

طراحی یک مدل یادگیری عمیق برای پیش‌بینی جایگاه‌های فسفریلاسیون

زینب ظهیری^۱، ناصر مهرشاد^{۲*}

۱- دانشجوی دکتری، گروه الکترونیک، دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران

۲- استاد، گروه الکترونیک، دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران

* صندوق پستی ۹۷۱۷۵/۶۱۵، بیرجند، ایران

nmehrshad@birjand.ac.ir

پذیرش: ۱۴۰۴/۰۵/۱۱

دریافت: ۱۴۰۳/۰۳/۲۱

چکیده

فسفریلاسیون مهم‌ترین نوع از تغییرات پس از ترجمه (Post-Translational Modification) است که نقش مهمی در مطالعات عملکرد پروتئین و طراحی تجربی دارد. با توجه به اهمیت فسفریلاسیون در پروتئین‌ها و افزایش روزافزون شمار توالی‌های پروتئین‌ها در پایگاه‌های داده، نیاز به بهبود روش‌های محاسباتی پیش‌بینی جایگاه‌های فسفریلاسیون، از نظر سرعت و دقت روز به روز با اهمیت‌تر می‌شود. با اینکه تاکنون ابزارهای پیش‌بینی کننده بسیار زیادی برای پیش‌بینی مکان‌های فسفریلاسیون با استفاده از روش‌های مختلف یادگیری ماشین معرفی گردیده است اما هنوز هم تا ابزار بسیار کارآمد فاصله زیادی است و تلاش‌ها برای دستیابی به چنین ابزاری ادامه دارد. اخیراً، چندین مطالعه ادعا کرده‌اند که روش‌های مبتنی بر یادگیری عمیق بهترین راه برای پیش‌بینی مکان‌های فسفریلاسیون هستند. زیرا یادگیری عمیق به عنوان یک روش یادگیری ماشینی پیشرفته می‌تواند به‌طور خودکار نمایش‌های پیچیده الگوهای فسفریلاسیون را از توالی‌های خام تشخیص دهد و بنابراین ابزاری قدرتمند برای بهبود پیش‌بینی جایگاه فسفریلاسیون ارائه می‌دهد. در این مطالعه، یک ساختار ترکیبی بر اساس روش یادگیری عمیق کانولوشنی (Convolutional Deep Learning) با نام ConvoPhos برای پیش‌بینی مکان‌های فسفریلاسیون معرفی شده است. به گونه‌ای که، بردار ویژگی CkSAAPair به‌دست آمده از روی توالی‌ها، به‌عنوان ورودی بخشی از طبقه‌بندی و تبدیل توالی‌ها به عکس، به‌عنوان ورودی بخش دیگری از شبکه‌های کانولوشنی استفاده شده است.

کپی‌رایت © ۲۰۲۵، انتشارات دانشگاه تربیت مدرس (TMU Press). این مقاله به‌صورت دسترسی آزاد منتشر شده و تحت مجوز بین‌المللی Creative Commons Attribution-NonCommercial 4.0 قرار دارد. بر اساس این مجوز، شما می‌توانید این مطلب را در هر قالب و رسانه‌ای کپی، بازنشر و بازآفرینی کنید و یا آن را ویرایش و بازسازی نمایید، به شرط آنکه نام نویسنده را ذکر کرده و از آن برای مقاصد غیرتجاری استفاده کنید.

نتایج حاصل از اعتبارسنجی متقابل ۱۰ باره، مقدار صحت ۹۴ درصدی را برای داده‌های phosphosite و AUC 90 درصدی را نشان می‌دهد، که در بین سایر روش‌های مقایسه شده بالاترین کارایی را دارد.

کلید واژگان: فسفریلاسیون پروتئین، استخراج ویژگی، پیش‌بینی کننده، شبکه یادگیری عمیق کانولوشنی.

۱- مقدمه

فسفریلاسیون پروتئین یکی از رایج‌ترین و مهم‌ترین تغییرات پس از ترجمه (PTMs) است [۱]. این مکانیسم برگشت‌پذیر از طریق پروتئین‌کینازها اتفاق می‌افتد و شامل افزودن یک گروه فسفات (PO_4) به گروه قطبی R از اسیدهای آمینه مختلف است. در نتیجه، این افزودن پروتئین را از آپولار آبگریز به قطبی آب‌دوست تغییر می‌دهد و به پروتئین اجازه می‌دهد در هنگام تعامل با مولکول‌های دیگر، ترکیب را تغییر دهد. یک آمینو اسید فسفریله می‌تواند مولکول‌هایی را که قادر به تعامل با پروتئین‌های دیگر و در نتیجه جمع‌آوری و جداسازی کمپلکس‌های پروتئینی هستند، متصل کند [۲]. فسفریلاسیون پروتئین مکانیزم تنظیمی است که در اکثر فرایندهای سلولی مانند سنتز پروتئین، تقسیم سلولی، انتقال سیگنال، رشد سلولی، رشد و پیری بسیار مهم است، زیرا بسیاری از آنزیم‌ها و گیرنده‌ها از طریق رویدادهای فسفریلاسیون و معکوس فسفریلاسیون به دلیل کینازهای خاص، فعال و غیرفعال می‌شوند [۳].

در سال‌های اخیر افزایش چند برابری توالی‌های پروتئین‌ها در بانک‌های اطلاعاتی نسبت به سال‌های قبل و هزینه بالای بررسی تجربی جایگاه‌های فسفریلاسیون، توجه محققان بسیاری را به استفاده از روش‌های محاسباتی برای کاهش جایگاه‌هایی که باید به صورت آزمایشگاهی تایید شوند، جلب کرده است. در روش‌های محاسباتی محققین با استفاده از اطلاعات حاصل از آنالیز داده‌های تجربی، ویژگی‌های مناسبی از پروتئین‌ها را برای آموزش روش‌های یادگیری ماشین برای پیش‌بینی، استفاده می‌کنند.

