13th International Conference on Information & Knowledge Technology
سیزدهمین کنفرانس بین المللی فناوری اطلاعات و دانش
۲۹ آذر ماه الی ۱ دی ماه ۱۴۰۱
"ایران هوشمند در پرتو فناوری اطلاعات و دانش"

شناسایی جایگاه مالونیلاسیون در پروتئین‌ها با بهره‌گیری از استخراج ویژگی و تکنیک‌های پردازش زبان طبیعی

حنانه رجبیون[۱]، محمد قاسم‌زاده[۲] و وحید رنجبر بافقی[۳]

[۱] دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه یزد، hananeh.rajabiun@stu.yazd.ac.ir

[۲] دانشیار دانشکده مهندسی کامپیوتر، دانشگاه یزد، m.ghasemzadeh@yazd.ac.ir

[۲] استادیار دانشکده مهندسی کامپیوتر، دانشگاه یزد، vranjbar@yazd.ac.ir

چکیده - یکی از مهمترین اصلاحات پس از ترجمه (*PTMs*) در پروتئین‌ها، مالونیلاسیون لیزین است که تأثیر بر عملکرد سلول‌ها می‌گذارد. برای آشکارسازی مکانیسم‌های عملکردهای سلولی لازم است محل مالونیلاسیون در پروتئین‌ها را شناسایی کرد. روش‌هایی مبتنی بر راه‌حل‌های یادگیری ماشینی پیشنهاد شده‌اند که باعث کاهش هزینه‌ها و پیچیدگی‌های زمانی و افزایش دقت می‌شود. با این حال، این رویکردها همچنان با کاستی‌هایی همراه هستند. این پژوهش نشان می‌دهد که در رابطه با یافتن جایگاه مالونیلاسیون در پروتئین‌ها، چگونه می‌توان با بهره‌گیری از رویکرد پردازش زبان طبیعی، فرکانس واژه - ضریب ارتباط دسته، به نتایجی مطلوب دست‌یافت. روش پیشنهادی توسط توابع تخصصی مربوطه و در محیط پایتون، پیاده‌سازی و اجرا گردید. نتایج اعتبار سنجی متقابل نشان‌دهنده عملکرد خوب رویکرد پیشنهادی است. علاوه بر این، طبقه‌بندی‌کننده‌ی *XGBOOST* و جنگل تصادفی بر دیگر طبقه‌بندی کننده‌ها برتر بود، که نشان‌دهنده اثربخشی ویژگی‌های تولیدشده توسط طرح پیشنهادی است.

کلید واژه- مالونیلاسیون، یادگیری ماشین، پردازش زبان طبیعی، استخراج ویژگی.

۱- مقدمه

توالی‌های پروتئین را می‌توان مانند زبان طبیعی، به‌صورت رشته‌ای از حروف نشان داد. علاوه بر این، مانند زبان طبیعی، پروتئین‌های تکامل‌یافته طبیعی معمولاً از عناصر مدولار استفاده مجدد تشکیل شده‌اند که تغییرات جزئی را نشان می‌دهند که می‌توانند به شیوه‌ای سلسله مراتبی بازآرایی و مونتاژ شوند. بنابراین، بلوک‌های سازنده عملکردی اساسی پروتئین‌ها که نقش‌ها و حوزه‌های پروتئینی هستند، شبیه کلمات، عبارات و جملات در زبان طبیعی هستند. کامل بودن اطلاعات یکی از ویژگی‌های اصلی مشترک پروتئین‌ها و زبان طبیعی است. عملکرد و ساختار پروتئین، پویا و وابسته به زمینه است (به‌عنوان مثال به وضعیت سلولی، مولکول‌های دیگر و PTMs)، ولیکن توسط توالی اسیدآمینه تعریف می‌شود. از دیدگاه تئوری اطلاعات، به این معناست که اطلاعات پروتئین (به‌عنوان مثال ساختار آن) در توالی آن وجود دارد [۱]. یکی از مکانیزم‌های اساسی برای تنظیم بسیاری از فرآیندهای بیولوژیکی، PTM می‌باشد. پروتئین‌ها همواره باهم در حال تعامل‌اند، انجام یک مطالعه دقیق بر روی PTMهای پروتئین می‌تواند منجر به آشکار کردن و شناسایی بیماری‌هایی ازجمله سرطان، دیابت، بیماری‌های خود ایمنی و مکانیزم‌های فعالیت‌های روزمره زندگی شود [۲]. یکی از PTM تازه کشف‌شده مالونیلاسیون است که نقش مهمی را در ایفای عملکردهای مختلف سلول، فرآیندهای بیولوژیکی و تنظیم پویایی سلول دارند. در آن اسیدهای آمینه لیزین با بار مثبت در یک پروتئین با افزودن یک گروه مالونیلاسیون با بار منفی اصلاح شیمیایی می‌شوند. لایه‌های مالونیلاسیون لیزین با توان بالا تأثیرات فوق‌العاده‌ای را بر روی سلول‌های یوکاریوتی و پروکاریوتی دارا هستند [۳،۴].

