای آموزش وجود دارد. مرحله پیش‌رو[۹] و مرحله پس انتشار:[۱۰]در مرحله اول تصویر ورودی به لایه کانولوشن شبکه تغذیه می‌شود و این عمل چیزی جز ضرب نقطه ای بین ماتریس تصویر ورودی و ماتریس فیلتر در هر لایه کانولوشن نیست. خروجی لایه های کانولوشن نشان‌دهنده ویژگی‌های سطح بالا در داده‌ها است، به عبارت ساده هدف لایه‌های کانولوشن در پردازش عکس ساختن ویژگی‌ها از داده های خام می‌باشد، آنها به دنبال اشیا و دنباله‌های با معنی موجود در عکس می‌گردند اما هیچ تصمیم گیری در مورد طبقه‌بندی انجام نمی‌دهند. فلت کردن این ویژگی‌ها در انتهای شبکه و اتصال آنها به دو لایه تماما متصل معمولا یک روش ارزان «از لحاظ بار محاسباتی» برای یادگیری ترکیبات غیرخطی این ویژگی‌ها است. ابعاد ماتریس وزن، برای تولید تعداد نرون‌های لازم در لایه تمام متصل برابر است با حاصل‌ضرب تعداد این نرون‌ها در تعداد نرون‌های لایه قبلی آنها. یکسوساز [۱۱]جهت

[6] Convolution
[7] Pooling
[8] Full Connected
[9] Feed-Forward
[10] Back Propagation
[11] RELU

صفر کردن مقادیر منفی ماتریس بدست آمده، استفاده می شود. لایه ادغام معمولا بعد از لایه کانولوشن قرار می گیرد و اندازه داده را کوچک می کنند. در ترکیب نورون‌ها، مکانیزم‌های مختلفی وجود دارد که معروف‌ترین آنها ادغام ماکسیمم[۲]است. در این مکانیزم پنجره هایی بر روی ماتریس بدست‌آمده مرحله قبل اعمال شده و با گام مشخصی حرکت می‌کند و وظیفه آن قرار دادن ماکسیمم اعداد موجود در پنجره به‌جای اعداد می باشد. لایه بیشینه هموار[۱۳] خروجی لایه تماما متصل[۱۴]را به توزیع احتمال کلاس‌ها تبدیل می-کند. سپس خروجی شبکه محاسبه می‌شود. به منظور تنظیم پارامترهای شبکه (مقادیر فیلترهای لایه کانولوشن و ماتریس‌های وزن لایه های تماما متصل)، در مرحله اول با استفاده از یک تابع خطا[۱۵] ، خروجی شبکه را با پاسخ صحیح مقایسه کرده و خطا محاسبه می‌شود. در مرحله بعدی بر اساس میزان خطای محاسبه شده مرحله پس‌انتشارخطا آغاز می‌شود. در این مرحله گرادیانت هر پارامتر، با توجه به قاعده زنجیره‌ای[۱۶]محاسبه می‌شود و تمامی پارامترها با توجه به تاثیری که بر خطای ایجادشده در شبکه دارند تغییر پیدا می‌کنند. بعد از بروز شدن پارامترها، مرحله بعدی پیش-رو شروع می‌شود. با تکرار تعداد مناسبی از این مراحل، آموزش شبکه پایان می‌یابد.

۴- فیلتر گابور

در کاربردهای مختلف بینایی کامپیوتر از قبیل آنالیز بافت و آشکارسازی لبه، توابع گابور بطور وسیعی استفاده شده‌اند . فیلتر گابور یک فیلتر خطی و محلی است. هسته کانولوشن فیلتر گابور حاصل‌ضرب یک تابع نمایی مختلط و گوسین است [۲۱]. مجموعه فیلترهای گابور از طریق رابطه (۱) بدست آورده می‌شوند.

(۱)

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = exp\left(-\frac{(xcos\theta+ysin\theta)^2+\gamma^2+(-xsin\theta+ycos\theta)^2}{2\sigma^2}\right) * exp\left(i\left(2\pi\frac{(xcos\theta+ysin\theta)}{\lambda}+\psi\right)\right)$$

در رابطه (۱) λ طول موج فرکانس سینوسی، θ زاویه چرخش فیلترهای گابور برای تعیین جهت نوارهای موازی تابع گابور ، ψ جابه‌جایی فاز برای تعیین تقارن تابع گابور، σ انحراف استاندارد پوشش تابع گاوسی، γ نسبت ابعاد فضایی و بیضوی، درصورتی که به‌طور مناسب و دقیق تنظیم شوند، عملکرد بسیار مناسبی در تشخیص ویژگی‌های بافت و لبه بافت دارند [۲۲]. ویژگی دیگر فیلترهای گابور درجه تفکیک مشترک بالای آنها است. این بدان معنی است که پاسخ آنها هم در حوزه مکان و هم در حوزه فرکانس کاملا محلی و قابل تنظیم کردن است.

در این مقاله زوایای مختلف چرخش فیلتر گابور با ۲ طول موج فرکانس سینوسی جهت استخراج ویژگی‌های مهمتر مورد مقایسه قرار می‌گیرند.

۵- روش‌های طبقه‌بندی توالی

سه روش متداول جهت طبقه‌بندی توالی وجود دارد. روش اول، طبقه‌بندی مبتنی بر ویژگی است. کل تعداد ویژگی‌های یک توالی پروتئین به طول n، شامل همه زیرمجموعه‌های ممکن برای مکان آمینواسیدها، به‌صورت رابطه ۲ تعریف می‌شود.

$$\sum_{k=0}^{n}\binom{n}{k} = \binom{n}{0}\binom{n}{1}+\cdots+\binom{n}{n} = 2^n \qquad (۲)$$

در رابطه ۲، n تعداد ویژگی‌های اولیه است. و k نیز زیر مجموعه انتخاب شده است. نشان داده شده است که پیدا کردن زیر مجموعه بهینه، یک مسئله NP-hard است [۲۳, ۲۴]. به‌عنوان روش‌های مبتنی بر ویژگی، در مرحله اول موقعیت حروف اسید آمینه به عنوان ویژگی تعریف شدند. در مرحله دوم، در حین حفظ یکپارچگی، حروف اسید آمینه توالی پروتئین به بردار دودویی تبدیل می‌شود؛ در این روش بازنمایی، به‌دلیل اینکه دنباله توالی‌ها باید حفظ شود، برای تک تک حروف اسید آمینه، در طول توالی پروتئین، اگر حرف مورد نظر ظاهر شود، مقدار ۱ و در غیر این‌صورت مقدار ۰ جایگزین می‌شود. جدول ۱ بازنمایی توالی پروتئین به شکل بردار دودویی را نشان می‌دهد.

---

۱ Max pooling ۲
۱ Softmax ۳
۱ fully connected ۴
۱ loss function ۵
۱ chain rule ۶

جدول ۱: نگاشت توالی پروتئین به بردار دودویی

توالی پروتئین	MKLIILQHVT…									
آمینو اسید	A	R	...	I	L	...	V			
بردار دودویی	0 0 0 0 0 0 0 0 0 0 0...	0 0 0 0 0 0 0 0 0 0 0...	...	0 0 0 1 1 0 0 0 0 0 0...	0 0 1 0 0 1 0 0 0 0 0...	...	0 0 0 0 0 0 0 0 1 0...			