برخی سیستم‌های پیش‌بینی محل فسفریلاسیون موجود از اطلاعات خاص کیناز جایگاه‌های فسفریله استفاده می‌کنند. در این شرایط، لازم است که پروتئین‌های بدون حاشیه‌نویسی کیناز از مجموعه داده‌های فسفریلاسیون مثبت که تا به امروز پیدا شده است، خارج شود. بررسی‌ها نشان می‌دهند که تنها ۳۰ درصد از جایگاه‌های فسفریلاسیون واقعی حاوی اطلاعات کیناز هستند، به عبارت دیگر بیش از ۷۰ درصد از مجموعه داده‌ها در طراحی سیستم پیش‌بینی کننده‌های مختص کیناز موجود حذف می‌شوند [۴]، این برش قابل توجه برخی از ویژگی‌های اساسی مکان‌های فسفریلاسیون را نادیده می‌گیرد.

همچنین، با پیشرفت‌های اخیر در فناوری توالی‌یابی، ژنوم‌های بسیاری از ارگانیسم‌های غیرمدل توالی‌یابی شده‌اند، و کینازهای بیشتری در آن ژنوم‌های بازسازی شده کشف شده‌اند که برخی از آنها اطلاعات بستر کافی برای آموزش الگوریتم‌های پیش‌بینی کیناز خاص ندارند. بنابراین، علاقه فزاینده‌ای به توسعه ابزارهای غیر اختصاصی کیناز برای انواع وسیع‌تری از گونه‌ها و دقت بالا برای حاشیه نویسی کل ژنوم وجود دارد [۵ و ۶].

تاکنون چندین مدل کلی پیش‌بینی محل فسفریلاسیون پیشنهاد شده است. اکثر روش‌های موجود در انتخاب الگوریتم‌های یادگیری ماشین و مهندسی استخراج ویژگی‌ها متفاوت هستند. پرکاربردترین روش‌های یادگیری ماشین در ابزارهای پیش‌بینی عمومی شامل، شبکه‌های عصبی مصنوعی (ANN)^۱، ماشین‌های بردار پشتیبان (SVM)^۲،

^۱ Support Vector Machine

^۲ Artificial Neural Network

رگرسیون خطی (LR)^۱ و جنگل تصادفی (RF)^۲ است. به عنوان مثال، NetPhos از شبکه‌های عصبی استفاده می‌کند [۷] و بیسواس و همکاران در PPRED اطلاعات تکاملی پروتئین‌ها را با SVM ها ترکیب می‌کنند [۸] تا مکان‌های فسفریلاسیون را پیش‌بینی کنند. در حالی که، DISPPOS از فرکانس اسید آمینه و اطلاعات اختلال برای آموزش یک مدل LR [۹] و Musite، سه مجموعه از پارامترها، از جمله K امتیاز نزدیکترین همسایه، امتیازدهنده اختلال پروتئین و فرکانس اسیدهای آمینه را به عنوان ویژگی‌هایی برای آموزش SVM استفاده کرده‌اند [۱۰]. Phospho-SVM که یکی از جدیدترین ابزارهای پیش‌بینی مبتنی بر SVM است، هشت تابع مختلف امتیاز دهی در سطح توالی را ادغام کرده و از SVM به منظور پیش‌بینی استفاده می‌کند [۵].

با توجه به بینشی که الگوریتم‌های RF می‌توانند در مورد اهمیت نسبی هر ویژگی ارائه دهند، طبقه‌بندی‌کننده‌های RF برای مسائل مختلف بیوانفورماتیک مورد استقبال قرار گرفته‌اند [۱۱]، به عنوان مثال، در RF-Phos ویژگی‌های توالی و ساختاری به عنوان ورودی جنگل تصادفی برای پیش‌بینی مکان عمومی فسفریلاسیون استفاده شده است [۱۲].

یکی از پیشرفت‌های بسیار مهم و موثر در یادگیری ماشین، معرفی شبکه‌های عصبی مصنوعی عمیق در سال‌های اخیر است. این روش پیشرفت‌های زیادی را در تشخیص تصویر و گفتار، پردازش زبان طبیعی و اخیراً زیست‌شناسی محاسباتی ایجاد کرده است [۱۳]. در مقایسه با تکنیک‌های یادگیری ماشین سنتی، یک شبکه عصبی عمیق داده‌های خام را در پایین‌ترین لایه (ورودی) دریافت، به طور خودکار نمایش‌های پیچیده را کشف می‌کند و انتزاع سطح بالا را به طور تطبیقی از داده‌های آموزشی برای طبقه‌بندی می‌گیرد. به دلیل نتایج مناسب روش‌های مبتنی بر یادگیری عمیق،

کاربرد آن‌ها برای تجزیه و تحلیل توالی بیولوژیکی در حال رشد است. به عنوان مثال، DeeperBind از یک شبکه عصبی کانولوشن (CNN) برای پیش‌بینی ویژگی‌های توالی پروتئین‌های متصل شونده به DNA و RNA استفاده می‌کند [۱۴]. MusiteDeep یک CNN با مکانیسم توجه دو بعدی را برای پیش‌بینی جایگاه‌ها انتخاب می‌کند [۱۵]، همچنین DeepNitro از شبکه عصبی عمیق لایه‌ای برای پیش‌بینی جایگاه‌های نیتراسیون و نیتروزیلاسیون استفاده می‌کند [۱۶]. DeepPhos، عملکرد MusiteDeep را با استفاده از معماری چند لایه CNN [۱۷] بهبود داده است، DeepPSP هر دو ویژگی محلی و سراسری را از توالی‌های پروتئین با دو ماژول موازی استخراج کرده و به عنوان ورودی یک سیستم یادگیری عمیق استفاده کرده است [۱۸].

