نشریه علمی علوم و فنون نقشه‌برداری، دوره ۱۵، شماره ۳، صفحه ۱ تا ۱۵، اسفند ماه ۱۴۰۴ - مقاله پژوهشی

مقایسه عملکرد الگوریتم‌های یادگیری ماشین در پیش‌بینی توزیع مکانی بیماری سالک بر رویکرد نقشه های دوگانه

زینب نیسانی سامانی [۱]، علی اصغر آل شیخ [۲*]

[۱] دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه نصیرالدین طوسی
zeinab.neisani@email.kntu.ac.ir

[۲] دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه نصیرالدین طوسی
alesheikh@kntu.ac.ir

(دریافت: آبان ۱۴۰۴، تصویب: آذر ۱۴۰۴)

چکیده

لیشمانیوز جلدی به‌عنوان یک بیماری مشترک انسان و حیوان، همچنان یکی از چالش‌های پایدار سلامت عمومی در مناطق اندمیک به‌شمار می‌رود. این پژوهش با تلفیق سامانه‌های اطلاعات مکانی و الگوریتم‌های یادگیری ماشین به بررسی تأثیر عوامل محیطی و مکانی بر الگوی پراکنش لیشمانیوز جلدی در استان ایلام (غرب کشور ایران) طی سال‌های ۲۰۱۴ تا ۲۰۱۹ پرداخته است. داده‌های وقوع بیماری با متغیرهای اقلیمی ترکیب شدند. به‌منظور غلبه بر محدودیت داده‌های فقط-حضور، یک چارچوب مدل‌سازی مقایسه‌ای با تولید داده‌های شبه‌عدم‌حضور و ترسیم نقشه‌های دوگانه مکانی توسعه یافت. سه الگوریتم شامل ماشین بردار پشتیبان، جنگل تصادفی و رگرسیون لجستیک پیاده‌سازی شدند. مدل جنگل تصادفی عملکرد برتری نسبت به سایر مدل‌ها نشان داد و به شاخص‌های ارزیابی شامل AUC-ROC برابر ۰/۹۹۹۵، Recall برابر ۰/۹۲، Precision برابر ۰/۸۸، F1-Score برابر ۰/۹۰ و Accuracy برابر ۰/۹۹۸۸ دست یافت. تحلیل اهمیت ویژگی‌ها، معیار بیشینه میانگین دما (TMax_M) را به‌عنوان مؤثرترین متغیر پیش‌بینی‌کننده شناسایی کرد. نقشه‌های خروجی نشان دادند که کانون‌های پرخطر عمدتاً در نواحی مرکزی و جنوب‌غربی استان متمرکز هستند. یافته‌های مکانی این پژوهش، ارتباط حیاتی بین محرک‌های خاص اقلیمی و کانون‌های بیماری را نشان می‌دهد. این مطالعه با ارائه نقشه‌های دوگانه (احتمال و ریسک)، شواهد کاربردی ارزشمندی را برای اولویت‌بندی نظارت و اقدامات پیشگیرانه در اختیار مقامات بهداشتی قرار می‌دهد.

واژگان کلیدی: سالک، یادگیری ماشین، سلامت مکانی، مدل‌سازی مکانی، چارچوب نقشه دوگانه، معیارهای اقلیمی.

* نویسنده رابط

۱- مقدمه

مطالعات جهانی نشان می‌دهند که بیماری لیشمانیوز پوستی[۱] بیش از ۸۰ کشور را درگیر کرده‌است و وقوع مکرر همه‌گیری‌ها، چالشی جدی برای سلامت عمومی محسوب می‌شود [۱]. ایران از جمله کشورهایی است که دارای شیوع بالای لیشمانیوز پوستی بوده و یکی از مناطق اندمیک اصلی این بیماری به شمار می‌رود. این بیماری توسط گونه‌های مختلف پشه خاکی در استان‌های متعدد کشور منتقل می‌شود و توزیع مکانی آن به‌شدت تحت تأثیر عوامل اقلیمی و محیطی قرار دارد. در سال‌های اخیر، سامانه‌های اطلاعات مکانی[۲] به‌عنوان ابزاری کلیدی برای مدل‌سازی پراکنش مکانی بیماری‌های منتقل‌شونده توسط ناقلین مانند CL مورد استفاده قرار گرفته‌اند؛ این سامانه‌ها امکان شناسایی مناطق پرخطر و پیش‌بینی بروز احتمالی بیماری را برای پژوهشگران فراهم می‌سازند. هم‌زمان، الگوریتم‌های یادگیری ماشین[۳] به دلیل توانایی در تحلیل مجموعه‌داده‌های پیچیده و شناسایی روابط غیرخطی و ظریف میان عوامل محیطی و شیوع بیماری، به‌طور فزاینده‌ای به رویکردهای مبتنی بر GIS افزوده شده‌اند [۲]. مدل‌های آماری سنتی یا مبتنی بر قواعد ثابت، غالباً قادر به درک کامل این روابط پیچیده نیستند؛ ازاین‌رو، نیاز به چارچوب‌های محاسباتی پیشرفته که بتوانند دقت پیش‌بینی‌های مکانی را بهبود داده و از اقدامات هدفمند بهداشت عمومی پشتیبانی کنند، بیش از پیش احساس می‌شود.

جدول ۱ مروری بر مطالعات مرتبط با پژوهش را ارائه می‌دهد. مطالعات پیشین از تحلیل و مدل‌سازی مکانی برای بررسی اثر متغیرهای محیطی بر شیوع لیشمانیوز پوستی بهره برده‌اند [۵–۳]. برخی پژوهش‌ها نیز با ترکیب تحلیل مکانی و هوش مصنوعی مکانی[۴] تلاش کرده‌اند تا ضمن مدل‌سازی و پیش‌بینی همه‌گیری بیماری، الگوهای فضایی مرتبط با انتشار آن را شناسایی کنند.