روش دوم، طبقه‌بندی مبتنی بر فاصله است که شباهت بین توالی‌ها را بررسی می‌کند. به‌دلیل تفاوت در طول توالی‌های انتخاب شده در مجموعه داده‌ها این روش مورد بررسی قرار نگرفت. روش سوم، طبقه مبتنی بر مدل است. در این مقاله، با نوشتن برنامه متلب از شبکه عصبی کانولوشنی در روش‌های پیشنهادی بازنمایی نگاشت توالی پروتئین، جهت طبقه بندی توالی مبتنی بر مدل استفاده شد.

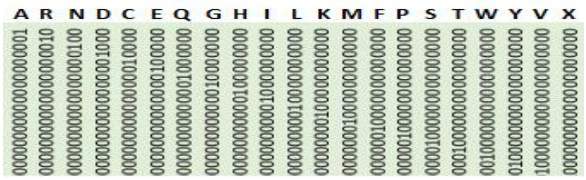
۶- روش‌های پیشنهادی بازنمایی نگاشت توالی پروتئین جهت طبقه‌بندی مبتنی بر مدل

در این بخش جهت ایجاد قالب ورودی معماری شبکه عصبی کانولوشنی طراحی شده در نرم‌افزار متلب، روش‌های پیشنهادی بازنمایی نگاشت توالی پروتئین به عدد ارائه می‌شود. برای ایجاد قالب ورودی طبقه‌بندی مبتنی بر مدل، در روش اول از با زنمایی توالی پروتئین به شکل آرایه یک‌بعدی از مقادیر عدد صحیح استفاده می‌شود. در این روش بازنمایی، حروف توالی اسید آمینه پروتئین، با استفاده از دستور aa2int نرم‌افزار متلب، به عدد صحیح مطابق جدول ۲ تبدیل می‌شوند. در نتیجه قالب ورودی شبکه عصبی کانولوشنی در این روش، یک بردار شامل ۲۸۹ عدد صحیح می‌باشد. برای ایجاد قالب ورودی طبقه‌بندی مبتنی بر مدل، در روش دوم از بازنمایی توالی پروتئین به شکل تصویر در قالب‌های بردار رنگی، تصویر رنگی مربعی، تصویر باینری و همچنین از تصویر باینری با اعمال فیلتر گابور به عنوان روش پیشنهادی، استفاده می‌شود.

جدول ۲: نگاشت توالی پروتئین به عدد صحیح با تبدیل آمینو اسید به عدد صحیح با تابع aa2int متلب

آمینواسید	کد	عدد صحیح
Alanine	A	۱
Arginine	R	۲
Asparagine	N	۳
Aspartic acid	D	۴
Cysteine	C	۵
Glutamic acid	E	۶
Glutamine	Q	۷
Glycine	G	۸
Histidine	H	۹
Isoleucine	I	۱۰
Leucine	L	۱۱
Lysine	K	۱۲
Methionine	M	۱۳
Phenylalanine	F	۱۴
Proline	P	۱۵
Serine	S	۱۶
Threonine	T	۱۷
Tryptophan	W	۱۸
Tyrosine	Y	۱۹
Valine	V	۲۰
Unknown	X	۰

در بازنمایی توالی پروتئین به شکل تصویر در ابتدا به ازای هر آمینو اسید، یک رشته بیتی شامل نوزده عدد ۰ و یک عدد ۱ ایجاد می‌شود. کد عدد صحیح آمینو اسید، محل قرار گرفتن ۱ را در رشته بیتی مطابق با شکل ۱ مشخص می‌نماید. به عنوان مثال رشته بیتی "۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۱۰"، کد R با عدد صحیح ۲ را نشان می‌دهد.



شکل ۱: نگاشت اسید آمینه به رشته بیتی

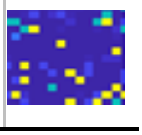
در قالب بردار رنگی، به‌ازای هر رشته بیتی یک کد رنگ RGB تولید می‌شود برای این‌کار از تبدیل کد باینری رشته بیتی به

دسیمال مطابق شکل ۲ استفاده شده است. در نتیجه به‌ازای هر حرف توالی پروتئین یک کد RGB تولید می‌شود. در این روش قالب ورودی یک بردار ۲۸۹ X ۱ خواهد شد.

توالی پروتئین	رشته بیتی	کد RGB
MKLIILQH...	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	4096 2048 1024 512 512 1024 ...

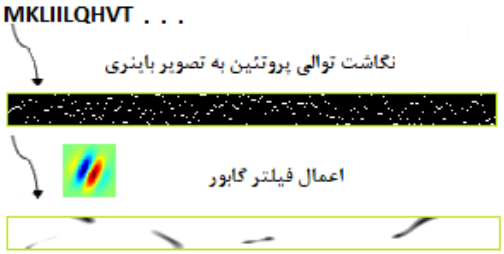
شکل ۲: مراحل بازنمایی توالی پروتئین به ۲۸۹ x ۱ کد RGB

در قالب تصویر رنگی مربعی هر ۱۷ کد RGB در یک سطر قرار می‌گیرد، شکل ۳ قالب ورودی به‌صورت تصویر رنگی ۱۷ X ۱۷ را نشان می‌دهد. سطر اول تصویر رنگی، بازنمایی ۱۷ حرف اول دنباله توالی پروتئین و سطر دوم نیز، ۱۷ حرف دوم، و سطرهای بعدی هم به همین ترتیب نگاشت می‌شوند.

توالی پروتئین با طول ۲۸۹	کد RGB	تصویر رنگی ۱۷x۱۷
MKLIILQH...	4096 2048 1024 512 512 1024	

شکل ۳: مراحل بازنمایی توالی پروتئین به تصویر رنگی ۱۷

برای ایجاد تصویر باینری ۲۸۹ X ۲۰ بر اساس رشته بیتی توالی پروتئین، در هر ستون تنها یک پیکسل سفید قرار می‌گیرد. مکان این پیکسل با توجه به مکان عدد یک در رشته بیتی مشخص می‌شود. مطابق با شکل ۴، هر حرف از توالی پروتئین، یک ستون از تصویر باینری را نشان می‌دهد.با توجه به اینکه طول تصویر باینری ایجاد شده با طول دنباله توالی یکسان می‌شود در نتیجه مفهوم دنباله در توالی حفظ می‌شود. عرض تصویر برابر ۲۰ یعنی تعداد اسید آمینه انتخاب شده‌است، لذا در هر ستون تنها یک پیکسل سفید وجود خواهد داشت که باعث می‌شود در این قالب ورودی، اسیدهای آمینه به صورت یکنواخت بازنمایی شوند. قالب پیشنهادی ورودی شبکه عصبی کانولوشنی در این مقاله جهت استخراج ویژگی‌های مهمتر، با اعمال فیلتر گابور در تصویر باینری، جهت بازنمایی توالی پروتئین مطابق با شکل ۴ معرفی می‌شود.



