تحلیل مقایسه‌ای شبکه عصبی عمیق و ماشین بردار پشتیبان بهینه‌شده برای برآورد مکانی پسلرزه با تأکید بر کالیبراسیون احتمال و کمّی‌سازی عدم‌قطعیت

چکیده

برآورد دقیق مکانی پسلرزه‌ها یکی از چالش‌های اساسی در زلزله‌شناسی آماری است. در این مطالعه، دو الگوریتم یادگیری ماشین شامل شبکه عصبی عمیق (DeepMLP) و ماشین بردار پشتیبان با هسته تابع پایه شعاعی (SVM-RBF) برای طبقه‌بندی دودویی سلول‌های شبکه ۵ کیلومتری به دو کلاس «دارای پسلرزه» و «بدون پسلرزه» مقایسه شدند. مجموعه داده شامل ۲۰۲٬۶۴۴ نمونه از ۱۹۹ رخداد زمین‌لرزه بود که با تقسیم در سطح رخداد (۱۴۲ رخداد آموزش، ۵۷ رخداد آزمون) از نشتی داده جلوگیری شد. SVM با پنج اصلاح کلیدی بهبود یافت: وزن‌دهی متعادل کلاس‌ها، جایگزینی معیار F1 با AUC، گسترش فضای جستجوی γ به بازه (۱، ۰/۰۰۰۱)، کالیبراسیون احتمال، و کاهش زیرنمونه به ۱۰٬۰۰۰. SVM بهبودیافته نه‌تنها توانایی رتبه‌بندی بهتری نشان داد (ΔAUC = +0.0011، DeLong p = 0.0011) بلکه کالیبراسیون احتمال آن نیز بهتر بود (Brier: 0.166 در برابر ۰.۱۷۲، بازه اطمینان ۹۵٪ بدون همپوشانی). با این حال، تصمیم‌های دودویی دو مدل در آستانه بهینه تفاوت معناداری نداشتند (McNemar p = 0.57). ارزیابی عدم‌قطعیت با AUSE سیگنال متوسطی (~۰/۱۱) برای هر دو مدل نشان داد، اما فیلتر کردن ۵۰٪ نمونه‌های با عدم‌قطعیت بالا، دقت را از ۰/۷۷ به ۰/۸۷ (+۱۰٪) رساند. منحنی‌های یادگیری حاکی از اشباع هر دو مدل در حدود ۵٬۰۰۰ تا ۱۰٬۰۰۰ نمونه بود؛ یعنی عامل محدودکننده، نمایش ویژگی است نه حجم داده. از نظر کارایی، SVM بهبودیافته حدود ۲۰ برابر سریع‌تر از DeepMLP آموزش می‌بیند (۱۹ در برابر ۳۷۶ ثانیه)، در حالی که DeepMLP در استنتاج ۵۶ برابر سریع‌تر عمل می‌کند. این یافته‌ها به مدیران بحران در انتخاب مدل مناسب بر اساس اولویت‌های عملیاتی (سرعت در برابر دقت مثبت) کمک می‌کند.

واژه‌های کلیدی: برآورد پس‌لرزه، یادگیری ماشین، ماشین بردار پشتیبان، شبکه عصبی عمیق، کالیبراسیون احتمال، کمی‌سازی عدم‌قطعیت، آزمون DeLong، AUSE.

۱- مقدمه

پسلرزه‌ها رایج‌ترین نشانه زمین‌لرزه‌های ناشی از تغییرات تنش هستند، اما پیش‌بینی توزیع مکانی آن‌ها همچنان دشوار است (Bath, 1965). پس از یک زمین‌لرزه بزرگ، توزیع مجدد تنش ایستا می‌تواند فعالیت لرزه‌ای بعدی را تشدید یا مهار کند. قوانین تجربی مانند قانون اوموری (Utsu, 1961) و رابطه گوتنبرگ–ریشتر، فروپاشی زمانی و توزیع بزرگی را توصیف می‌کنند، اما تبیین الگوهای مکانی چالش‌برانگیزتر است. معیار تغییر تنش شکست کولمب (ΔCFS) به‌عنوان معیار فیزیکی غالب مطرح بوده است (King et al., 1994)؛ با این حال، آزمون‌های مستقل AUC را حدود ۰/۸۳ گزارش کرده‌اند که نشان‌دهنده قدرت برآورد محدود آن در صورت عدم اطلاع از مشخصات گسل‌های گیرنده است (Zhao et al., 2022).

با وجود کاربرد گسترده معیار تغییر تنش کولمب (ΔCFS)، این معیار به دلیل حساسیت شدید به پارامترهای ناشناخته‌ای مانند جهت‌گیری صفحه گسل گیرنده، ضریب اصطکاک و فشار منفذی، با عدم‌قطعیت‌های فراوانی همراه است (Harris & Simpson, 1992; Hainzl et al., 2009). شواهد تجربی متعدد کارایی آن را زیر سؤال برده‌اند؛ برای نمونه، در زمین‌لرزه نورثریج (۱۹۹٤) ΔCFS نتوانست الگوی پس‌لرزه‌ها را بهتر از توزیع تصادفی پیش‌بینی کند (Hardebeck et al., 1998) و پس از زمین‌لرزه‌های لندرز (۱۹۹۲) و کوبه (۱۹۹۵)، افزایش لرزه‌خیزی در مناطقی رخ داد که کاهش تنش کولمب پیش‌بینی شده بود (Mallman & Zoback, 2007). علاوه‌براین، تأثیر چکانش تنش پویا (Felzer & Brodsky, 2006) و فقدان شواهد

مقاله در نوبت انتشار
نشریه علمی پژوهش
«اطلاعات جغرافیایی (سپهر)»

قاطع برای سایه‌های تنش (Harris & Simpson, 2002; Felzer & Brodsky, 2005)، همراه با ناایستایی وارون‌سازی لغزش (Hainzl et al., 2009)، تأثیر تنش‌های ثانویه (Helmstetter et al., 2005) و ناشناخته ماندن مکانیزم گسل‌های گیرنده، اعتبار مدل‌های مبتنی بر تنش ایستا را با تردید جدی روبرو کرده است. Meade و همکاران (۲۰۱۷) نیز نشان دادند هیچ معیار تنش اسکالر واحدی برتری قطعی ندارد. این محدودیت‌ها ضرورت به‌کارگیری رویکردهای داده‌محور و غیرخطی مانند یادگیری ماشین را آشکار می‌سازد. از سوی دیگر، در کاربردهای عملیاتی هشدار سریع، صرفاً معیارهای تفکیک‌پذیری (مانند AUC) کافی نیست و کالیبراسیون احتمال و کمّی‌سازی عدم‌قطعیت نیز نقش کلیدی دارند؛ جنبه‌هایی که در ادبیات موجود کمتر مورد توجه قرار گرفته‌اند.

ظهور یادگیری ماشین، افق‌های تازه‌ای را برای ارزیابی خطر لرزه‌ای مبتنی بر داده گشوده است. DeVries و همکاران (۲۰۱۸) با معرفی معماری شش‌لایه DeepMLP به AUC برابر ۰/۸۴۹ دست یافتند که به‌طور قابل‌توجهی از معیار کلاسیک ΔCFS بهتر بود، هرچند این معماری چالش‌هایی مانند تنظیم گسترده فراپارامترها، زمان آموزش طولانی و ماهیت جعبه‌سیاه را به همراه دارد. پژوهش‌های بعدی نشان دادند که پیچیدگی مدل لزوماً کلید موفقیت نیست: Mignan و Broccardo (۲۰۲۰) نشان دادند یک رگرسیون لجستیک ساده با تنها دو پارامتر می‌تواند به عملکردی قابل‌رقابت با شبکه ۱۳۰٤۵۱ پارامتری DeVriesدست یابد؛ Liu و همکاران (۲۰۲٤) با مدل ترکیبی XGBoost، LightGBM و CatBoost به دقت ۹۳/٤٤٪ دست یافتند و با SHAP تفسیرپذیری را نشان دادند؛ Zhao و همکاران (۲۰۲۲) نشان دادند انتخاب ویژگی‌های مناسب می‌تواند مدل‌های ساده را با شبکه‌های عصبی هم‌تراز کند؛ Karimzadeh و همکاران (۲۰۱۹) با SVM به دقت رقابتی دست یافتند؛ و Zhouو همکاران (۲۰۲٤) پتانسیل آسیب در توالی‌های اصلی–پسلرزه را با الگوریتم‌های یادگیری ماشین پیش‌بینی کردند. این یافته‌ها ضرورت بازنگری در برتری مطلق مدل‌های عمیق را مطرح کرد.

از این رو، در پژوهش حاضر SVM با هسته RBF به‌عنوان یک مدل ساده‌تر اما قدرتمند انتخاب شد. این الگوریتم بر پایه تئوری یادگیری آماری (Cortes & Vapnik, 1995; Vapnik, 1995) با یافتن ابرصفحه‌ای که حاشیه را بیشینه می‌کند، به دسته‌بندی بهینه می‌پردازد و عملکرد قوی روی داده‌های جدولی ساختاریافته نشان داده است. هسته RBF امکان مدل‌سازی روابط غیرخطی پیچیده با نگاشت ضمنی به فضای بی‌نهایت‌بعد را فراهم می‌کند، در حالی که بهینه‌سازی محدب همگرایی به جواب سراسری را تضمین می‌کند. با این حال، کاربرد SVM در مسائل بزرگ‌مقیاس به دلیل پیچیدگی $O(n^2)$ تا $O(n^3)$ و حساسیت به انتخاب فراپارامترهای C (منظم‌سازی) و γ (پهنای هسته) محدود بوده است.

در زمینه کالیبراسیون، Gneiting و Raftery (۲۰۰۷) مفاهیم قاعده امتیازدهی مناسب را معرفی کردند، Nixon و همکاران (۲۰۱۹) نشان دادند ECE استاندارد در داده‌های نامتوازن گمراه‌کننده است، و Sun و Xu (۲۰۱٤) پیاده‌سازی سریعی از آزمون DeLongبرای مقایسهAUCهای همبسته ارائه کردند. با این حال، در هیچ‌کدام از مطالعات پیشین، تحلیل جامع کالیبراسیون احتمال و کمّی‌سازی عدم‌قطعیت ارائه نشده است؛ این شکاف، انگیزه اصلی پژوهش حاضر را شکل می‌دهد.

این مطالعه با هدف ارائه یک تحلیل مقایسه‌ای جامع بین DeepMLP و SVM-RBF بهینه‌شده انجام شده است. اهداف اختصاصی عبارتند از (۱) بهبود SVM با پنج اصلاح کلیدی و بررسی تأثیر هر یک بر عملکرد نهایی؛ (۲) مقایسه توانایی رتبه‌بندی با آزمون DeLong که برای دو AUC همبسته طراحی شده است؛ (۳) ارزیابی کالیبراسیون احتمال با معیارهای قاعده امتیازدهی مناسب (Brier, NLL)؛ (٤) کمّی‌سازی عدم‌قطعیت با معیار (سطح زیر منحنی خطای پراکنده‌سازی) AUSE[1] و

[1] Area Under Sparsification Error

منحنی پیش‌بینی گزینشی؛ (۵) تحلیل منحنی یادگیری برای تعیین نقطه اشباع و شناسایی عامل محدودکننده؛ (٦) مقایسه کارایی محاسباتی دو مدل در سناریوهای عملیاتی.

۲- مبانی نظری

۲-۱- شبکه عصبی عمیق

معماری DeepMLP مورد استفاده در این مطالعه مطابق با پیکربندی DeVries18 است: شش لایه پنهان با ۵۰ نورون در هر لایه، توابع فعال‌سازی تانژانت هذلولوی (tanh)، و لایه‌های Dropout با نرخ ۰/۵ پس از هر لایه پنهان. لایه ورودی ۱۲ ویژگی (شش مؤلفه تانسور تنش و مقادیر منفی قدرمطلق آنها) را دریافت می‌کند و لایه خروجی دارای یک نورون با فعال‌سازی سیگموئید است که به عنوان احتمال برآورد شده وقوع پس‌لرزه تفسیر می‌شود (نگاره ۱-الف).

مدل با بهینه‌ساز Adadelta (که نرخ یادگیری را برای هر پارامتر به طور جداگانه تطبیق می‌دهد) و تابع هزینه آنتروپی متقاطع دودویی آموزش داده می‌شود. آموزش در ۲۰۰ دوره با اندازه دسته ۱۰۲٤ انجام می‌شود و بهترین وزن‌ها با نظارت بر دقت اعتبارسنجی از طریق فراخوانی ModelCheckpoint حفظ می‌شوند.

بر خلاف SVM که یک مسئله بهینه‌سازی محدب را حل می‌کند و جواب سراسری بهینه را تضمین می‌کند، شبکه‌های عصبی با استفاده از روش گرادیان تصادفی آموزش داده می‌شوند و ممکن است به بهینه‌های محلی همگرا شوند. با این حال، ظرفیت آن‌ها برای یادگیری بازنمایی‌های سلسله‌مراتبی ویژگی‌ها، آن‌ها را برای شناسایی الگوهای پیچیده در داده‌های با ابعاد بالا مناسب می‌سازد.