در این مطالعه، یک پیش‌بینی‌کننده جدید بر اساس یادگیری عمیق برای پیش‌بینی مکان‌های فسفریلاسیون با نام ConvoPhos، گزارش شده است. در ConvoPhos، ابتدا توالی‌های ورودی با روش‌های CkSAAPair تبدیل توالی‌ها به عکس و ترانهاده ماتریس عکس توالی‌ها، به کد تبدیل شده و سپس هر کدام به یک شبکه یادگیری عمیق کانولوشنی تحویل داده می‌شود و در نهایت یک لایه مسطح^۳ نتایج سه شبکه را به صورت پاسخ نهایی برای پیش‌بینی مکان‌های فسفریلاسیون تبدیل می‌کند. نتایج حاصل از اعتبارسنجی متقابل ۱۰ باره نشان می‌دهد که پیش‌بینی‌کننده ConvoPhos، بالاترین کارایی را در بین سایر روش‌های مقایسه شده دارد.

۲- مواد و روش‌ها

۲-۱ داده‌ها

پایگاه داده‌های متفاوتی در زمینه فسفریلاسیون وجود دارند که از بین آنها می‌توان به [۱۹] qphos، [۲۰] dbptm، [۲۱] dbpaf، [۲۲] phospho.ELM، [۲۳] Phosida، [۲۴]

^۳ Flatten

^۱ Random Forest

^۲ Random Forest

بررسی روی نتایج نهایی و یافتن بهترین نتایج از بین طول پنجره‌های بین ۱۷ تا ۴۱، انتخاب شد و با توجه به اینکه ابزارهای دیگری چون musitedeep [۱۵] و NetPhos 3.0 [۲۹] نیز همین طول پنجره را انتخاب کرده‌اند به نظر انتخاب مناسبی است. با توجه به تعداد بسیار زیاد نمونه‌های منفی، اضافه کردن همه آنها به مجموعه داده‌های آموزشی می‌تواند بایاس زیادی در نتایج و تفسیر آنها ایجاد کند و در نتیجه دقت مدل کاهش یابد، بنابراین نمونه‌های منفی به صورت تصادفی به نسبت ۱:۱، طوری انتخاب شده‌اند که تعداد آنها با تعداد برجسب‌های مثبت برابر باشد. بعد از به دست آوردن توالی‌های نمونه‌های مثبت و منفی با استفاده از نرم افزار CD-Hit [۳۰] نمونه‌هایی که بیشتر از ۷۰ درصد با هم شباهت داشتند از مجموعه داده‌های آموزشی حذف شدند تا شباهت بیش از حد توالی‌ها به یکدیگر باعث ایجاد جانبداری در نتایج نشود.

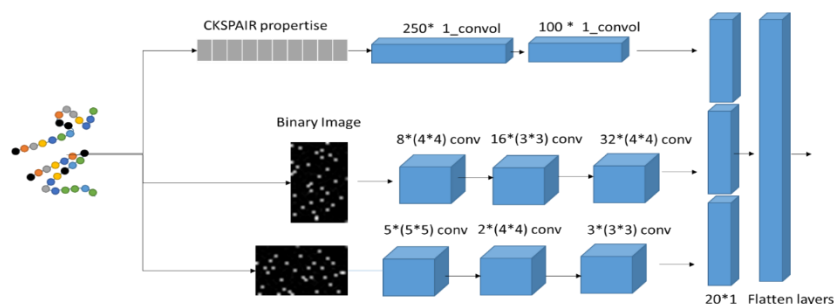
۳-۲ ساختار پیشنهادی

مدل یادگیری عمیق ارائه شده براساس مفاهیم شبکه‌های عمیق کانولوشنی و با ترکیب سه ساختار مختلف عمیق با ورودی‌های استخراج ویژگی CkSAApair و تبدیل توالی به عکس به صورت زیر می‌باشد (شکل ۱). همانگونه که از ساختار پیشنهادی موجود در این شکل قابل مشاهده است، در این مدل از توالی پروتینی، خاصیت CkSAApair، ماتریس عکس مربوط به نگاشت توالی و ترانزاده ماتریس عکس تولید شده در مرحله قبل استخراج می‌شود تا به عنوان ورودی استفاده شود.

PhoshoAt, dbpsp2 [۲۵] و phosphositeplus [۲۶] اشاره کرد. در این مطالعه برای اطمینان از کیفیت بالای داده‌ها، از داده‌های مربوط به پایگاه داده‌های phosphositeplus, dbpaf و P.ELM استفاده شده است، که همگی مربوط به داده‌های یوکاریوت‌ها می‌باشند. dbPAF یکی از کامل‌ترین و جدیدترین پایگاه داده‌ها در نوع خود می‌باشد در حالی که phosphositeplus پایگاه داده مرجع و پر استناد است و با توجه به اینکه در پیش‌بینی کننده‌های معرفی شده در سال‌های اخیر از جمله [۲۷] Deepkinzero, DeepPPSite [۲۸] و ابزارهای دیگر پر کاربرد بوده، به تنهایی نیز استفاده شد.

۲-۲ پیش پردازش داده‌ها

با توجه به اینکه فسفریلاسیون در یوکاریوت‌ها اغلب در باقیمانده‌های سرئین (S)، تیروزین (Y) و ترئونین (T) و در موارد خیلی محدودی در باقیمانده هیستیدین (H) اتفاق می‌افتد، برای ساخت داده‌های این مطالعه تنها از باقیمانده‌های S، T و Y استفاده شد. برای ساخت نمونه‌های با برجسب مثبت (فسفریله شده) در هر پایگاه داده، اگر هر یک از باقیمانده‌های S، T و Y برجسب مثبت داشت آن باقیمانده به همراه ۱۶ اسید آمینه قبل و بعد از آن که یک پنجره با طول ۳۳ آمینو اسیدی را تشکیل می‌دهد، به عنوان داده با برجسب مثبت برجسب گذاری می‌شود. سایر باقیمانده‌های (S، T و Y) که به عنوان جایگاه فسفریله گزارش نشده‌اند به عنوان نمونه‌های منفی در نظر گرفته شدند. قابل ذکر است که طول توالی استفاده شده پس از



شکل ۱ ساختار پیشنهادی بر مبنای شبکه‌های کانولوشنی استفاده شده به منظور پیش‌بینی جایگاه‌های فسفریلاسیون.