شکل ۴: تبدیل توالی پروتئین به تصویرباینری با ابعاد ۲۸۹ x ۲۰

۷- جامعه آماری

در این پژوهش، ۲۱۴۳ توالی اسید آمینه پروتئین از سایت مرکز ملی اطلاعات زیست فناوری معروف به NCBI که یکی از مراکز و شاخه‌های کتابخانه ملی پزشکی ایالات متحده آمریکا است و خود زیر مجموعه موسسه علمی سلامت NIH است، جمع آوری شده است. با توجه به اینکه طول توالی‌ها متفاوت بودند، در مرحله انتخاب داده‌ها، ۴۶۰ توالی پروتئین با طول بین ۲۳۵ و ۲۸۹ انتخاب شد .

۸- نمونه آماری

نمونه آماری شامل ۲ کلاس توالی پروتئینی است که عبارتند از:
سالم[۱۷]این کلاس شامل ۵۹ توالی پروتئینی که سرطان خون ندارند، می‌باشد.
سرطان خون[۱۸] این کلاس شامل ۴۰۱ توالی پروتئینی، از افرادی است که دچار بیماری سرطان خون می باشند .

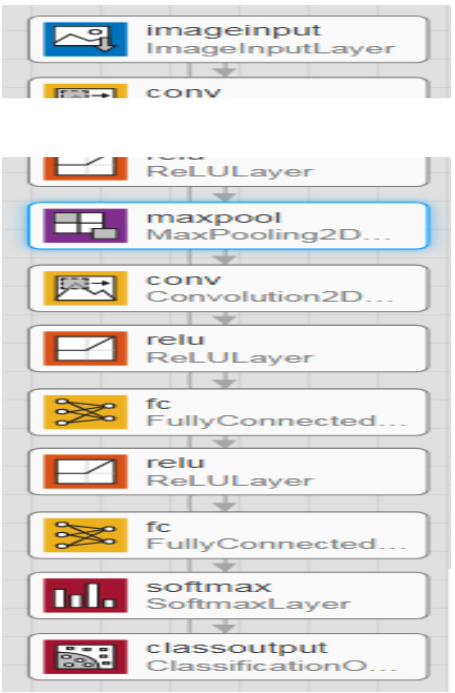
۹- ۶ مجموعه داده برای مدل‌های پیش‌بینی

مجموعه داده‌ها در این تحقیق به ۶ روش مختلف بازنمایی شده‌اند. شکل ۶ فلوچارت روش‌های طبقه‌بندی این مجموعه داده‌ها را نشان می‌دهد.
مجموعه داده اول شامل داده‌های اسمی، جهت طبقه بندی مدل‌های مبتنی بر ویژگی ایجاد شد، در این مجموعه داده موقعیت حروف اسیدآمینه به عنوان ویژگی تعریف شد، تعداد ۲۸۹ ویژگی برای هر توالی در نظر گرفته شد. در مجموعه داده دوم در حین حفظ یکپارچگی، حروف اسید آمینه توالی پروتئین به بردار دودویی تبدیل می‌شوند؛ در این روش بازنمایی، به‌دلیل اینکه دنباله توالی‌ها باید حفظ شود، برای تک تک حروف اسید آمینه، در طول توالی پروتئین، اگر حرف مورد نظر ظاهر شود، مقدار ۱ و در غیراین‌صورت مقدار ۰

---

[1] Leukemia [8]
[1] Normal [7]

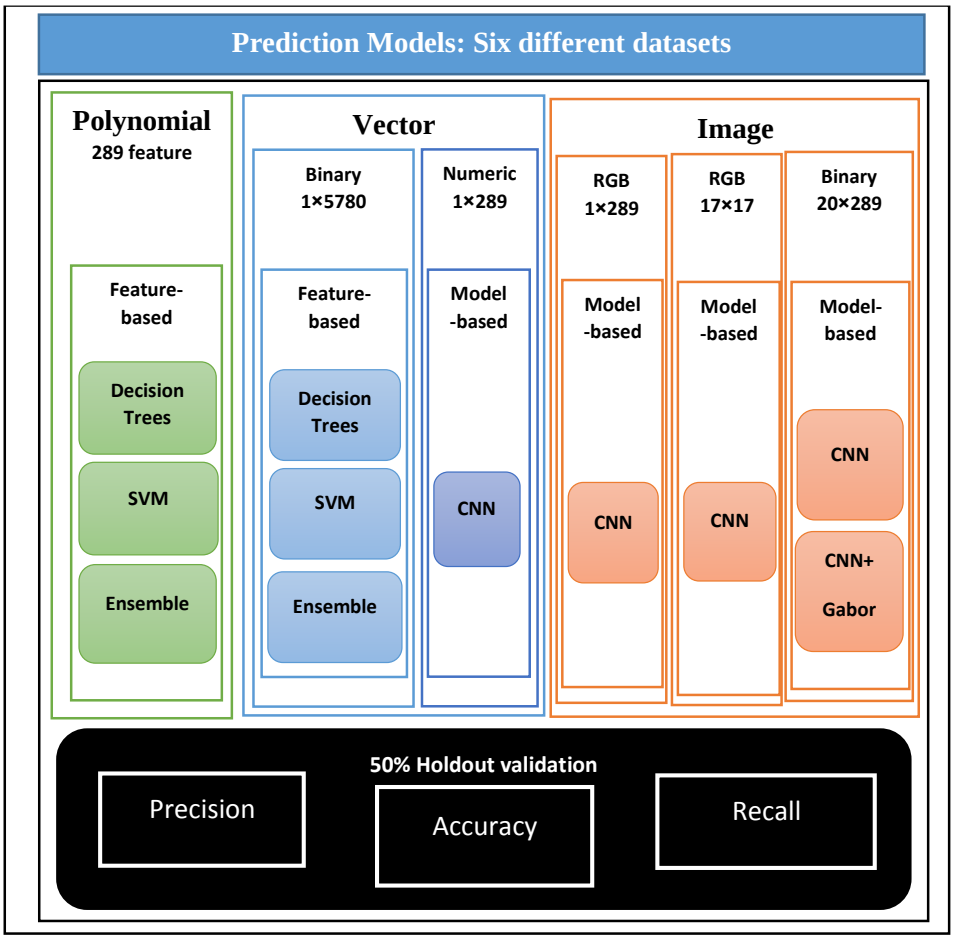
جایگزین می‌شود. تعداد کل ویژگی‌ها در این روش ۲۸۹*۲۰ برابر با ۵۷۸۰ خواهد شد. این مجموعه داده نیز جهت طبقه بندی مدل‌های مبتنی بر ویژگی ایجاد شد، روش‌های طبقه‌بندی مبتنی بر ویژگی در این ۲ مجموعه داده شامل درختان تصمیم، ماشین بردار پشتیبان، و روش‌های گروهی می‌باشد. مجموعه داده سوم، برداری از اعداد صحیح است، برای هر توالی به طول ۲۸۹، برداری شامل ۲۸۹ عدد صحیح به‌عنوان ویژگی، مطابق با اعداد صحیح جدول ۱ تعریف شد، شبکه عصبی کانولوشنی با معماری معرفی شده در شکل ۵ جهت طبقه بندی مبتنی بر مدل برای این مجموعه داده انتخاب شد. در معماری این شبکه عصبی کانولوشنی از ۱۱ لایه پس از لایه ورودی برای طبقه‌بندی استفاده شده است. لایه اول از لایه کانولوشن با ۳۰ فیلتر با ابعاد ۱X۳، لایه دوم از یکسوساز جهت صفر کردن اعداد منفی خروجی لایه قبل، لایه سوم لایه کانولوشن با ۶۰ فیلتر ۱X۲ و لایه چهارم نیز لایه یکسوساز، لایه پنجم لایه ادغام با ناحیه ادراکی ۱X۲ با عملگر حداکثر با گام ۲، لایه ششم لایه کانولوشن با ۸۰ فیلتر ۱X۳، لایه هفتم لایه یکسوساز، لایه هشتم و نهم تماما متصل به‌ترتیب با ۳۳ و ۲ نورون، لایه دهم لایه بیشینه هموار و لایه یازدهم، لایه طبقه‌بندی[9] تعریف شده است.



