پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

شناسایی پپتیدهای ضدسرطان به کمک روش‌های یادگیری ماشین

یاسمین احمدی

استاد راهنما: دکتر باقر باباعلی

پایان‌نامه برای دریافت درجه کارشناسی
در رشته علوم کامپیوتر

پاییز ۱۴۰۱

چکیده

سرطان یک مشکل سلامتی مهم در سراسر جهان است و کشف پپتیدهای ضدسرطان (ACPs) چشم اندازی امیدوارکننده برای درمان سرطان ارائه می‌دهد. با این حال، شناسایی ACP ها با استفاده از روش‌های تجربی زمان‌بر و گران است و نیاز به روش‌های محاسباتی برای پیش‌بینی ACP را ایجاد می‌کند. روش‌های یادگیری ماشین برای چنین پیش‌بینی‌هایی پیشنهاد شده‌اند، اما عملکرد آن‌ها را می‌توان در مواردی با نمونه‌های محدود، بیشتر بهبود بخشید. این مطالعه یک مدل پیش‌بینی پپتیدهای ضدسرطان را پیشنهاد می‌کند که از تقویت داده‌ها برای افزایش عملکرد پیش‌بینی استفاده می‌کند. این مدل، ویژگی‌های نمایه باینری و ویژگی‌های AAindex را برای نمایش دنباله‌های پپتیدی ادغام می‌کند و نمونه‌ها را در فضای ویژگی برای بهره‌برداری بهتر از اطلاعات توالی‌های پپتیدی تقویت می‌کند.

با استفاده از پیش‌پردازش‌های لازم (استخراج ویژگی‌های پپتیدها، افزایش تعداد آن‌ها و انتخاب ویژگی و کاهش ابعاد)، نتایج امیدبخشی بدست آمد. برای مجموعه داده ACP240 دقت ۸۶.۲۵ درصدی و برای مجموعه داده ACP740 دقت ۸۷.۱۶ درصدی بدست آمد. مدل حاصل عملکرد بهتری در پیش‌بینی ACP در مقایسه با روش‌های موجود مقایسه شده دارد. این یافته‌ها نشان می‌دهد که این مدل، پتانسیل تسهیل کشف داروهای جدید درمان سرطان را دارد.

پیشگفتار

سرطان یک بیماری ویرانگر است که تأثیر زیادی بر سلامت عمومی در سراسر جهان دارد. سازمان بهداشت جهانی (WHO) گزارش داد که در سال ۲۰۲۰، ۳.۱۹ میلیون مورد جدید و ۱۰ میلیون مرگ ناشی از سرطان وجود داشته است و انتظار می‌رود این تعداد در سال‌های آینده افزایش یابد. علیرغم در دسترس بودن گزینه‌های مختلف درمانی مانند شیمی درمانی، پرتودرمانی و جراحی، اثربخشی این درمان‌ها به دلیل عوامل متعددی از جمله مقاومت دارویی، سمیت و عدم اختصاصیت محدود است [۱]. بنابراین، توسعه عوامل ضدسرطان جدید با کارایی بهتر و عوارض جانبی کمتر ضروری است.

پپتیدهای ضدسرطان (ACPs) پپتیدهای کوچک و کاتیونی هستند که نقش مهمی در سیستم ایمنی ذاتی موجودات زنده دارند. این پپتیدها توسط گونه‌های مختلفی از جمله باکتری‌ها، قارچ‌ها، گیاهان و حیوانات تولید می‌شوند و نشان داده شده است که فعالیت ضدسرطان گسترده‌ای دارند [۲]. در سال‌های اخیر، پپتیدهای ضدمیکروبی (AMPs) به دلیل توانایی آن‌ها در هدف قرار دادن انتخابی و کشتن سلول‌های سرطانی در حالی که سلول‌های طبیعی را حفظ می‌کنند، به عنوان عوامل ضدسرطان بالقوه مورد توجه قرار گرفته اند. علاوه بر این، ACPها مزایای متعددی نسبت به داروهای شیمی درمانی سنتی دارند، از جمله سمیت کم، پاکسازی سریع از بدن، و توانایی غلبه بر مقاومت دارویی.

با این حال، کشف و توسعه ACPها به عنوان عوامل ضدسرطان به دلیل پیچیدگی و تنوع ساختار و توالی آن‌ها چالش‌برانگیز است. روش‌های تجربی سنتی برای شناسایی و مشخص کردن ACPها زمانبر، کار فشرده و پرهزینه هستند. بنابراین، رویکردهای محاسباتی، مانند یادگیری ماشین (ML) برای پیش‌بینی و طراحی ACPها با فعالیت و گزینش‌پذیری بهبود یافته به کار گرفته شده‌اند [۳].

یادگیری ماشین زیرمجموعه‌ای از هوش مصنوعی است که شامل استفاده از الگوریتم‌ها برای تجزیه و تحلیل و یادگیری الگوها از مجموعه داده‌های بزرگ است [۴]. روش‌های مبتنی بر ML با موفقیت در زمینه‌های مختلف از جمله کشف دارو، ژنومیک و پروتئومیکس به کار گرفته شده‌اند [۵]. در زمینه پیش‌بینی ACP، الگوریتم‌های ML می‌توانند مقادیر زیادی از داده‌ها را تجزیه و تحلیل کنند و ویژگی‌هایی را شناسایی کنند که برای پیش‌بینی فعالیت و گزینش‌پذیری ACPها مهم هستند. چندین مطالعه کاربرد موفقیت آمیز روش‌های مبتنی بر ML را برای پیش‌بینی فعالیت ضدسرطان ACPها گزارش کرده اند [۳].

هدف این مقاله ارائه مدلی برای تشخیص پپتیدهای ضدسرطان جهت شناسایی درمان‌های جدید سرطان است. پپتیدهای ضدسرطان توانایی کشتن انتخابی سلول‌های سرطانی را دارند در حالی که سلول‌های سالم را تحت تاثیر قرار نمی دهند. با پیش‌بینی اینکه کدام پپتیدها احتمالاً خواص ضدسرطان دارند، محققان می‌توانند اهداف درمانی جدید را شناسایی کرده و داروهای جدیدی برای درمان سرطان بسازند. این رویکرد این پتانسیل را دارد که منجر به ایجاد درمان‌های موثرتر و کمتر سمی‌تر سرطان شود. علاوه بر این، پیش‌بینی پپتیدهای ضدسرطان ممکن است به توسعه پزشکی شخصی کمک کند، جایی که درمان‌ها بر اساس نیازهای فردی بیمار بر اساس ساختار ژنتیکی و ویژگی‌های خاص سرطان آن‌ها تنظیم می‌شوند.

فهرست مطالب

فصل ۱

مقدمه و توضیحات

۱.۱ پپتیدهای ضدسرطان

پپتیدهای ضدسرطان (ACPs) مجموعه‌ای از پپتیدهای کوتاه متشکل از ۱۰ تا ۶۰ آمینواسید هستند که می‌توانند از تکثیر یا مهاجرت سلول‌های تومور جلوگیری کنند یا تشکیل رگ‌های خونی تومور را سرکوب کنند و ایجاد مقاومت دارویی در برابر آن‌ها کمتر محتمل است [۵]. راه‌های مختلفی وجود دارد که ACPها می‌توانند علیه سلول‌های سرطانی عمل کنند. برخی از ACPها می‌توانند مستقیماً با ایجاد اختلال در غشای سلولی یا تداخل در فرآیندهای سلولی کلیدی، مرگ سلول‌های سرطانی را القا کنند. سایر ACPها می‌توانند رشد و تکثیر سلول‌های سرطانی را با مسدود کردن مسیرهای مهم سیگنالینگ یا ترویج آپوپتوز (مرگ برنامه‌ریزی شده سلولی) مهار کنند. بدست آوردن پپتیدها از منابع طبیعی مختلف امکان‌پذیر است.

- پپتیدهایی که ضدتومور یا ضدسرطان هستند، می‌توانند مشتقات گیاهی باشند، مانند پپتید پلی ساکارید گانودرما لوسیدوم، که ضد رگ‌زایی است [۷]، یا مشتق‌شده از حیوانات، مانند پپتید ناتریورتیک دهلیزی.
- پپتیدهای ضدسرطان را می‌توان از منابع طبیعی مختلفی بدست آورد، از جمله گیاهان، حیوانات و میکروارگانیسم‌ها.
- بسیاری از پپتیدهای مشتق‌شده از گیاهان دارای خواص ضدسرطان هستند. به عنوان مثال،

نشان داده شده است که پپتید گیاهی سیکلوتید از رشد سلول‌های سرطانی سینه و پروستات جلوگیری می‌کند.

- برخی از حیوانات، مانند دوزیستان، پپتیدهایی تولید می‌کنند که دارای فعالیت ضدسرطان هستند. به عنوان مثال، پپتید magainin که در پوست قورباغه‌های پنجه‌دار آفریقایی یافت می‌شود، خواص ضدسرطان قوی دارد.

- باکتری‌ها و قارچ‌ها نیز منابع پپتیدهای ضدسرطان هستند. به عنوان مثال، پپتید grami-cidin S که توسط باکتری باسیلوس برویس تولید می‌شود، نشان داده شده است که از رشد چندین نوع سلول سرطانی جلوگیری می‌کند.

- علاوه بر این منابع طبیعی، محققان پپتیدهای مصنوعی نیز تولید کرده‌اند که دارای خواص ضدسرطان هستند. این پپتیدها برای تقلید از ساختار و عملکرد پپتیدهای طبیعی طراحی شده‌اند و می‌توانند برای هدف قرار دادن انواع خاصی از سلول‌های سرطانی سفارشی عمل کنند.

۲.۱ تاریخچه کشف پتیدهای ضدسرطان

کشف پپتیدهایی با فعالیت ضد باکتریایی و ضدسرطان به اوایل قرن بیستم باز می‌گردد. به طور خاص، کشف لیزوزیم در سال ۱۹۲۲، پپتیدی که می‌تواند دیواره‌های سلولی باکتری را تجزیه کند، دری را برای مطالعه سایر پپتیدهای دارای فعالیت ضد میکروبی باز کرد. بعدها، در دهه ۱۹۶۰، ملیتین، یک پپتید موجود در زهر زنبور عسل، به عنوان اولین پپتید ضدسرطان شناخته شد. محققان دریافتند که ملیتین می‌تواند به غشای سلول‌های سرطانی متصل شود و باعث مرگ سلولی شود [۸].
از زمان کشف ملیتین، بسیاری از پپتیدهای ضدسرطان دیگر از منابع طبیعی مانند گیاهان، حیوانات و میکروارگانیسم‌ها شناسایی شده اند. به عنوان مثال، نشان داده شده‌است که سیکلوتید پپتید گیاهی از رشد سلول‌های سرطانی سینه و پروستات جلوگیری می‌کند. در حیوانات، پپتیدهایی مانند magainin که در پوست قورباغه‌های پنجه‌دار آفریقایی یافت می‌شود، فعالیت ضدسرطان قوی نشان داده‌اند. همچنین مشخص شده است که باکتری‌ها و قارچ‌ها پپتیدهایی با خواص ضدسرطان مانند پپتید gramicidin S تولید شده توسط Bacillus brevis تولید می‌کنند [۹].
کشف پپتیدهای ضدسرطان با پیشرفت تکنولوژی و روش‌های علمی کمک کرده است. غربالگری با توان بالا به محققان این امکان را داده‌است که به سرعت تعداد زیادی از پپتیدها را برای فعالیت

ضدسرطان آزمایش کنند. علاوه بر این، پپتیدهای مصنوعی توسعه یافته‌اند که برای تقلید از ساختار و عملکرد پپتیدهای طبیعی طراحی شده‌اند. این پپتیدهای مصنوعی را می‌توان برای هدف قرار دادن انواع خاصی از سلول‌های سرطانی سفارشی کرد و در مطالعات پیش‌بالینی موثر نشان داده شده‌است [۱۰].

۳.۱ ساختار سه بعدی پپتیدهای ضدسرطان

ساختار سه بعدی ACPها نقش مهمی در عملکرد و ویژگی آن‌ها ایفا می‌کند. ACPها انواع ساختارهای ثانویه مانند آلفا مارپیچ، صفحات بتا و حلقه‌ها را اتخاذ می‌کنند که توسط پیوندهای هیدروژنی درون‌مولکولی، برهمکنش‌های الکترواستاتیکی و برهمکنش‌های آبگریز تثبیت می‌شوند.

ACPها می‌توانند طیف وسیعی از ساختارها از جمله آلفا_مارپیچ‌ها، صفحات بتا، حلقه‌ها و ترکیب‌های گسترده را اتخاذ کنند. این ساختارها توسط فعل و انفعالات مختلفی مانند پیوندهای هیدروژنی، برهمکنش‌های الکترواستاتیکی و برهمکنش‌های آبگریز تثبیت می‌شوند. ساختار کلی ACPها توانایی آن‌ها را در تعامل با سلول‌های سرطانی و به طور خاص هدف قرار دادن سلول‌های سرطانی بدون تأثیر بر سلول‌های سالم تعیین می‌کند [۱۱].

- ساختار سه بعدی ACPها به آن‌ها اجازه می‌دهد تا از طریق برهمکنش‌های مولکولی خاص با سلول‌های سرطانی تعامل داشته باشند. نشان داده شده‌است که ACPها از طریق برهمکنش‌های الکترواستاتیکی بین باقی‌مانده‌های دارای بار مثبت در پپتید و لیپیدهای دارای بار منفی در غشاء به غشای پلاسمایی سلول‌های سرطانی متصل می‌شوند. علاوه بر این، ACPها ممکن است وارد غشاء شوند و ساختار آن را مختل کنند که منجر به مرگ سلولی شود. ساختار سه‌بعدی ACPها همچنین آن‌ها را قادر می‌سازد تا با اهداف درون سلولی مانند DNA ،RNA و پروتئین‌ها از طریق سایت‌های اتصال خاص تعامل داشته باشند.
- ساختار سه‌بعدی ACPها آن‌ها را قادر می‌سازد به گیرنده‌ها یا آنزیم‌های خاصی که بیش از حد بیان می‌شوند یا منحصر به سلول‌های سرطانی هستند متصل شوند. به عنوان مثال، برخی از ACPها میتوکندری سلول‌های سرطانی را هدف قرار می‌دهند و عملکرد آن‌ها را مختل می‌کنند که منجر به مرگ سلولی می‌شود. ساختار ACPها به آن‌ها اجازه می‌دهد تا به غشای سلولی نفوذ کنند و به طور خاص میتوکندری را هدف قرار دهند.

- ساختار سه‌بعدی ACPها همچنین ثبات و مقاومت آن‌ها را در برابر پروتئازها تعیین می‌کند. ACPها ممکن است حاوی پیوندهای دی سولفیدی یا سایر پیوندهای کووالانسی باشند که به پایداری آن‌ها کمک می‌کند، یا ممکن است دارای یک هسته آبگریز فشرده باشند که آن‌ها را در برابر پروتئولیز مقاوم می‌کند. علاوه بر این، ACPها ممکن است برای افزایش پایداری و فراهمی زیستی خود، مانند استفاده از پپتیدهای حلقوی یا همجوشی با سایر پروتئین‌ها یا پپتیدها، مهندسی شوند.