مرور جامع منابع پیشین، چندین محدودیت اساسی در مطالعات گذشته را آشکار ساخت. هرچند پژوهش‌های قبلی داده‌های اقلیمی، اطلاعات بیماران، روش‌های هوش مکانی و تحلیل‌های فضایی را مورد توجه قرار داده بودند، اما اغلب آن‌ها به مجموعه‌داده‌هایی محدود بودند که تنها شامل نقاط ابتلای مثبت بوده و مناطق عاری از بیماری را در بر نمی‌گرفتند. برای رفع این محدودیت، در پژوهش حاضر داده‌های شبه عدم حضور[۵] از طریق نمونه‌برداری تصادفی در سراسر استان و با نسبت بهینه‌ی ۳ تا ۴ برابر داده‌های حضور واقعی تولید شد [۲۱, ۲۰]. به‌کارگیری این روش باعث بهبود قابل‌توجه عملکرد مدل در شاخص‌های مختلف از جمله Recall، Precision، F1-Score، AUC-ROC و Accuracy گردید. نوآوری اصلی این پژوهش در ادغام داده‌های شبه‌عدم‌حضور برای کاهش عدم‌قطعیت در مدل‌سازی‌های مبتنی بر داده‌های حضور و ترکیب آن با راهبرد چندمدلی و تولید نقشه‌های دوگانه‌ی مکانی نهفته است تا مدیریت هدفمند بیماری CL را تسهیل کند. هدف این مطالعه شناسایی الگوهای مکانی در توزیع بیماری با تمرکز بر استان ایلام، کشور ایران است. مجموعه‌داده‌های مورد استفاده، دوره‌ی زمانی ۱۳۹۲ تا ۱۳۹۷ (معادل سال میلادی ۲۰۱۴ تا ۲۰۱۹) را پوشش می‌دهد و شامل اطلاعات بیماران و متغیرهای اقلیمی است. سه الگوریتم یادگیری ماشین شامل ماشین بردار پشتیبان[۶]، جنگل تصادفی[۷] و رگرسیون لجستیک[۸] به‌کار گرفته شدند و با استفاده از روش اعتبارسنجی متقابل پنج‌تایی[۹] کالیبره شدند. خروجی مدل‌ها شامل نقشه‌های احتمال پیش‌بینی‌شده میانگین و نقشه‌های غالب خطر برای هر الگوریتم بود. یافته‌های این پژوهش از منظر سلامت عمومی و مدیریت بیماری‌ها اهمیت فراوانی دارد؛ زیرا می‌تواند در شناسایی مناطق پرخطر، تقویت نظام پایش بیماری، بهینه‌سازی تخصیص منابع و افزایش اثربخشی اقدامات پیشگیرانه مورد استفاده قرار گیرد. در مجموع، ترکیب روش‌های پیشرفته یادگیری ماشین با تحلیل مکانی و داده‌های شبه‌عدم‌حضور در این تحقیق، چارچوبی مناسب برای مدیریت داده‌محور بیماری CL فراهم کرده است.

۱ Cutaneous leishmaniasis (CL)
۲ Geographic Information Systems (GIS)
۳ Machine Learning (ML)
۴ GeoAI
۵ Pseudo-Absence
۶ Support Vector Machine (SVM)
۷ Random Forest (RF)
۸ Logistic Regression (LR)
۹ Five-Fold Cross Validation

جدول ۱ - مروری بر ویژگی‌های اصلی مطالعات ارزیابی‌شده.

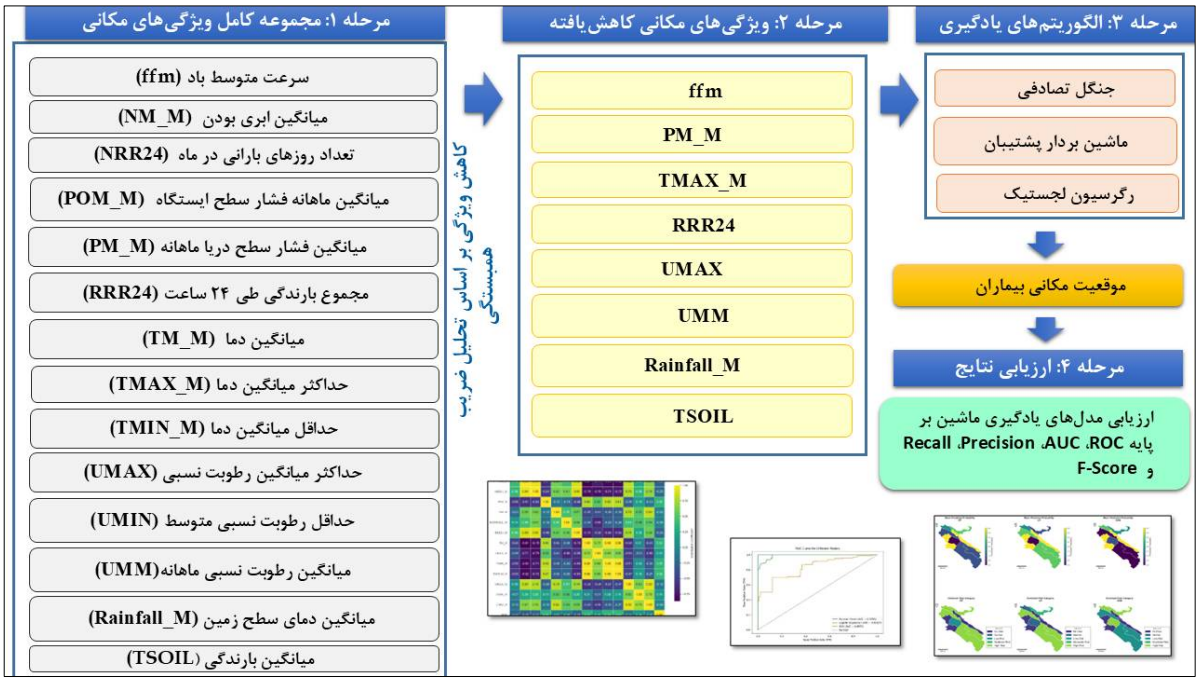
منبع	روش	منطقه	معیار
[۶]	شبکه عصبی مصنوعی ترکیبی [۱] و تقویت گرادیان شدید [۲]	پاکستان	داده‌های مبتلایان وزارت بهداشت پاکستان
[۷]	رگرسیون دوجمله‌ای منفی و TOPSIS	ایران، مشهد	عوامل اجتماعی-جمعیتی، محیطی، زمین‌شناسی
[۸]	شبکه عصبی پیشخور ساده [۳]، بازگشتی [۴]، پیشخور عمیق [۵] ، حافظه کوتاه مدت بلند مدت [۶]	برزیل	اطلاعات هواشناسی منطقه مورد مطالعه
[۹]	مدل بیزی مکانی-زمانی	برزیل	تغییرات پوشش و کاربری زمین، پوشش جنگل، کشاورزی، دامداری، استخراج گرایی، جنگل زدایی، جمعیت، آب و هوا، اجتماعی-اقتصادی، تصادفی مکانی-زمانی
[۱۰]	تحلیل مکانی-زمانی	برزیل	مبتلایان بیماری، متغیرهای محیطی/آب و هوایی شامل جنگل‌زدایی، دما، بارندگی، رطوبت.
[۱۱]	آمار توصیفی و فضایی	برزیل	توزیع فضایی و عوامل اجتماعی.
[۱۲]	دو الگوریتم یادگیری ماشین به نام‌های SVR و شبکه عصبی پرسپترون چند لایه [۷] ، روش ارزیابی شاخص RMSE [۸]	ایران، استان اصفهان	پارامترهای محیطی: سرعت باد، بارش، رطوبت، دمای هوا و ارتفاع
[۱۳]	تحلیل موقعیتی، تحلیل نقاط داغ [۹]، خودهمبستگی مکانی Moran's I، آمار فضایی Getis-Ord Gi	جزیره سریلانکا، اقیانوس هند	اطلاعات بالینی ثبت شده مبتلایان به لیشمانیوز جلدی از بیمارستان عمومی منطقه مورد مطالعه
[۱۴]	درخت تصمیم [۱۰]، SVM، رگرسیون خطی [۱۱]، فازی [۱۲]	ایران، اصفهان	عوامل محیطی، دما، رطوبت، بارندگی، ارتفاع، شیب، سرعت باد، NDVI ، تعداد روزهای آفتابی، تعداد روزهای یخبندان، فاصله تا رودخانه
[۴]	ANFIS و PCA-ANFIS	ایران، گلستان	آب و هوا، توپوگرافی، پوشش گیاهی و جمعیت انسانی
[۱۵]	آماره‌های نقطه کانونی، نقاط داغ، درون‌یابی IDW، داده‌های خوشه‌ای و پرت، تحلیل زمین‌آماری بیزی	پاکستان	اطلاعات بیمار برای مرکز بهداشت منطقه‌ای
[۱۶]	ترکیب GIS، مدل‌سازی مبتنی بر عامل [۱۳] و یک مدل اپیدمیولوژیک SEIR [14] بهبودیافته	ایران، استان گلستان	داده های مبتلایان از مرکز بهداشت استان گلستان، تقسیمات سیاسی ایران، نقشه های شبکه رودخانه ها، شاخص نرمال شده پوشش گیاهی (NDVI)، DEM، داده های آماری اطلاعات جمعیتی روستاها