شکل ۵: معماری شبکه عصبی کانولوشن با ورودی عدد آرایه یک‌بعدی .

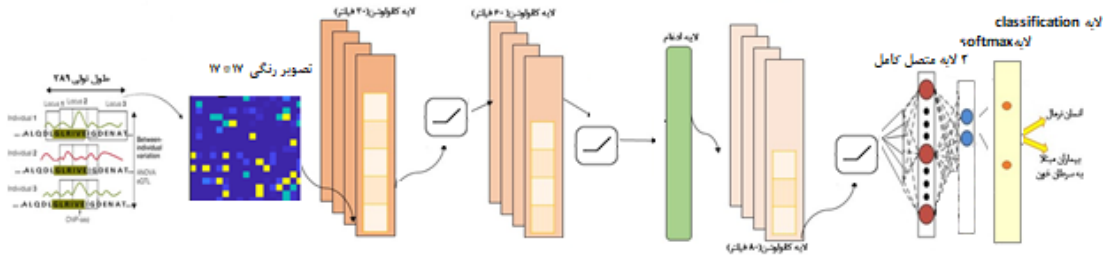
مجموعه داده چهارم با بردار کد رنگ ایجاد شد، در این مجموعه داده به‌ازای هر توالی ۲۸۹ کد رنگ به عنوان ویژگی در نظر گرفته شد. برای این مجموعه داده نیز جهت طبقه بندی مبتنی بر مدل از شبکه عصبی کانولوشنی با معماری شکل ۵ استفاده شد.

[1] classification [9]



شکل ۶: روش‌های طبقه‌بندی در ۶ مجموعه داده

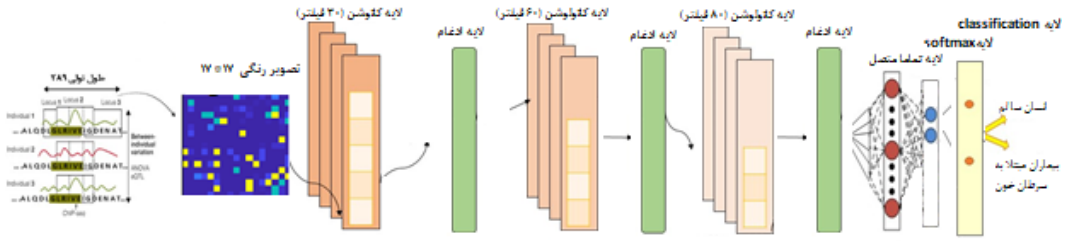
در مجموعه داده پنجم، برای هر توالی تصویر رنگی ۱۷X۱۷ تعریف شد. در این روش ۲ مدل معماری شبکه عصبی عمیق پیشنهاد داده شد. مطابق با شکل ۷ در مدل اول از شبکه عصبی کانولوشنی با ۱۱ لایه پس از لایه ورودی برای طبقه‌بندی تصاویر رنگی مربعی بازنمایی شده از توالی‌های پروتئین استفاده شد. لایه اول از لایه کانولوشنی با ۲۰ فیلتر با ابعاد ۵ X ۵، لایه دوم از یکسوساز جهت صفر کردن اعداد منفی خروجی لایه قبل، لایه سوم لایه کانولوشنی با ۶۰ فیلتر ۴ X ۴ و لایه چهارم نیز لایه یکسوساز، لایه پنجم لایه ادغام با ناحیه ادراکی ۲ X ۲ با عملگر ماکسیمم، لایه ششم لایه کانولوشنی با ۸۰ فیلتر ۳ X ۳، لایه هفتم لایه یکسوساز، لایه هشتم و نهم تماما متصل به ترتیب با ۲۰ و ۲ نورون، لایه دهم، لایه بیشینه هموار و لایه یازدهم، لایه طبقه‌بندی تعریف شده است.



شکل ۷: مدل اول شبکه عصبی کانولوشن جهت طبقه‌بندی تصویر رنگی ۱۷X۱۷

مطابق با شکل ۸ در مدل دوم، از ۱۰ لایه پس از لایه ورودی جهت طبقه‌بندی تصاویر رنگی مربعی بازنمایی شده از توالی‌های پروتئین استفاده شد. در لایه اول، از ۳۰ فیلتر کانولوشن با ابعاد ۳ X ۳، لایه دوم از لایه ادغام جهت کاهش ابعاد با ناحیه ارداکی ۲ X ۲، لایه

سوم شامل ۶۰ فیلتر کانولوشن با ابعاد ۲ X ۲، لایه چهارم مانند لایه دوم، لایه پنجم شامل ۸۰ فیلتر کانولوشن با ابعاد ۳ X ۳، لایه ششم نیز شبیه لایه چهارم و دوم از لایه ادغام با ناحیه ادراکی ۲ X ۲، لایه هفتم و هشتم لایه های تماما متصل به ترتیب با ۳۳ و ۲ نرون، لایه نهم، لایه بیشینه هموار و لایه دهم، لایه طبقه‌بندی تعریف شده است.



شکل ۸: مدل دوم شبکه عصبی کانولوشن در روش اول پیشنهادی

مطابق با شکل ۹ در روش پیشنهادی سوم از ۱۰ لایه شبکه عصبی کانولوشنی برای طبقه‌بندی خروجی فیلتر گابور تصاویر دودویی ۲۸۹ X ۲۰ استفاده می‌شود. لایه اول شامل ۱۰ فیلتر کانولوشن با ابعاد ۱۰۰ X ۱۰، لایه دوم: لایه یکسوساز، لایه سوم: لایه ادغام با ناحیه ادراکی ۵۰ X ۵، لایه چهارم شامل ۱۰ فیلتر کانولوشن با ابعاد ۸۰ X ۴، لایه پنجم: لایه یکسوساز، لایه ششم: لایه تماما متصل با ۱۰۰ نورون، لایه هفتم: لایه یکسوساز، لایه هشتم: لایه تماما متصل با ۲ نورون، لایه نهم: لایه بیشینه هموار، لایه دهم: لایه طبقه‌بندی در نظر گرفته شده است.