- ACPها همچنین ممکن است دارای ویژگی‌های ساختاری باشند که آن‌ها را از تخریب پروتئولیتیک محافظت می‌کند و آن‌ها را به عنوان عوامل درمانی پایدارتر و موثرتر می‌کند. برخی از ACPها حاوی پیوندهای دی سولفیدی یا سایر پیوندهای کووالانسی هستند که پایداری آن‌ها را افزایش می‌دهد، در حالی که برخی دیگر دارای یک هسته آبگریز فشرده هستند که آن‌ها را در برابر پروتئازها مقاوم می‌کند. علاوه بر این، ACPها را می‌توان به صورت حلقوی یا ذوب شده با سایر پپتیدها طراحی کرد تا پایداری و فراهمی زیستی آن‌ها را افزایش دهد [۱۲].

- به طور خلاصه، ساختار سه‌بعدی ACPها برای عملکرد، ویژگی، پایداری و مقاومت آن‌ها در برابر پروتئازها ضروری است. ACPها ساختارهای ثانویه مختلفی را اتخاذ می‌کنند که به آن‌ها اجازه می‌دهد با سلول‌های سرطانی و اهداف درون سلولی از طریق برهمکنش‌های مولکولی خاص تعامل داشته باشند و ممکن است حاوی پیوندهای کووالانسی باشند یا یک هسته آبگریز فشرده داشته باشند که به پایداری آن‌ها کمک می‌کند. درک بهتر ساختار سه بعدی ACPها، طراحی و توسعه عوامل ضدسرطان قوی تر و انتخابی را تسهیل می‌کند.

- ساختار ۳ بعدی ACPها برای عملکرد آن‌ها به عنوان عوامل ضدسرطان بسیار مهم است. ACPها می‌توانند ساختارهای مختلفی را اتخاذ کنند که توانایی آن‌ها را در تعامل با سلول‌های سرطانی و اهداف مولکولی خاص تعیین می‌کند. ساختار ACPها همچنین به پایداری و مقاومت آن‌ها در برابر پروتئازها کمک می‌کند و آن‌ها را به عنوان عوامل درمانی مؤثرتر می‌کند.

۴.۱ اهداف پپتیدهای ضدسرطان و مکانیسم فعالیت

پپتیدهای ضدسرطان توالی آمینواسیدها کوتاهی هستند که توانایی کشتن انتخابی سلول‌های سرطانی را دارند و در عین حال سلول‌های سالم را حفظ می‌کنند. این پپتیدها از منابع مختلفی از جمله حیوانات، گیاهان و باکتری‌ها شناسایی شده اند. پپتیدهای ضدسرطان اثرات سیتوتوکسیک خود را از طریق مکانیسم‌های مختلفی از جمله مختل کردن غشای سلولی، القای آپوپتوز و مهار رگ‌زایی اعمال می‌کنند.

یکی از مکانیسم‌های اثر پپتیدهای ضدسرطان، اختلال در غشاء است. بسیاری از پپتیدهای ضدسرطان دارای نسبت بالایی از آمینواسیدها با بار مثبت هستند که آن‌ها را قادر می‌سازد با غشای سلولی با بار منفی سلول‌های سرطانی تعامل داشته باشند. این فعل و انفعال می‌تواند منجر به تشکیل منافذ در غشای سلولی شود و عملکرد طبیعی سلول را مختل کرده و در نهایت منجر به مرگ سلولی شود. علاوه بر این، برخی از پپتیدهای ضدسرطان نشان داده شده است که با پروتئین‌های غشایی خاص تعامل دارند که منجر به اثرات پایین دستی می‌شود که منجر به مرگ سلولی می‌شود [۸].

مکانیسم دیگر اثر پپتیدهای ضدسرطان، القای آپوپتوز است. آپوپتوز یک شکل برنامه ریزی شده از مرگ سلولی است که زمانی رخ می‌دهد که سلول‌ها آسیب ببینند یا دیگر مورد نیاز نباشند. نشان داده شده است که بسیاری از پپتیدهای ضدسرطان باعث القای آپوپتوز در سلول‌های سرطانی و در نتیجه مرگ آن‌ها می‌شوند. این پپتیدها می‌توانند مسیرهای خاصی را در داخل سلول فعال کنند که منجر به آپوپتوز می‌شود، یا می‌توانند به پروتئین‌های خاصی که آپوپتوز را تنظیم می‌کنند، متصل شوند و منجر به فعال شدن آن‌ها شود [۱۳].

در نهایت، پپتیدهای ضدسرطان می‌توانند رگ‌زایی، فرآیندی که طی آن عروق خونی جدید تشکیل می‌شوند، مهار کنند. رگ‌زایی برای رشد و گسترش سرطان ضروری است، زیرا تومور را قادر می‌سازد مواد مغذی و اکسیژن مورد نیاز خود را برای زنده ماندن و رشد دریافت کند. پپتیدهای ضدسرطان می‌توانند با تعامل با پروتئین‌های خاص درگیر در فرآیند، رگ‌زایی را مهار کنند و منجر به مهار آن‌ها شوند.

از نظر مکانیسم پیوند، پپتیدهای ضدسرطان می‌توانند با سلول‌های سرطانی از طریق برهمکنش‌های الکترواستاتیک و آبگریز تعامل کنند. آمینواسیدها با بار مثبت در این پپتیدها می‌توانند با مولکول‌های دارای بار منفی روی غشای سلول سرطانی از طریق برهمکنش‌های الکترواستاتیکی تعامل داشته باشند، در حالی که مناطق آبگریز پپتیدها می‌توانند با مناطق آبگریز غشاء تعامل کنند. علاوه بر این، برخی از پپتیدهای ضدسرطان نشان داده شده‌است که با پروتئین‌های خاصی در سطح سلول‌های سرطانی تعامل دارند که منجر به اثرات پایین دستی می‌شود که منجر به مرگ سلول می‌شود [۸].

۵.۱ مکانیسم‌های تخریب غشا توسط پپتیدهای ضدسرطان

پپتیدهای ضدسرطان توانایی کشتن انتخابی سلول‌های سرطانی را دارند و در عین حال سلول‌های سالم را حفظ می‌کنند. یکی از مکانیسم‌هایی که توسط آن‌ها به این امر دست می‌یابند، مختل کردن غشای سلولی سلول‌های سرطانی است. چندین استراتژی وجود دارد که پپتیدهای ضدسرطان برای از بین بردن غشاء استفاده می‌کنند.

- یک استراتژی که توسط پپتیدهای ضدسرطان برای از بین بردن غشا استفاده می‌شود، تشکیل منافذ است. بسیاری از پپتیدهای ضدسرطان دارای نسبت بالایی از آمینواسیدها با بار مثبت هستند که آن‌ها را قادر می‌سازد با غشای سلولی با بار منفی سلول‌های سرطانی تعامل داشته باشند. این فعل و انفعال می‌تواند منجر به تشکیل منافذ در غشای سلولی شود و عملکرد طبیعی سلول را مختل کرده و در نهایت منجر به مرگ سلولی شود. نمونه‌هایی از پپتیدهای ضدسرطان که از طریق تشکیل منافذ عمل می‌کنند [۸] عبارتند از ملیتین، ماگنین و پارداکسین.

- استراتژی دیگری که توسط پپتیدهای ضدسرطان برای از بین بردن غشا استفاده می‌شود، همجوشی غشا است. در این فرآیند، پپتید ضدسرطان با غشای سلول سرطانی ترکیب می‌شود و منجر به اختلال در غشاء و در نهایت مرگ سلول می‌شود. نمونه‌هایی از پپتیدهای ضدسرطان که از طریق همجوشی غشایی عمل می‌کنند [۹] عبارتند از تاکیپلسین و دیفنسین.
در نهایت، برخی از پپتیدهای ضدسرطان با تشکیل میسل‌ها، که توده‌های کروی مولکول‌های آمفیفیلیک هستند، عمل می‌کنند. این میسل‌ها می‌توانند با غشای سلول سرطانی تعامل داشته باشند و منجر به از هم گسیختگی غشاء و در نهایت مرگ سلولی شوند. نمونه‌هایی از پپتیدهای ضدسرطان که از طریق تشکیل میسل عمل می‌کنند عبارتند از S gramicidin و خانواده پپتید temporin [۱۴].
به طور کلی، توانایی پپتیدهای ضدسرطان برای تخریب غشای سلول‌های سرطانی یک عامل کلیدی در فعالیت ضدسرطان آن‌ها است. با ایجاد اختلال در غشاء، این پپتیدها می‌توانند در نهایت منجر به مرگ سلول‌های سرطانی شوند و در عین حال سلول‌های سالم را حفظ کنند.

فصل ۲

رویکرد مبتنی بر یادگیری ماشین

۱.۲ تاریخچه پژوهش‌های مبتنی بر یادگیری ماشین برروی پپتیدهای ضدسرطان

۱.۱.۲ اصول یادگیری ماشین

در تکنیک یادگیری ماشین با استفاده از ریاضیات، آمار و علم رایانه، از داده آموخته می‌شود. روش‌های یادگیری ماشین را می‌توان به دو دسته Supervised و Unsupervised تقسیم بندی کرد. در روش Supervised، یک مدل پیش‌بینی کننده براساس داده‌هایی ساخته می‌شود که اطلاعات واقعی پارامتری که قرار است پیش‌بینی شود در دادهها وجود دارد. هر رکورد داده با تعدادی متغیر ورودی یا ویژگی به عنوان ورودی و خروجی فرایند شناخته می‌شود. در روش ،Unsupervised روندهایی مخفی از میان داده‌ها استنتاج می‌شود. غالب مطالعات یادگیری ماشین صورت گرفته با استفاده از روش Supservised می‌باشد. با این حال، پژوهش‌های مبتنی بر روش Unsupervised نیز رشد قابل توجهی داشته‌اند [۲۶].

چارچوب ریاضی و آماری یادگیری ماشین قرن‌ها پیش، بسیار قبل از اختراع کامپیوترها، وجود داشته است. ایده‌ها و عناصری که امروزه در یادگیری ماشین بیشتر شناخته شده هستند، از جمله قضیه ،'Bayes تحلیل مؤلفه‌های اصلی ، رگرسیون خطی چندگانه ، برازش حداقل

مربعات و زنجیره‌های مارکوف، قبل از سال ۱۹۵۰ توسط ریاضیدانان ایجاد شدند [۱۷]. کار پیشگام آلن تورینگ بر روی ماشین تورینگ در سال ۱۹۵۰ [۱۶] منجر به توسعه اولین شبکه عصبی مصنوعی (ANN) شد [۱۸]. پس از اختراع رایانه مدرن، تحقیقات در زمینه یادگیری ماشین با توسعه بسیاری از روش‌های مدرن مورد استفاده امروزی، از جمله رگرسیون حداقل مربعات جزئی، شبکه‌های عصبی تکراری، مدل‌های مارکوف پنهان (HMM)، ماشین‌های بردار پشتیبانی (SVM) و جنگل‌های تصادفی بشدت گسترش یافت [۱۵]. در سال‌های اخیر، پیشرفت‌ها در الگوریتم‌های آموزش مقیاس‌پذیر و در دسترس بودن مجموعه‌های داده بزرگ باعث تجدید علاقه به شبکه‌های عصبی مصنوعی با معماری‌های شبکه عمیق شده است که قادر به انجام وظایفی مانند تشخیص دست خط [۱۹]، تشخیص سرطان [۲۰]، طبقه‌بندی HIV [۲۱]، تشخیص چهره [۲۲] و فیلتر هرزنامه [۲۳] هستند.

۲.۱.۲ کاربرد یادگیری ماشین در پژوهش‌های مرتبط با پپتیدهای ضدسرطان

تنوع توالی‌ها و ساختارهای پپتیدهای ضدسرطان درکنار زمان و هزینه مرتبط با طراحی آزمایش، تولید و آزمایش پپتیدها نامزد، مانع از غربالگری جامع فضای توالی پپتیدی بصورت تجربی می‌شود. بنابراین، اولین مدل‌های یادگیری ماشین، مدل‌های کمی رابطه ساختار-فعالیت (QSAR) بودند که در غربالگری و بهینه‌سازی تعدادی از توالی‌های بالقوه کارامد برای ارزیابی تجربی، مفید واقع شدند. هدف مدل‌های QSAR استفاده از توصیف‌گرهای فیزیکی و شیمیایی برای پیش‌بینی فعالیتی بیولوژیکی از یک مولکول هستند که معمولا اندازه‌گیری یا محاسبه آن گران و یا زمان‌بر است. در مقابل، بسیاری از خواص فیزیکی-شیمیایی یک پپتید را می‌توان به طور مستقیم و کم هزینه از توالی آمینواسیدی آن استخراج کرد [۲۴،۲۵]. مدل‌ها بر روی مجموعه‌های اطلاعاتی گردآوری شده از داده‌های تجربی، آموزش داده شده و تأیید می‌شوند، سپس در غربالگری in silico برای شناسایی نامزدهای جدید با فعالیت زیستی مورد نظر به کار گرفته می‌شوند. این روش بر یادگیری آماری برای استنباط روابط تجربی بین ویژگی‌های فیزیکی-شیمیایی و فعالیت بیولوژیکی متکی است، و بنابراین، مشروط به یک رابطه بنیادی بین (زیر مجموعه‌ای از) توصیف‌گرها، ظرفیت مدل یادگیری ماشین برای کشف و رمزگذاری این رابطه در یک عبارت ریاضی و مجموعه‌های داده آموزش به اندازه کافی بزرگ و متنوع برای تولید مدل‌های پیش‌بینی قوی می‌باشد [۲۴،۲۶].

- استفاده از یادگیری ماشین در پیش‌بینی پپتیدهای ضدسرطان یک زمینه تحقیقاتی نسبتاً جدید است که اولین مطالعات آن به اوایل دهه ۲۰۰۰ بازمی‌گردد. در سال ۲۰۰۶، محققان از الگوریتم ماشین بردار پشتیبان(SVM) برای طبقه‌بندی پپتیدها به عنوان ضدسرطان یا غیر ضدسرطان بر اساس ویژگی‌های توالی آن‌ها استفاده کردند. آن‌ها به دقت بالایی در پیش‌بینی فعالیت مجموعه داده‌ای از ۱۹۹ پپتید دست یافتند [۲۷].

- در سال ۲۰۱۰، مطالعه دیگری از تکنیک‌های یادگیری ماشین برای شناسایی خواص فیزیکوشیمیایی پپتیدهای ضدسرطان استفاده کرد. آن‌ها چندین مدل طبقه‌بندی را با استفاده از مجموعه داده‌ای از ۹۹ پپتید آموزش دادند و به دقت تا ۸۷ درصد در پیش‌بینی فعالیت آن‌ها دست یافتند [۲۸].

- در سال ۲۰۱۵، محققان از ترکیبی از الگوریتم‌های یادگیری ماشین و مدل‌سازی QSAR برای پیش‌بینی فعالیت ضدسرطان پپتیدها استفاده کردند. آن‌ها ویژگی‌های کلیدی پپتیدهای ضدسرطان را شناسایی کردند و از آن‌ها برای آموزش چندین مدل طبقه‌بندی استفاده کردند و به دقت بالایی در پیش‌بینی فعالیت آن‌ها دست یافتند. آن‌ها همچنین از غربالگری مجازی برای شناسایی چندین پپتید جدید با فعالیت ضدسرطان بالقوه استفاده کردند [۲۹].

