Published by: Iranian Society for Biomedical Engineering / www.isbme.ir

Iranian Journal of Biomedical Engineering

www.ijbme.org / P-ISSN: 2008-5869 / E-ISSN: 8006-9685

Volume 19, Issue 3, Autumn 2025, 177 - 191

Generalizable protein druggability prediction with a two-stage classifier and time-constrained Bayesian optimization

Shiva Shekarchian[1] / Rezaee, Khosro[*2] / Eslami, Hossein [3]

[1] - Master of Science, Department of Biomedical Engineering, Faculty of Engineering, Meybod University, Meybod, Iran.
[2] - Associated Professor, Department of Biomedical Engineering, Faculty of Engineering, Meybod University, Meybod, Iran.
[3] - Associated Professor, Department of Biomedical Engineering, Faculty of Engineering, Meybod University, Meybod, Iran.

ARTICLE INFO

DOI: 10.22041/ijbme.2025.2039584.1922
Received: 26/8/2024 Revised: 20-9-2025 Accepted: 22-2-2026

KEYWORDS

Protein druggability
Two-stage classifier
Bayesian optimization
Accuracy–latency
Machine learning

ABSTRACT

Drug discovery demands rapid and reliable identification of protein targets, yet most automated approaches prioritize accuracy while overlooking inference latency, robustness, and reproducible tuning; sequence- and graph-based deep models are often costly and slow, and evolutionary hyperparameter search exhibits high variance. The central gap is the absence of a method that jointly optimizes accuracy and delay while intelligently routing easy and hard cases and setting control parameters correctly. We therefore propose a two-stage classifier built on CatBoost with early exit for high-confidence instances and a deeper pass for ambiguous ones, coupled with latency-aware Bayesian Hyperband (LABO-HB) for hyperparameter tuning, a single-objective target aligning accuracy and latency, and data-driven threshold selection. After a cost–benefit audit of deep versus boosted alternatives, we adopted this two-stage design because it preserves accuracy while reducing inference time and performance variance. Evaluation on balanced ProTar-II, unseen ProTar-II-Ind, and DPI-CDF compiled from DrugBank and Swiss-Prot used rigorous preprocessing, sequence-based feature extraction, strict leakage prevention, and fair train–validation–test splits, and generalization remained stable across internal and external benchmarks. The method achieved 96.6% overall accuracy and an F1 score of 96.5%, while significantly reducing prediction time relative to baselines. The combination of early exit and LABO-HB delivers high accuracy with lower latency and reduced variability, is robust to class imbalance, and yields reproducible outputs. The approach is practically useful in drug design, enabling agile target prioritization during early screening and more judicious allocation of laboratory resources.

[*]Corresponding Author

Address	Department of Biomedical Engineering, Faculty of Engineering, Meybod University, Meybod, Iran
E-Mail	*kh.rezaee@meybod.ac.ir*

ناشر: انجمن مهندسی پزشکی ایران / www.isbme.ir

مجله‌ی مهندسی پزشکی زیستی

شاپای چاپی: ۵۸۶۹-۲۰۰۸ / شاپای الکترونیکی: ۹۶۸۵-۸۰۰۶ / www.ijbme.org

دوره: ۱۹، شماره: ۳، پاییز ۱۴۰۴، ۱۷۷ - ۱۹۱

پیش‌بینی قابل‌تعمیم داروپذیری پروتئین با طبقه‌بند دو مرحله‌ای و بهینه‌سازی بیزی مقید به زمان

شکرچیان، شیوا[۱]/رضائی، خسرو [۲*]/ اسلامی، حسین[۳]

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

[۲] – دانشیار، گروه مهندسی پزشکی، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران

[۳] – دانشیار، گروه مهندسی پزشکی، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران

مشخصات مقاله

شناسه‌ی دیجیتال: 10.22041/ijbme.2025.2039584.1922

ثبت در سامانه: ۱۴۰۳/۶/۵ بازنگری: ۱۴۰۴/۶/۲۹ پذیرش: ۱۴۰۴/۱۲/۳

واژه‌های کلیدی

داروپذیری پروتئین
طبقه‌بند دومرحله‌ای
بهینه‌سازی بیزی
تراز دقت-زمان
یادگیری ماشینی

چکیده

کشف دارو به شناسایی سریع و معتبر اهداف پروتئینی نیاز دارد. روش‌های خودکار پیشین تنها بر بیشینه‌سازی دقت تمرکز می‌کنند و از زمان پیش‌بینی، پایداری و بازتولیدپذیری تنظیم چشم می‌پوشند. رویکردهای عمیق توالی و گراف نیز پرهزینه و کُند هستند و جست‌وجوی تکاملی در تنظیم ابرپارامترها نوسان بالا دارد. خلأ اصلی نبود بهینه‌سازی هم‌زمان دقت و تأخیر همراه با مسیردهی هوشمند نمونه‌های آسان و دشواری در تنظیم صحیح پارامترهای کنترلی است. بر همین اساس، در این پژوهش یک طبقه‌بند دومرحله‌ای بر پایه روش دسته‌بندی درختی *CatBoost* با خروج زودهنگام برای موارد با اطمینان بالا و گذر ژرف برای موارد مبهم ارائه می‌شود و تنظیم ابرپارامترها با اَبَرباند بیزی آگاه از زمان پاسخ موسوم به *LABO-HB* و نیز هدف تک‌معیاری تراز دقت و تأخیر و آستانه‌گذاری داده‌محور انجام می‌گیرد؛ پس از پایش هزینه و مصالحه میان گزینه‌های عمیق و تقویتی به این طرح دومرحله‌ای روی آوردیم، زیرا با حفظ دقت، زمان پاسخ و نوسان عملکردی را کاهش داد. ارزیابی با مجموعه داده‌های *ProTar-II* متعادل و *ProTar-II-Ind* دیده نشده و نیز *DPI-CDF* گردآوری‌شده از *DrugBank* و *Swiss-Prot* و با انجام پیش پردازش، استخراج ویژگی مبتنی بر توالی و پیشگیری از نشت و تفکیک منصفانه آموزش-اعتبارسنجی-آزمون انجام شد و تعمیم‌پذیری در هر دو محک داخلی و بیرونی پایدار ماند. روش پیشنهادی به درستی کل ۹۶/۶ درصد و نیز *F1* معادل ۹۶/۵ درصد دست یافت و نیز زمان پیش‌بینی را نسبت به مبنا به‌طور معنادار کاست. پیوند خروج زودهنگام با *LABO-HB* دقت بالا را با تأخیر کمتر و نوسان اندک همراه کرد و در برابر نامتوازنی، خروجی بازتولیدپذیر را ایجاد کرد. این روش در طراحی دارو مفید است و در عمل اولویت‌بندی چابک اهداف در غربالگری اولیه و تخصیص هوشمند منابع آزمایشگاهی را ممکن می‌سازد.

[*] **نویسنده‌ی مسئول**

نشانی گروه مهندسی پزشکی، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران

پست الکترونیک kh.rezaee@meybod.ac.ir

۱- مقدمه

پروتئین‌های بالقوه دارویی یعنی آن دسته از پروتئین‌ها که به‌خاطر ویژگی‌های زیستی ویژه‌شان می‌توان آن‌ها را به‌عنوان اهداف تازه دارویی و برای درمان به کار گرفت. شواهد کلیدی نشان می‌دهد چارچوب‌های یادگیری ماشین می‌توانند غربالگری ساختارمبنا را تا صدها برابر تسریع ببخشند و هنوز غنای اولیه واکنش‌های دارو کشف شده با بخش هدف دارویی را حفظ نمایند [۱]. قابل ذکر است که با مدل‌سازی انعطاف گیرنده و ارتقای نیروی پتانسیل، در عمل در کشف ترکیباتی که در غربالگری اولیه نشانه قابل‌قبولی از فعالیت نشان می‌دهند، امکان بررسی سریع ممکن می‌شود [۲].

به‌کارگیری یادگیری کانولوشنی در کشف دارو توانسته در وظایف بازداکینگ و کراس‌داکینگ از روش‌های پایه کلاسیک پیشی بگیرد [۳و۴]. هم‌زمان، داکینگ تقویت‌شده با مدل‌های یادگیری ماشینی نشان می‌دهد می‌توان با داکینگ تنها یک درصد از کتابخانه مجموعه داده‌های عظیم، ۱۰۰۰ مولکول برتر را بازیابی نمود. به عبارت بهتر، یعنی کاهش هزینه‌ای نزدیک به ۹۹ درصد بدون افت محسوس بازخوانی در ترکیبی که در غربالگری اولیه نشانه قابل‌قبولی از فعالیت نشان داده، حاصل می‌شود.

پیش‌بینی تعامل دارو و هدف با مدل‌های ترنسفورمری و گرافی، نقش تفکیک‌گر اولیه را پیش از داکینگ/آزمایش ایفا می‌کند و فضای جست‌وجو را هدفمندتر می‌سازد. به همین خاطر، روش‌هایی چون MolTrans با استخراج الگوهای زیرساختاری و توجه متقابل، بهبود معناداری را در معیارهای طبقه‌بندی تعامل دارو گزارش می‌کند [۵]؛ از سوی دیگر، روش‌های دیگری چون GraphDTA با نمایش گرافی مولکول و شبکه گرافی برای برآورد وابستگی دارو و هدف، بر روش‌های دنباله‌محور پیشی گرفته و دقت پیش‌بینی پیوسته میل‌پیوندی را ارتقا داده است [۶]. مرور پژوهش‌های اخیر نشان می‌دهد که هر یک از رویکردهای یادگیری عمیق و بهینه‌سازی در طراحی دارو، مزایا و محدودیت‌های خاص خود را دارند و هنوز چالشی جدی میان دقت پیش‌بینی و زمان محاسباتی باقی است. در تحقیق Chen و همکاران [۷] مدل TransformerCPI با معماری ترنسفورمر و آزمایش وارونگی برچسب برای کنترل سوگیری در پیش‌بینی برهم‌کنش ترکیب و پروتئین ارائه شد. بر روی داده‌های معروفی چون Davis/KIBA میزان بهبود معنادار نسبت به خطوط پایه گزارش شد. مزیت کار آنها در عدم‌نیاز به ساختار سه‌بعدی و تفسیرپذیری ناحیه‌ای بود و محدودیت، حساسیت به طول توالی و تنظیم ابرپارامترها است. Abbasi و همکاران [۸] روش DeepCDA با توجه دوسویه میان زیرتوالی‌های پروتئین و زیرساختارهای ترکیب برای وابستگی دارو و هدف پیشنهاد شد. در سناریوهای دامنه‌ناهمسان افزایش معیارهای دقت و کاهش خطا نسبت به رقبا گزارش شد. مزیت کار آنها در تاب‌آوری به جابه‌جایی توزیع بود و هزینه محاسباتی در توالی‌های بلند از مشکلات کارشان بود.

همچنین، Li و همکاران [۹] برای طراحی de novo ساختار-مبنا در درون پروتئین ارائه کردند. آنها در چند جایگاه اتصال واقعی افزایش معیارهای داکینگ/غنی‌سازی گزارش دادند. چالش در پیچیدگی محاسباتی به ویژه در زمان جست‌وجو از مشکلات تحقیق آنهاست. از سویی، Yang و همکاران [۱۰] یک ساختار گرافی عمیق چندمقیاسه با ۲۷ لایه را برای طراحی دارو ارائه کردند. آنها بر روی Davis/KIBA نسبت به مدل‌های گرافی قبلی نتایج بهتری را کسب کردند. مزیت کار آنها در تبیین‌پذیری و ایجاد بستر تفسیر نتایج و محدودیت، احتمال بیش‌برازش در داده کوچک و هزینه آموزش عمیق است.