شکل ۹: معماری شبکه کانولوشن در روش پیشنهادی تبدیل توالی پروتئین به تصاویر دودویی

۱۰- ارزیابی مدل‌های پیشنهادی طبقه‌بندی توالی پروتئین سالم و مریض(سرطانی)

با توجه به عدم توازن تعداد نمونه‌های هر دو کلاس، جهت ارزیابی دقت، از روش 2-fold-Cross-Validation (در مرحله اول ۵۰ درصد اول داده‌ها در هر کلاس به عنوان داده‌های آموزشی و ۵۰ درصد دوم نیز به عنوان داده‌های تست و در مرحله دوم ۵۰ درصد اول تست و ۵۰ درصد دوم آموزشی در نظر گرفته می‌شوند) جهت مقایسه عملکرد مدل‌های مخصوص طبقه‌بندی توالی با مدل‌های طبقه‌بندی مبتنی با تصویر استفاده شده است. در این ارزیابی از حساسیت[20](تقسیم تعداد درست تشخیص داده شده یک کلاس بر تعداد واقعی همان کلاس) و دقت[21](تقسیم تعداد درست تشخیص داده شده یک کلاس بر تعداد پیش‌بینی همان کلاس) و دقت‌کل[22](تقسیم تعداد درست تشخیص داده شده ۲ کلاس بر تعداد کل نمونه‌های تست) استفاده شده است.

۱۱- یافته‌ها

روش‌های طبقه‌بندی یادگیرنده نرم‌افزار متلب[23]شامل ۳ روش درخت تصمیم(Fine Tree, Medium Tree, Coarse Tree)، و ۷ روش ماشین بردار پشتیبان (Logistic Regression, Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM,) و ۳ روش گروهی (Ensemble Boosted Trees, Ensemble Bagged Trees, Ensemble RUS Boosted Trees) جهت طبقه‌بندی ۲ کلاس نامتوازن سالم و مریض در ۲ مجموعه داده مبتنی بر ویژگی معرفی شده در شکل ۵ مورد استفاده قرار گرفتند. جدول ۳، ارزیابی روش‌های طبقه‌بندی یادگیرنده نرم‌افزار متلب با داده اسمی در مجموعه داده مبتنی بر ویژگی را نشان می‌دهد. در این مجموعه داده، روش طبقه‌بندی Ensemble Bagged Trees بیشترین دقت کل را در طبقه‌بندی ۲ کلاس ارائه داد. جدول ۴ نیز ارزیابی روش‌های طبقه‌بندی یادگیرنده نرم‌افزار متلب با داده بردار دودویی در مجموعه داده مبتنی بر ویژگی را نشان می‌دهد. در این مجموعه داده، روش‌های طبقه‌بندی Quadratic SVM و Cubic SVM با دقت ۹۲,۰٪ بیشترین دقت کل را در طبقه‌بندی ۲ کلاس ارائه دادند. همچنین در مقایسه جدول ۳ با ۴ مشاهده شد در بیشتر روش‌های طبقه‌بندی یادگیرنده نرم‌افزار متلب در روش‌های مبتنی بر ویژگی، ورودی توالی به‌صورت بردار دودویی نتایج بهتری نسبت به ورودی داده‌های اسمی ارائه می‌دهند. علت افزایش دقت به ۲ دلیل می‌باشد. اول اینکه در روش بازنمایی با بردارودویی، دنباله توالی پروتئین به‌صورت یکپارچه کد شده‌اند، دوم اینکه پیوستگی در زیر دنباله در این روش بازنمایی حفظ شده است. جدول ۵ ارزیابی روش‌های طبقه مبتنی بر مدل را نشان می‌دهد. شبکه عصبی کانولوشنی مجموعه داده بردار عددی یک‌بعدی کلاس ۱ را تشخیص نداد. ارزیابی طبقه‌بندی شبکه عصبی کانولوشنی با ورودی به‌صورت بردار کد رنگ نسبت‌به بردار عددی بهتر شد. در هر ۲ روش بردار عدد صحیح و بردار کد رنگ به‌دلیل بازنمایی غیر یکپارچه حروف توالی، ارزیابی مناسبی نداشتند. شبکه عصبی کانولوشنی بر روی تصاویر رنگی مربعی در قالب (۱۷ X ۱۷)، در دو معماری مختلف جهت آموزش شبکه ارائه شد. هر یک از معماری‌ها با مقادیر و لایه‌های مختلف، مورد بررسی و تحلیل قرار گرفته شد. در این بررسی مشاهده شد که به‌دلیل قطعه قطعه شدن دنباله برای ایجاد تصویر مربعی در ۲ مدل شبکه عصبی کانولوشنی، و همچنین با توجه به یکپارچه نبودن بازنمایی اسیدهای آمینه به مقادیر رنگی به‌دلیل تفاوت بسیار زیاد اعداد رنگی ایجاد شده از اسیدهای آمینه، دقت و حساسیت کلاس ۱ کمتر از ۵۰٪ شد. اما در روش پیشنهادی با تعریف قالب ورودی به شکل تصاویر (۲۸۹ X ۲۰) دودویی و اعمال فیلتر گابور، دقت و حساسیت هر ۲ کلاس بیشتر از ۹۰٪ شد. جهت رسیدن به دقت کل بالاتر و تعیین تعداد دفعات لازم برای آموزش داده‌های آموزشی شبکه عصبی، ۴ زاویه (۰ و ۴۵ و ۹۰ و ۱۳۵) و ۲ طول موج (۵ و ۱۰) برای اعمال فیلتر گابور در نظر گرفته شد. نتایج جدول ۶ نشان می‌دهد که زاویه ۹۰ درجه از دو جنبه تعداد دفعات لازم برای آموزش داده‌های آموزشی شبکه عصبی و دقت کل داده‌های تست، با توجه به روش بازنمایی تصویر باینری و نگاشت حروف اسیدآمینه به‌صورت یکنواخت دودویی در عرض تصویر، مناسب‌تر است. همانطور که مشاهده می‌شود کمترین دقت و حسایت در تشخیص کلاس سالم، مربوط به روش‌های مبتنی بر ویژگی، با ورودی داده‌های اسمی توالی پروتئین در مدل Logistic regression و Ensemble Boosted Trees و با ورودی بردار دودویی توالی پروتئین در مدل Coarse Gaussian SVM و Ensemble Boosted Trees همچنین در مدل‌های مبتنی بر مدل با ویژگی‌ها به‌صورت بردار

---

20 Recall
21 Precision
22 Accuracy
23 Classification Learner - MathWorks

عددی یک‌بعدی در شبکه عصبی کانولوشنی با میانگین ۰٪ و بیشترین دقت و حساسیت در ورودی تصاویر دودویی با ابعاد ۲۸۹ X ۲۰ و اعمال فیلتر گابور با زاویه ۹۰ درجه در روش پیشنهادی با میانگین ۹۵,۴٪ را نشان دادند. در تشخیص کلاس مریض (سرطانی) کمترین دقت و حساسیت در مدل Logistic regression در روش‌مبتنی بر ویژگی در ورودی بردار دودویی با میانگین ۸۹٪ و بیشترین دقت و حساسیت با میانگین ۹۹,۳٪ با ورودی تصویر دودویی با ابعاد ۲۸۹ X ۲۰ و اعمال فیلتر گابور با زاویه ۹۰ درجه در روش پیشنهادی مشاهده شد. پایین‌ترین و بالاترین دقت کل در مجموع تشخیص هر ۲ کلاس نیز به‌ترتیب در مدل Logistic regression در روش‌مبتنی بر ویژگی در ورودی بردار دودویی با دقت کل ۸۱,۳٪ و در روش سوم پیشنهادی با ورودی تصویر دودویی با ابعاد ۲۸۹ X ۲۰ و اعمال فیلتر گابور (طول موج ۱۰ و زاویه ۹۰ درجه) با دقت کل ۹۸,۸٪ بدست آمد.