۲-۴ ویژگی CkSAApair

CkSAApair یک طرح رمزگذاری پراکنده است که توسط محققان در بخش‌های مختلف بیوانفورماتیک پروتئین‌ها [۳۱]، [۳۲] از جمله پیش‌بینی کننده‌های فسفریلاسیون استفاده شده است [۳۳ و ۳۴].

CkSAApair مخفف عبارت Composition of K-spaced Amino Acid Pairs است که در آن فرکانس جفت اسید آمینه محاسبه می‌شود که با k باقیمانده از هم جدا می‌شوند (مقدار K می‌تواند ۰، ۱، ۲، ۳، ۴ و ۵ باشد). این طرح اساساً نشان دهنده تعاملات کوتاه و بلند مدت بین باقیمانده‌ها در امتداد توالی است که منجر به عملکرد مناسبی در مسائل پیش‌بینی پروتئین می‌شود.

در این مطالعه، یک جایگاه فسفریله یا غیر فسفریله توسط یک قطعه توالی از ۳۳ اسید آمینه در نظر شده است. با توجه به اینکه، رمزگذاری CkSAApair به معنای ترکیب جفت‌های باقیمانده با فاصله k در قطعه است. با در نظر گرفتن $K=0$ ، ۴۰۰ جفت باقیمانده با فاصله صفر وجود دارد (به عنوان مثال، AA، AC، AD، ...، YY) به طوری که، بردار ویژگی را می‌توان به صورت رابطه‌ی ۱ تعریف کرد.

رابطه ۱

$$\left(\frac{N_{AA}}{N_{Total}}, \frac{N_{AC}}{N_{Total}}, \frac{N_{AD}}{N_{Total}}, \dots, \frac{N_{YY}}{N_{Total}} \right)_{400}$$

مقدار هر ویژگی نشان‌دهنده ترکیب جفت باقیمانده مربوطه در قطعه است. برای مثال، اگر جفت باقیمانده AA، تعداد m بار در قطعه ظاهر شود، ترکیب جفت باقیمانده AA برابر است با m تقسیم بر تعداد کل جفت‌های باقیمانده با فاصله صفر (NTotal) در توالی مورد بررسی. برای مقدار k، برابر با ۰، ۱، ۲، ۳، ۴ و ۵، مقدار NTotal به ترتیب ۳۰، ۳۱، ۳۲، ۳۳ و ۳۴ است [۳۵].

به طور کلی، مقادیر زیاد k می‌تواند منجر به افزایش تعداد ویژگی‌ها شده و به نتایج پیش‌بینی بهتر کمک کند، با این

حال، افزایش تعداد ویژگی‌ها در نهایت باعث افزایش پیچیدگی محاسبات می‌شود.

۲-۵ کدگذاری توالی‌ها به صورت عکس

با داشتن ۲۰ آمینو اسید و طول توالی ۳۳ می‌توان یک عکس ۲۱×۳۳ ایجاد کرد. در هر سطر به ازای آمینو اسید مربوطه مقدار یک و به ازای دیگر آمینو اسیدها مقدار صفر قرار داده می‌شود. برای مثال، اگر در سطر اول آمینو اسید A وجود داشته باشد آنگاه ستون مربوط به آن آمینو اسید برابر یک در نظر گرفته می‌شود. شکل ۲ این موضوع را نشان می‌دهد. تعداد ستون‌ها مساوی تعداد آمینو اسیدهای ممکن موجود در توالی است یعنی ۲۱ ستون و تعداد سطرها ۳۳ مورد است. سپس، ویژگی‌های CkSAApair به یک ساختار با دولایه کانولوشنی یک بعدی و ماتریس‌های نگاشت به عنوان ورودی به دو ساختار دوبعدی با سه لایه مختلف کانولوشنی وارد شده و در ادامه یک لایه مسطح ۲۰×۱، حاصل نهایی سه پیش‌بینی کننده را به صورت یک پاسخ نهایی با طبقه‌بندی فسریله و غیر فسفریله تحویل می‌دهد. قابل ذکر است که تابع فعال ساز استفاده شده در تمام بخش‌ها از نوع relu می‌باشد.

آمینو اسید

	A	C	P	L	T	F	G	H	I
A	1	0	0	0	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0
P	0	0	1	0	0	0	0	0	0
T	0	0	0	0	1	0	0	0	0

طول توالی

شکل ۲ ماتریس عکس مربوط به نگاشت توالی با طول ۳۳

۳- نتایج

برای ارزیابی عملکرد پیش‌بینی کننده‌های مختلف چهار پارامتر اصلی زیر تعریف می‌شوند.

مهم‌ترین ابزارهای ارزیابی پیش‌بینی کننده‌ها بر اساس موارد بالا به صورت حساسیت^۵، ویژگی^۶، صحت^۷، ضریب همبستگی متیو^۸ و دقت^۹ معرفی می‌شوند.

حساسیت: می‌توان به طور ساده حساسیت را به عنوان درصد مثال‌های مثبت تعریف کنیم، یعنی جایگاه‌های فسفریلاسیون که به درستی پیش‌بینی شده‌اند.

$$۲ \quad \text{حساسیت} = \frac{TP}{TP+FN}$$

ویژگی: این پارامتر درصد مثال‌های منفی درست پیش‌بینی شده را می‌دهد.

$$۳ \quad \text{ویژگی} = \frac{TN}{(TN+FP)} \times 100$$

صحت: در صورتی که اندازه کلاس‌ها یکسان باشد، درستی یا صحت پیش‌بینی کننده می‌تواند به عنوان میانگین ریاضی حساسیت و ویژگی بیان شود. هرچه صحت بیشتر باشد مقادیر اندازه‌گیری شده به مقدار واقعی نزدیک‌تر می‌باشند. مجموعه نتایج پیش‌بینی کننده‌ها در حالت تصادفی دارای صحت ۵۰ درصد می‌باشند.

$$۴ \quad \text{صحت} = \frac{TP+TN}{TP+FP+TN+FN} \times 100$$

دقت: این پارامتر میزان نزدیکی مقادیر اندازه‌گیری شده را نشان می‌دهد.

$$۵ \quad \text{دقت} = \frac{TP}{TP+FP}$$

TP^۱: جایگاه‌هایی که به عنوان مثبت شناسایی شده‌اند در حالی که مثبت هم بوده‌اند. (مثبت صحیح)

TN^۲: جایگاه‌های که به عنوان منفی پیش‌بینی شده‌اند و صحیح پیش‌بینی شده‌اند.