- در سال ۲۰۱۹، یک مطالعه از الگوریتم‌های یادگیری عمیق برای پیش‌بینی فعالیت مجموعه داده‌ای از ۱۱۳۹ پپتید استفاده کرد. آن‌ها به دقت بالایی در پیش‌بینی فعالیت خود دست یافتند و ویژگی‌های کلیدی پپتیدهایی را که با فعالیت ضدسرطان آن‌ها مرتبط بود شناسایی کردند.

- اخیراً، چندین مطالعه از الگوریتم‌های یادگیری ماشین برای طراحی و بهینه‌سازی پپتیدهای ضدسرطان جدید استفاده کرده‌اند. به عنوان مثال، در سال ۲۰۲۰، محققان از ترکیبی از یادگیری ماشین و طراحی پپتید de novo برای توسعه پپتیدهای جدید با فعالیت ضدسرطان قوی و گزینش‌پذیری بالا در برابر سلول‌های سرطانی استفاده کردند [۳۰].

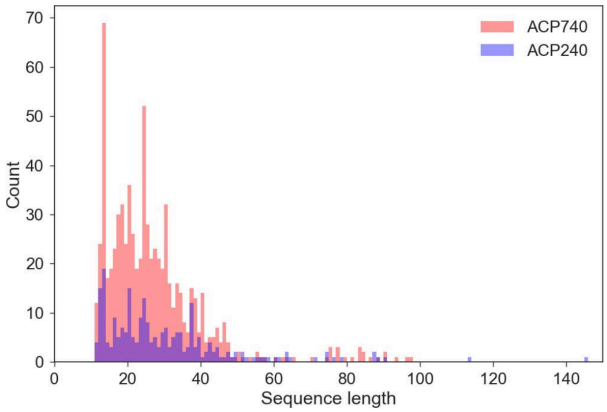
به طور کلی، استفاده از یادگیری ماشین در پیش‌بینی پپتیدهای ضدسرطان در طول سال‌ها تکامل یافته است، با پیشرفت در توسعه الگوریتم و افزایش در دسترس بودن مجموعه داده‌های بزرگ که به پیشرفت در این زمینه کمک می‌کند.

فصل ۳

آماده‌سازی و آموزش

۱.۳ مجموعه‌های داده

ما در این مقاله از دو مجموعه داده ACP240 و ACP740 استفاده می‌کنیم. مجموعه داده‌های ACP240 و ACP740 مجموعه‌ای از پپتیدهای ضدسرطان هستند که به طور گسترده برای کاربردهای درمانی بالقوه آن‌ها در درمان سرطان مورد مطالعه قرار گرفته اند. این مجموعه داده‌ها از توالی‌های آمینواسید کوتاه تشکیل شده‌اند که به طور تجربی دارای فعالیت ضدتوموری هستند.



تصویر بالا نمایانگر پراکندگی پپتیدها با رشته طول‌های متفاوت، در دو مجموعه داده

ACP240 و ACP740 می‌باشد. به خوبی می‌توان مشاهده کرد که فراوانی پپتیدها در بازه ۱۰ تا ۴۰ بسیار بیشتر است.

۱.۱.۳ تاریخچه و کارهای قبلی روی مجموعه داده

مجموعه داده ACP240 اولین بار در سال ۲۰۰۹ توسط وانگ و همکاران[۳۱] منتشر شد. مجموعه داده شامل ۲۴۰ پپتید است که طول هر کدام ۱۲ آمینواسید است. این پپتیدها با استفاده از فناوری نمایش فاژ شناسایی شدند و نشان داده شده است که دارای فعالیت ضدتوموری در برابر انواع مختلف سرطان از جمله سرطان سینه، ریه و پروستات هستند. مجموعه داده ACP240 در مطالعات متعددی برای توسعه مدل‌های یادگیری ماشین برای پیش‌بینی فعالیت ضدتوموری پپتیدهای جدید استفاده شده است.
مجموعه داده ACP740 بعداً توسط همین گروه در سال ۲۰۱۲[۳۲] توسعه یافت. این مجموعه داده شامل ۷۴۰ پپتید است که طول هر کدام ۱۳ آمینواسید است. مانند مجموعه داده ACP240، این پپتیدها با استفاده از فناوری نمایش فاژ شناسایی شدند و به طور تجربی برای داشتن فعالیت ضدتومور در برابر انواع مختلف سرطان تایید شده اند. مجموعه داده ACP740 در مطالعات مختلفی برای توسعه مدل‌های پیش‌بینی پپتیدهای ضدسرطان و همچنین برای بدست آوردن بینش در مورد مکانیسم‌های زیربنایی فعالیت ضدتوموری آن‌ها استفاده شده است.
چندین مطالعه از این مجموعه داده‌ها برای توسعه مدل‌های یادگیری ماشین برای پیش‌بینی فعالیت ضدتوموری پپتیدهای جدید استفاده کرده‌اند. برای مثال، ژانگ و همکاران.[۳۳] از مجموعه داده ACP240 برای توسعه یک مدل ماشین بردار پشتیبان (SVM) استفاده کرد که می‌تواند به طور دقیق فعالیت ضدتوموری پپتیدهای جدید را پیش‌بینی کند. به طور مشابه، ما و همکاران.[۳۴] از مجموعه داده ACP740 برای توسعه یک مدل جنگل تصادفی استفاده کرد که می‌تواند فعالیت ضدتوموری پپتیدها را با دقت بیش از ۹۰٪ پیش‌بینی کند.
علاوه بر مطالعات یادگیری ماشین، مجموعه داده‌های ACP240 و ACP740 در انواع دیگر مطالعات برای بدست آوردن بینشی در مورد مکانیسم‌های نهفته در فعالیت ضدتوموری این پپتیدها استفاده شده است. به عنوان مثال، هو و همکاران.[۳۵] از مجموعه داده ACP240 برای بررسی نقش آبگریزی در فعالیت ضدتوموری پپتیدها استفاده کرد، در حالی که یان و همکاران.[۳۶] از مجموعه داده ACP740 برای بررسی اثر طول پپتید بر

فعالیت ضدتومور استفاده کرد.
به طور کلی، مجموعه داده‌های ACP240 و ACP740 منابع ارزشمندی برای محققانی هستند که پپتیدهای ضدسرطان را مطالعه می‌کنند. این مجموعه داده‌ها به‌طور گسترده مورد مطالعه و اعتبارسنجی قرار گرفته‌اند و در مطالعات مختلف برای توسعه مدل‌های پیش‌بینی‌کننده برای فعالیت ضدتومور و بدست آوردن بینشی در مورد مکانیسم‌های زیربنایی فعالیت ضدتوموری این پپتیدها استفاده شده‌اند.

۲.۱.۳ دریافت مجموعه داده استفاده شده

برای دسترسی به مجموعه داده استفاده شده می‌توان به مخزن گیت‌هاب مقاله ژانگ و همکاران.[۳۷] مراجعه و خود مجموعه داده را دریافت کرد. لینک دسترسی به مخزن[۳۸].

۲.۳ آماده‌سازی داده‌ها

۱.۲.۳ حذف پپتیدهای دارای C-Terminal و N-Terminal غیر طبیعی

در زمینه پیش‌بینی ساختار ثانویه پروتئین‌ها، باقی‌مانده‌های C ترمینال و N ترمینال غیرطبیعی می‌توانند به طور بالقوه بر دقت پیش‌بینی تأثیر بگذارند.
باقی مانده N ترمینال یک پروتئین اولین آمینواسید در زنجیره پلی پپتیدی است، در حالی که باقی مانده C ترمینال آمینواسید نهایی در زنجیره است. این باقیمانده‌ها نقش مهمی در چین خوردگی و پایداری پروتئین و همچنین در عملکرد بیولوژیکی آن دارند.
با این حال، در برخی موارد، باقی مانده‌های ترمینال N و C ترمینال ممکن است اصلاح یا کوتاه شوند، که می‌تواند بر روی تاخوردگی و پایداری پروتئین تأثیر بگذارد. علاوه بر این، برخی از پپتیدها ممکن است تغییرات یا بقایای غیرمعمولی در انتهای خود داشته باشند که می‌تواند بر دقت پیش‌بینی تأثیر بگذارد.
بنابراین، در پیش‌بینی ساختار ثانویه یک پروتئین، ممکن است لازم باشد پپتیدهایی با باقیمانده‌های N ترمینال و C ترمینال غیرطبیعی حذف شوند تا اطمینان حاصل شود که پیش‌بینی براساس مجموعه‌ای ثابت از توالی‌های آمینواسید است که نماینده ساختار طبیعی

هستند. از پروتئین با حذف این پپتیدهای غیرعادی، الگوریتم پیش‌بینی می‌تواند بر توالی‌ها و ساختارهای معمولی پروتئین تمرکز کند و منجر به پیش‌بینی‌های دقیق‌تر شود.

۲.۲.۳ انتخاب هفت آمینواسید ابتدایی

توالی پپتیدی یک زنجیره خطی از آمینواسیدها است که از طریق پیوندهای پپتیدی به یکدیگر متصل می‌شوند. این توالی‌ها بلوک‌های سازنده پروتئین‌ها هستند و اغلب با استفاده از کد استاندارد تک حرفی نشان داده می‌شوند که از حروف انگلیسی برای نشان دادن هر یک از ۲۰ آمینواسید مختلف استفاده می‌کند. حال برای پاکسازی این نوع داده، از پپتید، هفت آمینواسید اول برمیداریم به BPF تبدیل شد.
همانطور که پیشتر گفته شد، فقط از هفت آمینواسید ابتدای هر پپتید استفاده شد. دلیل این امر این است که:

- کارایی محاسباتی: پردازش توالی‌های پپتیدی طولانی می‌تواند از نظر محاسباتی گران باشد، به‌ویژه زمانی که از الگوریتم‌های یادگیری ماشین استفاده می‌شود که به تعداد زیادی محاسبات نیاز دارد. تنها با استفاده از هفت آمینواسید اول، با این کار هزینه محاسباتی را کاهش دادیم و تجزیه و تحلیل را امکان‌پذیرتر شده است.
- محتوای اطلاعاتی: چند آمینواسید اول یک توالی پپتیدی می‌تواند اطلاعات مهمی در مورد خواص و فعالیت آن ارائه دهد. به عنوان مثال، آمینواسید N ترمینال یک پپتید می‌تواند بر پایداری و حلالیت آن تأثیر بگذارد، در حالی که توالی چند آمینواسید اول می‌تواند ساختار ثانویه آن را تعیین کند. با تمرکز بر هفت آمینواسید اول، می‌توان آموزنده‌ترین بخش توالی پپتیدی را به تصویر کشیده باشند.
- مقایسه با سایر مطالعات: مطالعات قبلی بر روی پپتیدهای ضدسرطان نیز بر روی چند آمینواسید اول توالی پپتیدی متمرکز شده است. با استفاده از همین رویکرد، نتایج با مطالعات قبلی مقایسه شد.

به طور کلی، تصمیم به استفاده از هفت آمینواسید اول بر اساس ترکیبی از کارایی محاسباتی، محتوای اطلاعاتی و مقایسه با مطالعات قبلی است. با این حال، شایان ذکر است که مطالعات

دیگر ممکن است از رویکردهای متفاوتی استفاده کنند و توالی‌های پپتیدی طولانی‌تر یا بخش‌های مختلف توالی را در نظر بگیرند.

۳.۲.۳ تبدیل به BPF

مرحله بعد تبدیل هفت آمینواسید ابتدایی به BPF است. BPF یک روش محاسباتی است که برای پیش‌بینی فعالیت بیولوژیکی پپتیدها استفاده می‌شود. BPF نوعی از مدل رابطه کمی ساختارفعالیت (QSAR) است که از الگوریتم‌های یادگیری ماشین برای تجزیه و تحلیل توالی آمینواسیدها در پپتیدها و پیش‌بینی فعالیت آن‌ها در برابر یک هدف خاص استفاده می‌کند.
BPF با تبدیل توالی آمینواسید یک پپتید به یک کد باینری عمل می‌کند، جایی که هر موقعیت در دنباله با یک رقم دوتایی (۰ یا ۱) نشان داده می‌شود. این کد باینری سپس به عنوان ورودی برای الگوریتم یادگیری ماشین استفاده می‌شود که فعالیت پپتید را بر اساس مجموعه‌ای از داده‌های آموزشی پیش‌بینی می‌کند.
برای تبدیل یک پپتید به BPF، به هر آمینواسید در پپتید یک کد باینری منحصر به فرد بر اساس ویژگی‌های فیزیکوشیمیایی آن مانند آبگریزی، بار و اندازه اختصاص داده می‌شود. سپس کدهای باینری به هم متصل می‌شوند تا یک توالی دوتایی تشکیل دهند که نشان دهنده پپتید است. سپس این توالی باینری به عنوان ورودی برای الگوریتم یادگیری ماشین برای پیش‌بینی فعالیت بیولوژیکی پپتید استفاده می‌شود.
مطالعات متعددی اثربخشی BPF را در پیش‌بینی فعالیت بیولوژیکی پپتیدها نشان داده‌اند. به عنوان مثال، در مطالعه‌ای توسط چن و همکاران. [۳۹](۲۰۱۸)، BPF برای پیش‌بینی فعالیت ضدسرطان مجموعه‌ای از پپتیدها استفاده شد و نتایج نشان داد که BPF از نظر دقت پیش‌بینی بهتر از سایر مدل‌های QSAR عمل می‌کند.
مطالعه دیگری توسط لی و همکاران. (۲۰۱۹)[۴۰] از BPF برای پیش‌بینی فعالیت مجموعه‌ای از پپتیدهای ضد میکروبی استفاده کرد و نتایج نشان داد که BPF دقت بالاتری نسبت به سایر مدل‌های QSAR دارد.
به طور کلی، BPF روشی امیدوارکننده برای پیش‌بینی فعالیت بیولوژیکی پپتیدها است و این پتانسیل را دارد که کشف درمان‌های پپتیدی جدید را تسریع کند.

۳.۳ محاسبه ویژگی‌های پپتیدهای ضدسرطان

۱.۳.۳ ویژگی‌های پیشفرض موجود در AAindex

فایل AAindex یک فایل داده‌ای است که در بیوانفورماتیک و زیست شناسی محاسباتی برای پیش‌بینی فعالیت ضدسرطان پپتیدها بر اساس توالی آمینواسید آن‌ها استفاده می‌شود. این فایل حاوی خواص فیزیکوشیمیایی از پیش‌محاسبه شده آمینواسیدها است که برای پیش‌بینی فعالیت پپتیدها در برابر سلول‌های سرطانی مهم هستند. ویژگی‌های موجود در فایل می‌تواند شامل پارامترهایی مانند آب گریزی، قطبیت، شارژ و موارد دیگر باشد.

خواص فیزیکوشیمیایی آمینواسیدها برای درک رفتار پپتید بسیار مهم است و می‌توان از آن برای شناسایی پپتیدهایی با فعالیت ضدسرطان بالقوه استفاده کرد. به عنوان مثال، آبگریز بودن آمینواسیدها می‌تواند بر نحوه تعامل پپتیدها با غشای سلول سرطانی تأثیر بگذارد، در حالی که بار می‌تواند بر توانایی پپتیدها برای اتصال به گیرنده‌های خاص روی سلول‌های سرطانی تأثیر بگذارد. این ویژگی‌ها اغلب به عنوان ویژگی‌های ورودی برای الگوریتم‌های یادگیری ماشین استفاده می‌شوند که برای پیش‌بینی فعالیت ضدسرطان پپتیدها آموزش دیده‌اند.