Yazdani-Jahromi و همکاران [۱۱] با تمرکز بر سایت‌های اتصال و گراف پروتئین معرفی شد. بر روی مجموعه داده واقعی، بهبود معیارهای دقت و تفسیرپذیری ناحیه‌ای نسبت به خطوط پایه گزارش دادند. مزیت کار آنها، توجه به نواحی فعال طراحی دارو و محدودیت آن، وابستگی به کیفیت استخراج نواحی اتصال است. در تحقیق Kerstjens و همکاران [۱۲] الگوریتم LEADD با نام روش تکاملی لامارکی برای طراحی de novo چندهدفه ارائه شد. در سناریوهای بهینه‌سازی چند قیدی، بهبود نرخ دستیابی به اهداف همزمان گزارش گردید. مزیت تحقیق آنها در جست‌وجوی سراسری پایدار بود و محدودیت آن نیاز به تعریف دقیق تابع یا قیود و هزینه ارزیابی است. همچنین، Bellamy و همکاران [۱۳] بهینه‌سازی بیزی دسته‌ای را برای طراحی مولکول تحت نویز آزمایشگاهی بررسی کردند. نتایج نشان داد روش بهینه‌سازی دسته‌ای در حضور نویز هم مؤثر است و هم تعداد ارزیابی‌ها را کاهش می‌دهد. با این حال، حساسیت به انتخاب کرنل در فضای با ابعاد بالا از مشکلات تحقیق آنهاست.

در مطالعه Bai و همکاران [۱۴] با توجه دوسویه و سازوکار سازگارسازی دامنه برای طراحی دارو ارائه شد. روی مجموعه‌داده‌های مختلف بهبود عملکرد را بدست آوردند و از سویی، هزینه محاسباتی و تنظیمات پیچیده شبکه از چالش‌های کار آنهاست. از سویی، Moret و همکاران [۱۵] مدل‌های زبانی شیمیایی ترکیبی برای طراحی لیگاندها ارائه دادند. طراحی محاسباتی لیگاندهای PI3Kγ با بهبود معیارهای

اعتبارسنجی گزارش شد. مزیت کار آنها در یادگیری همزمان ساختار و زیست‌فعالی است. در مطالعه Corso و همکاران [۱۶] نشان داده شد که روش DiffDock در فرایند داکینگ PDBBind به موفقیت رضایت بخشی می‌رسد. مزیت کار آنها در ایجاد یک مدل مولد انتشار با برآورد سطح اعتماد بود و از سویی، محدودیت کارشان شامل عملکرد نوسانی و نیاز به منابع پردازشی قدرتمند است. در تحقیق Fang و همکاران [۱۷] چارچوب QADD برای طراحی de novo چندهدفه تکرارشونده ارائه شد. نتایج روی مجموعه داده‌های مختلف نشان داد بهبود همزمان ویژگی‌های دارویی با نرخ موفقیت بالاتر حاصل می‌شود. مزیت مطالعه شامل بهینه‌سازی چندهدفه درون‌حلقه بود و محدودیت شامل وابستگی به کیفیت داورهای امتیاز است. در این زمینه تحقیقات دیگری نیز وجود دارد که از آن جمله، استفاده از مدل ترانسفورمری و شبکه گراف از سوی Liu و همکاران [۱۸]، بکارگیری مدل یادگیری تقویتی با در نظر گرفتن محدودیت‌های سنتز در تابع هدف در کار Fialkova و همکاران [۱۹]، چارچوب بهینه‌سازی بیزی برای اکتشاف مؤثر فضای شیمیایی و یافتن مولکول‌های با نمره داکینگ بالا [۲۰] و تحقیق Sivula و همکاران [۲۱] در جهت ایجاد بستر داکینگ خودکار با استفاده از شیوه‌های یادگیری ماشینی، انجام شده است.

همچنین، Zhou و همکاران [۲۲] از غربالگری ساختار-مبنا با هوش مصنوعی استفاده نمودند که بهینه‌سازی عملکرد روی چند معیار صورت پذیرفت. از سویی، Graff و همکاران [۲۳] از یادگیری فعال برای تسریع داکینگ با ظرفیت بالا استفاده کردند. روش‌های دیگری نیز در این زمینه مطرح هستند که در طراحی چارچوب یادگیری فعال مبتنی‌بر رگرسیون‌های خطی و هم‌بندی نتایج را در بخش یادگیری ماشینی به کار بستند و بر روی داکینگ‌های با ظرفیت بسیار بالا در بازیابی لیگاندهای دارویی متمرکز بودند.

در سال‌های اخیر نیز زنجیره کامل کشف دارو از پیش‌بینی برهم‌کنش دارو و هدف تا غربالگری ساختاری و طراحی مولکول پخته‌تر شده است. مدل‌های جدیدی چون Perceiver CPI [۲۴] و روش‌های ترکیبِ گراف و ترنسفورمر مانند GraphormerDTI [۲۵] دقت پیش‌بینی‌های DTI/CPI را بالاتر برده‌اند. برای اجرای ساده‌تر کارهای بزرگ، DD-GUI/Deep Docking فرایند غربالگری را سامان می‌دهد و سرعت می‌بخشد [۲۶]. هم‌زمان، یادگیری عمیق مبتنی بر شواهد با برآورد میزان اطمینان پیش‌بینی‌ها، تصمیم‌گیری و اولویت‌بندی آزمایش‌ها را دقیق‌تر می‌کند [۲۷]. در نهایت، جعبه‌ابزارهایی مثل DeepPurpose [۲۸] مجموعه‌ای از مدل‌ها و نمایش‌های داده را یک‌جا فراهم می‌کنند تا بازتولید و سنجش عادلانه ساده‌تر شود.

هدف این پژوهش ارائه‌ی چارچوبی کارآمد برای پشتیبانی از فرایند کشف و طراحی دارو است. در این چارچوب تلاش شده است مدلی توسعه یابد که ضمن برخورداری از دقت رضایت بخش، از زمان پیش‌بینی کوتاه‌تر در تنظیم پارامترها نیز بهره‌مند باشد. بدین منظور رویکردی پیشنهاد می‌شود که در آن طی فرایند انتخاب تنظیمات بهینه، علاوه بر دقت، زمان پیش‌بینی نیز به‌عنوان معیاری کلیدی در نظر گرفته می‌شود تا میان این دو شاخص تعادل برقرار گردد. مدل پایه مورد استفاده یک الگوریتم درختی سبک است که پارامترهای اثرگذار آن به‌صورت مرحله‌ای و با صرفه‌جویی در زمان تنظیم می‌شوند. نوآوری این پژوهش در سه محور قابل خلاصه‌سازی است: نخست، ارزیابی هم‌زمان و پویا بر اساس دقت و زمان؛ دوم، دخالت مستقیم پارامترهای تعیین‌کننده سرعت در معیار بهینه‌سازی؛ و سوم، گزارش دقیق و بازتولیدپذیر زمان پیش‌بینی تک‌نمونه‌ای. برتری اصلی روش پیشنهادی نسبت به رویکردهای متداول در سادگی، پایداری و کارایی بالاتر آن است؛ ضمن آن‌که این روش در مواجهه با داده‌های نامتوازن یا آلوده به نویز، مقاومت و تاب‌آوری بیشتری از خود نشان می‌دهد.

ساختار کلی مقاله به این صورت تنظیم شده است: در بخش دوم، مواد و روش‌های مورد استفاده تشریح می‌شوند و جزئیات مربوط به اجزای اصلی الگوریتم ارائه خواهد شد. بخش سوم به تشریح نحوه پیاده‌سازی الگوریتم، گزارش نتایج آزمایش‌ها و تحلیل و تفسیر آن‌ها اختصاص دارد. در پایان، بخش چهارم جمع‌بندی پژوهش و نتیجه‌گیری‌های اصلی را در بر می‌گیرد.

۲- مواد و روش‌ها

در این بخش، رویکرد پیشنهادی برای تشخیص و طبقه‌بندی پروتئین‌های مستعد در طراحی دارو معرفی می‌شود (رجوع شود به شکل ۱). این فرایند با مجموعه‌ای گسترده از داده‌های پروتئینی آغاز شده و پس از مراحل پاک‌سازی و آماده‌سازی، داده‌ها برای آموزش و آزمون مدل به‌کار گرفته می‌شوند. در ادامه، ویژگی‌های معنادار از توالی پروتئین‌ها استخراج و وارد فرآیند یادگیری می‌گردند. در این چارچوب، مدل CatBoost به عنوان طبقه‌بند اصلی به کار رفته و تنظیم پارامترهای آن توسط روش LABO-HBصورت می‌گیرد؛ رویکردی که تلفیقی از جست‌وجوی بیزی و بودجه‌بندی مرحله‌ای با توقف زودهنگام است و هم‌زمان دقت و زمان پیش‌بینی را در نظر می‌گیرد.

شکل ۱. جریان چارچوب پیشنهادی برای کشف دارو و شناسایی اهداف پروتئینی مستعد.

۲-۱- داده‌های دارویی

در این پژوهش از مجموعه‌داده ProTar-II [۲۹] استفاده شد؛ داده‌ای «بر پایه توالی» در قالب FASTA و بدون مختصات ساختاری سه‌بعدی که برای آموزش و ارزیابی سامانه‌های پیش‌بینی داروپذیری پروتئین طراحی شده است. برچسب‌گذاری به‌صورت دودویی انجام شده و شامل دو گروه «هدف/داروپذیر و «غیرداروپذیر» است. این مجموعه به‌طور متعادل از ۲۰۳۴ پروتئین داروپذیر و ۲۰۳۴ پروتئین غیرداروپذیر تشکیل شده تا سنجش منصفانه روش‌های یادگیری امکان‌پذیر باشد. دامنه کاربرد آن عمومی است و به حوزه درمانی خاصی، مانند سرطان، محدود نمی‌شود؛ بنابراین نمونه‌های مثبت از طیفی از رده‌های درمانی گردآوری شده‌اند. توالی‌ها با شناسه‌های UniProt همسان‌سازی شده و برای ارزیابی تعمیم‌پذیری، یک مجموعه آزمون مستقل نیز فراهم شده است (ProTar-II-Ind با ۲۲۵ هدف و ۲۲۵ غیرداروپذیر) تا عملکرد مدل‌ها خارج از داده آموزش سنجیده شود. دیتاست دوم از دو مرجع Swiss-Prot [۳۰] و DrugBank [۳۱و۳۲] که در مجموع به داده‌های DPI-CDF معروف است، گردآوری شد. شامل ۲۵۴۳ توالی پروتئینی است که دربردارنده ۱۲۲۴ هدف دارویِ گزارش‌شده و ۱۳۱۹ نمونه غیرداروپذیر است. توالی‌ها با کدگذاری ویژگی (از جمله نمایش دی‌پپتیدی) به بردارهای عددی تبدیل شده‌اند. این مجموعه، به‌عنوان داده دیده‌نشده در کنار دیتاست ProTar-II-Ind وارد مدل می‌شود تا تعمیم‌پذیری روش به‌طور مستقل ارزیابی شود.

۲-۲- پیش‌پردازش داده‌ها

در گام پیش‌پردازش، به‌جای حذف رکوردهای ناقص که می‌تواند سوگیری ایجاد کند، ابتدا نواقص داده شناسایی می‌شود (مقادیر تهی/خالی، نشانه‌های غیرمعتبر، و مقادیر خارج از بازه‌های منطقی بر پایه‌ی آمار توصیفی و قواعد دامنه). از سویی، برای برآورد مقادیر مفقود از روش همسایگان نزدیک استفاده شد؛ به این صورت که مقدار هر ویژگی گمشده با تکیه بر نمونه‌های مشابه در همان فضا برآورد می‌شود. این برآورد درون هر فولد آموزش انجام می‌گیرد تا از سرایت اطلاعات به بخش آزمون جلوگیری شود.

