Shahid Bahonar University of Kerman
Agricultural Biotechnology Journal
Print ISSN: 2228-6705 Online ISSN: 2228-6500
Iranian Biotechnology Society

Providing an efficient model to predict antimicrobial peptides using artificial intelligence algorithms

Mahin Rasani

PhD Student, Animal Science Department, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. E-mail address: rasanimahin@gmail.com

Keyvan Karami

Graduated of PhD Animal Science, Department of Animal Science, Faculty of Agriculture, Ferdowsi Mashhad University, Mashhad, Iran. E-mail address: karamikeyvan@yahoo.com

Mohammadreza Nassiri

Professor, Department of Animal Science, Faculty of Agriculture, Ferdowsi Mashhad University, Mashhad, Iran. E-mail address: nassiryr@um.ac.ir

Mojtabi Tahmorthpour

Professor, Department of Animal Science, Faculty of Agriculture, Ferdowsi Mashhad University, Mashhad, Iran. E-mail address: thmoores@um.ac.ir

Mohammad Hadi Sekhavati

*Corresponding author. Assistant Professor, Department of Animal Science, Faculty of Agriculture, Ferdowsi Mashhad University, Mashhad, Iran. E-mail address: hadisekhavati@gmail.com

Abstract

Objective

The aim of this study was to propose an efficient algorithm to predict antimicrobial peptides using artificial intelligence algorithms.

Materials and methods

In this study, an updated AMP and non-AMP data set including physico-chemical characteristics at the level of amino acids and protein sequence in three animal species and humans was extracted. After data exploration and data pre-processing steps, four methods Supervised learning including Decision Tree, Random Forest, Naive Bayes and SVM on the AMP dataset with 10-fold cross-validation to build models and predict the AMP label using the evaluation criteria of specificity, sensitivity, rate Accuracy, precision, recall, F1 score and area under the rock curve (AUC) were evaluated.

Results

In this study, using an up-to-date dataset, a machine learning model has been successfully trained to predict antimicrobial peptides. A comprehensive set of features has been subjected to feature selection to identify key features of antimicrobial peptides. After selecting the feature, among the different generated models, the model based on the RF model classifier showed the best performance with Accuracy (95 percent), Precision (96 percent), Recall (95 percent), F1 Score (95 percent). the four classification of algorithms, Random Forest algorithm and SVM are the most accurate. The Decision Tree classification algorithm had the least accuracy.

Conclusions

According to the obtained results, the proposed RF model has a better performance than other models for AMP prediction. This model predicted some peptides as peptides with antimicrobial properties. This predictive approach can be useful in extracting AMPs with antimicrobial properties from AMP libraries in useful clinical applications before moving on to experimental studies. On the other hand, several features in the final selection properties indicate that these features are critical determinants of peptide properties and should be considered in the development of models to predict peptide activity. The executable code is available in the attached file.

Keywords: Machine Learning Algorithms, Antimicrobial Peptides, Area under the rock curve, Confusion Matrix

Paper Type: Research Paper.

Citation: Rasani M, Karami K, Nassiri MR, Tahmorthpour M, Sekhavati MH (2024) Providing an efficient model to predict antimicrobial peptides using artificial intelligence algorithms. *Agricultural Biotechnology Journal* 16 (3), 89-110.

Agricultural Biotechnology Journal 16 (3), 89-110. DOI: 10.22103/jab.2024.23466.1571

Received: June 30, 2024. Received in revised form: August 13, 2024.

Accepted: August 14, 2024. Published online: September 30, 2024.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.

مجله بیوتکنولوژی کشاورزی
شاپا چاپی: ۶۷۰۵-۲۲۲۸ شاپا الکترونیکی: ۶۵۰۰-۲۲۲۸
دانشگاه شهید باهنر کرمان
انجمن بیوتکنولوژی ایران

ارائه مدلی کارآمد به‌منظور پیش‌بینی پپتیدهای ضدمیکروبی با استفاده از الگوریتم‌های هوش مصنوعی

مهین رسانی

دانشجوی دکتری ژنتیک و اصلاح دام، گروه علوم دامی، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه : rasanimahin@gmail.com

کیوان کرمی

دانش‌آموخته دکتری تخصصی علوم دامی، گروه علوم دامی، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه: karamikeyvan@yahoo.com

محمدرضا نصیری

استاد، گروه علوم دامی، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه:nassiryr@um.ac.ir

مجتبی طهمورث پور

استاد، گروه علوم دامی، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه:thmoores@um.ac.ir

محمد هادی سخاوتی

*نویسنده مسئول: دانشیار، گروه علوم دامی، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه: hadisekhavati@gmail.com

تاریخ دریافت: ۱۴۰۳/۰۴/۰۹ تاریخ دریافت فایل اصلاح شده نهایی: ۱۴۰۳/۰۵/۲۳ تاریخ پذیرش: ۱۴۰۳/۰۵/۲۴

چکیده

هدف: هدف از انجام این تحقیق پیشنهاد یک الگوریتم کارآمد به‌منظور پیش‌بینی پپتیدهای ضدمیکروبی با استفاده از الگوریتم‌های هوش مصنوعی می‌باشد.

مواد و روش‌ها: در این تحقیق، ابتدا یک مجموعه داده پپتیدهای ضدمیکروبی و پپتیدهای فاقد فعالیت ضدمیکروبی به روز شامل ویژگی‌های فیزیکوشیمیایی در سطح اسیدهای آمینه و توالی پروتئین در سه گونه حیوانی و انسان استخراج گردید. پس از کاوش داده‌ها و مراحل پیش‌پردازش داده، چهار روش یادگیری با نظارت شامل الگوریتم درخت تصمیم‌گیری ، الگوریتم جنگل تصادفی، الگوریتم بیز ساده و الگوریتم ماشین بردار پشتیبان بر روی مجموعه داده پپتیدهای ضدمیکروبی و پپتیدهای فاقد فعالیت

ضـدمیکروبی با اعتبارسـنجی متقابل ۱۰ برابری برای سـاخت مدل‌ها و پیش‌بینی برچسـب پپتیدهای ضـدمیکروبی با اسـتفاده از معیارهای ارزیابی اختصاصی بودن، حساسیت، نرخ صحت، معیار دقیق بودن، نرخ کامل بودن، معیار اف و سطح زیر منحنی راک ارزیابی گردید.

نتایج: در این تحقیق با اسـتفاده از یک مجموعه داده به‌روز، یک مدل یادگیری ماشـین. با موفقیت برای پیش‌بینی پپتیدهای ضـدمیکروبی آموزش داده شـده اسـت. مجموعه جامعی از ویژگی‌های تحت انتخاب ویژگی قرار گرفته‌اند تا ویژگی‌های کلیدی پپتیدهای ضـدمیکروبی را شـناسـایی کنند. پس از انتخاب ویژگی، در میان مدل‌های مختلف تولید شـده، مدل مبتنی بر طبقه‌بندی کننده مدل جنگل‌تصـادفی با نرخ صـحت (۹۵ درصـد)، معیاردقیق بودن (۹۶ درصـد)، نرخ کامل بودن (۹۵ درصـد) و معیار اف (۹۵ درصد)، بهترین عملکرد را نشان داد. از چهار الگوریتم دسته‌بندی، الگوریتم جنگل تصادفی و ماشین بردار پشتیبان بیشترین دقت را دارند. و الگوریتم دسته‌بند درخت تصمیم‌گیری کمترین دقت را داشت.

نتیجه‌گیری: با توجه به نتایج به دست آمده مدل پیشنهادی جنگل تصادفی عملکرد بهتری نسبت به سایر مدل‌ها برای پیش‌بینی پپتیدهای ضـدمیکروبی دارد، این مدل برخی از پپتیدها را به‌عنوان پپتید با خاصـیت ضـدمیکروبی پیش‌بینی کردند. این رویکرد پیش‌بینی می‌تواند در اسـتخراج پپتیدهای ضـدمیکروبی از کتابخانه‌های پپتیدهای ضـدمیکروبی در کاربردهای بالینی مفید قبل از حرکت به مطالعات تجربی مفید باشـد. از سـوی دیگر، چندین ویژگی در ویژگی‌های انتخابی نهایی نشـان می‌دهد که این ویژگی‌ها تعیین‌کننده حیاتی خواص پپتیدها هستند و باید در توسعه مدل‌هایی برای پیش‌بینی فعالیت پپتیدها در نظر گرفته شوند. کد اجرایی در فایل پیوست موجود است.

کلیدواژه‌ها: الگوریتم‌های یادگیری ماشـین، پپتیدهای ضـدمیکروبی، سـطح زیر منحنی‌های مشـخصـه عملکرد گیرنده، ماتریس درهم‌ریختگی

نوع مقاله: پژوهشی.