۱ Hybrid Artificial Neural Network (Hybrid ANN)
۲ Extreme Gradient Boosting (XGBoost)
۳ Simple Feedforward Neural Network (SFNN)
۴ Recurrent Neural Network (RNN)
۵ Deep Feedforward Neural Network (DFNN)
۶ Long Short-Term Memory (LSTM)
۷ Multilayer Perceptron (MLP)
۸ Root Mean Square Error
۹ Hot spot
۱۰ Decision Tree
۱۱ Linear Regression
۱۲ Fuzzy Logic
۱۳ Agent-Based Modeling (ABM)
۱۴ SEIR (Susceptible, Exposed, Infectious, Recovered)

[۱۷]	تحلیل‌های مکانی-آماری و ترکیب لایه‌های اطلاعاتی با مدل بولین	اصفهان، خراسان رضوی و جنوبی، مرکزی، فارس، کرمان، قم، تهران، قزوین و سمنان	لایه‌های اطلاعات مکانی و داده‌های بیماری برای سال‌های ۱۳۸۰ تا ۱۳۸۶ از مرکز مدیریت بیماری‌های وزارت بهداشت
[۱۸]	بررسی توزیع مکانی با GIS و نرم افزار آماری SPSS	ایران، استان مازندران	داده‌های جمعیت‌شناختی و اپیدمیولوژیک جمع‌آوری‌شده از بیماران مبتلا به لیشمانیوز جلدی در معاونت بهداشتی دانشگاه علوم پزشکی مازندران
[۱۹]	خودهمبستگی مکانی(Moran's I) و روش کریجینگ	ایران، استان قم	مبتلایان به بیماری لیشمانیوز

۲-روش‌شناسی تحقیق

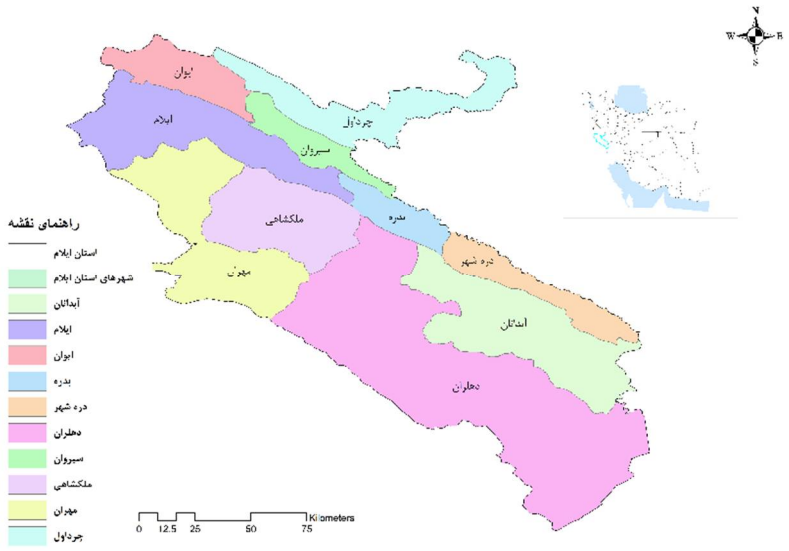
شکل ۱ نمای کلی از جریان روش اجرای پژوهش را نشان می‌دهد. در مرحله نخست، ویژگی‌های اقلیمی مرتبط با منطقه مورد مطالعه انتخاب می‌شوند. در مرحله دوم، پس از آماده‌سازی و پردازش داده‌های مورد نیاز، مؤثرترین ویژگی‌ها برای ورود به تحلیل شناسایی می‌گردند. در مرحله سوم، ویژگی‌های منتخب به سه مدل یادگیری ماشین [۲۲,۲۳] شامل RF [۲۶-۲۴]، SVM [۲۷]، LR [۲۸]، وارد می‌شوند تا خروجی‌های پیش‌بینی تولید شود. در نهایت، عملکرد هر مدل براساس نقشه‌های خروجی و شاخص‌های ارزیابی کمی مورد سنجش و مقایسه قرار می‌گیرد.



شکل ۱- نمودار روش اجرای پژوهش.