شناسایی دقیق جایگاه مالونیلاسیون بسیار مهم است و می‌تواند خبرهای مفیدی برای تحقیقات زیست پزشکی و درک بهتر عملکردهای مولکولی فراهم کند. در حال حاضر، به دلیل

محدودیت‌های زمانی و هزینه‌ای روش‌های تجربی، انجام این آزمایش‌ها دشوار است. از این رو، برای شناسایی درست جایگاه مالونیلاسیون لازم است از روش‌های محاسباتی استفاده شود. اخیراً در برخی از کارهای منتشرشده، از روش‌های یادگیری ماشین و یادگیری عمیق برای پیش‌بینی جایگاه مالونیلاسیون استفاده‌شده است. در روش (Mal-Lys) برای پیش‌بینی جایگاه (K-mal)، استخراج ویژگی‌ها بر اساس توالی تک و خواص فیزیکی و شیمیایی بر روی دیتاست (M.musculus) انجام شده است و در این روش با استفاده از حداکثر ارتباط افزونگی و روش ماشین بردار پشتیبان وجود مالونیلاسیون پیش‌بینی می‌شود [۵]. در رویکرد DeepMal، یک مدل شبکه یادگیری عمیق جدید را پیشنهاد می‌کند که ویژگی‌ها توسط ترکیب اسیدآمینه تقویت‌شده (EAAC)، ترکیب اسیدآمینه گروه‌بندی‌شده (EGAAC)، انحراف دی پپتیدی از میانگین مورد انتظار (DDE)، K نزدیک‌ترین همسایگان (KNN) و ماتریس BLOSUM62 استخراج می‌شوند و شبکه عصبی کانولوشن خطی برای استخراج ویژگی‌های خاص جایگاه مالونیلاسیون، انتخاب ویژگی‌های مربوطه و کاهش ابعاد ویژگی از طریق حداکثر ادغام استفاده می‌شود [۶]. وانگ و همکاران، که به‌طور جداگانه برای سه دیتاست M.musculus، H.sapiens، E.coli بر اساس ترکیب استخراج ویژگی‌ها شامل توالی تک، پروفایل‌های تکاملی توالی‌ها و خاصیت‌های آمینواسیدها به بررسی جایگاه مالونیلاسیون می‌پردازد [۷]. طاهرزاده و همکاران، استخراج ویژگی را با استفاده از ویژگی‌های توالی و ویژگی‌های ساختاری انجام دادند. این رویکرد اولین ابزار پیش‌بینی جایگاه مالونیلاسیون آنلاین است که خصوصیات ساختاری پروتئین‌ها را نیز در نظر گرفته می‌شود و از ماشین بردار پشتیبان جهت پیش‌بینی جایگاه مالونیلاسیون استفاده می‌شود [۸]. رجبیون و همکاران، یک روش انتخاب ویژگی ترکیبی برای پیش‌بینی مکان‌های مالونیلاسیون لیزین بر روی سه مجموعه داده مالونیلاسیون لیزین .M musculus، H. sapiens و E. coli پیشنهاد کردند [۹]. ژانگ و همکاران، با استفاده از یازده روش استخراج ویژگی، اطلاعات ویژگی‌های پروتئین را استخراج و از GainRatio برای تجزیه و تحلیل آنها استفاده می‌کنند. درنهایت از با استفاده از مدل‌های یادگیری RF، SVM، KNN، LR، Light Gradient Boosting Machine برای دسته‌بندی جایگاه مالونیلاسیون استفاده می‌شود [۱۰]. در سال ۲۰۱۸، چن و همکاران، یک طبقه‌بندی شبکه یادگیری عمیق (DL) بر اساس حافظه کوتاه‌مدت بلندمدت (LSTM) با جاسازی کلمه (LSTMWE) برای پیش‌بینی مکان‌های مالونیلاسیون پستانداران ارائه می‌دهد [۱۱]. پروتئین‌ها، که می‌توانند به‌عنوان رشته‌هایی از حروف اسیدآمینه نشان داده شوند، برای بسیاری از روش‌های پردازش زبان طبیعی مناسب هستند و می‌توان با روش‌های پردازش زبان طبیعی ویژگی‌های پروتئین‌ها را استخراج کرد. همچنین، برای درک بهتر مکانیسم مالونیلاسیون در پروتئین‌ها لازم است که از قبل به‌طور دقیق جایگاه مالونیلاسیون را مشخص کنیم. لذا می‌توان برای شناسایی مؤثرتر و دقیق جایگاه مالونیلاسیون از روش‌های محاسباتی استفاده کرد. در حال حاضر، رویکردهای محاسباتی موجود بیشتر به مهندسی ویژگی متکی هستند. ابزار موجود در حال حاضر فقط جایگاه مالونیلاسیون را در انسان ، موش و باکتری در نظر گرفته‌اند.

