Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
DOI: 10.22034/jscit.2026.548969.2172
Volume 15, Number 1, Spring 2026, pp. 38-57
Received: 09/24/2025, Revised: 01/02/2026, Accepted: 02/14/2026

# Boosting Predictive Performance for a Case Study in Bioinformatics by Applying a Learning-Based Predictor (Protein-Peptide Interactions at Binding Residues Level)

Shima Shafiee[1] and Abdolhossein Fathi[2*]
1-Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
2*- Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
[1] shafiee.shima@razi.ac.ir, [2*] a.fathi@razi.ac.ir

Corresponding author's address: Abdolhossein Fathi, Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran

**Abstract-** Predicting protein-peptide interactions plays an important role in understanding abnormal cellular behaviors, elucidating biological functions, designing drugs, and therapeutic strategies. Experimental and laboratory methods for detecting residues involved in protein-peptide interactions suffer from several limitations such as high labor costs, time-consuming procedures, and dependence on expensive instruments and equipment. To overcome these challenges, a computational predictor based on deep learning (deep gated recurrent unit) and machine learning (an improved support vector machine) was designed and implemented. The proposed model exploits various features extracted from protein structures, including evolutionary, sequence-based, structure-based, and physicochemical characteristics. The developed hybrid model uses two sampling techniques (under-sampling and over-sampling) to handle the over-prediction of the non-binding class caused by the inherent imbalance in protein data. The predictive performance of the proposed model was evaluated on two standard datasets (Sparks, Wei) derived from the BioLip database, and compared with several competing computational methods, achieving favorable results. These improvements include higher precision (approximately 1%), F-measure (at least 21.5%), balance between sensitivity and specificity (about 2.1%), enrichment factor, heat map analysis, and success rate based on protein domains.

**Keywords-** protein-peptide interaction, over-prediction, binding or non-binding, machine learning, deep learning.

مجله علمی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
شناسه دیجیتال: 10.22034/jscit.2026.548969.2172
جلد ۱۵، شماره ۱، بهار ۱۴۰۵، صفحه ۳۸-۵۷
دریافت: ۱۴۰۴/۰۷/۰۲، بازنگری: ۱۴۰۴/۱۰/۱۲، پذیرش: ۱۴۰۴/۱۱/۲۵

# تقویت عملکرد پیش‌بینی برای یک مطالعه موردی در بیوانفورماتیک با استفاده از یک پیش‌بینی کننده مبتنی بر یادگیری (برهمکنش‌های پروتئین-پپتید در سطح باقیمانده‌های پیوندی)

شیما شفیعی[۱] و عبدالحسین فتحی*[۲]
۱- گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه رازی، کرمانشاه، ایران.
*۲- گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه رازی، کرمانشاه، ایران.
[1] shafiee.shima@razi.ac.ir, [2*] a.fathi@razi.ac.ir

* نشانی نویسنده مسئول: عبدالحسین فتحی، کرمانشاه، طاق بستان، دانشگاه رازی، دانشکده فنی و مهندسی، گروه مهندسی کامپیوتر و فناوری اطلاعات.

**چکیده- پیش‌بینی برهمکنش‌های پروتئین-پپتید، در درک رفتارهای غیرطبیعی سلولی، تبیین عملکردهای بیولوژیکی، طراحی دارو و راهبردهای درمانی نقش مهمی دارد. روش‌های تجربی و آزمایشگاهی برای تشخیص باقیمانده‌های درگیر در برهمکنش پروتئین-پپتید، با محدودیت‌هایی مانند: هزینه‌های بالای نیروی انسانی، زمان‌بر بودن آزمایش‌ها و وابستگی به ابزارها و تجهیزات گران قیمت، همراه هستند. برای غلبه بر این محدودیت‌ها، یک پیش‌بینی‌کننده محاسباتی مبتنی بر بر یادگیری عمیق (واحد بازگشتی گیت‌دار عمیق) و یادگیری ماشین (ماشین بردار پشتیبان بهبود یافته) طراحی و پیاده‌سازی شده است. این مدل پیشنهادی، از انواع خصیصه‌های مستخرج شده از ساختارهای پروتئینی شامل ویژگی‌های تکاملی، مبتنی بر توالی، مبتنی بر ساختار و فیزیکوشیمیایی استفاده می‌کند. در مدل ترکیبی توسعه‌یافته از دو تکنیک نمونه‌برداری کاهشی و افزایشی برای مدیریت بیش پیش‌بینی کلاس غیرپیوندی ناشی از عدم توازن ذاتی داده‌های پروتئینی، استفاده شده است. عملکرد مدل پیشنهادی با دو مجموعه داده استاندارد (Sparks,Wei) برگرفته از پایگاه داده BioLip، سنجیده شده و در مقایسه با مدل‌های محاسباتی رقیب، موفق به أخذ نتایج مطلوب‌تر شده است. این نتایج شامل بهینگی از نظر معیارهای دقت (تقریباً۱درصد)، اندازه‌گیری-اف (حداقل۲۱/۵ درصد)، تعادل مابین حساسیت و خاصگی (تقریباً ۲/۱درصد)، فاکتور غنی‌سازی، نقشه گرمایی و نرخ موفقیت براساس دامنه‌های پروتئینی است.**

**واژه‌های کلیدی:** برهمکنش پروتئین-پپتید، بیش پیش‌بینی، پیوندی یا غیرپیوندی، یادگیری ماشین، یادگیری عمیق.

## ۱- مقدمه

پروتئین‌ها به عنوان یکی از مولکول‌های اساسی و فرمانروای کوچک‌ترین واحد حیات و بدن انسان (سلول) هستند که برای فرآیندهای بیولوژیکی و متابولیکی ضروری می‌باشند. عملکردهای بیولوژیکی درشت مولکول‌هایی مانند پروتئین به تعامل‌های آن‌ها با سایر مولکول‌ها بستگی دارد. بنابراین، پروتئین‌ها به ندرت به تنهایی عمل می‌کنند و اغلب با سایر لیگاندها تشکیل کمپلکس می‌دهند [۱, ۲]. برهمکنش‌های بین مولکولی مانند پروتئین-پپتید، ۱۵ تا ۴۰ درصد از برهمکنش‌های پروتئینی را تشکیل می‌دهند [۳]. ماهیت برهمکنش‌های پروتئین-پپتید غیرکووالانسی می‌باشد که در آن‌ها فاصله بین اتم‌ها می‌تواند متفاوت باشد. چنین برهمکنش‌هایی شامل پیوندهای هیدروژنی (حدود ۲/۵ تا ۳/۵ آنگستروم (Å))، نیروهای واندروالس (حدود۲ تا ۴ آنگستروم (Å))، برهمکنش‌های آب‌گریز

(معمولاً ۳ تا ۵ آنگستروم (Å)) و برهمکنش‌های الکترواستاتیک (تقریباً ۳ تا ۶ آنگستروم (Å)) می‌باشند. برخلاف پیوندهای کووالانسی که شامل اشتراک جفت الکترون می‌باشند، پیوندهای غیرکووالانسی برگشت‌پذیر هستند و شامل اشتراک یا انتقال الکترون بین اتم‌ها نیستند. لذا، یک برهمکنش غیرکووالانسی بین پروتئین و پپتید، تشکیل کمپلکس‌های گذرا[۱] را تسهیل می‌کند که برای فرآیندهای بیولوژیکی مختلف کاربرد دارد. این فرآیندها مانند: انتقال سیگنال، ایمنی سلول، فعالیت آنزیم‌ها و انتقال اطلاعات سلولی می‌باشند. انتقال اطلاعات باعث مرگ سلول‌ها یا تحریک ترشح هورمون‌ها برای مقاصد درمانی نوین و هدفمند می‌باشد. بنابراین بررسی و تشخیص برهمکنش‌های پروتئینی نقش مهمی در درمان بیماری‌ها دارند. این بیماری‌ها می‌توانند شامل: اختلالات نورودژنراتیو، بیماری آلزایمر [۴, ۵] و بیماری پارکینسون [۶, ۷] باشند. علاوه بر این، برهمکنش‌های پروتئینی با ترکیب پپتیدهای نفوذ کننده سلولی با لیگاندهای هدف در درمان سلول‌های سرطانی می‌توانند حائز اهمیت باشند. ساخت داروهایی مانند داروهای شیمی درمانی و واکسن‌های ضد ویروسی برای بیماری‌های عفونی مانند آنفولانزا [۸]، $HIV_1$ [۹] یا COVID-19 [۱۰]، اهمیت تشخیص برهمکنش‌های پروتئین-پپتید را مضاعف می‌کنند. در مثالی دیگر، می‌توان به کاربردهای متنوع پپتیدها برای درمان بیماری‌های متابولیک، بیماری‌های انکولوژیک، بیماری‌های اعصاب و روان، بهبود عملکرد سیستم ایمنی بدن، ترمیم عضلات و طراحی داروهای خاص برای شیمی‌درمانی اشاره نمود [۱۰]. همچنین کاربردهای داروهای پپتید در صنعت پزشکی و زیبایی مانند پی آر پی و بوتاکس می‌باشند. در باب اهمیت مطالعاتی پروتئین‌ها، می‌توان به زمینه‌ی مهندسی و طراحی پروتئین اشاره نمود. امروزه صاحب نظران با طراحی پروتئین‌هایی با جایگاه‌های پیوندی خاص می‌توانند پروتئین‌های ضد سرطانی را طراحی کنند [۸, ۹, ۱۱]. تا کنون، اهمیت و ارزش حل مسئله بیوانفورماتیک مورد هدف (پیش‌بینی باقیمانده‌های پیوندی[۲]و جایگاه‌های پیوندی[۳] در برهمکنش‌های پروتئین-پپتید) با توجه به تشریح مواردی مانند: اهمیت برهمکنش‌های پروتئین-پپتید، پپتیدها و پروتئین‌ها توجیه گردیده است. در همین راستا، مدل‌های آزمایشگاهی برای تشخیص برهمکنش‌های پروتئین- پپتید، موفقیت‌هایی را کسب کرده‌اند. علی‌رغم موفقیت‌های اکتسابی، مدل‌های مذکور از محدودیت‌هایی رنج می‌برند. این محدودیت‌ها: زمان بربودن، هزینه بربودن، وابستگی به نیروی انسانی، اندازه کوچک پپتیدها، میل پیوندی ضعیف[۴] و انعطاف‌پذیری پپتیدها می‌باشند. هدف کلی این مطالعه، طراحی و پیاده‌سازی یک پیش‌بینی کننده محاسباتی مبتنی بر یادگیری (یادگیری ماشین و یادگیری عمیق) برای بهبود یک مسئله بیوانفورماتیکی تحت عنوان پیش‌بینی برهمکنش پروتئین-پپتید است. به‌گونه‌ای که منظور از برهمکنش همان فعل و انفعال‌های میان درشت مولکول پروتئین و لیگاند پپتید است. پیش‌بینی محاسباتی برای برهمکنش زیست مولکولی مورد هدف با تمرکز بر دو مطالعه موردی یعنی پیش‌بینی باقی‌مانده‌های اسیدآمینه‌ای پیوندی و پیش‌بینی جایگاه‌های پیوندی برهمکنش مربوطه و با استفاده از ساختارهای اول و سوم پروتئینی می‌باشد. برحسب محدودیت‌های مدل‌های مبتنی بر ساختارهای سوم (عدم شناسایی ساختار کریستالی، اطلاعات نویزی و وابستگی به ابزاهای جانبی) و محدودیت‌های مدل‌های تجربی و آزمایشگاهی (هزینه‌بر، زمان‌بر، وابستگی به کاربر آزمایشگاه و پر زحمت)، توانستیم یک مدل پیشنهادی تحت عنوان GSPPIR[۵] طراحی و پیاده‌سازی کنیم که متکی بر مزایای مدل‌های مبتنی بر توالی و مبتنی بر یادگیری است. به بیانی دیگر، مزایای ماژول‌های طراحی شده برای معماری پیشنهادی، توانسته کارایی و أثربخشی مدل پیشنهادی GSPPIR را برای پیش‌بینی باقیمانده‌های پیوندی در برهمکنش‌های پروتئینی ارضا نماید. این مزایا مواردی شامل: استخراج خودکار ویژگی‌های معنادار، ضبط الگوها و وابستگی‌ها طولانی در داده‌های متوالی و ترتیب دار، کاهش منابع محاسباتی و متعاقبا کاهش زمان آموزش کمتر، مدیریت بیش‌برازش با لایه حذف تصادفی[۶]، مدیریت گرادیان محوشدگی با گیت‌های خاص (آپدیت و ریست) می‌باشند. مزایای مذکور توانسته‌اند أهمیت ماژول استخراج کننده ویژگی‌های معنادار و مبتنی بر یادگیری عمیق (DGRU[۷]) را دو چندان توجیه کنند. در پژوهش حاضر، استفاده از ماژول طبقه‌بند مبتنی بر یادگیری ماشین (SVM[۸]) بهبود یافته به عنوان طبقه‌بند دوکلاسه توانسته پیچیدگی مدل پیشنهادی را کاهش دهد. سایر مزایای طبقه‌بند ماشین بردار پشتیبان بهبود یافته، توانسته کارایی طبقه‌بند مبتنی بر یادگیری ماشین را بیش از پیش توجیه می‌کند. این مزایا مواردی همچون: پشتیبانی از هسته‌های متنوع برای تبدیل فضای ویژگی به یک فضای غیرخطی و دسته‌بندی داده‌های پیچیده زیستی، مرزبندی بهینه بین کلاس‌ها با أبرصفحه و کاهش خطای تعمیم[۹]، بهبود تفسیرپذیری، پایداری و مدیریت نویز داده‌های زیستی می‌باشند. با توجه به محدودیت‌های مطرح شده، نوآوری‌های اصلی این مطالعه عبارتند از:

الف) طراحی ماژول مبتنی بر یادگیری عمیق برای پردازش بهینه توالی‌های پروتئینی و استخراج ویژگی‌های معنادار: ماژول استخراج کننده ویژگی مبتنی بر یادگیری عمیق برای استخراج خودکار ویژگی‌های معنادار، بهبود بیش‌برازش، رسیدگی به گرادیان محوشدگی، مدیریت وابستگی‌ها طولانی برد در داده‌های ترتیب‌دار و

بهبود عملکرد پیش‌بینی، طراحی شده است.

ب) به کارگیری ماژول مبتنی بر یادگیر ماشین برای تعادل بهینه مابین کارایی و قدرت تعمیم پذیری: ماژول طبقه‌بند و مبتنی بر یادگیری ماشین بهبود یافته برای کاهش پیچیدگی محاسباتی، سهولت تنظیم و رفع خطا، بهبود خطای تعمیم، افزایش تفسیرپذیری و مدیریت نویز داده‌های زیستی طراحی شده است.