۲-۱- معرفی منطقه مورد مطالعه

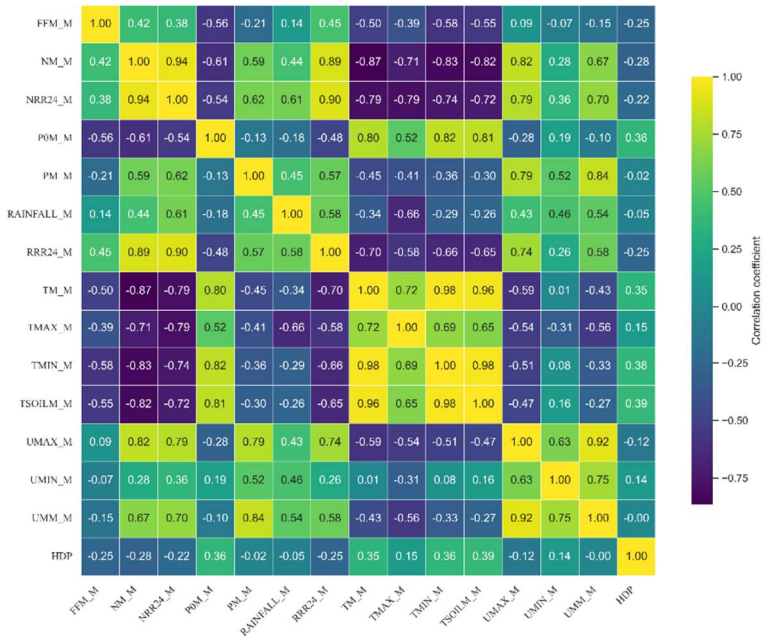
این پژوهش در استان ایلام (شکل ۲) واقع در غرب ایران انجام شده است. این استان با مساحت حدود ۲۰/۱۳۳ کیلومتر مربع، حدود ۱/۴ درصد از مساحت کل کشور را شامل می‌شود و جمعیتی در حدود ۵۵۷/۵۹۹ نفر دارد. از نظر جغرافیایی، ایلام در میان رشته‌کوه‌های زاگرس قرار گرفته و از سمت غرب با کشور عراق هم‌مرز است. با وجود ارتفاعی بیش از ۱/۲۰۰ متر از سطح دریا، اقلیم استان تحت تأثیر ناحیه‌های بیابانی غربی قرار دارد. میانگین دمای هوا در تابستان حدود ۲۹ تا ۳۰ درجه سانتی‌گراد و در زمستان حدود ۴ تا ۵ درجه سانتی‌گراد است. بخش‌های شمالی و شمال‌شرقی استان عمدتاً کوهستانی‌اند، در حالی‌که مناطق جنوبی و غربی بیشتر از دشت‌ها و نواحی پست تشکیل شده‌اند.



شکل ۲ - موقعیت منطقه مطالعه و توزیع مکانی و تراکم مبتلایان CL در استان ایلام، سال‌های ۲۰۱۴ تا ۲۰۱۹.

۲-۲- داده‌ها

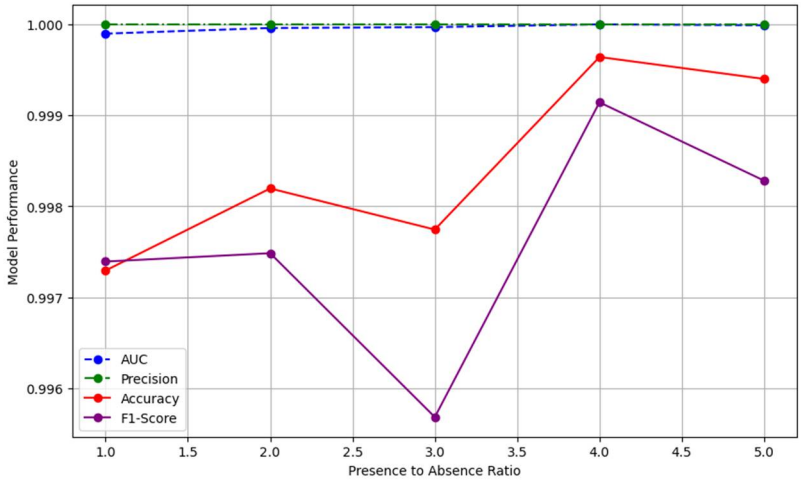
در این پژوهش از دو نوع مجموعه‌داده بیماری و داده‌های اقلیمی [۱۰] استفاده شد. داده‌های بیماری که شامل اطلاعات مربوط به توزیع و تراکم موارد ابتلا بودند، از مرکز بهداشت استان ایلام دریافت شدند. داده‌های اقلیمی نیز از اداره کل هواشناسی استان ایلام گردآوری گردیدند. پیش از انجام مدل‌سازی، به‌منظور کاهش هم‌خطی چندگانه، متغیرهایی که دارای ضریب همبستگی بالا بودند (r>۰/۸)، حذف شدند [۲۹,۳۰]. شکل ۳ ماتریس همبستگی بین متغیرهای مستقل و وابسته را نشان می‌دهد که در آن، ضرایب همبستگی برای هر جفت متغیر محاسبه شده‌اند. ویژگی‌هایی که با رنگ زرد مشخص شده‌اند، بیانگر ارتباط قوی‌تر با متغیر هدف هستند. پس از مرحله‌ی پیش‌پردازش داده‌ها، شامل نرمال‌سازی و مقیاس‌بندی، مقادیر بهینه فراپارامترهای کلیدی برای مدل‌های RF، SVM و LR با استفاده از روش GridSearch تعیین گردید. سپس مدل‌ها آموزش داده شدند و عملکرد آن‌ها با استفاده از شاخص‌های ارزیابی مختلف مورد بررسی قرار گرفت تا پیش‌بینی‌هایی دقیق و قابل اطمینان حاصل شود.



شکل ۳ - نقشه حرارتی همبستگی متغیرهای اقلیمی و محیطی کاربردی در پژوهش.

۲-۳- انتخاب داده‌های پس‌زمینه

در این مطالعه، از داده‌های حضور و شبه عدم حضور استفاده شد. داده‌های حضور شامل اطلاعاتی در مورد وقوع بیماری و شرایط اقلیمی در استان ایلام بود، در حالی که داده‌های شبه عدم حضور تولید شدند و به عنوان داده‌های پس‌زمینه نیز شناخته می‌شوند، که از طریق نمونه‌گیری تصادفی در منطقه مورد مطالعه به دست آمدند. این داده‌ها به کاهش عدم قطعیت و بهبود دقت پیش‌بینی‌ها در مورد شیوع بیماری کمک می‌کنند. داده‌های شبه عدم حضور به طور خاص در این تحقیق به دلیل توزیع غیر یکنواخت بیماری در سراسر منطقه مورد مطالعه و الگوهای مکانی نامشخص انتقال CL در مناطق مختلف استان ایلام مورد استفاده قرار گرفتند [۳۱,۳۲]. شکل ۴ عملکرد بهبود یافته مدل RF را تحت نسبت بهینه داده‌های حضور به داده‌های شبه عدم حضور نشان می‌دهد (شکل ۵، پیوست). مجموعه داده‌ها برای اعتبارسنجی متقابل به پنج بخش تقسیم شد که چهار بخش برای آموزش و یک بخش برای آزمایش در هر تکرار در نظر گرفته شد. در این مطالعه، نسبت‌های مختلف داده‌های حضور به نمونه‌های داده‌های شبه عدم‌حضور (۱:۱، ۱:۲، ۱:۳، ۱:۴ و ۱:۵) برای تعیین تعادل بهینه بین این دو ارزیابی شدند. مدل RF با استفاده از معیارهای Recall، Precision، F1-Score و Accuracy ارزیابی شد. نتایج نشان داد که وقتی نسبت حضور به نمونه‌های داده‌های شبه عدم‌حضور بین ۱:۳ و ۱:۴ باشد (یعنی تعداد نمونه‌های ۳ تا ۴ برابر بیشتر از تعداد نمونه‌های حضور باشد)، مدل RF در هر چهار معیار (Recall، Precision، F1-Score و Accuracy) عملکرد بهتری داشته است (جدول ۲).