استناد: رسانی مهین، کرمی کیوان، نصیری محمد رضا، طهمورث پور مجتبی، سخاوتی محمد هادی (۱۴۰۳) ارائه مدلی کارآمد به‌منظور پیش‌بینی پپتیدهای ضدمیکروبی با استفاده از الگوریتم‌های هوش مصنوعی. *مجله بیوتکنولوژی کشاورزی*، ۱۶(۳)، ۱۱۰-۸۹.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.

مقدمه

شبکه‌های عصبی مصنوعی برای کاهش محدودیت روش‌های رگرسیون سنتی پیشنهاد شده‌اند و می‌توانند برای مدیریت داده‌های غیرخطی و پیچیده، حتی زمانی که داده‌ها نادقیق و نویز هستند، استفاده شوند (Pour Hamidi et al., 2017). این شبکه‌ها شامل مجموعه‌ای از اجزای پردازش هستند که به‌عنوان نورون‌ها یا گره‌ها نیز شناخته می‌شوند که عملکرد آن‌ها بر اساس نورون‌های بیولوژیکی است (Khorshidi et al. 2019). این واحدها در لایه‌هایی تشکیل می‌شوند که اطلاعات ورودی را پردازش کرده و به لایه‌های بعدی منتقل می‌کنند. تولنایی شبکه در پردازش در نقاط قوت اتصال بین واحدی (یا وزن‌ها) جمع می‌شود و این توانایی از طریق فرآیند انطباق با مجموعه‌ای از الگوهای آموزشی به دست می‌آید (Ghotbaldini et al. 2019). علاوه بر این، روش شبکه عصبی مصنوعی کاملاً با رویکردهای آماری سنتی متفاوت است، که نیاز به یک الگوریتم مشخص برای تبدیل شدن توسط یک برنامه کامپیوتری دارد (Pour Hamidi et al., 2017). توسعه بسیار سریع فناوری‌های با کارایی بالا باعث تولید داده‌های حجیم در زیست‌شناسی و زیست‌فناوری شده است. این داده‌ها با مطالعه مولکول‌های زیستی، از قبیل متابولیت‌ها[1]، پروتئین‌ها[2]، RNA و DNA حاصل شده‌اند که برای درک و فهم نقش این مولکول‌ها در تعیین ساختار، عملکرد و دینامیک سیستم‌های زنده، مانند ارگانیسم، بافت یا سلول ضروری هستند (Mohammadabadi et al. 2024). داده‌های اومیکس می‌توانند به قدری بیش از حد بزرگ و پیچیده باشند که از طریق تجزیه و تحلیل بصری یا همبستگی‌های آماری قابل بررسی نیستند. این امر استفاده از به اصطلاح هوش ماشینی یا هوش مصنوعی[3] (AI) را ترویج داده است (Mohammadabadi et al. 2024). هوش مصنوعی نه تنها قادر به مدیریت حجم داده‌هایی است که برای ذهن انسان غیرقابل حل است، بلکه برای استخراج اطلاعاتی که فراتر از درک فعلی ما از سیستم تحت بررسی است نیز کاربرد دارد. نکته مهم این است که هوش مصنوعی به‌طور خودکار، از طریق تجربه به دست آورده از داده‌های آموزشی[4] عملکردش را بهبود می‌بخشد (Mohammadabadi et al. 2024). از سویی دیگر، امروزه مقاومت باکتری‌ها در برابر آنتی بیوتیک‌ها به یکی از چالش جدید دنیا تبدیل شده است. از این رو، تحقیقات گسترده‌ای بر روی ترکیبات طبیعی که مقاومت باکتریایی ایجاد نمی‌کنند و عوارض جانبی کمتری دارند متمرکز شده اهمیت ایجاد جایگزین‌های موثر برای آنتی بیوتیک‌های شناخته شده به دلیل افزایش مقاومت میکروبی در سال‌های اخیر شتاب بیشتری یافته است. بنابراین، پیش‌بینی، طراحی و مدل‌سازی محاسباتی پپتیدهای ضدمیکروبی[5] (AMP) از اهمیت بالایی برخوردار است (Nguyen et al. 2011). پپتیدهای ضدمیکروبی[6] (AMPs)، همچنین به‌عنوان "پپتیدهای دفاع میزبان" شناخته

[1] Metabolites
[2] Proteins
[3] Machine Intelligence or Artificial Intelligence
[4] Training data
[5] Antimicrobial peptide (AMP)
[6] Antimicrobial peptides (AMPs)

می‌شوند، پروتئین‌های کوچکی هستند که میکروب‌ها را مهار یا از بین می‌برند. شناخته شده‌ترین عملکرد آن‌ها به‌عنوان بخشی از سیستم ایمنی ذاتی است که در آن محافظت در برابر عوامل بیماری‌زا ارائه می‌کنند همچنین دارای طیف وسیعی از فعالیت‌های ضدمیکروبی بر علیه ارگانیسم‌های هدف از قبیل باکتری‌ها، قارچ‌ها، انگل‌ها و ویروس‌ها می‌باشند (Yeaman & Yount 2003). کشف پپتیدهای ضدمیکروبی به سال ۱۹۳۹ باز می‌گردد. تاکنون هزاران پپتیدهای ضدمیکروبی کشف شده است (Conlon 2010;Leippe 1999).

در پژوهشی Dubos et al. (1939) یک عامل ضدمیکروبی را از سویه باسیلوس خاک استخراج کردند. اولین پپتید ضدمیکروبی با منشا حیوانی گزارش شده دفنسین، است که در سال ۱۹۵۶ از لکوسیت‌های خرگوش جدا شد (Hirsch 1956). در سال‌های بعد، بومبینین از اپیتلیوم (Kiss & Michl 1962)، و لاکتوفرین از شیر گاو (Groves et al. 1965)، هر دو استخراج شدند. در همان زمان، همچنین ثابت شد که لکوسیت‌های انسانی حاوی پپتیدهای ضدمیکروبی (AMP) در لیزوزوم‌های خود هستند. از دهه ۱۹۸۰، مدل‌های روابط ساختار فعالیت کمی محاسباتی (QSARs) برای پیش‌بینی و بهینه‌سازی توالی برخی از فعالیت‌های بیولوژیکی استفاده شده‌اند. در طبیعت پپتیدهای ضدمیکروبی به‌طور انتخابی میکروب‌های مهاجم را بدون تداخل مضر با سلول‌های میزبان هدف قرار می‌دهند، و دلایل این امر به ساختار و ویژگی‌های فیزیکوشیمیایی آن‌ها مربوط می‌شود (Aronica et al. 2021). با توجه به اینکه کشف پپتیدهای ضدمیکروبی با استفاده از آزمایشات آزمایشگاهی زمان‌بر و پرهزینه می‌باشد، در سال‌های اخیر، بسیاری از روش‌های محاسباتی برای تسریع فرآیند کشف و طراحی داروی ضدمیکروبی با ارائه مبنایی منطقی برای انتخاب پپتیدهای ضدمیکروبی توسعه داده شده‌اند. الگوریتم‌های یادگیری ماشین را می‌توان به‌عنوان مدل‌هایی استفاده کرد که بین پپتیدهای ضدمیکروبی (AMPs) و پپتیدهای فاقد فعالیت ضدمیکروبی (Non-AMPs) فرایند تمایز و دسته‌بندی رکوردها را انجام می‌دهند (Lee et al. 2016; Su et al. 2019). اخیراً، مطالعات زیادی به توسعه مدل‌های پیش‌بینی با استفاده از تکنیک‌های یادگیری ماشین برای پیش‌بینی پپتیدهای ضدمیکروبی (AMP) بر اساس توالی آن‌ها اختصاص یافته است (Lee 2016; Su 2019). در پژوهشی Chaudhary et al. (2017) ابزاری برای پیش‌بینی فعالیت همولیتیک پپتیدها ایجاد کرد. ویژگی‌های مورد استفاده در کار آن‌ها بیشتر شامل ویژگی‌های فیزیکوشیمیایی بود. در پژوهشی Cherkasov et al. (2009) با استفاده از یک مدل شبکه عصبی مصنوعی[1] (ANN)، همراه با تکرارهای متعدد داده‌های جمع‌آوری شده، پپتید جدید با طول ۵۰ تا ۱۰۰ اسیدآمینه را با فعالیت در برابر باکتری‌های مقاوم به دارو، شناسایی کردند. در پژوهشی Wang et al. (2012) پیش‌بینی پپتیدهای ضدمیکروبی بر اساس توالی با روش‌های هم‌ترازی و انتخاب ویژگی را مورد مطلعه قرار دادند. در پژوهشی Maccari et al. (2013) جهت طراحی و اعتبار فعالیت ضدمیکروبی پپتیدهای طبیعی و یک پپتید با اسیدهای آمینه غیرطبیعی از مدل‌های جنگل تصادفی (RF) استفاده کردند. در پژوهشی Schneider et al. (2017) با موفقیت از رویکرد شبکه‌ای عمیق برای تجزیه

[1] Artificial Neural Networks

و تحلیل داده‌های پپتید ضدمیکروبی استفاده و بیان کردند یادگیری عمیق، رده‌ای از الگوریتم‌های یادگیری ماشین است. پپتیدهای ضدمیکروبی خاصیت ضدسرطان هم دارند. طبق پایگاه داده AMP (http://aps.unmc.edu/AP)، ۳۲۸۳ پپتیدهای ضدمیکروبی وجود دارد که تقریباً ۲۵۹ پپتید به‌عنوان پپتیدهای ضدسرطان فهرست شده است (Jafari et al. 2022). در سال‌های اخیر، کارهای زیادی با شکل‌گیری الگوریتم‌های یادگیری ماشین (ML)، برای کشف پپتیدهای ضدمیکروبی، ضدسرطانی و داروها توسعه یافته‌اند.