جدول ۳: حساسیت و دقت ارزیابی 2Fold-Cross-Validation روش‌های طبقه‌بندی یادگیرنده نرم‌افزار متلب با ورودی داده‌های اسمی توالی پروتئین مبتنی بر ویژگی

روش طبقه‌بندی	نوع کلاس	حساسیت Recall	دقت Precision	دقت کل Accuracy
Fine Tree	۱	۴۷%	۵۰%	۸۷,۲%
	۲	۹۳%	۹۲%	
Medium Tree	۱	۴۷%	۵۰%	۸۷,۲%
	۲	۹۳%	۹۲%	
Coarse Tree	۱	۳۷%	۵۰%	۸۷,۲%
	۲	۹۵%	۹۱%	
Logistic Regression	۱	۰%	۰%	۸۷,۲%
	۲	۱۰۰%	۸۷%	
Linear SVM	۱	۲%	۵۰%	۸۷,۲%
	۲	۹۹%	۸۷%	
Quadratic SVM	۱	۲%	۵۰%	۸۷,۲%
	۲	۹۹%	۸۷%	
Cubic SVM	۱	۲%	۵۰%	۸۷,۲%
	۲	۹۹%	۸۷%	
Fine Gaussian SVM	۱	۲%	۳۳%	۸۷,۰%
	۲	۹۹%	۸۷%	
Medium Gaussian SVM	۱	۲%	۵۰%	۸۷,۲%
	۲	۹۹%	۸۷%	
Coarse Gaussian SVM	۱	۲%	۳۳%	۸۷,۰%
	۲	۹۹%	۸۷%	
Ensemble Boosted Trees	۱	۰%	۰%	۸۷,۲%
	۲	۱۰۰%	۸۷%	
Ensemble Bagged Trees	۱	۲۴%	۱۰۰%	۹۰,۲%
	۲	۱۰۰%	۹۰%	
Ensemble RUS Boosted Trees	۱	۷۱%	۵۸%	۸۹,۶%
	۲	۹۲%	۹۶%	

جدول ۴: حساسیت و دقت ارزیابی 2Fold-Cross-Validation روش‌های طبقه‌بندی یادگیرنده نرم‌افزار متلب با ورودی روش ابتکاری بازنمایی یکپارچه دودویی توالی پروتئین مبتنی بر ویژگی

روش طبقه‌بندی	نوع کلاس	حساسیت Recall	دقت Precision	دقت کل Accuracy
Fine Tree	۱	۶۳%	۵۹%	۸۹,۶%
	۲	۹۴%	۹۴%	
Medium Tree	۱	۶۳%	۵۹%	۸۹,۶%
	۲	۹۴%	۹۴%	
Coarse Tree	۱	۴۴%	۵۵%	۸۸,۵%
	۲	۹۵%	۹۴%	
Logistic Regression	۱	۷۶%	۳۸%	۸۱,۳%
	۲	۸۲%	۹۶%	
Linear SVM	۱	۲۴%	۱۰۰%	۹۰,۲%
	۲	۱۰۰%	۹۰%	
Quadratic SVM	۱	۳۷%	۱۰۰%	۹۲,۰%
	۲	۱۰۰%	۹۲%	
Cubic SVM	۱	۳۷%	۱۰۰%	۹۲,۰%
	۲	۱۰۰%	۹۲%	
Fine Gaussian SVM	۱	۲۷%	۱۰۰%	۹۰,۷%
	۲	۱۰۰%	۹۰%	
Medium Gaussian SVM	۱	۲۹%	۱۰۰%	۹۰,۹%
	۲	۱۰۰%	۹۱%	
Coarse Gaussian SVM	۱	۰%	۰%	۸۷,۲%
	۲	۱۰۰%	۸۷%	
Ensemble Boosted Trees	۱	۰%	۰%	۸۷,۲%
	۲	۱۰۰%	۸۷%	
Ensemble Bagged Trees	۱	۲۹%	۱۰۰%	۹۰,۹%
	۲	۱۰۰%	۹۱%	
Ensemble RUS Boosted Trees	۱	۸۵%	۴۳%	۸۳,۵%
	۲	۸۳%	۹۷%	

جدول ۵: مقایسه حساسیت و دقت ارزیابی 2Fold-Cross-Validation روش‌های بازنمایی توالی پروتئین مبتنی بر مدل

نوع روش بازنمایی توالی پروتئین	نوع کلاس	حساسیت Recall	دقت Precision	دقت کل Accuracy
بردار اعداد صحیح ۱X۲۸۹	۱	۰,۰%	۰,۰%	۸۷,۲%
	۲	۱۰۰%	۸۷%	
بردار کد رنگ ۱X۲۸۹	۱	۴۸,۴%	۶۸,۲%	۹۰,۲%
	۲	۹۶,۶%	۹۲,۵%	
معماری اول تصویر رنگی ۱۷X۱۷	۱	۴۵,۲%	۴۵,۲%	۸۵,۵%
	۲	۹۱,۶%	۹۱,۶%	
معماری دوم تصویر رنگی ۱۷X۱۷	۱	۵۰%	۹,۷%	۸۶,۸%
	۲	۸۷,۷%	۹۸,۵%	
تصویر دودویی ۲۰X۲۸۹	۱	۱۷,۲%	۱۰۰%	۸۹,۵%
	۲	۱۰۰%	۸۹,۳%	
تصویر دودویی با فیلتر گابور	۱	۱۰۰%	۹۰,۷%	۹۸,۸%
	۲	۹۸,۶%	۱۰۰%	

جدول ۶: مقایسه حساسیت و دقت ارزیابی 2Fold-Cross-Validation و تعداد آموزش روش‌های اعمال فیلتر گابور در تصویر دودویی

نوع فیلتر گابور	تعداد آموزش	نوع کلاس	حساسیت Recall	دقت Precision	دقت کل Accuracy
بدون فیلتر	۳۵۰	۱	۱۷,۲%	۱۰۰%	۸۹,۵%
		۲	۱۰۰%	۸۹,۳%	
زاویه ۰ طول ۵	۲۰۰	۱	۵۸,۶%	۶۰,۷%	۹۰,۰%
		۲	۹۴,۵%	۹۴,۰%	
زاویه ۰ طول ۱۰	۳۵۰	۱	۴۸,۳%	۷۰,۰%	۹۰,۸%
		۲	۹۷,۰%	۹۲,۸%	
زاویه ۴۵ طول ۵	۲۵۰	۱	۶۶,۵%	۶۵,۵%	۹۱,۳%
		۲	۹۵,۰%	۹۵,۰%	
زاویه ۴۵ طول ۱۰	۳۵۰	۱	۵۵,۲%	۵۹,۳%	۸۹,۵%
		۲	۹۴,۵%	۹۳,۶%	
زاویه ۹۰ طول ۵	۲۰۰	۱	۸۲,۸%	۸۸,۹%	۹۶,۵%
		۲	۹۸,۵%	۹۷,۵%	
زاویه ۹۰ طول ۱۰	۲۰۰	۱	۱۰۰%	۹۰,۷%	۹۸,۸%
		۲	۹۸,۶%	۱۰۰%	
زاویه ۱۳۵ طول ۵	۲۵۰	۱	۳۱,۰%	۹۰,۰%	۹۰,۸%
		۲	۹۹,۵%	۹۰,۹%	
زاویه ۱۳۵ طول ۱۰	۲۵۰	۱	۳۴,۵%	۷۶,۹%	۹۰,۴%
		۲	۹۸,۵%	۹۱,۲%	