پ) اعمال تکنیک تقویت داده برای مدیریت محدویت ذاتی داده‌های پروتئینی (عدم تعادل شدید): برای مدیریت بیش پیش‌بینی[۱۰] کلاس غیرپیوندی، از دو تکنیک تقویت داده تحت عنوان نمونه‌برداری کاهشی[۱۱] و نمونه‌برداری افزایشی[۱۲] ، صرفا در مرحله‌ی آموزش استفاده شده است. در واقع، دو تکنیک نمونه برداری مذکور، برای کاهش سوگیری ناشی از عدم تعادل شدید مابین باقیمانده های اسید آمینه پیوندی و غیرپیوندی، بهبود توزیع واقعی داده و تعمیم‌پذیری مدل پیشنهادی استفاده شده است.تکنیک NearMiss، توزیع کلی کلاس اکثریت (غیرپیوندی) را مخدوش نمی‌کند. بلکه نمونه‌های دور از مرز را حذف می‌کند. یعنی مدل آموزش دیده، در تشخیص نمونه-های اکثریت (غیرپیوندی) جدید که در مناطق مختلف فضای ویژگی قرار دارند، قابلیت تعمیم مناسبی خواهد داشت. لذا بهبود مقدار اندازه گیری- اف و تعادل مابین حساسیت و خاصگی قابل توجیه است. تعمیم‌پذیری مبتنی بر SMOTE با یادگیری مرز تصمیم‌گیری قوی‌تر برای کلاس اقلیت (پیوندی)، قابل ارائه است. به گونه‌ای که این یادگیری از بیش برازش برای تعداد محدود نمونه‌های واقعی جلوگیری می نماید

ت) مدیریت محدودیت‌های مبتنی بر ساختار: استخراج ویژگی‌های متنوع (تکاملی، مبتنی بر توالی، مبتنی بر ساختار و فیزیکوشیمیایی) با هدف رسیدگی به محدودیت‌های مدل‌های مبتنی بر ساختار انجام شده است.

ث) طراحی یک مدل ترکیبی (یادگیری عمیق و یادگیری ماشین) هدفمند: یک مدل ترکیبی توسعه‌یافته با اهدافی همچون استخراج خودکار ویژگی‌های معنادار، بهبود فرآیند یادگیری مدل پیشنهادی ، مدیریت محدودیت بیش‌برازش، بهبود فرآیند تشخیص کلاس باقیمانده‌های اسیدآمینه و افزایش تعمیم‌پذیری مدل پیشنهادی طراحی شده است.

ج) مدیریت بیش-برازش برای مجموعه داده نامتعادل و مقاومت در برابر نویز و تعمیم‌پذیری مفید: مدیریت بیش-برازش، مقاومت در برابر نویز و تعمیم‌پذیری کارا از طریق لایه‌های حذف تصادفی، بردارهای پشتیبان و بیشینه حاشیه مابین ابر صفحه و بردارهای پشتیبان قابل ارائه است.

درادامه، این مقاله به صورت ذیل سازماندهی شده است: سابقه پژوهش در بخش ۲، ارائه شده است. بخش ۳، شامل مواد و روش می‌باشد. ارزیابی و مقایسه عملکرد مدل پیشنهادی در بخش ۴ نیز می‌باشد. انواع تحلیل‌ها و ارزیابی‌های بین رشته‌ای در بخش ۵، برای اثبات کارآمدی بیشتر مدل پیشنهادی، گزارش شده است. در نهایت در بخش ۶، مقاله با نتیجه‌گیری و پیشنهادهایی برای پژوهش‌های آینده نیز پایان می‌یابد.

## ۲- سابقه پژوهش

برحسب پیشینه‌ی تحقیق [۳, ۱۲, ۱۳]، پیش‌بینی بقایای پیوندی پروتئین- پپتید با تکنیک‌های محاسباتی مدرن، نوید پیشرفت و کارایی در زمینه‌های متنوع مانند: انفورماتیک سلامت، کشاورزی و محیط‌زیست را می‌دهد.

مدل‌های پیشگو به دو دسته‌ی تجربی و محاسباتی گروه‌بندی می‌شوند که با استفاده از توالی‌ها و ساختارهای پروتئینی، فرآیند شناسایی برهمکنش‌های پروتئین-پپتید را انجام می‌دهند. علی رغم پیشرفت‌های بیولوژیکی ساختاری برای تشخیص برهمکنش‌های پروتئین-پپتید، روش‌های تجربی و آزمایشگاهی همچنان دارای محدودیت‌های قابل توجهی می‌باشند. لذا، این روش‌ها می‌توانند زمانبر، هزینه‌بر و از نظر فنی به خاطر میل ترکیبی ضعیف، اندازه کوچک و انعطاف‌پذیری پپتیدها نیز چالش برانگیز و مستعد دقت پیش‌بینی ضعیف باشند [۱۲, ۱۳]. در همین راستا، توسعه مدل‌های محاسباتی مدرن مانند مدل‌های مبتنی بر یادگیری برای تشخیص کمپلکس‌های پروتئین-پپتید در سطح باقیمانده‌های پیوندی، قوت پیدا کرده است. مدل‌های یادگیری براساس استفاده از ساختارهای اول یا سوم پروتئینی به دو گروه دسته‌بندی شده‌اند. این گروه-ها همان مبتنی بر توالی و مبتنی بر ساختار می‌باشند. با وجود محبوبیت مدل‌های مبتنی بر ساختار، آن‌ها از محدودیت‌هایی مانند عدم شناسایی ساختارهای کریستالی، وابستگی به ابزارهای مختلف جانبی، ایجاد نویز و متعاقبا کاهش عملکرد پیش‌بینی رنج می‌برند. بنابراین، مدل‌های یادگیری و مبتنی بر توالی به عنوان یک ابزار جایگزین و ارزشمند برای هدایت مدل‌های تجربی، مدل‌های ساختاری و بهبود عملکرد تشخیص برهمکنش‌های پروتئینی توسعه یافته‌اند. به لطف محاسبات سنتی، اولین پیش‌بینی‌ها برای برهم‌کنش‌های پروتئین-پپتید با تکنیک‌هایی خاص همچون: ترازهای توالی چندگانه [۱۴]، پروب‌های نقشه‌برداری مبتنی بر دینامیک مولکولی به همراه ماتریس امتیازدهی خاص موقعیت[۱۳] [۱۵] وکاوشگرهای نقشه‌برداری برای تشخیص و خوشه‌بندی باقیمانده‌های پیوندی [۱۶] انجام شده‌اند. علی‌رغم دستیابی به نتایج معقول توسط

مدل‌های پیشگویی ابتدایی، کارایی آن‌ها به دلیل اتکای شدید به ویژگی‌های دست‌ساز[۱۴]، ناتوانی در مدیریت مجموعه داده‌های نامتعادل و کاهش عملکرد پیش‌بینی محدود شده است. در سال‌های اخیر مدل‌های یادگیری (یادگیری ماشین و یادگیری عمیق) توانسته‌اند برای انجام پیش‌بینی‌ها و استخراج اطلاعات مفید و در عین کاهش هزینه‌ها، کاهش وابستگی به تجهیزات آزمایشگاهی و کاهش مدت زمان به پیشرفت‌های چشم‌گیر دست یابند. پیش‌بینی برهمکنش‌های پروتئین-پپتید با مدل‌های یادگیری ماشین یک نقطه شروع برای پیش‌بینی‌های مبتنی بر محاسبات مدرن است. به گونه‌ای که، کارایی و اهمیت مدل‌های یادگیری ماشین با مزایایی مانند: توانایی پردازش حجم بالایی از داده‌های زیستی، شناسایی و تفسیر الگوهای پیچیده، سرعت و دقت مناسب برای پیش‌بینی برهمکنش‌های پروتئینی توجیه می‌شوند. نخستین مدل‌های یادگیری ماشین توانسته‌اند پیش‌بینی موقعیت و پیکربندی دقیق برهم‌کنش‌های پروتئینی را از طریق ترکیب با ابزارهای داکینگ، ارائه دهند [۱۷]. همچنین، این مدل‌های یادگیری توانسته‌اند نتایج حاصل از دینامیک مولکولی را با پیش‌بینی تغییرات انرژی پیوند، تقویت کنند [۱۸]. در ادامه، با توجه به ماهیت مدل پیشنهادی، جهت‌گیری‌های پیش‌بینی‌های مبتنی بر یادگیری با استفاده ساختارهای اول پروتئینی (مبتنی بر توالی) برای تشخیص برهمکنش‌های پروتئینی (باقیمانده‌های پیوندی و جایگاه‌های پیوندی) مشخص می‌شوند. در همین راستا، تکنیک‌های پیشگو و مبتنی بر یادگیری ماشین متنوع همچون: ماشین بردار پشتیبان به عنوان طبقه‌بند انفرادی[۱۹]، ساخت ویژگی مبتنی بر برنامه نویسی ژنتیک بهبود یافته به همراه طبقه‌بند ماشین بردار پشتیبان و خوشه‌بند ترتیب نقاط برای ساختار خوشه‌بندی[۱۵] [۲۰]، ساخت ویژگی مبتنی بر برنامه نویسی ژنتیک بهبود یافته توأم با طبقه‌بند تجمعی(کیسه‌بندی[۱۶]) [۲۱] و ترکیب ماشین بردار پشتیبان با دو الگوریتم مبتنی بر الگو[۱۱] توسعه یافته‌اند. مدل‌های مذکور، مزایایی مانند: افزایش استحکام در برابر نویز، مدیریت بایاس، سهولت تنظیم، مدیریت بیش برازش، کنترل داده گمشده، بهبود تفسیرپذیری و تشخیص را ارضا کرده‌اند. در چند سال اخیر، تکنیک‌های یادگیری عمیق به عنوان زیر شاخه‌ای از یادگیری ماشین، بنا به مزایایی برای پیش بینی برهمکنش‌های پروتئینی مورد استقبال واقع شده‌اند. این مزایا عبارتند از: خودکارسازی استخراج ویژگی‌ها و الگوهای سلسله مراتبی، معنادار، سطح بالا و پیچیده، بهبود قابلیت تطبیق‌پذیری، افزایش انعطاف‌پذیری، بهبود بیش‌برازش[۱۷] و تعمیم‌پذیری، رسیدگی به گرادیان محوشدگی[۱۸]، ضبط وابستگی‌های طولانی مدت[۱۹] و بهبود تشخیص برای برهمکنش پروتئین-پپتید. با وجود موفقیت‌های مدل‌های مبتنی بر یادگیری ماشین، مدل‌های مبتنی بر یادگیری عمیق از جمله Visual PepPFN ، [۲۵] PepCNN ،[۲۴] PepBCL ،[۲۳] PepNN-Seq ،[۲۲] [۳۸] ، E2Epep [۳۹] ، DPPBP [۴۰] ، توسعه یافته‌اند که از اطلاعات مبتنی بر توالی استفاده کرده‌اند. پژوهش‌های انجام شده برای تشخیص برهمکنش‌های پروتئین-پپتید با استفاده از معماری‌های متنوع مانند: شبکه‌های عصبی پیچشی [۲۲]، مدل‌های زبان مبتنی بر محتوا و معماری مبتنی بر توجه برای وظیفه خاص[۲۳]، مدل زبانی از پیش آموزش دیده پروتئینی [۲۴]، مدل زبانی از پیش آموزش دیده و یک ماژول یادگیری متقابل[۲۵]، مدل‌های زبانی پروتئینی [۳۹,۳۸]، و طبقه‌بند باینری دو طرفه [۴۰] طراحی و پیاده‌سازی شده‌اند.

با توجه به بررسی پیشینه تحقیق، محدودیت‌های بارز روش‌های رقیب عبارتند از: الف) زمانبر، هزینه‌بر بودن و وابستگی به کاربر انسانی، ب)وابستگی به اطلاعات ساختاری ناموجود، نویز مربوطه و کاهش عملکرد پیش‌بینی، پ)سوگیری به کلاس اکثریت، ت) عدم استخراج الگوهای معنادار و کاهش دقت پیش‌بینی. لذا در مطالعه حاضر، سعی نمودیم با نوآوری‌های مطرح شده (بخش ۱)، حدالامکان محدودیت‌های مذکور را رسیدگی کنیم.

## ۳- مواد و روش

### ۳-۱- مدل پیشنهادی

روند کلی مدل پیشنهادی در شکل ۱ قابل مشاهده است. مراحل مدل پیشنهادی GSPPIR با استفاده از توالی‌های پروتئینی به شرح زیر است: در ابتدا در بخش آماده‌سازی داده، یک توالی پروتئین از پایگاه داده BioLip [۲۶] با شرایط خاص اتخاذ می‌شود. شرایط خاص و عملیات آماده‌سازی داده‌ی ورودی در بخش مجموعه داده (بخش ۴-۱) توضیح داده شده است. براساس ماهیت مدل پیشنهادی، داده‌های ورودی دارای برچسب دودویی می‌باشند. سپس در مرحله‌ی مهندسی ویژگی، مراحل مختلفی مانند: استخراج ویژگی‌های پایه با ابزارهای بیوانفورماتیکی، بهبود کیفیت داده، تکنیک‌های نرمال‌سازی، تقویت داده[۲۰] و آماده سازی آرایه دو بعدی به ترتیب اجرا شده‌اند. در باب تقویت داده، بایستی مطرح نمود که تکنیک نمونه برداری افزایشی با تولید نمونه‌های مصنوعی برای کلاس اقلیت، مرز تصمیم‌گیری مرتبط با نمونه‌های پیوندی را تقویت می‌کند. مادامیکه تکنیک نمونه برداری کاهشی با حذف کنترل شده بخشی از نمونه‌های دور کلاس اکثریت و حفظ نمونه‌های نزدیک به مرز، تمایز بین نمونه‌های پیوندی و غیر پیوندی را برجسته کرده و بهبود متریک اندازه‌گیری-اف و تعادل مابین حساسیت و خاصگی را توجیه می‌کند[۳۷, ۳۱,۳۲]. در ادامه، مرحله‌ی آموزش شامل یک خط لوله تعبیه شده است. خط لوله پیشنهادی شامل یک ماژول استخراج ویژگی‌های معنادار مبتنی بر یادگیری عمیق (DGRU) و یک ماژول طبقه‌بند مبتنی بر یادگیری ماشین (SVM) می‌باشد. لازم به ذکر است که هایپرپارامترها در مرحله‌ی آموزش، با تکنیک جستجوی شبکه‌ای[۲۱] تنظیم می‌شوند و مدل با استفاده




**شکل ۱. چارچوب مدل پیشنهادی GSPPIR**

از آن‌ها آموزش دیده است. در نهایت، در مرحله‌ی تشخیص برهمکنش پروتئین-پپتید، عملکرد مدل آموزش دیده با استفاده از مجموعه اعتبارسنجی[۲۲] و معیارهای استاندار ارزیابی و بررسی می‌شود.

خروجی مدل آموزش دیده با استفاده از دو مجموعه آموزشی استاندارد (TR1116,TR640)، کلاس باقیمانده اسید آمینه (پیوندی/غیرپیوندی) در برهمکنش پروتئینی است.

## ۳-۲- ویژگی‌های بیوانفورماتیکی

انواع ویژگی‌های پایه با به‌کارگیری ابزارهای بیوانفورماتیک [۹, ۲۷-۲۹] از ساختارهای اول پروتئینی، استخراج شده‌اند. برحسب آزمایشات، ویژگی‌های پایه مورد استفاده برای تشخیص برهمکنش پروتئینی در جدول ۱ لیست شده‌اند.

