علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ...، شماره ...، .....
صفحات ...-....
مقاله پژوهشی

طبقه‌بندی پروتئین‌های بالقوه در طراحی دارو به‌کمک یادگیری بهینه و کاهش بُعد متکی‌بر خوشه‌بندی ویژگی‌ها و تحلیل مشارکتی

شیوا شکرچیان[۱]، حسین اسلامی[۲*]، خسرو رضائی[۲]

* نویسنده مسئول، دریافت: ۱۴۰۳/۰۹/۱۵، بازنگری: ۱۴۰۳/۱۱/۰۱، پذیرش: ۱۴۰۴/۰۲/۳۰

[۱] کارشناسی ارشد، گروه مهندسی پزشکی، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران
[۲] استادیار، گروه مهندسی پزشکی، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران

چکیده

در دهه‌های اخیر، پیشرفت‌های سریع در حوزه پروتئومیکس و طراحی دارو، نیاز به شناخت دقیق‌تر ساختار و عملکرد پروتئین‌ها را افزایش داده است. یکی از چالش‌های اصلی در این زمینه، پیش‌بینی دقیق پروتئین‌های بالقوه برای طراحی داروهای مؤثرتر است. این پژوهش با هدف بهبود دقت و کارایی پیش‌بینی پروتئین‌های بالقوه از طریق رویکردی کارآمد انجام شده است. در این مقاله، یک روش ترکیبی نوآورانه ارائه شده است که یادگیری مبتنی بر XGBoost بهینه‌شده، الگوریتم بهینه‌سازی ازدحام ذرات، و یک گام جدید انتخاب ویژگی مبتنی بر خوشه‌بندی و تحلیل پیچیدگی مشارکتی را با هم ترکیب می‌کند. پس از پیش پردازش و استخراج ویژگی، ویژگی‌های مهم با خوشه‌بندی و انتخاب نماینده‌های کلیدی شناسایی می‌شوند. سپس، مدل‌های پیش‌بینی با استفاده از داده‌های پروتئومیکس و اطلاعات ساختاری پروتئین‌ها آموزش داده می‌شوند. در نهایت، نسخه ارتقاءیافته الگوریتم ازدحام ذرات برای بهینه‌سازی پارامترهای مدل‌های یادگیری XGBoost استفاده می‌شود. داده‌های مورد استفاده شامل پروتئومیکس و ساختارهای پروتئینی از پایگاه‌های DrugBank و Swiss-Prot هستند. نتایج نشان می‌دهد این رویکرد باعث افزایش چشمگیر دقت پیش‌بینی‌ها شده و دقت مدل‌ها را به ۹۶/۶ درصد رسانده است. این روش نوین طراحی داروهای مؤثرتر را تسهیل کرده، هزینه و زمان را کاهش داده و تحقیقات آینده را تقویت می‌کند.

کلمات کلیدی: طراحی دارو، طبقه‌بندی پروتئین بالقوه، یادگیری XGBoost، خوشه‌بندی، تحلیل پیچیدگی مشارکتی.

۱- مقدمه

پروتئین‌های بالقوه دارویی، به پروتئین‌هایی گفته می‌شود که به دلیل ویژگی‌های زیستی خاص خود می‌توانند به عنوان اهداف جدیدی برای درمان‌ها مورد استفاده قرار گیرند. این پروتئین‌ها به واسطه نقش حیاتی‌شان در بدن، در توسعه داروهای نوین اهمیت بالایی دارند. هوش مصنوعی[۱]، و دانش پروتئومیکس[۲]، با شناسایی هوشمندانه این پروتئین‌ها، به کاهش هزینه‌های فرآیند توسعه دارو کمک می‌کنند. پروتئومیکس نیز شاخه‌ای از زیست‌شناسی مولکولی است که به مطالعه ساختار، نقش و عملکرد پروتئین‌ها در سطح سلولی می‌پردازد و زمینه‌ساز درک بهتر فرآیندهای زیستی است. بررسی فعالیت‌های بیولوژیکی و خواص جذب، توزیع، متابولیسم و دفع[۳] در طراحی دارو اهمیت بسزایی دارد. شروع فرآیند طراحی دارو با استفاده از هوش مصنوعی به دو روش شامل طراحی دارویی مبتنی بر ساختار و طراحی دارویی مبتنی بر لیگاند [۱] انجام می‌شود. یادگیری ماشین[۴] و یادگیری عمیق[۵] به عنوان زیرشاخه‌های هوش مصنوعی، به طور قابل توجهی حوزه داروسازی را تغییر داده‌اند. یادگیری ماشینی توانایی مقابله با چالش‌های شیمیایی پیچیده را داراست [۲] و می‌تواند فرآیند کشف دارو را پیش از مرحله تحقیقات پیش‌بالینی بهبود بخشد و هزینه‌های طراحی آن را نیز کاهش می‌دهد [۳]. دهه‌هاست که الگوریتم‌های متعددی در زمینه کشف دارو[۶] اجرا شده‌اند. الگوریتم‌های متداول شامل شیوه‌های نظیر نزدیک‌ترین همسایه‌ها، ماشین‌های بردار پشتیبان[۷] [۴]، طبقه‌بندی ساده بیز[۸] [۵]، رگرسیون خطی و لجستیک [۶]،

[1] Artificial Intelligence (AI)
[2] Proteomics
[3] Absorption, distribution, metabolism, and excretion (ADME)
[4] Machine Learning (ML)
[5] Deep Learning (DL)
[6] Drug discovery
[7] Support Vector Machine (SVM)
[8] Naive Bayes classifier

جنگل تصادفی، فرآیندهای گاوسی [۷]، بوستینگ تقویتی [۸] و درختان تصمیم [۹] هستند. یادگیری عمیق نیز، شیوه‌ای پیشرفته در یادگیری ماشین است که نشان داده به سرعت و با هزینه‌های محدود، نتایج قابل اطمینانی ارائه می‌دهد [۱۰]. همانند هر مسئله دیگری در طبقه‌بندی، پیچیدگی محاسباتی، بیش‌برازش، عدم قطعیت، و ابعاد بالای داده‌ها از چالش‌های اصلی یادگیری ماشینی در طراحی دارو محسوب می‌شوند.

در میان تحقیقات سابق، بر روی طبقه‌بندی پروتئین‌های بالقوه قالب دارو شدن، به‌صورت مستقیم تحقیقات کمی انجام شده است. پژوهش Liu و همکاران [۱۱] به بررسی نقش مدل‌های زبانی مبتنی بر یادگیری ماشینی در کشف و توسعه داروها پرداخته بودند. این مدل‌ها پتانسیل زیادی برای کمک به شناسایی اهداف جدید دارویی و طراحی بالینی داشتند. همچنین، در تحقیق Tang و همکاران [۱۲] از مدل‌های مولد مبتنی بر یادگیری ماشینی برای طراحی دارو استفاده شد. در مطالعه‌ای از سوی Lee و همکاران [۱۳]، کاربردهای یادگیری ماشینی در تحلیل داده‌های بزرگ و طراحی دارو مورد توجه قرار گرفته است. آن‌ها اشاره کرده‌اند که ارتباط عمیق یادگیری ماشینی و طراحی دارو در مرحله‌ای بسیار مهم به نام پیش‌پردازش اطلاعات نقش اساسی داشته است. در تحقیق Kolluri و همکاران [۱۴] نیز به بررسی کاربردهای یادگیری ماشین در تحقیقات و توسعه دارو پرداخته‌اند. مشابه دیگر تلاش‌های شکل‌گرفته در زمینه طراحی دارو و تحلیل آن با استفاده از روش‌های متکی بر یادگیری ماشینی و هوش مصنوعی، در تحقیقات Vora و همکاران [۱۵]، Ren و همکاران [۱۶]، تحقیق [۱۷] و مطالعه مروری Myrko و همکاران [۱۸] نیز دیده می‌شود. همچنین، Alghushairy و همکاران [۱۹] یک مدل به نام Drug-LXGB توسعه دادند که با استفاده از ترکیبی از ویژگی‌های پروتئین‌ها و الگوریتم Light XGBoostپروتئین‌های دارویی را شناسایی می‌کرد. Zhou و همکاران [۲۰] در مطالعه‌ای با تمرکز بر سرطان سینه، یک مدل ترکیبی شامل XGBoost، LightGBM و تصمیم‌گیری بر اساس ویژگی‌های ترکیبی ایجاد کردند که قادر بود به پیش‌بینی فعالیت و بهینه‌سازی ساختار داروهای ضد سرطان کمک کند. Zhang و همکاران [۲۱] مدل DrugFinder را معرفی کردند که با استفاده از داده‌های پروتئینی و الگوریتم XGBoost، به شناسایی پروتئین‌های دارویی با دقت بالا پرداخته و عملکرد برتری نسبت به روش‌های قبلی نشان داده است. Ramakrishnan و همکاران [۲۲] الگوریتمی برای پیش‌بینی نفوذپذیری داروها به سد خونی-مغزی با استفاده از XGBoostو ویژگی‌های فنوتیپی بالینی طراحی کردند که می‌توانست دقت پیش‌بینی داروها را برای کاربرد در پزشکی افزایش دهد.

مطالعات موثر در این زمینه نادر هستند؛ به طور مثال ترکیبی از شبکه‌های عصبی و ماشین بردار پشتیبان با استفاده از الگوریتم ازدحام ذرات[1] بهینه‌سازی شده و نتایج حاصل نشان‌دهنده بهبود قابل‌توجه دقت در پیش‌بینی تعاملات پروتئین‌ها بوده است [۲۳]. این رویکرد نه تنها باعث افزایش دقت پیش‌بینی می‌شود بلکه مقاوم بودن روش را نیز تضمین می‌کند. حتی در برخی تحقیقات، الگوریتم بهینه‌سازی ازدحام ذرات در پیش‌بینی برهم‌کنش‌های پروتئین-دارو به‌کار گرفته شده است [۲۴]. این روش توانسته است با دقت بالا و هزینه‌های کمتر نسبت به روش‌های متداول، به یکی از ابزارهای پیشرفته در طراحی دارو تبدیل شود.

با وجود تحقیقات متنوع، به کارگیری داده‌های واقعی و تعمیم‌پذیری در طراحی دارو محدود بوده است. بنابراین، تحقیقات دقیق‌تر و ارزیابی گسترده‌تر روش‌های یادگیری ماشین در کنار یادگیری ماشین و الگوریتم‌های بهینه‌سازی، ضروری است. نشان داده شده که الگوریتم بهینه‌سازی ازدحام ذرات در بهینه‌سازی مسائل غیرخطی و چندبعدی مؤثر است و به دلیل سرعت بالای همگرایی و توانایی فرار از بهینه‌های محلی، در طبقه‌بندی پروتئین‌ها پاسخگو است [۲۵]. لذا، در این تحقیق، نسخه‌ای بهبود داده شده از الگوریتم بهینه‌سازی ازدحام ذرات به عنوان یک رویکرد نوآورانه در ترکیب با طبقه‌بند تقویت گرادیانی پیشرفته[2] یا XGBoost مورد توجه قرار گرفته است. با توجه به اهمیت کاهش هزینه‌ها و افزایش دقت در فرآیند طراحی دارو، سوال اصلی این پژوهش این است که چگونه می‌توان با تلفیق الگوریتم‌های یادگیری ماشین و بهینه‌سازی، روشی کارآمد و مقاوم برای پیش‌بینی پروتئین‌های بالقوه دارویی ارائه داد که علاوه بر دقت بالا، از چالش‌های مرتبط با داده‌های پیچیده نیز عبور کند؟ بر این اساس، مسئله تحقیق در این مقاله به پیش‌بینی پروتئین‌های بالقوه در طراحی دارو بر اساس تلفیق یادگیری بهینه شده متشکل از گام جدیدی در انتخاب ویژگی و نیز نسخه بهبود داده شده از الگوریتم بهینه‌سازی ازدحام ذرات اشاره دارد. مشارکت‌های اصلی تحقیق به صورت زیر قابل بیان هستند:

الف) غلبه بر چالش پیچیدگی داده‌ها: طبقه‌بندی پروتئین‌های بالقوه اغلب با پیچیدگی ناشی از چند متغیره بودن پارامترهای اثرگذار همراه است. شیوه پیشنهادی تا حد زیادی توانسته از سطح این پیچیدگی بکاهد.