۲-۲- ماشین بردار پشتیبان

ماشین‌های بردار پشتیبان (SVM) مدل‌های یادگیری نظارت‌شده‌ای هستند که یک ابرصفحه جداکننده بهینه را در یک فضای ویژگی با ابعاد بالا می‌سازند (نگاره ۱-ب). برای یک مسئله طبقه‌بندی دودویی با داده‌های آموزشی $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$، که در آن $\mathbf{x}_i \in \mathbb{R}^d$ و $y_i \in \{-1, +1\}$ است، SVM به دنبال یافتن ابرصفحه‌ای است که حاشیه (Margin) بین دو کلاس را بیشینه کند. در فرمول‌بندی اولیه، این مسئله به بهینه‌سازی زیر تبدیل می‌شود:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \qquad (۱)$$

با قیود:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n$$

که در آن $\mathbf{w}$ بردار وزن، b مقدار بایاس و ξ_i متغیرهای لغزش[2] هستند که امکان خطای طبقه‌بندی در فرمول‌بندی حاشیه نرم را فراهم می‌کنند و $C > 0$ پارامتر منظم‌سازی است که مبادله بین بیشینه‌سازی حاشیه و کمینه‌سازی خطای آموزش را کنترل می‌کند.

فرمول‌بندی دوگان[3] که معمولاً در عمل حل می‌شود، به صورت زیر است:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (۲)$$

[2] Slack Variables
[3] Dual Formulation

با قیود:

$$0 \le \alpha_i \le C, \sum_{i=1}^{n} \alpha_i y_i = 0$$

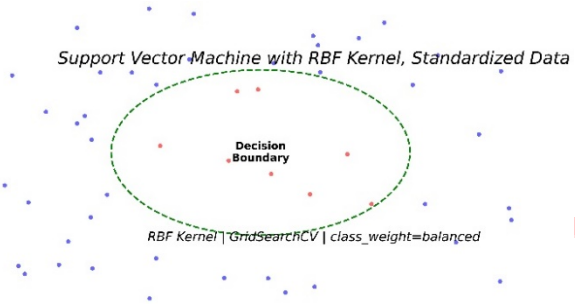
که در آن α_i ضرایب لاگرانژ هستند و $K(\mathbf{x}_i, \mathbf{x}_j)$ تابع هسته است که حاصل‌ضرب داخلی را در فضای ویژگی بدون انجام صریح تبدیل محاسبه می‌کند.

۲-۳- کالیبراسیون احتمال

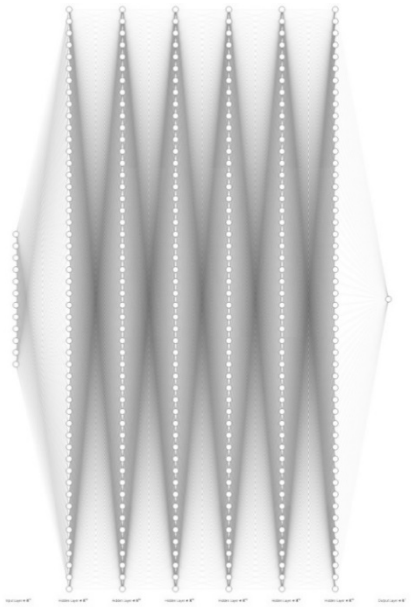
خروجی استاندارد ماشین بردار پشتیبان (SVM) یک مقدار کالیبره نشده[4] است که مستقیماً نشان‌دهنده احتمال تعلق نمونه به یک کلاس نیست (Platt, 1999). به عبارت دیگر، یک SVM استاندارد یک برچسب کلاس را پیش‌بینی می‌کند، اما درجه اطمینان خود را به صورت یک احتمال معتبر ارائه نمی‌دهد. برای رفع این مشکل، از روش کالیبراسیون پلات[5] استفاده شده است. این روش با برازش یک تابع لجستیک بر روی خروجی‌های مدل، آن‌ها را به احتمالات کالیبره‌شده تبدیل می‌کند. در این روش، دو پارامتر A و B با استفاده از روش بیشینه‌سازی درست‌نمایی[6] بر روی داده‌های آموزشی تخمین زده می‌شوند تا بهترین نگاشت را از خروجی خام مدل به یک احتمال معتبر فراهم آورند. رابطه این تبدیل به صورت زیر است:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp\left(A \cdot f(x) + B\right)} \qquad (۳)$$

که در آن $f(x)$ خروجی خام مدل است. این روش نه تنها امکان تفسیر خروجی مدل به عنوان احتمال را فراهم می‌کند، بلکه مزیت حفظ خلوت بودن[7] مدل SVM را نیز دارد. در پیاده‌سازی این پژوهش، از این روش برای تولید احتمالات خروجی SVMاستفاده شده است.



(ب)



(الف)

[4] uncalibrated value
[5] Platt Scaling
[6] maximum likelihood
[7] sparseness

نگاره ۱: معماری مدل‌ها: (الف) شبکه عصبی عمیق DeepMLP با پیکربندی ۱-۵۰-۵۰-۵۰-۵۰-۵۰-۵۰-۱۲. (ب) ماشین بردار پشتیبان با هسته RBF.

۲-٤- هسته RBF

هسته تابع پایه شعاعی (RBF) که به هسته گاوسی نیز معروف است، یکی از پرکاربردترین هسته‌ها برای مسائل غیرخطی محسوب شده و به صورت زیر تعریف می‌شود (Burges, 1998):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \quad (٤)$$

که در آن $\gamma > 0$ پارامتر کنترل‌کننده پهنای تابع گاوسی است و $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ مجذور فاصله اقلیدسی بین دو نمونه است. هسته RBF یک هسته Mercer است، به این معنا که شرط Mercer را برآورده می‌کند و با یک حاصل‌ضرب داخلی در یک فضای هیلبرت با هسته بازتولیدشونده (RKHS) مطابقت دارد. این هسته به طور ضمنی داده‌ها را به یک فضای ویژگی با ابعاد بی‌نهایت نگاشت می‌کند:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2} = \sum_{k=0}^{\infty} \frac{(-2\gamma)^k}{k!} (\mathbf{x}_i \cdot \mathbf{x}_j)^k \quad (٥)$$

این ویژگی به هسته RBF انعطاف‌پذیری فوق‌العاده‌ای برای مدل‌سازی مرزهای تصمیم پیچیده و غیرخطی می‌دهد. پارامتر γ محلی بودن تابع هسته را کنترل می‌کند. اگر γ مقدار کوچکی داشته باشد، شعاع تأثیر هسته وسیع خواهد بود و هر نمونه آموزشی بر ناحیه بزرگ‌تری از فضای ویژگی اثر می‌گذارد. این ویژگی منجر به مرزهای تصمیم نرم‌تر و سراسری‌تر می‌شود که می‌تواند از بیش‌برازش[8] جلوگیری کند و تعمیم‌پذیری مدل را بهبود بخشد. در مقابل، γ بزرگ شعاع تأثیر را باریک می‌کند و هر نمونه فقط بر همسایگی فوری خود تأثیر می‌گذارد؛ چنین حالتی امکان تشکیل مرزهای تصمیم پیچیده و محلی را فراهم می‌آورد، اما خطر بیش‌برازش به داده‌های آموزشی را افزایش می‌دهد.

پارامتر منظم‌سازی C نیز مبادله میان بیشینه‌سازی حاشیه و کمینه‌سازی خطای آموزش را تعیین می‌کند. مقدار بزرگ C جریمه سنگینی برای خطاهای طبقه‌بندی در نظر می‌گیرد و مدل را به سمت ایجاد حاشیه‌ای باریک سوق می‌دهد که ممکن است به نویز داده‌ها حساس شده و بیش‌برازش کند. در مقابل، مقدار کوچک C مدل را نسبت به خطاهای آموزشی آسان‌گیرتر می‌کند و با پذیرش خطاهای بیشتر، حاشیه‌ای پهن‌تر ایجاد می‌کند که معمولاً به تعمیم‌پذیری بهتر و کاهش خطر بیش‌برازش می‌انجامد.

۲-۵- ترفند هسته و پیچیدگی محاسباتی

ترفند هسته به SVMها امکان می‌دهد در فضاهای ویژگی با ابعاد بالا بدون محاسبه صریح تبدیل ویژگی عمل کنند. این کار با جایگزینی حاصل‌ضرب داخلی استاندارد $\mathbf{x}_i \cdot \mathbf{x}_j$ با تابع هسته $K(\mathbf{x}_i, \mathbf{x}_j)$ انجام می‌شود که می‌تواند به طور کارآمد در فضای ورودی اصلی محاسبه شود.

با این حال، پیچیدگی محاسباتی آموزش SVM به صورت درجه دوم ($O(n^2)$) تا مکعبی ($O(n^3)$) با تعداد نمونه‌های آموزشی افزایش می‌یابد، زیرا مسئله بهینه‌سازی نیاز به محاسبه و ذخیره‌سازی ماتریس هسته کامل $n \times n$ دارد. برای مجموعه داده‌های بزرگ – مانند ۱٤۷،۸٦۳ نمونه آموزشی در این مطالعه – این پیچیدگی غیرقابل‌قبول می‌شود. برای مقابله با این چالش، ما از

[8] overfitting

راهبرد زیرنمونه‌گیری استفاده می‌کنیم و SVM را روی یک زیرمجموعه تصادفی ۲۰٬۰۰۰ نمونه آموزش می‌دهیم که زمان آموزش را به شدت کاهش می‌دهد و در عین حال عملکرد رقابتی را حفظ می‌کند.

۳- روش‌شناسی

۳-۱- داده و پیش‌پردازش

داده‌های این مطالعه از همان مجموعه DeVries و همکاران (۲۰۱۸) برگرفته از پایگاه داده SRCMOD است و شامل ۱۹۹ زمین‌لرزه اصلی در مقیاس جهانی می‌شود. برای هر رخداد، شبکه‌ای با ابعاد ۵×۵×۵ کیلومتر تا ۱۰۰ کیلومتر مسطحاتی و ۵۰ کیلومتر عمودی گسترده شده و هر سلول با ۱۲ ویژگی ورودی (قدر مطلق شش مؤلفه تنش و منفی قدر مطلق آنها، نرمال‌شده با 10^{-6}) و یک برچسب دودویی (حضور یا عدم حضور پسلرزه) مشخص شده است. مجموعه نهایی شامل ۲۰۲٬۶۴۴ نمونه با نرخ مثبت ۵۲/۷٪ در آموزش و ۴۰/۳٪ در آزمون است.

تقسیم‌بندی در سطح رخداد انجام شد یعنی ۱۴۲ رخداد برای آموزش و ۵۷ رخداد برای آزمون، تا از نشتی داده جلوگیری شود، چرا که سلول‌های یک رخداد هرگز هم‌زمان در دو مجموعه حضور نمی‌یابند. همچنین ۱۰٪ از رخدادهای آموزشی به‌عنوان مجموعه اعتبارسنجی جدا شد. مشخصات مجموعه داده در جدول ۱ ارائه شده است.

جدول ۱. مشخصات مجموعه داده و تقسیم‌بندی در سطح رخداد

ویژگی	مقدار
تعداد رخدادهای آموزش	۱۴۲
تعداد رخدادهای آزمون	۵۷
نمونه‌های آموزش	۱۵۳٬۳۴۷
نمونه‌های اعتبارسنجی	۱۰٬۹۴۶
نمونه‌های آزمون	۳۸٬۳۵۱
نرخ مثبت آموزش	۰/۵۲۷
نرخ مثبت آزمون	۰/۴۰۳
تعداد ویژگی‌ها	۱۲

۳-۲- ماشین بردار پشتیبان بهبودیافته

ماشین بردار پشتیبان استاندارد در مواجهه با داده‌های پسلرزه به دلیل پنج ضعف ساختاری دچار فروپاشی عملکردی می‌شود: (۱) عدم تعادل شدید کلاس‌ها که باعث سوگیری مدل به سمت کلاس اکثریت می‌گردد؛ (۲) معیار بهینه‌سازی نامناسب (F1) که به آستانه تصمیم وابسته است و در جستجوی فراپارامترها به انتخاب‌های ناپایدار منجر می‌شود؛ (۳) محدودیت مصنوعی در فضای جستجوی پارامتر γ که مانع دستیابی به مرزهای تصمیم مناسب می‌شود؛ (۴) عدم کالیبراسیون احتمال که خروجی مدل را به یک مقدار کالیبره‌نشده تبدیل می‌کند و کاربرد آن را در تصمیم‌گیری مبتنی بر احتمال محدود می‌سازد؛ و (۵) حجم بیش از حد زیرنمونه که بدون بهبود معنادار در عملکرد، هزینه محاسباتی را چند برابر افزایش می‌دهد. بر این اساس، SVM در این مطالعه با پنج اصلاح کلیدی بهبود یافت: (۱) استفاده از class_weight='balanced' برای تعادل کلاس‌ها؛ (۲) تغییر معیار بهینه‌سازی از F1 به scoring='roc_auc' که مستقل از آستانه است؛ (۳) گسترش فضای جستجوی پارامتر γ به بازه [۱،

مقاله در دست انتشار
نشریه علمی پژوهشی
«اطلاعات جغرافیایی (سپهر)»

[۰/۰۰۰۱ ؛ (٤) کالیبراسیون احتمال با CalibratedClassifierCV (method='isotonic', cv=5, ensemble=True)؛ و (٥) کاهش حجم زیرنمونه به ۱۰۰٬۰۰۰ نمونه به‌جای ۲۰۰٬۰۰۰.