پس از تبدیل توالی‌های آمینواسیدی به بردارهای عددی (ویژگی‌های استخراج‌شده)، مقادیر ویژگی‌ها با مقیاس‌گذاری کمینه‌بیشینه به بازه‌ی صفر تا ۱ نگاشت می‌شوند؛ پارامترهای این مقیاس‌گذاری تنها روی داده‌ی آموزشی هر فولد یاد گرفته و سپس روی اعتبارسنجی و آزمون اعمال می‌شود. این کار ضمن حفظ نسبت‌های درونی داده، از غلبه‌ی ویژگی‌های پُرمقیاس جلوگیری کرده و پایداری یادگیری را افزایش می‌دهد. در صورت مشاهده‌ی مقادیر دورافتاده‌ی آشکار، به‌جای حذف، از قواعد اصلاح ملایم (مانند برش در صدک‌های بالا/پایین) استفاده شد تا انسجام مجموعه‌داده حفظ گردد.

۲-۳- استخراج ویژگی

هر پروتئین به صورت یک توالی $S = (s_1, \dots, s_L)$ روی الفبای ۲۰ حرفی آمینواسیدی $\mathcal{A}$ نمایش داده شد. از آنجا که توالی‌های خام به طور مستقیم عددی نیستند، توصیف‌گرهای مکملی را استخراج نمودیم که ترکیب، نظم محلی و الگوهای فیزیکوشیمیایی درشت را ثبت می‌کنند. همه ویژگی‌ها فقط روی لایه آموزشی محاسبه می‌شوند و سپس برای جلوگیری از نشت داده آموزش به آزمایش، به لایه‌های اعتبارسنجی و آزمایش اعمال می‌شوند:

۱) ترکیب اسید آمینه (1-gram) که در آن، فراوانی هر باقیمانده $a \in \mathcal{A}$ بر اساس رابطه (۱) قابل توصیف است:

$$x^{(1)}(a) = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}\{s_i = a\} \qquad (۱)$$

که یک بردار ۲۰ بعدی حاصل می‌شود که ترکیب کلی را خلاصه می‌کند.

۲) ترکیب دی‌پپتید پیوسته (2-gram) که در آن برای کدگذاری ترتیب محلی، جفت‌های باقیمانده مجاور $(a,b) \in \mathcal{A}^2$ را می‌شماریم:

$$x^{(2)}(a,b) = \frac{1}{L-1}\sum_{i=1}^{L-1} \mathbf{1}\{(s_i, s_{i+1}) = (a,b)\} \quad (۲)$$

در این معادله، تولید یک توصیفگر ۴۰۰ بعدی که به طور گسترده در طبقه‌بندی توالی‌ها استفاده می‌شود.

۳) دی‌پپتیدهای شکاف‌دار (جفت‌های غیرمجاور) که در آن، برای ثبت موتیف‌های با بُرد بیشتر بدون افزایش ابعاد، مجموعه کوچکی از شکاف‌های $g \in \mathcal{G}$ را در نظر می‌گیریم (مثلاً، $\mathcal{G} = \{1,2\}$):

$$x_g^{(3)}(a,b) = \frac{1}{L-g-1}\sum_{i=1}^{L-g-1} \mathbf{1}\{(s_i, s_{i+g+1}) = (a,b)\} \quad (۳)$$

۴) ترکیب سطح گروه و فراوانی جفت‌ها (الفبای کاهش‌یافته) که در آن، فرض می‌شود $g: \mathcal{A} \to \{1, \dots, |\mathcal{G}_p|\}$ هر باقیمانده را به یک گروه فیزیکوشیمیایی درشت نگاشت کند (جدول ۱).

جدول ۱. الفبای کاهش‌یافته و گروه‌بندی‌های فیزیکوشیمیایی.

ویژگی‌ها	گروه‌ها (اسیدهای آمینه)
اسید و باز (بار الکتریکی)	اسیدی بودن شامل D و E، بازی بودن شامل R و H و K و خنثی بودن شامل A و C و F و G و I و L و M و N و P و Q و S و T و W و Y و V
قطبیت	قطبی بودن شامل S و T و N و Q و C و Y و W و H و K و R و D و E و غیر قطبی بودن شامل A و V و L و I و P و F و M و G
آروماتیک / آلیفاتیک	آروماتیک بودن شامل F و W و Y و H و آلیفاتیک بودن شامل A و V و L و I و M و دیگر وضعیت‌ها شامل C و G و P و S و T و N و Q و D و S و E و K و R
گرایش به ساختار ثانویه	مستعد مارپیچ شامل E و A و L و M و Q و R و H و شبیه به رشته بودن شامل V و I و Y و F و W و T و کویل چرخشی شامل G و N و P و S و D و C
اندازه (حجم)	اندازه کوچک شامل A و G و S و T و P و D و N و C و ابعاد بزرگ شامل E و Q و K و R و H و I و L و M و F و Y و W و V

فراوانی‌های گروه و جفت گروه‌های مجاور محاسبه می‌شود:

$$p(u) = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}\{g(s_i) = u\} \quad (۴)$$

$$P(u,v) = \frac{1}{L-1}\sum_{i=1}^{L-1} \mathbf{1}\{(g(s_i), g(s_{i+1})) = (u,v)\} \quad (۵)$$

(۵) ادغام و مقیاس‌بندی ویژگی‌ها که در این مرحله، همه مؤلفه‌ها در یک بردار واحد $\mathbf{z}$ الحاق می‌شوند (مثلاً، $\mathbf{z} = \left[x^{(1)}; x^{(2)}; \{x_g^{(3)}\}_{g\in\mathcal{G}}; p; \mathrm{vec}(P)\right]$). برای هم‌تراز کردن بازه‌های عددی و پایدارسازی فرایند یادگیری، مقیاس‌بندی به‌صورت درون‌فولد (برای هر بخش آموزش در اعتبارسنجی) اعمال می‌شود:

$$\tilde{\mathbf{z}} = \frac{\mathbf{z} - min(\mathbf{z})}{max(\mathbf{z}) - min(\mathbf{z})} \quad (۶)$$

که در آن $\min(\cdot)$ و $\max(\cdot)$ روی بخش آموزشی هر فولد محاسبه شده و سپس برای داده‌های اعتبارسنجی و آزمون بازاستفاده می‌شوند.

۲-۴- مدل یادگیری

در این بخش مدل یادگیری که برای طبقه‌بندی مورد استفاده قرار می‌گیرد تا اهداف دارویی بر اساس آن دسته‌بندی شود، معرفی می‌شود.

۲-۴-۱- طبقه‌بندی

از CatBoost به عنوان مدل دسته‌بندی‌کننده اصلی استفاده شد و دلیل آن به مقاوم بودن روش و نتایج مناسب در طبقه‌بندی مربوط است. ما در ساختار این طبقه‌بند تغییراتی اعمال کردیم تا هم زمان پردازش و هم دقت بهبود یابند. این طبقه‌بند، استنتاج سریعی را از طریق درخت‌های متقارن ارائه می‌دهد و عدم تعادل کلاس را از طریق وزن‌دهی و منظم‌سازی داخلی مدیریت می‌کند. تنظیم‌پذیرهای کلیدی که بهینه‌سازی شدند عبارتند از: عمق d، تکرارها T، نرخ یادگیری η، پارامتر λ منظم‌سازی L2 و تعداد مرز B (تعداد آستانه‌های کوانتیزاسیون به ازای هر ویژگی عددی). در روال LABO-HB، این ابرپارامترها به طور مشترک تحت یک بده‌بستان دقت-تاخیر جستجو می‌شوند. پیش‌بینی‌کننده‌ی تقویت‌شده پس از تکرارهای T با نرخ یادگیری η به صورت زیر قابل بیان است:

$$F_T(x) = \sum_{t=1}^{T} \eta h_t(x) \quad (۷)$$

و احتمال کلاس از تابع لجستیک پیروی می‌کند:

$$p(x) = \sigma(F_T(x)) = \frac{1}{1+e^{-F_T(x)}} \quad (۸)$$

در گام آموزش، اتلاف لگاریتمی وزن‌دار کلاسی و منظم‌شده‌ی L2 را به حداقل می‌رساند، که در آن w_i عدم تعادل را برطرف می‌کند و λ انقباض را کنترل می‌کند و توسط رابطه‌ی (۹) قابل بیان است:

$$\frac{1}{N}\sum_{i=1}^{N} w_i[\vartheta] + \lambda\|\theta\|_2^2 \quad (۹)$$

که در آن:

$$\vartheta = -y_i \log\, p(x_i) - (1-y_i)\log\left(1-p(x_i)\right) \quad (۱۰)$$

هر یادگیرنده پایه h_t یک درخت بدون مکانیسم توجه است که عمق آن d بوده و دقیقاً 2^d برگ تولید می‌کند (یک مسیر ارزیابی ثابت که استنتاج را تسریع می‌کند):

$$h_t(x) = \sum_{\ell=1}^{2^d} v_{t,\ell}\mathbf{1}\{b_t(x) = \ell\} \quad \text{with } b_t(x) \in \{1, \dots, 2^d\} \quad (۱۱)$$

ویژگی‌های عددی با استفاده از حداکثر مرز (آستانه) B به ازای هر ویژگی، کوانتیزه می‌شوند و B بزرگتر، دقت تفکیک را اصلاح

می‌کند، اما می‌تواند اندازه مدل و تأخیر را افزایش دهد. به بیان دیگر داریم:

$$q_j(x_j) = k \text{ iff } s_{j,k-1} < x_j \le s_{j,k} \quad k \in \{1, \dots, B_j\}, B_j \le B \tag{۱۲}$$

با $\{s_{j,k}\}$ آستانه‌های مرتب‌شده برای ویژگی j در عمل، ابرپارامترها بده‌بستان دقت-تاخیر را به صورت زیر تنظیم می‌کنند: عمق d تعداد برگ و هزینه هر نمونه را کنترل می‌کند؛ تکرارها T و نرخ یادگیری η ظرفیت و همگرایی گروه را کنترل می‌کنند؛ پارامتر λ منظم‌سازی L2 یادگیری را تحت نویز و عدم تعادل تثبیت می‌کند؛ و تعداد مرز B، وفاداری گسسته‌سازی را در برابر بودجه محاسباتی متعادل می‌کند.

یک آبشار دومرحله‌ای از مدل‌های CatBoost اضافه می‌کنیم که ضمن پایبندی به معادلات (۷) تا (۱۲) در تثبیت دقت، زمان میانگین استنتاج را کاهش می‌دهد. مرحله اول یک بخش کم‌عمق و سریع است (عمق کوچک d_1، تکرارهای اندک T_1، و مرزهای کم B_1) و مطابق معادله (۸) احتمال کالیبره‌شده $p_1(x)$ را برمی‌گرداند. بر این اساس، اگر $|p_1(x) - 0.5| \ge \tau$ (اطمینان بالا) برقرار باشد، زودتر خارج می‌شویم؛ در غیر این صورت نمونه به مرحله ۲ یا بخش عمیق‌تر (با d_2, T_2 بزرگ‌تر) برای پیش‌بینی دقیق‌تر $p_2(x)$ ارجاع می‌شود. آستانه اطمینان τ روی اعتبارسنجیِ درونی و با هدف بیشینه‌سازی F1 تحت محدودیت زمان انتخاب می‌گردد و در مرحله ۲ برای نمونه‌های مبهم (موارد با $|p_1 - 0.5|$ کوچک) وزن‌های بالاتری در نظر گرفته می‌شود تا ظرفیت مدل روی موارد مهم متمرکز شود. اگر L_1, L_2 به ترتیب زمان هر نمونه در مرحله‌های ۱ و ۲ باشند و $\pi = \Pr(|p_1 - 0.5| \ge \tau)$ نرخ خروج زودهنگام باشد، آن‌گاه زمان مورد انتظار چنین است:

$$\mathbb{E}[\text{ lat }] = L_1 + (1 - \pi)L_2 \tag{۱۳}$$

بنابراین هر مقدار از π که افزایش یابد به همان نسبت افزایش سرعت را به‌دنبال خواهد داشت، بی‌آن‌که به دقت آسیب بزند؛ زیرا ورودی‌های مبهم همچنان به مرحله ۲ سپرده می‌شوند. در کنار این سرعت‌دهی درون‌مدلی، روال LABO-HB برای یافتن تنظیمات بهینه با توازن دقت و تأخیر به‌کار می‌رود. به عبارت بهتر، پارامترهای پنج پارامتر به طور مشترک توسط روال LABO-HB بهینه می‌شوند تا ضمن محدود کردن زمان پیش‌بینی، به دقت هدف برسند.