شکل ۴- بهبود عملکرد مدل RF در نسبت‌های مختلف PA

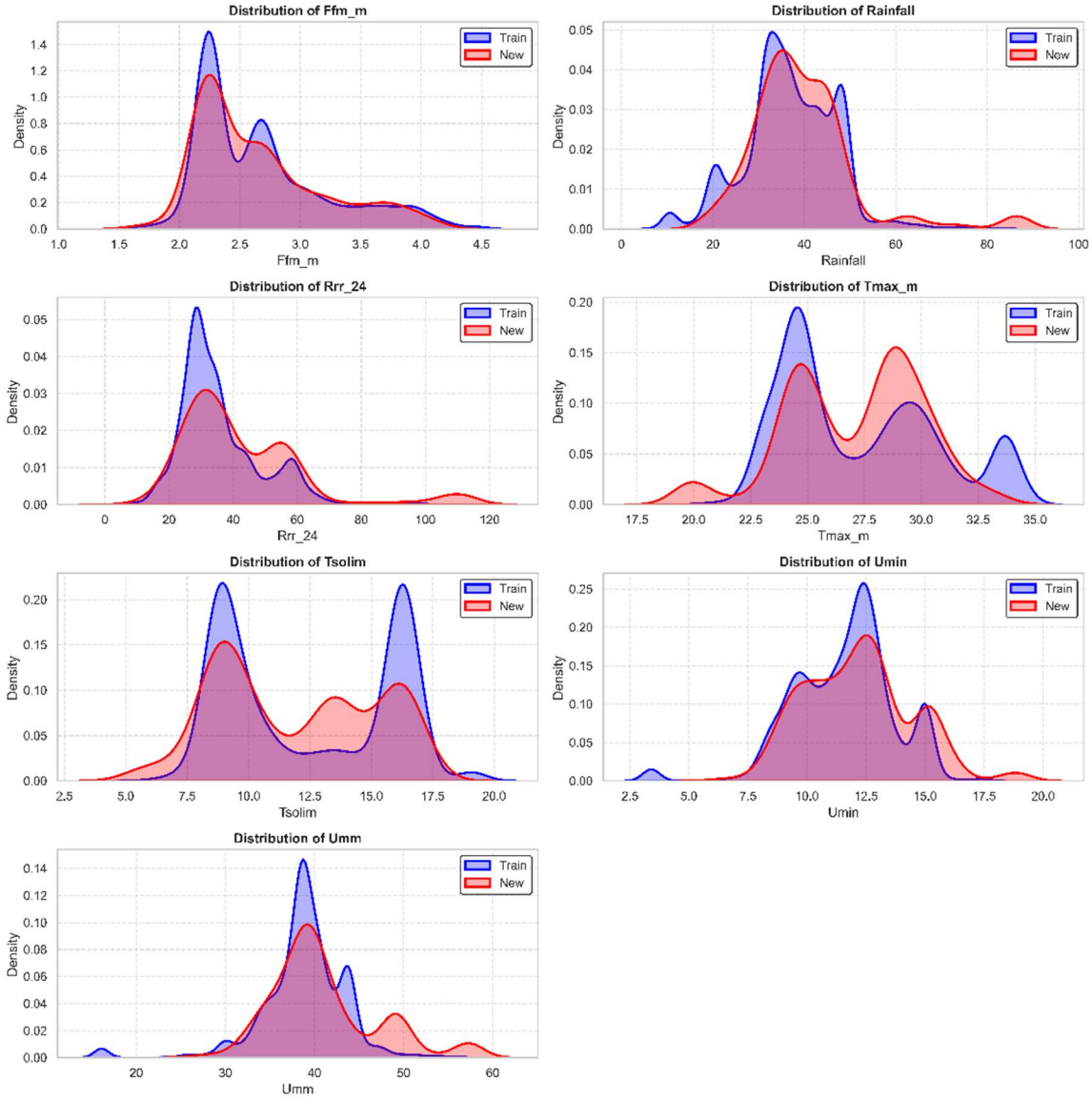
جدول ۲- نتایج ارزیابی الگوریتم RF.

AUC	F1-Score	Recall	Precision	Accuracy	Ratio
۰/۹۹۹۸۹۶	۰/۹۹۷۳۹۴	۰/۹۹۴۸۰۱	۱/۰	۰/۹۹۷۲۹۵	۰
۰/۹۹۹۹۵۸	۰/۹۹۷۴۸۵	۰/۹۹۴۹۸۳	۱/۰	۰/۹۹۸۱۹۶	۱
۰/۹۹۹۹۶۷	۰/۹۹۵۶۸۶	۰/۹۹۱۴۰۹	۱/۰	۰/۹۹۷۷۴۵	۲
۰/۹۹۹۹۹۸	۰/۹۹۹۱۴۲	۰/۹۹۸۲۸۵	۱/۰	۰/۹۹۹۶۳۹	۳
۰/۹۹۹۹۸۸	۰/۹۹۸۲۸۲	۰/۹۹۶۵۶۹	۱/۰	۰/۹۹۹۳۹۸	۴

۲-۴- نمودارهای تخمین چگالی هسته

هدف از نمودارهای KDE مقایسه توزیع ویژگی‌ها بین مجموعه داده‌های آموزشی و آزمایشی است که از سازگاری برای پیش‌بینی‌های دقیق نقشه‌برداری خطر بیماری اطمینان حاصل می‌کند. توزیع‌های چگالی برای مجموعه داده‌های آموزشی (آبی) و مجموعه داده‌های آزمایشی (قرمز) برای تسهیل این مقایسه روی هم قرار گرفته‌اند. در

این مرحله، از بین ۱۴ ویژگی (شکل ۶ پیوست)، مواردی که همبستگی بالای ۰/۸ داشتند از فرآیند مدل‌سازی حذف شدند. ویژگی‌های باقی‌مانده مورد استفاده برای مدل‌سازی شامل ffm ، TMax_m، RRR24، Umm، Umin، Rainfall و TSoil بودند (شکل ۷).



شکل ۷- نمودارهای KDE که توزیع ویژگی‌های انتخاب‌شده بین مجموعه داده‌های آموزشی و آزمایشی را نشان می‌دهند.

۳- یافته‌ها

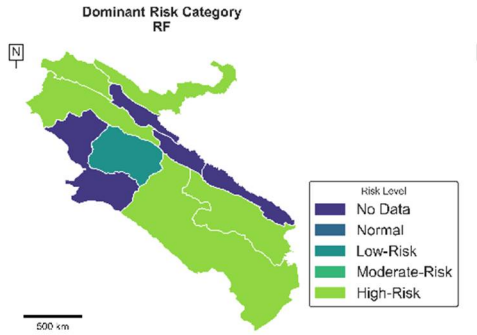
در این مطالعه به منظور پیش‌بینی خطر بیماری لیشمانیوز جلدی، دو نوع نقشه برای هر یک از سه مدل‌سازی تولید شد: نقشه‌های میانگین احتمال پیش‌بینی‌شده و نقشه‌های دسته ریسک غالب.