پس از اعمال این اصلاحات، روند بهینه‌سازی SVM در چهار مرحله کلیدی سازمان‌دهی شد. مرحله اول استانداردسازی داده‌هاست. برای SVM با هسته RBF، مقیاس‌بندی ویژگی‌ها حیاتی است، چرا که این هسته فاصله اقلیدسی $\| x_i - x_j \|^2$ را محاسبه می‌کند و ویژگی‌هایی با مقیاس بزرگ‌تر بر این فاصله غالب می‌شوند. برای اطمینان از مشارکت یکسان همه ویژگی‌ها، از StandardScaler برای تبدیل داده‌ها به توزیعی با میانگین صفر و واریانس یک استفاده شد:

$$x' = \frac{x - \mu}{\sigma} \quad (٦)$$

که در آن μ میانگین و σ انحراف معیار هر ویژگی است. اسکیلر تنها روی داده‌های آموزش برازش داده شد و سپس روی مجموعه‌های اعتبارسنجی و آزمون اعمال گردید تا از نشتی داده جلوگیری شود.

مرحله دوم زیرنمونه‌گیری برای کارایی محاسباتی است. با توجه به پیچیدگی محاسباتی $O(n^3) - O(n^2)$ در آموزش SVM، مدل روی یک زیرمجموعه تصادفی ۱۰۰٬۰۰۰ نمونه از مجموعه آموزش (از مجموع ۱۵۳٬۳٤۷ نمونه) آموزش داده شد. این انتخاب نه تنها زمان آموزش را از حدود ۳۷٦ ثانیه به ۱۹ ثانیه کاهش داد (صرفه‌جویی ۲۰ برابری)، بلکه عملکرد رقابتی را نیز حفظ کرد، زیرا SVM عمدتاً تحت تأثیر بردارهای پشتیبان قرار می‌گیرد و نه کل مجموعه داده.

مرحله سوم بهینه‌سازی فراپارامترهاست. عملکرد SVM-RBF به شدت به انتخاب پارامترهای C و γ حساس است. در این مطالعه، برای جستجوی سیستماتیک این فراپارامترها، از سه راهبرد مکمل استفاده شد: جستجوی شبکه‌ای با اعتبارسنجی متقابل سه‌تایی گروه‌بندی‌شده در سطح رخداد، جستجوی تصادفی با ۱۰۰ تکرار، و بهینه‌سازی بیزی با ۱۰۰ اجرا مبتنی بر الگوریتم TPE. فضای جستجو برای هر سه روش به‌صورت زیر تعریف شد:

$C \in [0.1, 10^3]$ (در مقیاس لگاریتمی)

$\gamma \in [10^{-4}, 10^0]$ (در مقیاس لگاریتمی)

class_weight ∈ {None,’balanced’}

گزینه class_weight='balanced' به‌طور خودکار وزن‌ها را به نسبت عکس فراوانی کلاس‌ها تنظیم می‌کند و به عدم توازن ذاتی در داده‌های پسلرزه می‌پردازد. معیار امتیازدهی در همه روش‌ها، سطح زیر منحنی ROC بر اعتبارسنجی متقابل انتخاب شد تا نتایج از نظر آماری قابل‌اعتماد و مستقل از آستانه تصمیم باشند.

در نهایت، مرحله چهارم کالیبراسیون احتمال است. پس از یافتن پارامترهای بهینه، SVM بر روی کل زیرنمونه ۱۰۰٬۰۰۰ تایی با CalibratedClassifierCV (method='isotonic', cv=5, ensemble=True) آموزش داده شد تا خروجی خام مدل (فاصله نمونه تا ابرصفحه) به احتمال‌های کالیبره در بازه صفر و یک تبدیل شود. این مرحله، خروجی SVM را برای کاربردهایی که به احتمال‌های معتبر نیاز دارند (مانند تصمیم‌گیری وزنی، محاسبه امتیاز Brier و تحلیل UQ) قابل استفاده می‌سازد.

۳-۳- بهینه‌سازی آستانه

آستانه تصمیم پیش‌فرض ۰/۵ لزوماً بهینه نیست، به ویژه زمانی که هزینه خطاهای مثبت کاذب و منفی کاذب نابرابر است. برای هر مدل، منحنی‌های دقت–بازیابی روی مجموعه اعتبارسنجی ساخته شد و آستانه‌ای که امتیاز F1 را بیشینه می‌کند، انتخاب گردید:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (٧)$$

سپس این آستانه بهینه روی احتمالات پیش‌بینی شده مجموعه آزمون اعمال شد تا دسته‌بندی‌های دودویی نهایی به دست آیند.

۳-٤- معیارهای ارزیابی

به منظور ارزیابی همه جانبه عملکرد مدل‌ها، معیارهای سنجش در پنج گروه اصلی دسته‌بندی شدند. گروه نخست به توانایی تفکیک مدل اختصاص دارد و شامل سطح زیر منحنی ROC (AUC)، میانگین دقت (AP) و آزمون DeLong برای مقایسه دو AUC همبسته است (Sun & Xu, 2014). گروه دوم کالیبراسیون احتمال را می‌سنجد و از معیارهای امتیاز Brier ، NLL[9] و RMSCE[10] تشکیل شده است. گروه سوم معیارهای طبقه‌بندی را در بر می‌گیرد که عبارتند از: دقت کلی[11]، دقت مثبت[12]، بازیابی[13]، ویژگی[14]، امتیاز F1 و آزمون McNemar. گروه چهارم به کمّی‌سازی عدم قطعیت (UQ) می‌پردازد و شامل AUSE و همبستگی میان عدم قطعیت و خطای پیشبینی می‌شود. در نهایت، گروه پنجم کارایی محاسباتی مدل را از طریق اندازه‌گیری زمان آموزش و زمان پیشبینی ارزیابی می‌کند.

۳-۵- کمّی‌سازی عدم‌قطعیت

برای ارزیابی قابلیت اطمینان پیش‌بینی‌های هر مدل، از دو راهبرد متفاوت برای کمّی‌سازی عدم‌قطعیت استفاده شد. در مورد شبکه عصبی عمیق، به دلیل حضور لایه‌های Dropout در معماری، از روش MC Dropout بهره گرفته شد. در این روش، ۱۰۰ بار پیش‌روی نمونه‌ها از شبکه با فعال نگه‌داشتن لایه‌های Dropout (یعنی training=True) انجام می‌شود و میانگین و انحراف معیار این ۱۰۰ پیش‌بینی به‌عنوان احتمال مرکزی و شاخص عدم‌قطعیت هر نمونه در نظر گرفته می‌شود. انحراف معیار بزرگ‌تر نشان‌دهنده عدم‌قطعیت بالاتر مدل در آن نمونه است. در مقابل، SVM چنین مکانیزم ذاتی برای UQ ندارد؛ بنابراین برای این مدل از روش بوت‌استرپ استفاده شد. بدین ترتیب که ۱۰ بار نمونه‌برداری تصادفی با جایگذاری از مجموعه آموزش (با نسبت ٪۷۰) انجام شد و در هر بار، یک نسخه کالیبره‌شده SVM روی زیرنمونه جدید آموزش دید. سپس پیش‌بینی‌های این ۱۰ مدل روی مجموعه آزمون محاسبه و میانگین و انحراف معیار آن‌ها به‌عنوان احتمال و عدم‌قطعیت هر نمونه گزارش شد. برای سنجش کیفیت این سیگنال‌های عدم‌قطعیت، از معیار AUSE استفاده شد که بر پایه مقایسه منحنی پراکنده‌سازی مدل با منحنی آرمانی (منحنی ایده‌آلی که خطاها را به‌طور کامل حذف می‌کند) تعریف می‌شود. در این معیار، هرچه عدم‌قطعیت مدل با خطای واقعی آن همبستگی بیشتری داشته باشد، فاصله منحنی مدل از منحنی آرمانی کمتر و در نتیجه AUSE کوچک‌تر خواهد بود؛ بنابراین مقادیر کمتر AUSE نشان‌دهنده کیفیت بالاتر کمّی‌سازی عدم‌قطعیت است.

۳-٦- نرم‌افزارها و ابزارها

[9] Negative Log-Likelihood
[10] Root Mean Square Calibration Error
[11] accuracy
[12] precision
[13] recall
[14] specificity

تحلیل‌ها با زبان Python 3.10 و کتابخانه‌های scikit-learn، TensorFlow، Optuna، و SHAP انجام شد. شبیه‌سازی‌های MC Dropout و بوت‌استرپ[15] با NumPy پیاده‌سازی شدند.

٤- نتایج

٤-١- بهینه‌سازی فراپارامترها

برای تعیین مقادیر بهینه پارامترهای SVM با هسته RBF، سه راهبرد جستجو شامل شبکه‌ای، تصادفی و بهینه‌سازی بیزی روی زیرمجموعه آموزشی SVM اجرا شد. در هر سه روش، معیار "سطح زیر منحنی ROC" بر اعتبارسنجی متقابل سه‌تایی گروه‌بندی‌شده در سطح رخداد انتخاب شد تا نتایج مستقل از آستانه تصمیم باشند. جدول ۲ خلاصه نتایج را نشان می‌دهد.

جدول ۲. نتایج بهینه‌سازی فراپارامترهای SVM با سه روش جستجو.

روش	Best CV AUC	زمان(s)	C	γ	class_weight
Grid Search	۰/۷۷۲۲	۱۵۲	۱/۰	۰/۰۱	balanced
Random Search	۰/۷۷۲۳	۳۵۷	۱/۸۱	۰/۰۱۰۸	balanced
Bayesian (TPE)	۰/۷۷۲٤	۱۰٤٥	۳/۳۱	۰/۰۰۸۱	balanced

هر سه راهبرد AUC بسیار نزدیک داشتند، اما تفاوت‌های چشمگیر در زمان اجرا و مقادیر پارامترهای بهینه، الگوهای متمایزی را آشکار می‌سازد. جستجوی شبکه‌ای سریع‌ترین اجرا را داشت اما به پارامترهای مرزی (C=1 و γ=0.01) همگرا شد. در مقابل، بهینه‌سازی بیزی با وجود هزینه محاسباتی بالا (حدود ۷ برابر)، بهترین AUC را داشت و پارامترهای بهینه نهایی (C=3.31 و γ=0.0081) را در فضای پیوسته و دور از مرزها شناسایی کرد. این مقادیر نشان می‌دهد که SVM بهبودیافته از یک تابع تصمیم نرم‌تر با منظم‌سازی متعادل استفاده می‌کند که به کالیبراسیون بهتر و پایداری بیشتر منجر می‌شود. بنابراین، پارامترهای نهایی SVM در این مطالعه بر پایه نتایج بهینه‌سازی بیزی تعیین شدند.

٤-٢- مقایسه عملکرد نهایی

پس از آموزش هر دو مدل و تنظیم آستانه تصمیم بر پایه بیشینه‌سازی F1 روی مجموعه اعتبارسنجی، ارزیابی نهایی روی ۳۸،۳۵۱ نمونه آزمون (از ۵۷ رخداد کاملاً مستقل) انجام شد. جدول ۳ خلاصه جامع این مقایسه را ارائه می‌دهد.

جدول ۳. معیارهای نهایی DeepMLP و SVM بهبودیافته روی مجموعه آزمون (۳۸،۳۵۱ نمونه).