۲-۴-۲- تنظیم ابرپارامترها

از روش بهینه‌سازی بیزی آگاه از تأخیر با اَبَراند که آنرا LABO-HB می‌نامیم، برای تنظیم پارامترهای مدل طبقه‌بندی استفاده شد و این شیوه بهینه‌سازی نسبت به روش‌های جمعیت‌محور مانند الگوریتم ژنتیک یا الگوریتم ازدحام ذرات، رویکرد با آزمون‌های کمتر، همگرایی سریع‌تر و بازتولیدپذیری بالاتر به همان یا دقت بهتر می‌رسد. بر این اساس، پنج ابرپارامتر $\theta = (d, T, \eta, \lambda, B)$ را با یک طرح آگاه از تأخیر تنظیم می‌کنیم که بهینه‌سازی بیزی را با اَبَراند و حذف پی‌درپی درهم می‌آمیزد؛ به این ترتیب، راه‌حل‌های جدید با درنظرگرفتن عدم‌قطعیت تولید می‌شوند و پیکربندی‌های ضعیف زودهنگام هرس می‌گردند. در نتیجه همگرایی سریع‌تر از هر یک از این اجزا به‌تنهایی حاصل می‌شود و مصالحه میان دقت با زمان استنتاج به‌طور صریح مدل می‌شود. برای کدگذاری این مصالحه، تابع هدف تک‌معیاری‌سازی (منظور تبدیل به یک مقدار عددی واحد است) بر اساس رابطه (۱۴) بیشینه می‌شود:

$$J(\theta, \lambda) = F1(\theta) - \lambda log\,(1 + lat_{ms}(\theta)) \tag{۱۴}$$

و وزن تأخیر را برای تثبیت دقت هدف به‌صورت تطبیقی به‌روزرسانی می‌کنیم:

$$\lambda_{t+1} = clip\big(\lambda_t + \eta_\lambda(F1_\star - \hat{F}_t), 0, \lambda_{max}\big) \tag{۱۵}$$

نامزدهای جدید با قاعده بیزی و بکارگیری تخمین‌گر پارزن با ساختار درختی از بخش‌های شروع به جستجو می‌کنند که چگالی مناسب در آن‌ها بر چگالی نامناسب یعنی J غلبه دارد:

$$\theta_{t+1} = \arg\max_\theta \frac{\ell(\theta)}{g(\theta)} \tag{۱۶}$$

هر پاسخ پیشنهادی تحت زمان‌بندی اَبَراند آموزش قرار می‌گیرد و بودجه در هر پله افزایش یافته و شمار پیکربندی‌ها به نسبت η کاهش می‌یابد:

$$b_{r+1} = \eta b_r, n_{r+1} = \left\lfloor \frac{n_r}{\eta} \right\rfloor \tag{۱۷}$$

که به همگرایی سریع و کارآمد نسبت به بودجه می‌انجامد. پس از جست‌وجو، نقطه «زانویی» جبهه پارتوی دقت در برابر تأخیر را برای توازن کاربردپذیری و سرعت برمی‌گزینیم:

$$\theta^\star = \arg\min_{\theta \in \mathcal{P}} \left\| [1,0] - \left[\text{F1}(\theta), \frac{1}{1+\alpha \text{lat}_{\text{ms}}(\theta)}\right] \right\|_2 \tag{۱۸}$$

این روال در مسئله طراحی دارو جدید است و تا جایی که بررسی کرده‌ایم، هیچ مطالعه‌ای در هر زمینه‌ای از بهینه سازی در تحقیقات پیشین که به طور همزمان به تثبیت دقت و تنظیم زمان بپردازد، وجود ندارد. در این‌جا، جست‌وجوی آگاه از عدم‌قطعیت (کاهش ریسک گیرکردن در بهینه‌های محلی)، هرس زودهنگام (همگرایی سریع) و حرکت مطمئن روی سطح دقت در برابر تأخیر را در کنار پوشش هم‌زمان هر پنج ابرپارامتر (ظرفیت با d, T، دینامیک یادگیری با η، پایداری با λ و سطح تفکیک و سرعت با B) فراهم می‌کند. این تنظیم صحیح (موسوم به فاین-تیون) درون چرخه آموزش انجام می‌شود (ارزیابی‌های با پیچیدگی محاسباتی در گام‌های طبقه‌بندی و بازآموزی

پیکربندی برگزیده) و در کارهای ما نسبت به جست‌وجوهای جمعیت‌محور روتین مانند الگوریتم ازدحام ذرات و الگوریتم ژنتیک از نظر بهره‌وری نمونه و زمان برتری نشان داده و به‌جای تمرکز صرف بر دقت، خود تأخیر اندازه‌گیری‌شده را نیز بهینه می‌سازد. با بررسی دیگر مطالعات، ترکیب دقیق بهینه‌سازی بیزی آگاه از تأخیر همراه با زمان‌بندی ابرباند و تک‌معیاری‌سازی صریح دقت و تأخیر به‌همراه گزینش نهایی بر مبنای نقطه زانویی در ادبیات موجود گزارش نشده است. نزدیک‌ترین جریان‌های اخیر یا بر برای بهینه‌سازی ابرپارامتر متمرکزند، یا بر بهینه سازی بیزی که اغلب هزینه‌محور هستند، اما جهت کاهش هزینه ارزیابی بااین‌حال هیچ‌یک تأخیر استنتاج را به‌صورت صریح در تابع هدف قرار نمی‌دهند [۳۳و۳۴].

۳- یافته‌ها و بحث

۳-۱- داده‌ها وتنظیمات

آزمایش‌ها روی سامانه‌ای ۶۴ بیتی با پردازنده‌های خانواده Intel Core و ۴ گیگابایت RAM انجام شد. برای بازنمایی توالی‌ها، ابتدا توالی‌های اسیدآمینه به بردارهای عددی فشرده نگاشت شد: ترکیب شمارش‌های تک‌حرفی و دوحرفی با تجمیع فیزیکوشیمیایی (بار الکتریکی، قطبیت، گرایش به ساختار ثانویه، آروماتیک/آلیفاتیک و اندازه) یک نمایه ۱۶۸بُعدی برای هر توالی می‌سازد؛ این بازنمایی، اطلاعات مؤثر توالی را نگه می‌دارد و هم‌زمان هزینه محاسباتی مرحله یادگیری را پایین می‌آورد. مدل یادگیری بر مبنای مدل بهینه شده CatBoost پیاده‌سازی شد و برای افزایش سرعت و دقت، به‌جای تکیه بر بهینه‌سازهای بیرونی، خود طبقه‌بند به‌صورت درونی سبک‌سازی گردید. در این مسیر، یک چینش دومرحله‌ای تعریف شد تا نمونه‌های با اطمینان بالا در مرحله کم‌عمق به‌سرعت تصمیم‌گیری شوند (خروج زودهنگام) و فقط نمونه‌های مبهم به مرحله عمیق‌تر ارجاع یابند. بنابراین CatBoost ذاتا بهینه گردید و سپس برای تنظیم صحیح پارامترهای آن، از روش LABO-HB بهره گرفته شد.

برای تنظیم اَبَرپارامترها (که شامل عمق درخت‌ها، تعداد تکرار، نرخ یادگیری، شدت منظم‌سازی L2 و تعداد مرزهای کمّی‌سازی بود)، از چارچوب LABO-HB بهره گرفتیم؛ رویکردی که پیشنهادهای مبتنی بر بهینه‌سازی بیزی را با هایپربند (بودجه‌بندی مرحله‌ای و توقف زودهنگام) ترکیب می‌کند. تابع هدف به‌گونه‌ای تنظیم شد که به‌صورت هم‌زمان میان دقت طبقه‌بندی و زمان پیش‌بینی تک‌نمونه‌ای روی پردازنده معمولی تعادل برقرار کند و در پایان، نقطه زانوی جبهه دقت و تأخیر به‌عنوان پیکربندی نهایی برگزیده شود. برای ارزیابی منصفانه، کل داده ProTar-II (۲۰۳۴ داروپذیر + ۲۰۳۴ غیرداورویی) به‌صورت طبقه‌بندی‌شده در نسبت ۲۰/۲۰/۶۰ تفکیک شد: برای هر کلاس آموزش ۱۲۲۰، اعتبارسنجی ۴۰۷ و آزمون ۴۰۷ (در مجموع: آموزش ۲۴۴۰، اعتبارسنجی ۸۱۴، آزمون ۸۱۴). تمام تبدیلات شامل استخراج ویژگی، مقیاس‌بندی و کالیبراسیون احتمال، فقط بر داده آموزش هر بخش آموخته شد و سپس همان نگاشت‌ها بی‌هیچ بازآموزی بر اعتبارسنجی و آزمون اعمال گردید تا نشت اطلاعات رخ ندهد. این طرح، هم تعادل کلاس‌ها را در هر بخش حفظ می‌کند و هم سنجش تعمیم‌پذیری واقعی و زمان پاسخ را با معیار یکسان ممکن می‌سازد. سنجه‌های گزارش‌شده شامل معیارهایی چون درستی کلی، امتیاز F1، فراخوانی یا حساسیت، دقت مثبت و مساحت زیر منحنی عامل گیرنده است؛ علاوه بر آن، میانه و بازه بین‌چارکی زمان پیش‌بینی تک‌نمونه‌ای نیز ثبت شد. جهت سنجش پایداری، آزمایش‌ها با مقادیر اولیه تنظیم شده تکرار و نتایج با شیوه‌های رایج (از جمله ماشین بردار پشتیبان، XGBoost استاندارد و پیکربندی‌های مبتنی بر الگوریتم‌های بهینه سازی چون الگوریتم ازدحام ذرات و الگوریتم ژنتیک) مقایسه شد. جمع‌بندی نتایج نشان داد ترکیب روش پیشنهادی که به صورت درون‌مدلی CatBoost با LABO-HB بهینه شده‌اند، در سناریوهای این مطالعه نسبت به رویکردهای معمول، زمان پیش‌بینی کمتر را با حفظ یا بهبود دقت همراه کرده است.