FP^۳: جایگاه‌هایی که اشتباهاً به عنوان جایگاه مثبت پیش‌بینی شده‌اند.

FN^۴: جایگاه‌هایی که اشتباهاً به عنوان جایگاه منفی پیش‌بینی شده‌اند. به عبارت دیگر می‌توان بیان داشت که در تأیید آزمایشگاهی محل‌های محتمل فسفریلاسیون (از نظر محاسباتی)، درستی موقعیت‌های پیش‌بینی شده و پیش‌بینی درست موقعیت‌ها دو شاخصی هستند که به طور مستقیم مورد توجه هستند. به صورت غیرمستقیم، مکمل یا معادل، از درستی موقعیت‌های رد شده و رد درست موقعیت‌ها نیز می‌توان صحبت کرد. انتخاب ضربدری دو شاخص از این چهار شاخص هم به نوعی توصیف کننده رفتار پیش‌بینی کننده است. این چهار شاخص و ارتباط آن‌ها در جدول ۱ نشان داده شده‌اند. در تشخیص محل فسفریلاسیون به صورت محاسباتی و تأیید آزمایشگاهی آن مهم است که حساسیت پیش‌بینی کننده به تشخیص درست بالا بوده (موارد تشخیصی، همه موارد واقعی را شامل شود) و علاوه بر آن اشتباه در تشخیص کم بوده و تشخیص از ارزش بالایی برخوردار باشد.

جدول ۱ نحوه توصیف کامل از کارایی تخمین‌گر محل فسفریلاسیون در شکل نشان داده شده است.

مثبت واقعی =	FN	+	TP
	+		+
منفی واقعی =	TN	+	FP
	منفی پیش‌بینی شده =		مثبت پیش‌بینی شده =
			پیش‌بینی شده

^۶ Specificity

^۷ Accuracy

^۸ Matthew's correlation coefficient

^۹ Precision

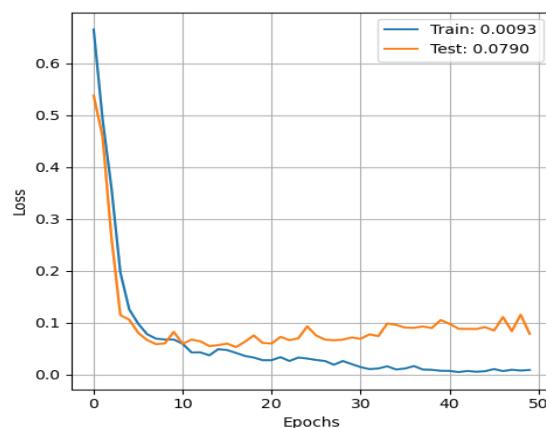
^۱ True Positive

^۲ True Negative

^۳ False Positive

^۴ False Negative

^۵ Sensitivity



شکل ۳ مراحل آموزش این مدل را نمایش می‌دهد. همانطور که مشاهده می‌شود مقدار خطای loss تا epoch شماره ۱۵ کاهش یافته و بعد از آن افزایش پیدا می‌کند. بنابراین، نتایج را تا epoch 15 می‌توان معتبر دانست.

جدول ۲ نتایج حاصل از پیش‌بینی جایگاه‌های فسفریلاسیون برای پایگاه داده‌های مختلف و با استفاده از روش پیشنهادی شکل ۱