در این پژوهش یک رویکرد پردازش زبان طبیعی برای شناسایی جایگاه مالونیلاسیون راه‌اندازی کردیم. یک طبقه‌بندی کننده، ویژگی‌های جدید به‌دست‌آمده از اطلاعات دامنه عملکردی پروتئین‌ها را با استفاده از یک رویکرد پردازش زبان طبیعی، فرکانس واژه - ضریب ارتباط دسته بکار گرفت. برای این منظور، در این مقاله روشی چهار مرحله‌ای ارائه شده است که ابتدا استخراج ویژگی از پروتئین‌ها صورت می‌پذیرد. سپس پیش‌پردازش انجام می‌شوند. درنهایت در آخرین مرحله، انتخاب ویژگی از ویژگی‌های استخراج شد و عمل دسته‌بندی انجام می‌شود.

۲- تعاریف و مقدمات موردنیاز

در این بخش به بررسی روش استخراج ویژگی از توالی پروتئین، مجموعه داده‌های مورد استفاده و رویکردهای ارزیابی مدل می‌پردازیم.

۲-۱- استخراج ویژگی

استخراج ویژگی یک گام مهم در طراحی طبقه‌بندی کننده‌های کارآمد است. در این پژوهش باید از هر پروتئین، ویژگی‌هایی استخراج‌کنیم که می‌تواند خواص اساسی پروتئین‌ها را حفظ کند. دامنه عملکردی به‌طور گسترده، برای بررسی مشکلات مختلف مرتبط با پروتئین، ازجمله پیش‌بینی جایگاه مالونیلاسیون استفاده می‌شود. الگوریتم‌های استخراج ویژگی، اطلاعات کاراکترهای یک توالی را به یک بردار عددی تبدیل می‌کند. در این پژوهش، به‌منظور استخراج ویژگی از توالی پروتئین، از الگوریتم، فرکانس واژه - ضریب ارتباط دسته[۱۲] استفاده شده است.