در این رویکرد پیشنهادی، از زبان برنامه نویسی پایتون که دربخش مواد و روش‌ها ارائه شده است استفاده شد. پایتون با مجموعه‌ای از کتابخانه‌های علم داده و یادگیری ماشین عرضه می‌شود که می‌توان از آن‌ها برای پیش‌بینی براساس ویژگی‌های مختلف یک مجموعه داده استفاده کرد. پایتون به خوبی با الگوریتم‌های یادگیری ماشین سازگار است. برخی کتابخانه‌های مفید و کاربردی پایتون که دراین پژوهش استفاده شد در جدول شماره ۱ آورده شده است. الگوریتم‌های بیز ساده[1] ترکیبی از الگوریتم‌های طبقه‌بندی براساس قضیه بیز هستند (Aburomman & Reaz 2017). الگوریتم بیز ساده به‌طور کامل به‌عنوان احتمال شرطی و حداکثر احتمال وقوع تعریف شده است (Kumar & Bhatnagar 2021). الگوریتم ماشین بردار پشتیبانی[2] (SVM) به‌عنوان الگوریتم، یادگیری ماشین با نظارت[3] مشخص می‌شود که ادغام الگوریتم‌های یادگیری برای طبقه‌بندی، رگرسیون و تجزیه و تحلیل داده‌ها است. الگوریتم ماشین بردار پشتیبانی با جداسازی ابرصفحه‌ها با استفاده از طبقه‌بندی کننده‌های متمایز تعریف می‌شود (Yin et al. 2017). الگوریتم درخت تصمیم‌گیری[4] (DT) پارتیشن‌بندی بازگشتی را انجام می‌دهد، این الگوریتم به‌طور مکرر مقادیر ویژگی‌های مورد استفاده برای استخراج الگوهای ممکن را بین متغیرهای هدف و معمولی تقسیم می‌کند (Safavian & Landgrebe 1991; Jamali et al. 2016). الگوریتم جنگل تصادفی[5] (RF) همان‌طور که از نام آن پیداست، یک مجموعه مبتنی بر درخت است که هر درخت به مجموعه‌ای از متغیرهای تصادفی بستگی دارد (Cutler et al. 2012). این الگوریتم یک الگوریتم یادگیری تحت نظارت است که از آن هم برای طبقه‌بندی و هم رگرسیون استفاده می‌شود. اما به‌طور کلی برای مسائل طبقه‌بندی از آن استفاده می‌شود. الگوریتم درخت تصادفی روی نمونه‌های داده، درختان تصمیم‌گیری می‌سازد و سپس از هر کدام از آن‌ها پیش‌بینی می‌گیرد و در نهایت به واسطه رای‌گیری، بهترین راه حل را انتخاب می‌کند (Breiman 2001). هدف از این پروژه بکارگیری و استفاده از روش یادگیری ماشین و ارائه مدلی دقیق به‌منظور پیش‌بینی پپتیدهای ضدمیکروبی با استفاده از الگوریتم‌های هوش مصنوعی پپتیدهای ضدمیکروبی جدید در سه دام استراتژیک اهلی (گاو، گوسفند و مرغ) و انسان با استفاده خصوصیات فیزیکوشیمیایی در سطح اسیدهای آمینه و پروتئین می‌باشد.

[1] Algorithms Naive Bayes
[2] Algorithm Support Vector Machine
[3] Supervised Machine Learning
[4] Algorithm Decision tree
[5] Algorithm Random Forest

جدول ۱. معرفی برخی از کتابخانه‌های پایتون مفید و کاربردی در این فرآیند

Table 1. Introducing some useful and practical Python libraries in this process

کتابخانه سایکیت لرن Sklearn library	این کتابخانه دارای الگوریتم‌های یادگیری ماشین مانند ماشین بردار پشتیبان (SVM)، درخت تصمیم و k نزدیک‌ترین همسایه [1]است. از کتابخانه‌های عددی و آماری پایتون مانند کتابخانه نامپای[2] و سای پای [3] پشتیبانی می‌کند
learn	This library has machine learning algorithms such as support vector machine (SVM), decision tree and k nearest neighbor. It supports Python numerical and statistical libraries such as NumPy and SciPy
کتابخانه پانداس Pandas library	برای مدیریت و تحلیل داده‌های ساختار یافته و پردازش و پیش‌پردازش پایگاه داده‌های مختلف در تحقیقات مرتبط با پپتیدهای ضدمیکروبی
	Pandas' library is used for managing and analyzing structured data, as well as processing and pre-processing various databases in the studies related to antimicrobial peptides.
کتابخانه نامپای NumPy library	این کتابخانه منبع باز قدرتمند، به خوبی بهینه شده و رایگان برای زبان برنامه نویسی پایتون است که از آرایه‌های بزرگ و چند بعدی پشتیبانی می‌کند
	It is a powerful, well-optimized, and free open-source library for the Python programming language that supports large, multidimensional arrays

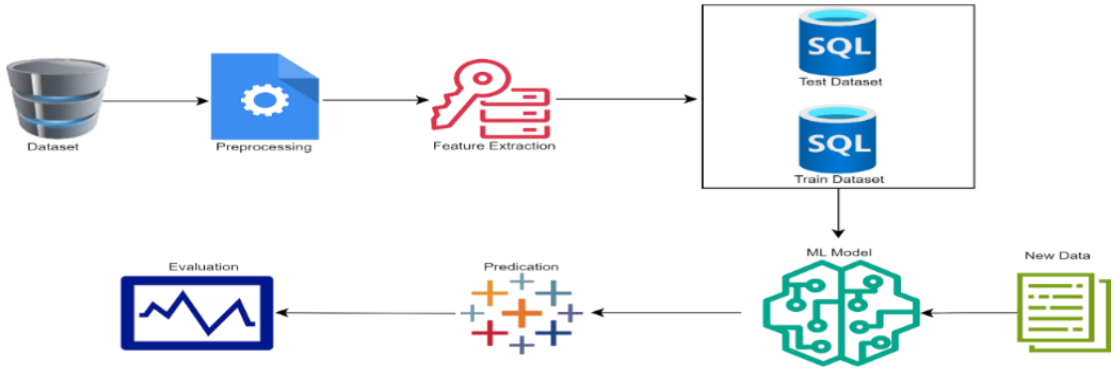
مواد و روش‌ها

فرایند پیش‌بینی پپتیدهای ضدمیکروبی با استفاده از استخراج ویژگی‌ها و به کمک روش‌های یادگیری با نظارت در شکل ۱ نشان داده شده است. هر کدام از مراحل فرایند پیاده سازی رویکرد پیشنهاد شده در ادامه توضیح داده‌اند.

جمع‌آوری داده‌ها: در پژوهش حاضر، داده‌های مثبت پپتیدهایی هستند که خاصیت ضدمیکروبی در آن‌ها به صورت آزمایشگاهی تایید شده است. برای ساخت مجموعه داده مثبت، توالی‌های پپتیدی ضدمیکروبی از پایگاه‌های داده (یا مجموعه‌های داده) در دسترس عموم جمع‌آوری شدند به‌طور خاص، پپتیدهای ضدباکتری از پایگاه داده DBAASP [4] جمع‌آوری شد. داده‌های منفی پپتیدهایی هستند که خاصیت ضدمیکروبی ندارند. برای ساخت مجموعه داده منفی، توالی‌های فاقد فعالیت ضدمیکروبی معمولاً از پایگاه داده UniProtKB / Swiss-Prot به دست می‌آید. مجموعه داده نهایی مورد استفاده در این مطالعه شامل ۱۵۸

[1] K-Nearest Neighbors
[2] NumPy library
[3] SciPy library
[4] Database of Antimicrobial Activity and Structure of Peptides

پپتید با خصوصیات ضدمیکروبی و ۱۱۰ پپتید فاقد فعالیت ضدمیکروبی بود که یک مجموعه داده دو کلاسه را تشکیل می‌داد. در ارتباط با پپتیدهای کوتاه زنجیر، طول این پپتیدها دراین مطالعه ۱۰۰ یا کمتر از ۱۰۰ اسیدآمینه در نظر گرفته شد. کلاس پپتیدهای ضدمیکروبی (AMPs) با ۱ و کلاس پپتیدهای فاقد فعالیت ضدمیکروبی (Non-AMPs) با ۰ برچسب‌گذاری شد. با استفاده از یک تنظیم کننده مقیاس این بار از نوع استاندارد اسکایلر[1] مقادیر ویژگی‌ها استاندارد شدند در یک واحد قرار گرفتند. ویژگی‌های نهایی در هر دسته ویژگی و زیر مجموعه در جدول ۲ آورده شده است.