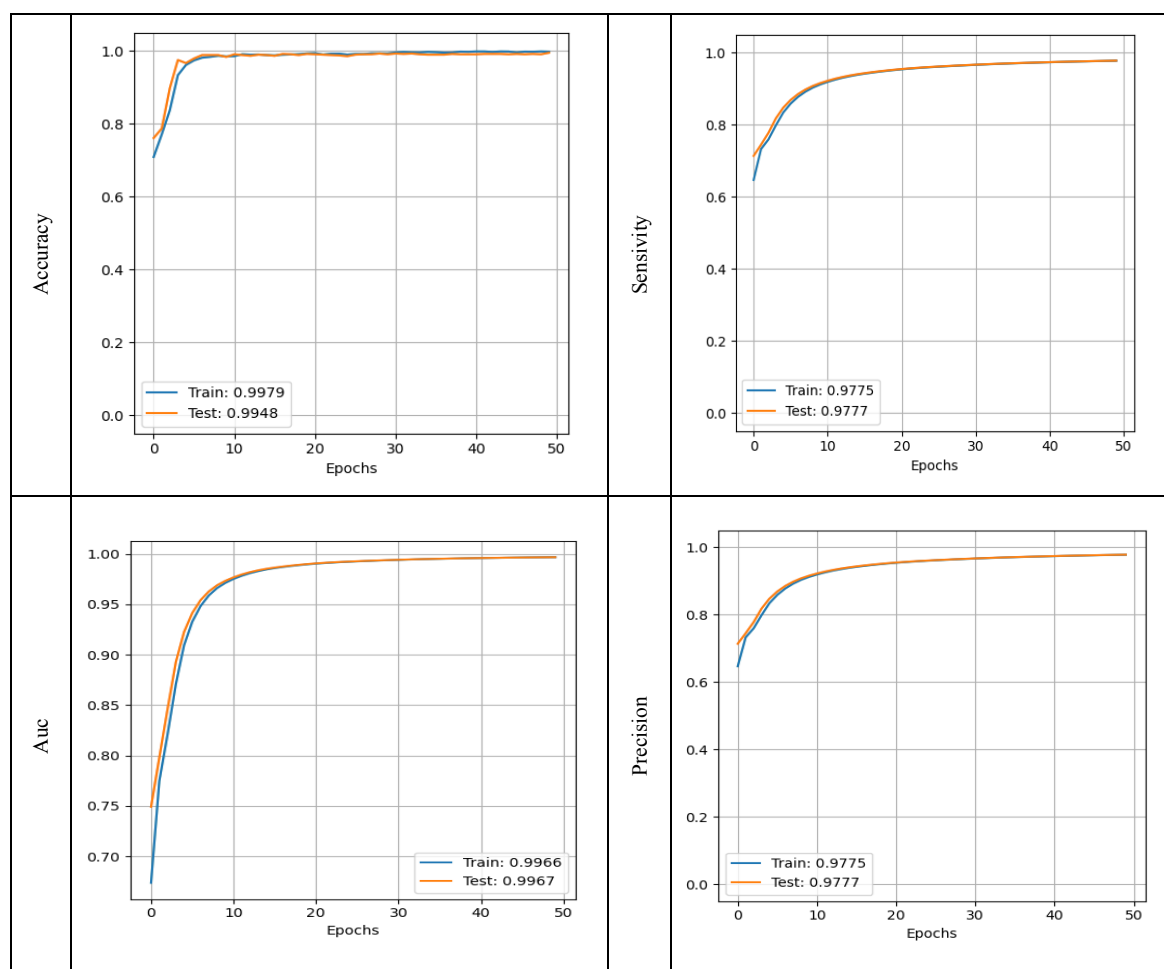
زمان اجرا	امتیاز AUC	دقت	حساسیت	صحت	پایگاه داده
۵ دقیقه	۰/۹	۰/۹	۰/۹	۰/۹۴	Phosphositeplus
۱۰ دقیقه	۰/۹۱	۰/۸۹	۰/۸۹	۰/۸۸	مجموع پایگاه داده‌ها

اینکه از تکرار ۱۵ به بعد شاهد افزایش تلفات داده‌های ارزیابی هستیم، نتایج برای دیگر پارامترها را محدود به همین تعداد تکرار در نظر گرفتیم.

در ادامه، پیش‌بینی‌کننده پیشنهادی از طریق معیارهای آماری مختلف، از جمله منحنی حساسیت، دقت، صحت و منحنی AUC برحسب تعداد تکرار مورد بررسی قرار گرفتند و نتایج در شکل ۴ قابل بررسی است.

بررسی‌ها نشان داد، در صورتی‌که برای پایگاه داده Phosphosite تنها از ماتریس عکس‌ها استفاده شود دقت مدل به‌دست آمده برابر با ۶۸ درصد و اگر تنها توالی آمینو اسیدی CkSAAPair به‌عنوان ورودی به‌کار گرفته شود مدل دارای دقت ۷۵ درصدی خواهد شد.

علاوه بر پارامترهای بیان شده در بالا، امتیاز AUC یک معیار طبقه‌بندی رایج یادگیری ماشین است و مخفف منطقه زیر منحنی ROC است. ROC نموداری است که رابطه بین نرخ مثبت واقعی و مثبت کاذب را ترسیم می‌کند. به‌طور کلی امتیاز AUC را می‌توان توانایی مدل برای طبقه‌بندی دقیق کلاس‌ها در مقیاس ۰ تا ۱ تفسیر کرد که در آن ۱ بهترین است و ۰/۵ به خوبی انتخاب تصادفی است. برای ارزیابی مدل از آزمون اعتبار سنجی متقابل استفاده کردیم، که یک فرایند آماری قوی برای فرار از مشکل بیش‌برازش است و روشی مناسب و پر کاربرد برای الگوریتم‌های طبقه‌بندی مختلف می‌باشد. ابتدا نمودار تلفات آموزش/ ارزیابی را برای ۱۰۰ تکرار رسم کردیم (شکل ۳) و سپس با توجه به



شکل ۴ نتایج مربوط به حساسیت، صحت، دقت و AUC به ازاء تکرارهای مختلف برای داده‌های مربوط به Phosphositeplus.

جدول ۳ مقایسه نتایج روش‌های معروف و روش ConvoPhos.

پیش‌بینی کننده	حساسیت	صحت	دقت	ضریب همبستگی متیو	معیار F1
DeepPhos71	۵۲/۱	۷۱/۵	۸۳/۵	۳۳/۳	۶۴/۳
PhosIDNSeq	۴۰/۲۵	۶۵/۱۵	۷۹/۵	۳۴/۶۵	۵۳/۰۵
Netphos 3.0	۱۹/۸۵	۵۴/۹	۶۵/۶	۱۳/۴	۳۰/۳
PPSP	۲۷/۸	۵۸/۹	۷۳/۱۵	۲۲/۵۵	۴۰/۱۵
Musite	۱۶	۵۵/۷	۷۶/۷	۱۸/۴	۲۶/۳۵
MusiteDeep	۳۴/۷	۶۲/۳۵	۷۶/۵	۲۹/۲	۴۷/۲۵
DeepPhos	۴۷/۸	۶۴/۵	۸۷/۷	۴۱/۴	۴۸/۷۷
PhosIDN	۵۲/۱	۷۱/۱	۸۳/۹	۴۵/۵	۶۴/۳

۸۰	۵۵	۸۹	۸۸	۸۹	ConvoPhos
----	----	----	----	----	-----------

هنوز هم تا ابزار بسیار کارآمد فاصله زیادی است و تلاش‌ها برای دستیابی به چنین ابزاری ادامه دارد.

در این مطالعه از یک معماری ترکیبی شامل معماری عمیق کانولوشنی با ورودی‌های مختلف ماتریس عکس ورودی، ترانهاده ماتریس عکس ورودی و ماتریس مربوط به استخراج ویژگی CkSAAPair از روی توالی‌های ورودی برای پیش‌بینی مکان‌های فسفریلاسیون استفاده شد.