۲-۲- مجموعه داده‌ها

در این پژوهش، مجموعه داده از پایگاه دادهUniProt [۱۳] و CD-HIT [۱۴] تهیه شده است و جهت جلوگیری از خطاها مدل آموزش و کاهش شباهت و همسانی توالی‌ها و حفظ توالی‌های پروتئینی با هویت توالی <۳۰٪ استفاده می‌شود. این مجموعه داده شامل ۱۷۴۶ مالونیلاسیون از ۵۹۵ پروتئین E. coli، ۳۴۳۵ مالونیلاسیون از ۱۱۷۴ پروتئین در M. musculus و ۴۵۷۹ مالونیلاسیون از ۱۶۶۰ پروتئین

در H. sapiens است. طول هر توالی پروتئین به ۲۵ آمینه اسید کاهش داده شده است که لیزین (K) واقع در مرکز توالی است. اگر K مرکز آن مالونیلاسیون باشد به عنوان یک نمونه مثبت تعریف می‌شود ، در غیر این صورت به عنوان یک نمونه منفی تعریف می‌شود.

۲-۳- ارزیابی مدل

در این مقاله از روش اعتبار سنجی متقابل با مقدار ۱۰ (۱۰-fold cross-validation) برای تعیین پارامترهای مدل بر اساس مجموعه داده‌های آموزشی و از مجموعه داده مستقل برای ارزیابی عملکرد مدل استفاده می‌شود . جهت ارزیابی مدل از حساسیت (Sn) رابطه (۱)، ویژگی (Sp) رابطه (۲)، دقت (Pre) رابطه (۳)، صحت (ACC) رابطه (۴) و ضریب همبستگی (MCC) رابطه (۵) استفاده‌شده است. این شاخص‌ها به شرح زیر تعریف می‌شوند:

$$Sn = \frac{TP}{TP+FN} \quad (۱)$$

$$Sp = \frac{TN}{TN+FP} \quad (۲)$$

$$Pre = \frac{TP}{TP+FP} \quad (۳)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (۴)$$

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (۵)$$

که TP به‌درستی مثبت می‌باشد یعنی تعداد برچسب‌های مثبتی که به‌درستی توسط طبقه‌بندی کننده مثبت پیش‌بینی‌شده است، TN به‌درستی منفی هست، تعداد برچسب‌های منفی که توسط طبقه‌بندی کننده به‌درستی منفی پیش‌بینی‌شده است، FP نادرست مثبت، تعداد برچسب‌های منفی که توسط طبقه‌بندی کننده به اشتباه مثبت پیش‌بینی‌شده است و FN نشان‌دهنده نادرست منفی است و به تعداد برچسب‌های مثبت که توسط طبقه‌بندی کننده به اشتباه منفی پیش‌بینی‌شده است اشاره می‌کند.

۳- روش پیشنهادی

در این مقاله یک مدل به منظور پیش‌بینی جایگاه مالونیلاسیون پیشنهادشده است. روش پیشنهادی از پنج مرحله اصلی شامل انتخاب مجموعه داده‌ها، استخراج ویژگی، نرمال‌سازی ویژگی‌ها، انتخاب ویژگی و در نهایت ارزیابی عملکرد مدل تشکیل‌شده است:

مرحله اول) انتخاب مجموعه داده‌ها: از سه مجموعه داده‌ها، یعنی (E. coli) و (M. musculus) و (H. sapiens) برای آموزش و آزمایش روش پیشنهادی استفاده‌شده است. مجموعه داده به‌طور تصادفی به مجموعه‌های آموزش و مجموعه داده آزمایش تقسیم می‌شود. برای تجزیه و تحلیل کارآمد، یک استراتژی اعتبار دهی متقابل انجام‌شده است.

مرحله دوم) استخراج ویژگی: در این مرحله به‌منظور استخراج اطلاعات دامنه عملکردی پروتئین‌ها را از طریق یک رویکرد پردازش زبان طبیعی استفاده‌شده است. روش‌های استخراج ویژگی استفاده‌شده در این مقاله TF-CRF یا طبیعی فرکانس واژه - ضریب ارتباط دسته می‌باشد. استخراج ویژگی با استفاده از TF-CRF که منجر به وزن دهی دقیق‌تر در دو فاکتور positiveRF (فاکتور ارتباط مثبت) و negativeRF (فاکتور ارتباط منفی) می‌شود. طول بردار ویژگی حاصل ۲۰ است.