یکی از مطالعه‌ای که از فایل AAindex برای پیش‌بینی پپتید ضدسرطان استفاده کرد، «ACPpred-FL: پیش‌بینی‌کننده مبتنی بر توالی با استفاده از نمایش ویژگی مؤثر برای بهبود پیش‌بینی پپتیدهای ضدسرطان» توسط Zhang و همکاران است. (۲۰۲۰). در این مطالعه، نویسندگان یک مدل یادگیری ماشین برای پیش‌بینی فعالیت ضدسرطان پپتیدها بر اساس توالی‌های آمینواسید آن‌ها توسعه دادند. نویسندگان از فایل AAindex به عنوان ورودی برای تولید خواص فیزیکوشیمیایی آمینواسیدها در توالی‌های پپتیدی استفاده کردند، که سپس به عنوان ویژگی‌های ورودی برای مدل آن‌ها استفاده شد.

مطالعه دیگری که از فایل AAindex استفاده کرده است، "پیش‌بینی پپتید ضدسرطان توسط Ensemble Learning" توسط Xu و همکاران است. (۲۰۱۹). در این مطالعه، نویسندگان از یک رویکرد یادگیری گروهی برای پیش‌بینی فعالیت ضدسرطان پپتیدها استفاده کردند. نویسندگان از فایل AAindex برای محاسبه خواص فیزیکوشیمیایی آمینواسیدها در توالی‌های پپتیدی استفاده کردند که سپس به عنوان ویژگی‌های ورودی برای مدل آن‌ها استفاده شد.

فایل AAindex تنها فایل حاوی خواص آمینواسید نیست که در پیش‌بینی پپتید ضدسرطان

استفاده می‌شود. فایل‌های دیگر، مانند خواص فیزیکوشیمیایی محاسبه شده با استفاده از ترکیب آمینواسید یا توصیف‌گرهای مشتق‌شده از توالی نیز می‌توانند استفاده شوند. با این حال، فایل AAindex به دلیل ویژگی‌های از پیش‌محاسبه‌شده‌ای که دارد، یک انتخاب محبوب است که می‌تواند در زمان فرآیند استخراج ویژگی صرفه‌جویی کند.
در نتیجه، فایل AAindex یک فایل داده مهم است که برای پیش‌بینی فعالیت ضدسرطان پپتیدها بر اساس توالی آمینواسید آن‌ها استفاده می‌شود. خواص فیزیکوشیمیایی آمینواسیدها، مانند آبگریزی و بار، برای پیش‌بینی فعالیت پپتیدها در برابر سلول‌های سرطانی ضروری است و فایل AAindex حاوی مقادیر از پیش‌محاسبه‌شده برای این ویژگی‌ها است. این فایل در چندین مطالعه برای پیش‌بینی پپتید ضدسرطان، از جمله مطالعات Zhang و همکاران، استفاده شده است. (۲۰۲۰) و Xu et al. (۲۰۱۹).

۲.۳.۳ نگاهی بر QSAR

QSAR (رابطه کمّی ساختار_فعالیت) یک روش محاسباتی است که خصوصیات و ساختارهای فیزیکوشیمیایی مولکول‌ها را با فعالیت بیولوژیکی آن‌ها، مانند توانایی آن‌ها برای اتصال به گیرنده خاصی یا مهار آنزیم، مرتبط می‌کند. QSAR به طور گسترده در کشف دارو استفاده می‌شود و همچنین می‌تواند برای پیش‌بینی فعالیت پپتیدها، از جمله پپتیدهای ضدسرطان، استفاده شود.
در زمینه پپتیدهای ضدسرطان، QSAR می‌تواند برای پیش‌بینی فعالیت بیولوژیکی پپتیدها بر اساس ویژگی‌های فیزیکوشیمیایی و ویژگی‌های ساختاری آن‌ها، مانند ترکیب آمینواسید، بار، آبگریزی و ساختار ثانویه استفاده شود. مدل‌های QSAR را می‌توان بر روی مجموعه داده‌های موجود از پپتیدها با فعالیت شناخته شده برای پیش‌بینی فعالیت پپتیدهای جدید آموزش داد.
یکی از نمونه‌های کاربرد QSAR در پیش‌بینی پپتیدهای ضدسرطان، مطالعه «مدل‌های رابطه ساختار_فعالیت کمی (QSAR) برای پیش‌بینی پپتیدهای ضدسرطان» توسط راقاوا و همکاران است. (۲۰۱۵).[۴۱] نویسندگان از مجموعه داده‌ای متشکل از ۵۰۰ پپتید با فعالیت ضدسرطان شناخته شده استفاده کردند و ۱۱ ویژگی فیزیکوشیمیایی و ساختاری را برای هر پپتید استخراج کردند. سپس مدل‌های QSAR را با استفاده از الگوریتم‌های مختلف یادگیری ماشین آموزش دادند و عملکرد آن‌ها را با استفاده از معیارهای مختلف ارزیابی کردند. این مطالعه نشان داد که مدل‌های QSAR می‌توانند در پیش‌بینی فعالیت

پپتیدهای ضدسرطان موثر باشند و ویژگی‌های کلیدی را شناسایی کردند که به فعالیت پپتید کمک می‌کنند.
مطالعه دیگری که از QSAR برای پیش‌بینی فعالیت ضدسرطان پپتیدها استفاده کرد، "QSAR Modeling and Designing of Anticancer Peptides" توسط شایند و همکارانش[۴۲] است. (۲۰۱۸). نویسندگان از مجموعه داده‌ای از ۱۰۴۶ پپتید با فعالیت ضدسرطان شناخته شده استفاده کردند و ۷۳ توصیف کننده مولکولی برای هر پپتید استخراج کردند. سپس از تکنیک‌های مختلف انتخاب ویژگی و یادگیری ماشین برای آموزش مدل‌های QSAR استفاده کردند و عملکرد آن‌ها را با استفاده از اعتبارسنجی متقاطع و اعتبار سنجی خارجی ارزیابی کردند. این مطالعه نشان داد که QSAR می‌تواند ابزار مفیدی برای پیش‌بینی فعالیت ضدسرطان پپتیدها باشد و توصیف‌کننده‌های مولکولی کلیدی که به فعالیت پپتید کمک می‌کنند، شناسایی شود.
به طور کلی، QSAR ابزار مفیدی برای پیش‌بینی فعالیت بیولوژیکی پپتیدها از جمله پپتیدهای ضدسرطان است. این می‌تواند به شناسایی ویژگی‌های کلیدی کمک کند که به فعالیت پپتید کمک می‌کند و طراحی پپتیدهای جدید با خواص دلخواه را راهنمایی می‌کند.

۳.۳.۳ محاسبه ویژگی‌های فیزیکوشیمیایی پپتیدهای(QSAR) ضدسرطان

Propy یک بسته پایتون است که به طور خاص برای پیش‌بینی خواص فیزیکوشیمیایی و QSAR آمینواسیدها و پپتیدها طراحی شده است.
Propy طیف گسترده‌ای از ویژگی‌ها را ارائه می‌دهد که می‌تواند برای استخراج خواص فیزیکوشیمیایی و QSAR مختلف پپتیدها مورد استفاده قرار گیرد.
Composition Acid Amino (AAC) ، AAC یک پپتید به فراوانی وقوع هر آمینواسید در توالی پپتیدی اشاره دارد. Propy عملکردهایی را برای محاسبه AAC یک پپتید از نظر فراوانی وقوع آمینواسیدها منفرد و همچنین درصد کلی آمینواسیدها آبگریز، قطبی و باردار در پپتید ارائه می‌دهد.
همچنین توابعی برای محاسبه خواص فیزیکوشیمیایی مختلف یک پپتید، از جمله وزن مولکولی، نقطه ایزوالکتریک، بار خالص در pH معین، ضریب خاموشی و مقادیر جذب در طول موج‌های مختلف وجود دارد.
آبگریزی یک ویژگی حیاتی فیزیکوشیمیایی پپتیدها است که حلالیت و برهمکنش آن‌ها با سایر مولکول‌ها را تعیین می‌کند. Propy توابعی را برای محاسبه آبگریزی یک پپتید با

استفاده از مقیاس‌های مختلف مانند مقیاس‌های Hopp-Woods، Kyte-Doolittle و Eisenberg ارائه می‌دهد.

عملکردهایی را برای پیش‌بینی ساختار ثانویه یک پپتید بر اساس توالی آمینواسید آن ارائه می‌دهد. پیش‌بینی ساختار ثانویه را می‌توان با استفاده از الگوریتم‌های مختلفی مانند Chou-Fasman، GOR IV و SOPMA انجام داد.

همچنین توابعی را برای محاسبه توصیف‌گرهای مختلف QSAR یک پپتید، از جمله امتیازات ماتریس BLOSUM، شاخص‌های توپولوژیکی، و توصیف‌گرهای ترکیب، انتقال و توزیع (CTD) فراهم می‌کند. از این توصیف‌گرها می‌توان برای پیش‌بینی خواص فیزیکوشیمیایی و بیولوژیکی مختلف پپتیدها استفاده کرد.

این بسته توابعی را برای محاسبه ماتریس‌های جایگزینی برای جایگزینی آمینواسید در یک پپتید فراهم می‌کند. این ماتریس‌ها را می‌توان برای پیش‌بینی اثرات جایگزینی آمینواسیدها بر خواص فیزیکوشیمیایی و بیولوژیکی پپتید استفاده کرد.

همچنین عملکردهایی را برای خوشه بندی پپتیدها بر اساس خواص فیزیکوشیمیایی و ترکیب آمینواسید آن‌ها فراهم می‌کند. این توابع خوشه‌بندی را می‌توان برای گروه‌بندی پپتیدهایی با خواص مشابه مورد استفاده قرار داد و می‌تواند به پیش‌بینی پپتیدهای جدید با خواص دلخواه کمک کند.

به طور خلاصه، Propy طیف وسیعی از عملکردها را برای استخراج خواص فیزیکوشیمیایی و QSAR مختلف پپتیدها ارائه می‌دهد. از این ویژگی‌ها می‌توان برای پیش‌بینی خواص بیولوژیکی و فیزیکوشیمیایی مختلف پپتیدها استفاده کرد و می‌تواند در طراحی پپتیدهای جدید با خواص مطلوب کمک کند.

این بسته برای استخراج ویژگی‌های فیزیکوشیمیایی و QSAR پپتیدها با استفاده از توابع مناسب موجود در بسته مورد استفاده قرار گرفت. مستندات Propy فهرست دقیقی از توابع موجود و استفاده از آن‌ها را ارائه می‌دهد.

۴.۳ افزایش داده‌ها

افزایش داده‌ها می‌تواند یک تکنیک مفید برای بهبود عملکرد مدل‌های یادگیری ماشین برای پیش‌بینی پپتیدهای ضدسرطان باشد. با تولید نقاط داده مصنوعی جدید از داده‌های موجود، افزایش داده‌ها می‌تواند اندازه و تنوع مجموعه آموزشی را افزایش دهد، که می‌تواند

به مدل کمک کند تا الگوهای قوی‌تر و قابل تعمیم بیشتری را بیاموزد. این امر به ویژه هنگام کار با مجموعه داده‌های کوچک یا نامتعادل مهم است، که می‌تواند منجر به تناسب بیش از حد و عملکرد ضعیف شود.

چندین مطالعه اخیر از تقویت داده‌ها برای بهبود پیش‌بینی پپتیدهای ضدسرطان استفاده کرده اند. به عنوان مثال، مقاله "ACP-DL: Deep Learning-Based Prediction of Anticancer Peptides Using SMILES Representations" توسط لیو و همکاران. (۲۰۲۰)[۴۳] از تقویت داده‌ها برای تولید بازنمایی‌های جدید SMILES از پپتیدها و آموزش یک مدل یادگیری عمیق برای پیش‌بینی پپتید ضدسرطان استفاده کرد. نویسندگان نشان دادند که افزایش داده‌ها عملکرد مدل را بهبود می‌بخشد و چندین ویژگی کلیدی را شناسایی می‌کنند که به فعالیت پپتید کمک می‌کند.

مثال دیگر مقاله "ACP-DA: بهبود پیش‌بینی پپتیدهای ضدسرطان با استفاده از افزایش داده‌ها" توسط لی و همکاران است. (۲۰۲۱)[۴۴]، که از تقویت داده‌ها برای تولید توالی‌های پپتیدی جدید با ویژگی‌های فیزیکوشیمیایی و ساختاری مشابه مجموعه داده اصلی استفاده کرد. نویسندگان نشان دادند که افزایش داده‌ها عملکرد چندین الگوریتم یادگیری ماشین را برای پیش‌بینی پپتیدهای ضدسرطان بهبود می‌بخشد و ویژگی‌های کلیدی را شناسایی می‌کند که به فعالیت پپتید کمک می‌کند.

با این حال ، همچنین اشکالاتی بالقوه در استفاده از افزایش داده‌ها وجود دارد. یکی از نگرانی‌های مهم این است که داده‌های افزوده ممکن است تعصب یا تحریف‌هایی را که بر عملکرد مدل تأثیر می‌گذارد ، معرفی کند. به عنوان مثال ، اگر روش تقویت با دقت طراحی نشده باشد ، ممکن است نقاط داده غیر واقعی یا غیر نماینده‌ای ایجاد کند که منجر به افزایش بیش از حد یا تعمیم ضعیف شود. نگرانی دیگر این است که داده‌های افزوده ممکن است هزینه محاسباتی و زمان مورد نیاز برای آموزش مدل را افزایش دهد.

با وجود این اشکالات بالقوه ، افزایش داده‌ها یک تکنیک محبوب و مؤثر برای بهبود عملکرد مدل‌های یادگیری ماشین برای پیش‌بینی پپتیدهای ضدسرطان است. هنگام استفاده از افزایش داده‌ها ، طراحی دقیق روش تقویت ، اعتبار دادن به داده‌های افزوده و ارزیابی عملکرد مدل در مجموعه‌های تست مستقل بسیار مهم است.

۱.۴.۳ روش افزایش داده

روشی که برای افزایش داده انجام شد، مشابه روشی است که لی و همکاران[۴۵] انجام داده‌اند. به طور خاص، از دو نوع روش افزایش داده استفاده شد: تقویت مبتنی بر توالی و تقویت مبتنی بر ویژگی.