با استفاده از اعتبارسنجی متقابل ۱۰ برابری، عملکرد ConvoPhos مورد بررسی قرار گرفت و حساسیت ۸۹ درصد، دقت ۸۹ درصد، و صحت ۸۸ درصد برای همه داده‌ها اندازه‌گیری شد. نتایج نشان داد که استفاده از ترکیب ورودی‌ها و شبکه‌های کانولوشنی، عملکرد بهتری نسبت به استفاده از هر یک از ورودی‌ها به تنهایی دارد. همچنین با مقایسه نتایج پیش‌بینی کننده معرفی شده در این مقاله با نتایج مربوط به دیگر روش‌ها، مشاهده شد که روش معرفی شده عملکرد بسیار بهتری از لحاظ دقت، صحت، ویژگی و AUC نسبت به سایر روش‌های مورد بررسی دارد.

رویکرد ارائه شده در این مقاله روشی کارآمد برای شناسایی مکان‌های فسفریلاسیون در یک توالی اولیه پروتئین ارائه می‌کند و به نظر می‌رسد که روش استخراج ویژگی CkSAAPair برای زیست‌شناس‌های مولکولی که بر روی جایگاه‌های فسفریلاسیون پروتئین و برای دانشمندانی که در حال توسعه سیستم‌های پیش‌بینی برای تغییرات پس از ترجمه مانند گلیکوزیلاسیون، نیتراسیون و فسفریلاسیون کار می‌کنند، اطلاعات ارزشمندی از توالی‌های پروتئینی می‌دهد.

۵- منابع

[1] H. Xu et al., "PTMD: A Database of Human Disease-associated Post-translational Modifications," *Genomics Proteomics Bioinformatics*, vol. 16, no. 4, pp. 244–251, Aug. 2018, doi: 10.1016/j.gpb.2018.06.004.

این در حالی است که طبق نتایج موجود در جدول ۲، دقت مدل پیشنهادی شکل ۱ بر روی همان داده‌ها ۹۴ درصد خواهد بود. این نتایج نشان می‌دهد که استفاده توأم از روش‌های فوق یعنی ترکیب ماتریس عکس از توالی آمینواسیدی و استخراج ویژگی CkSAAPair می‌تواند دقت مدل را تا حدود ۲۰ درصد بر روی این پایگاه داده افزایش دهد. همچنین، در این جدول نتایج به دست آمده از روش فوق برای تمام داده‌های سه پایگاه داده نیز آورده شده است. در ادامه در جدول ۳، عملکرد روش ConvoPhos را با چند پیش‌بینی کننده که از بهترین روش‌های موجود برای پیش‌بینی مکان‌های اتفاق افتادن فسفریلاسیون هستند مقایسه کردیم. این روش‌ها شامل روش‌های مبتنی بر یادگیری ماشین سنتی (Musite، Netphos3.0) و روش‌های مبتنی بر یادگیری عمیق (DeepPhos، PhosIDNseq، DeepPhos71، PhosIDN، PPSP و MusiteDeep) می‌باشند. همانطور که در جدول ۳ مشاهده می‌شود، ConvoPhos در شاخص‌های حساسیت، صحت، ضریب همبستگی متیو و معیار F1 نسبت به سایر روش‌ها بهترین نتایج را دارد.

۴- بحث

با توجه به کشف مشکلات موجود در بسیاری از فرایندهای موجودات زنده به دلیل انجام شدن یا نشدن فسفریلاسیون در پروتئین‌های خاصی از آن موجود، اهمیت فسفریلاسیون در پروتئین‌ها و افزایش بسیار زیاد توالی‌های جدید کشف شده و نیاز به کم کردن جایگاه‌هایی که نیاز به تایید روش‌های تجربی دارند، همگی بهبود روش‌های محاسباتی پیش‌بینی جایگاه‌های فسفریلاسیون را از نظر سرعت و دقت روز به روز با اهمیت‌تر می‌کند. با اینکه تاکنون ابزارهای پیش‌بینی کننده بسیار زیادی معرفی شده است اما