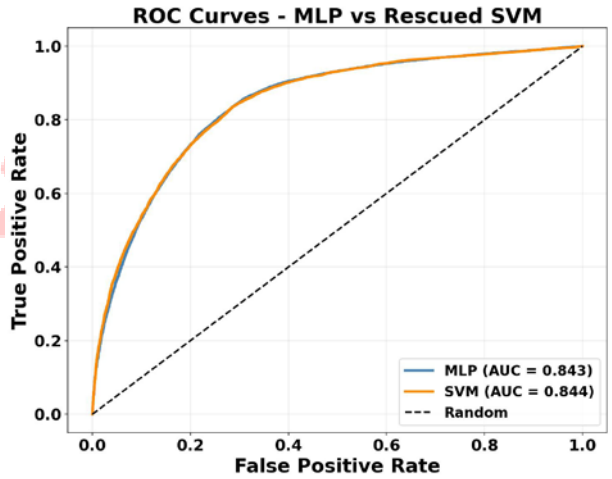
معناداری	برنده	SVM بهبودیافته	DeepMLP	معیار
DeLong p= ۰/۰۰۱	SVM	۰/۸٤٤۳	۰/۸٤۳۲	AUC
–	SVM	۰/۷۷۸۰	۰/۷۷٤۱	AP
غیرمعنادار	SVM	۰/۷۷۱۹	۰/۷۷۱٦	Accuracy
غیرمعنادار	SVM	۰/۷۱٦۳	۰/۷۱۳۳	F۱
–	SVM	۰/۷۱۳۷	۰/۷۰٤۳	Recall
غیرمعنادار	DeepMLP	۰/۷۱۸۹	۰/۷۲۲٥	Precision

[15] Bootstrap

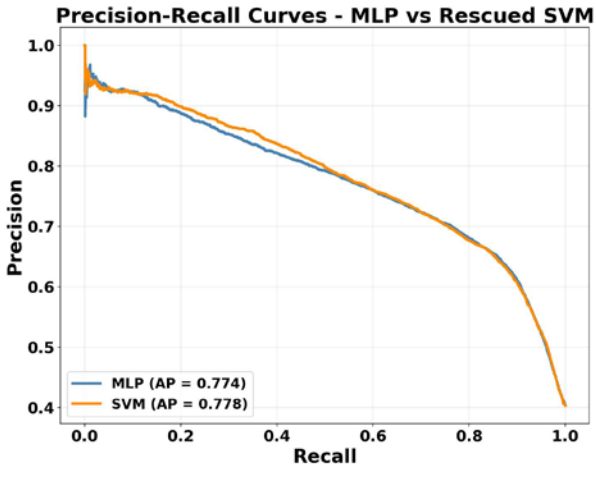
–	DeepMLP	۰/۸۱۱۱۲	۰/۸۱۷۰	Specificity
جدا CI	SVM	۰/۱۶۵۹	۰/۱۷۲۰	Brier ↓
معنادار	SVM	۰/۵۱۰۰	۰/۵۲۵۳	NLL ↓
نامناسب	DeepMLP	۰/٤٦۳۸	۰/٤٤۱۵	RMSCE ↓
۲۰×	SVM	۱۹	۳۷٦	زمان آموزش (s)
٥٦×	DeepMLP	۱۱۸	۲/۱	زمان پیش‌بینی (s)

سه یافته کلیدی از جدول ۳ استخراج می‌شود. نخست، برتری معنادار SVM در تفکیک‌پذیری: AUC آن ۰/۸٤٤۳ در برابر ۰/۸٤۳۲ است که با آزمون DeLong در سطح ۱٪ معنادار تشخیص داده شد؛ این برتری در AP نیز تکرار می‌شود. دوم، برتری معنادار SVM در کالیبراسیون احتمال: SVM با Brier = ۰/۱۶۵۹ در برابر ۰/۱۷۲۰ و NLL = ۰/۵۱۰۰ در برابر ۰/۵۲۵۳ بهتر عمل می‌کند، و به‌ویژه Brier بازه اطمینان ۹۵٪ کاملاً جداگانه‌ای دارد که نشان‌دهنده برتری آماری قاطع است. سوم، کارایی محاسباتی مکمل: SVM با زمان آموزش ۱۹ ثانیه حدود ۲۰ برابر سریع‌تر از DeepMLP است، در حالی که DeepMLP با زمان پیش‌بینی ۲/۱ ثانیه ٥٦ برابر سریع‌تر از SVM عمل می‌کند. این تفکیک، انتخاب مدل را کاملاً وابسته به سناریوی عملیاتی می‌سازد.

از سوی دیگر، DeepMLP در دو معیار دقت مثبت و ویژگی برتری نسبی دارد، اما این اختلاف‌ها بر اساس بازه اطمینان ۹۵٪ غیرمعنادار هستند. نکته حائز اهمیت آن است که RMSCE تنها معیار کالیبراسیونی که DeepMLP را برنده نشان می‌دهد، یک قاعده امتیازدهی مناسب نیست و همانند ECE به آستانه درون بازه حساس است؛ بنابراین تناقض ظاهری آن با Brier و NLL نه ناشی از ضعف SVM، بلکه ناشی از محدودیت ذاتی خود RMSCE در داده‌های نامتوازن است. نگاره ۲ منحنی‌های ROC دو مدل را نشان می‌دهد؛ SVM بهبودیافته در ناحیه FPR پایین که برای کاربردهای هشدار سریع حیاتی است، برتری محسوسی دارد. نگاره ۳ منحنی‌های دقت–بازیابی را ارائه می‌دهد؛ هرچند اختلاف AP دو مدل در سطح عددی کوچک است، SVM در کل محدوده بازیابی به‌ویژه در بازیابی بالای ۰٫۵ که برای سیستم‌های هشدار عملیاتی اهمیت دارد، عملکرد بهتری نشان می‌دهد.



نگاره ۲: منحنی‌های ROC دو مدل. SVM بهبود یافته با AUC = ۰/۸٤٤ در ناحیه FPR پایین بهتر عمل می‌کند.



نگاره ۳: منحنی‌های دقت–بازیابی. SVM بهبود یافته با AP = ۰/۷۷۸ بهتر از DeepMLP با AP = ۰/۷۷٤ است.

٤–۳– آزمون‌های آماری

برای ارزیابی معناداری آماری تفاوت دو مدل، سه آزمون مکمل به کار گرفته شد :آزمون DeLong برای مقایسه دو AUC همبسته، آزمون McNemar برای مقایسه تصمیم‌های دودویی، و بازه اطمینان بوت‌استرپ برای برآورد عدم‌قطعیت معیارها. این انتخاب بر این منطق استوار است که این سه معیار، سه لایه متفاوت عملکرد یعنی تفکیک‌پذیری، تصمیم دودویی، و کالیبراسیون را می‌سنجند و ممکن است به نتایج متفاوتی منجر شوند.

آزمون DeLong به‌طور اختصاصی برای مقایسه معناداری دو AUC همبسته طراحی شده است (Sun & Xu, 2014) و برخلاف روش‌های مبتنی بر بوت‌استرپ، کوواریانس میان دو مدل را در محاسبه عدم‌قطعیت لحاظ می‌کند. به همین دلیل، این آزمون برای مقایسه دو مدل آموزش‌دیده روی یک مجموعه‌داده مشترک، از حساسیت آماری بالاتری برخوردار است. نتایج در جدول ٤ ارائه شده است.

جدول ٤. نتایج آزمون DeLong برای مقایسه AUC دو مدل.

پارامتر	مقدار
AUC(DeepMLP)	۰/۸٤۳۲
AUC(SVM)	۰/۸٤٤۳
ΔAUC	+۰/۰۰۱۱
z-statistic	−۳/۲۷٤
p-value	۰/۰۰۱۱
نتیجه	**تفاوت معنادار در سطح ۱٪**

مقدار z منفی نشان می‌دهد که AUC مدل SVM به‌طور معنادار بزرگ‌تر از DeepMLP است، هرچند این تفاوت در سطح عددی ناچیز است. با توجه به حجم نمونه بزرگ (۳۸٬۳۵۱ نمونه) و ساختار همبسته دو AUC، این تفاوت کوچک از نظر آماری معنادار است (p = ۰/۰۰۱۱) و نشان می‌دهد SVM بهبودیافته در تمام آستانه‌های ممکن، توانایی رتبه‌بندی بهتری دارد. این یافته با منحنی ROC (نگاره ٤) نیز سازگار است، به‌ویژه در ناحیه FPR پایین که برای کاربردهای هشدار سریع حیاتی است.

در حالی که آزمون DeLong توانایی رتبه‌بندی را در سراسر دامنه آستانه‌ها می‌سنجد، آزمون McNemar فقط تصمیم‌های دودویی نهایی در آستانه بهینه هر مدل را با یکدیگر مقایسه می‌کند. نتایج در جدول ٥ ارائه شده است.

جدول ٥. نتایج آزمون McNemar برای تصمیم‌های دودویی.

پارامتر	مقدار
هر دو درست	۲۹٬۳۳۰
فقط DeepMLP درست	۲٦۰
فقط SVM درست	۲۷٤
هر دو غلط	۸٬٤۸۷
p-value	۰/۵۷۳۷
نتیجه	**تفاوت غیرمعنادار**

تعداد مواردی که SVM تنها درست پیش‌بینی کرده و مواردی که DeepMLP تنها درست پیش‌بینی کرده تفاوت معناداری با یکدیگر ندارند؛ این تقارن تقریبی، مقدار p = ۰/۵۷۳۷ را توضیح می‌دهد. به عبارت دیگر، در سطح تصمیم دودویی، دو مدل

مقاله در نوبت انتشار
نشریه علمی پژوهشی
«اطلاعات جغرافیایی (سپهر)»

به‌طور آماری معادل عمل می‌کنند. این نتیجه در ظاهر با یافته آزمون DeLong در تناقض است، اما در واقع تصویری دقیق‌تر از ماهیت تفاوت دو مدل ارائه می‌دهد: مزیت SVM در رتبه‌بندی احتمال‌ها است، نه در مرز تصمیم نهایی.

برای ارزیابی معناداری تفاوت‌های کوچک، از روش بوت‌استرپ با ۵۰۰ تکرار استفاده شد. جدول ٦ نتایج را برای معیارهای کلیدی نشان می‌دهد.

جدول ٦. بازه اطمینان ٪۹۵ (بوت‌استرپ، ۵۰۰ تکرار) برای معیارهای کلیدی.

معیار	DeepMLP	SVM	همپوشانی CI
Accuracy	[۰/۷۶۷٦، ۰/۷۷۵۷]	[۰/۷۶۷۷، ۰/۷۷٦۱]	**کامل**
AUC	[۰/۸۳۹۳، ۰/۸٤۷۱]	[۰/۸٤۰۲، ۰/۸٤۸۱]	**کامل**
AP	[۰/۷٦۷۹، ۰/۷۸۰٦]	[۰/۷۷۱٤، ۰/۷۸٤٤]	**کامل**
Brier	[۰/۱۷۰٤، ۰/۱۷۳۵]	[۰/۱٦٤۱، ۰/۱٦۷۵]	**هیچ**

تحلیل جدول ٦ الگویی متمایز را آشکار می‌سازد: در سه معیار دقت کلی، AUC و AP، بازه‌های اطمینان دو مدل کاملاً همپوشانی دارند و تفاوت‌ها از نظر آماری معنادار نیستند. اما در معیار Brier، بازه‌های اطمینان دو مدل هیچ همپوشانی ندارند، قوی‌ترین شاهد آماری برای برتری SVM در کالیبراسیون احتمال. به بیان دیگر، اگرچه دو مدل در تفکیک‌پذیری و تصمیم دودویی معادل عمل می‌کنند، SVM بهبودیافته به‌طور معنادار احتمال‌های معتبرتری تولید می‌کند، کیفیتی که در سیستم‌های هشدار احتمالاتی و تصمیم‌گیری مبتنی بر ریسک نقشی حیاتی دارد.

٤-٤- تحلیل کالیبراسیون احتمال

کالیبراسیون احتمال، یعنی میزان تطابق بین احتمال پیش‌بینی‌شده و فراوانی واقعی رخداد، یکی از جنبه‌های حیاتی در کاربردهای هشدار سریع است که در ادبیات پیشین کمتر مورد توجه قرار گرفته است. برای ارزیابی این جنبه، از سه معیار مکمل استفاده شد: Brier و NLL که هر دو از خانواده قاعده امتیازدهی مناسب هستند و کمینه آن‌ها تنها در نقطه کالیبراسیون کامل رخ می‌دهد، و RMSCE که یک معیار مبتنی بر بازه‌بندی است. نگاره ٤ نمودار قابلیت اطمینان دو مدل را به‌صورت بصری نشان می‌دهد و جدول ۷ مقایسه کمّی سه معیار را ارائه می‌کند.

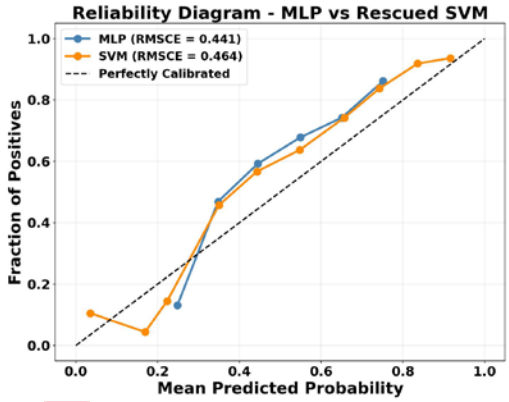
جدول ۷. مقایسه معیارهای کالیبراسیون دو مدل.