در نهایت، یک بردار ویژگی پایه شامل ویژگی‌های تکاملی، مبتنی بر ساختار، مبتنی بر توالی و فیزیکوشیمیایی ساخته می‌شود. لذا، بردار ویژگی‌های پایه (سطح پایین) مستخرج شده از ساختارهای اول پروتئینی، ازمعادله (۱) پیروی می‌کند:

$$\overrightarrow{\mathrm{BF}} = \{\mathrm{PSSM, Entropy\ and\ CNCC, ASA, SS, LBA, PP, AAO}\} \in \mathrm{R}^{1\times 44} \qquad (۱)$$

در ادامه، برای یکنواخت‌تر شدن گرادیان‌ها، تسریع فرآیند آموزش، به حداقل رساندن بیش برازش و پایداری مدل، مقادیر اکتسابی باید مابین بازه‌ی [۰,۱] مقیاس‌بندی شوند. بدین گونه، برای تنظیم مقادیر در بازه‌ی مذکور، معادله (۲) اجرا شده است.

$$\mathrm{x'} = \frac{(\mathrm{x} - \min(\mathrm{x}))}{\max(\mathrm{x}) - \min(\mathrm{x})} \qquad (۲)$$

در معادله‌ی (۲)، $x$ نشان‌دهنده‌ی ویژگی‌های پایه و $x'$ نشان‌دهنده‌ی

**جدول ۱. گروه‌های ویژگی استخراج‌شده از ساختارهای اول پروتئینی برای مدل پیشنهادی GSPPIR**

| تشریح | اختصار و نام ویژگی | ابزار استخراج‌کننده | ماهیت |
|---|---|---|---|
| • پیش بینی ساختارهای ثانویه مانند آلفا- هلیکس<br>• ارائه‌ی اطلاعات تکمیلی از آرایش فضایی اسیدهای آمینه<br>• بیانگر زوایای موضعی ستون فقرات پروتئین<br>• تسهیل در فرآیند مدلسازی خصیصه‌های هندسی ساختارهای کریستالی | SS: Secondary Structure | SPINE-X [28] | مبتنی بر توالی |
| • نمایش ویژگی‌های فیزیکی-شیمیایی اسیدهای آمینه<br>• مؤثر در تشخیص نواحی عملکردی | LBA: Local Backbone Angle | SPIDER3. 0 [30] | مبتنی بر ساختار |
| • اندازه گیری میزان در دسترس بودن اسیدهای آمینه برای حلال<br>• مؤثر در تشخیص نواحی فعال پروتئین‌ها | PP: Physicochemical | DisPredict2 [29] | تکاملی |
| • استخراج اطلاعات تکاملی- حفاظتی از ساختارهای اول<br>• نمایش احتمال وقوع اسیدهای آمینه در هر موقعیت از توالی | ASA: Accessible Surface Area | SPINE-X [28] | مبتنی بر توالی |
| • نشان دهنده میزان تنوع یا حفاظت تکاملی در هر موقعیت از ساختار های پروتئینی<br>• مدلسازی همبستگی میان موقعیت‌های مجاور توالی | PSSM: Position Specific Scoring Matrix | PSI-BLAST [25] | تکاملی |
| • بیانگر فراوانی نسبی هر اسید آمینه در توالی | Entropy, CNCC | PSI-BLAST [27] | تکاملی و مبتنی بر توالی |
| | AAO: Amino acid Occurrence | Ref. [9] | مبتنی بر توالی |

ویژگی‌های نرمال شده است. سپس تکنیک نمونه‌برداری[۲۳] [۳۱] SMOTE و تکنیک NearMiss [۳۲] برای ویژگی‌های مقیاس بندی شده، اعمال شده است.

### ۳-۳- تبدیل داده

در این مطالعه، برای ماژول مبتنی بر یادگیری عمیق، بایستی داده پروتئینی یک بعدی تبدیل به داده دو بعدی شود. لذا، تکنیک قطعه‌بندی مبتنی بر پنجره، بر روی توالی‌های پروتئینی اجرا شده است. این تکنیک براساس اندازه پنجره کشویی[۲۴] و خاصیت محلیت[۲۵]برای هر باقی‌مانده اسیدآمینه، است.

مزایای اعمال خاصیت محلیت همان شناسایی الگوهای محلی در توالی، متعادل‌سازی برهمکنش پروتئین-پپتید، بهینگی براساس اطلاعات باقیمانده‌های اسیدآمینه‌ای همسایه و بهبود عملکرد پیش‌بینی می‌باشند [۳۳]. بدین منظور، K تا از نزدیک‌ترین همسایه‌های اسیدآمینه مرکزی (مورد هدف) در هر دو طرف (چپ و راست) در نظر گرفته می‌شوند. لذا، پنجره‌ای به نام $W_i$ و با اندازه ۱ +۲k مشخص می‌شود. تمرکز پنجره‌ی مذکور برای تشخیص اسیدآمینه مرکزی ($R_i$) است که شامل یک بازه‌ی محلی است. رابطه‌ی (۳)، نمونه‌ای از یک اندازه پنجره با اندازه ۷ است:

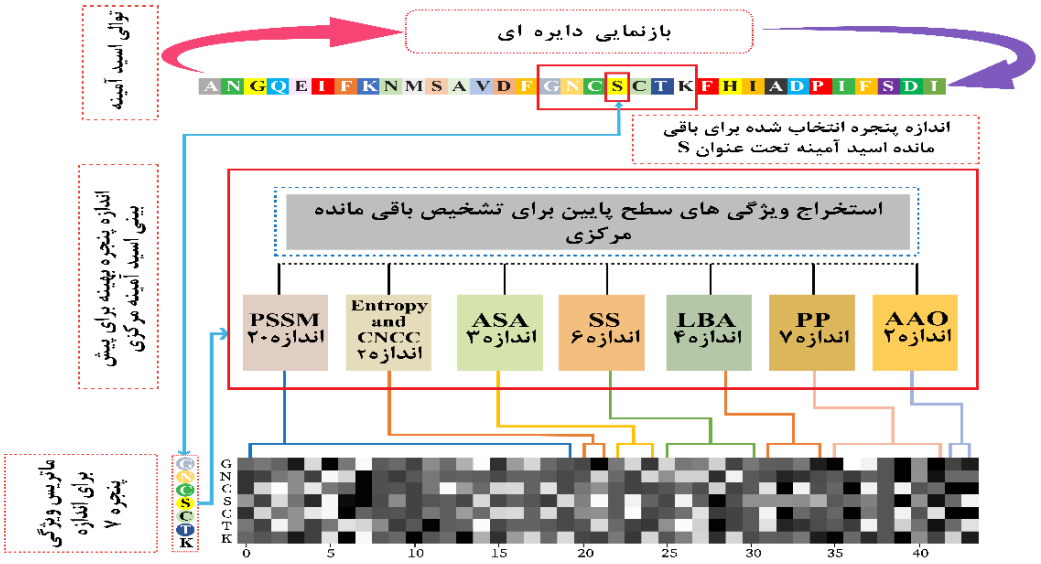
$$W_i = \{R_{i-3}, \dots, R_i, \dots, R_{i+3}\} \quad (۳)$$

در ادامه، ۴۴ نوع ویژگی متنوع (مبتنی بر ساختار، مبتنی بر توالی، تکاملی و فیزیکو شیمیایی) برای هر اسیدآمینه در این پنجره استخراج می‌شود. نتیجتاً یک آرایه ۴۴×(۲k+۱) ایجاد می‌شود که ۴۴ همان انباشتگی ویژگی‌ها به‌صورت افقی و در قالب بردار ویژگی است. ۷ بیانگر بازه‌ی محلی (۲k+۱) بهینه برای مدل پیشنهادی است. قابل‌توجه است که تکنیک دایره‌ای (حلقه‌ای) [۳۳] برای رسیدگی به باقیمانده‌های اسیدآمینه لبه‌ای در پیمایش توالی است که در ابتدا یا انتهای توالی پروتئینی می‌باشند. شکل ۲ بیانگر یک بازنمایی مبتنی بر آرایه‌ی دوبعدی (تصویر با مقیاس خاکستری[۲۶]) با اندازه پنجره ۷ است که برای تشخیص باقی‌مانده اسیدآمینه مرکزی سرین(S) طراحی شده است.

## ۴- ارزیابی مدل پیشنهادی

### ۴-۱- مجموعه داده

به‌منظور طراحی، پیاده‌سازی و ارزیابی تشخیص باقی‌مانده‌های اسیدآمینه در برهمکنش پروتئین-پپتید، دو مجموعه داده استاندارد، استفاده شده است. هر دو مجموعه داده از پایگاه داده [۲۶] BioLip[۲۷] جمع‌آوری شده‌اند. این پایگاه داده نوعی پایگاه داده نیمه دستی برای




**شکل ۲. بازنمایی آرایه دوبعدی با اندازه پنجره ۷ و ۴۴ ویژگی مستخرج شده**

انواع برهمکنش‌های لیگاند و پروتئین می‌باشد. اولین مجموعه داده استاندارد توسط آزمایشگاه Sparks [۳۵] در سال ۲۰۱۶ معرفی گردید. دومین مجموعه داده هم توسط آزمایشگاه Wei [۲۴] در سال ۲۰۲۲ ارائه شده است. شایان ذکر است برای مقایسه و ارزیابی منصفانه، شرط‌های معین برای پیش‌پردازش اجرا شده‌اند. شرایط اعمال‌شده برای پیش‌پردازش عبارتند از: الف) مجموعه داده اولیه (ساختارهای پروتئینی) از پایگاه داده BioLip [۲۶] گرفته شده است. ب) پپتید به‌عنوان یک زنجیره در نظر گرفته‌شده که طول آن کوتاه‌تر از ۳۰ باقی‌مانده اسیدآمینه است. پ) پروتئین‌های پیوندی به پپتید که طول کمتر از سه باقی‌مانده اسیدآمینه پیوندی دارند نیز حذف شده‌اند. ت) تعریف یک باقیمانده پیوندی برحسب فاصله اتمی ۳/۵ آنگستروم (Å) می‌باشد. ث) پروتئین‌هایی با بیش از ٪۳۰ شباهت حذف شده‌اند. عملیات حذف با استفاده از جعبه‌ابزار blastclust [۳۴] انجام شده است. ث)مجموعه داده اول دارای نسبت عدم تعادل ۱:۱۷ ما بین باقیمانده‌های اسیدآمینه پیوندی و غیر پیوندی می‌باشد. ج) مجموعه داده دوم دارای نسبت عدم تعادل ۱:۱۸ ما بین باقیمانده‌های اسیدآمینه پیوندی و غیر پیوندی می‌باشد. چ) تکنیک ارزیابی متقابل (K=10) صرفا برای تقسیم‌بندی مجموعه آموزش اتخاذ شده است. ح) هر یک از دو مجموعه داده استاندارد، توسط آزمایشگاه‌های Wei [۲۴] و Sparks [۳۵] به دودسته آموزش و آزمون با نسبت ۹۰ درصد و ۱۰ درصد (به ترتیب) تفکیک شده‌اند. خ) تکنیک جداسازی فولد مبتنی بر پروتئین[۳۵] برای اثبات عدم وجود پروتئین یکسان در دو مجموعه آموزش و آزمون و کاهش پدیده بیش آموزش[۲۸] استفاده شده است. در این مطالعه، همین شرایط عینا دنبال شده‌اند. همچنین مجموعه داده‌ها و کدهای مربوط به معماری پیشنهادی در لینک گیت‌هاب تعریف شده در دسترس می‌باشند.

https://github.com/shsh95/GSPPIR-at-Bioinformatics

با توجه به توضیحات، جزئیات دو مجموعه داده استاندارد برای آموزش و آزمون مدل پیشنهادی به تفکیک عبارتند از:

**جدول ۲. مشخصات عددی مجموعه داده استاندارد Sparks**

| نام | تعداد پروتئین | تعداد باقیمانده‌ها | تعداد باقیمانده‌های پیوندی | تعداد باقیمانده‌های غیر پیوندی |
|---|---|---|---|---|
| مجتمع پروتئین-پپتید[۲۹] | ۱۲۴۱ | ۲۹۷۵۹۸ | ۱۶۶۷۸ | ۲۸۰۹۲۰ |
| مجموعه آموزش-TR1116 | ۱۱۱۶ | ۲۶۶۵۹۹ | ۱۴۹۵۹ | ۲۵۱۷۶۹ |
| مجموعه آزمون مستقل[۳۰] TS125 | ۱۲۵ | ۳۰۸۷۰ | ۱۷۱۹ | ۲۹۱۵۱ |

**جدول ۳. مشخصات عددی مجموعه داده استاندارد Wei**

| نام | تعداد پروتئین | تعداد باقیمانده‌ها | تعداد باقیمانده‌های پیوندی | تعداد باقیمانده‌های غیر پیوندی |
|---|---|---|---|---|
| مجتمع پروتئین-پپتید | ۱۲۷۹ | ۳۰۷۶۹۲ | ۱۶۷۴۹ | ۲۹۰۹۴۳ |
| مجموعه آموزش – TR640 | ۶۴۰ | ۱۵۷۳۶۲ | ۸۲۵۹ | ۱۴۹۱۰۳ |
| مجموعه آزمون مستقل- TS639 | ۶۳۹ | ۱۵۰۳۳۰ | ۸۴۹۰ | ۱۴۱۸۴۰ |

## ۴-۲- معیارهای ارزیابی

برای مقایسه‌های منصفانه، باید مجموعه داده‌ها و همچنین شرایط و روند آزمایش‌های در مراحل آموزش و آزمون یکسان باشند. برای مقایسه عملکرد عادلانه مابین مدل پیشنهادی و سایر مدل‌های رقیب و همچنین نتایج قابل اطمینان، از معیارهای ارزیابی مبتنی بر ماتریس آشفتگی[۳۱]، استفاده شده است. اجزای ماتریس آشفتگی عبارتند از:

True Positive (TP): تعداد باقیمانده‌های پیوندی واقعی که به درستی شناسایی شده‌اند.

True Negative (TN): تعداد باقیمانده‌های غیر پیوندی که به درستی به صورت باقیمانده‌های غیر پیوندی، شناسایی شده‌اند.

False Positive (FP): تعداد باقیمانده‌های غیرپیوندی که به اشتباه به صورت باقیمانده‌های پیوندی دسته‌بندی شده‌اند.

False Negative (FN): تعداد باقیمانده‌های پیوندی که به اشتباه به صورت باقیمانده‌های غیر پیوندی دسته‌بندی شده‌اند.

بر اساس تعاریف برگرفته از ماتریس آشفتگی، معیارهای مختلف ارزیابی مدل عبارتند از:

$$\text{F} - \text{Measure (F} - \text{M)} = \frac{2 \times TP}{(2 \times TP + FP + FN)} \quad (۴)$$

$$\text{Specificity (SPE)} = \frac{\text{TN}}{(\text{FP} + \text{TN})} \quad (۵)$$

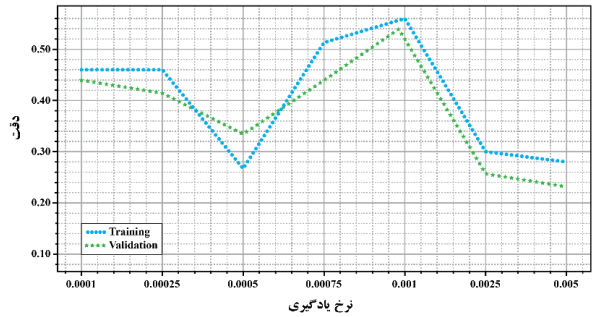
$$\text{Sensitivity (SEN)} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (۶)$$

$$\text{Precision (PRE)} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (۷)$$

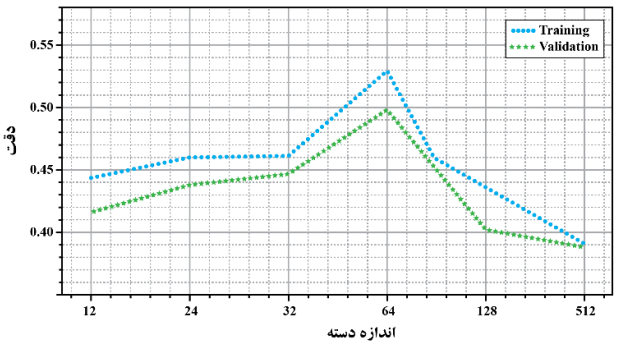
Area Under Curve (AUC or AUROC) همان مساحت زیر منحنی ROC است.