ب) معرفی یک گام نوآورانه در انتخاب ویژگی: یکی از چالش‌های کلیدی در طراحی دارو، حذف ویژگی‌های زائد و شناسایی اطلاعات کلیدی از میان حجم بالای داده‌های پروتئومیکس است. در این تحقیق، از یک روش انتخاب ویژگی مبتنی بر خوشه‌بندی و تحلیل پیچیدگی مشارکتی استفاده شده است. این روش ابتدا ویژگی‌ها را بر اساس ماتریس همبستگی پویا خوشه‌بندی می‌کند و سپس از هر خوشه نماینده‌ای انتخاب می‌شود که بیشترین تأثیر را بر پیش‌بینی هدف دارد. این گام نه تنها حجم داده‌ها را کاهش داده و دقت مدل را افزایش داده است، بلکه منجر به کاهش زمان محاسباتی و بهبود عملکرد مدل در تحلیل داده‌های پیچیده پروتئومیکس شده است.

ج) بهبود پارامترهای مدل استاندارد XGBoost و افزایش دقت طبقه‌بندی: پارامترهای مدل XGBoost که خود به‌صورت انطباقی طراحی خواهد شد، بهینه‌سازی شده‌اند تا دقت و کارایی مدل در تحلیل پروتئین‌های بالقوه در طبقه‌بندی افزایش یابد. نسخه بهینه شده از الگوریتم ازدحام ذرات (موسوم به oPSO) به‌گونه‌ای بهینه‌سازی شده که پارامترهای کلیدی XGBoost را به‌صورت تطبیقی و بر اساس مرحله اجرای الگوریتم تنظیم کند. این تنظیمات پویا شامل تغییرات وزن اینرسی و ضرایب یادگیری است که امکان کنترل دقیق‌تری بر کاوش و همگرایی در مراحل مختلف بهینه‌سازی فراهم می‌کند. با تلفیق روش بهینه سازی ازدحام ذرات بهینه شده و رویکرد XGBoost انطباقی، امکان افزایش دقت و قابلیت طبقه‌بندی مدل در موضوع طراحی دارو و پروتئین‌های بالقوه افزایش می‌یابد.

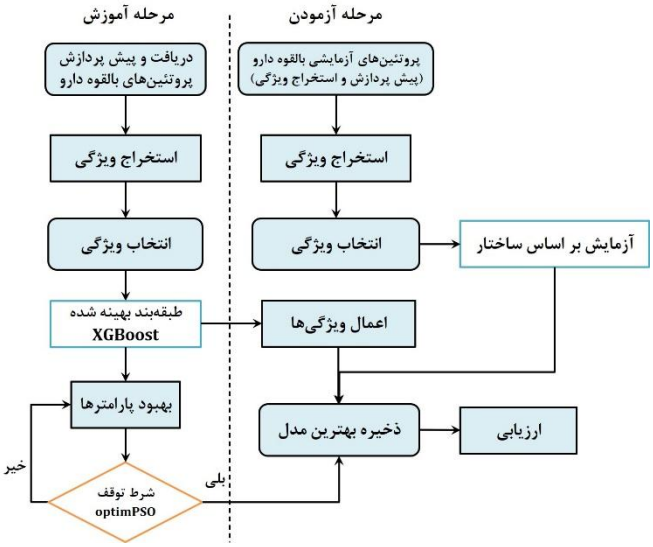
سایر قسمت‌های مقاله به صورت زیر سازماندهی یافته‌اند: در بخش ۲، مواد و روش‌ها به همراه توضیحات مربوط به اجزای مختلف الگوریتم بیان شده است. بخش ۳ به پیاده‌سازی الگوریتم، نتایج تجربی و بحث و تفسیر یافته‌ها اختصاص دارد. در نهایت، بخش ۴ به نتیجه‌گیری از تحقیق خواهد پرداخت.

۲- مواد و روش‌ها

در این بخش، رویکرد پیشنهادی برای پیش‌بینی پروتئین‌های بالقوه در طراحی دارو بر اساس ادغام یادگیری XGBoost و الگوریتم بهینه‌سازی ازدحام ذرات ارائه می‌شود (به شکل ۱ رجوع شود). این رویکرد با استفاده از یک مجموعه داده جامع از پروتئین‌ها آغاز می‌شود که پس از پیش‌پردازش، برای آموزش و ارزیابی مدل‌ها مورد استفاده قرار می‌گیرد. سپس، ویژگی‌های کلیدی از رشته‌های پروتئینی استخراج شده و به منظور طبقه‌بندی وارد فرآیندهای بعدی می‌شوند.

[1] Particle Swarm Optimization (PSO)

[2] Extreme Gradient Boosting (XGBoost)



شکل ۱- روندنمای اجرای روش پیشنهادی برای طراحی دارو و شناسایی پروتئین‌های بالقوه.

در این روش، مدل‌های یادگیری XGBoost خود به صورت انطباقی تنظیم می شود و سپس پارامترهای تنظیمی آن برای بهبود عملکرد، توسط نسخه‌ای بهبودیافته از الگوریتم بهینه‌سازی ازدحام ذرات بهینه می‌شود. هدف نهایی این رویکرد، ارائه ابزاری دقیق و قدرتمند برای شناسایی پروتئین‌های بالقوه در طراحی دارو است.

۲-۱- داده‌ها

جهت طبقه‌بندی پروتئین‌های بالقوه دارویی، داده‌های پروتئینی از پایگاه داده جامع و معتبر DrugBank [۲۶و۲۷] و پایگاه داده Swiss-Prot [۲۸] جمع‌آوری شدند. لازم به ذکر است که DrugBank شامل اطلاعات ساختاری و دارویی درباره داروها و اهداف آن‌هاست، در حالی‌که Swiss-Prot زیرمجموعه‌ای از UniProt است که پروتئین‌های انسانی با تایید تجربی و حاشیه‌نویسی دقیق را شامل می‌شود. از پایگاه DrugBank تعداد ۱۲۲۴ توالی پروتئینی استخراج شد که به عنوان اهداف دارویی شناخته شده‌اند. برای ایجاد تعادل در کلاس‌ها، ۱۳۱۹ پروتئین غیرهدف نیز به‌صورت تصادفی از میان پروتئین‌های انسانی Swiss-Prot انتخاب شدند. به‌منظور حذف افزونگی و جلوگیری از یادگیری مغرضانه توسط مدل، از الگوریتم CD-HIT با آستانه شباهت ۹۰ درصد استفاده شد تا توالی‌های بسیار مشابه فیلتر شوند و تنها نماینده‌های منحصربه‌فرد در مجموعه باقی بمانند. به این ترتیب، مجموعه‌داده نهایی شامل ۲۵۴۳ توالی یکتا از پروتئین‌های انسانی است. توالی‌های پروتئینی به‌صورت رشته‌ای از آمینواسیدها (مانند"MKWVTFISLL") هستند و در حالت خام، فاقد ساختار عددی قابل استفاده برای مدل‌های یادگیری ماشین می‌باشند. به همین دلیل، استخراج ویژگی به‌عنوان مرحله‌ای حیاتی در تبدیل این داده‌های متنی به بردارهای عددی انجام شد. در این پژوهش، از روش ترکیب دی‌پپتیدی برای استخراج فراوانی جفت‌های آمینواسیدی استفاده شد که منجر به تولید ۴۰۰ ویژگی عددی اولیه برای هر نمونه گردید. سپس با استفاده از مدل‌های زبانی پیش‌آموزش‌دیده مانند ProtBERT، جاسازی برداریی از توالی‌ها تولید شده و با اطلاعات تکاملی حاصل از هم‌ترازی چندگانه توالی‌ها ادغام شدند. در نهایت، شایان ذکر است که این ویژگی‌ها به طور مستقیم از داده‌های واقعی و معتبر استخراج شده‌اند و توسط نویسندگان ساخته نشده‌اند. برای خوانندگانی که با زیست‌مولکول‌ها آشنا نیستند، می‌توان این روند را مشابه پردازش زبان طبیعی در نظر گرفت: توالی‌های پروتئینی همانند جملات متنی هستند که هدف، تشخیص «داروشدنی» بودن آن‌ها با استفاده از مدل‌های طبقه‌بندی است.

۲-۲- پیش‌پردازش

در مرحله پیش‌پردازش داده‌ها، به‌جای حذف مقادیر از دست رفته، که می‌تواند دقت مدل را کاهش دهد، از روش Kنزدیکترین همسایگی برای بازیابی آن‌ها استفاده می‌شود. در این روش، مقادیر از دست‌رفته هر ویژگی با استفاده از نزدیک‌ترین همسایگان آن در داده‌های موجود برآورد می‌شود. برای این کار، ابتدا فاصله هر نمونه از دست‌رفته با دیگر نمونه‌ها محاسبه می‌شود و سپس مقادیر ویژگی‌های همسایه‌های نزدیک برای پرکردن جای خالی استفاده می‌شوند. این روش به دلیل بهره‌گیری از ساختار داده‌ها در جایگزینی مقادیر، توانایی حفظ انسجام و دقت داده‌ها را افزایش می‌دهد. لازم به ذکر است که برای شناسایی خودکار داده‌های از دست‌رفته، فرآیندی با استفاده از شیوه بررسی داده و شناسایی نواقص پیاده‌سازی شده است. داده‌های از دست‌رفته به‌عنوان مقادیر تهی[1]، مقادیر خالی یا مقادیر خارج از بازه‌های منطقی و آماری تعریف‌شده در داده‌ها تشخیص داده می‌شوند. همچنین، هر ویژگی با توجه به بازه‌های عددی و معیارهای آماری بررسی می‌شود تا داده‌های خارج از بازه شناسایی گردند. برای شناسایی این مقادیر، از شیوه آماری مانند توزیع نرمال و انحراف معیار از میانگین بهره گرفته شد. در گام بعدی، داده‌ها با نرمال‌سازی کمینه-بیشینه مقیاس‌بندی می‌شوند تا یکپارچگی و دقت در طبقه‌بندی حفظ شود. این کار پس از تبدیل توالی‌های آمینواسیدی به بردارهای عددی ممکن می‌شود. گام هنجارسازی ویژگی‌ها بر اساس رابطه (۱) صورت می‌پذیرد:

$$Y_{\text{norm}} = \frac{Y_s - Y_{s_{min}}}{Y_{s_{max}} - Y_{s_{min}}} \quad (۱)$$

روش بیشینه-کمینه به حفظ توزیع نسبی داده‌ها کمک می‌کند و مانع از تأثیر منفی ویژگی‌هایی با مقیاس‌های بزرگ‌تر روی مدل‌های یادگیری ماشین می‌شود. در این معادله، Y_{norm} مقدار هنجار شده، Y_s مقدار اصلی ویژگی یا ویژگی حال حاضر، Y_{smin} کمترین مقدار ویژگی در مجموعه داده و Y_{smax} نیز بزرگترین مقدار ویژگی در مجموعه داده است. این رابطه مقادیر اصلی را به بازه صفر تا یک منتقل می‌کند.