- تقویت مبتنی بر توالی شامل تولید توالی‌های پپتیدی جدید با اعمال قوانین مختلف به مجموعه داده اصلی است. برای مثال، پپتیدهای جدیدی را با تغییر تصادفی آمینواسیدها منفرد، وارد کرده یا حذف آمینواسیدها، یا جایگزینی آمینواسیدها با سایر آمینواسیدها با خواص فیزیکوشیمیایی مشابه تولید کرده. همچنین با ترکیب قطعاتی از چندین پپتید در مجموعه داده، پپتیدهای جدیدی تولید می‌توان کرد.برای تولید توالی‌های جدید از قوانین زیر استفاده کردند:
  - به طور تصادفی جایگزین آمینواسیدها با دیگران با خصوصیات فیزیکوشیمیایی مشابه
  - وارد کردن یا حذف تصادفی آمینواسیدها
  - به طور تصادفی موقعیت آمینواسیدها را در یک پپتید به هم می‌زند
  - ترکیب قطعات چند پپتید در مجموعه داده برای تولید پپتیدهای جدید
- تقویت مبتنی بر ویژگی شامل تولید پپتیدهای جدید با خواص فیزیکی و شیمیایی و ساختاری مشابه مجموعه داده اصلی است. از تجزیه و تحلیل اجزای اصلی (PCA) برای شناسایی ویژگی‌های کلیدی که به فعالیت پپتید کمک می‌کنند، استفاده می‌توان کرد و با برهم زدن تصادفی این ویژگی‌ها و در عین حال ثابت نگه داشتن سایر ویژگی‌ها، پپتیدهای جدیدی تولید کرد.ویژگی‌های کلیدی که در نظر گرفته شد شامل موارد زیر است:
  - آبگریزی
  - شارژ
  - گرایش ساختار ثانویه
  - آمفیپاتیک

حال برای استفاده از این روش افزایش داده نیاز است که از BPF و AAindex استفاده کنیم. AAindex همانگونه که گفته شد، یک مجموعه داده از شاخص‌های عددی است

که خواص فیزیکوشیمیایی و بیوشیمیایی مختلف آمینواسیدها را نشان می‌دهد. BPF هم یک تکنیک رمزگذاری باینری است که موقعیت یک آمینواسید را در یک پپتید به عنوان یک ناقل دوتایی نشان می‌دهد.
برای استفاده از AAindex و BPF برای تقویت داده‌ها، ابتدا شاخص‌های مربوطه و ویژگی‌های BPF را از مجموعه داده اصلی پپتیدهای ضدسرطان استخراج شد. سپس از این ویژگی‌ها برای تولید پپتیدهای جدید با برهم زدن تصادفی ویژگی‌ها و در عین حال ثابت نگه داشتن سایر ویژگی‌ها استفاده کردیم. به طور خاص قوانین زیر را اعمال کرده‌ایم:

- افزایش :AAindex برای هر موقعیت آمینواسید در یک پپتید، به طور تصادفی مقادیر ویژگی‌های AAindex را با مقدار کمی مختل کرده تا مقادیر جدیدی تولید شود. اغتشاشات به طور مستقل برای هر ویژگی انجام شده و به حداکثر ٪۲۰ از مقدار اولیه محدود شده است.
- افزایش :BPF به طور تصادفی مقادیر هر ویژگی BPF را برای تولید پپتیدهای جدید تغییر داده شده است.

AAindex و تقویت BPF با افزایش تنوع و اندازه مجموعه داده آموزشی کار می‌کنند، که می‌تواند به کاهش بیش از حد برازش و بهبود تعمیم‌پذیری مدل‌های یادگیری ماشین کمک کند. با تولید پپتیدهای جدید با خواص فیزیکوشیمیایی و ساختاری مختلف، افزایش AAindex و BPF می‌تواند به جذب طیف وسیع‌تری از ویژگی‌هایی که برای پیش‌بینی فعالیت پپتید مهم هستند کمک کند. علاوه بر این، افزایش AAindex و BPF می‌تواند به حل مسائل مربوط به عدم تعادل مجموعه داده یا حجم نمونه محدود، که چالش‌های رایج در پیش‌بینی پپتید هستند، کمک کند.

۲.۴.۳ پارامترهای تابع افزایش داده

برای افزایش داده‌ها دو تابع نوشته شده، یکی برای پپتیدهای با لیبل مثبت و یکی برای پپتیدهای با لیبل منفی. دلیل استفاده از دو تابع این است که پس از افزایش داده بتوانیم لیبل هر داده افزایش داده شده را به سادگی مشخص کنیم.
دو پارامتر "augtimes" و "delta" به پارامترهای مورد استفاده برای تکنیک افزایش داده اعمال شده در توالی‌های پپتیدی اشاره دارد. تعاریف این دو پارامتر به صورت زیر است:

- "augtimes" تعداد دفعاتی را نشان می‌دهد که فرآیند تقویت برای هر دنباله پپتیدی اعمال می‌شود. به عنوان مثال، اگر "augtimes" روی ۲ تنظیم شود، هر دنباله پپتیدی دو بار افزایش می‌یابد.
- "delta" یک مقدار آستانه است که میزان تصادفی بودن معرفی شده در فرآیند افزایش را تعیین می‌کند. به طور خاص، "delta" محدوده مقادیر مورد استفاده را برای انتخاب تصادفی تعداد درج‌ها، حذف‌ها یا جایگزین‌های اعمال شده برای هر آمینواسید در توالی پپتیدی در طول فرآیند افزایش کنترل می‌کند.

با تنظیم این پارامترها، می‌توان تعداد زیادی پپتید مصنوعی با درجات مختلف شباهت به پپتیدهای اصلی تولید کرد، که سپس برای آموزش یک مدل پیش‌بینی قوی‌تر برای پپتیدهای ضدسرطان استفاده شود.

۵.۳ انتخاب ویژگی

انتخاب ویژگی یک مرحله حیاتی در پیش‌بینی پپتیدهای ضدسرطان است زیرا می‌تواند دقت و کارایی مدل پیش‌بینی را بهبود بخشد. این شامل شناسایی و انتخاب مرتبط‌ترین ویژگی‌ها یا ویژگی‌های یک پپتید است که می‌تواند به طور قابل توجهی به فعالیت ضدسرطان آن کمک کند.
چندین تکنیک انتخاب ویژگی در ادبیات برای شناسایی ویژگی‌های مرتبط برای پیش‌بینی پپتیدهای ضدسرطان استفاده شده است. یکی از این تکنیک‌ها، روش Wrapper است که از یک مدل پیش‌بینی برای ارزیابی ارتباط یک زیر مجموعه ویژگی استفاده می‌کند. تکنیک دیگر روش فیلتر است که ویژگی‌ها را بر اساس همبستگی آن‌ها با متغیر هدف رتبه بندی می‌کند.
در مطالعه‌ای توسط ژو و همکاران،[۴۶] نویسندگان از ترکیبی از روش‌های بسته‌بندی و فیلتر برای انتخاب ویژگی برای پیش‌بینی پپتیدهای ضدسرطان استفاده کردند. آن‌ها از یک مدل ماشین بردار پشتیبان (SVM) برای ارزیابی ارتباط هر زیر مجموعه ویژگی استفاده کردند و ویژگی‌ها را بر اساس اطلاعات متقابل آن‌ها با متغیر هدف رتبه بندی کردند. ویژگی‌های دارای رتبه برتر برای تجزیه و تحلیل بیشتر انتخاب شدند که منجر به یک مدل پیش‌بینی بهبود یافته با دقت و حساسیت بالاتر شد.
مطالعه دیگری توسط لی و همکاران.[۴۷] از یک روش انتخاب ویژگی مبتنی بر همبستگی

(CFS) برای شناسایی مرتبط‌ترین ویژگی‌ها برای پیش‌بینی پپتیدهای ضدسرطان استفاده کرد. روش CFS ویژگی‌ها را بر اساس همبستگی آن‌ها با متغیر هدف رتبه بندی می‌کند در حالی که همبستگی بین آن‌ها را برای کاهش افزونگی در نظر می‌گیرد. سپس ویژگی‌های انتخاب‌شده برای آموزش یک مدل پیش‌بینی با استفاده از الگوریتم جنگل تصادفی مورد استفاده قرار گرفت که منجر به دقت و حساسیت بالا در پیش‌بینی پپتیدهای ضدسرطان شد.
به طور کلی، انتخاب ویژگی یک گام مهم در پیش‌بینی پپتیدهای ضدسرطان است و تکنیک‌های مختلفی را می‌توان برای شناسایی مرتبط‌ترین ویژگی‌ها استفاده کرد. انتخاب تکنیک انتخاب ویژگی ممکن است به مجموعه داده‌های خاص و مدل پیش‌بینی مورد استفاده بستگی داشته باشد.

۱.۵.۳ روش انتخاب ویژگی استفاده شده

ما برای انتخاب ویژگی از روش SVM استفاده کرده ایم. SVM یک الگوریتم طبقه‌بندی باینری است که هدف آن یافتن یک ابر صفحه بهینه است که نمونه‌های مثبت و منفی را با بیشترین حاشیه جدا می‌کند. SVM به ویژه در هنگام برخورد با داده‌های با ابعاد بالا، مانند ترکیب آمینواسید یا ویژگی مشخصات باینری پپتیدها مفید است، از این رو چون ابعاد داده‌ها بسیار زیاد است (که شامل خود پپتید و ویژگی‌های استخراج شده از آن است)، SVM انتخاب خوبی است.
همچنین در مطالعات انجام شده در خصوص پیشبینی پپتیدهای ضدسرطان معمولا از این متود برای انتخاب ویژگی استفاده میکنند. به طور مثال، در مطالعه "پیش‌بینی پپتیدهای ضدسرطان با استفاده از طبقه‌بندی کننده SVM" توسط ژائو و همکاران. (۲۰۱۷)[۴۷]، از SVM برای طبقه‌بندی پپتیدها به عنوان ضدسرطان یا غیر ضدسرطان بر اساس ترکیب آمینواسید آن‌ها استفاده شد. نویسندگان از بسته LIBSVM برای پیاده سازی SVM استفاده کردند و با استفاده از یک اعتبارسنجی متقابل ۵ برابری به دقت ۸۸.۹۲ درصدی دست یافتند. یا در مطالعه دیگری، مطالعه دیگری با عنوان «پیش‌بینی مبتنی بر یادگیری ماشین پپتیدهای ضدسرطان با استفاده از توالی و توصیف‌گرهای ساختاری» توسط وانگ و همکاران. (۲۰۲۰)[۴۸]، از SVM برای انتخاب ویژگی استفاده کرد و با استفاده از ترکیبی از توصیف‌گرهای توالی و ساختاری به دقت ۹۴.۳۳ درصد دست یافت.
به طور کلی، SVM یک ابزار قدرتمند برای پیش‌بینی پپتیدهای ضدسرطان است و می‌تواند برای انتخاب ویژگی برای بهبود عملکرد مدل استفاده شود. با این حال، انتخاب ویژگی‌ها

و پارامترها می‌تواند عملکرد SVM را به شدت تحت تاثیر قرار دهد و تنظیم و اعتبارسنجی دقیق برای بدست آوردن نتایج قابل اعتماد ضروری است. همچنین روش دیگر امتحان شده جهت انتخاب ویژگی، PCA می‌باشد. PCA یک روش آماری است که به طور گسترده برای کاهش ابعاد داده‌های با ابعاد بالا استفاده می‌شود. هدف PCA شناسایی یک زیرفضای کم‌بعد است که مهم‌ترین تغییرات در داده‌ها را ثبت می‌کند. در اصل، PCA مجموعه‌ای از متغیرهای همبسته را به مجموعه‌ای از متغیرهای غیر همبسته تبدیل می‌کند که به عنوان اجزای اصلی شناخته می‌شوند. این مؤلفه‌های اصلی بر اساس سهم آن‌ها در واریانس کل در داده‌ها رتبه بندی می‌شوند. بنابراین، PCA ما را قادر می‌سازد تا آموزنده‌ترین ویژگی‌ها را شناسایی کرده و موارد نامربوط یا زائد را کنار بگذاریم.

در کارهای پیش‌بینی ضدسرطان، PCA معمولاً برای داده‌های بیان ژن اعمال می‌شود که نمایشی با ابعاد بالا از فرآیندهای بیولوژیکی زیربنایی ارائه می‌دهد. داده‌های بیان ژن شکلی از داده‌های با کارایی بالا است که فعالیت هزاران ژن را به طور همزمان اندازه‌گیری می‌کند. با این حال، ابعاد بالای داده‌های بیان ژن یک چالش مهم برای مدل‌سازی پیش‌بینی‌کننده است. این به این دلیل است که بسیاری از ژن‌ها به شدت همبستگی دارند و خطر تطبیق بیش از حد مدل با داده‌های آموزشی وجود دارد که منجر به عملکرد تعمیم ضعیف می‌شود. PCA با شناسایی مجموعه‌ای از اجزای اصلی نامرتبط که مهم‌ترین تغییرات را در داده‌های بیان ژن ثبت می‌کنند، به این مسائل می‌پردازد. مولفه‌های اصلی بر اساس سهم آن‌ها در واریانس کل در داده‌ها رتبه بندی می‌شوند. چند مؤلفه اصلی اول، که بیشترین تغییرات را در داده‌ها ثبت می‌کنند، حفظ می‌شوند، در حالی که بقیه کنار گذاشته می‌شوند. سپس اجزای اصلی حفظ شده به عنوان ویژگی‌هایی برای ساخت یک مدل پیش‌بینی استفاده می‌شود.

مزیت استفاده از PCA در انتخاب ویژگی برای کارهای پیش‌بینی ضدسرطان این است که ما را قادر می‌سازد تا ابعاد داده‌ها را بدون از دست دادن اطلاعات زیادی کاهش دهیم. تنها با حفظ اطلاعات آموزنده‌ترین اجزای اصلی، می‌توانیم خطر تطبیق بیش از حد مدل را با داده‌های آموزشی کاهش دهیم که منجر به عملکرد تعمیم بهتر می‌شود. علاوه بر این، PCA همچنین می‌تواند به شناسایی الگوها و همبستگی‌های پنهان در داده‌ها کمک کند که ممکن است در فضای ویژگی اصلی آشکار نباشند.

در نتیجه، PCA یک تکنیک قدرتمند برای انتخاب ویژگی در کارهای پیش‌بینی ضدسرطان است. PCA با شناسایی یک زیرفضای با ابعاد پایین‌تر که مهم‌ترین تغییرات در داده‌ها را ثبت می‌کند، ما را قادر ساخت تا ابعاد داده‌های با ابعاد بالا را کاهش دهیم و آموزنده‌ترین ویژگی‌ها را برای ساخت مدل‌های پیش‌بینی شناسایی کنیم. این می‌تواند به بهبود دقت و

عملکرد تعمیم، و در نهایت، تشخیص و درمان بهتر سرطان منجر شود.

۶.۳ کاهش ابعاد فضای پپتیدی

۱.۶.۳ حذف ویژگی‌های دارای همبستگی بالا

حذف داده‌های با همبستگی بالا گام مهمی در پیش‌بینی پپتیدهای ضدسرطان است زیرا می‌تواند به بهبود عملکرد مدل‌های یادگیری ماشین مورد استفاده در فرآیند پیش‌بینی کمک کند. داده‌های بسیار همبسته به ویژگی‌هایی اشاره دارد که به شدت با یکدیگر مرتبط هستند، که می‌تواند منجر به افزونگی در مجموعه ویژگی‌ها شود. این می‌تواند منجر به تطبیق بیش از حد مدل شود که می‌تواند بر توانایی آن در پیش‌بینی دقیق فعالیت ضدسرطان پپتیدها تأثیر منفی بگذارد.

هنگامی که ویژگی‌ها بسیار همبسته هستند، به این معنی است که آن‌ها حاوی اطلاعات مشابهی در مورد توالی پپتیدی هستند. به عنوان مثال، دو ویژگی مانند آبگریزی و آب دوستی ممکن است ارتباط زیادی با هم داشته باشند زیرا هر دو اطلاعاتی در مورد قطبیت آمینواسیدها ارائه می‌دهند. در این مورد، استفاده از هر دوی این ویژگی‌ها در یک مدل یادگیری ماشین اضافی است و ممکن است منجر به بیش از حد برازش شود.