معیار	DeepMLP	SVM بهبودیافته	تفاوت نسبی	نوع معیار
Brier ↓	۰/۱۷۲۰	۰/۱٦۵۹	−٪۳/۵	**قاعده امتیازدهی مناسب**
NLL ↓	۰/۵۲۵۳	۰/۵۱۰۰	−٪۲/۹	**قاعده امتیازدهی مناسب**
RMSCE ↓	۰/٤٤۱۵	۰/٤٦۳۸	+٪۵	**نامناسب**

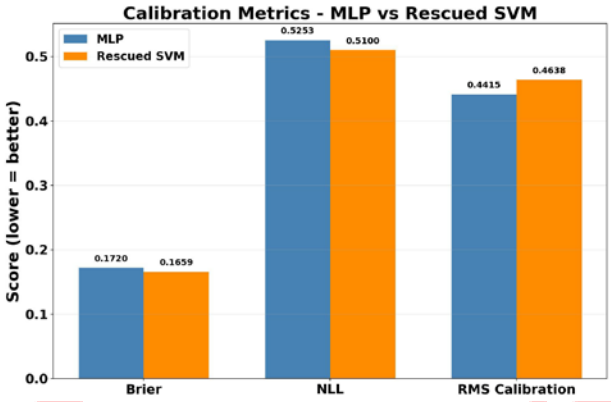
همان‌طور که در جدول ۷ ملاحظه می‌شود، Brier و NLL به‌طور همسو برتری SVM بهبودیافته را نشان می‌دهند و این برتری از نظر آماری نیز معنادار است، چرا که بازه اطمینان ۹۵٪ بوت‌استرپ برای Brier در دو مدل هیچ همپوشانی‌ای ندارد. در مقابل، RMSCE تصویر معکوسی ارائه می‌دهد؛ اما این تناقض ریشه در ماهیت این معیار دارد: RMSCE یک قاعده امتیازدهی مناسب نیست و چون در محاسبه دقت درون هر بازه از آستانه ثابت ۰/۵ استفاده می‌کند، در داده‌های با نرخ پایه نامتوازن

مقاله در نوبت انتشار
نشریه علمی پژوهشی
« اطلاعات جغرافیایی (سپهر) »

(نرخ مثبت ۰/٤۰ در این مطالعه) به‌طور سیستماتیک بیش‌برآورد می‌شود و مدل‌هایی با احتمال‌های دور از ۰/۵ را به‌ناحق برتر نشان می‌دهد. این محدودیت در ادبیات کالیبراسیون به‌طور مستند گزارش شده است (Nixon et al., 2019; Gneiting & Raftery, 2007). بنابراین در این مطالعه وزن اصلی بر Brier و NLL گذاشته شده است و نگاره ۵ این تناقض را به‌صورت بصری آشکار می‌سازد.



نگاره ٤. نمودار نمودار قابلیت اطمینان دو مدل.



نگاره ۵. مقایسه معیارهای کالیبراسیون (Brier، NLL، RMSCE).

نتیجه عملی این تحلیل آن است که SVM بهبودیافته، برخلاف نسخه اولیه خود که در آن امتیاز Brier تا ۰/۳۲٤ بیش‌برآورد شده بود، اکنون به سطح قابل‌قبولی از کالیبراسیون دست یافته و حتی از DeepMLP نیز پیشی گرفته است. این یافته نشان می‌دهد که ضعف‌های کالیبراسیون گزارش‌شده در مطالعات پیشین درباره SVM، بیش از آنکه ناشی از ماهیت ذاتی الگوریتم باشد، به انتخاب نادرست معیار بهینه‌سازی و عدم کالیبراسیون احتمال پس از آموزش مربوط می‌شود.

٤-۵- کمّی‌سازی عدم‌قطعیت

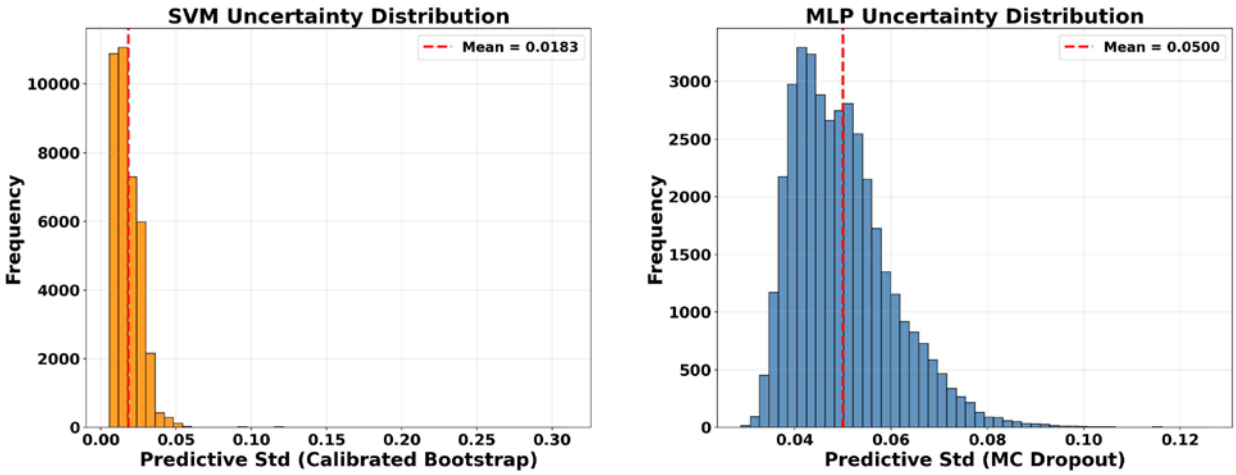
کمّی‌سازی عدم‌قطعیت (UQ) یکی از ارکان اساسی در کاربردهای عملیاتی پیش‌بینی پسلرزه است، زیرا در سامانه‌های هشدار سریع، دانستن میزان اطمینان مدل به پیش‌بینی خود به‌اندازه خود پیش‌بینی اهمیت دارد. در این مطالعه، عدم‌قطعیت DeepMLP با روش MC Dropout و در SVM بهبودیافته با بوت‌استرپ کالیبره‌شده محاسبه شد. کیفیت این عدم‌قطعیت‌ها با دو معیار ارزیابی شد: AUSE و همبستگی میان عدم‌قطعیت و خطا. جدول ۸ خلاصه این معیارها را نشان می‌دهد.

جدول ۸. معیارهای کمّی‌سازی عدم‌قطعیت (UQ).

مدل	AUSE ↓	Mean Std	Unc-Error Pearson	Unc-Error Spearman	Acc(low unc, 50%)
DeepMLP	۰/۱۱۰٤	۰/۰۵۰۰	۰/۲۰۳	۰/۲۳۸	۰/۸۷۰٤
SVM	۰/۱۰۹۲	۰/۰۱۸۳	۰/۱۳۵	۰/۲۳۵	۰/۸٦٤۹

آنچه در جدول ۸ جلب توجه می‌کند، تفاوت بنیادین در ماهیت عدم‌قطعیت دو مدل است. میانگین انحراف معیار پیش‌بینی‌ها در DeepMLP حدود ۲/۷ برابر SVM است که ناشی از ماهیت متفاوت روش‌های UQ است؛ اما هر دو مدل به AUSE تقریباً یکسان (اختلاف کمتر از ۱٪) دست می‌یابند. این بدان معناست که SVM با یک‌سوم تنوع عدم‌قطعیت، همان قدرت تفکیک‌پذیری را حفظ می‌کند، یعنی هر واحد انحراف معیار در SVM اطلاعات بیشتری حمل می‌کند. نگاره‌های ٦-الف و ٦-ب توزیع عدم‌قطعیت دو مدل را به‌صورت جداگانه نشان می‌دهند.

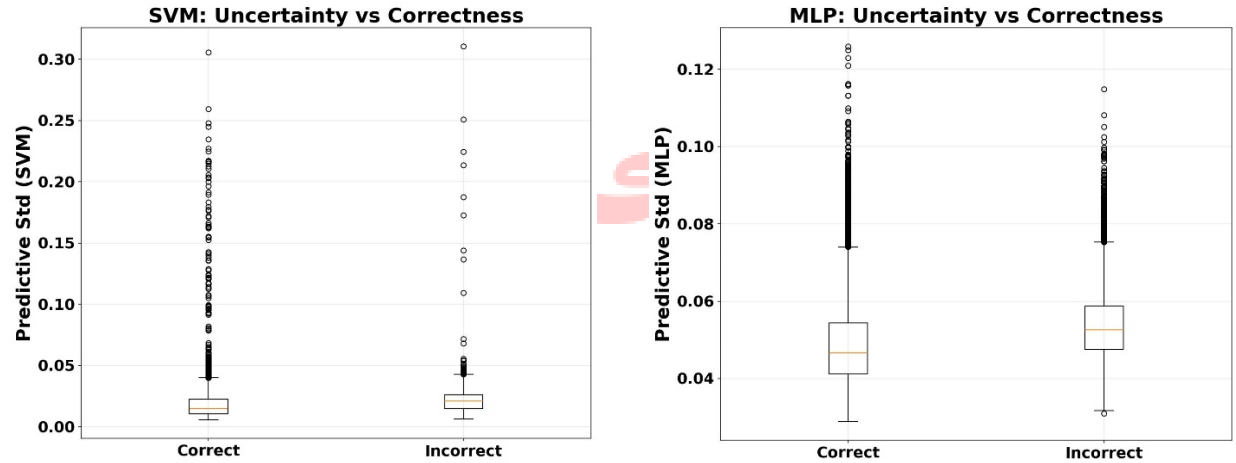
نشریه علمی پژوهشی
«اطلاعات جغرافیایی (پژوهش)»



(الف) (ب)

نگاره ٦. توزیع عدم‌قطعیت (الف) DeepMLP (MC Dropout)، (ب) SVM (Calibrated Bootstrap)

معیار دوم، بررسی رابطه میان عدم‌قطعیت و خطای واقعی است. همبستگی پیرسون[16] برای DeepMLP و SVM به‌ترتیب ۰/۲۰۳ و ۰/۱۳۵ است، اما همبستگی اسپیرمن[17] در هر دو مدل تقریباً یکسان (۰/۲۳۸ در برابر ۰/۲۳۵) است، یعنی توانایی رتبه‌بندی نمونه‌ها بر اساس احتمال خطا در دو مدل مشابه است. نگاره‌های ۷–الف و ۷–ب این رابطه را در قالب نمودار جعبه‌ای نشان می‌دهند و تأیید می‌کنند که نمونه‌های اشتباه به‌طور سیستماتیک عدم‌قطعیت بالاتری دارند.



(الف) (ب)

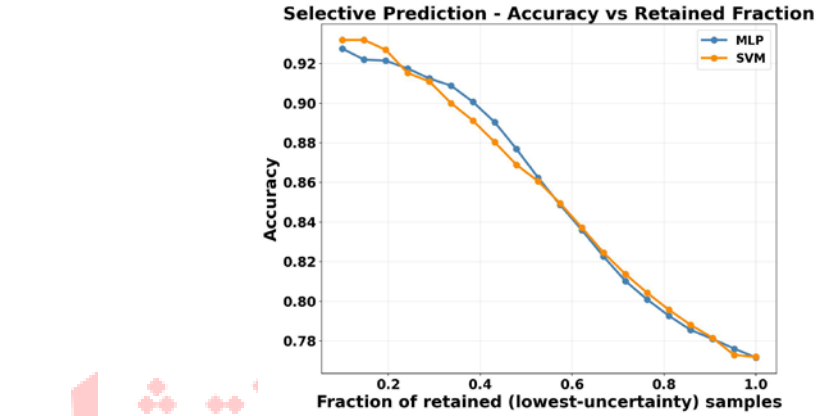
نگاره ۷. مقایسه عدم‌قطعیت و درستی برای (الف) DeepMLP و (ب) SVM

کاربرد عملی UQ در منحنی پیش‌بینی گزینشی (نگاره ۸) نمایان می‌شود: با فیلتر کردن ۵۰٪ نمونه‌های با عدم‌قطعیت بالا، دقت از ۰/۷۷~ به ۰/۸۷~ افزایش می‌یابد (۱۰+ واحد درصد). این بهبود در هر دو مدل تقریباً یکسان است (۰/۸۷۰٤ در برابر ۰/۸٦٤۹) که یک‌بار دیگر هم‌ارزی عملی دو مدل را تأیید می‌کند.

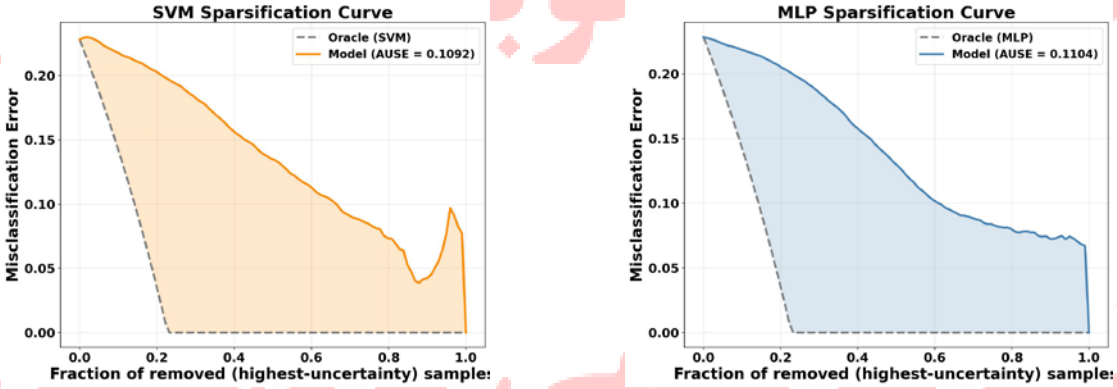
برای تحلیل دقیق‌تر، منحنی پراکنده‌سازی هر مدل به‌طور جداگانه رسم شد (نگاره‌های ۹–الف و ۹–ب) که مساحت بین منحنی مدل و منحنی آرمانی همان AUSE است. در مجموع، هر دو مدل از سیگنال عدم‌قطعیت "متوسط اما مفید" برخوردارند: مقدار AUSE ≈۰/۱۱ فاصله قابل‌توجهی از حالت ایده‌آل دارد، اما همین سطح از UQ کافی است تا با فیلتر کردن نمونه‌های

[16] Pearson
[17] Spearman

با عدم‌قطعیت بالا، دقت تصمیم‌گیری ۱۰٪ بهبود یابد. این یافته برای کاربردهای عملیاتی اهمیت بالایی دارد، چرا که نشان می‌دهد حتی UQ متوسط می‌تواند در سامانه‌های هشدار سریع ارزش افزوده چشمگیری ایجاد کند.