۲-۳- استخراج ویژگی

ویژگی‌های متعددی از داده‌های دارویی استخراج شده‌اند که در فرآیند طبقه‌بندی اهمیت بالایی دارند. برای افزایش دقت، از تکنیک ترکیب دی‌پپتیدی استفاده شده است، که یک روش ۲-گرمی است و احتمال وقوع دو باقی‌مانده متوالی اسید آمینه در توالی را پیش‌بینی می‌کند. در این فرآیند، با روش پنجره کشویی، الگوهای متوالی اسیدهای آمینه در توالی‌های پروتئینی رمزگذاری و فراوانی رخدادها محاسبه می‌شود. برخلاف ترکیب اسید آمینه‌های منفرد، ترکیب دی‌پپتیدی به جفت‌شدن باقی‌مانده‌های مجاور توجه کرده و اطلاعاتی درباره ترکیب و ویژگی‌های توالی اسیدهای آمینه ارائه می‌دهد. ترکیب دی‌پپتید به عنوان یک رویکرد کارآمد برای استخراج ویژگی در نظر گرفته می‌شود و اغلب به صورت یک بردار ویژگی جامع ۴۰۰ بعدی توصیف می‌شود:

$$\bar{q} = [q_1, q_2, q_3, \ldots, q_{400}] \quad (۲)$$

در این معادله، عبارت q_i ضریب احتمال جفت‌های باقی‌مانده i فرض می‌شود و به صورت رابطه (۳) قابل تعریف است:

$$q_i = \frac{m_i}{M} (i = 1,2,3,\ldots,400) \quad (۳)$$

در این معادله، ضریب m_i تعداد جفت‌های باقی‌مانده از i است و از سویی، مقدار M هم تعداد همه احتمال‌های مرتبط با جفت‌های باقی‌مانده در نظر گرفته می‌شود. در اینجا، ویژگی‌های استخراج‌شده از توالی‌های پروتئینی برای هر رشته به‌صورت یک بردار ۴۰۰ بُعدی ساخته می‌شوند. این ویژگی‌ها در پنج دسته شامل قطبیت،

[1] Null

اسیدیته، بار الکتریکی، ساختار ثانویه و دی‌هیدروپیریدین خلاصه شده‌اند. این دسته‌بندی بر اساس خصوصیات فیزیکوشیمیایی ۲۰ اسید آمینه صورت گرفته و به‌طور خاص برای نمایش ساختارهای متنوع و عملکردهای فیزیولوژیکی خاص آن‌ها تنظیم شده است. در نهایت، طبقه‌بندی توالی‌های کوتاه‌شده پروتئین‌ها بر اساس این ویژگی‌های کلیدی، مطابق جدول ۱ انجام می‌گیرد [۲۹].

جدول ۱ نقش کلیدی ویژگی‌های فیزیکوشیمیایی، از جمله قطبیت و اسیدیته، را در تحلیل توالی‌های پروتئین نشان می‌دهد. قطبیت پروتئین‌ها بر تعاملات آن‌ها با محیط اطراف و مولکول‌های دیگر اثر می‌گذارد و اسیدیته نیز می‌تواند به‌طور مستقیم بر پایداری ساختاری و فعالیت بیولوژیکی پروتئین‌ها تأثیر بگذارد.

جدول ۱- نمایش توالی آمینو اسیدهای بدست آمده از پایگاه داده که ویژگی‌ها از مجموعه بدست آمده است.

ویژگی‌های مربوط به پپتیت	دسته‌بندی رشته‌ها
دی هیدروپیریدین	KR AVNCQGHILMFPSTWY DE
بار	KR AVNCQGHILMFPSTWY DE
ساختار ثانویه	EHALMQKR VTIYCWF GDNPS
اسیدیته	DE KHR ACFGILMNPQSTVWY
قطبیت / اسیدیته	DE RHK WYF SCMNQT GAVLIP

این ویژگی‌ها با ارائه اطلاعات اساسی درباره خواص پیوندهای شیمیایی و ساختار سه‌بعدی پروتئین‌ها، نقش حیاتی در طبقه‌بندی و تحلیل توالی‌های پروتئینی دارند. ۲۰ اسید آمینه را می‌توان به طور سیستماتیک به پنج دسته گروه بندی کرد: اسیدی (D، E)، بازی (R، H، K)، معطر (W، Y، F)، قطبی خنثی (S، C، M، N، Q، T)، و غیر قطبی (G، A، V، L، I، P). به عنوان مثال، در تجزیه و تحلیل پروتئین I = {MATRTQARGAVVELLYAFESGNEEIKKIASSMLEE}، ۲۰ اسید آمینه را می توان در یک توالی فشرده کرد و بر قطبیت و اسیدیته پروتئین تأکید کرد.

با کاهش پیچیدگی توالی و تمرکز بر ویژگی‌های فیزیکوشیمیایی مربوطه، تجزیه و تحلیل و تفسیر داده‌های پروتئین قابل دسترس‌تر و قابل درک‌تر می‌شود. بر اساس تعریف I = {PNPKPPNKNNNNNRNAPNPAANKKNNPPPNAA} فرض می‌شود.

تجزیه و تحلیل فرکانس های کاراکتر در یک توالی پروتئین متراکم شامل بررسی ترکیباتی مانند $|R_i|$ و $|R_iR_j|$، که در آن R_i و R_j نشان دهنده انواع اسیدهای آمینه i و j ام در توالی کاهش یافته است. برای نمایش اسید آمینه k-امین و کاراکتر i-امین در کاهش دنباله، علامت $|a_{ki}|$ بکارگرفته شده است. در اینجا، k به موقعیت مشخصی از یک اسید آمینه در توالی پروتئینی کاهش‌یافته اشاره دارد. همچنین، دو بردار شاخص V_1 و V_2 بکار گرفته شده است که در آن بردار V_1 به صورت معادله (۴) تعریف می شود. در این معادله، هر عنصر از بردار مربوطه به جنبه خاصی از توالی پروتئین اشاره دارد که رخدادها و روابط بین انواع اسیدهای آمینه و تکرار آنها را نشان می‌دهد. این رویکرد جامع امکان تجزیه و تحلیل دقیق الگوهای اساسی و ویژگی‌های ساختاری در توالی پروتئین متراکم را فراهم می‌کند:

$$V_1 = \left(\frac{|a_{k1}|}{|R_1|}, \frac{|a_{k2}|}{|R_2|}, \ldots, \frac{|a_{km}|}{|R_m|}\right) \quad (۴)$$

بر این اساس، بردار حاصل از V_2 مطابق رابطه (۵) قابل توصیف است:

$$V_2 = \left(\frac{|R_1, R_j|}{|R_1|}, \frac{|R_2, R_j|}{|R_2|}, \ldots, \frac{|R_m, R_j|}{|R_m|}\right) \quad (۵)$$

در هر دو معادله (۴) و (۵)، V_1 برداری است که شامل مقادیر فراوانی جفت شدن‌های متوالی باقی‌مانده‌های اسید آمینه در یک توالی پروتئینی است، در حالی که V_2 نیز برداری مشابه است اما به فراوانی جفت‌های غیرمتوالی و با فاصله بین باقی‌مانده‌های اسید آمینه در همان توالی اشاره دارد.

۲-۴- انتخاب ویژگی

برای کاهش ابعاد ویژگی‌های استخراج شده از انتخاب ویژگی مبتنی بر خوشه‌بندی ویژگی‌ها و تحلیل مشارکتی استفاده می‌شود که در تحقیق دیگر از شیوه مشابه آن بهره گرفته نشده است. در این روش، به جای ارزیابی هر ویژگی به صورت مستقل، مجموعه‌ای از ویژگی‌ها به صورت خوشه‌ای تحلیل می‌شوند. خوشه‌بندی ویژگی‌ها بر اساس همبستگی و نقش آن‌ها در توضیح تنوع داده انجام می‌شود. سپس، از هر خوشه، نماینده‌ای به عنوان ویژگی اصلی انتخاب می‌شود که بیشترین اطلاعات را درباره هدف ارائه می‌دهد. این روش از مفاهیمی مانند ماتریس همبستگی پویا و شاخص پیچیدگی داده استفاده می‌کند. در این روش، پیشنهاد کرده‌ایم که در ابتدا محاسبه ماتریس همبستگی پویا صورت پذیرد و برای این منظور، ابتدا ماتریس همبستگی C برای تمام ویژگی‌ها محاسبه می‌شود:

$$C_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_{X_i} \cdot \sigma_{X_j}} \quad (۶)$$

که در آن C_{ij} میزان همبستگی بین ویژگی X_i و X_j است، Cov کوواریانس و σ انحراف معیار است. سپس، ماتریس همبستگی به صورت پویا تنظیم می‌شود تا اثر ویژگی‌های غیرخطی نیز در نظر گرفته شود:

$$C'_{ij} = C_{ij} \cdot \exp\left(-\frac{\Delta(X_i, X_j)}{\tau}\right) \quad (۷)$$

در اینجا نیز، $\Delta(X_i,X_j)$ فاصله پیچیدگی بین ویژگی‌ها و τ پارامتر تنظیمی است که حساسیت را کنترل می‌کند. ما در گام بعدی از خوشه بندی تجمعی[1] استفاده می‌کنیم. بر اساس ماتریس همبستگی پویا C'، ویژگی‌ها خوشه‌بندی می‌شوند و هر خوشه شامل مجموعه‌ای از ویژگی‌های مرتبط است که اطلاعات مشابهی ارائه می‌دهند. رای هر خوشه، نماینده‌ای به عنوان ویژگی اصلی انتخاب می‌شود. این انتخاب بر اساس معیار زیر انجام می‌شود:

$$S_j = \frac{\mathrm{Var}(X_j \cap Y)}{\mathrm{Var}(X_j)} \quad (۸)$$

که در آن S_j نمره مشارکت ویژگی X_j در پیش‌بینی هدف دارویی شدن و از طرفی هم $\mathrm{Var}(X_j \cap Y)$ واریانس مشترک بین ویژگی X_j و قابلیت دارویی شدن است. به منظور ارزیابی کیفیت خوشه‌بندی و انتخاب نماینده‌ها، شاخص پیچیدگی مشارکتی CC محاسبه می‌شود:

$$CC = \sum_{k=1}^{K} \left(\frac{|C_k|}{N} \cdot \frac{\text{Diversity } (C_k)}{\text{Redundancy } (C_k)}\right) \quad (۹)$$

در اینجا، Diversity میزان تنوع اطلاعاتی ویژگی‌های خوشه، Redundancy میزان همبستگی داخلی ویژگی‌های خوشه، N تعداد کل ویژگی‌ها و در نهایت $|C_k|$ تعداد ویژگی‌ها در خوشه kام تعریف می‌شوند. نماینده‌های منتخب از خوشه‌ها به عنوان ویژگی‌های نهایی برای مدل اصلی استفاده می‌شوند. روش پیشنهادی مزایای قابل توجهی در پردازش داده‌های پیچیده دارد. نخست آنکه، با استفاده از خوشه‌بندی ویژگی‌های مشابه، هم‌خطی بین ویژگی‌ها کاهش می‌یابد و ویژگی‌های تکراری حذف می‌شوند، که این امر در داده‌های با ابعاد بالا که ممکن است باعث انحراف عملکرد مدل شوند، بسیار مفید است. علاوه بر این، این روش پیچیدگی‌های غیرخطی را نیز در نظر می‌گیرد؛ استفاده از ماتریس همبستگی پویا و شاخص پیچیدگی مشارکتی، روابط غیرخطی بین ویژگی‌ها را تجزیه و تحلیل می‌کند که ممکن است توسط روش‌های متداول نادیده گرفته شوند. این ویژگی باعث افزایش دقت و تفسیرپذیری مدل می‌شود؛ زیرا انتخاب نماینده‌های خوشه‌ها باعث می‌شود که ویژگی‌های منتخب نمایانگر دقیق‌تری از کل داده‌ها باشند. این مسئله به بهبود عملکرد مدل کمک می‌کند چرا که ویژگی‌های انتخاب‌شده، اطلاعات مفید و مرتبط‌تری را فراهم می‌کنند. علاوه بر این، این روش انعطاف‌پذیر است و می‌تواند در انواع داده‌ها (زیستی، تصویری، عددی) و مسائل مختلف به کار گرفته شود. در مقایسه با روش‌های کلاسیک مانند آنالیز اجزای اصلی یا الگوریتم انتخاب ویژگی مبتنی بر کوچک‌سازی مطلق، این روش به‌طور گروهی به همبستگی و تنوع

[1] Agglomerative Clustering

اطلاعاتی میان ویژگی‌ها توجه می‌کند و بنابراین موثرتر از آنها واقع می‌شود. در داده‌های زیستی که ویژگی‌ها معمولاً ارتباطات پیچیده و غیرخطی دارند، این روش می‌تواند اطلاعات مفیدتری استخراج کرده و به ساخت مدل‌های پیش‌بینی بهتری منجر شود. با کاهش هم‌خطی و انتخاب ویژگی‌های اصلی، عملکرد مدل پیشنهادی oPSO-oXGBoost در طبقه‌بندی داده‌های نمونه به‌طور قابل توجهی بهبود می‌یابد.