۳-۱- نقشه‌های دسته‌بندی ریسک غالب

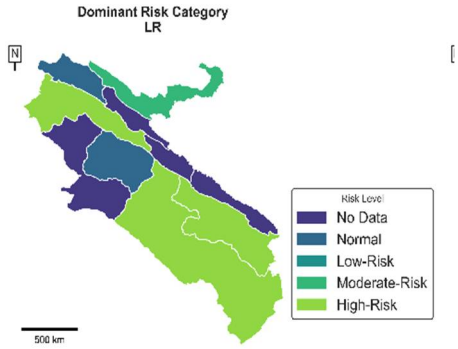
نقشه‌های دسته‌بندی ریسک غالب، شکل ۸ (الف- ب – ج) استان را به چهار سطح ریسک متمایز، همانطور که در راهنمای نقشه تعریف شده است، طبقه‌بندی می‌کنند: عادی (آبی تیره)، کم‌خطر (سبز فیروزه‌ای/سبز آبی)، پرخطر (سبز) و پرخطر (سبز روشن/زرد-سبز). این سه مدل در طبقه‌بندی نهایی ریسک خود اختلاف نظر قابل توجهی نشان می‌دهند. مدل RF، شکل ۸ (الف)، بسیار حساس است و اکثریت قریب به اتفاق استان را به عنوان "پرخطر" (سبز روشن) طبقه‌بندی می‌کند، و مناطق مرکزی و شمالی کوچک‌تر را به عنوان "متوسط-خطر" یا "کم-خطر" طبقه‌بندی می‌کند. مدل LR، شکل ۸ (ب) طبقه‌بندی متنوع‌تری ارائه می‌دهد و مناطق بزرگ "پرخطر" و "متوسط-خطر" و همچنین مناطق

"عادی" قابل توجهی (آبی تیره) را در مرکز شناسایی می‌کند. مدل SVM، شکل ۸ (ج)، الگوی دیگری را ارائه می‌دهد که در آن "ریسک متوسط" (سبز) گسترده‌ترین دسته است و در میان آن مناطق "ریسک پایین" (سبز فیروزه‌ای) و "ریسک بالا" (سبز روشن) قرار دارند.

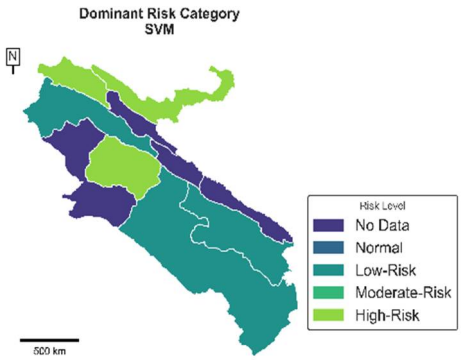
۳-۲- نقشه‌های احتمال پیش‌بینی‌شده میانگین



(الف)



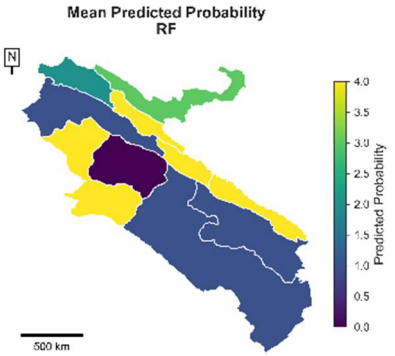
(ب)



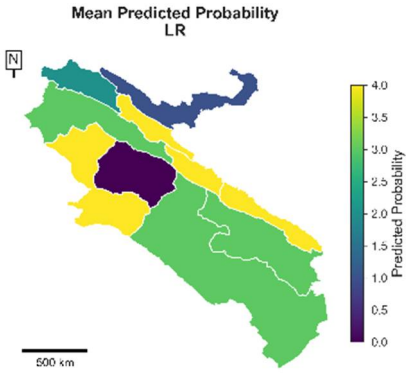
(ج)

شکل ۸- خروجی اجرای الگوریتم: نقشه‌های دسته‌بندی ریسک غالب

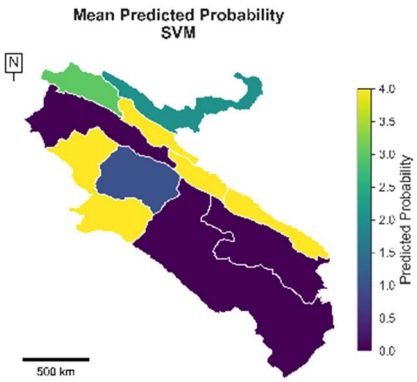
نقشه‌های احتمال پیش‌بینی‌شده میانگین، شکل ۹ (الف-ب – ج) احتمال مکانی وقوع CL را با استفاده از یک رمپ رنگی پیوسته، از احتمال کم (آبی تیره) تا احتمال زیاد (زرد روشن) نشان می‌دهند. مدل RF ، شکل ۹ (الف)، بالاترین احتمالات را پیش‌بینی می‌کند و مناطق وسیعی را در مناطق مرکزی و جنوب غربی به عنوان مناطق پرخطر (زرد و بنفش) شناسایی می‌کند. در مقابل، مدل LR شکل۹(ب)، محافظه‌کارترین مدل است و اکثریت قریب به اتفاق استان را به عنوان مناطق با احتمال کم (آبی تیره) با تنها خوشه‌های کوچک با احتمال متوسط طبقه‌بندی می‌کند. مدل SVM شکل ۹ (ج)، توزیع متوسطی را ارائه می‌دهد و مناطق با احتمال بالای قابل توجهی را در غرب مرکزی شناسایی می‌کند، اما گستردگی آن کمتر از مدل RF است.