نگاره ۸. منحنی پیش‌بینی گزینشی. با فیلتر ۵۰٪ نمونه‌های با عدم قطعیت بالا، دقت به≈۰/۸۷ می‌رسد.



(ب) (الف)

نگاره ۹. منحنی پراکنده‌سازی (AUSE) برای (الف) DeepMLP و (ب) SVM.

٤-٦- منحنی یادگیری

برای بررسی اینکه حجم داده، عامل محدودکننده عملکرد مدل‌ها است یا خیر، منحنی‌های یادگیری برای هر دو مدل DeepMLP و SVM به‌طور جداگانه ترسیم شد. در این تحلیل، مدل‌ها روی زیرمجموعه‌های آموزشی با اندازه‌های متفاوت دوباره آموزش دیدند و عملکرد آن‌ها روی مجموعه‌های اعتبارسنجی و آزمون ارزیابی شد. جدول ۹ نتایج DeepMLP و جدول ۱۰ نتایج SVM را نشان می‌دهند.

جدول ۹. منحنی یادگیری DeepMLP (۳۰ اپوک برای هر اندازه).

اندازه نمونه	Val AUC	Val Brier	زمان آموزش(s)
۵۰،۰۰۰	۰/۸٤۰٤	۰/۱۷٤۷	۵/٦
۱۰۰،۰۰۰	۰/۸٤۱۲	۰/۱۷٦۰	٦/۸
۲۵۰،۰۰۰	۰/۸۳۹۹	۰/۱۷۱۵	۱۱/۱
۵۰۰،۰۰۰	۰/۸۳۹۹	۰/۱۷۲۲	۲۰/۵
۱۰۰۰،۰۰۰	۰/۸۳۸۸	۰/۱۷۲۲	۳۵/۵
۱۵۰۰،۰۰۰	۰/۸٤۱۲	۰/۱۷۰۷	۵۳/٤

جدول ۱۰. منحنی یادگیری SVM (زیرنمونه‌های مختلف).

اندازه نمونه	Val AUC	Test AUC	Test Brier	زمان آموزش(s)

۵۰٬۰۰۰	۰/۸٤۰۲	۰/۸٤۲۰	۰/۱٦٦۲	٤/٤
۱۰۰٬۰۰۰	۰/۸٤۱۵	۰/۸٤۳۲	۰/۱٦٦۳	۱۹/۰
۲۰۰٬۰۰۰	۰/۸٤۰۸	۰/۸٤۱۲	۰/۱٦۷۲	۸۹/۵
۳۰۰٬۰۰۰	۰/۸۳۹۳	۰/۸۳۸٦	۰/۱٦۷۹	۲۰۷/۲
۵۰۰٬۰۰۰	۰/۸٤۱٦	۰/۸٤۰۳	۰/۱٦۷۲	٦۳۲/٦

همان‌طور که در جدول ۹ و ۱۰ ملاحظه می‌شود، هر دو مدل در حدود ۵۰٬۰۰۰ تا ۱۰۰٬۰۰۰ نمونه اشباع می‌شوند. در مورد DeepMLP، افزایش ۳۰ برابری در حجم داده آموزشی (از ۵۰٬۰۰۰ به ۱۵۰۰٬۰۰۰ نمونه) کمتر از ۰/۳٪ بهبود در AUC ایجاد می‌کند، در حالی که زمان آموزش ۱۰ برابر می‌شود. در مورد SVM نیز، بهترین عملکرد در ۱۰۰٬۰۰۰ نمونه حاصل می‌شود و پس از آن، با افزایش داده تا ۵۰۰٬۰۰۰ نمونه نه‌تنها بهبودی مشاهده نمی‌شود، بلکه در برخی اندازه‌ها افت جزئی نیز رخ می‌دهد، در حالی که زمان آموزش ۳۳ برابر می‌شود.

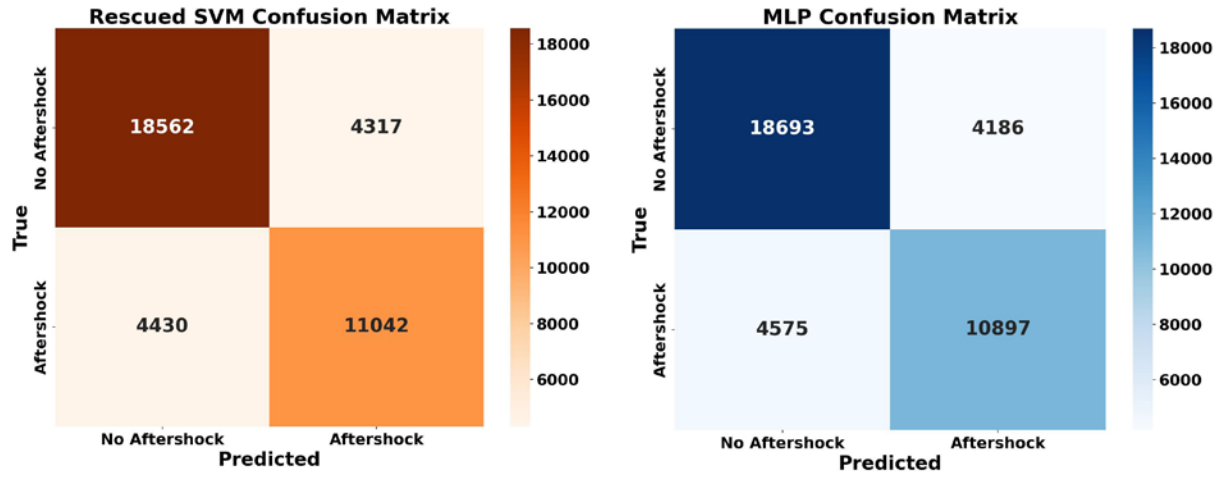
این یافته مشترک، یک نتیجه عملی مهم دارد: عامل محدودکننده عملکرد مدل‌های فعلی، حجم داده نیست بلکه نمایش ویژگی است. یعنی ۱۲ مؤلفه تانسور ممان، حداکثر اطلاعات قابل استخراج را برای این مسئله فراهم می‌کنند و افزودن نمونه‌های بیشتر از همان فضای ویژگی، سقف عملکرد را جابه‌جا نمی‌کند. برای عبور از این سقف، باید ویژگی‌های جدید (مانند تنش دینامیکی، ویژگی‌های هندسی گسل، یا اطلاعات پوسته‌ای منطقه‌ای) به مدل معرفی شوند. این نتیجه با یافته‌های Zhao و همکاران (۲۰۲۲) هم‌خوانی دارد. از منظر کارایی محاسباتی نیز، انتخاب زیرنمونه ۱۰۰٬۰۰۰ نمونه برای SVM کاملاً توجیه‌پذیر است: این اندازه کوچک‌ترین حجمی است که به عملکرد کامل می‌رسد و در نتیجه ۱۵ برابر سریع‌تر از آموزش روی ۵۰۰٬۰۰۰ نمونه است، بدون هیچ افت قابل‌اندازه‌گیری در معیارهای عملکرد.

٤-۷- ماتریس‌های درهم‌ریختگی

برای درک الگوی خطاهای دو مدل در سطح هر کلاس، ماتریس‌های درهم‌ریختگی DeepMLP و SVM بهبودیافته روی مجموعه آزمون محاسبه شد (نگاره‌های ۱۰-الف و ۱۰ب). دو مدل الگوهای خطای مکمل و متمایزی نشان می‌دهند: DeepMLP با دقت بالاتر در تشخیص سلول‌های بدون پسلرزه (ویژگی ۰/۸۱۷)، به‌سمت محافظه‌کاری بیشتر در اعلام پسلرزه تمایل دارد، در حالی که SVM بهبودیافته با حساسیت بالاتر در کشف سلول‌های دارای پسلرزه (بازیابی ۰/۷۱٤)، موازنه متفاوتی بین حساسیت و ویژگی برقرار می‌کند. این تفاوت، پیامد مستقیم آستانه تصمیم متفاوت دو مدل است (۰/۳۸۵ برای DeepMLPدر برابر ۰/٤۰۳ برای SVM).

از منظر عملیاتی، این تفکیک نقش به دو سناریوی متفاوت پاسخ می‌دهد: در سامانه‌های هشدار سریع که کاهش هشدارهای اشتباه (FP) اولویت دارد، DeepMLP با FP کمتر گزینه مناسب‌تری است، چرا که هر هشدار نادرست هزینه‌های اجتماعی و اقتصادی قابل‌توجهی ایجاد می‌کند. در مقابل، در تحلیل‌های علمی که شناسایی حداکثری مناطق پرخطر (کمینه‌سازی FN) اهمیت دارد، SVM بهبودیافته ترجیح داده می‌شود. این تفکیک نقش، همراه با معیارهای تفکیک‌پذیری و کالیبراسیون که پیش‌تر بررسی شد، تصویری کامل از مزیت‌ها و محدودیت‌های هر مدل ارائه می‌دهد و به کاربران امکان انتخاب آگاهانه بر اساس اولویت‌های خاص خود را می‌دهد.

مقاله در نوبت انتشار
نشریه علمی پژوهشی
« اطلاعات جغرافیایی (سپهر) »



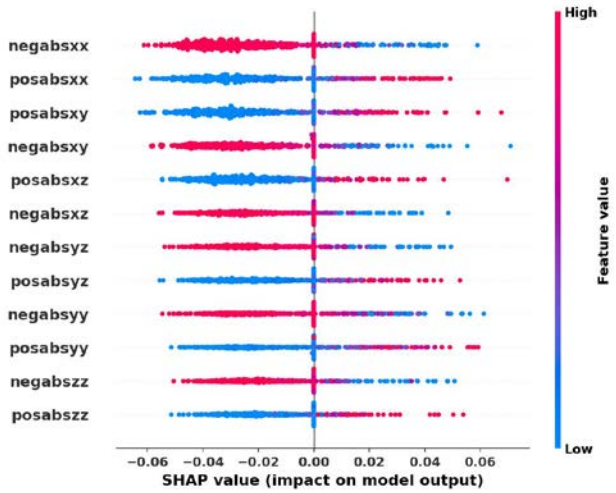
(الف) (ب)

نگاره ۱۰. ماتریس درهم‌ریختگی برای (الف) DeepMLP و (ب) SVM بهبودیافته.

٤-٨- تحلیل تفسیرپذیری با SHAP

برای شناسایی ویژگی‌های مؤثر در تصمیم‌گیری SVM بهبودیافته، از تحلیل مقادیر SHAP با استفاده از KernelExplainer روی ۵۰۰ نمونه از مجموعه آزمون و ۱۰۰ نمونه پس‌زمینه بهره گرفته شد. نگاره ۱۱ خلاصه گرافیکی این تحلیل را ارائه می‌دهد و نتایج نشان می‌دهد که مدل SVM به‌طور غالب بر مؤلفه‌های برشی (به‌ویژه negabsxy و posabsxy) تکیه می‌کند، در حالی که مؤلفه‌های نرمال (مانند posabsyy و negabsyy) کمترین سهم را در تصمیم‌گیری دارند. این یافته با مبانی فیزیکی پسلرزه‌ها هم‌خوانی دارد: مؤلفه‌های برشی تنش نقش مستقیم‌تری در لغزش گسل و فعال‌سازی پسلرزه‌ها ایفا می‌کنند، در حالی که مؤلفه‌های نرمال عمدتاً بر باز و بسته شدن شکاف‌ها اثر می‌گذارند و تأثیر کمتری بر چکانش دارند.

برتری نسبی مؤلفه‌های برشی افقی بر برشی قائم و نرمال نشان می‌دهد که SVM هندسه سه‌بعدی تانسور ممان را به شکلی سازگار با سامانه‌های گسل‌خیز در مقیاس جهانی که در آن‌ها سازوکارهای امتدادلغز با مؤلفه‌های برشی قابل توجه غالب هستند یاد گرفته است. از آنجا که داده‌ها شامل طیف متنوعی از رژیم‌های زمین‌ساختی (از گسل‌های امتدادلغز قاره‌ای تا مناطق فرورانش) است، این یافته نشان می‌دهد که SVM توانسته است الگوهای مشترک میان انواع مختلف گسل‌ها را کشف کند، نه فقط الگوهای خاص یک منطقه محدود. توزیع نسبتاً متعادل SHAP در میان ده ویژگی نیز حاکی از آن است که مدل از یک راهبرد «رأی‌گیری توزیع‌شده» بر روی چندین مؤلفه استفاده می‌کند که به پایداری مدل در برابر نوفه و افزایش تفسیرپذیری لرزه‌ای آن کمک می‌کند.



نگاره ۱۱. خلاصه SHAP برای SVM بهبودیافته.