۲-۵- طبقه‌بندی

در بخش طبقه‌بندی از رویکردی نوآورانه استفاده می‌شود که oPSO-oXGBoost نام دارد و در آن هر دو الگوریتم PSO و XGBoost به صورت بهینه‌سازی‌شده طراحی و اجرا شده‌اند. این روش ترکیبی از قابلیت‌های قدرتمند الگوریتم PSO برای بهینه‌سازی پارامترها و توانایی مدل XGBoost در ایجاد مدل‌های دقیق است. در ادامه، ابتدا بخش‌های بهینه‌شده PSO و XGBoost توضیح داده می‌شوند و سپس ترکیب این دو بخش به عنوان یک مدل هیبریدی توصیف می‌گردد.

۲-۵-۱- نسخه بهبود یافته از الگوریتم ازدحام ذرات

در نسخه بهینه‌شده الگوریتم ازدحام ذرات، تغییراتی در جهت بهبود دقت و سرعت همگرایی اعمال شده است. هدف از این تغییرات، کاهش احتمال گیر افتادن در بهینه‌های محلی و بهبود همگرایی بهینه در فضای جستجو است. برای مثال، در مراحل اولیه، وزن اینرسی بالاتر است که اجازه می‌دهد الگوریتم فضای جستجو را به‌صورت گسترده‌تری کاوش کند، و در مراحل پایانی، این وزن به‌تدریج کاهش می‌یابد تا الگوریتم به سمت بهینه‌های دقیق‌تر همگرا شود. درابتدا به جای استفاده از پارامترهای ثابت برای کنترل گام‌های کاوش و بهره‌برداری، پارامترها به صورت پویا و بر اساس مرحله فعلی الگوریتم تنظیم می‌شوند. این کار باعث می‌شود که الگوریتم بتواند در مراحل ابتدایی به خوبی فضای جستجو را کاوش کند و در مراحل پایانی به بهینه‌های محلی نزدیک‌تر شود. گام بعدی شامل بهبود مکانیزم به‌روزرسانی موقعیت ذرات است که معادله به‌روزرسانی موقعیت ذرات در الگوریتم به صورت (۱۰) و (۱۱) تنظیم می‌شود:

$$v_i^{(t+1)} = \omega v_i^{(t)} + c_1 r_1 \left(p_i^{best} - x_i^{(t)}\right) + c_2 r_2 \left(g_i^{best} - x_i^{(t)}\right) \quad (۱۰)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \quad (۱۱)$$

که در این معادلات، $v_i^{(t)}$ و $x_i^{(t)}$ به ترتیب سرعت و موقعیت ذره iام در تکرار t ام هستند. همچنین، ω وزن اینرسی است که می‌تواند به صورت پویا تنظیم می‌شود و c_1 و c_2 ثابت‌های یادگیری در نظر گرفته می‌شوند. به همین ترتیب، r_1 و r_2 ضرایب تصادفی و در نهایت p_i^{best} و g^{best} به ترتیب بهترین موقعیت ذره iام و بهترین موقعیت هستند. با این حال، در الگوریتم oPSO، به جای استفاده از مقادیر ثابت برای ω، c_1 و c_2، این پارامترها به‌صورت پویا تنظیم می‌شوند تا سازگاری الگوریتم افزایش یابد. برای این منظور، ω طوری تعریف می‌شود که با گذشت زمان مقدار آن کاهش یابد؛ به عبارت بهتر در مراحل اولیه، کاوش گسترده‌تری انجام می‌شود و در مراحل پایانی همگرایی بهتری رخ می‌دهد:

$$\omega^{(t)} = \omega_{max} - \frac{\omega_{max}-\omega_{min}}{T} \times t \quad (۱۲)$$

که در آن ω_{max} و ω_{min} به ترتیب بیشینه و کمینه وزن‌های اینرسی هستند و از سویی، T تعداد نهایی تکرارها در نظر گرفته می‌شوند. کاهش تدریجی ω کمک می‌کند که در ابتدا کاوش گسترده‌ای صورت گیرد و در مراحل بعدی همگرایی بهینه‌ای حاصل شود. همچنین، در الگوریتم بهینه شده oPSO، c_1 و c_2 نیز به صورت پویا تغییر می‌کنند و بر این اساس، روابط زیر برای تنظیم این ضرایب مورد استفاده قرار می‌گیرند:

$$c_1^{(t)} = \left(c_1^{\text{final}} - c_1^{\text{initial}}\right) \cdot \frac{t}{T} + c_1^{\text{initial}} \quad (۱۳)$$

$$c_2^{(t)} = \left(c_2^{\text{final}} - c_2^{\text{initial}}\right) \cdot \left(1 - \frac{t}{T}\right) + c_2^{\text{initial}} \quad (۱۴)$$

بر اساس این معادلات، روند الگوریتم به گونه‌ای پیشرفت می‌کند که اگر در مراحل اولیه، ضریب یادگیری شخصی c_1 بیشتر باشد و کاوش را تشویق کند، در حالی که ضریب یادگیری جمعی c_2 در مراحل بعدی افزایش می‌یابد تا ذرات را به سمت بهترین راه‌حل‌ها هدایت کند.

بر اساس این معادلات، الگوریتم به گونه‌ای پیشرفت می‌کند که در مراحل اولیه، ضریب یادگیری شخصی c_1 بیشتر است و کاوش را تشویق می‌کند، در حالی که ضریب یادگیری جمعی c_2 در مراحل بعدی افزایش می‌یابد تا ذرات را به سمت بهترین راه‌حل‌ها هدایت کند. علاوه بر بهینه‌سازی در پارامترها و مکانیسم بروزرسانی، بهینه‌سازی همزمان چند هدفه نیز مد نظر است. در این بخش از بهینه‌سازی الگوریتم، بهینه‌سازی همزمان چندین تابع هدف نیز به کار گرفته می‌شود تا الگوریتم بتواند به صورت همزمان چندین معیار مختلف (مانند دقت، سرعت، و دقت تعمیم عملکرد) را بهینه کند.

۲-۵-۲- گام یادگیری

طبقه‌بند XGBoost یکی از الگوریتم‌های توانمند در تقویت مدل‌های درخت تصمیم است. این الگوریتم از رویکرد تقویت‌سازی گرادیانی [1] استفاده می‌کند که در آن مدل‌های ضعیف‌تر به‌صورت پی‌درپی ساخته می‌شوند تا خطاهای مدل‌های قبلی را اصلاح کنند. بخش اول بهینه‌سازی این طبقه‌بند افزایش کارایی با استفاده از نمونه‌گیری تطبیقی است که در آن روش به مدل اجازه می‌دهد تا وزن بیشتری به نمونه‌هایی که طبقه‌بندی آن‌ها دشوارتر است بدهد. به عبارت دیگر، نمونه‌هایی که به درستی طبقه‌بندی نشده‌اند یا مدل در آن‌ها دچار اشتباه شده است، در تکرارهای بعدی آموزش وزن بیشتری دریافت می‌کنند. این کار باعث می‌شود که مدل بتواند بهتر بر روی نقاط ضعف خود تمرکز کند و عملکرد کلی را بهبود دهد.

در بخش بعدی بهینه‌سازی این رویکرد، منظم‌سازی [2] یکی از ابزارهای مهم در جلوگیری از بیش‌برازش [3] در مدل‌های یادگیری ماشین است. بر این اساس، XGBoost به‌طور پیش‌فرض از تکنیک‌های منظم‌سازی L1 یا Lasso و L2 یا Ridge استفاده می‌کند. در oXGBoost، علاوه بر این شیوه‌ها، از روش Elastic Net نیز استفاده می‌شود که ترکیبی از L1 و L2 است که باعث کاهش همبستگی بین ویژگی‌ها و جلوگیری از بیش‌برازش می‌شود. این روش به مدل کمک می‌کند تا همزمان از مزایای هر دو نوع منظم‌سازی بهره ببرد و به تعمیم بهتری دست یابد.

در XGBoost استاندارد، نمونه‌ها با وزن ثابت در تکرارهای آموزشی پردازش می‌شوند، اما در نسخه بهینه‌شده‌ی oXGBoost، نمونه‌گیری تطبیقی به مدل این امکان را می‌دهد تا بر روی نمونه‌هایی تمرکز کند که طبقه‌بندی آن‌ها دشوارتر است. به این ترتیب، هر نمونه‌ای که در تکرارهای قبلی به درستی طبقه‌بندی نشده یا به عنوان یک نمونه چالشی شناسایی شده، در تکرارهای بعدی وزن بیشتری دریافت می‌کند. فرض برآن است که w_i وزن اولیه نمونه i باشد. اگر نمونه i در تکرار t به اشتباه طبقه‌بندی شود، وزن آن در تکرار t+1 به صورت زیر تنظیم می‌شود:

$$w_i^{(t+1)} \& = w_i^{(t)} \cdot (1 + \alpha) \quad (۱۵)$$

در اینجا α یک مقدار افزایشی است که بر اساس میزان خطای مدل تنظیم می‌شود. این فرایند به مدل کمک می‌کند تا تمرکز بیشتری بر روی نمونه‌هایی که به خطا منجر می‌شوند داشته باشد و نقاط ضعف خود را بهتر شناسایی کند. به‌عبارت دیگر، با این روش تطبیقی، مدل در هر تکرار تلاش می‌کند با توجه بیشتر به نمونه‌های دشوار، عملکرد کلی خود را بهبود دهد. در XGBoost استاندارد، منظم‌سازی L1 و L2 به‌طور پیش‌فرض برای جلوگیری از بیش‌برازش استفاده می‌شوند. این تکنیک‌های منظم‌سازی به مدل کمک می‌کنند که از ایجاد وابستگی

[1] Gradient Boosting
[2] Regularization
[3] Overfitting

بیش از حد به ویژگی‌های خاص اجتناب کرده و تعمیم‌پذیری بهتری داشته باشد. با این حال، در نسخه بهینه‌شده‌ی oXGBoost، علاوه بر این تکنیک‌ها، از منظم‌سازی Elastic Net نیز استفاده می‌شود. بر این اساس، Elastic Net ترکیبی از L1 و L2 است که به شکل زیر تعریف می‌شود:

$$\Omega(\theta)\& = \lambda_1 \sum_{j=1}^{p} |\theta_j| + \lambda_2 \sum_{j=1}^{p} \theta_j^2 \qquad (۱۶)$$

که در آن، λ_1 و λ_2 ضرایب منظم‌سازی هستند که میزان تاثیر هر یک از تکنیک‌ها را کنترل می‌کنند. از سویی، $|\theta_j|$ میزان جریمه L1 است که برای انتخاب ویژگی‌ها و کاهش تعداد آن‌ها موثر واقع می‌گردد. در نهایت، θ_j^2 میزان جریمه L2 است که وابستگی بین ویژگی‌ها را کاهش داده و موجب کاهش بیش‌برازش می‌شود. ترکیب دوگانه به oXGBoost این امکان را می‌دهد که از مزایای هر دو نوع منظم‌سازی بهره‌مند شود. در این روند، Elastic Net علاوه بر کاهش همبستگی بین ویژگی‌ها، به مدل کمک می‌کند تا ویژگی‌های غیرضروری را حذف کرده و تعمیم‌پذیری مدل را افزایش دهد. این منظم‌سازی بهینه‌شده در oXGBoost، مقاومت مدل در برابر بیش‌برازش را بیشتر می‌کند، به ویژه در مسائل پیچیده‌ای که داده‌ها همبستگی بالایی دارند.

۲-۵-۳- مدل بهینه نهایی

در این مرحله، ترکیب دو روش oPSO و oXGBoost برای طبقه‌بندی پروتئین‌های بالقوه دارویی توضیح داده می‌شود. این ترکیب به عنوان یک مدل هیبریدی عمل می‌کند که از قدرت بهینه‌سازی PSO و توانایی‌های XGBoost برای رسیدن به دقت و کارایی بالا بهره می‌برد. یکی از چالش‌های اصلی در استفاده از XGBoost، تنظیم بهینه پارامترها برای بهبود عملکرد مدل است. پارامترهایی نظیر نرخ یادگیری[1]، تعداد درخت‌ها، عمق درخت‌ها و پارامترهای منظم‌سازی نقش مهمی در کارایی مدل دارند. در oXGBoost، این پارامترها به‌طور پویا و با استفاده از oPSO تنظیم می‌شوند. این فرآیند تنظیم به صورت پویا به این معناست که الگوریتم PSO به‌طور مداوم پارامترهای بهینه را در طول آموزش به‌روزرسانی می‌کند. این امر منجر به افزایش دقت و کاهش خطای مدل نهایی می‌شود.