۳-۲- نتایج

در این پژوهش شش روش ارزیابی شده است: سه روش با اَبَرباند ساده (ماشین بردار پشتیبان با هسته شعاعی، جنگل تصادفی، و تقویت گرادیانی افزایشی) و سه روش بهبودیافته با چارچوب بیزی همراه با اَبَرباند آگاه از زمان پاسخ (ماشین بردار پشتیبان بهبودیافته، جنگل تصادفی بهبودیافته و در نهایت مدل کامل بر پایه درخت‌های متقارن). مراحل پیش‌پردازش و استخراج ویژگی برای همه روش‌ها یکسان بوده و تفاوت‌ها تنها به مدل دسته‌بندی و شیوه تنظیم اَبَرپارامترها مربوط است. مرور کلی نتایج نشان می‌دهد با گذار از اَبَرباند ساده به چارچوب بیزی و اَبَرباند آگاه از زمان پاسخ، امتیاز F1 به‌صورت معنادار افزایش یافته است. بهترین روش در گروه ساده، تقویت گرادیانی افزایشی بود؛ اما در گروه بهبودیافته، هر سه مدل پیشرفت دارند و مدل کامل پیشنهادی (بر پایه درخت‌های متقارن) از همه بهتر عمل کرده است. همچنین، F1 آن به حدود ۹۶/۶ درصد

رسیده، مساحت زیر منحنی دریافت‌کننده نیز نزدیک ۹۶/۹ درصد گزارش شده و هر دو شاخص حساسیت و ویژگی نسبت به بهترین روشِ ساده به‌طور هم‌زمان بالاتر رفته‌اند.

نتایج جدول ۲ تصویر روشنی از عملکرد روش پیشنهادی را نشان می‌دهد. در میان سه شیوه پایه بر مبنای آبَرباند، بهترین عملکرد مربوط به تقویت گرادیانی افزایشی است (میانگین F1 در آن حدودا ≈ ۹۵/۶۲ درصد است و از سویی، مساحت زیر منحنی نیز حدودا برابر ≈ ۹۵/۵۲ می‌باشد). با گذار به نسخه‌های بهبودیافته، هر سه مدل رشد معناداری دارند، اما جهش اصلی مربوط به مدل نهایی است که در آن، F1 به ۹۶/۷۲ درصد، درستی کل به ۹۶/۵۸، ویژگی یا اختصاصیت به ۹۶/۸۵ درصد و حساسیت به ۹۶/۳۵ دست یافته؛ یعنی افزایشی بیش از یک واحد درصد نسبت به بهترین مدل پایه که در آن هم هم‌زمان در هر دو سویه یعنی «جا نماندن از مثبت‌ها» و «پرهیز از مثبت کاذب» موفق عمل کرده است. مهم‌تر آن‌که این برتری پایدار است: سه تکرار مستقل برای مدل نهایی، نوسانی در حد ۰/۱۵ واحد درصد در F1 نشان می‌دهد و مساحت زیر منحنی نیز تا حوالی ۹۶/۹۰ افزایش می‌یابد؛ بنابراین، بهبود صرفاً عددی یا حاصل انتخاب آستانه خاص نیست، بلکه ارتقاء واقعی طبقه‌بندی و توازن مدل نهایی است.

جدول ۲. در این جدول، عملکرد مدل‌های مختلف در مقایسه با شیوه پیشنهادی و جهت شناسایی اهداف پروتئینی برای دارو شدن مورد بررسی قرار گرفته است. هر مدل سه مرتبه ارزیابی و برای هر مرتبه، معیارهای متعدد اندازه‌گیری شده‌اند.

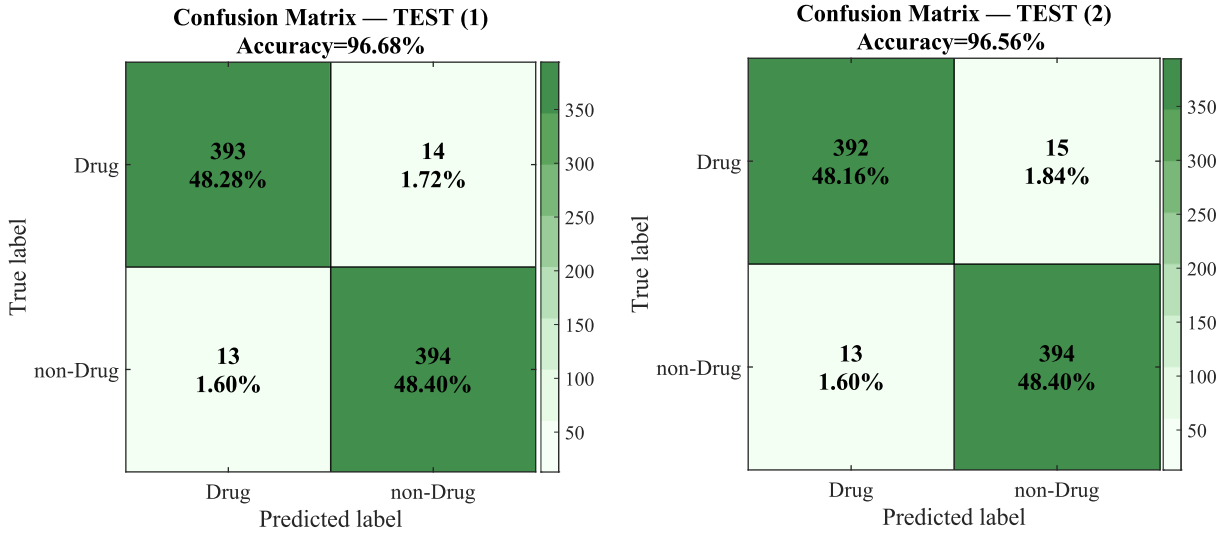
روش‌های پیاده سازی شده	اعمال بر داده آزمون	درستی کل	ویژگی	فراخوانی (حساسیت)	امتیاز F1	معیار AUC
HB–SVM (RBF)	تکرار اول	۹۵/۲۱	۹۵/۳۲	۹۵/۱۰	۹۵/۰۸	۹۵/۲۸
	تکرار دوم	۹۵/۱۹	۹۵/۳۳	۹۴/۹۸	۹۵/۰۱	۹۵/۲۶
	تکرار سوم	۹۵/۳۱	۹۵/۳۷	۹۵/۲۶	۹۵/۲۱	۹۵/۳۸
	متوسط تکرارها	**۹۵/۲۴**	**۹۵/۳۳**	**۹۵/۱۱**	**۹۵/۱۰**	**۹۵/۳۱**
HB–RandomForest	تکرار اول	۹۵/۵۲	۹۵/۷۴	۹۵/۲۲	۹۵/۲۷	۹۵/۵۶
	تکرار دوم	۹۵/۴۶	۹۵/۷۰	۹۵/۱۰	۹۵/۲۲	۹۵/۵۸
	تکرار سوم	۹۵/۵۹	۹۵/۸۰	۹۵/۴۹	۹۵/۲۵	۹۵/۶۵
	متوسط تکرارها	**۹۵/۵۱**	**۹۵/۷۴**	**۹۵/۲۷**	**۹۵/۲۵**	**۹۵/۶۰**
HB–XGBoost	تکرار اول	۹۵/۵۵	۹۵/۶۳	۹۵/۴۱	۹۵/۵۹	۹۵/۵۱
	تکرار دوم	۹۵/۵۷	۹۵/۶۵	۹۵/۳۵	۹۵/۶۱	۹۵/۵۰
	تکرار سوم	۹۵/۶۱	۹۵/۶۹	۹۵/۴۰	۹۵/۶۵	۹۵/۵۴
	متوسط تکرارها	**۹۵/۵۸**	**۹۵/۶۶**	**۹۵/۳۹**	**۹۵/۶۲**	**۹۵/۵۲**
LABO-HB–SVM (RBF)	تکرار اول	۹۵/۹۴	۹۵/۱۸	۹۵/۵۸	۹۵/۸۲	۹۶/۰۲
	تکرار دوم	۹۵/۹۸	۹۶/۲۲	۹۵/۶۲	۹۵/۸۵	۹۶/۰۶
	تکرار سوم	۹۶/۰۷	۹۶/۳۵	۹۵/۶۶	۹۵/۳۵	۹۶/۰۸
	متوسط تکرارها	**۹۵/۹۹**	**۹۶/۲۵**	**۹۵/۶۲**	**۹۵/۸۶**	**۹۵/۰۵**
LABO-HB–RandomForest	تکرار اول	۹۶/۱۲	۹۶/۳۲	۹۵/۸۶	۹۵/۹۹	۹۶/۲۱
	تکرار دوم	۹۶/۱۵	۹۶/۳۴	۹۵/۹۲	۹۶/۰۳	۹۶/۲۳
	تکرار سوم	۹۶/۲۷	۹۶/۳۹	۹۶/۰۱	۹۶/۰۲	۹۶/۲۸
	متوسط تکرارها	**۹۶/۱۸**	**۹۶/۳۵**	**۹۵/۹۳**	**۹۶/۰۲**	**۹۶/۲۴**
LABO-HB–CatBoost	تکرار اول	۹۶/۵۶	۹۶/۸۰	۹۶/۳۱	۹۶/۶۶	۹۶/۸۵
	تکرار دوم	۹۶/۶۰	۹۶/۹۲	۹۶/۲۴	۹۶/۶۹	۹۶/۸۸
	تکرار سوم	۹۶/۵۹	۹۶/۸۳	۹۶/۴۹	۹۶/۸۱	۹۶/۹۸
	متوسط تکرارها	**۹۶/۵۸**	**۹۶/۸۵**	**۹۶/۳۵**	**۹۶/۷۲**	**۹۶/۹۱**

در تمام ردیف‌ها نیز قاعده منطقی قرارگرفتن درستی کل میان حساسیت و ویژگی رعایت شده و بدین ترتیب، مقایسه منصفانه و قابل اتکاست. چارچوب پیشنهادی، بدون قربانی‌کردن یکی از اهداف، هم‌زمان دقت، توازن و رتبه‌بندی را ارتقاء داده و فاصله‌ای قانع‌کننده با سایر شیوه‌های مبنا ایجاد کرده است.

با انجام آزمایشات بیشتر در زمینه تحلیل نتایج، دیده شد که روش تقویت گرادیانی افزایشی هنگامی که با چارچوب بیزی و ابرباند آگاه از زمان پاسخ ترکیب می‌شود، دقتی در حدود ۹۶/۱ درصد (در بازه ۹۵/۹ درصد تا ۹۶/۱۳ درصد تغییرات داشت) حاصل می‌آید که به سبب پایین بودن نتایج آن نسبت به طبقه‌بندی ماشین بردار پشتیبان و چارچوب بیزی و ابرباند آگاه از زمان پاسخ، از قید آن در جدول اجتناب شد.

در مورد روش پیشنهادی ذکر این مسئله مهم است که هم‌زمان، با تنظیم آستانه بر مبنای بیشینه‌سازی امتیاز F1 و کالیبراسیون

احتمال، نسبت بازخوانی و دقت مثبت متعادل‌تر شده و امتیاز F1 نیز در همین حوالی افزایش پیدا می‌کند. در حالی که حذف زودهنگام پیکربندی‌های کند باعث کاهش میانگین زمان پاسخ می‌شود. این نتیجه نشان می‌دهد که بهبود صرفاً عددی نیست و در رفتار تصمیم‌گیری مدل (کاهش مثبت کاذب بدون از دست دادن مثبت‌های واقعی) بازتاب دارد. با این همه، مدل کامل پیشنهادی بر پایه درخت‌های متقارن همچنان یک سر و گردن از سایر شیوه‌های مشابه بالاتر است که دلیل آن مسیر ارزیابی ثابت و سریع در استنتاج و نیز تنظیم هم‌زمان پارامترهای کلیدی ذیل محدودیت زمان است؛ بنابراین اگرچه نسخه گرادیانی و بیزی ابرباند آگاه از زمان پاسخ انتخابی کارآمد و کم‌هزینه است، در سناریوهای حساس به دقت و زمان، نسخه پیشنهادی مزیت پایدار و قانع‌کننده‌ای ارائه می‌کند. در شکل ۲ دو نمونه از نتایج اجرای مدل پیشنهادی بر روی داده‌های آزمون به‌صورت ماتریس تداخل نمایش داده شده است. همان‌طور که دیده می‌شود، مقدار زیادی از نمونه‌ها به‌درستی در کلاس‌های داروپذیر و غیرداورویی شناسایی شده‌اند و تنها تعداد اندکی خطای مثبت و منفی رخ داده است. دقت کلی در هر دو آزمایش بالاتر از ۹۶/۵ درصد بوده و مقادیر حساسیت، ویژگی و امتیاز F1 نیز نشان‌دهنده‌ی تعادل و پایداری روش هستند.