شکل ۱. نمایش شماتیک روش‌های یادگیری ماشین نظارت شده در پیش‌بینی پپتیدهای ضدمیکروبی

Figure 1. Schematic representation of supervised machine learning methods in the prediction of antimicrobial peptides

جدول ۲. ویژگی‌های نهایی در هر دسته ویژگی و زیر مجموعه

Table2. Final features in each feature category and Sub-category

Feature Sub-category زیر مجموعه ویژگی	Feature Category دسته ویژگی
Charge density چگالی بار	Physico-chemical فیزیکوشیمیایی
Physicochemical composition ترکیب فیزیکوشیمیایی	
Physicochemical distribution توزیع فیزیکوشیمیایی	
Physicochemical transition انتقال فیزیکوشیمیایی	
Pseudo amino acid composition ترکیب شبه آمینو اسید	Pseudo-amino acid composition ترکیب شبه آمینو اسید
Amino acid composition ترکیب اسید آمینه	Amino acid composition ترکیب اسید آمینه

پیش‌پردازش مجموعه داده: یکی از موارد بدیهی موجود در هر مجموعه داده وجود داده‌های ناکامل ناسازگار و تکراری می‌باشد که باعث عدم انجام یک تصمیم‌گیری درست و صحیح توسط دسته‌بندها می‌شود. در واقع اگر مجموعه داده اولیه‌ای که برای آموزش و ساخت کلاسیفایر[2] یا دسته‌بند استفاده شده است دارای کیفیت و کمیت لازم نباشد خود باعث ایجاد و شکل‌گیری

[1] Standard Scaler
[2] Classifier

مدل‌های دسته‌بند بی‌دقت خواهد شد. از آنجا که برخی از ویژگی‌های انتخاب شده دارای مقادیر نویزی بوده و مقادیر ویژگی برچسب با رویکرد انتخابی ناسازگار می‌باشد از پیش‌پردازش به‌منظور بهبود دقت مدل استفاده گردید.

استخراج ویژگی: یکی از الزامات اساسی یادگیری ماشین و آموزش مدل، شناسایی و استخراج ویژگی‌ها[1] است. در این تحقیق برای محاسبه ویژگی‌ها از نرم‌افزار پایتون[2] نسخه ۳.۱۲.۳ استفاده شد (Cao et al. 2013). در این تحقیق، برای ۲۶۸ توالی (توالی‌های پپتیدهای ضدمیکروبی و همچنین توالی‌های کدکننده پروتئین‌های بیولوژیکی خاصیت ضدمیکروبی ندارند) در سطح پروتین و اسیدهای آمینه در گونه انسان و حیوانات استراتژیک (گاو، گوسفند، طیور) با استفاده از بسته رایگان پروپی پایتون[3] تعداد زیادی ویژگی استخراج گردید (Cao et al. 2013). که تنها ۱۹۸ ویژگی به‌عنوان ویژگی، با اهمیت شناسایی گردید که در ادامه به‌عنوان ویژگی ورودی استفاده خواهند شد. درسال ۲۰۱۰، .Yizeng et al و در سال ۲۰۱۳، .Cao et al پروژه propy را توسعه دادند. برخی از این ویژگی‌ها عددی و برخی رشته‌ای بودند بخشی از این ویژگی‌ها در جدول ۳ شده است.

ساخت مدل: با استفاده از عبارت دستوری، ترین اسپلیت[4] ویژگی‌های ۱ تا ۱۹۸ را به‌عنوان ویژگی ورودی در نظر گرفته شد. ویژگی ۱۹۸ را به‌عنوان ویژگی Y یا همان ویژگی هدف در نظر گرفتیم که در ادامه می‌خواهیم مدل را طوری آموزش دهیم که برحسب ویژگی‌های X بتواند مقدار ویژگی Y را پیش‌بینی کند.

آموزش مدل‌های یادگیری ماشینی: در اینجا، چندین مدل از جمله الگوریتم ماشین بردار پشتیبان، الگوریتم بیز ساده، الگوریتم جنگل تصادفی، الگوریتم درخت تصمیم روی مجموعه داده‌ها که شامل ۲۶۸ رکود هست برای پیش‌بینی پپتیدهای ضدمیکروبی آموزش داده شدند. در این مطالعه مبتنی بر روش‌های یادگیری ماشینی، از طریق اعتبارسنجی متقاطع ده برابری برای آموزش و تایید استفاده شد. مجموعه داده آموزشی به ده زیر گروه غیرهمپوشانی با اندازه‌های تقریباً مساوی تقسیم شد. در هر دور از نه زیرگروه برای آموزش[5] و یک زیر گروه برای تست[6] استفاده شد.

تست مدل یادگیری ماشینی: فرآیند اعتبارسنجی ده بار تکرار شد در هر مرحله اعتبارسنجی، عملکرد مدل‌های آموزش دیده در مرحله قبل با استفاده از، این معیارهای ارزیابی مورد بررسی قرار گرفت. از خروجی فرایند ارزیابی ماتریس درهم‌ریختگی[7] را به‌دست آوردیم. که از ماتریس درهم‌ریختگی چهار پارامتر مثبت حقیقی[8] (TP)، منفی حقیقی[9] (TN)، مثبت کاذب[10] (FP) و منفی

[1] Feature Extraction
[2] Python software
[3] Propy python package
[4] Train test split
[5] Train
[6] Test
[7] Confusion matrix
[8] True Positive
[9] True Negative
[10] False Positive

کاذب[1] (FN) محاسبه گردید. در گام بعدی براساس مقادیر ماتریس درهم‌ریختگی، شش معیار ارزیابی کارایی الگوریتم دسته‌بندی، یعنی اختصاصی بودن[2]، حساسیت[3]، صحت[4]، معیار دقیق بودن[5]، نرخ کامل بودن[6]، و معیار اف[7] سطح زیر منحنی‌های مشخصه عملکرد گیرنده[8] (AUC) محاسبه شدند.

جدول ۳. دسته‌بندی ویژگی‌های محاسبه شده برای هر پپتید

Table 3. Feature categories calculated for each peptide

Name نام	Category دسته‌بندی	Name نام	Category دسته‌بندی
_PolarityC1	Physico-chemical	_SecondaryStrD1050	Physico-chemical
قطبیت	فیزیکوشیمیایی	ثانوی	فیزیکوشیمیایی
charge Density	Physico-chemical	_HydrophobicityT12	Physico-chemical
چگالی شارژ	فیزیکوشیمیایی	آب‌گریزی	فیزیکوشیمیایی
_SecondaryStrD2001	Physico-chemical	_ChargeD1001	Physico-chemical
ثانوی	فیزیکوشیمیایی	شارژ	فیزیکوشیمیایی
_PolarityT23	Physico-chemical	_SecondaryStrD1100	Physico-chemical
قطبیت	فیزیکوشیمیایی	ثانوی	فیزیکوشیمیایی
_PolarityC3	Physico-chemical	_ChargeT12	Physico-chemical
قطبیت	فیزیکوشیمیایی	شارژ	فیزیکوشیمیایی
_HydrophobicityD3001	Physico-chemical	_PolarizabilityT23	Physico-chemical
آب‌گریزی	فیزیکوشیمیایی	قطبی‌پذیری	فیزیکوشیمیایی
_PolarizabilityD2050	Physico-chemical	_NormalizedVDWVD1075	Physico-chemical
قطبی‌پذیری	فیزیکوشیمیایی	نرمال شده	فیزیکوشیمیایی
_NormalizedVDWVC3	Physico-chemical	_HydrophobicityD2075	Physico-chemical
نرمال شده	فیزیکوشیمیایی	آب‌گریزی	فیزیکوشیمیایی
_SolventAccessibilityD1075	Physico-chemical	_SecondaryStrT13	Physico-chemical
قابلیت دسترسی به حلال	فیزیکوشیمیایی	ثانوی	فیزیکوشیمیایی
_PolarizabilityC2	Physico-chemical	_NormalizedVDWVT13	Physico-chemical
قطبی‌پذیری	فیزیکوشیمیایی	نرمال شده	فیزیکوشیمیایی
_NormalizedVDWVD2001	Physico-chemical	_HydrophobicityC2	Physico-chemical
نرمال شده	فیزیکوشیمیایی	آب‌گریزی	فیزیکوشیمیایی
_PolarizabilityD2001	Physico-chemical	_PolarizabilityT13	Physico-chemical
قطبی‌پذیری	فیزیکوشیمیایی	قطبی پذیری	فیزیکوشیمیایی
L	Amino acid composition	W	Amino acid Composition
لوسین	ترکیب اسید آمینه	تریپتوفان	ترکیب اسید آمینه