هر ذره در این جمعیت نماینده‌ای از یک مجموعه پارامترهای ممکن برای طبقه‌بند بهینه شده oXGBoost است. در فرآیند بهینه‌سازی پارامترهای oXGBoost توسط نسخه بهبود داده شده از الگوریتم ازدحام ذرات، هدف یافتن مجموعه‌ای از پارامترها است که دقت و عملکرد مدل را به حداکثر برساند. به همین دلیل، oPSO این فرایند را با استفاده از تابع هدفی که بر اساس معیارهایی مانند دقت طبقه‌بندی و F1-score تعریف شده، انجام می‌دهد. هر ذره در oPSO نمایانگر ترکیبی از پارامترهای کلیدی مانند نرخ یادگیری، تعداد درخت‌ها و عمق درخت‌ها است. در هر تکرار، موقعیت ذرات بر اساس نتایج به‌دست‌آمده به‌روزرسانی می‌شود تا مدل به تدریج به بهینه‌ترین ترکیب پارامترها دست یابد و در برابر داده‌های متنوع تعمیم‌پذیری بهتری داشته باشد.

بر این اساس، oPSO به صورت پویا پارامترها را به‌روزرسانی می‌کند تا بهترین مجموعه پارامترها برای oXGBoost به دست آید. این به‌روزرسانی شامل تغییرات در سرعت و موقعیت ذرات است که در معادلات (۱۰) و (۱۱) توضیح داده شد. همچنین، oPSO به تدریج ذرات را به سمت بهینه‌های محلی یا جهانی هدایت می‌کند. پس از اتمام فرآیند بهینه‌سازی، مدل نهایی برای طبقه‌بندی oPSO-oXGBoost با پارامترهای بهینه برای طبقه‌بندی پروتئین‌های بالقوه دارویی استفاده می‌شود. این مدل قادر است با دقت بالایی پروتئین‌هایی را که پتانسیل دارویی دارند شناسایی کند. از آنجا که این روش ترکیبی از بهترین پارامترها و ساختار مدل است نتایج حاصل از آن بهبود قابل توجهی نسبت به روش‌های استاندارد دربرداشته است.

[1] Learning rate

۳- یافته‌ها و بحث

در این بخش به ارائه نتایج و یافته‌های تحقیق پرداخته می‌شود و سپس بحث و بررسی آنها مد نظر خواهد بود.

۳-۱- داده‌ها وتنظیمات

یک سیستم‌عامل ۶۴ بیتی با ۴ گیگابایت حافظه RAM برای پردازنده‌های Intel Core و Core i7 مورد استفاده قرار گرفت. برای کاهش توالی‌ها و ایجاد بردارهای نشانگر برای توالی‌های پروتئین، از الگوریتم انتخاب ویژگی پیشنهادی که در بخش پیش توضیح داده شد، بهره گرفته شد. این الگوریتم شامل مراحل محاسبه ماتریس همبستگی پویا، خوشه‌بندی ویژگی‌ها، انتخاب نماینده از هر خوشه و تحلیل پیچیدگی مشارکتی است. در نتیجه، بردار ویژگی ۱۶۸ بعدی برای هر توالی پروتئین ایجاد شد، که پیش‌تر به عنوان یک توالی ۴۰۰ تایی از دنباله پروتئین فرض شده بود. این بردار ویژگی با استفاده از تحلیل تعداد رخدادها و جفت رخدادهای آمینواسیدها در پنج ویژگی فیزیکوشیمیایی شامل قطبیت، اسیدیته، بار، ساختار ثانویه و دی‌هیدروپیریدین به دست آمده است. این روش امکان توصیف دقیق ساختارهای متنوع و عملکردهای پروتئین را فراهم می‌کند و با کاهش ویژگی‌های غیرضروری، دقت و کارایی مدل را بهبود می‌بخشد.

برای اجرای روش پیشنهادی oPSO-oXGBoost، تعداد ذرات الگوریتم ازدحام ذرات بهینه‌شده به ۵۰ و تعداد تکرارها به ۱۰۰ تنظیم شد. وزن اینرسی یا ω در بازه ۰/۴ تا ۰/۹ و ثابت‌های یادگیری یعنی c_1 و c_2 در بازه ۱/۵ تا ۲/۵ تنظیم شدند. پارامترهای اولیه XGBoost شامل نرخ یادگیری ۰.۱، تعداد درخت‌ها ۱۰۰، عمق درخت‌ها ۶ و پارامتر منظم‌سازی L2 برابر با ۱ تنظیم شد. معیار بهینه‌سازی برای الگوریتم ازدحام ذرات بهینه‌شده، دقت مدل انتخاب شد.

داده‌ها به دو بخش آموزشی و آزمایشی تقسیم و از روش اعتبارسنجی متقابل ۱۰ برابری استفاده شد. عملکرد مدل با استفاده از معیارهای دقت، F-score، دقت (Precision)، فراخوانی (Recall) و نیز در نهایت، مساحت زیر منحنی عامل گیرنده (AUC-ROC) ارزیابی گردید. مدل پیشنهادی oPSO-oXGBoost بر روی مجموعه داده اجرا شد و نتایج به دست آمده برای اطمینان از پایداری و تکرارپذیری چندین بار بررسی شدند. علاوه بر این، نتایج مدل پیشنهادی با روش‌های متداولی مانند ماشین بردار پشتیبان و نسخه‌های استاندارد XGBoost مقایسه شد. در نهایت، تنظیمات بهینه مدل ارائه و نتایج با تحقیقات مشابه در این حوزه مقایسه شده و جمع‌بندی‌های کلیدی گزارش گردیدند.

۳-۲- نتایج

برای ارزیابی روش و تحلیل صحیح عملکرد، معیارهای مختلفی در شناسایی پروتئین‌های بالقوه مورد توجه قرار گرفت. علاوه بر این، به منظور توجیه عملکرد روش پیشنهادی، مقایسه‌ای میان این روش و دیگر شیوه‌های مشابه صورت گرفت که شامل ۶ مدل مختلف است. لازم به ذکر است که مراحل پیش‌پردازش و استخراج ویژگی‌ها، از جمله استفاده از الگوریتم انتخاب ویژگی مبتنی بر خوشه‌بندی و تحلیل پیچیدگی مشارکتی، در تمامی این شیوه‌ها به صورت مشترک اعمال شده است. این روش‌ها عبارت‌اند از:

الف) PSO-SVMRBF یا طبقه‌بندی با ماشین بردار پشتیبان دارای کرنل تابع پایه شعاعی، همراه با تنظیم پارامترها توسط الگوریتم ازدحام ذرات استاندارد.

ب) PSO-RandForest یا طبقه‌بندی با جنگل تصادفی، با تنظیم پارامترها توسط الگوریتم ازدحام ذرات استاندارد.

ج) PSO-XGBoost یا طبقه‌بندی با XGBoost استاندارد و تنظیم پارامترها توسط الگوریتم ازدحام ذرات استاندارد.

د) oPSO-SVMRBF یا طبقه‌بندی با ماشین بردار پشتیبان دارای کرنل تابع پایه شعاعی، همراه با تنظیم پارامترها توسط الگوریتم ازدحام ذرات بهینه‌شده.

و) oPSO-RandForest یا طبقه‌بندی با جنگل تصادفی و تنظیم پارامترها توسط الگوریتم ازدحام ذرات بهینه‌شده.

ی) oPSO-oXGBoost یا طبقه‌بندی با XGBoost بهینه‌شده و تنظیم پارامترها توسط الگوریتم ازدحام ذرات بهینه‌شده.

مقایسه عملکرد روش پیشنهادی oPSO-oXGBoost با مدل‌هایی مشابه و توانمندتری چون SVM-RBF و Random Forest اهمیت زیادی دارد، زیرا این مدل‌ها نمایانگر رویکردهای متفاوتی در طبقه‌بندی هستند و هر یک نقاط قوت و ضعف خاص خود را دارند. این مقایسه به درک بهتر قابلیت‌های مدل پیشنهادی در مواجهه با داده‌های پیچیده و چندمتغیره کمک می‌کند و نقاط تمایز آن را برجسته می‌سازد. نتایج این مقایسه‌ها، در کنار ارزیابی عملکرد با معیارهایی نظیر دقت، F-score، دقت (Precision)، فراخوانی (Recall)، و مساحت زیر منحنی (AUC-ROC)، نشان می‌دهند که روش پیشنهادی می‌تواند با دقت و پایداری بالایی در شناسایی پروتئین‌های بالقوه عمل کند و از روش‌های مشابه برتری یابد.

مدل PSO-SVMRBF با دقت کلی ۹۵/۱۳٪، اختصاصیت ۹۵/۲۱٪، و حساسیت ۹۵/۰۳٪ عملکرد مناسبی داشته، اما در مقایسه با مدل‌های پیشرفته‌تر مانند oPSO-oXGBoost در معیارهایی مانند F-score و AUC که به ترتیب برابر ۹۵/۱۷٪ و ۹۵/۱۲٪ است، ضعیف‌تر عمل کرده است. مدل PSO-RandForest، دقت بهتری با میانگین ۹۵/۴۲٪ نشان داده و با اختصاصیت ۹۵/۵۰٪ و حساسیت ۹۵/۲۵٪، تعادل بهتری برقرار کرده است. این مدل با امتیاز F معادل ۹۵/۴۶٪ و میزان AUC معادل ۹۵/۳۶٪ همچنان عملکرد بهتری نسبت به PSO-SVMRBF دارد، اما در مقایسه با مدل‌های بهینه‌شده مانند oPSO-RandForest و oPSO-oXGBoost همچنان محدودیت‌هایی دارد.

مطابق جدول ۲، مدل oPSO-oXGBoost بهترین عملکرد را با دقت میانگین ۹۶/۵۹٪، اختصاصیت ۹۷/۳۰٪، و حساسیت ۹۵/۸۸٪ نشان داده است. همچنین، امتیاز F-score این مدل ۹۶/۲۳٪ و مقدار AUC آن ۹۶/۶۱٪ بوده که نشان‌دهنده توازن بالا میان تشخیص مثبت و منفی و حداقل کردن خطاها است. مدل oPSO-RandForest نیز با دقت ۹۵/۹۳٪ و اختصاصیت ۹۶/۰۲٪ عملکرد مناسبی داشته، اما از oPSO-oXGBoost ضعیف‌تر عمل کرده است.

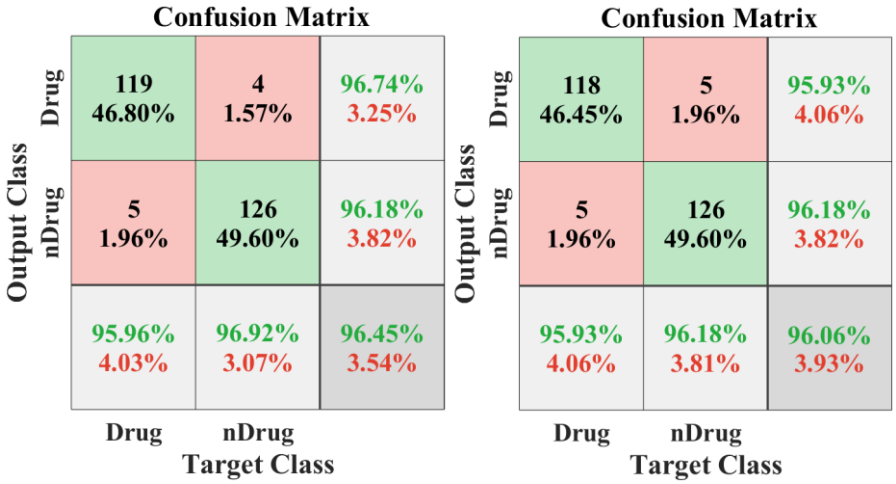
جدول ۲- نتایج مقایسه عملکرد مدل‌های مختلف برای شناسایی پروتئین‌های بالقوه در طراحی دارو. هر مدل در سه تکرار آزمایش مورد ارزیابی قرار گرفته و معیارهای چندگانه برای هر تکرار محاسبه شده است.