Receiver operating characteristics (ROC) نوعی منحنی است که عملکرد مدل را با در نظر گرفتن نرخ مثبت واقعی (TPR) و نرخ مثبت کاذب (FPR) در آستانه‌های مختلف پیش‌بینی نشان می‌دهد. تنظیمات مدل پیشنهادی براساس هایپرپارامترهای کاربردی عبارتند از:

در ادامه برای شفافیت بیشتر فرآیند بهینه‌سازی و تأثیر مستقیم تنظیمات بر عملکرد مدل پیشنهادی، تغییرات نرخ یادگیری و اندازه دسته (بچ) بر عملکرد مدل پیشنهادی در شکل‌های ۳ و ۴ آورده شده است.




**شکل ۳. تأثیر نرخ یادگیری برای فرآیندهای آموزش و ارزیابی**




**شکل ۴. تأثیر اندازه دسته برای فرآیندهای آموزش و ارزیابی**

بر طبق شکل ۳، مقادیر نرخ یادگیری مختلف از ۰/۰۰۰۱ تا ۰/۰۰۵ در جستجوی شبکه‌ای آزمایش شده است. نرخ یادگیری ۰/۰۰۱ بهترین عملکرد را با دقت ۵۳٪ روی مجموعه تعلیم نشان داده است. بر حسب شکل ۴، اندازه‌های دسته (بچ)های مختلف از ۱۲ تا ۵۱۲ از طریق جستجوی شبکه‌ای بررسی شده‌اند. اندازه دسته ۶۴ ، بهینه‌ترین تعادل مابین پایداری آموزش و سرعت همگرایی را فراهم کرد لازم به ذکر است، اختلاف اندک برای دو منحنی آموزش و ارزیابی، حاکی از تعمیم‌پذیری مناسب و تعادل مناسب پیچیدگی (عدم کم-برازش و عدم بیش-برازش) برای مدل پیشنهادی است.

در این مطالعه برای آموزش مدل‌های پیشنهادی و انجام آزمایشات

**جدول۴. هایپرپارامترهای کاربردی برای مدل GSPPIR**

| أبر پارامتر | نوع/ مقدار |
|---|---|
| تنظیم‌کننده | جستجوی شبکه‌ای (گرید) |
| لایه ورودی | Embedding (1024) |
| لایه پیچشی | Padding=same, activation=ReLU, stride=1,kernel=3 |
| تعداد لایه‌های پیچشی با فیلتر ۶۴ | ۲ |
| لایه نرمالساز دسته‌ای | ۲ |
| لایه ادغام با فیلتر ۱۲۸ | Step size=2, activation=ReLU, kernel=3 |
| تعداد لایه ادغام بیشینه | ۱ |
| لایه پیچشی | Padding=same, activation=ReLU, stride=1,kernel=3 |
| تعداد لایه پیچشی با فیلتر ۲۵۶ | ۲ |
| لایه نرمالساز دسته‌ای | ۲ |
| لایه ادغام با فیلتر ۵۱۲ | Step size=2, activation=ReLU, kernel=3 |
| تعداد لایه ادغام بیشینه | ۱ |
| لایه GRU | Unit=256 |
| تعداد لایه GRU | ۲ |
| لایه مسطح | تبدیل خروجی به بردار ۱ بعدی |
| تعداد لایه مسطح | ۱ |
| لایه کاملا متصل | Activation= Rectified Linear Unit (ReLU) |
| تعداد لایه کاملا متصل | ۲ |
| لایه حذف تصادفی | Dropout Rate = 0. 05 |
| تعداد لایه حذف تصادفی | ۲ |
| بهینه‌ساز | Adam |
| نرخ یادگیری | 0.001 |
| اندازه دسته | ۶۴ |
| تابع ضرر | Cross entropy |
| بسته | LibSM |
| تابع ارزیاب | ماشین بردار پشتیبان |
| تعداد دوره‌ها | ۱۰۰ |
| تابع هسته | $K(X_i, X_j) = \exp(-\gamma . D_{ij}^2)$ |
| گاما | 0.05 |
| تابع پیش‌بینی | $F(x) = \text{sign}\left(\sum_{i=1}^{n_{support}} \alpha_i y_i K(X, X_i) + b\right)$ |
| پارامتر منظم سازی: سی | ۱ |
| هسته | Radial Basis Function (RBF) |
| Γ | $\gamma = \frac{1}{n_{-features} . var\ (X)}$ |
| تنظیم حاشیه | 0.1 |
| مکانیزم انتخاب جفت‌ها در SMO | Boundary pair selection |

از یک سیستم با واحد پردازشگر اصلی مدل - Intel i7 10750H، ۳۲ گیگابایت حافظه اصلی و واحد پردازشگر گرافیکی مدلTi۲۰۶۰ GPU NVIDIA GeForce GTX استفاده شده است.

## ۴-۳- ارزیابی استحکام نتایج مدل پیشنهادی

این مرحله شامل اثبات استحکام نتایج اکتسابی و تعمیم‌پذیری مدل پیشنهادی ((GSPPIR (Ours by NearMiss می‌باشد. لذا، مقایسه‌هایی مبتنی بر تکنیک اعتبارسنجی متقابل[۳۲] و مجموعه آزمون مستقل انجام شده است. این مقایسه صرفا برای حالت بهینه یعنی NearMiss گزارش شده است.

جدول ۵ بیانگر استحکام روش پیشنهادی از طریق اختلاف اندک برای عملکردهای مبتنی بر مجموعه آزمون مستقل و اعتبارسنجی متقابل است. شایان ذکر است که برای ثبات و مدیریت پراکندگی عملکرد پیش‌بینی، نتایج برای اعتبارسنجی متقابل برحسب میانگین

و انحراف معیار[۳۳] گزارش شده‌اند. همچنین برای بررسی اهمیت و سهم ویژگی‌ها، عملکرد مدل پیشنهادی با استفاده از به کارگیری تکنیک‌های متنوع و مبتنی بر ویژگی همچون[۲۱,۲۰], Wrapper Feature Selection,Wrapper evolutionary, Integrated Gradient,SHAP, LIME,Feature importance Feature importance، بررسی و ارزیابی شده است. با توجه به اهداف این مطالعه (تعادل مابین نمونه‌های پیوندی و غیرپیوندی، مدیریت بیش-برازش و مدیریت سوگیری به سمت کلاس اکثریت)، مدل پیشنهادی مبتنی بر تکنیک Integrated Gradient توانسته بهترین اندازه‌گیری-اف، دقت پیش‌بینی بهینه و بهترین تعادل مابین خاصگی و حساسیت را فراهم نماید.

نتایج اکتسابی جدول ۶، اهمیت و کارایی گروه‌های ویژگی منتخب را برای دستیابی به اهداف مذکور، بیش از پیش اثبات می‌نماید.عملکرد مدل پیشنهادی GSPPIRبا تکنیک‌های متنوع و مبتنی بر ویژگی همچون ,Wrapper Feature Selection,Wrapper evolutionary, Integrated Gradient,SHAP, LIME,Feature importance Feature importance,، ارزیابی شده است. بهترین مقدار مربوط به معیارهای سطح زیر نمودار و خاصگی و با استفاده از تکنیک Feature importance گزارش شده است و مقادیر روش پیشنهادی با مقادیر بهینه مذکور نیز قابل مقایسه می‌باشند. با توجه به اهداف این مطالعه (تعادل مابین نمونه‌های پیوندی و غیرپیوندی ، مدیریت بیش‌برازش و مدیریت سوگیری به سمت کلاس اکثریت)، مدل پیشنهادی با استفاده از تکنیک Integrated Gradient توانسته بهترین اندازه‌گیری-اف، دقت پیش‌بینی بهینه و بهترین تعادل مابین خاصگی و حساسیت را فراهم نماید.

### ۴-۴- تأثیر استفاده از مدل‌های مختلف برای ساخت مدل نهایی

در این مطالعه، در شکل ۵ انواع ارزیابی‌های تجربی با استفاده از انواع اندازه پنجره‌های مختلف شامل:۹,۱۱، ۷، ۵، ۳، نشان داده شده است. به گونه‌ای که انواع لیگاندهای متنوع (پپتید، دئوکسی‌ریبونوکلئیک اسید، ریبونوکلئیک اسید، کربوهیدرات) اعمال شده‌اند. بر حسب آزمایش‌های انجام شده، یک جستجوی شبکه‌ای برای تعیین تعداد بهینه باقیمانده‌ها در اطراف یک باقیمانده هدف (اندازه پنجره) اعمال کرده‌ایم. تکنیک مذکور می‌تواند برهمکنش بین پروتئین و پپتید را تعدیل کرده و یک پیش‌بینی بهینه ایجاد نماید. همچنین اندازه پنجره مناسب، مزایایی مانند: تعادل مطلوب‌تری بین ثبت زمینه محلی کافی، اجتناب از نویز و مدیریت بیش‌برازش را فراهم می‌کند.

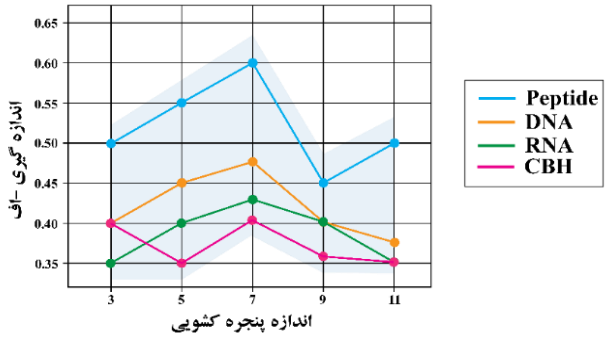
**جدول ۵ . مقایسه عملکرد مدل پیشنهادی بر روی مجموعه آزمون مستقل و اعتبارسنجی متقابل برای مجموعه داده استاندارد** Sparks

| مجموعه داده | تکنیک نمونه برداری | دقت (٪) | خاصگی (٪) | حساسیت (٪) | اندازه‌گیری-اف (٪) | سطح زیر نمودار(٪) |
|---|---|---|---|---|---|---|
| مجموعه آزمون مستقل | Near Miss | ۵۵/۷۰ | ۷۰/۱۰ | ۶۸/۰۰ | ۶۱/۲۰ | ۸۱/۱۱ |
| اعتبارسنجی متقابل | Near Miss | ۵۳/۴۳ | ۶۹/۱۱ | ۶۷/۵۷ | ۵۹/۶۷ | ۷۸/۰۷ |
| | | ±۰/۰۱ | ±۰/۰۲ | ±۰/۱۱ | ±۰/۴۱ | ±۱/۰۱ |

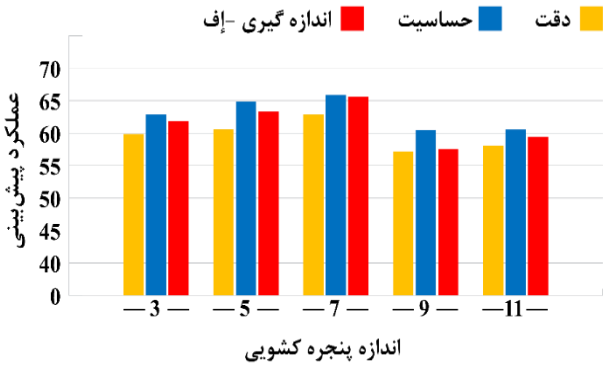
**جدول ۶. مقایسه عملکرد مدل پیشنهادی با استفاده از مجموعه آموزشی** TR1116 **و با تکنیک‌های مبتنی بر ویژگی**

| تکنیک مبتنی بر ویژگی | دقت (٪) | خاصگی (٪) | حساسیت (٪) | اندازه‌گیری-اف (٪) | سطح زیر نمودار(٪) |
|---|---|---|---|---|---|
| Wrapper Feature Selection | ۴۶/۲۶ | ۶۵/۰۷ | ۶۳/۹۱ | ۵۳/۶۷ | ۷۳/۸۹ |
| | ±۰/۰۷ | ±۰/۰۵ | ±۰/۰۲ | ±۰/۰۱ | ±۰/۰۷ |
| Feature importance | ۴۹/۲۶ | ۶۷/۰۰ | ۶۶/۱۱ | ۵۶/۴۰ | ۷۵/۱۹ |
| | ±۰/۰۵ | ±۰/۰۷ | ±۰/۰۳ | ±۰/۰۱ | ±۰/۰۲ |
| Wrapper evolutionary | ۴۹/۰۶ | ۶۶/۱۱ | ۶۵/۱۰ | ۵۵/۹۰ | ۷۵/۰۷ |
| | ±۰/۰۳ | ±۰/۰۲ | ±۰/۰۷ | ±۰/۰۱ | ±۰/۰۵ |
| Integrated Gradient | ۵۱/۰۸ | ۶۸/۲۱ | ۶۷/۱۷ | ۵۸/۴۷ | ۷۶/۰۷ |
| | ±۰/۰۱ | ±۰/۰۲ | ±۰/۰۷ | ±۰/۰۶ | ±۰/۴۰ |
| Integrated Gradient | ۵۳/۴۳ | ۶۹/۱۱ | ۶۷/۵۷ | ۵۹/۶۷ | ۷۸/۰۷ |
| | ±۰/۰۱ | ±۰/۰۲ | ±۰/۱۱ | ±۰/۴۱ | ±۱/۰۱ |
| SHAP | ۴۳/۵۵ | ۶۳/۴۴ | ۶۱/۶۶ | ۴۸/۹۵ | ۷۱/۱۱ |
| | ±۰/۰۹ | ±۰/۰۳ | ±۰/۰۲ | ±۰/۰۱ | ±۰/۰۵ |
| LIME | ۴۴/۲۲ | ۶۶/۴۳ | ۵۹/۰۰ | ۵۰/۰۵ | ۷۱/۵۷ |
| | ±۰/۰۲ | ±۰/۰۹ | ±۰/۰۷ | ±۰/۷۱ | ±۰/۰۳ |
| Feature importance | ۵۱/۳۸ | ۷۰/۲۱ | ۶۶/۶۱ | ۵۷/۰۹ | ۷۹/۸۸ |
| | ±۱/۰۰ | ±۰/۰۲ | ±۰/۰۱ | ±۰/۲۱ | ±۰/۰۱ |

لذا، اندازه پنجره‌ای که مدل پیشنهادی به ازای آن دارای عملکرد پیش بینی بهینه است، به عنوان بهترین اندازه پنجره انتخاب شده است. نتایج حاکی از عملکرد پیش‌بینی بهینه با استفاده از لیگاند پپتید و اندازه پنجره ۷ توسط مدل پیشنهادی است. با توجه به توضیحات، شکل ۶ حاوی تحلیل عملکرد مدل پیشنهادی براساس اندازه پنجره کشویی است.




**شکل۵. مقایسه عملکردهای پیش بینی با لیگاندها و اندازه پنجره‌های مختلف**




**شکل۶. أثر عملکردهای پیش بینی با اندازه پنجره‌های مختلف**

بر حسب آزمایشات، با اندازه‌های پنجره کوچکتر از ۷، میدان پذیرنده اطراف هر باقیمانده مرکزی بسیار محدود می‌شود. بنابراین مدل پیشنهادی نمی‌تواند زمینه ساختاری و فیزیکوشیمیایی محلی مناسب و کافی (مانند محیط زنجیره جانبی، ساختار ثانویه مجاور)را که بر برهمکنش پپتید-پروتئین تأثیر می‌گذارد، ثبت نماید. در مقابل، اندازه‌های پنجره بزرگتر از ۷ می‌توانند یک ناحیه جانبی (همسایگی) وسیع را معرفی می‌کنند که اغلب به طور مستقیم به برهمکنش مورد هدف تعلق ندارند. این أمر موجب می‌شود که اطلاعات نویزی اضافه شوند و تعداد پارامترها هم افزایش یابند. همچنین اندازه‌های پنجره بزرگتر از ۷، موجب می‌شود که خطر بیش‌برازش به ویژه تحت محدودیت عدم تعادل کلاس، افزایش یابد. در نتیجه، اندازه پنجره ۷ نسبت به سایر اندازه‌های کاندید، می‌تواند تعادل مطلوب‌تری بین ثبت زمینه محلی کافی، اجتناب از نویز و مدیریت بیش‌برازش فراهم نماید. در همین راستا، عملکرد پیش‌بینی بهینه برای مدل پیشنهادی بر حسب سه معیار ارزیابی بهینه در شکل زیر آورده شده است.