[1] False Negative
[2] Specificity
[3] Sensitivity
[4] Accuracy
[5] Precision
[6] Recall
[7] F1score
[8] Area under the ROC curve

به منظور مقایسه کارایی دسته‌بندها، منحنی مشخصه عملکرد سیستم[1] (ROC) ترسیم گردید. سطح زیر منحنی عمل گیرنده اغلب به‌عنوان شاخصی برای تعیین میزان قدرت یک مدل مورد استفاده قرار می‌گیرد. نمودار ROC از فرمول‌های ۵ و ۶ به‌دست می‌آیند. منحنی مشخصه عملکرد سیستم (ROC) نقشه حساسیت در برابر میزان مثبت کاذب می‌باشد. در یک پیش‌بینی دقیق و قابل اعتماد، اختصاصی بودن و حساسیت نزدیک به ۱ می‌باشند. سطح زیر منحنی‌های مشخصه عملکرد گیرنده (AUC) به‌عنوان یک معیار معتبر در بحث مدلینگ مطرح است. هر چه میزان سطح زیر منحنی‌های مشخصه عملکرد گیرنده (AUC) به ۱ نزدیک‌تر باشد، پیش‌بینی قابل اعتمادتر می‌باشد (Bradley 1997). فرمول محاسبه انواع پارامترهای ارزیابی مدل‌های یادگیری ماشین که با استفاده از ماتریس درهم‌ریختگی به دست می‌آید در ادامه تشریح شدند.

نرخ صحت: یعنی از کل پیش‌بینی‌های انجام شده توسط کلاسیفایر چند مورد درست بوده است. معادله ۱ بیان ریاضی دقت را نشان می‌دهد.

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{FN} + \boldsymbol{TN} + \boldsymbol{FP})} \ (\mathbf{1})$$

معیار دقیق بودن: برابر است با تقسیم تعداد مواردی که توسط مدل درست تشخیص داده شده است بر تعداد مواردی که واقعاً درست هستند، درست تشخیص داده شده‌اند (Marne et al. 2020). معادله ۲ بیان ریاضی دقت را نشان می‌دهد.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})} \ (\mathbf{2})$$

نرخ کامل بودن: برابر است با تقسیم تعداد مواردی که توسط مدل درست تشخیص داده شده‌اند بر تعداد کل مواردی که توسط مدل ایجاد شده است کامل بودن مدل را اندازه گیری می‌کند که چگونه می‌تواند پپتیدها با خاصیت ضدمیکروبی را شناسایی کند (Marne et al. 2020). معادله ۳ بیان ریاضی یادآوری را نشان می‌دهد.

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \ (\mathbf{3})$$

معیار اف: یک معیار مناسب برای ارزیابی دقت یک آزمایش است. این معیار، معیار دقیق بودن و نرخ کامل بودن را با هم در نظر می‌گیرد. معیار F1 در بهترین حالت، یک و در بدترین حالت صفر است (Marne et al. 2020).

$$\boldsymbol{F1\ score} = \mathbf{2}\left(\frac{\boldsymbol{Precision} \times \boldsymbol{recall}}{\boldsymbol{Precision} + \boldsymbol{recall}}\right) \ (\mathbf{4})$$

$$\boldsymbol{Sensitivity} = \boldsymbol{TPR} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})} \ (\mathbf{5})$$

$$\boldsymbol{Specificity} = \boldsymbol{FPR} = \frac{\boldsymbol{TN}}{(\boldsymbol{TN} + \boldsymbol{FP})} \ (\mathbf{6})$$

[1] Receiver Operating Characteristic

نتایج و بحث

مجموعه داده برای مدل‌های آموزشی: یک مشکل جدی بهداشت عمومی، شکست آنتی‌بیوتیک‌های مرسوم در کشتن باکتری‌های بیماری‌زا به دلیل ایجاد مقاومت چند دارویی است. روش‌های محاسباتی که می‌توانند به سرعت و با دقت پپتیدهای کاندید را به‌عنوان AMP برای سنجش‌های آزمایشی بعدی شناسایی کنند، برای کوتاه کردن فرآیند کشف دارو ضروری هستند. در مطالعه خود از ۱۹۸ ویژگی فیزیکوشیمیایی که در بالا به آن‌ها اشاره شد استفاده کرده‌ایم. دراین مطالعه ویژگی‌های فیزیکوشیمیایی را در سه گونه حیوانی در سطح پروتئین مورد بررسی قرار دادیم. در اینجا، از روش اعتبارسنجی متقاطع ۱۰ برابری[1] استفاده کردیم. خواص فیزیکوشیمیایی را محاسبه و از آن‌ها به‌عنوان ویژگی استفاده شد. در نهایت در این تحقیق، با استفاده از یک مجموعه داده به‌روز، یک مدل یادگیری ماشینی برای پیش‌بینی پپتیدهای ضدمیکروبی با عملکرد عالی ایجاد کردیم.

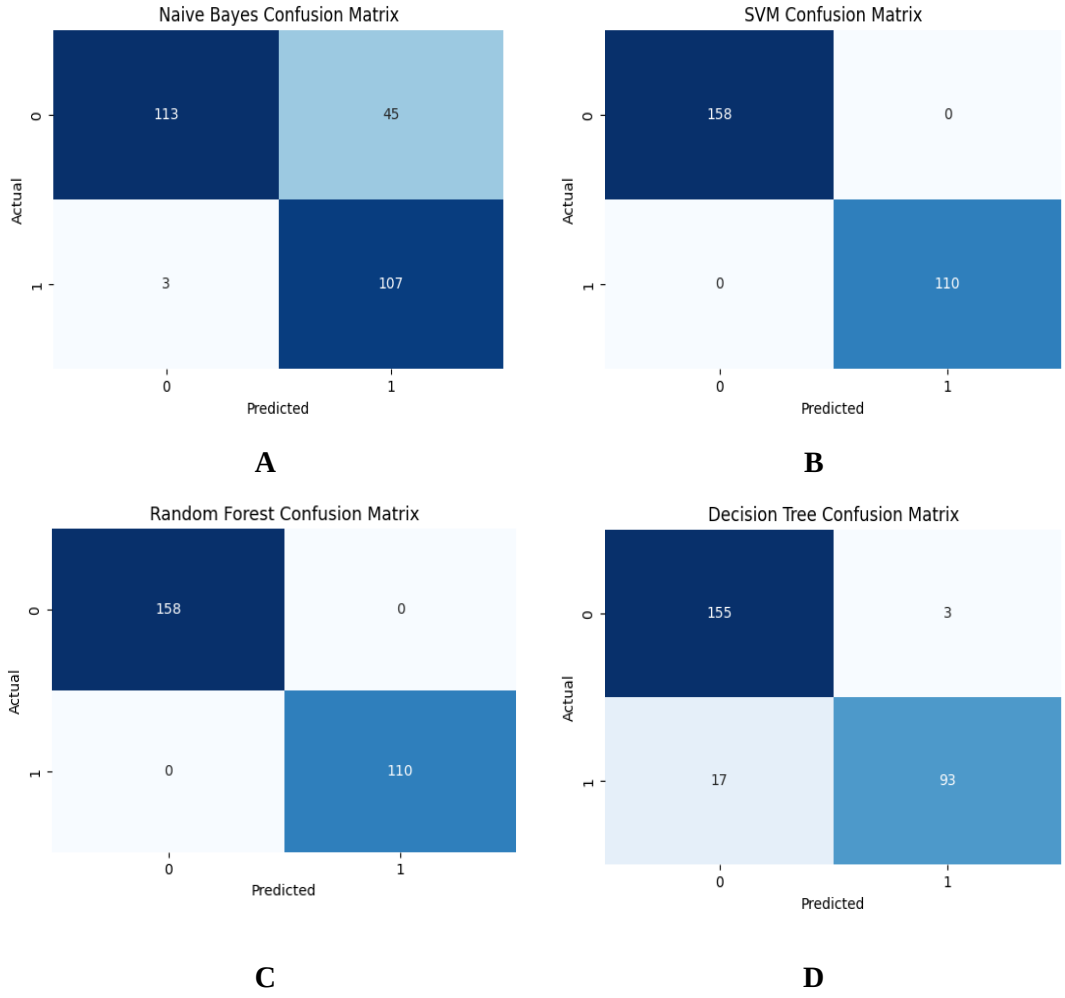
تحلیل ویژگی‌ها در بین پپتیدهای ضدمیکروبی و پپتیدهای فاقد فعالیت ضدمیکروبی: در این مطالعه به‌جای استفاده از یک الگوریتم یادگیری ماشین از چهار الگوریتم یادگیری ماشین استفاده شد و الگوریتمی که بهترین توجیه را داشت برای پیش‌بینی داده‌های جدید استفاده کردیم. همچنین نشان داده شده است که خواص فیزیکوشیمیایی نقش مهمی در عملکرد پپتید دارند و بنابراین باید در آموزش مدل‌های جدید در نظر گرفته شوند. در این تحقیق از ترکیبات شبه آمینواسیدها هم استفاده کردیم که در مطالعات گذشته ترکیب شبه آمینو اسید به‌جای ترکیب اسید آمینه چندین بار برای پیش‌بینی پپتیدهای ضدمیکروبی[2] (AMP) استفاده شده است زیرا اطلاعات توالی کمتری را از دست می‌دهد و بنابراین عملکرد پیش‌بینی AMP را افزایش خواهد داد (Wang et al. 2011;Khosravian et al. 2013; Zare et al. 2015; Lin & Xu 2016; Meher et al. 2017). ما در این مطالعه از ترکیب اسیدهای آمینه به‌عنوان ویژگی استفاده کردیم. مطالعات گذشته تفاوت ویژگی‌های ترکیبی پپتیدهای ضدمیکروبی (AMP) و پپتیدهای فاقد فعالیت ضدمیکروبی[3] (Non-AMP) را نشان داد به‌طور خاص، لوسین[4] "L"، گلیسین[5] "G"،لیزین[6] "K"، و اسیدهای آمینه فراوان برای پپتیدهای ضدمیکروبی (AMP) بودند، در حالی که لوسین "L"، آلانین[7] "A" و گلیسین "G" اسیدهای آمینه فراوان برای پپتیدهای فاقد فعالیت ضدمیکروبی (Non-AMP) بودند. علاوه بر این، تفاوت آشکاری در ترکیب سیستئین[8] "C" بین AMP و غیر AMP وجود دارد که دلیل این امر باید به دلیل غلبه مولکول‌های با پیوند دی‌سولفیدی و دفاعی باشد (Mishra 2012). ترکیب لیزین "K" بین پپتیدهای ضدمیکروبی (AMPs) و پپتیدهای فاقد فعالیت ضدمیکروبی (Non-AMPs) متفاوت بود، زیرا هسته‌های ساختاری AMP عمدتا دارای بار خالص مثبت بودند (Chang 2015).