مدل	تکرار آزمایش	دقت	اختصاصیت	حساسیت	امتیاز F	مقدار AUC
PSO-SVMRBF	آزمایش ۱	۹۵/۰۲	۹۵/۱۰	۹۴/۹۵	۹۵/۰۶	۹۵/۰۳
	آزمایش ۲	۹۵/۱۳	۹۵/۲۱	۹۵/۰۰	۹۵/۱۷	۹۵/۱۰
	آزمایش ۳	۹۵/۲۵	۹۵/۳۲	۹۵/۱۵	۹۵/۲۸	۹۵/۲۳
	میانگین	**۹۵/۱۳**	**۹۵/۲۱**	**۹۵/۰۳**	**۹۵/۱۷**	**۹۵/۱۲**
PSO-RANDFORCE	آزمایش ۱	۹۵/۳۵	۹۵/۴۳	۹۵/۲۰	۹۵/۳۹	۹۵/۳۲
	آزمایش ۲	۹۵/۴۴	۹۵/۵۲	۹۵/۲۵	۹۵/۴۸	۹۵/۳۹
	آزمایش ۳	۹۵/۴۸	۹۵/۵۶	۹۵/۳۰	۹۵/۵۲	۹۵/۴۳
	میانگین	**۹۵/۴۲**	**۹۵/۵۰**	**۹۵/۲۵**	**۹۵/۴۶**	**۹۵/۳۸**
PSO-XGBOOST	آزمایش ۱	۹۵/۵۵	۹۵/۶۳	۹۵/۴۱	۹۵/۵۹	۹۵/۵۱
	آزمایش ۲	۹۵/۵۷	۹۵/۶۵	۹۵/۳۵	۹۵/۶۱	۹۵/۵۰
	آزمایش ۳	۹۵/۶۱	۹۵/۶۹	۹۵/۴۰	۹۵/۶۵	۹۵/۵۴
	میانگین	**۹۵/۵۸**	**۹۵/۶۶**	**۹۵/۳۹**	**۹۵/۶۲**	**۹۵/۵۲**
OPSO-SVMRBF	آزمایش ۱	۹۵/۷۰	۹۵/۷۹	۹۵/۵۵	۹۵/۷۴	۹۵/۶۷
	آزمایش ۲	۹۵/۸۲	۹۵/۹۰	۹۵/۶۰	۹۵/۸۶	۹۵/۷۵
	آزمایش ۳	۹۵/۷۵	۹۵/۸۳	۹۵/۵۵	۹۵/۷۹	۹۵/۶۹
	میانگین	**۹۵/۷۶**	**۹۵/۸۴**	**۹۵/۵۷**	**۹۵/۸۰**	**۹۵/۷۰**
OPSO-RANDFORCE	آزمایش ۱	۹۶/۰۰	۹۶/۰۸	۹۵/۸۰	۹۶/۰۴	۹۵/۹۴
	آزمایش ۲	۹۵/۹۱	۹۶/۰۱	۹۵/۷۰	۹۵/۹۵	۹۵/۸۴
	آزمایش ۳	۹۵/۸۸	۹۵/۹۶	۹۵/۷۰	۹۵/۹۲	۹۵/۸۳
	میانگین	**۹۵/۹۳**	**۹۶/۰۲**	**۹۵/۷۳**	**۹۵/۹۷**	**۹۵/۸۷**
OPSO-OXGBOOST	آزمایش ۱	۹۶/۴۸	۹۶/۷۵	۹۶/۲۱	۹۶/۳۴	۹۶/۴۸
	آزمایش ۲	۹۶/۵۶	۹۷/۴۶	۹۵/۶۶	۹۶/۱۱	۹۶/۵۶
	آزمایش ۳	۹۶/۷۳	۹۷/۶۸	۹۵/۷۸	۹۶/۲۵	۹۶/۸۴
	میانگین	**۹۶/۵۹**	**۹۷/۳۰**	**۹۵/۸۸**	**۹۶/۲۳**	**۹۶/۶۱**

یکی از نکات کلیدی روش پیشنهادی، تأثیر انتخاب ویژگی است که با استفاده از تحلیل خوشه‌بندی و انتخاب نماینده‌ها توانسته است دقت مدل را به‌طور معناداری بهبود بخشد. مقایسه بین PSO-XGBoost و oPSO-XGBoost نشان می‌دهد که بهینه‌سازی پارامترها و انتخاب ویژگی‌ها توانسته است دقت را به‌طور متوسط ۲ درصد افزایش و خطای کل را ۱ درصد کاهش دهد.

برای روش پیشنهادی در آزمایش‌های انجام‌شده، دقت میانگین مدل‌ها حدود ۹۶/۶ درصد با نوسانی در حدود ۰/۴۳٪ بوده است. کمترین دقت ثبت‌شده ۹۵/۹۶٪ و بیشترین آن ۹۶/۸۱٪ با انحراف استاندارد ۰/۱۷٪ بوده است. این نتایج نشان‌دهنده پایداری و ثبات روش پیشنهادی است. همچنین، مقاومت روش در برابر پیچیدگی و نویز داده‌ها، توانایی آن را برای گسترش به حوزه‌های متنوع و استفاده در داده‌های جدید نشان می‌دهد. استفاده از تکنیک‌هایی مانند Elastic Net نیز با

کاهش بیش‌برازش و افزایش تعمیم‌پذیری، به بهبود عملکرد مدل در شرایط پیچیده کمک کرده است. به طور کلی، نتایج نشان می‌دهند که روش پیشنهادی با انعطاف‌پذیری بالا و بهبود معیارهای عملکرد، گزینه‌ای قدرتمند برای شناسایی پروتئین‌های بالقوه دارویی است. به دلیل محدود بودن تعداد نمونه‌های مرتبط با مطالعه، مدل‌ها انحراف استاندارد بیشتری را بین ورودی‌ها نشان می‌دهند. شکل ۲ عملکرد مدل بر روی مجموعه داده‌های نمونه دارویی و پیش‌بینی نتایج بهینه را به صورت یک ماتریس تداخل نشان می‌دهد. انعطاف‌پذیری به ویژه در مواجهه با پیچیدگی‌ها و تغییرات ذاتی داده‌ها در شرایط واقعی اهمیت پیدا می‌کند.



شکل ۲- دو نمونه از نتایج طبقه‌بندی حاصل از ساختار معرفی‌شده (مدل ۶) که به صورت ماتریس تداخل نمایش داده شده‌اند.

۳-۳- بحث و تفسیر

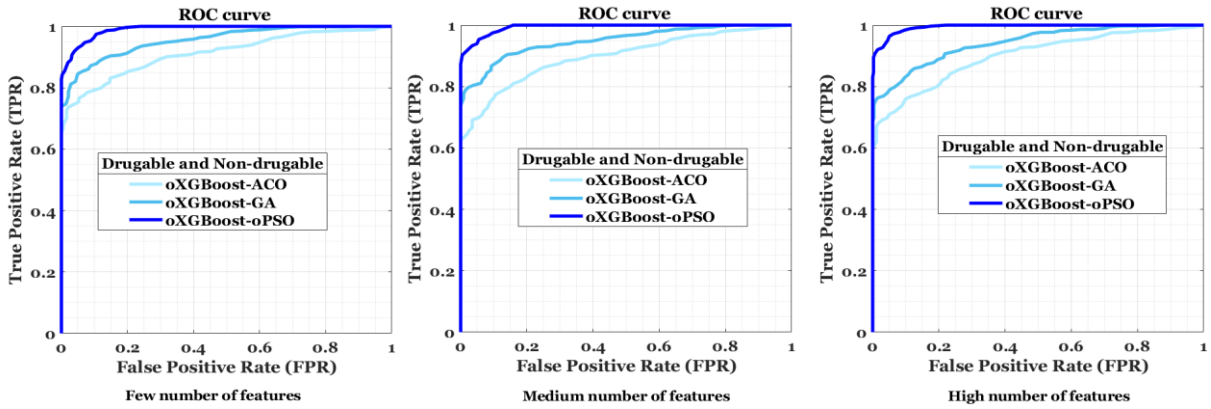
آزمایشات برای محاسبه و ارزیابی عملکرد منحنی ROC با استفاده از داده‌های آموزش و اعتبارسنجی انجام شدند. منحنی ROC که در شکل ۳ نمایش داده شده، نشان می‌دهد که چگونه یک سیستم طبقه‌بندی دودویی می‌تواند کلاس‌ها را در سطوح مختلف آستانه برای تعداد ویژگی‌ها و بررسی آنها تفکیک کند. در این شکل، سه سطح مختلف از تعداد ویژگی‌ها در نظر گرفته شده است تا تأثیر کاهش ویژگی‌های غیرضروری بر دقت مدل و تعادل میان نرخ‌های مثبت و منفی صحیح بررسی شود. نمودارهای ROC نشان می‌دهند که روش پیشنهادی، به دلیل بهره‌گیری از الگوریتم انتخاب ویژگی مبتنی بر خوشه‌بندی و تحلیل پیچیدگی مشارکتی، عملکرد بهتری در تمام سطوح ویژگی‌ها دارد. این امر نشان‌دهنده توانایی روش در بهینه‌سازی مدل حتی با تعداد ویژگی‌های کمتر است

در این میان، دقت ۹۶/۶ درصد برای سیستم طبقه‌بندی گزارش شده است. همچنین، AUC یا مساحت زیر منحنی ROC نشان‌دهنده توانایی مدل در تمایز بین کلاس‌ها است و هر چه AUC بیشتر باشد، عملکرد مدل در تفکیک پروتئین‌های بالقوه برای داروی موثر بهتر است. روش پیشنهادی در این تحقیق، با استفاده از الگوریتم oPSO برای بهینه‌سازی پارامترها، سطح AUC بیشتری نسبت به روش‌های دیگر به دست آورده است. همگرایی الگوریتم در نیل به بهترین پاسخ ممکن و تنظیم پارامترهای شبکه‌های طبقه‌بندی نیز در شکل ۴ نشان داده شده است، که نشان‌دهنده موفقیت الگوریتم در یافتن مقادیر بهینه و کاهش تغییرات پارامتری تا حد ممکن است. این روش با سرعت بیشتری به مقادیر بهینه همگرا می‌شود و دقت بالاتری را در مقایسه با ساختارهای ساده‌تر نشان می‌دهد. اگرچه پیچیدگی محاسباتی این رویکرد تا حدودی بیشتر است، اما تفاوت آن ناچیز است و با تکرارهای بیشتر، اثربخشی بهبود یافته آن آشکار می‌شود. مضاف بر این، تنظیم پارامترهای oXGBoost در فاز آموزش صورت می‌پذیرد و نیاز نیست در گام آزمایش، پارامتری تنظیم گردد. در طراحی و ارزیابی روش پیشنهادی، نیاز به ایجاد یک مصالحه میان دقت، زمان و عملکرد کلی مدل وجود داشت. با توجه به چالش‌های موجود در مسائل پیچیده مانند تشخیص پروتئین‌های دارویی، تمرکز اصلی ما بر بهبود تعمیم‌پذیری و غلبه بر چالش عدم قطعیت و همچنین کاهش خطر بیش‌برازش بود. هدف روش پیشنهادی آن است که با افزایش دقت پیش‌بینی‌ها، عملکرد مدل در مواجهه با داده‌های جدید بهینه شود، حتی اگر زمان محاسباتی کمی افزایش یابد. بالا بودن پیچیدگی محاسباتی روش به‌دلیل تنظیمات پویا و بهینه‌سازی پارامترها در مرحله آموزش است که نتایج قابل‌اعتمادی را در کاربردهای پیچیده فراهم می‌آورد. در جدول ۳، معیارهای زمان آموزش، زمان آزمایش، پاسخ زمانی برای هر نمونه، پیچیدگی زمانی کل و تعداد پارامترهای تنظیم‌شده در مدل‌های مختلف بررسی شده‌اند. برای سنجش این معیارها، زمان‌های محاسباتی با اجرای مدل‌ها بر روی مجموعه داده‌ای از نمونه‌های متنوع اندازه‌گیری شد و پاسخ زمانی برای هر نمونه به‌صورت میانگین محاسبه گردید. مشاهده می‌شود که روش پیشنهادی oPSO-oXGBoost، با وجود پیچیدگی زمانی متوسط، دقت طبقه‌بندی و عملکرد تعمیم‌پذیر بالایی را نیز ارائه می‌دهد که نشان‌دهنده‌ی موفقیت مدل در توازن میان دقت و زمان است. این روش، به ویژه در مسائلی که نیاز به تحلیل‌های دقیق و پایدار دارند، عملکرد مناسبی از خود نشان داده است و این مزایا استفاده از آن را توجیه می‌کنند. در خصوص چالش ابعاد ویژگی‌ها در مورد تعداد و اثر ویژگی‌های استخراج شده در شکل ۳ یا منحنی‌های عامل گیرنده مورد بررسی قرار گرفت.