(الف)



(ب)

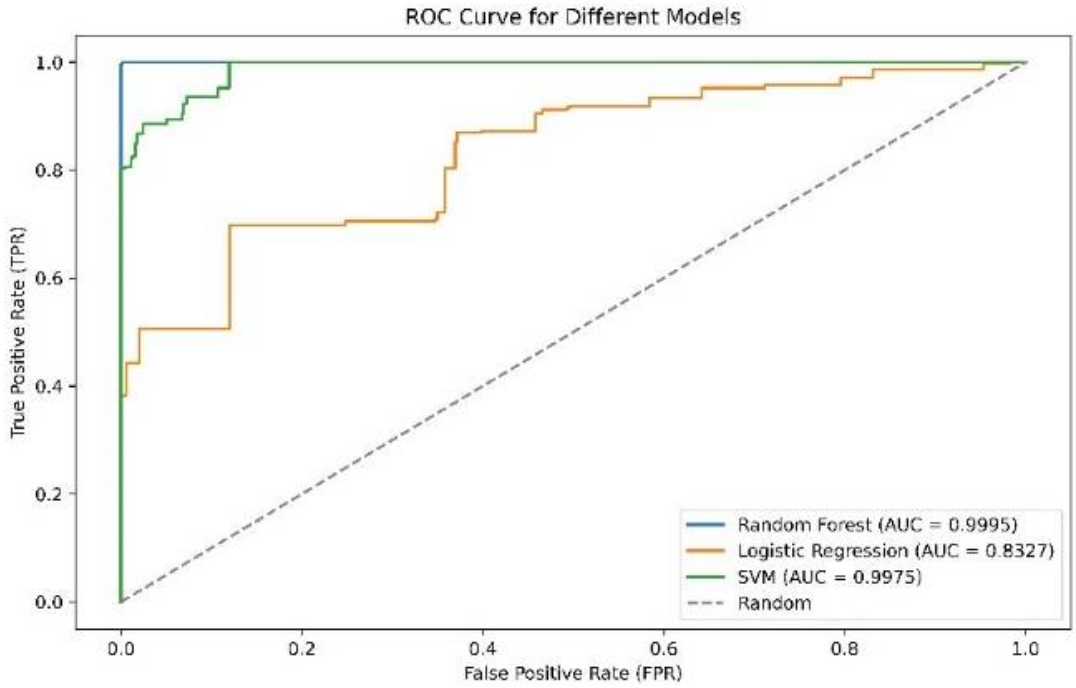


(ج)

شکل ۹ - نقشه‌های احتمال پیش‌بینی‌شده میانگین

۳-۳- مقایسه الگوریتم‌ها

هنگام ارزیابی مدل‌ها، استفاده از معیارهای چندگانه برای ارزیابی جامع عملکرد آنها بسیار مهم است. در این مطالعه، الگوریتم‌های SVM، LR و RF با استفاده از یک مجموعه داده اقلیمی توسعه داده شدند و عملکرد پیش‌بینی‌کننده آنها برای شناسایی مناطق مستعد بیماری مقایسه شد. شکل ۱۰ منحنی‌های ROC مدل‌های توسعه‌یافته را براساس مجموعه داده تجربی نشان می‌دهد. طبق این تحلیل، مدل RF بهترین عملکرد را با مقدار AUC نزدیک به مقدار ۱ نشان داد که نشان‌دهنده جدایی تقریباً کامل بین کلاس‌ها است. SVM همچنین عملکرد قوی‌ای را نشان می‌دهد، در حالی‌که LR کمترین اثربخشی را در این مطالعه داشت، و نشان می‌دهد ممکن است نیاز به بهینه‌سازی پارامتر یا اتخاذ رویکردهای مدل‌سازی پیچیده‌تر داشته باشد.



شکل ۱۰- منحنی‌های ROC الگوریتم‌های مختلف یادگیری ماشین، شامل RF، SVM و LR.

شکل ۱۱ نمایش نتایج مقایسه تحلیل سه الگوریتم پیاده‌سازی‌شده را براساس معیارهای accuracy، recall، F1 Score- و AUC-ROC ارائه می‌دهد. همانطور که نشان داده شده است، الگوریتم RF در مقایسه با سایر روش‌ها، عملکرد برتر را در تمام معیارهای ارزیابی نشان می‌دهد.

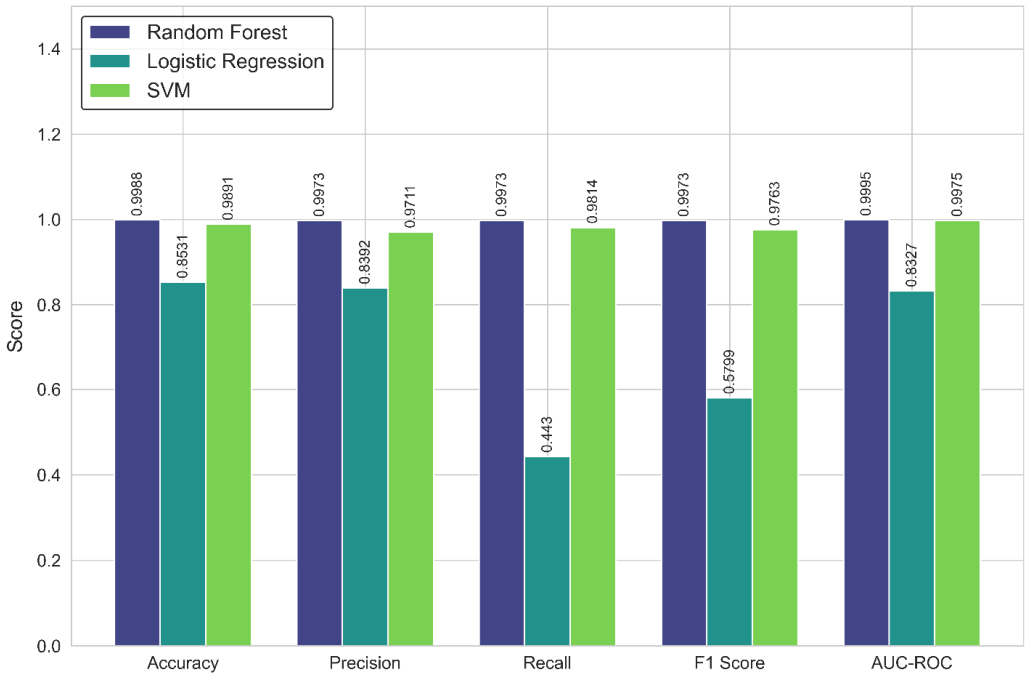
Accuracy: مدل RF با ۰/۹۹۸۸ بالاترین دقت را دارد که نشان‌دهنده عملکرد کلی در سطح بالا است. در مقابل، LR دقت بسیار پایین‌تری معادل ۸۵/۳۱٪ را نشان داد که ممکن است برای مجموعه داده‌های نامتوازن مناسب نباشد.

Precision: دقت بالای RF برابر ۹۹٪/۷۳ توانایی بالای آن را در پیش‌بینی صحیح موارد مثبت واقعی نشان می‌دهد. LR کمترین دقت را با ۸۳/۹۲٪ نشان داد.

Recall: الگوریتم RF همچنین با ۹۹/۷۳٪ به بالاترین میزان دست یافت که نشان می‌دهد اکثر موارد مثبت (مناطق پرخطر) را با موفقیت شناسایی کرده است. LR در این معیار عملکرد ضعیفی داشت و مقدار آن ۰/۴۴۳٪ است و بسیاری از موارد مثبت را از دست داد.