مرحله سوم) نرمال‌سازی ویژگی‌ها: با استخراج ویژگی‌ها از توالی‌های پروتئینی، آنها در محدوده‌های مختلفی خواهند بود. تفاوت در مقادیر ویژگی تأثیر برخی از ویژگی‌های مهم را به شدت کاهش می‌دهد. همچنین، ویژگی‌هایی با دامنه وسیع نوسانات، کارایی مدل‌های یادگیری را بدتر می‌کند. بر این اساس، داده‌ها باید نرمال‌سازی شوند تا کارایی بهبود یابد. در کار حاضر از نرمال‌سازی Z-Score برای این منظور استفاده‌شده است. در واقع، Z-score یک استراتژی عادی‌سازی است که از داده‌ها و ویژگی‌های پرت جلوگیری می‌کند. Z-Score یک استراتژی نرمال‌سازی داده‌ها می‌باشد که از داده‌های پرت (outlier) جلوگیری می‌کند. رابطه (۶) نرمال‌سازی Z-Score به شرح زیر است:

$$z = \frac{x-\mu}{\sigma} \quad (۶)$$

در این رابطه ، μ مقدار متوسط (میانگین) ویژگی و σ انحراف استاندارد ویژگی می‌باشد. اگر مقدار یک ویژگی دقیقاً برابر با میانگین تمام مقادیر ویژگی باشد، به صفر نرمال‌سازی می‌شود. اگر این مقدار زیر میانگین باشد، عدد منفی خواهد بود و اگر بالاتر از میانگین باشد، عدد مثبت خواهد بود. اندازه این اعداد منفی و مثبت بر اساس انحراف معیار ویژگی اصلی تعیین می‌شود. اگر ویژگی‌ها انحراف معیار بزرگی داشته باشند، مقادیر نرمال‌سازی به صفر نزدیک‌تر می‌شوند.

مرحله چهارم) انتخاب ویژگی: ویژگی‌های استخراج‌شده برای پیش‌بینی محل مالونیلاسیون مفید است. بااین‌حال، همه ویژگی‌ها ممکن است کارآمد نباشند. برخی از آنها ممکن است نامربوط و برخی دیگر اضافی باشند. چنین ویژگی‌هایی منجر به تطبیق بیش‌ازحد (overfitting) مدل، ایجاد افزونگی و نویز که باعث تضعیف عملکرد مدل می‌شود. بنابراین، حفظ ویژگی‌های مربوطه ضروری است. برای شناسایی ویژگی‌های مفید (زیرمجموعه) از مجموعه ویژگی‌های اولیه،

از روش F-Score که یک روش انتخاب ویژگی است، استفاده می‌شود [۱۵]. معیار F-score برای ویژگی i ام به‌صورت رابطه (۷) زیر محاسبه می‌شود:

(۷)