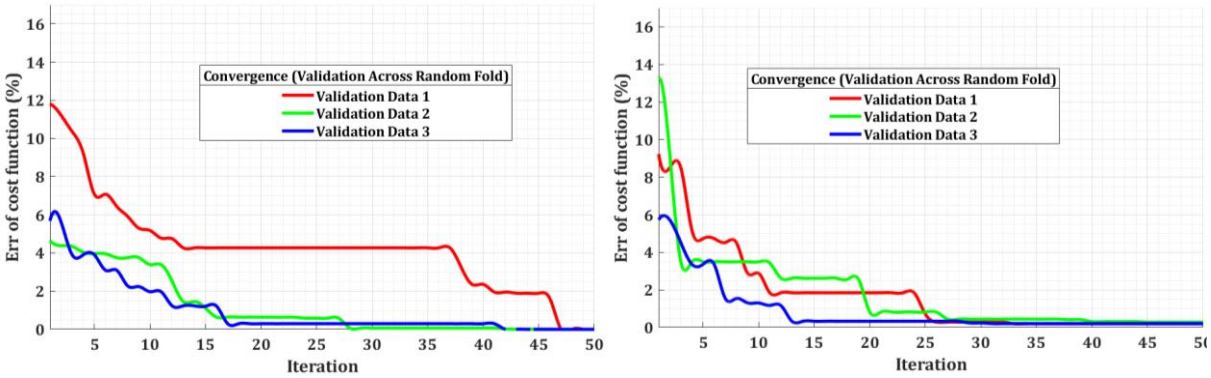
جدول ۳ نشان‌دهنده مقایسه عملکرد زمانی مدل‌ها بر اساس معیارهای مختلف شامل زمان آموزش، زمان آزمایش، پاسخ زمانی برای هر نمونه، پیچیدگی زمانی کل، و تعداد پارامترهای تنظیم‌شده است. در این جدول، مدل oPSO-oXGBOOST به عنوان روش پیشنهادی، علی‌رغم اینکه از نظر صرف زمانی در جایگاه سوم قرار دارد، مصالحه‌ای رضایت بخشی بین دقت و زمان انجام داده است. این روش توانسته با بهره‌گیری از بهینه‌سازی پارامترهای متعددی (۶ پارامتر)، عملکردی کارآمد و رقابتی را ارائه دهد. هرچند همه روش‌ها نیز زیر یک ثانیه پاسخ زمانی داشته‌اند، می‌توان روش‌ها و در کنار آن شیوه پیشنهادی را بلادرنگ دانست. هرچند انتظار می‌رفت که زمان این مدل به سبب پیچیدگی جستجوی پویا و فرآیندهای سنگین XGBoost بسیار بیشتر باشد، اما تنظیمات مناسب و بهینه‌سازی‌های دقیق باعث شده که زمان صرف شده برای آموزش و آزمایش در محدوده مطلوبی باقی بماند. از سوی دیگر، سرعت بالای این روش با ترکیب کارآمد oXGBoost و الگوریتم جستجوی ذرات حاصل شده است که در مقایسه با سایر روش‌ها، عملکرد متعادل‌تری از لحاظ دقت و زمان ارائه می‌دهد.

معیارهای ارزیابی شامل امتیاز F1، دقت، حساسیت و اختصاصیت برای طبقه‌بندی، یک تقسیم‌بندی از ویژگی‌ها در شکل ۵ نشان داده شده است. در این پژوهش، از

یک الگوریتم نوآورانه انتخاب ویژگی مبتنی بر خوشه‌بندی و تحلیل پیچیدگی مشارکتی استفاده شده است که به کاهش تعداد ویژگی‌های غیرضروری و افزایش کارایی مدل کمک کرده است. این روش توانسته است با حفظ دقت بالا، پیچیدگی محاسباتی را نیز به میزان قابل توجهی کاهش دهد. ویژگی‌ها به شش دسته تقسیم شدند و معیارهای اساسی برای هر دسته اندازه‌گیری شد. اگرچه بالاترین دقت در گروه ۵ (ویژگی‌های ۱ تا ۳۲۵) به دست آمد، اما الگوریتم پیشنهادی با بهره‌گیری از تحلیل خوشه‌بندی توانسته است اهمیت واقعی هر ویژگی را ارزیابی کرده و با ترکیب مؤثر ویژگی‌های کلیدی، دقت بالایی را بدون نیاز به استفاده از همه ویژگی‌ها فراهم کند.



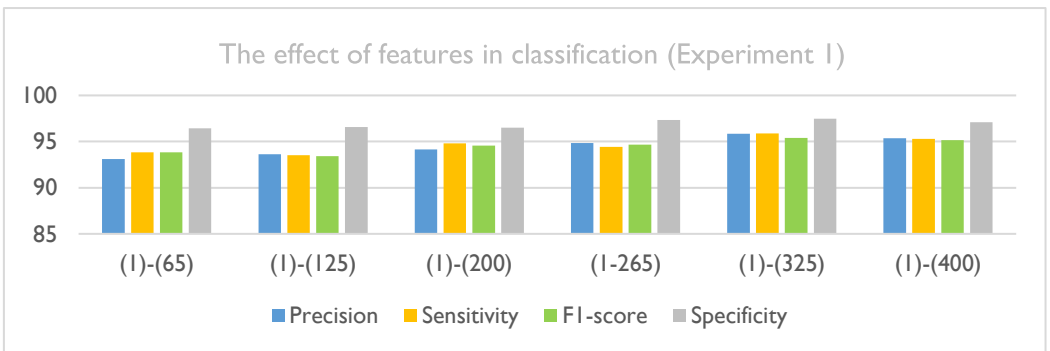
شکل ۳- نشان‌دهنده تأثیر انتخاب ویژگی بر عملکرد مدل پیشنهادی در مقایسه با دو روش بهینه‌سازی دیگر است.

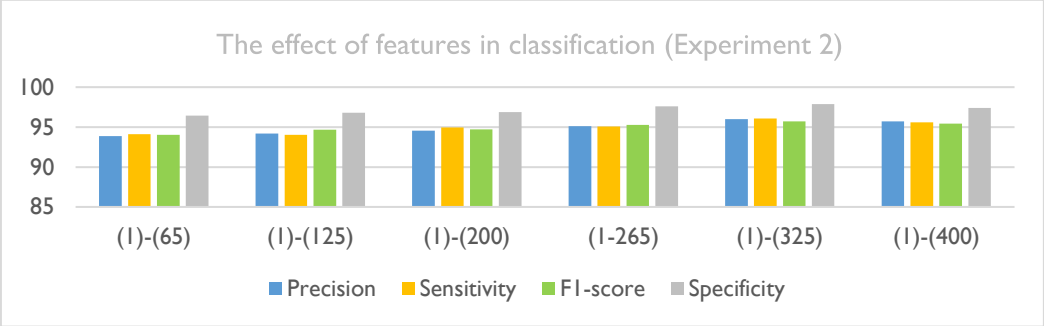


شکل ۴- همگرایی و خطای کم روش oPSO برای شش دسته داده تصادفی شامل داده‌های اعتبار در دو فولد تصادفی و برای تنظیم عملکرد oXGBoost در طبقه‌بندی پروتئین‌های بالقوه دارو به نمایش درآمده است.

جدول ۳- مقایسه مدل‌های مختلف بر اساس زمان آموزش، زمان آزمایش، پاسخ زمانی برای هر نمونه، پیچیدگی زمانی کل، و تعداد پارامترهای تنظیم‌شده

مدل	زمان آموزش (ثانیه)	زمان آزمایش (ثانیه)	پاسخ زمانی برای هر نمونه (میلی‌ثانیه)	پیچیدگی زمانی کل	تعداد پارامترهای تنظیم‌شده
PSO-SVMRBF	۴۹۱/۷۸۱	۹/۰۱	۳۰۳	متوسط	۵
PSO-RANDFOREST	۲۱۲/۱۹۵	۷/۶۲	۱۸۶	کم	۴
PSO-XGBOOST	۴۶۸/۹۲۶	۸/۸۵	۲۵۵	متوسط	۶
OPSO-SVMRBF	۵۱۳/۶۵۸	۱۲/۷۶	۲۸۴	متوسط	۵
OPSO-RANDFOREST	۲۰۸/۸۹۳	۶/۱۲	۲۰۳	کم	۴
OPSO-OXGBOOST	۴۹۴/۰۸۴	۸/۳۲	۲۳۸	متوسط	۶





شکل ۵- استفاده از روش انتخاب ویژگی پیشنهادی ارزیابی اثر آنها در ۶ دسته متفاوت برای بررسی دقت طبقه‌بندی مطابق با ۶ معیار ارزیابی

در محدوده ۱ تا ۴۰۰ ویژگی استخراج‌شده از نمونه‌های پروتئین بالقوه برای توسعه دارو، ویژگی‌ها برای هر دسته به شرح زیر در نظر گرفته شدند: دسته یک از ویژگی ۱ تا ۶۵، دسته دو از ویژگی ۱ تا ۱۲۵، دسته سه از ویژگی ۱ تا ۲۰۰، دسته چهار از ویژگی ۱ تا ۲۶۵، دسته پنج از ویژگی ۱ تا ۳۲۵ و دسته شش از ویژگی ۱ تا ۴۰۰. آزمایش دو بار تکرار شد و در هر دو تکرار، دسته پنج (ویژگی‌های ۱ تا ۳۲۵) منجر به خطاهای کمتری شد. الگوریتم انتخاب ویژگی توانسته است ضمن حذف ویژگی‌های اضافی، پیچیدگی محاسباتی ناشی از پردازش کل داده‌ها را کاهش داده و عملکرد کلی مدل را بهبود بخشد.

در مدل نهایی، روش پیشنهادی با ترکیب بهینه ویژگی‌های کلیدی از دسته‌های مختلف، نه تنها دقت بالایی را حفظ کرده، بلکه پیچیدگی محاسباتی را نیز کنترل کرده است. استفاده از این روش در کاهش تعداد ویژگی‌ها به حدود ۳۲۵ مورد، منجر به کاهش ۲۰ درصدی زمان محاسباتی شده است، در حالی که دقت مدل حتی در مقایسه با مدل‌هایی که از تمام ۴۰۰ ویژگی استفاده کرده‌اند، افزایش یافته است.

بر اساس جدول ۴، روش پیشنهادی این تحقیق در مقایسه با سایر روش‌های موجود، عملکرد برتری را در شناسایی پروتئین‌های بالقوه برای طراحی دارو نشان داده است. برخلاف روش‌های Jamali و همکاران [۲۷] و Lin و همکاران [۳۰] که از استخراج ویژگی‌های پراهمیت با پیچیدگی محاسباتی متوسط و تقسیم داده‌ها به روش ۵-فولد استفاده کرده‌اند، روش پیشنهادی با بهینه‌سازی انتخاب ویژگی توانسته است دقت بالاتری ارائه دهد و از چالش‌هایی مانند بیش‌برازش و عملکرد ضعیف در داده‌های حجیم جلوگیری کند. در حالی که تحقیقات پیشین با محدودیت‌هایی چون عدم تعمیم‌پذیری و تأثیر حجم زیاد ویژگی‌ها مواجه بوده‌اند، این روش توانسته است با کاهش مؤثر تعداد ویژگی‌ها، کارایی و تعمیم‌پذیری بهتری را ارائه دهد و به عنوان یک رویکرد پیشرو در شناسایی پروتئین‌های بالقوه دارویی مطرح شود.