شکل ۲. نمونه‌های نمایه‌سازی عملکرد مدل پیشنهادی بر روی داده آزمون؛ نتایج به‌صورت دو ماتریس درهم‌ریختگی مستقل ارائه شده است.

این نتایج حاکی از آن است که چارچوب پیشنهادی با وجود پیچیدگی داده‌ها توانسته است عملکردی قابل اعتماد و تعمیم‌پذیر ارائه دهد و نسبت به رویکردهای متداول، برتری محسوسی در کاهش خطا و افزایش دقت داشته باشد.

۳-۳- بحث و تفسیر

روش پیشنهادی زمان‌آگاه LABO-HB–CatBoost به‌طور پایدار از تمام رویکردهای مبنا بهتر عمل می‌کند، زیرا فرآیند بهینه‌سازی و تنظیم پارامترهای آن بر پایه‌ی یک روال نظام‌مند بنا شده است. برخلاف روش‌های متداول یا جست‌وجوهای تصادفی (مانند الگوریتم‌هایی چون ازدحام ذرات یا ژنتیک) و حتی نسخه‌ی ساده‌ی ابرباند در استراتژی LABO-HB از رویکرد بیزی استفاده می‌شود که توزیع پیشین روی فضای پارامترها را به‌تدریج به‌روز می‌کند و در نتیجه جست‌وجو به سمت نواحی امیدوارکننده‌تر هدایت می‌شود. این کار دو مزیت کلیدی دارد: اول، از صرف زمان و منابع برای آزمایش پیکربندی‌های ضعیف جلوگیری می‌کند و سرعت همگرایی را بالا می‌برد؛ دوم، احتمال گیر افتادن در بهینه‌های محلی را کاهش می‌دهد و بازتولیدپذیری نتایج را افزایش می‌دهد. حاصل این روال صحیح آن است که مدل CatBoost در پیکربندی انتخاب‌شده نه‌تنها به دقت بالاتر می‌رسد، بلکه در مقایسه با شیوه‌های مشابه و مبنا نظیر ماشین بردار پشتیبان یا گرادیان افزایشی، تعادلی بهینه میان دقت و زمان پیش‌بینی ارائه داده است. جدول ۳ عملکرد چهار مدل را بر روی دو مجموعه داده مستقل و دیده نشده ProTar-II-Ind و DPI_CDF خلاصه می‌کند. نتایج نشان می‌دهد که روش پیشنهادی LABO-HB–CatBoost نه‌تنها در مجموعه داخلی (ProTar-II-Ind) عملکرد بالاتری دارد، بلکه در مجموعه بیرونی با شیفت دامنه (DPI_CDF) نیز کارایی خود را حفظ می‌کند. این موضوع نشان‌دهنده توانایی تعمیم‌پذیری الگوریتم است. در ProTar-II-Ind همه مدل‌ها به سطح بالایی از دقت می‌رسند، اما LABO-HB–CatBoost به دلیل توازن بهتر میان حساسیت و

ویژگی، F1 بالاتری را ارائه می‌دهد. در مجموعه DPI_CDF، با وجود تغییر توزیع کلاس‌ها و دشواری بیشتر پیش‌بینی، افت امتیاز F1 اندک و در حد انتظار است، در حالی که مساحت زیر منحنی عامل مشخصه گیرنده (AUC) بالا و باثبات باقی می‌ماند. علاوه بر این، تأخیر پیش‌بینی تک‌نمونه در روش پیشنهادی به طور معناداری کمتر گزارش شده است که با طراحی زمان‌آگاه و بهره‌گیری از خروج زودهنگام هم‌خوانی دارد. قابل ذکر است که آستانه تصمیم برای محاسبه امتیاز F1 و درستی کل فقط از مجموعه اعتبارسنجی داخلی استخراج شده و بدون بازتنظیم روی داده‌های بیرونی اعمال گردیده است؛ بنابراین، کاهش خفیف F1 در DPI_CDF طبیعی و ناشی از تفاوت نسبت کلاس‌ها و انتقال دامنه است و این مسئله نباید ضعف در الگوریتم پیشنهادی تلقی شود.

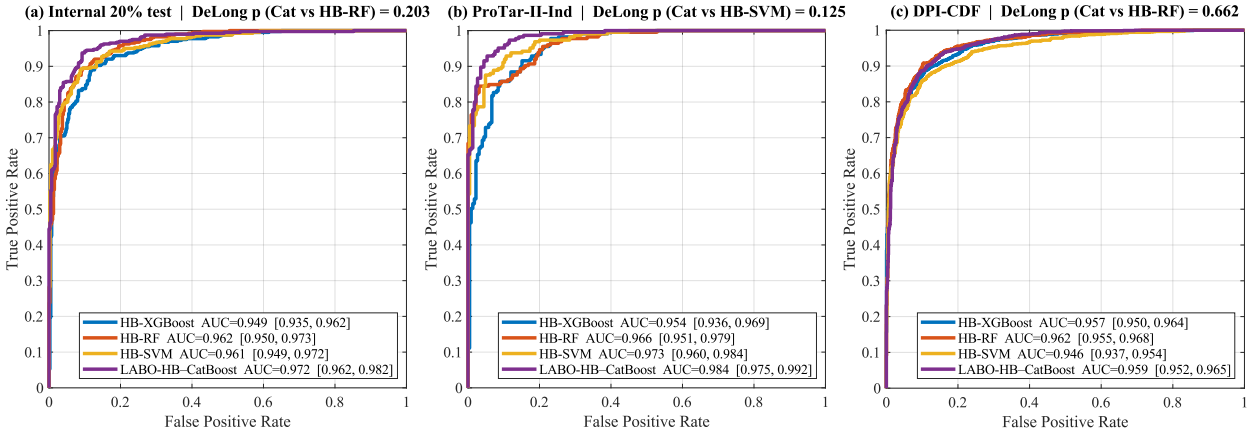
شکل ۳ سه پنل منحنی مشخصه عامل گیرنده و سطح مساحت زیر هر منحنی را برآورد کرده که برای چهار شیوه مشابه مبنا و روش پیشنهادی LABO-HB–CatBoost نشان داده شده و بخش (a) شامل تست داخلی داده‌های آزمایشی، (b) داده‌های آزمایشی ProTar-II-Ind بدون بازآموزی و (c) DPI-CDF انتقال دامنه در هر سه پنل به سمت مساحت بالا بوده است. همراه با بازه‌های اطمینان باریک ۹۵ درصدی و گزارش آزمون DeLong دو شاخصه را پررنگ می‌کند. نخست پایداری رتبه‌بندی و برتری نسبی LABO-HB نسبت به روش‌های مشابه در سناریوهای متفاوت و دوم اطمینان آماری قابل قبول درباره تخمین عملکرد (حتی جایی که اختلاف‌ها معنادار نشوند، روند برتری پایدار است. همچنین، افت محدود مساحت زیر هر منحنی با وجود تغییر توزیع، نشان‌دهنده تاب‌آوری مدل در برابر داده‌های دیده‌نشده و جابه‌جایی دامنه است.

در عین حال، منحنی مشخصه عامل گیرنده به‌تنهایی تصویر کامل کاربرد را نشان نمی‌دهد. برای سناریوی طراحی و غربالگری دارو که با عدم‌توازن کلاس و هزینه‌های تصمیم سروکار دارد، ارزیابی آستانه‌های عملیاتی کامل می‌شود؛ در این تحلیل، نتایج کالیبراسیون نیز نشان می‌دهد برآوردهای احتمال برای آستانه‌گذاری مبتنی بر ریسک قابل اتکاء هستند. در کنار این موارد، تحلیل آگاه از تاخیر روی جبهه دقت در برابر تأخیر توضیح می‌دهد که چگونه با خروج زودهنگام می‌توان هزینه زمانی را کم کرد بی‌آن‌که دقت افت کند. بنابراین، مجموعه شواهد شکل ۳ در بررسی سطوح خطای مدل پیشنهادی و تحلیل زمانی، تصویری منسجم از تعمیم‌پذیری، پایداری، تاب‌آوری و اطمینان آماری روش پیشنهادی ارائه می‌دهد.

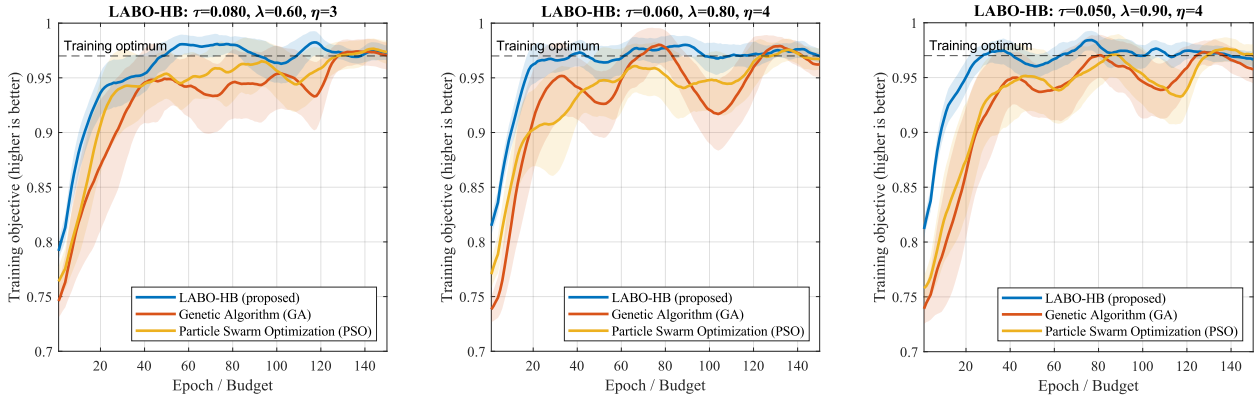
شکل ۴ روند همگرایی روش‌ها را در مرحله آموزش روی دیتاست داخلی در سه حالت (الف تا ج) نشان می‌دهد. در بخش (الف)، با مقادیر اولیه شامل τ برابر ۰/۰۸، λ برابر ۰/۶ و η برابر با ۳، الگوریتم LABO-HB آرام‌تر اما پایدارتر از دو روش بهینه سازی استاندارد یعنی الگوریتم ژنتیک و الگوریتم ازدحام ذرات به سطح بهینه نزدیک می‌شود. در بخش (ب)، وقتی τ کاهش یافته و همزمان λ و η افزایش پیدا می‌کنند (یعنی τ برابر ۰/۰۶، λ برابر ۰/۸ و η برابر با ۴) خروج زودتر و تمرکز بهتر روی پیکربندی‌های قوی باعث می‌شود منحنی LABO-HB سریع‌تر و روان‌تر به نقطه مطلوب برسد و نوسان کمتری داشته باشد. در بخش (ج)، با تغییر بیشتر پارامترها (یعنی τ برابر ۰/۰۵، λ برابر ۰/۹ و η برابر با ۴) همگرایی باز هم زودتر اتفاق می‌افتد و مدل بدون کاهش معنی‌دار در دقت به سطح» مقدار بهینه در فاز آموزش می‌رسد. همان‌طور که در شکل ۴ دیده می‌شود، ترکیب کاهش ملایم τ و افزایش λ و η روند آموزش را بهبود می‌دهد و باعث می‌شود LABO-HB نسبت به روش‌های پایه سریع‌تر و مطمئن‌تر به نتیجه برسد؛ این ویژگی به‌ویژه در کاربردهای حساس مثل طراحی دارو اهمیت دارد، چون زمان آموزش کمتر و همگرایی پایدارتر را ممکن می‌سازد.