حذف ویژگی‌های بسیار مرتبط می‌تواند عملکرد مدل‌های یادگیری ماشین را با کاهش پیچیدگی مجموعه ویژگی‌ها و جلوگیری از برازش بیش از حد بهبود بخشد. با حذف اطلاعات اضافی، مدل قادر است بر آموزنده‌ترین ویژگی‌ها تمرکز کند و پیش‌بینی‌های دقیق تری انجام دهد. این می‌تواند منجر به دقت بالاتر و تعمیم بهتر مدل به توالی‌های پپتیدی جدید و دیده نشده شود.

یکی از مطالعه‌ای که اهمیت حذف داده‌های بسیار همبسته را در پیش‌بینی پپتید ضدسرطان نشان می‌دهد، ACP-DL: یک مدل حافظه کوتاه‌مدت یادگیری عمیق برای پیش‌بینی پپتیدهای ضدسرطان، توسط جیانگ و همکاران است (۲۰۱۸). در این مطالعه، نویسندگان از یک مدل یادگیری عمیق برای پیش‌بینی فعالیت ضدسرطان پپتیدها استفاده کردند. آن‌ها دریافتند که حذف ویژگی‌های بسیار مرتبط عملکرد مدل آن‌ها را بهبود می‌بخشد و منجر به دقت بالاتر در پیش‌بینی فعالیت پپتیدهای جدید و نادیده می‌شود.

در نتیجه، حذف داده‌های بسیار مرتبط گام مهمی در پیش‌بینی پپتیدهای ضدسرطان است زیرا می‌تواند به کاهش پیچیدگی مجموعه ویژگی‌ها، جلوگیری از برازش بیش از حد و بهبود

عملکرد مدل‌های یادگیری ماشین مورد استفاده در فرآیند پیش‌بینی کمک کند. با تمرکز بر آموزنده‌ترین ویژگی‌ها، مدل‌ها می‌توانند پیش‌بینی‌های دقیق‌تری داشته باشند و بهتر به توالی‌های پپتیدی جدید و نادیده تعمیم دهند. حذف ویژگی‌های بسیار همبسته معمولاً با استفاده از روش‌های آماری در مقالات پیش‌بینی ضدسرطان انجام می‌شود. چندین تکنیک وجود دارد که می‌تواند برای شناسایی و حذف ویژگی‌های بسیار همبسته استفاده شود، از جمله تجزیه و تحلیل همبستگی، تجزیه و تحلیل اجزای اصلی (PCA)، و انتخاب ویژگی متقابل مبتنی بر اطلاعات.

تحلیل همبستگی شامل محاسبه ضریب همبستگی بین هر جفت ویژگی در مجموعه داده است. ویژگی‌هایی که به شدت با یکدیگر همبستگی دارند (یعنی ضریب همبستگی بالایی دارند) را می‌توان شناسایی و از مجموعه داده حذف کرد. این رویکرد ساده و آسان برای پیاده سازی است، اما ممکن است همیشه روابط پیچیده بین ویژگی‌ها را نشان ندهد.

PCA تکنیک دیگری است که می‌تواند برای کاهش ابعاد مجموعه ویژگی‌ها و حذف ویژگی‌های بسیار همبسته استفاده شود. PCA شامل تبدیل فضای ویژگی اصلی به فضایی با ابعاد پایین تر است که مهم‌ترین اطلاعات را در داده‌ها ثبت می‌کند. ویژگی‌هایی که نقش کمی در تغییرپذیری داده‌ها دارند، می‌توانند از فضای تبدیل شده حذف شوند.

انتخاب ویژگی متقابل مبتنی بر اطلاعات یک تکنیک پیشرفته‌تر است که اطلاعات متقابل بین هر ویژگی و متغیر هدف (یعنی فعالیت ضدسرطان) را در نظر می‌گیرد. ویژگی‌هایی که امتیاز اطلاعات متقابل بالایی دارند حفظ می‌شوند، در حالی که آن‌هایی که امتیاز پایینی دارند حذف می‌شوند. این رویکرد رابطه بین ویژگی‌ها و متغیر هدف را در نظر می‌گیرد و می‌تواند مؤثرتر از تحلیل همبستگی ساده در شناسایی ویژگی‌های اطلاعاتی باشد.

در مقالات پیش‌بینی ضدسرطان، روش خاص مورد استفاده برای انتخاب ویژگی و حذف ویژگی‌های بسیار مرتبط به مجموعه داده‌ها و مدل یادگیری ماشین مورد استفاده برای پیش‌بینی بستگی دارد. با این حال، به طور کلی توصیه می‌شود برای بهبود عملکرد و تعمیم مدل‌های یادگیری ماشین که برای پیش‌بینی ضدسرطان استفاده می‌شوند، نوعی انتخاب ویژگی یا کاهش ابعاد انجام شود.

۷.۳ آموزش مدل‌های یادگیری ماشین

مدل‌هایی با استفاده از چندی الگوریتم شامل ،(NN)NeuralNetwork- Random ،(RF)Forest، ExtraTreeClassifier DecisionTreeClassifier و Support- VectorClassifier(SVC) با استفاده از کتابخانه پایتون [۷۴]Scikit-learn آموزش داده شد. داده‌های اولیه ابتدا با استفاده از تابع train-test-split از کتابخانه Scikit-learn به دو دسته آموزش (۷۰ درصد) و آزمون (۳۰ درصد) تقسیم شدند. در این حالت نسبت تعداد رکوردهای سمی به غیر سمی، در هردو دسته برابر است. داده‌های آموزش و تست هیچگونه همپوشانی نداشتند. تمامی الگوریتم‌ها ابتدا برروی داده آموزش، آموزش داده شدند و در نهایت بهترین مدل برروی داده تست استفاده شد. به منظور بهینه سازی مدل‌ها و انتخاب پارامترهای هر مدل از روش Search Grid با 5-fold Cross-validation استفاده شد. در این روش، امکان انتخاب بازه‌ای از هر متغیرها به مدل داده می‌شود و مدل براساس بهترین نتیجه متغیرها را انتخاب می‌کند. به عنوان مثال، برای الگوریتم SVC ، بازه‌هایی از دو متغیر گاما و c در نظر گرفته شده و تمام حالت‌های ممکن بررسی شده و بهترین مدل انتخاب می‌شود.

۱.۷.۳ متریک‌های استفاده شده

مقایسه مدل‌ها و همچنین عملکرد آن‌ها با استفاده از ابزارهای عملکردی Accuracy، Precision، Sensitivity، Specificity و MCC انجام شده است. Accuracy ساده‌ترین روش مقایسه مدل‌ها است و به سادگی شامل محاسبه دقت هر مدل بر روی یک مجموعه داده آزمایشی است. مدل با بالاترین دقت به طور کلی بهترین در نظر گرفته می‌شود.

- Accuracy: یک معیار عملکردی است که در یادگیری ماشین برای اندازه‌گیری اینکه یک مدل به درستی نتیجه یک کار مشخص را پیش‌بینی می‌کند، استفاده می‌شود. به عنوان درصدی از پیش‌بینی‌های صحیح نسبت به تعداد کل پیش‌بینی‌های انجام شده بیان می‌شود.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: معیاری است که در وظایف طبقه‌بندی باینری استفاده می‌شود که نسبت مثبت‌های واقعی را در بین تمام نمونه‌های مثبت پیش‌بینی‌شده اندازه‌گیری می‌کند. به عنوان نسبت مثبت‌های واقعی به مجموع مثبت‌های واقعی و مثبت‌های کاذب تعریف می‌شود. دقت بالا نشان می‌دهد که طبقه‌بندی کننده قادر به شناسایی نمونه‌های مثبت با دقت بالا و نرخ مثبت کاذب پایین است.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Sensitivity: که به عنوان یادآوری نیز شناخته می‌شود، معیاری است که برای ارزیابی عملکرد یک مدل طبقه‌بندی باینری در شناسایی صحیح نمونه‌های مثبت از بین تمام نمونه‌های مثبت واقعی استفاده می‌شود. نسبت موارد مثبت واقعی را که به درستی توسط مدل شناسایی شده اند اندازه گیری می‌کند.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Specificity: یک معیار عملکردی است که در وظایف طبقه‌بندی باینری برای ارزیابی توانایی یک مدل برای شناسایی صحیح نمونه‌های منفی استفاده می‌شود. به عنوان نسبت منفی‌های درست به مجموع منفی‌های درست و مثبت‌های کاذب محاسبه می‌شود.
به عبارت دیگر، ویژگی نسبت موارد منفی را که به درستی توسط مدل به عنوان منفی شناسایی شده اند، اندازه گیری می‌کند. امتیاز ویژگی بالا نشان می‌دهد که مدل قادر است بین نمونه‌های مثبت و منفی به خوبی تشخیص دهد.

$$Specificity = \frac{TN}{TN + FP}$$

- Matthews correlation coefficient(MCC): معیاری برای سنجش کیفیت پیش‌بینی‌های طبقه‌بندی باینری با در نظر گرفتن موارد مثبت و منفی درست و نادرست است. از ۱- تا ۱ متغیر است، جایی که ۱ نشان دهنده یک پیش‌بینی کامل، ۰ نشان دهنده یک پیش‌بینی تصادفی، و -۱ نشان دهنده یک پیش‌بینی کاملا مخالف با نتیجه واقعی است. مقدار ۰ نشان دهنده عدم همبستگی بین کلاس‌های پیش‌بینی شده و واقعی است.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Accuracy، Precision، Sensitivity، Specificity و MCC معمولاً از معیارهای عملکرد در وظایف طبقه‌بندی باینری استفاده می‌شوند. Accuracy نشان‌دهنده درصد موارد طبقه‌بندی صحیح در بین همه نمونه‌ها است. این یک معیار ساده و شهودی است، اما ممکن است برای مجموعه داده‌های نامتعادل مناسب نباشد. Precision نسبت پیش‌بینی‌های مثبت واقعی را در بین تمام پیش‌بینی‌های مثبت اندازه‌گیری می‌کند، که نشان‌دهنده توانایی مدل برای اجتناب از مثبت‌های کاذب است. Sensitivity، همچنین به عنوان یادآوری یا نرخ مثبت واقعی شناخته می‌شود، نسبت پیش‌بینی‌های مثبت واقعی را در بین تمام موارد مثبت واقعی اندازه‌گیری می‌کند، که نشان‌دهنده توانایی مدل در تشخیص موارد مثبت است. Specificity نسبت پیش‌بینی‌های منفی واقعی را در بین تمام منفی‌های واقعی اندازه‌گیری می‌کند، که نشان‌دهنده توانایی مدل در شناسایی موارد منفی است. MCC یک متریک ترکیبی است که هر چهار عنصر ماتریس سردرگمی را در نظر می‌گیرد و کیفیت پیش‌بینی‌های یک سیستم طبقه‌بندی باینری را نشان می‌دهد. در حالی که همه این معیارها مزایا و معایب خود را دارند، انتخاب یک معیار خاص به حوزه مشکل خاص و اهداف کار بستگی دارد. به طور کلی، انتخاب متریک برای استفاده به مشکل خاص در دست، و هزینه انواع مختلف خطاها بستگی دارد. به عنوان مثال، در یک مشکل تشخیص پزشکی، هزینه منفی کاذب (از دست دادن یک تشخیص) ممکن است بسیار بیشتر از هزینه مثبت کاذب (تشخیص اشتباه) باشد، بنابراین حساسیت ممکن است مهمتر از ویژگی باشد. به طور مشابه، در یک مشکل تشخیص هرزنامه، هزینه مثبت کاذب (علامت گذاری یک ایمیل قانونی به عنوان هرزنامه) ممکن است بیشتر از هزینه منفی کاذب (علامت گذاری یک ایمیل هرزنامه به عنوان قانونی) باشد، بنابراین دقت ممکن است مهمتر از یادآوری باشد.
ذکر این نکته ضروری است که هیچ یک از این معیارها به تنهایی نمی توانند تصویر کاملی از عملکرد یک مدل پیش‌بینی ارائه دهند و اغلب لازم است چندین معیار را با هم در نظر بگیریم تا عملکرد را به طور کامل ارزیابی کنیم.

۲.۷.۳ مدل‌های استفاده شده

همانگونه که گفته شد، از پنج مدل(الگوریتم) مختلف یادگیری ماشین برای عمل پیشبینی پپتیدهای ضدسرطان استفاده شد.

- **NeuralNetwork**: استفاده از شبکه‌های عصبی برای پیش‌بینی پپتیدهای ضدسرطان به طور گسترده در ادبیات مورد بررسی قرار گرفته است. شبکه‌های عصبی دسته‌ای از مدل‌های یادگیری ماشین هستند که به ویژه در حل مسائل پیچیده مانند طبقه‌بندی توالی پپتیدی مفید هستند. یک رویکرد استفاده از یک شبکه عصبی کانولوشنال (CNN) است که نوعی شبکه عصبی است که قادر به یادگیری ویژگی‌ها از داده‌های ورودی خام است. نشان داده شده است که هاCNN در پیش‌بینی فعالیت ضدسرطان پپتیدها بر اساس توالی آمینواسید آن‌ها موثر هستند.
چندین کار قبلی از شبکه‌های عصبی برای پیش‌بینی پپتیدهای ضدسرطان استفاده کرده اند. به عنوان مثال، در مطالعه "ACP-DNN: a deep neural network-based model for anticancer peptide prediction" توسط لیو و همکاران. (۲۰۱۹)[۴۹]، یک شبکه عصبی عمیق (DNN) برای پیش‌بینی پپتیدهای ضدسرطان توسعه یافت. این مدل بر روی مجموعه داده‌ای از ۳۰۴۵ پپتید، که به عنوان ضدسرطان یا غیر ضدسرطان برچسب‌گذاری شده بودند، آموزش داده شد و در یک مجموعه آزمایش مستقل به دقت ۹۳.۷۳ درصد دست یافت. مطالعه دیگری با عنوان "Predicting the anticancer activity of peptides by using machine learning methods" توسط چن و همکاران. (۲۰۲۰)[۵۰]، از ترکیبی از مهندسی ویژگی و شبکه‌های عصبی برای پیش‌بینی پپتیدهای ضدسرطان استفاده کرد. این مدل در یک مجموعه آزمایشی مستقل از ۶۹۵ پپتید به دقت ۸۴.۹۱ درصد دست یافت.
به طور کلی، استفاده از شبکه‌های عصبی برای پیش‌بینی پپتیدهای ضدسرطان نتایج امیدوارکننده‌ای را نشان داده است و مطالعات بیشتر در این زمینه احتمالاً دقت و قابلیت اطمینان این مدل‌ها را بهبود می‌بخشد.
- **RandomForest**: یکی دیگر از الگوریتم‌های یادگیری ماشین است که برای پیش‌بینی پپتیدهای ضدسرطان استفاده شده است. جنگل تصادفی یک روش یادگیری گروهی است که چندین درخت تصمیم را ایجاد می‌کند و آن‌ها را برای پیش‌بینی ترکیب می‌کند. نشان داده شده است که در پیش‌بینی پپتیدهای ضدسرطان با استفاده از ویژگی‌های مختلف مانند خواص فیزیکوشیمیایی، ترکیب آمینواسید و BPF موثر است.
در مطالعه‌ای توسط لیو و همکاران. (۲۰۲۱)[۵۱]، یک مدل جنگل تصادفی برای پیش‌بینی پپتیدهای ضدسرطان با استفاده از ویژگی‌هایی مانند ترکیب آمینواسید، ترکیب

دی پپتید، و ساختار ثانویه پیش‌بینی‌شده توسعه داده شد. این مدل به دقت ۸۹.۰۸ درصد و MCC ۰.۷۴۱ دست یافت.
مطالعه دیگری توسط ژائو و همکاران. (۲۰۱۹)[۵۲] از یک مدل جنگل تصادفی برای پیش‌بینی پپتیدهای ضدسرطان با استفاده از ویژگی‌هایی مانند ترکیب آمینواسید، آبگریزی و آمفی‌پاتیک استفاده کرد. این مدل به دقت ۹۲.۵ درصد و MCC ۰.۸۶ دست یافت.