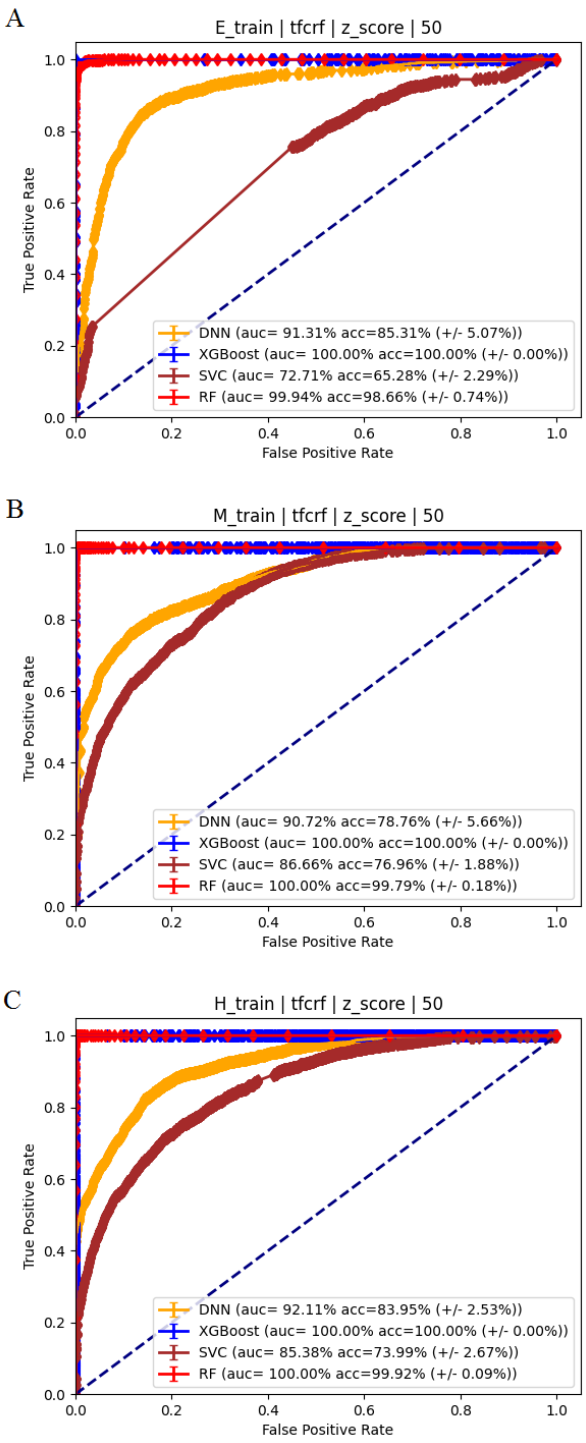
$$F-Score(i)=\frac{\sum_{k=1}^{m}\left(\bar{x}_i^k-\bar{x}_i\right)^2}{\sum_{k=1}^{m}\frac{1}{n^k-1}\sum_{j=1}^{n^k}\left(x_{j,i}^k-\bar{x}_i^k\right)^2}$$

در این رابطه m تعداد کلاس‌ها، n^k تعداد نمونه‌های کلاس k ام، $\bar{x}_i$ میانگین ویژگی i ام در کل داده‌ها، $\bar{x}_i^k$ میانگین ویژگی i ام در کلاس k، $x_{j,i}^k$ مقدار ویژگی i ام در j امین نمونه از کلاس k می‌باشد. اگر مقدار F-score یک ویژگی زیاد باشد، نشان می‌دهد که آن ویژگی دارای اطلاعات مناسبی جهت دسته‌بندی می‌باشد.

مرحله پنجم) ارزیابی عملکرد مدل: از روش اعتبار سنجی متقابل با مقدار ۱۰ برای ارزیابی عملکرد پیش‌بینی مدل طبقه‌بندی استفاده شد .طبقه بندهای مختلف مورداستفاده در این پژوهش شامل (XGBoost)، (SVM)، (RF)، (DNN) است که در آنها مقادیر Pre، ACC، Sn، Sp و MCC محاسبه شدند. همچنین برای ارزیابی دقیق‌تر به‌رسم منحنی ROC و نمودار میله خطا پرداخته شد. از مدل‌های ساخته‌شده برای آزمایش مدل با استفاده از یک مجموعه آزمون استفاده می‌شود.