۱۲- بحث

روش‌های طبقه بندی مخصوص توالی مبتنی بر ویژگی، به‌دلیل در نظر نگرفتن ارتباط بین زیر مجموعه‌های توالی به‌عنوان ویژگی، که بر اساس موقعیت زیر مجموعه در توالی معنی پیدا می‌کنند، و همچنین روش‌های طبقه بندی مخصوص توالی مبتنی بر فاصله، به‌دلیل تفاوت در طول توالی پروتئین، و در نهایت روش‌های طبقه بندی مخصوص توالی مبتنی بر مدل با ورودی بردار یک‌بعدی به‌ازای هر توالی، به‌دلیل عدم یکپارچگی در کد کردن حروف اسید آمینه، جهت طبقه بندی ۴۶۰ توالی پروتئین با طول بین ۲۳۵ و ۲۸۹ در ۲ کلاس سالم و سرطانی موفق نبودند. جهت بهبود دقت طبقه‌بندی در روش‌های یادگیری ماشین، با رویکرد ابتکاری، در حین حفظ یکپارچگی، حروف اسید آمینه توالی پروتئین به بردار دودویی تبدیل شدند. جهت بهبود دقت در طبقه‌بندی مبتنی بر مدل در شبکه عصبی کانولوشنی، ۳ روش برای بازنمایی توالی پروتئین به تصویر ارائه شد. روش اول، در تبدیل حروف توالی پروتئین به کد رنگ به‌دلیل بازنمایی غیریکپارچه برای هر اسید آمینه موفق نبود. روش دوم با تغییر ساختار دنباله با کنار هم قرار دادن هر ۱۷ اسید آمینه برای ایجاد تصویر رنگی مربعی، در طبقه‌بندی توالی‌ها، موفق نبود. روش سوم، بازنمایی تصویر دودویی به‌دلیل بازنمایی یکپارچه اسید آمینه و عدم تغییر در ساختار اصلی دنباله و اعمال فیلتر گابور برای استخراج ویژگی‌های مناسب مانند بافت و لبه بافت بسیار موفق عمل کرد. با توجه به اینکه عرض تصاویر دودویی تعیین کننده نوع اسید آمینه موجود در توالی پروتئین می‌باشد زاویه فیلتر گابور ۹۰ درجه به بهترین شکل ویژگی‌های موثر را استخراج کردند. تجزیه و تحلیل انجام شده بر روی روش‌های بازنمایی توالی پروتئین، بیان کننده موفقیت نگاشت توالی به‌صورت ماتریس ۲ بعدی نسبت‌به بردار یک‌بعدی را نشان می‌دهد. ماتریس ۲ بعدی به شکل دودویی، به‌دلیل حفظ یکپارچگی در بازنمایی حروف اسید آمینه در عرض ماتریس، نسبت به حالت یک‌بعدی که حروف اسید آمینه بدون رعایت یکپارچگی به عدد نگاشت می‌شوند، بازنمایی مناسب‌تری دارند. علت مهم عدم موفقیت طبقه‌بندی در نگاشت توالی پروتئین به عدد ارایه یک بعدی، جایگزین کردن داده‌های اسمی با داده‌های عددی می‌باشد، اختلاف این اعداد برای شبکه عصبی معنی دار خواهد بود و لایه های مختلف شبکه عصبی نمی‌توانند بازنمایی خوبی از این اعداد ایجاد نمایند و پارامترهای وزن لایه تمام متصل نیز برای پیش‌بینی کلاس درست، به خوبی آموزش نخواهد دید. از طرف دیگر به‌دلیل یکسان بودن طول ماتریس با طول توالی، ساختار توالی در بازنمایی به‌صورت ماتریس دودویی حفظ می‌شود و وابستگی‌های موجود بین حروف اسید آمینه که با ترتیب قرار گرفتن

آنها در دنباله تعریف می شود نیز برقرار می‌ماند، این در حالی است که برخی از این وابستگی‌ها، در روش‌های مخصوص توالی مبتنی بر ویژگی نادیده گرفته می‌شود زیرا که آنها بخشی از زیر مجموعه‌های توالی را به عنوان ویژگی‌ها در نظر می‌گیرند. فیلنر گابور همانطور که در تصاویر، به‌عنوان تشخیص لبه و بافت تعریف می‌شود. در تصویر دودویی بازنمایی شده از توالی، جهت تشخیص وابستگی‌های مهم در بین حروف اسید آمینه اعمال شده است، لذا بازنمایی توالی به صورت ماتریس دودویی در درجه اول، یکنواختی در تبدیل حروف اسید آمینه را با استفاده از اعداد دودویی برقرار می‌کند، و در درجه دوم، وابستگی بین حروف اسید آمینه توالی که در ترتیب این حروف تعریف می شود، را حفظ می‌کند. در نتیجه ماتریس دودویی به‌عنوان ورودی شبکه عصبی کانولوشنی، بازنمایی مناسبی از توالی پروتئین را ارائه می‌دهد. رویکرد ارائه شده یادگیری عمیق با اعمال فیلتر گابور به تصویر دودویی و معماری شبکه عصبی کانولوشنی شامل ۱۱ لایه معرفی شد. لایه‌های کانولوشن در این معماری با اعمال فیلترهای مختلف برای پیچش[۴]کردن تصویر دودویی ورودی و همینطور برای نگاشت ویژگی‌های[۵] میانی استفاده می‌شود. اتصال محلی فیلترها با تصویر دودویی ورودی ارتباط موجود در زیر دنباله توالی را یاد می‌گیرند و با نگاشت ویژگی که انجام می‌شود تعداد پارامترها تصویر دودویی ورودی بسیار کاهش می‌یابد. ابعاد ۱۰x۱۰۰ فیلتر با گام ۱ در لایه اول کانولوشن به این دلیل انتخاب شده است که عرض تصویر دودویی ورودی برابر ۲۰ و طول تصویر دودویی ورودی برابر ۲۸۹ می‌باشد و بتواند وابستگی تقریبا نیمی از زیر دنباله را در توالی استخراج و نگاشت کند. لایه یکسوساز بعد از این لایه مقادیر منفی را صفر می‌کند، لایه ادغام نیز برای کاهش ابعاد و لایه کانولوشن دوم با ۱۰ فیلتر ۴۰x۸۰ نیز طوری انتخاب شده است که تقریبا نیمی از ابعاد خروجی لایه قبل را پوشش دهد. لایه یکسوساز مجددا جهت صفر کردن اعداد منفی و ۲ لایه تماما متصل جهت آموزش وزن‌ها برای نگاشت داده‌ها ابتدا به ۱۰۰ عدد و سپس به ۲ عدد تعریف شده است. لایه بیشینه هموار جهت تعیین احتمال پیش‌بینی هر کلاس و لایه طبقه‌بندی برای تعیین کلاس خروجی در نظر گرفته شده است.