۵- بحث و یافته‌های تحقیق

۵-۱- پارادوکس McNemar و DeLong

یکی از یافته‌های کلیدی این مطالعه، تفکیک بنیادین میان توانایی رتبه‌بندی و تصمیم‌گیری دودویی است. دو آزمون آماری DeLongو McNemar تصویری متفاوت از برتری SVM ارائه می‌دهند: SVM بهبود یافته به‌طور معنادار رتبه‌بندی احتمال‌ها را بهتر انجام می‌دهد، اما وقتی این احتمال‌ها به تصمیم‌های دودویی در آستانه بهینه تبدیل می‌شوند، مزیت آن ناپدید می‌شود.

این تفاوت ظاهری، در واقع پارادوکس نیست، بلکه بازتاب دو مفهوم متفاوت در ارزیابی است. آزمون DeLong تفاوت را در کل فضای آستانه‌ها می‌سنجد، در حالی که McNemar تنها تصمیم نهایی در یک آستانه مشخص را مقایسه می‌کند. وقتی SVMدر رتبه‌بندی بهتر عمل می‌کند اما این برتری به افزایش معنادار تصمیم‌های درست نمی‌رسد، دو آزمون به نتایج متفاوتی می‌رسند.

بسیاری از مطالعات پیشین تنها به McNemar یا دقت کلی بسنده می‌کنند و از آزمون‌های مبتنی بر رتبه غافل می‌مانند. این رویکرد می‌تواند ظرفیت واقعی مدل را پنهان کند. به‌ویژه در کاربردهای حساس مانند سامانه‌های هشدار سریع لرزه‌ای که ترتیب و کالیبراسیون احتمال‌ها اهمیت مستقیم دارند، گزارش همزمان DeLong و McNemar تصویری کامل‌تر ارائه می‌دهد.

۵-۲- کالیبراسیون احتمال

کالیبراسیون احتمال یکی از جنبه‌های حیاتی در سامانه‌های پیش‌بینی پسلرزه است که در ادبیات موجود کمتر مورد توجه قرار گرفته است؛ چرا که یک مدل با AUC بالا می‌تواند هم‌زمان احتمال‌هایی بسیار دور از واقعیت تولید کند که برای تصمیم‌گیری عملیاتی غیرقابل‌اعتماد باشند. به همین دلیل، در این مطالعه سه معیار متمایز کالیبراسیون یعنی امتیاز score، NLL و RMSCE روی مجموعه آزمون ارزیابی شد.

SVM بهبودیافته در هر دو معیار Brier و NLL برتری واضحی نسبت به DeepMLP دارد، به‌گونه‌ای که بازه اطمینان ۹۵٪ این دو معیار برای دو مدل هیچ همپوشانی ندارند، نشانه‌ای قوی از معناداری آماری برتری SVM در کالیبراسیون. این یافته در تناقض ظاهری با نتیجه RMSCE است که DeepMLP را برنده معرفی می‌کند. ریشه این تناقض در ماهیت آماری RMSCEنهفته است: برخلاف Brier و NLL که قواعد امتیازدهی مناسب هستند یعنی کمینه آن‌ها تنها در نقطه کالیبراسیون کامل رخ می‌دهد و هیچ مدل دیگری نمی‌تواند امتیاز بهتری بگیرد، RMSCE این ویژگی را ندارد. RMSCE با تقسیم فضای احتمال به بازه‌های احتمال مساوی و محاسبه دقت درون هر بازه در آستانه ثابت ۰/۵، به دو عامل حساس می‌شود: تعداد بازه‌ها و نرخ پایه داده. در داده حاضر با نرخ مثبت ۰/٤۰، RMSCE حتی برای یک مدل کاملاً کالیبره نیز بیش‌برآورد می‌شود. این مسئله پیش‌تر توسط Nixon و همکاران (۲۰۱۹) و Vaicenavicius و همکاران (۲۰۱۹) به‌تفصیل بررسی شده و توصیه صریح آن‌ها این است که ECE و مشتقات آن (از جمله RMSCE) تنها به‌عنوان مکمل Brier و NLL گزارش شوند، نه به‌عنوان معیار اصلی.

مقاله در نوبت انتشار
نشریه علمی پژوهشی
« اطلاعات جغرافیایی (سپهر) »

بر این اساس، در این مطالعه Brier و NLL را به‌عنوان معیارهای اصلی کالیبراسیون انتخاب کرده‌ایم و RMSCE را صرفاً برای شفافیت و کامل بودن گزارش آورده‌ایم، با تأکید بر محدودیت تفسیری آن.

۵-۳- اشباع سریع منحنی یادگیری

منحنی‌های یادگیری هر دو مدل نشان می‌دهند که عملکرد آن‌ها در حجم نسبتاً محدودی از داده به اشباع می‌رسد؛ به‌گونه‌ای که افزایش چشمگیر حجم داده آموزشی نه‌تنها بهبود محسوسی در معیارهای تفکیک‌پذیری ایجاد نمی‌کند، بلکه در برخی موارد به افت جزئی عملکرد نیز منجر می‌شود. این رفتار به‌روشنی بیانگر آن است که عامل محدودکننده اصلی در این مجموعه داده، نمایش ویژگی است، نه حجم داده. به بیان دیگر، مدل‌های مورد استفاده توانسته‌اند تقریباً تمام سیگنال قابل استخراج از مؤلفه‌های تانسور ممان را در همان حجم محدود اولیه جذب کنند و افزودن داده‌های بیشتر تنها به افزایش هزینه محاسباتی بدون بهبود عملکرد منجر می‌شود.

این یافته با نتیجه‌گیری Zhao و همکاران (۲۰۲۲) هم‌خوانی کامل دارد که نشان دادند در پیش‌بینی توزیع مکانی پسلرزه‌ها، انتخاب و مهندسی ویژگی‌ها نقش بسیار مهم‌تری از افزایش حجم داده یا پیچیدگی مدل ایفا می‌کند. بر این اساس، مسیر بهبود عملکرد در پژوهش‌های آینده نه در جمع‌آوری داده‌های بیشتر، بلکه در ساخت ویژگی‌های ترکیبی جدید، به‌ویژه از تلفیق تنش ایستا و پویا و همچنین استخراج ویژگی‌های مبتنی بر ناهمگنی محیطی و هندسه گسل‌های گیرنده، نهفته است.

از منظر عملی، این یافته پیامد مستقیمی برای طراحی نحوه آموزش SVM دارد: انتخاب اندازه زیرنمونه در این مطالعه انتخابی بهینه بوده است، چرا که ضمن حفظ عملکرد کامل مدل، زمان آموزش را به‌طور چشمگیری کاهش داد. این کاهش قابل توجه، امکان به‌کارگیری SVM را در سناریوهای عملیاتی با منابع محدود یا نیاز به آموزش‌های مکرر فراهم می‌آورد.

۵-٤- کارایی کمّی‌سازی عدم‌قطعیت

تحلیل UQ نشان می‌دهد که هر دو مدل کیفیت کمّی‌سازی عدم‌قطعیت متوسطی دارند؛ مقدار AUSE برای هر دو مدل تقریباً یکسان و در حدود ۰/۱ است که نشان می‌دهد منحنی پراکنده‌سازی مدل‌ها فاصله محسوسی از منحنی آرمانی دارد. با این حال، همین سیگنال متوسط در کاربردهای عملی ثمربخش است: با فیلتر کردن ۵۰٪ نمونه‌های با عدم قطعیت بالا، دقت هر دو مدل حدود ۱۰ درصد افزایش می‌یابد. این بهبود، در کاربردهایی مانند غربالگری اولیه مناطق پرخطر یا اولویت‌بندی سلول‌ها برای بازرسی دقیق‌تر، ارزش عملی قابل‌توجهی دارد.

نکته متمایزکننده دو مدل در نحوه حمل این اطلاعات نهفته است. DeepMLP دامنه گسترده‌تری از عدم‌قطعیت را تجربه می‌کند، در حالی که SVM بهبودیافته توزیع عدم‌قطعیت بسیار متمرکزتری دارد، تا حدی که میانگین انحراف معیار آن به حدود یک‌سوم DeepMLP می‌رسد. با وجود این تفاوت چشمگیر در مقیاس، AUSE دو مدل تقریباً یکسان است. این مشاهده به یک نتیجه مهم منجر می‌شود: هر واحد انحراف معیار در SVM اطلاعات بیشتری حمل می‌کند، یعنی SVM با تغییرات کوچک‌تر در عدم‌قطعیت، همان قابلیت تمایزگذاری بین نمونه‌های سخت و آسان را فراهم می‌آورد. این ویژگی از دو منظر حائز اهمیت است: نخست، از منظر عملی، آستانه‌گذاری بر روی عدم‌قطعیت SVM ساده‌تر و پایدارتر است؛ دوم، از منظر نظری، نشان می‌دهد که کالیبراسیون ایزوتونیک در SVM توزیع احتمال را به‌گونه‌ای تنظیم کرده که بخش عمده اطلاعات مرتبط با اطمینان در بازه کوچکی متمرکز شود.

مقاله در نوبت انتشار
نشریه علمی پژوهشی
« اطلاعات جغرافیایی (سپهر) »

از سوی دیگر، همبستگی بین عدم‌قطعیت و خطای پیش‌بینی نیز تصویر متفاوتی برای دو مدل ترسیم می‌کند. هر دو مدل همبستگی مثبت و معنادار بین عدم‌قطعیت و خطا نشان می‌دهند (شاهدی بر معتبر بودن UQ)، و همبستگی اسپیرمن آن‌ها تقریباً یکسان است، که حاکی از توانایی مشابه دو مدل در رتبه‌بندی نمونه‌ها بر اساس عدم‌قطعیت است. در مجموع، با وجود متوسط بودن کیفیت UQ در هر دو مدل، SVM بهبودیافته به دلیل فشردگی توزیع عدم‌قطعیت و توانایی حمل اطلاعات متراکم‌تر در هر واحد Std، گزینه مناسب‌تری برای استقرار در سامانه‌های عملیاتی است که به فیلترکردن کارآمد نمونه‌های پرخطر نیاز دارند.

۵-۵- کارایی محاسباتی

کارایی محاسباتی دو مدل در سه مرحله کلیدی یعنی آموزش، استنتاج و کمّی‌سازی عدم‌قطعیت ارزیابی شد. الگوی کارایی دو مدل کاملاً متضاد است و هر مدل در مرحله‌ای متفاوت برتری چشمگیری دارد. در مرحله آموزش، SVM بهبودیافته به‌دلیل زیرنمونه‌گیری هوشمندانه، حدود ۲۰ برابر سریع‌تر از DeepMLP آموزش می‌بیند. این برتری در سناریوهایی که نیاز به آموزش‌های مکرر (مانند به‌روزرسانی دوره‌ای مدل یا اجرای متعدد جستجوی فراپارامتر) وجود دارد، از ارزش عملی بالایی برخوردار است.

در مرحله استنتاج، الگو معکوس می‌شود. SVM به‌دلیل ماهیت محاسباتی خود که برای هر نمونه آزمون باید مقدار تابع هسته را نسبت به تمام بردارهای پشتیبان محاسبه کند، حدود ٥٦ برابر کندتر از DeepMLP پیش‌بینی می‌کند. این تفاوت فاحش ناشی از پیچیدگی محاسباتی $O(n_{SV} \cdot d)$ در SVM در برابر عبور یک‌باره از شبکه عصبی در DeepMLP است.

در مرحله کمّی‌سازی عدم‌قطعیت نیز DeepMLP با اختلاف حدود ۸ برابر سریع‌تر عمل می‌کند، چرا که MC Dropout تنها به ۱۰۰ بار عبور از شبکه آموزش‌دیده نیاز دارد، در حالی که بوت‌استرپ برای SVM مستلزم آموزش مجدد ده‌ها مدل کالیبره‌شده روی زیرنمونه‌های مختلف است.

بر پایه این یافته‌های متضاد، توصیه عملیاتی به‌شرح زیر است: برای استقرار در سیستم‌های هشدار سریع که نیاز به پردازش حجم وسیعی از سلول‌های شبکه در کوتاه‌ترین زمان ممکن دارند، DeepMLP به‌دلیل برتری چشمگیر در زمان استنتاج، گزینه ترجیحی است. در مقابل، برای محیط‌های با منابع محاسباتی محدود یا کاربردهایی که در آن‌ها زمان آموزش و به‌روزرسانی مدل اهمیت دارد، SVM بهبودیافته به‌دلیل برتری قابل توجه در زمان آموزش، انتخاب مناسب‌تری محسوب می‌شود.

۵-٦- مقایسه با ادبیات موجود

مقایسه یافته‌های این مطالعه با پژوهش‌های پیشین، تصویری چندلایه از جایگاه SVM بهبودیافته در ادبیات موجود ارائه می‌دهد. از یک سو، نتایج ما با یافته‌های DeVries و همکاران (۲۰۱۸) هم‌خوانی قابل‌توجهی دارد؛ آنان با استفاده از معماری DeepMLPبه AUC نزدیک به ۰/۸۵ روی مجموعه آزمون مستقل دست یافتند و ما نیز در این مطالعه به مقادیر مشابهی برای DeepMLP و SVM بهبودیافته رسیدیم. این هم‌خوانی، اعتبار داخلی نتایج ما را تأیید می‌کند و نشان می‌دهد پیاده‌سازی ما با همان استانداردهای دقت انجام شده است.