[1] 10-fold cross-validation
[2] Antimicrobial Peptides
[3] Non-Antimicrobial Peptides
[4] Leucine
[5] Glycine
[6] lysine
[7] Alanine
[8] Cysteine

ما در این مطالعه از ویژگی آب‌گریزی استفاده کردیم. مطالعات گذشته نشان داد آب‌گریزی آشکارا بین پپتیدهای ضدمیکروبی (AMP) و پپتیدهای فاقد فعالیت ضدمیکروبی (Non-AMP) برای کلاس قطبی متفاوت بود در نتیجه می‌تواند به دلیل برهمکنش آب‌گریز با بخش‌های لیپیدی غشاها، که باعث اتصال پپتید به سلول نیز می‌شود (Matsuzaki 2009). در پژوهشی Torrent et al. (2011) یک رویکرد شبکه عصبی مصنوعی را براساس ویژگی‌های فیزیکوشیمیایی پپتیدهای ضدمیکروبی (AMP) توصیف کردند که نه تنها قادر به شناسایی پپتیدهای فعال است بلکه می‌تواند قدرت ضدمیکروبی آن را نیز ارزیابی کند. خواص فیزیکوشیمیایی در نظر گرفته شده مستقیماً از توالی پپتیدی مشتق شده است و مجموعه کاملی از پارامترها را تشکیل می‌دهد که پپتیدهای ضدمیکروبی (AMP) را به دقت توصیف می‌کند. در پژوهشی Bhadra et al. (2017) روش محاسباتی برای پیش‌بینی AMP با مدل الگوریتم جنگل تصادفی، الگوهای توزیع خواص فیزیکوشیمیایی اسیدآمینه در طول توالی پیش‌بینی کردند مجموعه‌های بزرگ و متنوع داده‌های AMP و غیرAMP با ۱۹ طبقه‌بندی جنگل تصادفی با اعتبارسنجی متقاطع ۱۰ برابری ارزیابی شدند. در پژوهشی Söylemez et al. (2023) یک تکنیک محاسباتی برای پیش‌بینی AMP با استفاده از ویژگی‌های فیزیکوشیمیایی در سطح اسیدآمینه یک مدل با استفاده از شبکه‌های عصبی عمیق[1](DNN) آموزش دادند. و مدل را با اعتبارسنجی متقاطع ۱۰ برابری بر روی یک مجموعه داده معیار ارزیابی کرد. در پژوهشی Lin et al. (2021) یک پیش‌بینی کننده پپتید ضدمیکروبی با استفاده از روش رمزگذاری مبتنی بر ویژگی فیزیکی و یادگیری عمیق هم مجموعه داده‌های به‌روز را جمع‌آوری کردند، که بر اساس آن روش‌های رمزگذاری پروتئین و مدل یادگیری عمیق برای پپتیدهای ضدمیکروبی (AMP) مورد بررسی قرار داد در نتیجه مدل آموزش دیده به دقت ۹۰ درصد دست یافته است.

ارزیابی عملکرد مدل برای طبقه‌بندی پپتیدهای ضدمیکروبی و پپتیدهای فاقد فعالیت ضدمیکروبی: پس از مرحله آموزش و تست مدل، نتایج مرحله تست مبنای اساسی ارزیابی عملکرد مدل برای مدل‌های طبقه‌بندی باینری مانند پیش‌بینی‌کننده‌های پپتیدهای ضدمیکروبی (AMP)، ماتریس درهم‌ریختگی است که چهار نتیجه ممکن را هنگام مقایسه نشان می‌دهد. با مشخص شدن پارامترهای ماتریس درهم‌ریختگی، می‌توان مدل‌های یادگیری ماشین را مورد ارزیابی قرار داد. در بررسی کارایی معیارهای ارزیابی، دقت، حساسیت، ویژگی، معیار دقیق بودن و سطح زیر منحنی مشخصه عملکرد گیرنده (AUC) مورد استفاده قرار گرفت. مقادیر این معیارها با استفاده از ماتریس درهم‌ریختگی[2] (جدول۴) نشان داده شده است. نتایج پیش‌بینی به مقادیر واقعی کلاس می‌رسد. ماتریس‌های درهم‌ریختگی هر نسخه از مدل طبقه‌بندی مورد مطالعه در این تحقیق در شکل ۲ نشان داده شده است.

[1] Deep learning neural network
[2] Confusion matrix



شکل ۲. ماتریس‌های درهم‌ریختگی هر نسخه از مدل طبقه‌بندی: ماتریس درهم‌ریختگی یک جدول ۲*۲ هست که مقادیر مثبت حقیقی، منفی کاذب ، منفی حقیقی و مثبت کاذب را نشان می‌دهد که هرکدام از مدل‌های روش یادگیری با نظارت شامل درخت تصمیم‌گیری، جنگل تصادفی، بیز ساده و ماشین بردار پشتیبان این چهار پارامتر را نشان می‌دهند که با استفاده از پارامترهای ماتریس درهم‌ریختگی مقادیر معیارهای ارزیابی اختصاصی بودن، حساسیت، نرخ صحت، معیاردقیق بودن، کامل‌بودن، معیار اف و سطح زیر منحنی راک محاسبه گردید. به‌عنوان مثال شکل ۳ (C) می‌توان این تفسیر را ارائه کرد. ماتریس درهم‌ریختگی، ۱۵۸ رکود که AMP هستند را به درستی AMP پیش‌بینی کرد و ۱۱۰ رکود Non-AMP را به درستی Non-AMP پیش‌بینی کرد