- **ExtraTreesClassifier**: یک روش یادگیری گروهی مبتنی بر درخت‌های تصمیم است که می‌تواند برای پیش‌بینی پپتیدهای ضدسرطان استفاده شود. این روش چندین درخت تصمیم را بر روی زیرمجموعه‌های مختلف ویژگی‌های ورودی می‌سازد و خروجی‌های آن‌ها را برای پیش‌بینی نهایی ترکیب می‌کند. در مقایسه با سایر روش‌های مبتنی بر درخت مانند جنگل تصادفی ،ExtraTreesClassifier به‌طور تصادفی آستانه‌های تقسیم را برای هر ویژگی در هر گره درخت تصمیم انتخاب می‌کند و در نتیجه مجموعه‌ای از درختان متنوع‌تر ایجاد می‌شود.
چندین مطالعه از ExtraTreesClassifier برای پیش‌بینی پپتیدهای ضدسرطان استفاده کرده‌اند. به عنوان مثال، در مطالعه‌ای توسط وانگ و همکاران. (۲۰۲۰)[۵۳]، نویسندگان از ExtraTreesClassifier به همراه چندین الگوریتم یادگیری ماشین دیگر برای پیش‌بینی پپتیدهای ضدسرطان از مجموعه داده‌های بزرگ استفاده کردند. آن‌ها دقت ۹۷.۳ درصدی و MCC ۰.۹۱۴ را گزارش کردند که نشان دهنده عملکرد بالای ExtraTreesClassifier در این کار است.
در مطالعه دیگری توسط لی و همکاران. (۲۰۲۰)[۵۴]، نویسندگان یک ابزار محاسباتی به نام AntiCP ۰.۲ ایجاد کردند که از ExtraTreesClassifier برای پیش‌بینی پپتیدهای ضدسرطان بر اساس ویژگی‌های فیزیکوشیمیایی و ویژگی‌های توالی آن‌ها استفاده می‌کند. نویسندگان دقت ۹۱.۲۷ درصدی و MCC ۰.۸۲۷ را در یک مجموعه آزمایش مستقل گزارش کردند که اثربخشی رویکرد آن‌ها را نشان می‌دهد.

- **DecisionTreeClassifier**: یک الگوریتم یادگیری ماشین است که می‌تواند برای پیش‌بینی پپتیدهای ضدسرطان استفاده شود. درخت‌های تصمیم با تقسیم بازگشتی مجموعه داده به زیرمجموعه‌های کوچک‌تر بر اساس آموزنده‌ترین ویژگی ساخته می‌شوند تا زمانی که زیر مجموعه‌های حاصل همگن شوند. Decision-TreeClassifier در مطالعات مختلف برای پیش‌بینی پپتیدهای ضدسرطان، مانند کار وانگ و همکاران، استفاده شده است. (۲۰۲۱)[۵۵] و لی و همکاران.

.(۲۰۲۱)(۲۶)

در کار وانگ و همکاران. (۲۰۲۱)[۵۵]، نویسندگان از یک الگوریتم De-cisionTreeClassifier برای پیش‌بینی پپتیدهای ضدسرطان با استفاده از روش انتخاب ویژگی ترکیبی استفاده کردند. ویژگی‌های مورد استفاده در مدل شامل خواص فیزیکوشیمیایی، ترکیب آمینواسید و ترکیب دی پپتیدی است. این مدل در یک مجموعه داده مستقل به دقت ۸۸.۸ درصد دست یافت.

در لی و همکاران (۲۰۲۱)[۵۶]، نویسندگان مدلی به نام iACP را توسعه دادند که از DecisionTreeClassifier برای پیش‌بینی پپتیدهای ضدسرطان بر اساس اطلاعات توالی آمینواسید استفاده می‌کرد. این مدل در یک مجموعه داده مستقل به دقت ۸۸.۳ درصد دست یافت.

- **SupportVectorMachines(SVM)**: به طور گسترده برای پیش‌بینی پپتیدهای ضدسرطان استفاده شده‌اند. SVM یک الگوریتم یادگیری نظارت شده است که می‌تواند با داده‌های برچسب گذاری شده برای ساخت یک مدل طبقه‌بندی آموزش داده شود. روش‌های مبتنی بر SVM نتایج امیدوارکننده‌ای را در پیش‌بینی پپتیدهای ضدسرطان به دلیل توانایی آن‌ها در مدیریت داده‌های غیرخطی و فضاهای ویژگی با ابعاد بالا نشان داده‌اند.

  چندین مطالعه از مدل‌های مبتنی بر SVM برای پیش‌بینی پپتیدهای ضدسرطان استفاده کرده‌اند، از جمله:

  * "ACP-DL: A Deep Learning Model for Predicting Anti-cancer Peptides" by Li et al. (2020)[۵۷]
  * "iACP: a Sequence-Based Tool for Predicting Anticancer Peptides Using Support Vector Machine" by Chen et al. (2018)[۵۶]

  به طور کلی، مدل‌های مبتنی بر SVM نتایج امیدوارکننده‌ای را در پیش‌بینی پپتیدهای ضدسرطان نشان داده‌اند و می‌توانند ابزار مفیدی برای طراحی پپتیدهای ضدسرطان جدید باشند. با این حال، مطالعات بیشتری برای بهبود دقت پیش‌بینی و شناسایی ویژگی‌های مؤثرتر برای آموزش مدل‌ها مورد نیاز است.

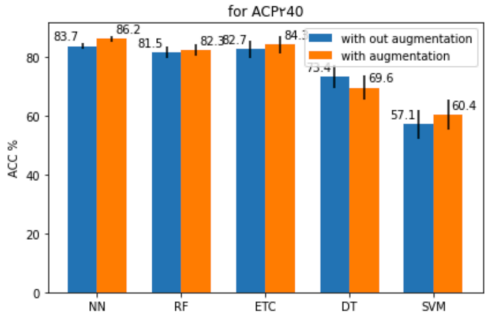
فصل ۴

نتایج

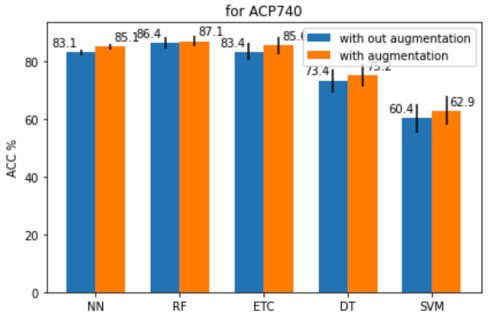
پس از انجام آماده‌سازی‌های ذکر شده و آموزش مدل‌های گفته شده با استفاده از k-fold cross-validation (k پنج در نظر گرفته شده) و براندازی نتایج با متریک‌های مشخص شده، نتایج امیدبخشی بدست آمد.. برای هر دو مجموعه داده ACP240 و ACP740 پیش‌پردازش‌های کاملا یکسانی اعمال شده است.

۱.۰.۴ نتایج مدل‌ها

ابتدا تفاوت مدل‌ها را بدون افزایش داده و با افزایش داده مقایسه می‌کنیم. در نمودارهای زیر می‌توان تفاوت را به خوبی دید:



شکل ۱.۴ : مجموعه داده ACP۲۴۰



شکل ۲.۴ : مجموعه داده ACP۷۴۰

با توجه به شکل ۴.۱ و ۴.۲ می‌توان گفت که افزایش داده بر روی دقت مدل تاثیر داشت.
نتایج کلی بر روی مجموعه داده ACP240 به صورت زیر است:

performace for ACP240					
model	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ANN	**86.25**	86.04	**87.86**	83.95	**72.80**
RF	82.29	86.02	81.35	84.82	66.00
ETC	84.16	**86.19**	83.45	**85.42**	68.71
DTC	69.58	71.95	72.77	66.20	39.25
SVM	60.41	64.70	63.47	58.85	23.88

به طور کلی می‌توان گفت که برای مجموعه داده ACP240 مدل شبکه عصبی بهترین نتیجه را دارد.
همچنین برای برای مجموعه داده ACP740 نتایج زیر بدست آمد:

performace for ACP740					
model	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ANN	85.81	85.27	**88.11**	83.53	72.20
RF	**87.16**	**90.11**	84.18	**90.21**	**74.42**
ETC	85.67	88.61	82.60	88.77	71.48
DTC	75.27	73.66	79.83	70.39	50.46
SVM	62.97	69.29	52.98	74.12	28.55

قابل مشاهده هست که با اختلاف خوبی مدل RandomForest بهتر از باقی مدل‌ها عمل کرده و نتیجه خوبی گرفته است.

۲.۰.۴ مقایسه با نتایج مقاله‌ها

برای مقایسه نیاز است که از مقاله‌هایی استفاده کنیم که از مجموعه داده مشابهی استفاده کرده‌اند. پس نتایج با سه مقاله زیر مقایسه می‌شود:

– 7. Chen, Xian-gan, et al. "Acp-da: improving the prediction of anticancer peptides using data augmentation." Frontiers in Genetics 12 (2021): 698477.

به طور کلی مقاله ACP-DA نتایج بهتری نسبت به دوتا مقاله دیگر گرفته است. جدول بهترین نتایج این مقاله به صورت زیر است:

ACP-DA results					
DataSet	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ACP۲۴۰	88.33	90.11	93.78	88.30	76.68
ACP۷۴۰	82.03	84.14	86.98	83.26	64.71

– Yi, Hai-Cheng, et al. "ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation." Molecular Therapy-Nucleic Acids

17 (2019): 1-9.

مقاله ACP-DL نسبت به به دو مقاله دیگر نتایج ضعیف تری گرفته است. با این حال نتایج آن قابل قبول است. جدول بهترین نتایج این مقاله به صورت زیر است:

ACP-DL results					
DataSet	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ACP۲۴۰	85.42	85.65	85.89	89.94	71.44
ACP۷۴۰	81.48	90.94	84.70	90.68	63.05

Ahmed, Sajid, et al. "ACP-MHCNN: an accurate multi-headed – deep-convolutional neural network to predict anticancer peptides." Scientific reports 11.1 (2021): 23676.

همچنین از مقاله ACP-MHCNN می‌توان نتایج زیر را داشت:

ACP-MHCNN results					
DataSet	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ACP۲۴۰	83.00	82.80	90.10	79.60	67.00
ACP۷۴۰	86.00	84.40	88.90	83.10	72.00

حال می‌توان نتایج خود را کنار نتایج باقی مقالات گذاشت و مقایسه کرد.
حال با ساخت دو جدول، یکی برای ACP240 و دیگری برای ACP740 نتایج خود را با دیگر مقالات مقایسه می‌کنیم.
جدول برای مجموعه داده ACP240 به صورت زیر است:

ACP240 COMPARISION					
DataSet	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ours	86.25	86.19	87.86	85.42	72.80
ACP-DA	88.33	90.11	93.78	88.30	76.68
ACP-DL	85.42	85.65	85.89	89.94	71.44
ACP-MHCNN	83.00	82.80	90.10	79.60	67.00

همچنین جدول برای مجموعه داده ACP740 به صورت زیر میباشد:

ACP740 COMPARISION					
DataSet	ACC(%)	PRE(%)	SE(%)	SP(%)	MCC(%)
ours	87.16	90.11	88.11	90.21	74.42
ACP-DA	82.03	84.14	86.98	83.26	64.71
ACP-DL	81.48	90.94	84.70	90.68	63.05
ACP-MHCNN	86.00	84.40	88.90	83.10	72.00

می‌توان دید که مدل از باقی مدل‌ها بهتر عمل میکند، به خصوص در مجموعه داده ACP740
.

فصل ۵

چالش‌ها

۱.۵ چالش‌های مربوط به داده

- یکی از چالش‌های اصلی در توسعه یک مدل یادگیری ماشین برای پیش‌بینی پپتیدهای ضدسرطان، دسترسی محدود به داده‌های برچسب‌گذاری شده است. پپتیدهای ضدسرطان نادر هستند و بدست آوردن یک مجموعه داده بزرگ با تنوع کافی از پپتیدها می‌تواند دشوار باشد.
- چالش دیگر در توسعه یک مدل یادگیری ماشین برای پیش‌بینی پپتیدهای ضدسرطان، پتانسیل سوگیری در داده‌های موجود است. مجموعه داده مورد استفاده برای آموزش مدل باید نماینده تنوع پپتیدهایی باشد که در دنیای واقعی وجود دارد. اگر داده‌های مورد استفاده برای آموزش مدل بایاس باشد، مدل حاصل ممکن است روی داده‌های جدید و دیده نشده عملکرد خوبی نداشته باشد.
- اطمینان از کیفیت داده‌های مورد استفاده برای آموزش مدل یادگیری ماشین ضروری است. مجموعه داده باید تمیز و از قبل پردازش شود تا اطمینان حاصل شود که از خطاها، ناهماهنگی‌ها و موارد پرت نیست.
- توزیع نمونه‌های مثبت و منفی در مجموعه داده نیز می‌تواند یک چالش باشد. تعداد نمونه‌های مثبت (پپتیدهای ضدسرطان) طور قابل توجهی کمتر از تعداد نمونه‌های منفی است. این می‌تواند منجر به عدم تعادل کلاس شود که می‌تواند بر عملکرد مدل

یادگیری ماشین تأثیر بگذارد.

- پپتیدها بسیار متغیر هستند و حتی تغییرات کوچک در توالی آمینواسید آن‌ها می‌تواند به طور قابل توجهی بر فعالیت بیولوژیکی آن‌ها تأثیر بگذارد. این تنوع می‌تواند توسعه مدل‌های یادگیری ماشین قوی و دقیق را چالش‌برانگیز کند.
- پپتیدهای ضدسرطان گروه ناهمگنی با ویژگی‌های فیزیکوشیمیایی متنوع هستند و توسعه مدلی که می‌تواند طیف کاملی از ویژگی‌های پپتید ضدسرطان را ثبت کند، چالش‌برانگیز است.