در ادامه، برای یافتن بهترین معماری برای مدل پیشنهادی، ترکیبات مختلف مبتنی بر یادگیری عمیق و یادگیری ماشین ارزیابی شده‌اند. این امر استفاده از الگورتیم یادگیری عمیق پیشنهادی (DGRU) و الگورتیم یادگیری ماشین پیشنهادی (Improved SVM) را به وضوح اثبات می‌کند. مقایسه‌ها مبتنی بر ترکیبات مختلف یادگیری در جدول ۷ با استفاده از مجموعه آزمون مستقل (TS125) گزارش شده است. بر حسب مطالعات و آزمایشات پیشین ، مجموعه ای از الگورتیم‌های متداول یادگیری ماشین و یادگیری عمیق به صورت تجربی برای مورد مطالعاتی بیوانفورماتیک مورد ارزیابی قرار گرفتند. مدل‌ها تحت شرایط یکسان و با استفاده از مجموعه داده یکسان آموزش و ارزیابی شدند تا مقایسه‌ای منصفانه حاصل شود. عملکرد هر یک از مدل‌های مبتنی بر یادگیری، براساس معیارهای ارزیابی محاسبه و نتایج کمی حاصل در جدول ۷ ارائه شده است. براساس این نتایج، الگورتیم یادگیری ماشین (Improved SVM) و یادگیری عمیق (DGRU) دارای عملکرد بهینه بر حسب معیارهای ارزیابی می‌باشند و به عنوان مدل نهایی انتخاب شده‌اند.

لازم به ذکر است، مزایای ماژول‌های طراحی شده برای معماری پیشنهادی، توانسته کارایی و أثربخشی مدل پیشنهادی GSPPIR را برای پیش‌بینی باقیمانده‌های پیوندی در برهمکنش‌های پروتئینی ارضا نماید. این مزایا برای DGRU، مواردی شامل: استخراج خودکار ویژگی‌های معنادار، ضبط الگوها و وابستگی‌ها طولانی در داده‌های متوالی و ترتیب دار، کاهش منابع محاسباتی و متعاقبا کاهش زمان آموزش کمتر، مدیریت بیش‌برازش با لایه حذف تصادفی، مدیریت گرادیان محوشدگی با گیت‌های خاص (آپدیت و ریست) می‌باشند. مزایای مذکورتوانسته‌اند، أهمیت ماژول استخراج کننده ویژگی‌های معنادار مبتنی بر یادگیری عمیق (DGRU) را دو چندان توجیه کنند. در پژوهش حاضر، استفاده از ماژول طبقه‌بند مبتنی بر یادگیری ماشین (SVM) بهبود یافته به عنوان طبقه‌بند دوکلاسه توانسته پیچیدگی مدل پیشنهادی را کاهش دهد. سایر مزایای طبقه‌بند ماشین بردار پشتیبان بهبود یافته، توانسته کارایی طبقه‌بند مبتنی بر یادگیری ماشین را بیش از پیش توجیه می‌کند. این مزایا، مواردی همچون: پشتیبانی از هسته‌های متنوع برای تبدیل فضای ویژگی به یک فضای غیرخطی و دسته‌بندی داده‌های پیچیده زیستی، مرزبندی بهینه بین کلاس‌ها با أبرصفحه و کاهش خطای تعمیم، بهبود تفسیرپذیری، پایداری و مدیریت نویز داده‌های زیستی می‌باشند.

**جدول ۷: مقایسه عملکرد مدل‌های مختلف یادگیری در مرحله آزمون**

| روش | سطح زیر نمودار | اندازه گیری-اف | دقت | خاصگی | حساسیت |
|---|---|---|---|---|---|
| DRNN+KNN (ours by NearMiss) | ۷۴/۵ | ۵۰/۴ | ۵۰/۱ | ۵۲/۵ | ۵۰/۹ |
| DRNN+KNN (ours by SMOT E) | ۷۴/۲ | ۵۰/۰ | ۵۰/۱ | ۵۱/۰ | ۵۰/۲ |
| DRNN+RF (ours by NearMiss) | ۷۵/۷ | ۵۲/۰ | ۵۰/۲ | ۵۲/۷ | ۵۲/۰ |
| DRNN+RF (ours by SMOTE) | ۷۵/۱ | ۵۱/۶ | ۵۰/۱ | ۵۱/۴ | ۵۱/۲ |
| DRNN+SVM (ours by NearMiss) | ۷۶/۳ | ۵۴/۱ | ۵۱/۹ | ۵۳/۱ | ۵۲/۳ |
| DRNN+SVM (ours by SMOTE) | ۷۶/۱ | ۵۳/۵ | ۵۱/۰ | ۵۳/۰ | ۵۱/۱ |
| DLSTM+KNN (ours by NearMiss) | ۷۶/۵ | ۵۲/۶ | ۵۲/۰ | ۵۴/۲ | ۵۳/۴ |
| DLSTM+KNN (ours by SMOT E) | ۷۶/۱ | ۵۲/۰ | ۵۲/۰ | ۵۴/۰ | ۵۲/۲ |
| DLSTM+RF (ours by NearMiss) | ۷۷/۰ | ۵۳/۱ | ۵۲/۲ | ۵۴/۹ | ۵۳/۹ |
| DLSTM+RF (ours by SMOTE) | ۷۶/۹ | ۵۳/۰ | ۵۲/۱ | ۵۴/۷ | ۵۴/۱ |
| DLSTM+SVM (ours by NearMiss) | ۷۷/۲ | ۵۳/۸ | ۵۲/۹ | ۵۵/۰ | ۵۴/۸ |
| DLSTM+SVM (ours by SMOTE) | ۷۷/۱ | ۵۳/۶ | ۵۲/۷ | ۵۴/۹ | ۵۴/۷ |
| DCNN+KNN (ours by NearMiss) | ۷۷/۷ | ۵۲/۱ | ۵۳/۱ | ۵۵/۵ | ۵۳/۳ |
| DCNN+KNN (ours by SMOTE) | ۷۸/۱ | ۵۲/۷ | ۵۳/۰ | ۵۷/۳ | ۵۳/۸ |
| DCNN+RF (ours by NearMiss) | ۷۸/۸ | ۵۳/۱ | ۵۳/۳ | ۵۷/۷ | ۵۴/۳ |
| DCNN+RF (ours by SMOTE) | ۷۹/۰ | ۵۴/۱ | ۵۳/۲ | ۵۹/۱ | ۵۵/۵ |
| DCNN+SVM (ours by NearMiss) | ۷۹/۰ | ۵۴/۵ | ۵۳/۹ | ۵۹/۳ | ۵۶/۱ |
| DCNN+SVM (ours by SMOTE) | ۷۹/۵ | ۵۵/۳ | ۵۳/۳ | ۵۹/۵ | ۵۷/۶ |
| DGRU+KNN (ours by NearMiss) | ۷۹/۹ | ۵۶/۵ | ۵۴/۱ | ۶۲/۲ | ۶۰/۲ |
| DGRU+KNN (ours by SMOTE) | ۷۹/۴ | ۵۶/۲ | ۵۳/۹ | ۶۲/۰ | ۶۰/۱ |
| DGRU+RF (ours by NearMiss) | ۸۰/۰ | ۵۸/۷ | ۵۴/۳ | ۶۶/۹ | ۶۴/۰ |
| DGRU+RF (ours by SMOTE) | ۷۹/۳ | ۵۸/۵ | ۵۴/۰ | ۶۶/۶ | ۶۳/۹ |
| DGRU+SVM (ours by NearMiss) | **۸۱/۱** | **۶۱/۲** | **۵۵/۷** | **۷۰/۱** | **۶۸/۰** |
| DGRU+SVM (ours by SMOTE) | ۸۰/۰ | ۶۰/۱ | ۵۴/۷ | ۶۸/۸ | ۶۷/۱ |

برحسب آزمایش‌ها برای طراحی مدل نهایی، از سه مدل یادگیری ماشین شامل SVM، RF[۳۴]، KNN[۳۵]، و چهار مدل یادگیری عمیق شامل: DRNN[۳۶], DGRU ,DCNN[۳۷], DLSTM[۳۸] و اندازه پنجره‌های متنوع (۱۱,۹,۷,۵,۳) استفاده شده است. در مقایسه با سایر معماری‌های رقیب، بهترین نتایج اکتسابی مربوط به معماری مبتنی بر ترکیب DGRU و SVM در تمام معیارهای ارزیابی می‌باشد. این معیارهای مقایسه شده شامل معیارهای دقت[۳۹] (۵۵/۷ درصد)، خاصگی[۴۰] (۷۰/۱۰ درصد)، حساسیت[۴۱] (۶۸/۰۰ درصد)، اندازه‌گیری-اف[۴۲] (۶۱/۲۰ درصد) و سطح زیر منحنی (۸۱/۱۰ درصد) با استفاده از مجموعه آزمون مستقل TS125 در مرحله آزمون می‌باشند. بهینگی‌های معیارهای دقت(حداقل ۱ درصد)، خاصگی (حداقل ۱/۳ درصد)، حساسیت (تقریباً ۱ درصد)، اندازه‌گیری-اف (حداقل ۱/۱ درصد) و سطح زیر منحنی (حداقل ۱/۱ درصد) با اندازه هفت برای مدل نهایی تحت عنوانGSPPIR، گزارش شده است.

## ۴-۵- مقایسه با مدل‌های رقیب

در این مرحله، آنالیز و ارزیابی نتایج مابین مدل پیشنهادی GSPPIR و سایر مدل‌های رقیب با استفاده از مجموعه آزمون مستقل (TS125) انجام شده است. برای ارزیابی منصفانه، صرفاً مدل‌های یادگیری، مبتنی بر ساختار اول پروتئینی و با استفاده از مجموعه داده یکسان، ارزیابی و بررسی شده‌اند. با توجه به شرایط مذکور، نتایج مقایسه مابین مدل پیشنهادی و دیگر مدل‌های رقیب در جدول ۸ گزارش شده است.

جدول ۸ بیانگر مقادیر عملکرد پیش‌بینی مدل پیشنهادی GSPPIR (ours by NearMiss) با استفاده از مجموعه آزمون مستقل TS125 برای معیارهای ارزیابی دقت (۵۵/۷درصد)، خاصگی (۷۰/۱۰درصد)، حساسیت (۶۸/۰درصد)، اندازه‌گیری-اف (۶۱/۲۰درصد) و سطح زیر منحنی (۸۱/۱۰ درصد) است. برحسب ارزیابی‌های همه جانبه، روش پیشنهادی می‌تواند برچسب کلاس پیش‌بینی‌شده را برای هر

**جدول ۸. مقایسه نتایج مدل پیشنهادیGSPPIR با سایر مدل‌های رقیب با مجموعه آزمون مستقل TS125**

| مدل | دقت (٪) | خاصگی (٪) | حساسیت (٪) | اندازه‌گیری-اف (٪) | سطح زیر نمودار (٪) |
|---|---|---|---|---|---|
| Pepsite [14] | ۲۷/۹۰* | ۹۷/۰۰ | ۱۸/۰۰ | ۲۱/۹۰ | ۶۱/۰۰ |
| Peptimap [15] | ۲۷/۱۰* | ۹۵/۰۰ | ۳۲/۰۰ | ۲۹/۴۰ | ۶۳/۰۰ |
| SPRINT-Seq [19] | ۲۳/۳۰* | ۹۶/۰۰ | ۲۱/۰۰ | ۲۲/۱۰ | ۶۸/۰۰ |
| PepBind [11] | ۴۶/۹۰ | – | ۳۴/۴۰ | ۳۹/۶۰* | ۷۹/۳۰ |
| Visual [22] | – | ۶۸/۰۰ | ۶۷/۰۰ | – | ۷۳/۰۰ |
| GP+SVM [20] | ۵۱/۵۰* | ۹۴/۰۰ | ۲۷/۹۰ | ۳۶/۲۰ | ۶۹/۹۰ |
| PepNN-Seq [23] | – | – | – | – | ۷۹/۴۰ |
| SPPPred [21] | ۵۲/۹۰* | ۹۵/۹۰ | ۳۱/۵۰ | ۳۹/۵۰ | ۷۱/۰۰ |
| PepBCL [24] | ۵۴/۰۰ | ۹۸/۴۰ | ۳۱/۵۰ | ۳۹/۷۰* | ۸۱/۵۰ |
| PepCNN [25] | ۵۵/۰۰ | ۹۸/۸۰ | ۲۵/۴۰ | ۳۴/۷۰* | ۸۴/۳۰ |
| PepPFN[38] | – | **۹۹/۲۰** | ۱۹/۵ | – | – |
| E2EPep[39] | ۵۲/۵ | – | ۳۸/۱ | – | **۸۵/۱** |
| DPPBP[40] | ۵۱/۳ | ۹۸/۰ | ۳۶/۳ | – | ۸۲/۱۰ |
| GSPPIR (ours by NearMiss) | **۵۵/۷۰** | ۷۰/۱۰ | **۶۸/۰۰** | **۶۱/۲۰** | ۸۱/۱۰ |

**نماد * بیانگر محاسبه توسط نگارنده و عدم گزارش توسط نویسندگان است.**

باقیمانده اسیدآمینه با بهترین دقت (تقریباً۱درصد)، حساسیت (حداقل۱درصد)، اندازه‌گیری-اف (حداقل۲۱/۵درصد) و تعادل (کمینه اختلاف) ما بین حساسیت و خاصگی (۲/۱درصد) در مقایسه با مدل‌های رقیب پیش‌بینی کند. به بیانی دیگر، نتایجی با استفاده از ارزیابی مبتنی بر مجموعه آزمون مستقل (TS125)، استنتاج می‌گردند که عبارتند از: الف) بهترین مقدار دقت توسط مدل پیشنهادی، گزارش شده است. لذا مزایایی با مقدار دقت بهینه تضمین می‌گردد. این مزایا عبارتند از:کمینه کردن مثبت کاذب، استحکام مدل پیشنهادی و افزایش اطمینان در پیش‌بینی‌ها. ب) بهترین اندازه‌گیری-اف مربوط به مدل پیشنهادی است که تعادل مابین معیارهای دقت و حساسیت را نشان می‌دهد که منجر به نتایج قابل‌اعتمادتر و آموزنده‌تر می‌شود. پ) با ایجاد تعادل مابین حساسیت (نرخ مثبت واقعی) و خاصگی (نرخ منفی واقعی)، به درک کامل‌تری از عملکرد مدل پیش‌بینی می‌رسیم و از معیارهای مغرضانه اجتناب می‌شود. این بدین معناست که ایجاد تعادل مابین حساسیت و خاصگی به هنگام مواجهه با مجموعه داده‌های نامتعادل، ضروری است. در ادامه برحسب ارزیابی جامع، معیارهای اکتسابی غیر بهینه هم توسط مدل توسعه یافته بررسی و ارزیابی‌شده‌اند. این معیارها عبارتند از: الف) بهترین مقدار خاصگی توسط مدل رقیب PepPFN [۳۸] گزارش شده است. اگرچه این مدل دارای مقدار خاصگی بهینه است. اما به سمت کلاس غیر پیوندی(اکثریت) سوگیری دارد. این سوگیری موجب خاصگی زیاد (۹۹/۲۰) و حساسیت کم (۱۹/۵) را فراهم می‌کند. ب) تیم E2EPep[39]، معتقدند که معیار سطح زیر منحنی قادر است اطمینان مناسبی را برای ارزیابی عملکرد با مجموعه داده نامتعادل فراهم آورد. لذا، مدل E2EPep[39] توانسته معیار سطح زیر منحنی بهینه (۸۵/۱) را نسبت به سایر مدل‌های رقیب گزارش دهد.