۴- نتایج آزمایشی و تحلیل

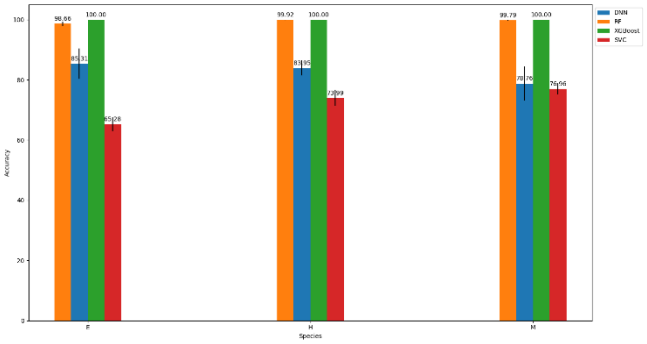
در این مقاله از روش‌های طبقه‌بندی ازجمله شامل افزایش شدید گرادیان (XGBoost)، ماشین بردار پشتیبان (SVM)، جنگل تصادفی (RF)، شبکه‌های عصبی عمیق (DNN) بر اساس معیارهای ارزیابی مختلف شامل حساسیت (Sn)، ویژگی (Sp)، دقت (Pre)، صحت (ACC) و ضریب همبستگی (MCC Matthew) با هم مقایسه شده‌اند. الگوریتم (XGboost) و (RF) در برای سه مجموعه داده، عملکرد بهتری داشته و صحت بالاتری نسبت به سایر الگوریتم‌های بررسی‌شده دارد. به‌منظور تحلیل و آنالیز بهتر در شکل ۱، منحنی ROC [۱۶] بر روی دسته‌بندهای مختلف نشان داده‌شده است. منحنی ROC بهترین ویژگی در سه مجموعه داده آموزشی (H. sapiens) و (E. coli) و (M. musculus) نشان داده‌شده است. همان‌طور که در منحنی‌های ROC مربوط به طبقه‌بندی کننده‌های SVM، XGBoost، RF و DNN مشخص است، مساحت زیر منحنی ROC در طبقه‌بندی XGBoost به‌طور قابل‌توجهی بالاتر از سه الگوریتم طبقه‌بندی دیگر قرارگرفته است که نشان می‌دهد الگوریتم طبقه‌بندی (XGBoost) دارای توانایی تعمیم قوی و عملکرد پیش‌بینی خوبی برای جایگاه مالونیلاسیون و غیر مالونیلاسیون پروتئین است.



شکل ۱: منحنی ROC برای روش پیشنهادی در الگوریتم‌های مختلف طبقه بندی. نمودارهای (A)، (B) و (C) به ترتیب به مجموعه داده‌های M. musculus ,E. coli وH. sapiens مربوط می‌شوند.

تجزیه و تحلیل خطا به‌منظور نشان دادن میزان مقاومت و پایداری مدل است. میله‌ی خطا، با نشان دادن خطای برآورد شده یا عدم قطعیت، درک عمیق‌تری از دقیق بودن اندازه‌گیری‌ها منتقل می‌کند. این کار با استفاده از نشانگرهای رسم شده بر روی نمودار اصلی که به‌طور معمول، برای نمایش انحراف استاندارد، خطای استاندارد، فواصل اطمینان یا حداقل و حداکثر مقادیر در یک مجموعه

داده محدود استفاده می‌شود. در شکل ۲، مقایسه‌ای بین الگوریتم‌های (DNN)، (RF)، (XGboost) و (SVC) انجام شد. دقت الگوریتم‌ها از طریق یک استراتژی اعتبار سنجی متقابل با مقدار ۱۰ در برای سه مجموعه داده‌های نشان داده‌شده است. همان‌طور که در شکل مشخص است الگوریتم (XGboost) و (RF) نسبت به سایر الگوریتم‌ها عملکرد بهتری داشته است و الگوریتم (DNN) میزان خطای بیشتری را با توجه به میله خطا نشان می‌دهد. هرچه مقدار میله خطا کمتر باشد دقت بالاتر الگوریتم و واریانس کمتر دقت مدل را نشان می‌دهد درنتیجه با توجه به شکل ۶ می‌توان گفت نتایج اعتبار سنجی متقابل با مقدار ۱۰ در الگوریتم (XGboost) و (RF) نزدیک به هم بوده است و باعث شده میزان خطا تقریباً برابر با صفر شود بنابراین این مدل عملکرد تعمیم بالایی دارد.بااین‌حال، در الگوریتم (DNN) عکس این موضوع اتفاق افتاده است و به این موضوع اشاره می‌کند که نتایج اعتبار سنجی متقابل با مقدار ۱۰ نزدیک هم نیست که منجر به واریانس بالاتر در دقت، و درنتیجه، عملکرد تعمیم پایین‌تر می‌شود.



شکل ۲: بررسی مدل های دسته بند با استفاده از میله ی خطا برای E. coli، M.musculus و H. sapiens.