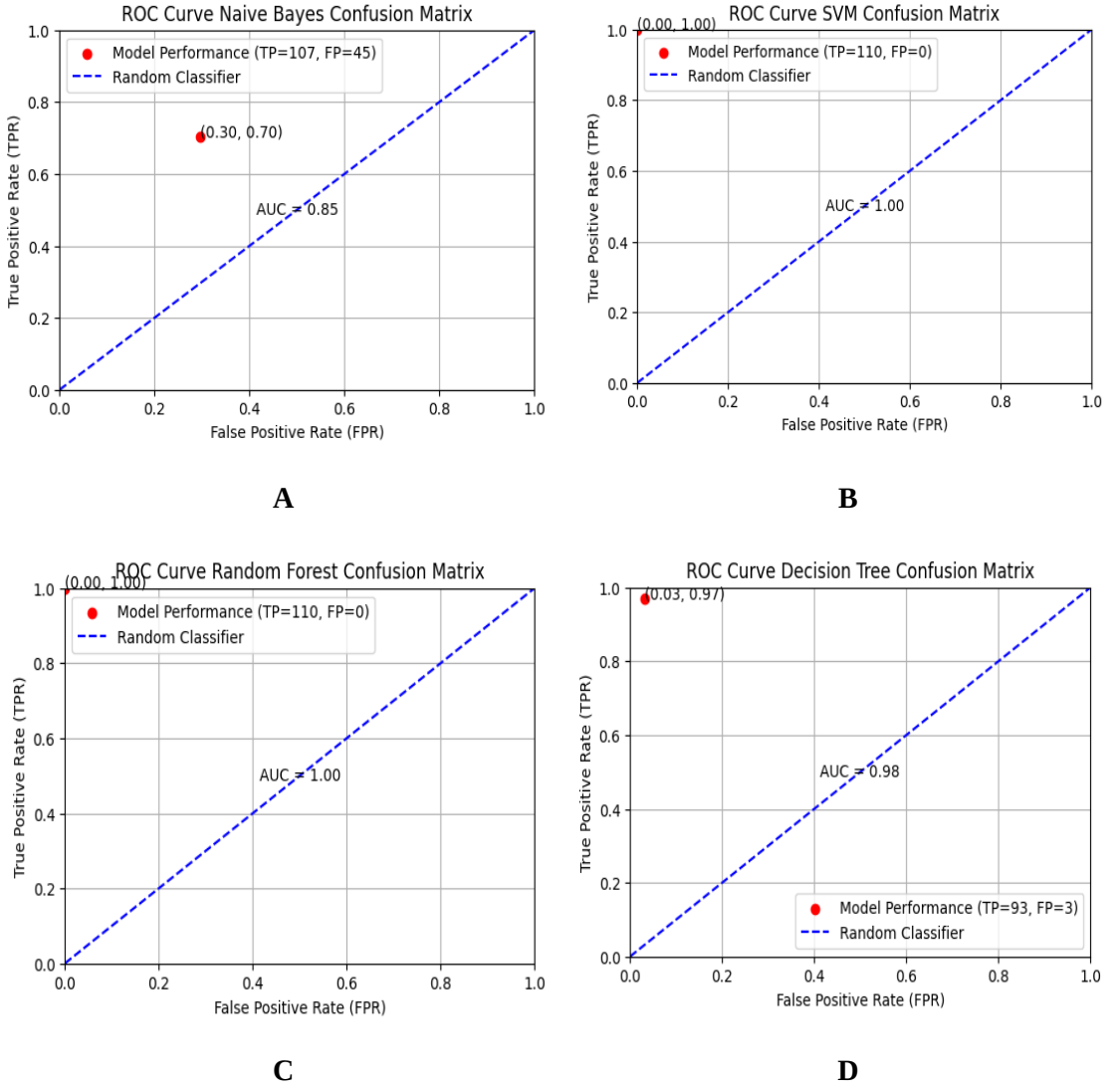
Figure 2. The confusion matrices of each version of the classification model: The confusion matrix is a 2x2 table that shows the values of true positives, false negatives, true negatives, and false positives that each model of the supervised learning method includes a decision tree. Giri, random forest, simple Bayes and support vector machine show these four parameters that use the parameters of the confusion matrix to evaluate the values of specificity, sensitivity, accuracy rate, accuracy, completeness, F-criterion and level. It was calculated under the rock curve. For example, Figure 3 (C) can provide this interpretation. The confusion matrix correctly predicted 158 AMP recessions as AMP and 110 Non-AMP recessions as non-AMP

منحنی‌های مشخصه عملکرد گیرنده (Gasteiger et al. 2005)، برای تعیین و مقایسه عملکرد مدل‌ها استفاده شد. سطح زیر منحنی‌های مشخصه عملکرد گیرنده (AUC) برای اندازه‌گیری توانایی یک طبقه‌بندی کننده برای تمایز بین کلاس‌ها استفاده می‌شود. هر چه AUC بالاتر باشد، عملکرد بهتری دارد. شکل ۳ نتایج ROC را نشان می‌دهد. همان‌طور که در شکل ۳ (c) مشاهده می‌شود، مدل جنگل تصادفی به امتیازات AUC بالاتری دست‌یافته است. در نتیجه مدل جنگل تصادفی به‌عنوان بهترین مدل شناسایی شد، این مدل چهار پپتید را پیش‌بینی کرد که می‌توانستند خاصیت ضدمیکروبی داشته باشند (جدول۵). در نهایت نتایج رویکرد پیشنهادی نشان داد الگوریتم جنگل تصادفی با نرخ صحت ۹۵ درصد در رتبه اول، الگوریتم ماشین بردار پشتیبان با نرخ دقت ۹۲ درصد در رتبه بعدی و الگوریتم‌های بیز ساده و درخت تصمیم با نرخ دقت‌های به ترتیب ۸۲، ۷۹ درصد در رتبه‌های بعدی قرار گرفتند. دقت به شناسایی و طبقه‌بندی صحیح نمونه‌ها (Khourdifi & Bahaj 2018) مربوط می‌شود. الگوریتم دسته‌بند جنگل تصادفی در دو سال گذشته حداقل چهار پیش‌بینی کننده AMP پیاده‌سازی کرده‌اند (Burdukiewicz 2020; Lin & Xu 2016; Bhadra 2018). هدف نهایی پیش‌بینی پپتیدهای ضدمیکروبی (AMP) طراحی توالی‌های پپتیدی جدید با اثرات ضدمیکروبی و درمانی مطلوب است.

جدول ٤. دقت مدل‌ها (هر کدام با شش معیار ارزیابی نرخ صحت، دقیق بودن، نرخ کامل بودن، معیار اف، حساسیت و ویژگی) بر روی مجموعه داده‌های محاسبه شده با اعتبارسنجی ۱۰ برابری. در جدول ٤ به وضوح نشان می‌دهد که دقت الگوریتم جنگل تصادفی و ماشین بردار پشتیبان بالاتر از سایر روش یادگیری ماشین است،. در نهایت الگوریتم جنگل تصادفی و ماشین بردار پشتیبان از نظر مقایسه سطح زیر منحنی راک بهترین عملکرد را دارند

Table 4. The accuracy of the models (each with six evaluation criteria of accuracy rate, precision, Recall, F1Score, sensitivity and specificity) on the dataset calculated with 10-fold validation. Table 4 clearly shows that the accuracy of RF and SVM t is higher than the other machine learning method is. Finally, RF and SVM perform better in comparison AUC performs best

Algorithms الگوریتم	Accuracy نرخ صحت	Precision دقیق بودن	Recall کامل‌بودن	F1 Score معیار اف	Sensitivity حساسیت	Specificity ویژگی	AUC سطح زیرمنحنی راک
SVM ماشین بردار پشتیبان	0.922	0.927	0.922	0.921	1	1	1
Random Forest جنگل تصادفی	0.959	0.963	0.959	0.958	1	1	1
Naive Bayes مدل بیز ساده	0.829	0.838	0.829	0.825	0.972	0.715	0.85
Decision Tree درخت تصمیم‌گیری	0.791	0.83	0.791	0.791	0.845	0.981	0.98



شکل ۳. منحنی‌های مشخصه عملکرد گیرنده (ROC)

Figure 3. Receiver operating characteristic curves (ROC)

نتیجه‌گیری: در نهایت لازم به ذکر است که در عصر اطلاعات گسترده زیستی، روش‌های محاسباتی و فنون رایانه‌ای علاوه بر کاهش هزینه‌های آزمایش به تسریع استخراج دانش از اطلاعات می‌پردازد و به گونه‌ای سرعت تولید داده و کسب دانش کاربردی از آن را متعادل می‌سازد. الگوریتم‌های یادگیری ماشین بردار پشتیبان (SVM)،الگوریتم بیز ساده (NB)، الگوریتم جنگل تصادفی (RF)، الگوریتم درخت تصمیم‌گیری (DT)، با استفاده از یک مجموعه داده به‌روز، برای پیش‌بینی پپتیدهای ضدمیکروبی مورد ارزیابی قرار گرفته‌اند. پس از برازش مدل‌های RF، SVM، NB، DT به مجموعه داده‌ها، معیار صحت برای مدل‌های مذکور به ترتیب برابر ۰/۹۵، ۰/۹۲، ۰/۸۲، ۰/۷۹ درصد و معیار دقیق بودن ۰/۹۶، ۰/۹۲، ۰/۸۳، ۰/۸۳ درصد و نرخ کامل بودن ۰/۹۵، ۰/۹۲، ۰/۸۲، ۰/۷۹ درصد و سطح زیر منحنی راک به ترتیب ۱، ۱، ۰/۸۵، ۰/۹۸ محاسبه گردید.