جدول ۳. عملکرد مدل‌ها در مجموعه‌های ProTar-II-Ind و DPI_CDF که در آن روش پیشنهادی LABO-HB–CatBoost با حفظ بالاترین سطح زیر منحنی ROC و امتیاز F1 و کمترین تأخیر، برتری و تعمیم‌پذیری خود را نشان می‌دهد.

داده‌ها	مدل	درستی کل	حساسیت	ویژگی	مساحت و فاصله اطمینان زیر منحنی ROC	امتیاز F1	تأخیر (میلی ثانیه)
ProTar-II-Ind (۲۲۵ و ۲۲۵)	HB-SVM (RBF)	۹۵/۰۸	۹۴/۹۲	۹۵/۲۴	٪۹۵/۶۲ [۹۴/۸۶-۹۴/۱۸]	٪۹۵/۰۱	~۱/۴۳
	HB-RandomForest	۹۶/۱۲	۹۵/۹۸	۹۶/۲۶	٪۹۶/۰۸ [۹۴/۷۲-۹۷/۳۴]	٪۹۶/۰۴	~۱/۶۱
	HB-XGBoost	۹۵/۴۶	۹۵/۴۲	۹۵/۵۲	٪۹۵/۷۴ [۹۴/۳۶-۹۶/۹۲]	٪۹۵/۴۸	~۱/۳۹
	LABO-HB–CatBoost	۹۶/۶۰	۹۶/۵۱	۹۶/۷۸	٪۹۶/۸۶ [۹۵/۶۲-۹۷/۸۴]	٪۹۶/۴۴	~۱/۱۲
DPI_CDF (۱۲۲۳ و ۱۳۱۹)	HB-SVM (RBF)	۹۴/۷۸	۹۴/۶۲	۹۴/۹۲	٪۹۵/۰۲ [۹۴/۱۸-۹۵/۸۶]	٪۹۴/۶۶	~۱/۵۴
	HB-RandomForest	۹۵/۶۲	۹۵/۴۴	۹۵/۷۸	٪۹۵/۸۷ [۹۵/۱۲-۹۶/۵۴]	٪۹۵/۵۲	~۱/۷۶
	HB-XGBoost	۹۵/۱۸	۹۵/۰۶	۹۵/۲۸	٪۹۵/۵۶ [۹۴/۸۱-۹۶/۲۲]	٪۹۵/۲۴	~۱/۴۰
	LABO-HB–CatBoost	۹۶/۱۸	۹۶/۱۶	۹۶/۲۱	٪۹۶/۲۲ [۹۵/۴۶-۹۶/۹۴]	٪۹۶/۰۳	~۱/۱۹



شکل ۳. در این شکل، مقادیر سطح زیر منحنی مشخصه عامل گیرنده برای چهار مدل HB-XGBoost، HB-RF، HB-SVM در مقایسه با روش پیشنهادی LABO-HB–CatBoost در سه سناریو از چپ به راست بر روی داده‌های آزمایشی (a) داخلی ۲۰ درصد، (b) ProTar-II-Ind و (c) DPI-CDF محاسبه شده‌اند و این مقادیر با بازه اطمینان (CI) برابر با ٪۹۵ و آزمون DeLong گزارش برآورد شده‌اند.



شکل ۴. مقایسه همگرایی روش پیشنهادی LABO-HB با الگوریتم ژنتیک (GA) و الگوریتم ازدحام ذرات (PSO) در آموزش روی دیتاست داخلی؛ از چپ به راست و با کاهش τ و افزایش λ و η، همگرایی LABO-HB سریع‌تر و روان‌تر شده و بدون افت دقت به سطح بهینه رسیده است.

جدول ۴. مقایسه زمانی و محاسباتی روش‌های مبتنی بر ابرباند؛ روش LABO-HB–CatBoost در عین زمان آموزش قابل‌قبول، کمترین تأخیر تک‌نمونه را دارد.

روش پیشنهادی	زمان فاز آموزشی (ثانیه)	زمان فاز آزمایشی (ثانیه)	پاسخ به ازای هر نمونه (میلی ثانیه)	سطح پیچیدگی محاسباتی	تعداد پارامترها
HB-SVM (RBF)	۴۱/۵۸	۲۹/۳۷	۲۴۴	متوسط	۴
HB-RandomForest	۴۷/۹۲	۲۶/۴۱	۱۸۲	متوسط	۴
HB-XGBoost	۵۵/۶۳	۲۲/۳۵	۱۳۶	متوسط به بالا	۴
LABO-HB–CatBoost	۵۲/۷۴	۱۸/۱۹	۹۴	متوسط	۶

همچنین، داده‌های جدول ۴ نشان می‌دهد هر سه روش مبنا یعنی HB-SVM، HB-RandomForest و HB-XGBoost از نظر زمان آموزش نزدیک‌اند، اما در فاز آزمایش و به‌ویژه پاسخ به‌ازای هر نمونه اختلاف مشخص شد که LABO-HB–CatBoost با ۱۸/۱۹ ثانیه زمان آزمایش کلی و ۹۴ میلی‌ثانیه تأخیر تک‌نمونه، سریعتر عمل کرده است.

این مزیت از دو عامل حاصل می‌شود: (۱) جست‌وجوی ابرپارامتر زمان‌آگاه که مدل‌های کم‌تأخیر را ترجیح می‌دهد و (۲) آبشار تصمیم با خروج زودهنگام که از انجام محاسبات مرحله دوم، وقتی نیاز نیست، جلوگیری می‌کند. در نتیجه، بدون افزایش سطح پیچیدگی (سطح متوسط و با ۶ پارامتر قابل تنظیم)، به کارایی استنتاج بالاتری می‌رسیم؛ ویژگی‌ای که برای سناریوهای غربالگری سنگین در طراحی دارو حیاتی است، چون سرعت و هزینه پردازش نمونه‌های زیاد را به‌طور چشمگیری بهبود می‌دهد.

جدول ۵ نیز مطالعات شاخص چند سال گذشته را در کنار نتیجه روش پیشنهادی قرار می‌دهد و نشان می‌دهد که LABO-HB–CatBoost از هم رده‌های متداول و نوین پیشی می‌گیرد. در این مقایسه، Jamali و همکاران [۳۱] با ویژگی‌محور کلاسیک به ۸۹/۷۸ درصد رسیده، Lin و همکاران

[۳۵] با الگوریتم GA–Bagging–SVM به ۹۳/۷۸ درصد و نسخه‌های XGBoost همچون Sikander و همکاران [۳۷] دقتی معادل ۹۴/۶۴ درصد را گزارش کرده‌اند. نسل جدیدتر شامل مدل ترنسفورمری QuoteTarget از Chen و همکاران [۳۸] و شیوه Cascade Deep Forest از سوی Arif و همکاران [۳۹] به دقت‌هایی نزدیک به ۹۵٪ را بدست آورده‌اند. در همین چارچوب، روش پیشنهادی روی DPI_CDF به ۹۶/۱۷ درصد دست یافت و بهبودی در حدود ۱/۱۶ درصد را شاهد بود.این بهبود نسبت به روش‌های دیگری چون [۳۷] و یا [۳۸] به ترتیب ۱/۵۳ درصد و ۲/۳۹ درصد بهبود حاصل آمده است. هرچند روش پیشنهادی بر روی مجموعه داده ProTar-II آموزش دید، با این حال بر روی مجموعه داده‌های دیده نشده دیگری آزمایش شد و معیارهای ارزیابی رضایت بخشی را برای طبقه‌بندی نمونه‌های دارویی به همراه داشت. این مسئله نشان می‌دهد که برتری محدود به یک مجموعه نیست و در سناریوهای داخلی، دیده نشده و مستقل نیز پایدار باقی می‌ماند که بر اساس آن، تعمیم‌پذیری مدل پیشنهادی قابل اثبات است. از دید روش‌شناسی، این ارتقاء با دو عامل توضیح‌پذیر است؛ نخست آنکه بهینه‌سازی زمان‌آگاه در جست‌وجوی ابرپارامترها به میزان تاخیر وزن می‌دهد و پیکربندی‌های کم‌تاخیر و باثبات را ارجح می‌کند. در درجه دوم، مدل طبقه‌بندی آبشاری دومرحله‌ای با خروج زودهنگام که بودجه محاسباتی را زودتر روی نامزدهای قوی‌تر متمرکز می‌سازد، عامل تثبیت کننده دقت محسوب می‌شود. ضمن آن‌که پروتکل ارزیابی ثابت (قفل‌کردن آستانه بر مبنای اعتبارسنجی داخلی و عدم بازآموزی روی مجموعه بیرونی) باعث می‌شود رقم ۹۶/۱۷ درصد برای پایگاه داده DPI_CDF که در مطالعات پیشین بیشتر مورد استفاده قرار گرفته، نتیجه‌ای منصفانه از عملکرد بیرونی باشد. مدل پیشنهادی با وجود عملکرد بالا، محدودیت‌هایی هم دارد. نخست این‌که بیشتر بر ویژگی‌های توالی تکیه می‌کند و از اطلاعات ساختاری سه‌بعدی یا داده‌های تعاملی زیستی کمتر استفاده شده است؛ این موضوع می‌تواند باعث سوگیری به سمت خانواده‌های پروتئینی پرنمونه شود و پیش‌بینی روی کلاس‌های کم‌نمونه را دشوار کند. دوم، در شرایطی که نسبت کلاس‌ها تغییر می‌کند یا توزیع داده دچار جابه‌جایی می‌شود، شاخص‌هایی مثل F1 به آستانه تصمیم حساس هستند و احتمال خطای کالیبراسیون در داده‌های بیرونی وجود دارد. سوم، برچسب‌های گرفته‌شده از پایگاه‌هایی مانند DrugBank یا Swiss-Prot همیشه به‌طور کامل بدون خطا نیستند و ممکن است بخشی از داده‌ها نویزی باشند. همچنین اگرچه طراحی زمان‌آگاه، سرعت پیش‌بینی را بالا می‌برد، اما فرآیند آموزش پیچیده‌تر و پرهزینه‌تر می‌شود. در نهایت، مدل در توضیح‌پذیری زیستی هنوز جای پیشرفت دارد. برای آینده می‌توان چند مسیر را دنبال کرد. استفاده از اطلاعات ساختاری و داده‌های چند بعدی برای غنی‌سازی ویژگی‌ها، به‌کارگیری روش‌های تصحیح جابجایی دامنه و کالیبراسیون پس از انتقال برای افزایش پایداری، گزارش سنجه‌هایی مثل صحت یا F1 در یک نقطه عملیاتی ثابت برای کنترل تأثیر نسبت کلاس‌ها از آن جمله است. همچنین موارد دیگری نظیر به‌کارگیری شیوه‌های استنتاجی برای ارائه بازه اطمینان در سطح نمونه، کاهش هزینه آموزش با استفاده از جست‌وجوی ابرپارامتر سبک‌تر و در نهایت افزایش توضیح‌پذیری با مطالعات حذفی و تحلیل زیستی دقیق‌تر نیز می‌تواند در آینده مطرح شود.