- [13] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," pp. 1–16, 2016.
- [14] H. R. Hassanzadeh and M. D. Wang, "DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins," *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, pp. 178–183, 2017.
- [15] D. Wang et al., "MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction," *Bioinformatics*, vol. 33, no. 24, pp. 3909–3916, 2017.
- [16] Y. Xie et al., "DeepNitro: Prediction of Protein Nitration and Nitrosylation Sites by Deep Learning," *Genomics Proteomics Bioinformatics*, vol. 16, no. 4, pp. 294–306, 2018.
- [17] F. Luo, M. Wang, Y. Liu, X. M. Zhao, A. Li, and J. Hancock, "DeepPhos: Prediction of protein phosphorylation sites with deep learning," *Bioinformatics*, vol. 35, no. 16, pp. 2766–2773, 2019.
- [18] L. Guo et al., "DeepPSP: A Global-Local Information-Based Deep Neural Network for the Prediction of Protein Phosphorylation Sites," *J Proteome Res*, vol. 20, no. 1, pp. 346–356, 2021.
- [19] K. Yu et al., "qPhos: a database of protein phosphorylation dynamics in humans," *Nucleic Acids Res*, vol. 47, no. D1, pp. D451–D458, Jan. 2019.
- [20] K.-Y. Huang et al., "dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications," *Nucleic Acids Res*, vol. 47, no. D1, pp. D298–D308, Jan. 2019.
- [21] S. Ullah et al., "DbPAF: An integrative database of protein phosphorylation in animals and fungi," *Sci Rep*, vol. 6, no. January, pp. 1–9, 2016.
- [22] H. Dinkel et al., "Phospho.ELM: A database of phosphorylation sites-update 2011," *Nucleic Acids Res*, vol. 39, no. SUPPL. 1, pp. 261–267, 2011.
- [23] F. Gnäd, J. Gunawardena, and M. Mann, "PHOSIDA 2011: the posttranslational modification database," *Nucleic Acids Res*, vol. 39, no. suppl_1, pp. D253–D260, 2010.
- [2] F. Ardito, M. Giuliani, D. Perrone, G. Troiano, and L. Lo Muzio, "The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review)," *Int J Mol Med*, vol. 40, no. 2, pp. 271–280, Aug. 2017, doi: 10.3892/ijmm.2017.3036.
- [3] K. McCance and S. Huether, *Pathophysiology: The Biologic Basis for Disease in Adults and Children*, 7th ed. Elsevier, 2014.
- [4] Z. Zahiri, N. Mehrshad, and M. Mehrshad, "DF-Phos: Prediction of Protein Phosphorylation Sites by Deep Forest," *The Journal of Biochemistry*, Dec. 2023.
- [5] Y. Dou, B. Yao, and C. Zhang, "PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine," *Amino Acids*, vol. 46, no. 6, pp. 1459–1469, 2014.
- [6] B. Trost and A. Kusalik, "Computational prediction of eukaryotic phosphorylation sites," *Bioinformatics*, vol. 27, no. 21, pp. 2927–2935, 2011.
- [7] N. Blom, S. Gammeltoft, and S. Brunak, "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites," *J Mol Biol*, vol. 294, no. 5, pp. 1351–62, 199.
- [8] A. K. Biswas, N. Noman, and A. R. Sikder, "Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information," *BMC Bioinformatics*, vol. 11, no. 1, p. 273, May.2010.
- [9] L. M. Iakoucheva et al., "The importance of intrinsic disorder for protein phosphorylation," *Nucleic Acids Res*, vol. 32, no. 3, pp. 1037–1049, 2004.
- [10] J. Gao, J. J. Thelen, a K. Dunker, and D. Xu, "Musite, a tool for global prediction of general and kinase-specific phosphorylation sites," *Mol Cell Proteomics*, vol. 9, no. 12, pp. 2586–2600, 2010.
- [11] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] A. Jones, H. Ismail, J. H. Kim, R. Newman, and B. K. Dukka, "RF-Phos: Random forest-based prediction of phosphorylation sites," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Nov. 2015, pp. 135–140.

- 1633–1649, 2004.
- [30] W. Li and A. Godzik, “Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
- [31] Z. Ju and S.-Y. Wang, “Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou’s general pseudo amino acid composition,” *Gene*, vol. 664, pp. 78–83, 2018.
- [32] Md. M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song, and Z. Zhang, “Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of k-Spaced Amino Acid Pairs,” *PLoS One*, vol. 10, no. 6, p. e0129635, Jun. 2015.
- [33] X. Zhao, W. Zhang, X. Xu, Z. Ma, and M. Yin, “Prediction of Protein Phosphorylation Sites by Using the Composition of k-Spaced Amino Acid Pairs,” *PLoS One*, vol. 7, p. e46302, Jan. 2012.
- [34] S. Liu, C. Cui, H. Chen, and T. Liu, “Ensemble learning-based feature selection for phosphorylation site detection,” *Front Genet*, vol. 13, Oct. 2022.
- [35] Y.-Z. A. N. D. W. X.-F. A. N. D. W. C. A. N. D. Y. R.-X. A. N. D. Z. Z. Chen Zhen AND Chen, “Prediction of Ubiquitination Sites by Using the Composition of k-Spaced Amino Acid Pairs,” *PLoS One*, vol. 6, no. 7, pp. 1–8, Jan. 201.
- [24] Y. Shi et al., “dbPSP 2.0, an updated database of protein phosphorylation sites in prokaryotes,” *Sci Data*, vol. 7, no. 1, pp. 1–9, Dec. 202.
- [25] J. L. Heazlewood et al., “PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor,” *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D1015–21, Jan. 2008.
- [26] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek, “PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations,” *Nucleic Acids Res*, vol. 43, no. D1, pp. D512–D520, 2015.
- [27] I. Deznabi, B. Arabaci, M. Koyutürk, and O. Tastan, “DeepKinZero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases,” *Bioinformatics*, vol. 36, no. 12, pp. 3652–3661, Jun. 2020.
- [28] S. Ahmed, M. Kabir, M. Arif, Z. U. Khan, and D.-J. Yu, “DeepPPSite: A deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information,” *Anal Biochem*, vol. 612, p. 113955, 2021.
- [29] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, and S. Brunak, “Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence,” *Proteomics*, vol. 4, no. 6, pp.

Designing a deep learning model to predict phosphorylation sites

Zeynab Zahiri¹, Nasser Mehrshad^{2*}

1. Ph.D. Candidate, Department of Electronics, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

2. Professor, Department of Electronics, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Receipt: 2024/06/10

Accepted: 2025/08/02

Abstract

Phosphorylation is one of the most important types of post-translational modification (PTM) that plays an important role in protein function studies and experimental design. Considering the importance of phosphorylation in proteins and the increasing number of protein sequences in databases, the need to improve computational methods for predicting phosphorylation sites becomes increasingly important in terms of both speed and accuracy. Although many predictive tools have been introduced to predict phosphorylation sites using various machine learning methods, there is still a long way to go to develop a highly efficient tool, and efforts to achieve this continue. Recent studies have shown that deep learning-based methods are the best approach for predicting phosphorylation sites. This is because deep learning, as an advanced machine learning method, can automatically recognize complex representations of phosphorylation patterns from raw sequences, therefore providing a powerful tool for improved phosphorylation site prediction.

In this study, a hybrid structure based on the convolutional deep learning method named ConvoPhos has been introduced for predicting phosphorylation sites. In such a way that the CkSAApair feature vector obtained from sequences is used as the input for part of the classifier and the conversion of sequences to images, as the input for another part of the convolutional networks. The results of 10-fold cross-validation show an accuracy value of 94% for phosphosite data and an AUC of 90%, which is the highest performance compared to the other methods.

Keywords: Protein phosphorylation, feature extraction, predictor, convolutional deep learning network