۱۳- نتیجه گیری

استفاده از ابزارهای کامپیوتری مکانیزه کننده، مخصوصا در یادگیری عمیق به منظور تسهیل آنالیزهای پزشکی و تشخیص، یک عرصه مهم و امید بخش می‌باشد. در این تحقیق نشان دادیم که چگونه یادگیری مشخصه با ناظر، می‌تواند برای داده توالی پروتئین اسیدهای آمینه، با استفاده از یادگیری عمیق مورد استفاده قرار گیرد. مزیت اصلی روش پیشنهادی نسبت به روش‌های قبلی طبقه‌بندی داده توالی اسیدهای آمینه پروتئین، انتخاب بازنمایی متناسب با نگاشت ژنوم با رویکرد یادگیری عمیق بود. به‌دلیل تفاوت در ابعاد تصاویر بازنمایی شده در سه روش پیشنهادی، از معماری‌های متفاوتی برای ایجاد شبکه عصبی کانولوشنی برای هر روش پیشنهادی استفاده شد. روش پیشنهادی موفق ما در این تحقیق، نگاشت توالی پروتئین اسیدهای آمینه به‌صورت تصاویر دودویی با ابعاد ۲۸۹ X ۲۰ و اعمال فیلتر گابور با زاویه ۹۰ درجه روی تصاویر و همچنین طبقه‌بندی تصاویر با شبکه عصبی کانولوشنی با معماری پیشنهادی، ذکر شده در این مقاله می‌باشد. نتایج بدست آمده از تجزیه و تحلیل بر روی داده‌های توالی پروتئینی اسیدهای آمینه افراد سالم و افراد دارای سرطان خون ، نشان داد که روش پیشنهادی، جهت طبقه‌بندی افراد سالم و افراد دارای سرطان خون با وجود عدم توازن در تعداد نمونه داده‌های ۲ کلاس به‌دلیل استفاده از رمزگذاری یکپارچه حروف اسید آمینه و عدم تغییر ساختار دنباله و همچنین اعمال فیلتر گابور با زاویه ۹۰ درجه بسیار موفقیت آمیز صورت گرفته است و بنابراین نوید دهنده دستیابی کلی‌تر و جامع‌تری برای طبقه‌بندی داده توالی اسیدهای آمینه پروتئین خواهد بود.

مراجع

[1] A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," 2015, pp. 1328-1335.

[2] Y. Liu, S. Zhou, and Q. Chen, "Discriminative deep belief networks for visual data classification," Pattern Recognition, vol. 44, pp. 2287-2296, 2011.

[3] J. Chen, R. Swofford, J. Johnson, B. B. Cummings, N. Rogel, K. Lindblad-Toh, et al., "A quantitative framework for characterizing the evolutionary history of mammalian gene expression," Genome research, vol. 29, pp. 53-63, 2019.

[4] T. Hardy, J. Feng, D. Lawrence, T. Fullston, and H. Scott, "Application of Artificial Intelligence To Analysis of The Embryonic Genome For Preimplantation Genetic Diagnosis," Pathology, vol. 51, p. S65, 2019.

۴ Convolve

۵ Feature mapping

[5] C. S. Boddy and S. Ma, "Frontline therapy of CLL: evolving treatment paradigm," Current hematologic malignancy reports, vol. 13, pp. 69-77, 2018.

[6] K. He, D. Ge, and M. He, "Big data analytics for genomic medicine," International journal of molecular sciences, vol. 18, p. 412, 2017.

[7] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," Molecular systems biology, vol. 12, p. 878, 2016.

[8] M. Leung, H. Xiong, L. Lee, and B. Frey, "Deep learning of the tissueregulated splicing code," Bioinformatics 30, pp. i121 – i129, 2014.

[9] H. Xiong, B. Alipanahi, L. Lee, H. Bretschneider, D. Merico, R. Yuen, et al., "The human splicing code reveals new insights into the genetic determinants of disease," Science 347, p. 1254806, 2015.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," Advances in Neural Information Processing Systems 27, pp. 3320-3328, 2014.

[11] B. Alipanahi, A. Delong, M. Weirauch, and B. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," Nat Biotechnol 33, pp. 831 – 838, 2015.

[12] J. Zhou and O. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," Nat Methods 12, pp. 931 – 934, 2015.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," 2018, pp. 512-519.

[14] W. Sun, T.-L. B. Tseng, J. Zhang, and W. Qian, "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data," Computerized Medical Imaging and Graphics, vol. 57, pp. 4-9, 2017.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine learning, vol. 46, pp. 389-422, 2002.

[16] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, 2014, pp. 818-833.

[17] M. Biswas, A. Tiwari, M. Turk, J. Laird, C. Asare, L. Saba, et al., "A Review on a Deep Learning Perspective in Brain Cancer Classification," Cancers, vol. 11, 2019.

[18] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85-117, 2015.

[19] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, et al., "Recent advances in convolutional neural networks," Pattern Recognition, vol. 77, pp. 354-377, 2018.

[20] M. A. Jafri, S. A. Ansari, M. H. Alqahtani, and J. W. Shay, "Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies," Genome medicine, vol. 8, p. 69, 2016. "

[21] X. Chu and K. L. Chan, "Rotation and scale invariant texture analysis with tunable Gabor filter banks," in Pacific-Rim Symposium on Image and Video Technology, 2009, pp. 83-93.

[22] R. C. González, R. E. Woods, and S. L. Eddins, Digital Image Processing Using MATLAB: Pearson, 2004.

[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, pp. 1157-1182, 2003.

[24] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge & Data Engineering, pp. 491-502, 2005.

Converting protein sequence to image for classification with convolutional neural network

Reza Ahsan, PhD Student of Information Technology [1], Mansour Ebrahimi, Associate Professor [2], Rouhollah Dianat, Assistant Professor [3]

1- Faculty of Engineering, University of Qom, Qom, Iran, Email: ahsan@qom-iau.ac.ir
2- Faculty of Basic sciences, University of Qom, Qom, Iran, Iran, Email: mansour@future.edu
3- Faculty of Engineering, University of Qom, Qom, Iran, Email: rdianat@qom.ac.ir

Abstract

Since methods for sequencing machine learning sequences were not successful in classifying healthy and cancerous proteins, it is imperative to find a way to represent these sequences to classify healthy and ill individuals with deep learning approaches. In this study different methods of protein sequence representation for classification of protein sequence of healthy individuals and leukemia have been studied. Results showed that conversion of amino acid letters to one-dimensional feature vectors in classification of 2 classes was not successful and only one disease class was detected. By changing the feature vector to colored numbers, the accuracy of the healthy class recognition was slightly improved. The binary protein sequence representation method was more efficient than the previous methods with the initiative of sequencing the sequences in both one-dimensional and two-dimensional (image by Gabor filtering). Protein sequence representation as binary image was classified by applying Gabor filter with 100% accuracy of the protein sequence of healthy individuals and 98.6% protein sequence of those with leukemia. The findings of this study showed that the representation of protein sequence as binary image by applying Gabor filter can be used as a new effective method for representation of protein sequences for classification.

Keywords: Converting protein sequence to image, Gabor filter, Convolution Neural Network, Protein classification.