از سوی دیگر، یافته اصلی این مطالعه با نتیجه‌گیری Mignan و Broccardo (2020) تفاوت معناداری دارد. آنان نشان دادند یک رگرسیون لجستیک ساده با تنها دو پارامتر می‌تواند به عملکردی قابل رقابت با شبکه‌های عصبی عمیق دست یابد و نتیجه گرفتند که پیچیدگی معماری عمیق برای داده‌های تنش ایستا ضرورتی ندارد. در مطالعه ما، اگرچهDeepMLP و SVM در

تصمیم‌های دودویی خود تفاوت معناداری نداشتند، اما SVM بهبودیافته به‌طور معنادار توانایی رتبه‌بندی و کالیبراسیون بهتری نشان داد. این یافته بیانگر آن است که برخلاف تصور رایج در ادبیات پیشین، SVM با کالیبراسیون مناسب و انتخاب درست معیار بهینه‌سازی می‌تواند نه‌تنها با DeepMLP رقابت کند، بلکه در برخی جنبه‌های کلیدی از آن پیشی بگیرد.

مهم‌ترین پیام نظری این مقایسه آن است که بخش قابل‌توجهی از شکست‌های SVM در مطالعات پیشین، نه به دلیل ضعف ذاتی این الگوریتم، بلکه ناشی از دو خطای روش‌شناختی رایج بوده است: نخست، انتخاب نادرست معیار بهینه‌سازی فراپارامترها (مانند F1 که به آستانه تصمیم حساس است)، و دوم، عدم کالیبراسیون احتمال خروجی SVM که منجر به احتمال‌های نزدیک به صفر یا یک می‌شود. در این مطالعه، رفع همزمان این دو مشکل، به‌همراه وزن‌دهی متعادل کلاس‌ها، گسترش فضای جستجوی پارامتر هسته و کاهش هوشمندانه حجم زیرنمونه توانست SVM را به مدلی رقابتی و در برخی معیارها برتر تبدیل کند. این یافته، درس روش‌شناختی مهمی برای پژوهش‌های آینده مبتنی بر SVM فراهم می‌آورد: قضاوت درباره عملکرد یک الگوریتم، بدون توجه به روند پیش‌پردازش و کالیبراسیون آن، می‌تواند به نتیجه‌گیری‌های گمراه‌کننده منجر شود.

مقاله در نوبت انتشار

۵-۷- توصیه‌های انتخاب مدل

انتخاب میان DeepMLP و SVM بهبودیافته، برخلاف بسیاری از مقایسه‌های پیشین که به دنبال یک «برنده مطلق» هستند، ماهیتی وابسته به زمینه دارد. دو مدل از نظر آماری در تصمیم‌های دودویی معادل‌اند، اما در ابعاد تفکیک‌پذیری، کالیبراسیون، کارایی محاسباتی و الگوی خطا الگوهای مکمل و متمایزی دارند؛ بنابراین تصمیم‌گیرنده باید مصالحه میان این ابعاد را بر اساس کاربرد مورد نظر ارزیابی کند.

نشریه علمی پژوهشی

در سامانه‌های هشدار سریع که کاهش هشدارهای اشتباه اولویت دارد، DeepMLP گزینه مناسب‌تری است، چرا که هم دقت مثبت و هم ویژگی بالاتری دارد و در نتیجه هشدارهای نادرست کمتری تولید می‌کند. در چنین کاربردهایی مانند اطلاع‌رسانی عمومی، تخلیه موقت مناطق مسکونی یا توقف فعالیت‌های صنعتی که هزینه اجتماعی و اقتصادی هر هشدار نادرست قابل توجه است، این مزیت می‌تواند تعیین‌کننده باشد. در مقابل، برای کاربردهایی که رتبه‌بندی احتمال یا کشف حداکثری پسلرزه اهمیت دارد، SVM بهبودیافته ترجیح داده می‌شود؛ چرا که هم توانایی رتبه‌بندی آن به‌طور معنادار بهتر است و هم کالیبراسیون احتمال آن با اطمینان بالا از DeepMLP بهتر است. این ویژگی‌ها SVM را به گزینه‌ای مناسب برای تحلیل‌های علمی، مطالعات مخاطره و سامانه‌هایی که به احتمال‌های کالیبره‌شده برای تصمیم‌گیری وزنی نیاز دارند، تبدیل می‌کند.

« اطلاعات جغرافیایی (سپهر) »

از منظر کارایی محاسباتی، مصالحه‌ای مشابه اما معکوس مشاهده می‌شود. SVM بهبودیافته به دلیل زیرنمونه‌گیری هوشمندانه، حدود ۲۰ برابر سریع‌تر از DeepMLP آموزش می‌بیند؛ این مزیت در سناریوهایی که نیاز به آموزش‌های مکرر (اعتبارسنجی متقابل، تنظیم مجدد فراپارامترها یا به‌روزرسانی دوره‌ای) وجود دارد، صرفه‌جویی قابل توجهی ایجاد می‌کند. در مقابل، DeepMLP در مرحله استنتاج به‌طور چشمگیری سریع‌تر عمل می‌کند؛ این تفاوت در کاربردهایی که نیاز به پیش‌بینی روی میلیون‌ها سلول شبکه در زمان واقعی دارند، به یک مزیت تعیین‌کننده به نفع DeepMLP تبدیل می‌شود.

در نهایت، در محیط‌های با منابع محاسباتی محدود (ایستگاه‌های دورافتاده، سامانه‌های تعبیه‌شده یا سرورهای بدون GPU)، SVM بهبودیافته به دلیل مصرف حافظه کمتر و آموزش سریع‌تر عملی‌تر است، در حالی که برای کاربردهایی با نیاز به استنتاج پرحجم و مکرر، DeepMLP به دلیل سرعت بالای پیش‌بینی برتری دارد. به‌طور خلاصه، هیچ‌کدام از دو مدل را نمی‌توان «برنده مطلق» معرفی کرد؛ مدل مناسب، تابعی از اولویت‌های کاربرد است: دقت در برابر حساسیت، کالیبراسیون در

برابر سرعت، و هزینه آموزش در برابر هزینه استنتاج. این تنوع گزینه‌ها به مدیران بحران و توسعه‌دهندگان سامانه‌ها امکان می‌دهد بر اساس محدودیت‌های واقعی خود، آگاهانه‌ترین انتخاب را انجام دهند.

٦- نتیجه‌گیری

این مطالعه یک تحلیل مقایسه‌ای جامع بین DeepMLP شش‌لایه و SVM بهبود یافته با هسته RBF برای برآورد مکانی پسلرزه ارائه داد. یافته‌های اصلی نشان می‌دهد SVM بهبودیافته به‌طور معنادار توانایی رتبه‌بندی بهتری دارد. با این حال، آزمون McNemarنشان داد دو مدل در تصمیم‌های دودویی معادل‌اند؛ این پارادوکس ظاهری، نکته روش‌شناختی مهمی را آشکار می‌سازد: مزیت SVM در رتبه‌بندی و کالیبراسیون نهفته است، نه در مرز تصمیم نهایی.

تحلیل منحنی‌های یادگیری نشان داد هر دو مدل در حدود ۵۰٬۰۰۰ تا ۱۰۰٬۰۰۰ نمونه به اشباع می‌رسند و افزایش سی‌برابری داده، بهبود قابل‌توجهی ایجاد نمی‌کند. این یافته حاکی از آن است که عامل محدودکننده اصلی، نمایش ویژگی است نه حجم داده. در زمینه کمّی‌سازی عدم‌قطعیت، هر دو مدل سیگنال متوسطی ارائه دادند، اما فیلتر کردن نیمی از نمونه‌های با عدم‌قطعیت بالا، دقت را حدود ۱۰ واحد درصد افزایش داد که ارزش عملی UQ را تأیید می‌کند. SVM با عدم‌قطعیت کم‌تنوع‌تر و حفظ همان سطح AUSE، نشان داد هر واحد انحراف معیار آن اطلاعات بیشتری حمل می‌کند. از منظر کارایی محاسباتی نیز، SVM بسیار سریع‌تر آموزش می‌بیند، در حالی که DeepMLP در استنتاج سریع‌تر است.

بر اساس این یافته‌ها، سه راهکار عملیاتی پیشنهاد می‌شود: (۱) برای سیستم‌های هشدار سریع که کاهش هشدارهای اشتباه اولویت دارد، DeepMLP توصیه می‌شود؛ (۲) برای تحلیل‌های احتمالاتی و کاربردهایی که رتبه‌بندی خطر اهمیت بیشتری دارد، SVM بهبودیافته گزینه برتر است؛ (۳) فیلتر UQ در روند سامانه‌ها قرار گیرد تا با حذف نمونه‌های با عدم‌قطعیت بالا، اطمینان تصمیم‌گیری به‌طور معناداری افزایش یابد.

برای کارهای آینده، پنج محور اصلی پیشنهاد می‌شود: (۱) مهندسی ویژگی با تلفیق تنش ایستا و پویا؛ (۲) یادگیری انتقالی برای تعمیم مدل به مناطق جدید؛ (۳) ترکیب مدل‌ها با وزن‌دهی مبتنی بر عدم‌قطعیت؛ (٤) روش‌های تفسیرپذیری نظیر SHAP و LIME؛ و (۵) در نظر گرفتن همبستگی فضایی سلول‌های مجاور از طریق معماری‌های CNN یا GNN.

مقاله در دست انتشار
نشریه علمی پژوهشی
«اطلاعات جغرافیایی (سپهر)»

منابع

Bath, M. (1965). Lateral inhomogeneities of the upper mantle. *Tectonophysics*, 2(6), 483–514.

DeVries, P. M., Viégas, F., Wattenberg, M., & Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720), 632–634.

Felzer, K. R., & Brodsky, E. E. (2006). Decay of aftershock density with distance indicates triggering by dynamic stress. *Nature*, 441(7094), 735–738.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Hainzl, S., Enescu, B., Cocco, M., Woessner, J., Catalli, F., Wang, R., & Roth, F. (2009). Aftershock modeling based on uncertain stress calculations. *Journal of Geophysical Research*, 114, B05309.

Hardebeck, J. L., Nazareth, J. J., & Hauksson, E. (1998). The static stress change triggering model: Constraints from two southern California aftershock sequences. *Journal of Geophysical Research*, 103(B10), 24427–24437.

Harris, R. A., & Simpson, R. W. (1992). Changes in static stress on southern California faults after the 1992 Landers earthquake. *Nature*, 360(6401), 251–254.

Karimzadeh, S., Matsuoka, M., Kuang, J., & Ge, L. (2019). Spatial prediction of aftershocks triggered by a major earthquake: A binary machine learning perspective. *ISPRS International Journal of Geo-Information*, 8(10), 462.
King, G. C. P., Stein, R. S., & Lin, J. (1994). Static stress changes and the triggering of earthquakes. *Bulletin of the Seismological Society of America*, 84(3), 935–953.
Liu, B., Wen, H., Di, M., Huang, J., Liao, M., Yu, J., & Xiang, Y. (2024). Mapping and interpretability of aftershock hazards using hybrid machine learning algorithms. *Journal of Rock Mechanics and Geotechnical Engineering*, 16(2), 245–261.
Mallman, E. P., & Zoback, M. D. (2007). Assessing elastic Coulomb stress transfer models using seismicity rates in southern California and southwestern Japan. *Journal of Geophysical Research*, 112, B03304.
Meade, B. J., DeVries, P. M., Faller, J., Viegas, F., & Wattenberg, M. (2017). What is better than Coulomb failure stress? A ranking of scalar static stress triggering mechanisms from 105 mainshock-aftershock pairs. *Geophysical Research Letters*, 44(22), 11409–11416.
Mignan, A., & Broccardo, M. (2020). Neural network applications in earthquake prediction: A review and a new approach. *Frontiers in Earth Science*, 8, 573–586.
Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., & Tran, D. (2019). Measuring calibration in deep learning. *CVPR Workshops*, 2(5), 1–8.
Okada, Y. (1992). Internal deformation due to shear and tensile faults in a half-space. *Bulletin of the Seismological Society of America*, 82(2), 1018–1040.
Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61–74.
Sun, X., & Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11), 1389–1393.
Utsu, T. (1961). A statistical study on the occurrence of aftershocks. *Geophysical Magazine*, 30, 521–605.
Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. B. (2019). Evaluating model calibration in classification. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 3459–3467.
Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.
Zhao, S., Wang, H., Xue, Y., Wang, Y., Li, S., Liu, J., & Pan, G. (2022). What are more important for aftershock spatial distribution prediction, features, or models? A case study in China. *Journal of Seismology*, 26(1), 181–196.
Zhou, Z., Wang, M., Han, M., Yu, X., & Lu, D. (2024). Prediction of damage potential in mainshock-aftershock sequences using machine learning algorithms. *Earthquake Engineering and Engineering Vibration*, 23(4), 919–938.

مقاله در نوبت انتشار
نشریه علمی پژوهشی
«اطلاعات جغرافیایی (سپهر)»