در ادامه، برای اثبات أثربخشی بیشتر مدل پیشنهادی برای پیش‌بینی برهمکنش پروتئینی، ارزیابی عملکرد پیش‌بینی با مجموعه داده استاندارد Wei انجام شده است.

جدول ۹، عملکرد پیش‌بینی برای مدل پیشنهادی ( GSPPIR (Ours by NearMiss) با استفاده از مجموعه آزمون مستقل T639 برحسب معیارهای ارزیابی را گزارش کرده است. این معیارهای ارزیابی شامل ۴۹/۰۰ درصد دقت، ۵۹/۴۳ درصد خاصگی، ۵۵/۲۲ درصد حساسیت، ۵۱/۹۰ درصد اندازه‌گیری-إف و ۸۱/۰۱ درصد سطح زیر منحنی می‌باشد. برحسب ارزیابی‌های انجام‌شده، روش پیشنهادی می‌تواند کلاس هر باقیمانده اسید آمینه را با بهترین دقت (حداقل ۱/۱درصد)، بهترین اندازه گیری-إف (حداقل ۱۴/۷درصد) و بهترین تعادل مابین معیارهای حساسیت و خاصگی (حداقل ۴/۱ درصد) در مقایسه با سایر مدل‌های رقیب، پیش‌بینی نماید. با ارزیابی انجام‌شده، مزایایی با استفاده از مجموعه آزمون مستقل TS639، نیز استنتاج می‌گردد. این مزایا عبارتند از: الف) مدل پیشنهادی به دقت بهینه، دست‌یافته است.

**جدول ۹. مقایسه نتایج مدلGSPPIR با سایر مدل‌های رقیب با مجموعه آزمون مستقل TS 639**

| مدل پیشنهادی | دقت (٪) | خاصگی (٪) | حساسیت (٪) | اندازه‌گیری-اف (٪) | سطح زیر نمودار (٪) |
|---|---|---|---|---|---|
| PepBind [11] | ۴۵/۰۰ | – | ۳۱/۷۰ | ۳۷/۱۹* | ۷۶/۷۰ |
| PepNN-Seq [19] | – | – | – | – | ۷۹/۵۰ |
| PepBCL [25] | ۴۷/۰۰ | ۹۸/۳۰ | ۲۵/۲۰ | ۳۲/۸۰* | ۸۰/۴۰ |
| PepCNN [25] | ۴۷/۹۰ | **۹۸/۶۰** | ۲۱/۷۰ | ۲۹/۸۶* | **۸۲/۶۰** |
| E2EPep[39] | ۴۳/۸۰ | – | ۳۳/۱۰ | ۳۷/۷۱* | – |
| DPPBP[40] | ۴۹/۳۰ | ۹۸/۳۰ | ۲۷/۵۰ | – | – |
| GSPPIR (ours by NearMiss) | **۴۹/۳۱** | ۵۹/۴۳ | **۵۵/۲۲** | **51/90** | ۸۱/۰۱ |

نماد * بیانگر محاسبه توسط نگارنده و عدم گزارش توسط نویسندگان است.

مقدار دقت بهینه بیانگر این است که مدل طبقه‌بند خطاهای مثبت کاذب کمتری را انجام می‌دهد. ب) بیشینه مقدار اندازه‌گیری-اف مربوط به مدل پیشنهادی است. با توجه به اینکه هر دو مجموعه داده‌ی استاندارد، نامتعادل هستند. معیار اندازه‌گیری- اف تضمین می‌کند که عملکرد طبقه‌بند نسبت به کلاس اکثریت (غیرپیوندی) تعصب ندارد. به بیانی دیگر، مدل پیشنهادی توانسته محدودیت بیش پیش‌بینی کلاس غیرپیوندی را نسبت به سایر مدل‌های رقیب، بهتر مدیریت نماید. پ) مقدار حساسیت توسط مدل پیشنهادی بهینه است. به بیان جزئی‌تر، در مجموعه داده‌هایی با نرخ عدم تعادل زیاد، حساسیت بالا کمک می‌کند تا مطمئن شویم که کلاس پیوندی به اندازه کافی در پیش‌بینی‌های مدل بررسی شده است. ت) با مدل پیشنهادی به بهترین تعادل مابین معیارهای حساسیت و خاصگی دست یافته‌ایم. در مجموعه داده‌های نامتعادل، کمینه اختلاف مابین حساسیت و خاصگی شامل معاوضه مابین شناسایی صحیح نمونه‌های مثبت و اجتناب از مثبت کاذب است. لذا، تمرکز برافزایش مثبت واقعی و کاهش مثبت کاذب بنا نهادیم. این تمرکز می‌تواند موجب بهینگی هزینه و کاهش وابستگی به نیروی انسانی گردد. در ادامه برحسب ارزیابی جامع، معیارهای اکتسابی غیربهینه توسط مدل پیشنهادی در مرحله آزمون با استفاده از مجموعه TS639 نیز گزارش شده‌اند: الف) بهترین مقدار خاصگی برای مدل [۲۴] PepCNN ارائه شده است. این بهینگی با توجه به مدل پیشنهادی (شبکه عصبی پیچشی بهبود یافته) و به‌کارگیری مجموعه داده نامتعادل کاملاً قابل توجیه است. علی‌رغم این بهینگی، مقدار حساسیت اکتسابی توسط مدل پیشنهادی [۲۴] PepCNN، کمینه است. با توجه به مقادیر حساسیت و خاصگی و اختلاف زیاد مابین دو معیار مذکور، مدیریت بیش پیش‌بینی کلاس غیر پیوندی توسط مدل [۲۴] PepCNN جای تأمل دارد. ب) مقدار بهینه برای سطح زیر منحنی توسط مدل [۲۴] PepCNN گزارش شده است. این یکی ازتمرکزهای بهینه توسط مدل [۲۴] PepCNN نیز است. در مقایسه‌ی نتایج، این نکته حائز اهمیت است که هیچ یک از مدل‌های توسعه‌یافته، دارای بهینگی برای تمام معیارهای ارزیابی نمی‌باشند. لذا در این مطالعه، معیارهای ارزیابی بهینه و غیر بهینه مورد بحث واقع شدند.

در نهایت، برای پشتیبانی از عملکرد روش پیشنهادی با استفاده از دو مجموعه داده استاندارد، أثربخشی آن در طی چند مرحله مورد بررسی قرار گرفته است. نخست، مطابق با پیشینه تحقیق [۴۱،۲۲]، اندازه گیری-اف به عنوان یک معیار بده‌بستان مابین حساسیت و دقت است. لذا، می‌تواند دیدگاه جامع‌تری نسبت به عملکرد مدل پیشنهادی ارائه دهد. دوم، معیار دقت به این مربوط می‌شود که باقیمانده‌های پیوندی واقعی چقدر مناسب و کارا به عنوان باقیمانده‌های پیوندی شناسایی می‌شوند. طبق نتایج اکتسابی، روش پیشنهادی توانسته به بهترین اندازه‌گیری-اف و دقت دست یابد. سوم، به دلیل مجموعه داده‌های نامتوازن و اهمیت پیش‌بینی کلاس غیرپیوندی و پیوندی، خاصگی انتخاب شده است. لذا، تعادل مابین خصیصه (نرخ منفی واقعی) و حساسیت (نرخ مثبت واقعی) معیار مناسبی برای اندازه‌گیری توانایی یک روش در پیش‌بینی منفی‌ها (غیرپیوندی‌ها) و مثبت‌های (پیوندی‌های) واقعی هر گروه می‌باشد. بر حسب ارزیابی، مدل‌های پیشنهادی رقیب E2EPep[39] و PepCNN [25] به ترتیب به بالاترین مقادیر (۸۵.۱ و ۸۲.۶۰) در ناحیه زیر منحنی، دست یافته‌اند. این بهبودها را می‌توان با به کارگیری داده‌های مبتنی بر ساختار و مدل زبان پروتئینی توجیه کرد. در حالیکه مقادیر ناحیه زیر منحنی به‌دست‌آمده توسط روش پیشنهادی با روش‌های رقیب (مبتنی بر توالی و مبتنی بر ساختار) قابل مقایسه می‌باشند. روش‌های PepPFN[38] و PepCNN [25] در دو مجموعه داده استاندارد دارای بیشینه مقدار خاصگی می‌باشند. مادامیکه هر دو روش دارای مقدار حساسیت پایینی می‌باشند. به گونه‌ای که اختلاف این دو متریک برای مجموعه داده های Sparks(حداقل ۷۷ درصد) و Wei( حداقل۷۵ درصد) زیاد می‌باشد. به بیانی دیگر، اگر چه این مدل‌ها در تشخیص نمونه‌های

غیر پیوندی کارا می‌باشند. اما در تشخیص و یادگیری کلاس‌های پیوندی ناتوان می‌باشند و بسیاری از نمونه های مثبت واقعی را از دست می‌دهند.در مقایسه‌ی نتایج، این نکته حائز اهمیت است که هیچ یک از مدل‌های توسعه‌یافته، دارای بهینگی برای تمام معیارهای ارزیابی نمی‌باشند. لذا در این مطالعه، معیارهای ارزیابی بهینه و غیر بهینه مورد بحث واقع شدند.

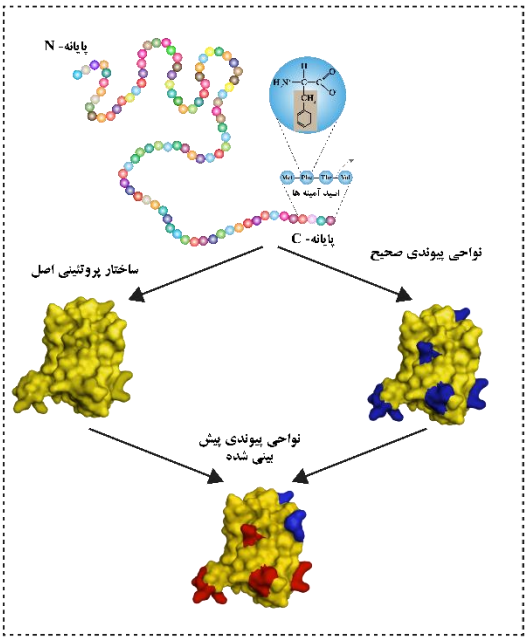
## ۵- ارزیابی بین رشته‌ای نتایج مدل پیشنهادی

ارزیابی مدل از هر دو دیدگاه زیست‌شناسی و انفورماتیکی تضمین می‌کند که، مدل پیشنهادی از نظر زیستی معنادار است و می‌تواند به عنوان مکمل مدل‌های تجربی پیشنهاد گردد. همچنین، مدل پیشنهادی از نظر محاسباتی کارآمد و دقیق است و می‌تواند در مقیاس‌های محاسباتی هم کاربرد داشته باشد.

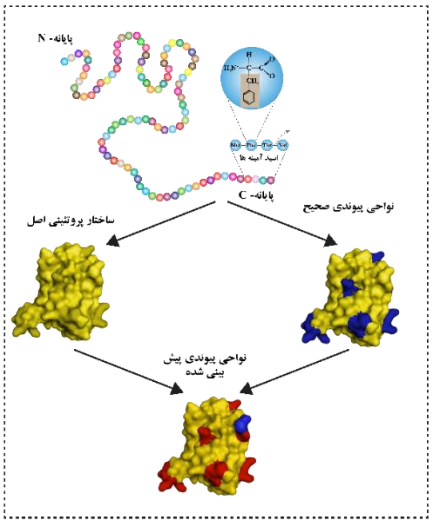
### ۵-۱- از نگاه زیست‌شناسی

شکل‌های ۸ و ۷، شامل آنالیز و مقایسه از دیدگاه انفورماتیک مابین مدل پیشنهادی با مدل ساختاری رقیب(SPRINT-Str [۳۵]) است. در این مرحله، برای اطمینان بیشتر از نتایج پیش‌بینی‌شده، به‌طور تصادفی یک پروتئین تحت عنوان PDBID: 2bugA را از مجموعه آزمون مستقل TS125 انتخاب کرده‌ایم. حالت‌های پیوندی واقعی و پیش‌بینی شده توسط مدل پیشنهادی و مدل SPRINT-Str [۳۵] به ترتیب در شکل‌های ۸ و ۷ نشان داده شده‌اند. رنگ آبی در دو شکل مذکور نشانگر باقی‌مانده‌های پیوندی واقعی و رنگ قرمز بیانگر باقی‌مانده‌های پیوندی پیش‌بینی‌شده، می‌باشند. لذا، از نظر تعداد جایگاه‌های پیوندی، مدل پیشنهادی توانسته، ۹ جایگاه پیوندی از ۱۱ جایگاه پیوندی را به‌درستی پیش‌بینی کند. مادامی‌که مدل رقیب (SPRINT-Str [۳۵]) دارای نرخ مثبت واقعی ۷ است. به‌عبارتی‌دیگر نرخ منفی کاذب برابر با ۲ و ۴ برای مدل پیشنهادی و مدل رقیب گزارش شده است (به ترتیب). این مقایسات مصور شده نیز اثبات می‌کنند که نتایج توسط مدل پیشنهادی بسیار نزدیک به نتایج توسط مدل آزمایشگاهی می‌باشند.

دامنه پروتئینی در زیست‌شناسی مولکولی، ناحیه‌ای از ساختار پروتئین است که به‌طور مستقل دارای فعالیت‌های خاص مانند تاشدگی، تغییر و تکامل و عملکرد زیستی منحصربه‌فرد است [۳۵]. برحسب مطالعه‌ی طاهرزاده و همکاران [۳۵]، دامنه‌های پروتئینی MHC, PDZ, SH2, SH3 [۳۵]، نیز از مجموعه داده استاندارد Sparks و از نوع پیوندی می‌باشند. همچنین این تیم، یک باقیمانده پیوندی را با فاصله اتمیک ۳/۵ آنگستروم (Å) تعریف کرده است[۳۵].