جدول ۵. پپتیدهای ضدمیکروبی پیش‌بینی شده با روش یادگیری ماشین نظارت شده

Table 5. Antimicrobial peptides predicted by supervised machine learning method

Uniport پایگاه داده	Gene ژن	SVM ماشین بردار پشتیبان	Random Forest جنگل تصادفی	Naive Bayes مدل بیز ساده	Decision Tree درخت تصمیم‌گیری
'P07470'	COX7A1[1] سیتوکروم c اکسیداز	AMP	Non- AMP	Non- AMP	AMP
'P02820'	BGLAP[2] پروتئین گاما کربوکسی گلوتامات	AMP	AMP	Non- AMP	AMP
'Q3SZ47'	MRPL33[3] پروتئین ریبوزومی میتوکندریL33	Non- AMP	Non- AMP	AMP	Non- AMP
'A8NN94'	MRPL34[4] پروتئین ریبوزومی میتوکندریL34	AMP	AMP	AMP	AMP
'Q32PC3'	MRPL27[5] پروتئین ریبوزومی میتوکندریL27	Non- AMP	Non- AMP	Non- AMP	Non- AMP
'Q3ZBI7'	ATP5MD[6]	AMP	Non- AMP	AMP	Non- AMP
'Q5ZLR5'	UQCRFS1[7] یوبی کینول–سیتوکروم سی ردوکتاز	Non- AMP	Non- AMP	Non- AMP	Non- AMP
'Q9BZL1'	UBL5[8] یوبیکوئیتین	Non- AMP	Non- AMP	Non- AMP	Non- AMP
'P53803'	POLR2K[9] آران‌ای پلی‌مراز ۲	Non- AMP	AMP	Non- AMP	AMP

[1] cytochrome c oxidase subunit 7A1
[2] bone gamma-carboxyglutamate protein
[3] mitochondrial ribosomal protein L33
[4] mitochondrial ribosomal protein L34
[5] mitochondrial ribosomal protein L27
[6] ATP synthase membrane subunit k
[7] ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1
[8] ubiquitin like 5
[9] RNA polymerase II, I and III subunit K

با توجه به معیارهای صحت، دقیق بودن، نرخ کامل بودن و سطح زیر منحنی راک، مدل برتر معرفی شد. مدل RF بالاترین صحت و بهترین عملکرد را داشت. و منحنی مشخصه عملکرد سیستم (ROC) نشان می‌دهد که در مقایسه با سایر الگوریتم‌ها، الگوریتم RF مورد استفاده در این تحقیق دقت بیشتری به دست می‌دهد. .با توجه به اطمینان بالای ما نسبت به صحت و دقت بالای مدل ساخته شده، انتظار می‌رود که پیش‌بینی مذکور انحراف کمی با واقعیت داشته باشد. بنابراین، می‌توان نتیجه گرفت که این ویژگی‌ها، در مجموع، دارای اطلاعات ضروری برای پیش‌بینی خاصیت ضدمیکروبی یک پپتید ضدمیکروبی (AMP) است.

سپاسگزاری: نگارندگان بر خود لازم می‌دانند از داوران و سردبیر محترم مجله به خاطر ارائه نظرهای ساختاری و علمی سپاسگزاری نمایند.

منابع

محمدآبادی محمدرضا، خیرالدین حمید، آفاناسنکو ولودیمیر، بابنکو اولنا، کلوپنکو ناتالیا، کلاشنیک الکساندر، ایوستافیوا یولیا، بوچکوفسکا ویتا (۱۴۰۳) نقش هوش مصنوعی در ژنومیکس. مجله بیوتکنولوژی کشاورزی، ۱۶(۲)، ۲۷۹-۱۹۵.

References

Aburomman AA, Reaz MBI (2017) A survey of intrusion detection systems based on ensemble and hybrid classifiers. J Comput Secur 65**,** 135-152.

Aronica PG, Reid LM, Desai N, et al. (2021) Computational methods and tools in antimicrobial peptide research. J Chem Inf Model 61, 3172-3196.

Bhadra P, Yan J, Li J, et al. (2018) AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. J Sci Rep 8, 1697

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. J Pattern Recognit 30**,** 1145-1159

Breiman L (2001) Random forests. J Mach Learn 45**,** 5-32

Cao DS, Xu QS, Liang YZ (2013) propy: a tool to generate various modes of Chou's PseAAC. J Bioinform 29(7), 960–962

Chaudhary K, Kumar R, Singh S, et al. (2016) A web server and mobile app for computing hemolytic potency of peptides.J Sci Rep 6**,** 22843

Cherkasov A, Hilpert K, Jenssen H, et al. (2009) Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs.J ACS Chem Biol 4, 65-74

Cutler A, Cutler DR, Stevens JR (2012) Random Forests. In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9326-7_5.

Dubos RJ (1939) Studies on a bactericidal agent extracted from a soil bacillus: I. Preparation of the agent. Its activity in vitro. J Exp Med 70, 1-10.

Feng Y, Xu J, Shi M, et al. (2022) COX7A1 enhances the sensitivity of human NSCLC cells to cystine deprivation-induced ferroptosis via regulating mitochondrial metabolism. J Cell Death Dis 13, e988

Gasteiger E, Hoogland C, Gattiker A, et al. (2005) Protein identification and analysis tools on the ExPASy server. J Springer https://doi.org/10.1385/1-59259-890-0:571.

Ghotbaldini H, Mohammadabadi MR, Nezamabadi-pour H, et al. (2019) Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed. Acta Scientiarum Anim Sci 41, e45282.

Groves M, Peterson R, Kiddy C (1965) Polymorphism in the Red Protein isolated from Milk of Individual Cows. J Nature 207, 1007-1008.

Hirsch JG (1956) Phagocytin: a bactericidal substance from polymorphonuclear leucocytes. J Exp Med 103, 589.

Jafari A,Babajani A,Sarrami Forooshani R, et al. (2022) Clinical applications and anticancer effects of antimicrobial peptides: from bench to bedside. J Front Oncol 12, 819563.

Jamali AA, Ferdousi R, Razzaghi S, et al. (2016) DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. J Drug Discov Today Technol 21, 718-724.

Khorshidi M, Mohammadabadi MR, Esmailizadeh AK, et al. (2019) Comparison of artificial neural network and regression models for prediction of body weight in Raini Cashmere goat. Iran J Appl Anim Sci 9 (3), 453-461.

Khosravian M, Kazemi Faramarzi F, Mohammad Beigi M, et al. (2013) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. J Protein Pept Lett 20, 180-186

Khourdifi Y, Bahaj M (2018) Applying best machine learning algorithms for breast cancer prediction and classification. In: International conference on electronics, control, optimization and computer science.J ICECOCS. IEEE. pp. 1-5

Kiss G, Michl H (1962) Uber das Giftsekret der Gelbbauchunke, Bombina variegata L. J Toxicon 1, 33-34.

Kohavi R (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. J IJCAI, Montreal, Canada. pp. 1137-1145.

Kumar K, Bhatnagar V (2021) Machine Learning Algorithms Performance Evaluation for Intrusion Detection. J Inf Technol Manag 13, 42-61.

Lee EY, Fulan BM, Wong GC, Ferguson AL (2016) Mapping membrane activity in undiscovered peptide sequence space using machine learning. J Proc Natl Acad Sci 113**,** 13588-13593.

Lin W, Xu D (2016) Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. J Bioinform 32**,** 3745-3752.

Marne S, Churi S, Marne M (2020) Predicting breast cancer using effective classification with decision tree and k means clustering technique. International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE. pp. 39-42.

Meher PK, Sahu TK, Saini V, Rao AR (2017) Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. J Sci Rep 7, 42362.

Mohammadabadi M, Kheyrodin H, Afanasenko V, Babenko O, Klopenko N, Kalashnyk O, Ievstafiieva Y, Buchkovska V (2024) The role of artificial intelligence in genomics. Agricultural Biotechnology Journal 16 (2), 195-279 (In Persian.)

Nguyen LT, Haney EF, Vogel HJ (2011) The expanding scope of antimicrobial peptide structures and their modes of action. J Trends Biotechnol 29**,** 464-472.

Pour Hamidi S, Mohammadabadi MR, Asadi Foozi M, Nezamabadi-pour H (2017) Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. J Livestock Sci Technol 5 (2), 53-61.

Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. J IEEE Trans Syst Man Cybern 21, 660-674

Schneider P, Müller AT, Gabernet G, et al. (2017) Hybrid network model for "deep learning" of chemical data: application to antimicrobial peptides. J Mol Inform 36**,** e1600011.

Söylemez UG, Yousef M, et al. (2023) Prediction of Antimicrobial Peptides Using Deep Neural Networks. In Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 3: BIOINFORMATICS, pages 188-194.

Su X, Xu J, Yin Y, et al. (2019) Antimicrobial peptide identification using multi-scale convolutional network. J BMC Bioinform 20**,** 1-10.

Wang G, Mishra B (2012) The importance of amino acid composition in natural AMPs: an evolutional, structural, and functional perspective. J Front Immunol 3, e31946

Wang P, Hu L, Liu G, et al. (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. J PloS one 6**,** e18476

Yeaman MR, Yount NY (2003) Mechanisms of antimicrobial peptide action and resistance. J Pharmacol Rev 55, 27-55.

Yin C, Zhu Y, Fei J, He X (2017) A deep learning approach for intrusion detection using recurrent neural networks. J IEEE Access 5, 21954-21961

Zare M, Mohabatkar H, Faramarzi FK, et al. (2015) Using Chou's Pseudo Amino Acid Composition and Machine Learning Method to Predict the Antiviral Peptides. J Open Bioinform 9, 13-19.