جدول ۵. مقایسه کلی عملکرد روش‌های پیشین و روش پیشنهادی در طبقه‌بندی دارو برای چند پایگاه داده

مطالعه	شیوه اجرا شده	داده‌های مورد استفاده	درستی کل (٪)
Jamali و همکاران [۳۱]	مقایسه چند الگوریتم که شبکه عصبی بهترین نتیجه را داشته	۴۴۳ ویژگی از داده‌های DPI_CDF	۸۹/۷۸
Lin و همکاران [۳۵]	روش ترکیبی شامل GA-Bagging-SVM	۱۴۳ ویژگی از داده‌های DPI_CDF	۹۳/۷۸
Sikander و همکاران [۳۷]	XGB-DrugPred (XGBoost با ویژگی‌های بهینه)	۱۵۵ ویژگی از داده‌های DPI_CDF و کشف دارو برای دو نوع بیماری	۹۴/۶۴
Chen و همکاران [۳۸]	QuoteTarget شامل مدل ترانسفورمری و ساختار گرافی	داده‌های DPI_CDF	۹۵/۰۰
Arif و همکاران [۳۹]	مدل Cascade Deep Forest	داده‌های DPI_CDF	۹۵/۰۱
روش پیشنهادی	LABO-HB–CatBoost (زمان-آگاه و خروج زودهنگام)	ProTar-II	۹۶/۶۴
		ProTar-II-Ind	۹۶/۴۵
		DPI_CDF	۹۶/۱۷

در نهایت، انجام ارزیابی‌های چندمرکزی و آینده‌نگر همراه با انتشار شفاف روش محاسباتی می‌تواند به رفع این کاستی‌ها و کاربرد گسترده‌تر مدل در طراحی و کشف دارو کمک کند.

۴- نتیجه‌گیری

در این پژوهش چارچوبی قابل تعمیم و بازتولیدپذیر برای پیش‌بینی داروپذیری پروتئین ارائه شد که با تکیه بر یک طبقه‌بند دومرحله‌ای مبتنی بر CatBoost و خروج زودهنگام برای نمونه‌های مطمئن و گذر ژرف برای موارد مبهم و نیز تنظیم ابرپارامتر با ابرباند بیزی آگاه از زمان پاسخ، تراز دقیق میان کارایی پیش‌بینی و هزینه زمانی را به‌طور صریح محقق می‌کند. این شیوه در عین حفظ دقت، تأخیر و نوسان عملکرد را می‌کاهد. سنجش بر ProTar-II، ProTar-II-Ind و DPI-CDF نشان داد که روش پیشنهادی علاوه بر دستیابی به درستی کل حدود ۹۶/۶ درصد و F1 حدود ۹۶/۵ درصد نسبت به شیوه‌های مبنا، زمان پیش‌بینی را به شکل معناداری کاهش می‌دهد، در برابر نامتوازنی و نویز پایا می‌ماند و تعمیم‌پذیری خود را در محک‌های داخلی و بیرونی حفظ می‌کند. نتایج با فراهم کردن احتمال‌های تنظیم‌شده و آستانه‌گذاری داده‌محور، اولویت‌بندی چابک اهداف و طراحی بهتر پروتئین دارای پتانسیل دارو شدن را در غربالگری اولیه ممکن می‌سازد و بدین‌سان شکاف دیرپای میان دقت و زمان را در کاربردهای واقعی پر می‌کند. مسیرهای آتی شامل بهره‌گیری از سرنخ‌های ساختاری و چند وجهی، تطبیق دامنه برای داده‌های جدید و ارزیابی پیش‌نگر در همکاری‌های آزمایشگاهی است تا اثر این رویکرد در طراحی دارو بیش از پیش تثبیت شود.

۵- مراجع

[1] X. Zhang, C. Shen, H. Zhang, Y. Kang, C.-Y. Hsieh, and T. Hou, “Advancing Ligand Docking through Deep Learning: Challenges and Prospects in Virtual Screening,” Accounts of Chemical Research, vol. 57, no. 10, pp. 1500–1509, 2024.

[2] X. Qi, Y. Zhao, Z. Qi, S. Hou, and J. Chen, “Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges,” Molecules, vol. 29, no. 4, Art. no. 903, Feb. 2024.

[3] H. Nada, N. A. Meanwell, and M. T. Gabr, “Virtual Screening: Hope, Hype, and the Fine Line in Between,” Expert Opinion on Drug Discovery, vol. 20, no. 2, pp. 145–162, 2025.

[4] V. Itälinna, H. Kaltiainen, N. Forss, M. Liljeström, and L. Parkkonen, “Using normative modeling and machine learning for detecting mild traumatic brain injury from magnetoencephalography data,” PLOS Computational Biology, vol. 19, no. 11, Art. no. e1011613, Nov. 2023.

[5] K. Huang, C. Xiao, Lucas M. Glass, and J. Sun, “MolTrans: Molecular Interaction Transformer for drug–target interaction prediction,” Bioinformatics, vol. 37, no. 6, pp. 830–836, 2021.

[6] T. Nguyen et al., “GraphDTA: predicting drug–target binding affinity with graph neural networks,” Bioinformatics, vol. 37, no. 8, pp. 1140–1147, 2021.

[7] L. Chen et al., “TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments,” Bioinformatics, vol. 36, no. 16, pp. 4406–4414, Aug. 2020.

[8] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, and A. Masoudi-Nejad, “DeepCDA: Deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks,” Bioinformatics, vol. 36, no. 17, pp. 4633–4642, Sep. 2020.

[9] Y. Li, J. Pei, and L. Lai, “Structure-based de novo drug design using 3D deep generative models,” Chemical Science, vol. 12, no. 41, pp. 13664–13675, 2021.

[10] Z. D. Yang, W. H. Zhong, L. Zhao, and C. Y.-C. Chen, “MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction,” Chemical Science, vol. 13, no. 3, pp. 816–833, 2022.

[11] M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, E. Kolanthai, C. J. Neal, S. Seal, et al., “AttentionSiteDTI: an interpretable graph-based model for drug–target interaction prediction using NLP sentence-level relation classification,” Briefings in Bioinformatics, vol. 23, no. 4, bbac272, 2022.

[12] A. Kerstjens and H. De Winter, “LEADD: Lamarckian evolutionary algorithm for de novo drug design,” Journal of Cheminformatics, vol. 14, no. 1, p. 3, 2022.

[13] H. Bellamy, A. A. Rehim, O. I. Orhobor, and R. King, “Batched Bayesian optimization for drug design in noisy environments,” Journal of Chemical Information and Modeling, vol. 62, no. 17, pp. 3970–3981, 2022.

[14] P. Bai, F. Miljković, B. John, and H. Lu, “Interpretable bilinear attention network with domain adaptation improves drug–target prediction,” Nature Machine Intelligence, vol. 5, no. 2, pp. 126–136, 2023.

[15] M. Moret, I. P. Angona, L. Cotos, S. Yan, et al., “Leveraging molecular structure and bioactivity with chemical language models for de novo drug design,” Nature Communications, vol. 14, no. 1, p. 114, 2023.

[16] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, “DiffDock: diffusion steps, twists, and turns for molecular docking,” in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[17] X. Fang, W. Zhong, J. Cao, Z. Zhang, S. Kar, and X. Pan, “Bayesian active learning for protein–ligand docking,” Proceedings of the National Academy of Sciences of the USA (PNAS), 2023.

[18] X. Liu et al., “DrugEx v3: scaffold-constrained drug design with graph transformer-based

gene ontologies," Bioinformatics, vol. 41, no. 7, Art. no. btaf360, Jul. 2025.
[30] Swiss-Prot: The manually annotated and reviewed protein sequence database, UniProt, 2023. [Online]. Available: https://www.uniprot.org/
[31] A. A. Jamali, R. Ferdousi, S. Razzaghi, J. Li, R. Safdari, and E. Ebrahimie, "DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins," Drug Discovery Today, vol. 21, no. 5, pp. 718-724, 2016.
[32] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," Nucleic Acids Res., vol. 36, suppl. 1, pp. D901-D906, 2008.
[33] Z. Z. Foumani, M. Shishehbor, A. Yousefpour, and R. Bostanabad, "Multi-fidelity cost-aware Bayesian optimization," *Computer Methods in Applied Mechanics and Engineering*, vol. 407, p. 115937, Mar. 2023.
[34] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and efficient hyperparameter optimization at scale," in Proc. 35th Int. Conf. Machine Learning (ICML), PMLR, 2018.
[35] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier," Artificial Intelligence in Medicine, vol. 98, pp. 35-47, Jul. 2019.
[36] C. Chen, Q. Zhang, B. Yu, Z. Yu, P. J. Lawrence, Q. Ma, and Y. Zhang, "Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier," *Computers in Biology and Medicine*, vol. 123, p. 103899, 2020.
[37] R. Sikander, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," Scientific Reports, vol. 12, no. 1, p. 5505, Apr. 2022.
[38] J. Chen, Z. Gu, Y. Xu, M. Deng, L. Lai, and J. Pei, "QuoteTarget: A sequence-based transformer protein language model to identify potentially druggable protein targets," *Protein Science*, vol. 32, no. 2, p. e4555, 2023.
[39] M. Arif, G. Fang, A. Ghulam, S. Musleh, and T. Alam, "DPI_CDF: druggable protein identifier using cascade deep forest," BMC Bioinformatics, vol. 25, no. 1, Art. no. 145, 2024.
reinforcement learning," Journal of Cheminformatics, vol. 15, no. 1, article 24, 2024.
[19] Vendy Fialková et al., "LibINVENT: Reaction-based generative scaffold decoration for in silico library design," Journal of Chemical Information and Modeling, vol. 62, no. 9, pp. 2046–2063, 2022.
[20] S. Mehta, S. Laghuvarapu, Y. Pathak, A. Sethi, M. Alvala, and U. D. Priyakumar, "MEMES: machine learning framework for enhanced molecular screening," Chemical Science, vol. 12, no. 22, pp. 11710–11721, 2021.
[21] T. Sivula, H. Käpylä, Z. Li, V. V. Rantanen, and T. Kallio, "Machine learning-boosted docking enables the efficient structure-based virtual screening of giga-scale enumerated chemical libraries," Journal of Chemical Information and Modeling, 2023.
[22] G. Zhou, H. J. Park, A. S. Tamkin, B. K. Smith, Y. Shen, F. DiMaio, et al., "An artificial intelligence accelerated virtual screening platform for drug discovery," Nature Communications, vol. 15, no. 1, 2024.
[23] D. E. Graff, E. I. Shakhnovich, and C. W. Coley, "Accelerating high-throughput virtual screening through molecular pool-based active learning," Chemical Science, vol. 12, no. 22, pp. 7866–7881, 2021.
[24] N-Q. Nguyen, G. Jang, H. Kim, and J. Kang, "Perceiver CPI: a nested cross-attention network for compound–protein interaction prediction," Bioinformatics, vol. 39, no. 1, btac731, 2023.
[25] M. Gao et al., "GraphormerDTI: A graph transformer-based approach for drug–target interaction prediction," Computers in Biology and Medicine, vol. 173, 108339, 2024.
[26] J. C. Yaacoub et al., "DD-GUI: A graphical user interface for deep learning-accelerated virtual screening of large chemical libraries (Deep Docking)," Bioinformatics, vol. 38, no. 4, pp. 1146–1148, 2022.
[27] A. P. Soleimany et al., "Evidential deep learning for guided molecular property prediction and discovery," ACS Central Science, vol. 7, no. 8, pp. 135–144, 2021.
[28] K. Huang et al., "DeepPurpose: a deep learning library for drug–target interaction prediction," Bioinformatics, vol. 36, no. 22–23, pp. 5545–5547, 2020.
[29] N. Borhani, I. Izadi, A. Motahharynia, M. Sheikholeslami, et al., "DrugTar improves druggability prediction by integrating large language models and