۵- نتیجه‌گیری و پیشنهادها

در این مقاله، یک روش مبتنی بر یادگیری ماشین و پردازش زبان طبیعی برای شناسایی جایگاه مالونیلاسیون راه‌اندازی کردیم. از یک رویکرد پردازش زبان طبیعی، (TF-CRF)، برای استخراج اطلاعات دامنه عملکردی پروتئین‌ها استفاده شد. همچنین برای جلوگیری از تطبیق بیش‌ازحد (overfitting) مدل، کارآمدترین ویژگی‌ها از روش انتخاب ویژگی انتخاب‌شده‌اند. نتایج اعتبارسنجی متقاطع نشان داد که XGboost و RF بر اساس ویژگی‌های استخراج‌شده و انتخاب‌شده، بهتر از دیگر طبقه‌بندی عمل می‌کند. نتایج همچنین نشان‌دهنده برتری ویژگی‌های استخراج‌شده توسط طرح پیشنهادی است.

مراجع

[1] Ofer D, Brandes N, Linial M, "The language of proteins: NLP, machine learning & protein sequences," Comput Struct Biotechnol J, vol. 19, pp. 1750-1758, 2021.

[2] Wang, Minghui, Cui, Xiaowen, Yu, Bin, Chen, Cheng, Ma, Qin, Zhou, Hongyan, "SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting," *Neural Computing and Applications*, vol. 32, no. 17, p. 13843–13862, 2020.

[3] Xie Z, Dai J, Dai L, Tan M, Cheng Z, Wu Y, Boeke JD, Zhao Y, "Lysine succinylation and lysine malonylation in histones," Mol Cell Proteomics, vol. 11, pp. 100-7, 2012.

[4] Olsen CA, "Expansion of the lysine acylation landscape," *Angew Chem Int Ed Engl*, vol. 51, no. 16, pp. 3755-6, 2012.

[5] Xu Y, Ding YX, Ding J, Wu LY, Xue Y, "Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection," *Sci Rep*, vol. 6, 2016.

[6] Wang, M., Xiaowen Cui, Shan Li, Xin-hua Yang, Anjun Ma, Yusen Zhang and Bin Yu, "DeepMal: Accurate prediction of protein malonylation sites by deep neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 207, 2020.

[7] Wang LN, Shi SP, Xu HD, Wen PP, Qiu JD, "Computational prediction of species-specific malonylation sites via enhanced characteristic strategy," *Bioinformatics*, vol. 33, pp. 1457-1463, 2017.

[8] Ghazaleh Taherzadeh, Yuedong Yang,Haodong Xu,Yu Xue,Alan Wee-Chung Liew,Yaoqi Zhou, "Predicting lysine-malonylation sites of proteins using sequence and predicted structural features," *Journal of Computational Chemistry*, vol. 39, no. 22, pp. 1757-1763, 2018.

[9] Hananeh Rajabiun, Mahdis MohammadHoseini, Hadi Zarezadeh, Mehdi Delkhosh,, "A hybrid feature selection method for predicting lysine malonylation sites in proteins via machine learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 222, no. 0169-7439, 2022.

[10] Zhang, Y., Xie, R., Wang, J., Leier, A., Marquez-Lago, T. T., Akutsu, T., Webb, G. I., Chou, K. C., & Song, J, "Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework," *Briefings in bioinformatics*, vol. 20, pp. 2185-2199, 2019.

[11] Chen Z, He N, Huang Y, Qin WT, Liu X, Li L, "Integration of A Deep Learning Classifier with A Random Forest Approach for Predicting Malonylation Sites," *Genomics Proteomics Bioinformatics*, vol. 16, pp. 451-459, 2018.

[12] M. Maleki, A. Abdollahzadeh, "TFCRF: A Novel Feature Weighting Method Based on Class Information in Text Categorization," in *International Conference on Computer, Information and Systems Science and Engineering*, Bangkok, 2007.

[13] U. Consortium, "Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Res*, no. 39(Database issue), 2011.

[14] Huang Y, Niu B, Gao Y, Fu L, Li W, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, pp. 680-682, 2010.

[15] M.M. Gromiha, M. Suwa, "A simple statistical method for discriminating outer membrane proteins with better accuracy," *Bioinformatics* , vol. 21, pp. 961-968, 2005

[16] Wang X, Yu B, Ma A, Chen C, Liu B, Ma Q, "Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique.," Bioinformatics, vol. 39, pp. 239