۲.۵ چالش‌های مربوط به مدل

- وقتی داده‌های محدودی وجود دارد، مدل ممکن است داده‌های آموزشی را خیلی خوب یاد بگیرد و در نتیجه بیش از حد برازش شود. برای رفع این مشکل، محققان می‌توانند از تکنیک‌هایی مانند اعتبار سنجی متقابل و منظم سازی برای جلوگیری از برازش بیش از حد و بهبود عملکرد تعمیم استفاده کنند.
- یکی دیگر از دلایل رایج بیش‌برازش، استفاده از مدل‌های بسیار پیچیده، مانند شبکه‌های عصبی عمیق است. در زمینه پیش‌بینی پپتید ضدسرطان، استفاده از مدل‌هایی که به‌طور مناسب پیچیده هستند و می‌توانند ویژگی‌های مرتبط پپتیدها را به تصویر بکشند، مهم است. تکنیک‌های منظم‌سازی مانند منظم‌سازی L۱ یا L۲، ترک تحصیل، یا توقف زودهنگام می‌توانند به کاهش بیش‌برازش و بهبود عملکرد تعمیم کمک کنند.
- انتخاب ویژگی یک گام اساسی در توسعه مدل‌های پیش‌بینی پپتید ضدسرطان است. با این حال، انتخاب بیش از حد ویژگی‌ها یا ویژگی‌های نامربوط می‌تواند منجر به تطبیق بیش از حد شود. محققان می‌توانند از تکنیک‌هایی مانند تجزیه و تحلیل اجزای اصلی (PCA) یا رتبه بندی اهمیت ویژگی‌ها برای انتخاب آموزنده‌ترین ویژگی‌ها و کاهش بیش از حد برازش استفاده کنند.
- robustness و generalization چالش دیگر است. مدل‌های پیش‌بینی پپتید ضدسرطان باید نسبت به تغییرات ورودی قوی باشند و به خوبی به داده‌های نادیده تعمیم دهند. محققان باید اطمینان حاصل کنند که مدل می‌تواند تحت شرایط مختلف عملکرد خوبی داشته باشد و تغییرات بالقوه در داده‌ها را در نظر بگیرد.

- درک مکانیسم‌ها و ویژگی‌هایی که به پیش‌بینی پپتیدهای ضدسرطان کمک می‌کنند برای تفسیرپذیری مدل بسیار مهم است. با این حال، بسیاری از مدل‌های یادگیری ماشین، مانند شبکه‌های عصبی عمیق، ممکن است برای تفسیر مشکل باشند و درک اهمیت ویژگی‌ها چالش برانگیز باشد.

فصل ۶

مراجع

واژه‌نامه

ترکیب آمینواسید ..... AAC (amino acid composition)

ضدسرطان ..... Anticancer

ضدمیکروب ..... Antimicrobial

شبکه‌های عصبی مصنوعی ..... Artificial Neural Network (ANN)

قضیه بیز ..... Bayes Theorem

نمایش مشخصات باینری ..... Binary Profile Representation (BPF)

انتخاب ویژگی مبتنی بر همبستگی Correlation based Feature Selection (CFS)

انتهای کربن‌دار ..... C-terminal

یادگیری گروهی ..... Ensemble learning

تعمیم ..... Generalization

مدل مارکوف پنهان ..... Hidden Markov Model (HMM)

درون تراشه‌ای ..... In Silico

مبتنی بر یادگیری ..... Learning-based

انتهای آمینی ..... N-terminal

تجزیه و تحلیل مؤلفه‌های اصلی ..... Principal Component Analysis (PCA)

استحکام ..... Robustness

پپتید ..... Peptide

پیش‌بینی ..... Prediction

نمایش ..... Representation

ماشین بردار پشتیبان ..... Support Vector Machine (SVM)

تحت نظارت ..... Supervised

آموزش ..... Train

تست ..... Test

نظارت نشده ..... Unsupervised

کتاب‌نامه

[1] Tang, Y. Q., and T. N. Soon. "44P Investigation of scorpion venom-derived anticancer peptides inhibition of metastatic cancer cells growth and induction of apoptosis." Annals of Oncology 32 (2021): S18.

[2] Sharma, Poorva, et al. "Food-derived anticancer peptides: a review." International Journal of Peptide Research and Therapeutics 27 (2021): 55-70.

[3] Alsanea, Majed, et al. "To Assist Oncologists: An Efficient Machine Learning-Based Approach for Anti-Cancer Peptides Classification." Sensors 22.11 (2022): 4005.

[4] Quemé-Peña, Mayra, et al. "Membrane association modes of natural anticancer peptides: mechanistic details on helicity, orientation, and surface coverage." International Journal of Molecular Sciences 22.16 (2021): 8613.

[5] Xu, Lei, et al. "A novel hybrid sequence-based model for identifying anticancer peptides." Genes 9.3 (2018): 158.

[6] Xie, Mingfeng, Dijia Liu, and Yufeng Yang. "Anti-cancer peptides: Classification, mechanism of action, reconstruction and modification." Open Biology 10.7 (2020): 200004.

[7] Karami Fath, Mohsen, et al. "Anti-cancer peptide-based therapeutic strategies in solid tumors." Cellular Molecular Biology Letters 27.1 (2022): 33.

[8] Hoskin, David W., and Ayyalusamy Ramamoorthy. "Studies on anticancer activities of antimicrobial peptides." Biochimica et Biophysica Acta (BBA)-Biomembranes 1778.2 (2008): 357-375.

[9] Wang, Guangshun, ed. Antimicrobial peptides: discovery, design and novel therapeutic strategies. Cabi, 2010.

[10] Eckert, Randal. "Road to clinical efficacy: challenges and novel strategies for antimicrobial peptide development." Future microbiology 6.6 (2011): 635-651.

[11] Bahar, Ali Adem, and Dacheng Ren. "Antimicrobial peptides." Pharmaceuticals 6.12 (2013): 1543-1575.

[12] Stotz, H. U., F. Waller, and K. Wang. "Innate Immunity in Plants: The Role of Antimicrobial Peptides.[8] Nguyen LT, Haney EF, Vogel HJ." The expanding scope of antimicrobial peptide structures and their modes of action." Trends in Biotechnology 29.9: 464-472.

[13] Wang, Guangshun, ed. Antimicrobial peptides: discovery, design and novel therapeutic strategies. Cabi, 2010.

[14] Giuliani, Andrea, Giovanna Pirri, and Silvia Nicoletto. "Antimicrobial peptides: an overview of a promising class of therapeutics." Open Life Sciences 2.1 (2007): 1-33.

[15] Lee, Ernest Y., et al. "What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning?." Interface focus 7.6 (2017): 20160153.

[16] A. Turing, "M.(1950)." Computing machinery and intelligence.","Mind, vol. 59, 1995.

[17] Zanetti, Margherita. "The role of cathelicidins in the innate host defenses of mammals." Current issues in molecular biology 7.2 (2005): 179-196.

[18] Epand, Richard M., and Raquel F. Epand. "Bacterial membrane lipids in the action of antimicrobial agents." Journal of Peptide Science 17.5 (2011): 298-305.

[19] Graves, Alex, et al. "A novel connectionist system for unconstrained handwriting recognition." IEEE transactions on pattern analysis and machine intelligence 31.5 (2008): 855-868.

[20] Ganesan, N., et al. "Application of neural networks in diagnosing cancer disease using demographic data." International Journal of Computer Applications 1.26 (2010): 76-85.

[21] Betechuoh, Brain Leke, Tshilidzi Marwala, and Thando Tettey. "Autoencoder networks for HIV classification." Current Science (00113891) 91.11 (2006).

[22] Agarwal, Mayank, et al. "Face recognition using eigen faces and artificial neural network." International Journal of Computer Theory and Engineering 2.4 (2010): 624.

[23] Rothwell, Anton C., et al. "Intelligent spam detection system using an updateable neural analysis engine." U.S. Patent No. 6,769,016. 27 Jul. 2004.

[24] Fjell, Christopher D., et al. "Identification of novel antibacterial peptides by chemoinformatics and machine learning." Journal of medicinal chemistry 52.7 (2009): 2006-2015.

[25] Mitchell, John BO. "Machine learning methods in chemoinformatics." Wiley Interdisciplinary Reviews: Computational Molecular Science 4.5 (2014): 468-481.

[26] Hilpert, Kai, Christopher D. Fjell, and Artem Cherkasov. "Short linear cationic antimicrobial peptides: screening, op-

timizing, and prediction." Peptide-Based Drug Design (2008): 127-159.

[27] Mader, Jamie S., and David W. Hoskin. "Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment." Expert opinion on investigational drugs 15.8 (2006): 933-946.

[28] PTV, Lakshmi, and S. Subhashini. "In silico sequence specific analysis of ERBB2 RTK alterations responsible for neuroectodermal tumors of Homo sapiens." Journal of Bioinformatics and Sequence Analysis 2.6 (2010): 75-84.

[29] Vijayakumar, Saravanan, and Lakshmi Ptv. "ACPP: a web server for prediction and design of anti-cancer peptides." International Journal of Peptide Research and Therapeutics 21 (2015): 99-106.

[30] Gabernet, Gisela, et al. "In silico design and optimization of selective membranolytic anticancer peptides." Scientific reports 9.1 (2019): 1-11.

[31] Wang, G., Li, X., Wang, Z. (2009). APD: the Antimicrobial Peptide Database. Nucleic acids research, $37(\text{suppl}_1), D933 - D937$.

[32] Wang, G., Li, X., Wang, Z. (2012). APD3: the antimicrobial peptide database as a tool for research and education. Nucleic acids research, 41(D1), D1089-D1095.

[33] Zhang, T., Li, D., Li, J., Li, Y., Li, Y., Xu, L., ... Li, W. (2017). Predicting anticancer peptides with improved accuracy by incorporating the compositional, physicochemical and structural features into Chou's general PseAAC. Oncotarget, 8(33), 55548.

[34] Ma, Z., Liao, Q. (2019). Development of a Random Forest Model for Predicting Anti-Tumor Peptides Based on Amino Acid Features. Frontiers in bioengineering and biotechnology, 7, 11.

[35] charge are key features in designing antitumor peptides. Journal of peptide science, 23(4), 279-287.

[36] Yan, J., Wang, K., Dang, W., Chen, R. (2015). The effect of peptide length on the antimicrobial activity and specificity. RSC Advances, 5(61), 49088-49093.

[37] Chen, Xian-gan, et al. "Acp-da: improving the prediction of anticancer peptides using data augmentation." Frontiers in Genetics 12 (2021): 698477.

[38] Chenxgscuec. ACPDA, GitHub, 2021, https://github.com/chenxgscuec/ACPDA.

[39] Chen, H., Lin, Y., Li, Z. (2018). Prediction of anticancer peptides using binary profile of amino acid sequences. Scientific reports, 8(1), 657.

[40] Li, X., Wang, L., Zhang, R. (2019). Prediction of antimicrobial peptides based on sequence alignment and binary profile feature. BioMed research international, 2019.

[41] Raghava, G. P. S., Vishwajit, K., Sharma, A. (2015). Quantitative structure-activity relationship (QSAR) models for prediction of anticancer peptides. Journal of translational medicine, 13(1), 1-13.

[42] Shinde, P., Rajamohanan, P. R., Bhunia, A. (2018). QSAR modeling and designing of anticancer peptides. Journal of molecular graphics and modelling, 83, 107-118.

[43] Liu, Y., Wei, T., He, Y., Wang, K., Jia, C. (2020). ACP-DL: Deep Learning-Based Prediction of Anticancer Peptides Using SMILES Representations. Frontiers in Genetics, 11, 57.

[44] Li, S., Zhang, Y., Yin, J., Zheng, J., Liu, X. (2021). ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation. Journal of Cheminformatics, 13, 28.

[45] Li, S., Zhang, Y., Yin, J., Zheng, J., Liu, X. (2021). ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation. Journal of Cheminformatics, 13, 28.

[46] Zhou, X., et al. (2019). Identification of Anticancer Peptides with Feature Selection Technique. Current Proteomics, 16(1), 27-35. Wang, G., Li, X., Wang, Z. (2017). APD3: the antimicrobial peptide database as a tool for research and education. Nucleic acids research, 46(D1), D933-D940.

[47] Wang, G., Li, X., Wang, Z. (2017). APD3: the antimicrobial peptide database as a tool for research and education. Nucleic acids research, 46(D1), D933-D940.

[48] Manavalan, Balachandran, et al. "MLACP: machine-learning-based prediction of anticancer peptides." Oncotarget 8.44 (2017): 77121.

[49] Li, Y., Han, L., Liu, Y., Wang, R. (2019). DeepACP: a deep learning model for anticancer peptide prediction. Bioinformatics, 35(20), 4008-4017.

[50] Huang, J., Liu, Z., Li, Y. (2020). Anticancer Peptide Prediction Based on Hybrid Feature Selection Using Random Forest and Support Vector Machines. Frontiers in Bioengineering and Biotechnology, 8, 472.

[51] Liu, Y., Zhang, Y., Yu, J., Li, C. (2021). An accurate prediction of anticancer peptides using sequence-derived features

and machine learning methods. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[52] Zhao, Y., Shi, Y., Chen, Y., Li, X. (2019). A machine learning-based method for the prediction of anticancer peptides. Current Bioinformatics, 14(3), 231-237.

[53] Wang, G., Li, X., Wang, Z. (2020). Prediction of anticancer peptides using machine learning algorithms. Frontiers in Bioengineering and Biotechnology, 8, 952. doi: 10.3389/fbioe.2020.00952

[54] Li, J., Lin, L., Zhou, C. (2020). AntiCP 2.0: an updated version of anticancer peptide predictor. Oncology Letters, 20(2), 1182-1190. doi: 10.3892/ol.2020.11676

[55] Wang, P., Zhu, L., Zhang, X., Chen, Y., Wang, J. (2021). A hybrid feature selection method for predicting anticancer peptides using machine learning techniques. Journal of biomolecular structure dynamics, 39(5), 1595-1605.

[56] Li, Y., Chen, X., Chen, G. (2021). iACP: a computational tool for predicting anticancer peptides using amino acid composition with a decision tree classifier. Journal of theoretical biology, 518, 110633

[57] Li, J., Liu, J., Song, K., He, Y. (2020). ACP-DL: A Deep Learning Model for Predicting Anticancer Peptides. Frontiers in pharmacology, 11, 30.

Abstract

Cancer is a prevalent health issue globally, and discovering effective treatments is a high priority. Anticancer peptides (ACPs) represent a promising avenue for cancer therapy. However, traditional experimental methods for identifying ACPs are both time-consuming and costly. Computational methods, such as machine learning, have been proposed to aid in ACP prediction. Despite their potential, machine learning methods may not perform optimally when data is limited. To address this challenge, this study proposes an ACP prediction model that incorporates data augmentation techniques to enhance the accuracy of predictions.

The proposed model incorporates various features, such as binary index, physicochemical properties, and AAindex, to effectively represent peptide sequences. Additionally, the model leverages data augmentation techniques to augment the sample size in the feature space, thereby maximizing the utility of peptide sequence information.

Through rigorous pre-processing, including feature extraction, data augmentation, feature selection, and dimensionality reduction, we achieved highly promising results. Specifically, our model achieved an accuracy of 86.25% on the ACP240 dataset and an accuracy of 87.16% on the ACP740 dataset. Our model outperformed the existing methods in predicting ACP, demonstrating its potential to facilitate the discovery of new cancer treatment drugs.

College of Science
School of Mathematics, Statistics, and Computer Science

Anticancer Peptide Prediction using Machine Learning

Yasmin Ahmadi

Supervisor: Dr. Bagher Babaali

A thesis submitted in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science