جدول ۴. مقایسه نتایج مطالعات مختلف در طبقه‌بندی داروها در مقایسه با روش پیشنهادی با استفاده از ویژگی‌های استخراج‌شده از پایگاه داده مشابه

نویسنده	روش	داروهای مورد بررسی	دقت (٪)
JAMALI و همکاران [۲۷]	طبقه‌بندی با شیوه ماشین بردار پشتیبان	۴۴۳ ویژگی از داده‌های DrugBank	۸۹/۷۸
LIN و همکاران [۳۰]	طبقه‌بندی با شیوه ماشین بردار پشتیبان	۱۴۳ ویژگی از داده‌های DrugBank	۹۳/۷۸
CHEN و همکاران [۳۱]	طبقه‌بندی با روش XGBoost	۱۵۵ ویژگی از داده‌های DrugBank و کشف دارو برای دو نوع بیماری	۹۴/۶۴
SIKANDER و همکاران [۳۲]	طبقه‌بندی با روش DrugPred XGB-	تعداد بالای ویژگی‌های استخراج شده از پایگاه داده DrugBank	۹۴/۸۶
CHEN و همکاران [۳۳]	روش دسته‌بندی گراف‌های کانولوشنی عمیق	تعداد بالای ویژگی‌های استخراج شده از پایگاه داده DrugBank	۹۵/۰۰
روش پیشنهادی	انتخاب ویژگی با استفاده از خوشه‌بندی و تحلیل پیچیدگی مشارکتی و طبقه‌بندی با مدل oXGBoost بهینه‌شده توسط الگوریتم oPSO	۲۵۴۳ نمونه از داده‌های DrugBank و Swiss-Prot	۹۶/۶۰

در مقابل، روش‌های مدرن‌تری مانند XGBoost و XGB-DrugPred که توسط Chen و همکاران [۳۱] و Sikander و همکاران [۳۲] به‌کار گرفته شده‌اند، با وجود پیچیدگی محاسباتی متوسط و تقسیم داده‌ها به روش ۱۰-فولد، با مشکلاتی نظیر برون‌یابی مقادیر هدف و بیش‌برازش مواجه‌اند. همچنین، روش Chen و همکاران [۳۳] با استفاده از گراف کانولوشنی دقت بالایی را ارائه می‌دهد و توانایی پیش‌بینی با حجم زیاد داده‌ها را داراست، اما همچنان دارای چالش‌هایی در محاسبات و تولید گراف‌ها و احتمال بیش‌برازش است. در این میان، روش پیشنهادی این تحقیق که ترکیبی از XGBoost بهینه‌شده و الگوریتم ازدحام ذرات است، با حفظ پیچیدگی محاسباتی در سطح متوسط و ارائه دقت ۹۶/۱۰٪، تعادل مناسبی بین دقت و کارایی برقرار کرده است. با همه این تفاسیر، استفاده از شبکه‌های عصبی عمیق و تکنیک‌های یادگیری عمیق برای تحلیل داده‌های زیستی نیازمند توان پردازشی بالا و تنظیم مناسب پارامترهاست، که خود یکی از چالش‌های کلیدی در طراحی دارو محسوب می‌شود [۳۴].

۴- نتیجه‌گیری

در این مقاله، رویکردی نوآورانه برای بهبود دقت و کارایی پیش‌بینی پروتئین‌های بالقوه در فرآیند طراحی دارو معرفی شد. با ادغام انتخاب ویژگی با استفاده از خوشه‌بندی و تحلیل پیچیدگی مشارکتی و طبقه‌بندی با مدل oXGBoost

بهینه‌شده توسط الگوریتم oPSO، بهبود قابل توجهی در دقت پیش‌بینی‌ها به دست آمد. نتایج نشان می‌دهد که روش پیشنهادی می‌تواند به عنوان ابزاری موثر در شناسایی پروتئین‌های درمانی مورد استفاده قرار گیرد و به توسعه داروهای جدید و کارآمدتر کمک کند. این رویکرد پتانسیل دارد تا فرآیند کشف دارو را با کاهش زمان و هزینه‌ها و حفظ دقت بالا، بهینه کند. پژوهش‌های آینده در این زمینه می‌توانند بر روی بهبود بیشتر الگوریتم‌های یادگیری، افزایش دقت مدل‌ها و بررسی امکان‌پذیری گسترش این روش برای مدیریت مجموعه داده‌های بزرگتر و ساختارهای پروتئینی پیچیده‌تر تمرکز کنند. همچنین، ادغام شیوه‌های پیشرفته دیگر یادگیری ماشین با این راهکار می‌تواند به بینش‌های بیشتری منجر شود و در نهایت به ایجاد یک زنجیره طراحی داروی قوی‌تر و کارآمدتر کمک کند.

مراجع

[1] A. Lavecchia, "Deep learning in drug discovery: opportunities, challenges and future prospects," *Drug Discov. Today*, vol. 24, no. 10, pp. 2017-2032, Oct. 2019.

[2] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Fut. Healthc. J.*, vol. 6, no. 2, pp. 94-98, 2019.

[3] J. Peña-Guerrero, P. A. Nguewa, and A. T. García-Sosa, "Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 11, no. 5, p. e1513, 2021.

[4] B. Suay-Garcia, et al., "Quantitative structure–activity relationship methods in the discovery and development of antibacterials," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 10, no. 6, p. e1472, 2020.

[5] D. Jiang et al., "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *J. Cheminform.*, vol. 13, no. 1, pp. 1-23, 2021.

[6] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of artificial intelligence for computer-assisted drug discovery," *Chem. Rev.*, vol. 119, no. 18, pp. 10520-10594, 2019.

[7] D. Chowell et al., "Improved prediction of immune checkpoint blockade efficacy across multiple cancer types," *Nat. Biotechnol.*, vol. 40, no. 4, pp. 499-506, 2022.

[8] V. Svetnik, et al., "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947-1958, 2003.

[9] F. Rayhan et al., "iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting," *Sci. Rep.*, vol. 7, no. 1, pp. 1-18, 2017.

[10] M. Maniruzzaman and A. Nokhodchi, "Continuous manufacturing via hot-melt extrusion and scale up: regulatory matters," *Drug Discov. Today*, vol. 22, no. 2, pp. 340-351, 2017.

[11] Z. Liu et al., "AI-based language models powering drug discovery and development," *Drug Discov. Today*, vol. 26, no. 11, pp. 2593-2607, Nov. 2021.

[12] B. Tang, J. Ewalt, and H. L. Ng, "Generative AI models for drug discovery," in *Biophysical and Computational Tools in Drug Discovery*, Cham: Springer International Publishing, 2021, pp. 221-243.

[13] J. W. Lee et al., "Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD)," *Biochem. Soc. Trans.*, vol. 50, no. 1, pp. 241-252, Feb. 2022.

[14] S. Kolluri et al., "Machine learning and artificial intelligence in pharmaceutical research and development: a review," *AAPS J.*, vol. 24, Feb. 2022.

[15] L. K. Vora et al., "Artificial intelligence in pharmaceutical technology and drug delivery design," *Pharmaceutics*, vol. 15, no. 7, p. 1916, Jul. 2023.

[16] F. Ren et al., "AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor," *Chem. Sci.*, vol. 14, no. 6, pp. 1443-1452, 2023.

[17] A. López-Cortés, et al, "Unraveling druggable cancer-driving proteins and targeted drugs using artificial intelligence and multi-omics analyses," *Scientific Reports*, vol. 14, no. 1, pp. 19359, 2024.

[18] I. Myrko et al., "Current trends of chemoinformatics and computer chemistry in drug design: A review," *Curr. Chem. Lett.*, vol. 13, no. 1, pp. 151-162, 2024.

[19] O. Alghushairy, et al., "Machine learning-based model for accurate identification of druggable proteins using light extreme gradient boosting," *Journal of Biomolecular Structure & Dynamics*, vol. 2023, pp. 1-12, 2023.

[20] S. Zhou, Y. Li, and X. Zhang, "Optimization modeling of anti-breast cancer candidate drugs," *Biotechnology & Genetic Engineering Reviews*, vol. 2023, pp. 1-19, 2023.

[21] M. Zhang, F. Wan, and T. Liu, "DrugFinder: Druggable protein identification model based on pre-trained models and evolutionary information," *Algorithms*, vol. 16, no. 6, pp. 263, 2023.

[22] M. Subha Ramakrishnan and N. Ganapathy, "Extreme gradient boosting based improved classification of blood-brain-barrier drugs," *Studies in Health Technology and Informatics*, vol. 294, pp. 872-873, 2022.

[23] A. Ghosh, M. Talukdar, and U. K. Roy, "Stable drug designing by minimizing drug protein interaction energy using PSO," *arXiv preprint arXiv:1507.08408*, Jul. 2015.

[24] X. Zhan et al., "Prediction of drug-target interactions by ensemble learning method from protein sequence and drug fingerprint," *IEEE Access*, vol. 8, pp. 185465-185476, Sep. 2020.

[25] Z. U. Haq et al., "Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction," *Bioresour. Technol.*, vol. 363, p. 128008, Nov. 2022.

[26] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, suppl. 1, pp. D901-D906, 2008.

[27] A. A. Jamali, R. Ferdousi, S. Razzaghi, J. Li, R. Safdari, and E. Ebrahimie, "DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins," *Drug Discovery Today*, vol. 21, no. 5, pp. 718-724, 2016.

[28] Swiss-Prot: The manually annotated and reviewed protein sequence database, UniProt, 2023. [Online]. Available: https://www.uniprot.org/

[29] C. Xu, L. Ge, Y. Zhang, M. Dehmer, and I. Gutman, "Computational prediction of therapeutic peptides based on graph index," *J. Biomed. Inform.*, vol. 75, pp. 63-69, 2017.

[30] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier," *Artificial Intelligence in Medicine*, vol. 98, pp. 35-47, Jul. 2019.

[31] C. Chen, Q. Zhang, B. Yu, Z. Yu, P. J. Lawrence, Q. Ma, and Y. Zhang, "Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier," *Computers in Biology and Medicine*, vol. 123, p. 103899, 2020.

[32] R. Sikander, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," *Scientific Reports*, vol. 12, no. 1, p. 5505, Apr. 2022.

[33] J. Chen, Z. Gu, Y. Xu, M. Deng, L. Lai, and J. Pei, "QuoteTarget: A sequence-based transformer protein language model to identify potentially druggable protein targets," *Protein Science*, vol. 32, no. 2, p. e4555, 2023.

[34] S. S. Masoomkhah, K. Rezaee, M. Ansari, and H. Eslami, "Deep Learning in Drug Design—Progress, Methods, and Challenges," *Frontiers in Biomedical Technologies*, vol. 11, no. 3, pp. 492–508, Summer 2024.

Classification of Potential Proteins in Drug Design Using Optimized Learning and Dimensionality Reduction Based on Feature Clustering and Collaborative Analysis

Shiva Shekarchian[1], Hossein Eslami[2], Khosro Rezaee[2]*

* Corresponding Author | Received: 12/05/2024 | Revised: 19/01/2025 | Accepted: 20/05/2025

[1] Master's Graduate, Department of Biomedical Engineering, Faculty of Engineering, Meybod University, Meybod, Iran
[2] Assistant Professor, Department of Biomedical Engineering, Faculty of Engineering, Meybod University, Meybod, Iran

Abstract

In recent decades, rapid advancements in the fields of proteomics and drug design have heightened the need for a more precise understanding of protein structure and function. One of the primary challenges in this domain is the accurate prediction of potential proteins for designing more effective drugs. This study aims to enhance the accuracy and efficiency of predicting potential proteins through an efficient approach. A novel hybrid method is presented in this paper, which combines optimized XGBoost-based learning, the particle swarm optimization algorithm, and a new feature selection step based on feature clustering and collaborative complexity analysis. Following preprocessing and feature extraction, significant features are identified through clustering and the selection of key representatives. Subsequently, predictive models are trained using proteomic data and structural information about proteins. Finally, an enhanced version of the particle swarm optimization algorithm is utilized to optimize the parameters of XGBoost learning models. The data used in this study include proteomics and protein structures obtained from the DrugBank and Swiss-Prot databases. The results demonstrate that this approach significantly improves prediction accuracy, achieving a model accuracy of 96.6%. This innovative method facilitates the design of more effective drugs, reduces costs and time, and strengthens future research in the field.

Keywords: Drug Design, Potential Protein Classification, XGBoost Learning, Clustering, Collaborative Complexity Analysis.