**شکل ۷. مقایسه عملکرد پیش‌بینی جایگاه باقیمانده‌های پیوندی توسط مدل تجربی و مدل ( SPRINT-Str [۳۵])**

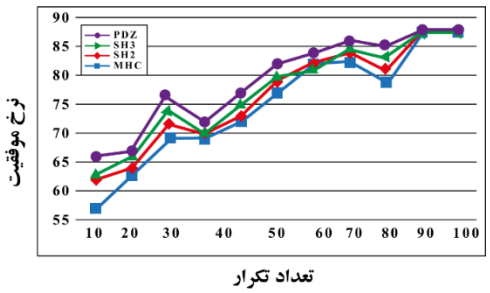



**شکل ۸. مقایسه عملکرد پیش‌بینی جایگاه باقیمانده‌های پیوندی توسط مدل تجربی و مدل پیشنهادیGSPPIR**

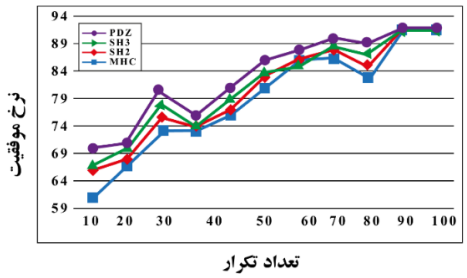
از آنجایی‌که مدل پیشنهادی مبتنی بر توالی است. لذا، با در نظر گرفتن دامنه‌های پروتئینی، قادر هستیم که به ارزیابی منصفانه و مبتنی بر ساختارهای پروتئینی شناخته شده مابین مدل پیشنهادی(مبتنی بر توالی) و مدل رقیب (مبتنی بر ساختار) دست یابیم. با توجه به توضیحات فوق، نرخ موفقیت از معادله (۸) تبعیت می‌کند:

$$\text{Success rate} = \frac{\text{Number of regions having DCC} \leq 3.5\ \text{Å}}{\text{Total number of amino acid residues}} \quad (۸)$$

بر حسب معادله‌ی (۸)، مفهوم DCC[۴۳] به فاصله‌ی مابین مراکز برای دو دامنه‌ی پروتئینی اشاره دارد . اگر این فاصله کمتر یا مساوی ۳/۵ آنگستروم (Å) باشد، باقی‌مانده پیش‌بینی شده به عنوان باقی‌مانده صحیح پیش‌بینی شده در نظر گرفته می‌شود. در ارزیابی عملکرد وابسته به دامنه، بیشینه نرخ موفقیت مربوط دامنه PDZ است. در ادامه، برحسب شکل‌های ۹ و ۱۰، برتری عملکرد پیش‌بینی توسط مدل پیشنهادی با استفاده از دامنه‌های تعیین‌شده در مقایسه با عملکرد پیش‌بینی مدل SPRINT-Seq[۱۹] توجیه می‌گردد.




**شکل ۹. ارزیابی نرخ موفقیت مدل SPRINT-Seq[۱۹] در مرحله آزمون**




**شکل۱۰. ارزیابی نرخ موفقیت مدل پیشنهادی GSPPIR در مرحله آزمون**

أثربخشی مدل پیشنهادی با معیار فاکتور غنی‌سازی[۴۴] در برابر امتیاز احتمالاتی ناحیه پیوندی [۳۵]، برای برهمکنش پروتئین-پپتید اثبات شده است. فاکتور غنی‌سازی برحسب معادله(۹) عبارت است از:

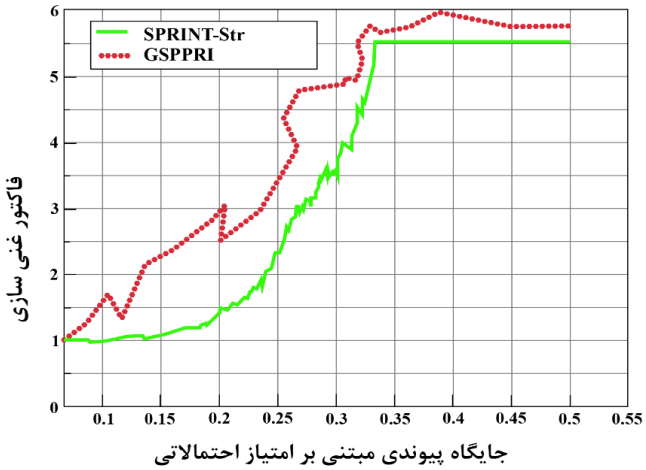
$$EF_i = \frac{\frac{B_i}{\sum B_i}}{\frac{N_i}{\sum N_i}} \qquad (۹)$$

در معادله‌ی (۹)، $B_i$ تعداد پروتئین‌هایی را نشان می‌دهد که نواحی پیوندی آن‌ها از یک آستانه[۴۵] خاص فراتر رفته است و مبتنی بر امتیاز قابلیت اطمینان است. $N_i$ نشان‌دهنده‌ی تعداد کل پروتئین‌ها با یک پیوند معین i است. امتیاز احتمالاتی ناحیه پیوندی بیانگر این است که یک ناحیه مشخص در مولکول با چه احتمالی، ناحیه پیوند است. در همین راستا و برحسب شکل ۱۱، فاکتور غنی‌سازی بیشتر نشان‌دهنده ارتباط قوی‌تر مابین معیارهای امتیاز قابلیت اطمینان و مناطق پیوندی پیش‌بینی‌شده است. به بیان جزئی‌تر، اگر فاکتور غنی‌سازی بیشتر از آستانه ۰/۴ باشد، آنگاه احتمال بیشتری برای شناسایی جایگاه‌های پیوندی بر روی ساختار کریستالی وجود دارد. اگر فاکتور غنی‌سازی کمتر از آستانه ۰/۲ باشد، آنگاه احتمال یافتن جایگاه‌های پیوندی هم کاهش می‌یابد. بر حسب شکل ۱۱، منحنی مدل پیشنهادی (GSPPIR) به‌طور مداوم بالاتر از منحنی روش SPRINT-Str [۳۵] است. بدین معناکه، مدل پیشنهادی نسبت به مدل رقیب، در شناسایی نواحی پیوند با احتمال بالا، موفق‌تر است.

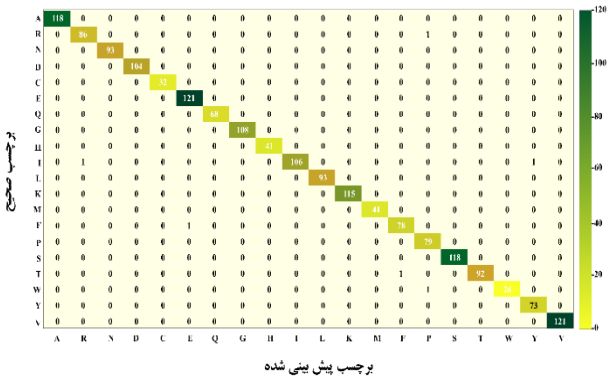
## ۵-۲- از نگاه انفورماتیکی

در این مرحله، مقایسه و ارزیابی از دیدگاه انفورماتیک برای اثبات کارایی و أثربخشی مدل پیشنهادی اجرا شده است. این مقایسه‌ی مبتنی بر نقشه گرمایی[۴۶] و با استفاده از مجموعه آزمون مستقل (TS125) است. در این نقشه رنگی، سلول‌هایی با مقادیر عددی بیشتری نسبت به سلول‌هایی با مقادیر عددی کمتر، دارای رنگ‌های تیره‌تری می‌باشند. برحسب شکل های ۱۲ و ۱۳، هر عنصر روی قطر اصلی نشان‌دهنده تعداد نمونه‌هایی است که به درستی به همان کلاس خودشان اختصاص داده شده‌اند. در این مطالعه، مجموعه آزمون مستقل (TS125)، شامل ۱۷۱۹ باقیمانده پیوندی واقعی از مجموع ۳۰۸۷۰ باقیمانده اسیدآمینه (پیوندی و غیرپیوندی) است. بنابراین، مدل پیشنهادیGSPPIR توانسته ۱۷۱۱ باقی‌مانده را به درستی به‌عنوان باقی‌مانده پیوندی پیش‌بینی نماید. مادامی‌که مدل SPRINT-Seq [۳۵] در مرحله آزمون ۱۷۰۸ باقی‌مانده پیوندی را به درستی تشخیص داده است.

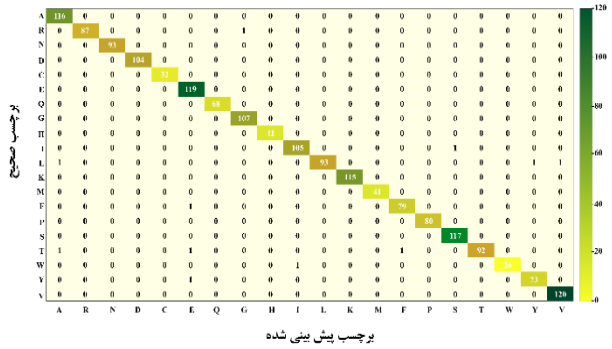
.




**شکل۱۱. فاکتور غنی‌سازی برحسب امتیاز احتمالاتی نواحی پیوند**




**شکل۱۲. شماتیک باقی‌مانده‌های پیوندی واقعی و باقی‌مانده‌های پیش‌بینی شده در مرحله آزمون برای مدل پیشنهادی GSPPIR**

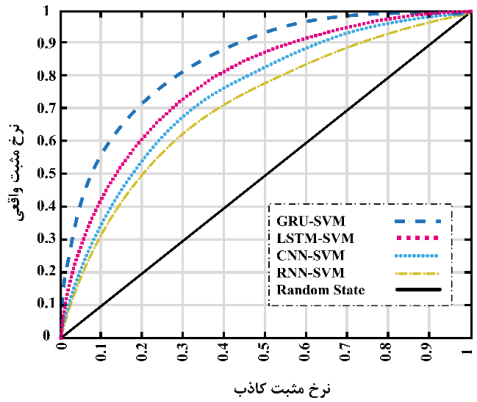



**شکل۱۳. شماتیک باقی‌مانده‌های پیوندی واقعی و باقی‌مانده‌های پیش‌بینی‌شده در مرحله آزمون برای مدل [۱۹]SPRINT-Seq**

مرحله بعدی، بیانگر انتخاب مدل نهایی از طریق بررسی و ارزیابی مدل‌های یادگیری متنوع است. لذا یک منحنی خصیصه عملیاتی گیرنده(ROC) اجرا شده است که عملکرد چهار مدل مختلف یادگیری: GRU-SVM، LSTM-SVM، CNN-SVM، RNN-SVM، را برای ساخت مدل پیشنهادی ارزیابی و مقایسه می‌کند.

این ارزیابی براساس نرخ مثبت واقعی در برابر نرخ مثبت کاذب است. همچنین، خط مورب سیاه رنگ بیانگر عملکرد تصادفی می‌باشد. برحسب مقایسه در شکل ۱۴، مدل GRU-SVM نسبت به دیگر مدل‌های کاندید شده دارای عملکرد پیش‌بینی بهتری می‌باشد. لذا مدل پیشنهادی، بیشینه مساحت زیر منحنی(۸۱/۰۱) را دارد. در نتیجه مدل GRU-SVM به عنوان مدل پیشنهادی GSPPIR،برای پیش‌بینی باقیمانده‌های پیوندی در برهمکنش‌های پروتئین-پپتید شناخته می‌شود.

گام آخر، دربرگیرنده ارزیابی اهمیت آماری برای بررسی عملکرد مدل پیشنهادی می‌باشد. هدف از این ارزیابی نیز پاسخ‌گویی به این سؤال است که آیا تفاوت معناداری مابین مدل پیشنهادی و سایر مدل‌های رقیب وجود دارد یا خیر؟




**شکل۱۴. ارزیابی عملکرد پیش بینی توسط مدل‌های مبتنی بر یادگیری**

بر حسب جدول ۱۰، تمام مقادیر P-Value کمتر از ۰/۰۵ می‌باشند و به صورت مستقیم فرض صفر ($H_0$) را رد می‌کنند. منظور از فرض صفر همان برابر بودن عملکرد مدل پیشنهادی (GSPPIR) و سایر مدل‌های رقیب است. لازم به ذکر است که روند کاهشی مقادیر، حاکی از برتری آماری است.

مقادیر صعودی برای T-test[36]، بیانگر فاصله زیاد و غیر قابل انکار توزیع تفاوت‌ها از صفر و شواهد قوی علیه شانس است. به بیانی دیگر، برتری مدل پیشنهادی با رد فرض صفر و شواهد قوی علیه فرض صفر مرتبط است.

مقادیر صعودی برای میانگین تفاوت($\bar{d}$)، بیانگر بهینگی مدل پیشنهادی در عمل به صورت چشمگیر (بیش از نه واحد) و قدرتمند است.

**جدول ۱۰. ارزیابی آماری مبتنی بر آزمون تی دو طرفه**

| معیار بهینگی | مدل | P-Value | T-test | $\bar{d}$ | CI ٪۹۹ |
|---|---|---|---|---|---|
| | Pepsite [14] | ۰/۰۳۸ | ۳۰/۹۹۹ | ۹/۳۰ | [8/51,10/09] |
| | Peptimap [15] | ۰/۰۳۵ | ۳۲/۰۰۰ | ۹/۶۰ | [8/81,10/39] |
| | SPRINT-Seq [19] | ۰/۰۳۲ | ۳۳/۱۱۱ | ۹/۹۳ | [9/14,10/72] |
| | PepBind [11] | ۰/۰۳۰ | ۳۴/۲۵۰ | ۱۰/۲۸ | [9/49,11/07] |
| | Visual [22] | ۰/۰۲۸ | ۳۵/۵۰۰ | ۱۰/۶۵ | [9/86,11/44] |
| اندازه‌گیری-اف | PepNN-Seq [23] | ۰/۰۲۷ | ۳۶/۸۰۰ | ۱۱/۰۴ | [10/25,11/83] |
| | GP+SVM [20] | ۰/۰۲۶ | ۳۷/۵۰۰ | ۱۱/۲۵ | [10/46,12/04] |
| | SPPPred [21] | ۰/۰۲۵۵ | ۳۷/۹۰۰ | ۱۱/۳۷ | [10/58,12/16] |
| | PepBCL [24] | ۰/۰۲۵۲ | ۳۸/۰۰۰ | ۱۱/۴۰ | [10/61,12/19] |
| | PepCNN [25] | ۰/۰۲۵۱ | ۳۸/۰۵۰ | ۱۱/۴۳ | [10/63,12/21] |
| | PepPFN[38] | ۰/۰۲۵۰ | ۳۸/۰۸۸ | ۱۱/۴۲ | [10/64,12/22] |
| | E2EPep[39] | ۰/۰۲۴۸ | ۳۸/۲۰۰ | ۱۱/۵۰ | [10/71,12/29] |
| | DPPBP[40] | ۰/۰۲۴۵ | ۳۸/۳۵۰ | ۱۱/۵۵ | [10/76,12/34] |
| | GSPPIR(our method) | ۰/۰۲۴۰ | ۳۸/۵۰۰ | ۱۱/۶۰ | [10/81,12/39] |

مقادیر موجود در ستون فاصله اطمینان(CI) ٪۹۹، بیانگر بازه‌های کاملا مثبت، همپوشانی‌های مثبت و رو به رشد و عدم وجود صفر در هیچ بازه‌ای است. عدم وجود صفر در هیچ بازه‌ای بیانگر رد قاطع فرض صفر است. به بیانی جزئی‌تر، زمانی که صفر در فاصله اطمینان (CI) جای نمی‌گیرد، به معنای آن است که عدم تفاوت یک احتمال غیر قابل قبول است.

با توجه به توضیحات، برتری مدل پیشنهادی (GSPPIR) ، یک یافته‌ی تصادفی نیست، بلکه سیستماتیک، کمی شده و از بالاترین سطوح اطمینان آماری برخوردار است. لذا، برتری مدل پیشنهادی GSPPIR از نظر آماری با مقادیر موجود در ستون‌های جدول ۱۰ توجیه می‌گردد.

P-Value,T-test,$\bar{d}$ , CI، به ترتیب بیانگر معناداری آماری تفاوت مابین مدل‌ها، قدرت شواهد قوی علیه فرض صفر، اندازه عملی بهبود و برآورد فاصله‌ای با اطمینان بالا است. با توجه به مقدار P-Value ،تفاوت معناداری و با مقدار T-test شواهد بسیار قوی برای مدل پیشنهادی اثبات می‌گردد. مقادیر CI, $\bar{d}$ بیانگر بهبود بزرگ و دقت بالا می‌باشند.

در نهایت، جدول ۱۱شامل مشخصات محاسباتی مدل پیشنهادی است که کارایی مدل را تأیید می‌کند.

**جدول ۱۱. مشخصات محاسباتی مدل پیشنهادی**

| فرآیند | مقدار |
|---|---|
| زمان آموزش (برای ۱ فولد) | ۱:۲۲:۱۱ (hr:min: sec) |
| زمان استنتاج | ۰۰:۱۱:۰۷ (hr:min: sec) |
| تعداد پارامترها | ۶۹۲/۴۳ (million) |
| واحد پردازش گرافیکی | RTX 2060 Ti (۶/۵ TFLOPs) |
| پردازنده مرکزی | Intel Core i7-10750H |

## ۶- نتیجه‌گیری

در این مطالعه، یک مدل ترکیبی مبتنی بر یادگیری عمیق و ماشین برای پیش‌بینی باقیمانده‌های پیوندی برهمکنش‌های پروتئین-پپتید معرفی شده است. کارآمدی مدل پیشنهادی براساس مزایایی همچون تقویت عملکرد طبقه‌بندی با ساخت خودکار ویژگی‌های متمایز و معنادار توسط ماژول یادگیری عمیق، پشتیبانی از کرنل‌های متنوع، مرزبندی بهینه بین کلاس‌ها، بهبود پایداری و قابلیت تعمیم بالا، اعمال خاصیت محلیت با نگاشت آرایه‌های یک بعدی(اسید آمینه‌های پیوندی) به آرایه دو بعدی، پیاده‌سازی مدل پیش‌بینی‌کننده مبتنی بر الگوریتم‌های یادگیری توأم با استخراج انواع خصیصه‌های بیوانفورماتیکی (اطلاعات تکاملی، مبتنی بر توالی، مبتنی بر ساختار و فیزیکوشیمیایی) و مدیریت بیش پیش‌بینی کلاس غیرپیوندی قابل توجیه می‌باشد. برای اثبات استحکام و پایداری مدل پیشنهادی، ارزیابی‌های استاندارد و بین رشته‌ای (زیستی و انفورماتیکی) انجام شده است. این ارزیابی‌ها نشان‌دهنده‌ی برتری روش پیشنهادی نسبت به مدل‌های مبتنی بر یادگیری موجود با استفاده از پایگاه داده BioLip [۲۶] و برحسب معیارهای ارزیابی: دقت، اندازه‌گیری-إف، تعادل بهینه(کمینه اختلاف) بین حساسیت و خاصگی، فاکتور غنی‌سازی، نقشه گرمایی و نرخ موفقیت براساس دامنه‌های پروتئینی می‌باشد. در نهایت، انتظار داریم که مدل پیشنهادی به عنوان مکمل مدل‌های تجربی و گسترش دانش درباره‌ی درک عملکرد پروتئین‌ها، ایمونولوژی و تشخیص بیومارکرها نقش بسزایی داشته باشد.

برحسب نتایج، علی رغم کسب دقت پیش‌بینی بهینه توسط مدل پیشنهادی، هنوز هم نمونه‌های مثبت کاذب در فضای ویژگی وجود دارند که با نمونه‌های پیوندی واقعی همپوشانی دارند. به گونه‌ای که مدل پیشنهادی در تمایز این نواحی مبهم دارای محدودیت است و این أمر مستقیما بر کاهش مقدار سطح زیر منحنی نیز أثر دارد. همچنین، مدل پیشنهادی توانسته، بهبود قابل توجه در معیارهای ارزیابی اندازه‌گیری-إف و تعادل (کمینه اختلاف) مابین حساسیت و خصیصه ایجاد نماید. به بیانی دیگر، اگرچه مدل پیشنهادی، در بهبود محدودیت سوگیری شدید به کلاس اکثریت (غیرپیوندی) توانسته نسبت به سایر رقیبان موفق‌تر عمل نماید. اما هنوز هم نیازمند به بهبود دقت پیش‌بینی در شناسایی نمونه مثبت (پیوندی) واقعی است. با نگاهی به آینده، برای رفع محدودیت‌های مذکور و توسعه مدل پیش‌بینی کننده، طراحی یک شبکه کانولوشن گرافی با ماژول توجه چند سر و با استفاده از ساختارهای کریستالی، در تحقیقات آتی دنبال می‌شود.

## مراجع


[1] M. Y. Cao, S. Zainudin, and K. M. Daud, "Protein features fusion using attributed network embedding for predicting protein-protein interaction," *BMC genomics*, vol. 25, no. 1, p. 466, 2024, doi: 10.1186/s12864-024-10361-8.

[2] Z. Hekmati, J. Zahiri, and A. Aalami, "Computational prediction of protein–protein interactions' network in Arabidopsis thaliana," *Acta Physiologiae Plantarum*, vol. 45, no. 12, p. 142 ,2023, doi: 10.1007/s11738-023-03623-7.

[3] S. Shafiee, A. Fathi, and F. Abdali-Mohammadi, "A Review of the Uses of Artificial Intelligence in Protein Research," in *the Fourth National Conference on Proteins and Peptide science*, University of Isfahan,Isfahan,2019.

[4] M. H. Viet *et al.*, "In silico and in vitro study of binding affinity of tripeptides to amyloid β fibrils: implications for Alzheimer's

[24] R. Wang, J. Jin, Q. Zou, K. Nakai, and L. Wei, "Predicting protein–peptide binding residues via interpretable deep learning," *Bioinformatics*, vol. 38, no. 13, pp. 3351-3360, 2022, doi: 10.1093/bioinformatics/btac352.

[25] A. Chandra, A. Sharma, I. Dehzangi, T. Tsunoda, and A. Sattar, "PepCNN deep learning tool for predicting peptide binding residues in proteins using sequence, structural, and language model features," *Scientific Reports*, vol. 13, no. 1, p. 20882, 2023, doi: 10.1038/s41598-023-47624 5.35.

[26] zhanglab. "BioLip 2 for Ligand-protein binding database." http://zhanglab.ccmb.med.umich.edu/BioLiP (accessed 2024).

[27] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403-410, 1990, doi: 10.1016/S0022-2836(05)80360-2.

[28] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *Journal of computational chemistry*, vol. 33, no. 3, pp. 259-267, 2012, doi: 10.1002/jcc.21968.

[29] S. Iqbal, A. Mishra, and M. T. Hoque, "Improved prediction of accessible surface area results in efficient energy function application," *Journal of theoretical biology*, vol. 380, pp. 380-391, 2015, doi: 10.1016/j.jtbi.2015.06.012.

[30] C. Fang, Y. Shang, and D. Xu, "Prediction of protein backbone torsion angles using deep residual inception neural networks," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 3, pp. 1020-1028, 2018, doi: 10.1109/TCBB.2018.2814586.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[32] S. J. Yen and Y. S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," In *Lecture Notes in Control and Information Science* , 2006, doi: 10.1007/978-1-4939-6406-2_6.

[33] N. Sapoval, A. Aghazadeh, M. G. Nute, and et al, "Current progress and open challenges for applying deep learning across the biosciences," *Nat Commun*, vol. 13, p. 1728, 2022, doi: 10.1038/s41467-022-29268-7.

[34] A. Biegert, C. Mayer, M. Remmert, J. Soding, and A. Lupas, "The MPI Bioinformatics Toolkit for protein sequence analysis," *Nucleic acids research*, vol. 2, no. suppl_2, pp. W335-W339, 2006, doi: 10.1093/nar/gkw348.

[35] G. Taherzadeh, Y. Zhou, A. W. C. Liew, and Y. Yang, "Structure-based prediction of protein–peptide binding regions using Random Forest," *Bioinformatics*, vol. 34, no. 3, pp. 477-484, 2018, doi: 10.1093/bioinformatics/btx614.

[36] S. Y. Novak, "On the T-test," *Statistics & Probability Letters*, vol. 189, 2022, doi: 1016/j.spl.2022.109562.

[37] A. A. Mamun, A. Enan, D.A. Indah, J. Mwakalonge, G. Comert, and M. Chowdhury, "Crash severity risk modeling strategies under data imbalance," *arxiv.org/abs/2412.02094*, 2025,doi: 10.48550/arXiv.2412.02094.

[38] X. Li, B. Cao, H. Ding, J. Mwakalonge, N. Kang, & T. Song, "PepPFN:protein-peptide binding residues prediction via pre-trained module-based Fourier Network," In *2024 IEEE Conference on Artificial Intelligence (CAI)*, Singapore, 2024,doi: 10.1109/CAI59869.2024.00195.

[39] J. Hu, K.X Chen, B. Rao, J. YuanNi, M.A. Thafar, S. Albaradei & M. Arif, "Protein-Peptide Binding Residue Prediction Based on Protein Language Models and Cross-Attention Mechanism". *Analytical Biochemistry*, vol. 694 ,p. 115637, 2024,doi: 10.1016/j.ab.2024.115637.

[40] Y. Yang, Y. Hua, W. Zhang, & X. Song, "DPPBP: Dual-stream Protein-Peptide Binding Sites Prediction Based on Region Detection". In *2025 International Conference on Intelligence* , 2025, doi: 10.65286/icic.v21i2.56686.

[41] D. Chicco, G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", *BMC Genomics* ,vol.21 ,no.1, p.6 ,2020, doi:10.101186/s12864-019-6413-7.

disease," *The Journal of Physical Chemistry B*, vol. 119, no. 16, pp. 5145-5155, 2015, doi: 10.1021/acs.jpcb.5b00006.

[5] P. Scheltens et al., "Alzheimer's disease," *The Lancet*, vol. 397, no. 10284, pp. 1577-1590, 2021, doi: 10.1016/S0140-6736(20)32205-4.

[6] E. Giusto, T. A. Yacoubian, E. Greggio, and L. Civiero, "Pathways to Parkinson's disease: a spotlight on 14-3-3 proteins," *npj Parkinson's Disease*, vol. 7, no. 1, p. 85 2021, doi: 10.1038/s41531-021-00230-6.

[7] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284-2303, 2021, doi: 10.1016/S0140-6736(21)00218-X.

[8] E. C. Hutchinson, "Influenza virus," *Trends in microbiology*, vol. 26, no. 9, pp. 809-810, 2018, doi: 10.1016/j.tim.2018.05.013.

[9] A. Fathi and R. Sadeghi, "A genetic programming method for feature mapping to improve prediction of HIV-1 protease cleavage site," *Applied Soft Computing*, vol. 72, pp. 56-64, 2018, doi: 10.1016/j.asoc.2018.06.045.

[10] L. Yang et al., "COVID-19: immunopathogenesis and Immunotherapeutics," *Signal transduction and targeted therapy*, vol. 5, no. 1, p. 128, 2020, doi: 10.1038/s41392-020-00243-2.

[11] Z. Zhao, Z. Peng, and J. Yang, "Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method," *Journal of Chemical Information and Modeling*, vol. 58, no. 7, pp. 1459-1468, 2018, doi: 10.1021/acs.jcim.8b00019.

[12] S. Shafiee, A. Fathi, and F. A. Mohammadi, "Prediction of protein–peptide binding residues using classification algorithms," In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020,doi: 10.1109/BIBE50027.2020.00013.

[13] S. Shafiee and A. Fathi, "Prediction of protein–peptide-binding amino acid residues regions using machine learning algorithms," In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 2021,doi: 10.1109/CSICC52343.2021.9420568.

[14] E. Petsalaki, A. Stark, E. García-Urdiales, and R. B. Russell, "Accurate prediction of peptide binding sites on protein surfaces," *PLoS computational biology*, vol. 5, no. 3, p. e1000335, 2009, doi: 10.1371/journal.pcbi.1000335.

[15] A. Lavi *et al.*, "Detection of peptide ‐ binding sites on protein surfaces: The first step toward the modeling and targeting of peptide ‐ mediated interactions," *Proteins: Structure, Function, and Bioinformatics*, vol. 81, no. 12, pp. 2096-2105, 2013, doi: 10.1002/prot.24422.

[16] T. Kambe, B. E. Correia, M. J. Niphakis, and B. F. Cravatt, "Mapping the protein interaction landscape for fully functionalized small-molecule probes in human cells," *Journal of the American Chemical Society*, vol. 136, no. 30, pp. 10777-10782, 2014, doi: 10.1021/ja505517t.

[17] H. Lee, L. Heo, M. S. Lee, and C. Seok, "GalaxyPepDock: a protein–peptide docking tool based on interaction similarity and energy optimization," *Nucleic acids research*, vol. 4, no. W1, pp. W431-W435, 2015, doi: 10.1093/nar/gkv495.

[18] D. E. Shaw *et al.*, "Atomic-level characterization of the structural dynamics of proteins," *Science*, vol. 330, no. 6002, pp. 341-346, 2010, doi: 10.1126/science.1187409.

[19] G. Taherzadeh, Y. Yang, T. Zhang, A. W. C. Liew, and Y. Zhou, "Sequence ‐ based prediction of protein‐peptide binding sites using support vector machine," *Journal of computational chemistry*, vol. 37, no. 13, pp. 1223-1229, 2016, doi: 10.1002/jcc.24314.

[20] S. Shafiee and A. Fathi, "Combination of genetic programming and support vector machine-based prediction of protein-peptide binding sites with sequence and structure-based features," *Journal of Computing and Security*, vol. 8, no. 1, pp. 45-63, 2021, doi: 10.22108/jcs.2021.126817.1062.

[21] S. Shafiee, A. Fathi, and G. Taherzadeh, "Spppred: Sequence-based protein-peptide binding residue prediction using genetic programming and ensemble learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 3, pp. 2029-2040, 2022, doi: 10.1109/tcbb.2022.3230540.

[22] W. Wardah et al., "Predicting protein-peptide binding sites with a deep convolutional neural network," *Journal of Theoretical Biology*, vol. 496, p. 110278, 2020, doi: 10.1016/j.jtbi.2020.110278.

[23] O. Abdin, S. Nim, H. Wen, and P. M. Kim, "PepNN: a deep attention model for the identification of peptide binding sites," *Communications biology*, vol. 5, no. 1, p. 503, 2022, doi: 10.1038/s42003-022-03445-2.

**پاورقی‌ها:**

[1] Transient
[2] Binding residues
[3] Binding regions
[4] Weak binding affinity
[5] deep Gated recurrent unit and Support vector machine Protein-Peptide Interaction Residue
[6] Dropout
[7] Deep Gated recurrent unit
[8] Support vector machine
[9] Generalization Error
[10] Over-predicting
[11] Near Miss
[12] SMOTE
[13] Position-specific scoring matrix
[14] Hand-craft
[15] Ordering points to identify the clustering structure
[16] Bagging
[17] Overfitting
[18] Gradient vanishing
[19] Long dependencies
[20] Data Augmentation
[21] Grid search
[22] Validation set
[23] Sampling
[24] Sliding window size
[25] Locality
[26] Gray scale
[27] http://zhanglab. ccmb. med. umich. edu/BioLiP
[28] Over train
[29] Protein-peptide complex
[30] Independent test set
[31] Confusion matrix
[32] Cross-validation
[33] Standard deviation
[34] Random Forest
[35] K-Nearest Neighbor
[36] Deep Recurrent Neural Network
[37] Deep Convolutional Neural Network
[38] Deep Long Short-Term Memory
[39] PRE: Precision
[40] SPE: Specificity
[41] SEN: Sensitivity
[42] F-M: F-Measure
[43] Distance Center to Center
[44] Enrich factor
[45] Cutoff
[46] Heat Map