F1-Score: که ترکیبی از Recall و Precision است، برای RF (۹۹/۷۳٪) بالاترین بود. LR کمترین F1-Score را با ۵۷/۹۹ ٪ دارد که عمدتاً به دلیل Recall بسیار پایین آن بود.

AUC-ROC: مدل RF به AUC-ROC برابر با ۹۹۹۵ /۰ دست یافت که از سایر مدل‌ها عملکرد بهتری داشت. در مقایسه، LR دارای AUC-ROC برابر با ۰/۸۳۲۷ بود که نشان‌دهنده تفکیک طبقاتی نسبتاً ضعیف است.



شکل ۱۱- مقایسه مدل‌های پیاده‌سازی شده بر اساس معیارهای ارزیابی.

۴- بحث

این پژوهش با هدف ایجاد نقشه خطر بیماری CL در استان ایلام و ارزیابی کارایی الگوریتم‌های یادگیری ماشین در تلفیق با متغیرهای اقلیمی انجام شد. یکی از چالش‌های بنیادین در مدل‌سازی اپیدمیولوژیک، ماهیت صرفاً حضور داده‌های بیماری است که منجر به عدم قطعیت ذاتی در تشخیص مناطق امن واقعی می‌شود؛ چرا که فقدان رکورد بیماری در یک منطقه لزوماً به معنای عدم وجود خطر نیست، بلکه ممکن است ناشی از نقص در گزارش‌دهی باشد. در این مطالعه، برای غلبه بر این محدودیت، از رویکرد تولید داده‌های شبه‌عدم‌حضور استفاده شد تا مدل‌ها بتوانند تمایز دقیق‌تری میان شرایط محیطی کانون‌های بیماری و پس‌زمینه عمومی استان قائل شوند.

۴-۱- تحلیل عوامل تأثیرگذار اقلیمی

بر اساس نمودار اهمیت ویژگی‌ها در مدل جنگل تصادفی (شکل ۱۲)، متغیر بیشینه میانگین دما (Tmax_m) با اختلاف قابل‌توجهی به‌عنوان مؤثرترین عامل تعیین‌کننده خطر بیماری شناسایی شد. پس از آن، متغیرهای میانگین دمای سطح خاک (Tsolim) و حداقل رطوبت نسبی (Umin) در رتبه‌های بعدی قرار گرفتند. این یافته‌ها از منظر اکولوژیکی بسیار حائز اهمیت است؛ زیرا چرخه زندگی پشه خاکی (ناقل بیماری) وابستگی شدیدی به آستانه‌های حرارتی دارد. دمای هوا و دمای خاک مستقیماً بر نرخ رشد حشرات ناقل تأثیر می‌گذارند. همچنین، نقش پررنگ متغیر حداقل رطوبت، نشان می‌دهد که خشکی هوا یا رطوبت نسبی پایین می‌تواند به‌عنوان یک عامل محدودکننده یا تسهیل‌کننده در بقای ناقل عمل کند. در مقابل، متغیرهایی مانند میانگین رطوبت ماهانه (Umm) کمترین تأثیر را در مدل‌سازی داشتند که نشان می‌دهد مقادیر حدی (حداقل/حداکثر) نقش تعیین‌کننده‌تری نسبت به میانگین‌ها در اکولوژی این بیماری ایفا می‌کنند.

۴-۲-عملکرد الگوریتم‌ها و تفسیر نقشه‌های دوگانه

در این پژوهش، سه الگوریتم RF، SVM و LR به کار گرفته شدند که نتایج ارزیابی (شکل ۸ و ۹) برتری مدل RF را در تمامی شاخص‌ها دقت، حساسیت و AUC نشان داد. تولید نقشه‌های دوگانه (احتمال و ریسک) دیدگاه مکملی را برای مدیریت بیماری فراهم کرد:

مدل RF: این مدل با حساسیت بالا، مناطق وسیعی در مرکز و جنوب‌غربی استان را به‌عنوان نواحی پرخطر (با رنگ سبز روشن در نقشه ریسک) شناسایی کرد. انطباق این نواحی با کانون‌های دمایی استان (تأثیر Tmax_m) نشان‌دهنده دقت

مدل در کشف روابط غیرخطی است. از آنجایی که نتایج به‌دست‌آمده از اجرای الگوریتم جنگل تصادفی، بهترین عملکرد را طبق معیارهای ارزیابی الگوریتم نشان می‌دهد، شکل ۱۲ معیارهای پیش‌بینی مؤثر به‌دست‌آمده از اجرای الگوریتم RF را نشان می‌دهد.

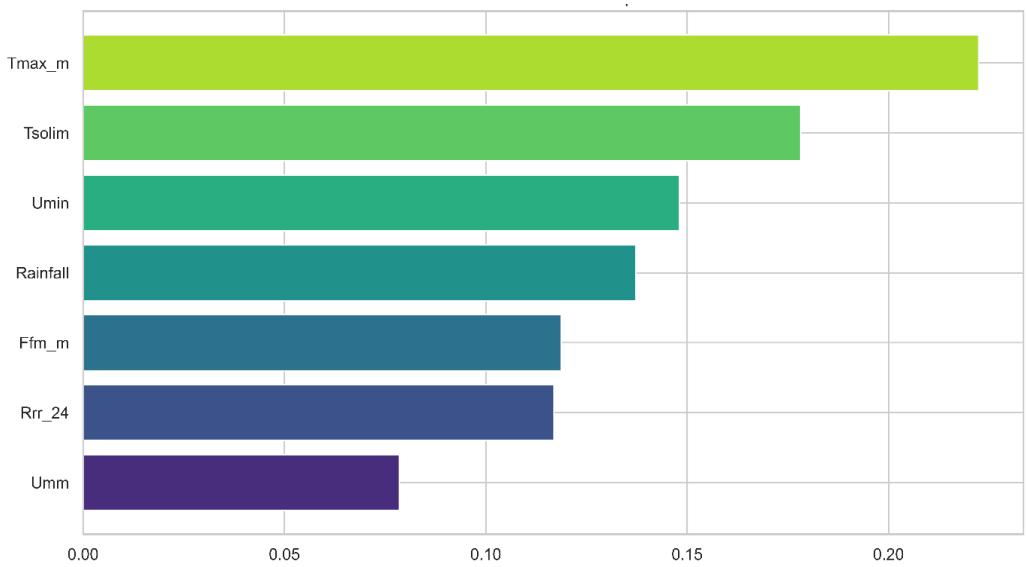
مدل LR: برخلافRF، مدل رگرسیون لجستیک رویکردی بسیار محافظه‌کارانه داشت و تنها لکه‌های کوچکی را پرخطر دانست. این ضعف عملکرد (Recall پایین) تأیید می‌کند که مدل‌های خطی توانایی کافی برای درک پیچیدگی‌های اکولوژیک لیشمانیوز را ندارند.

مدل SVM: این مدل الگوی میانه‌ای را ارائه داد و اگرچه نواحی خطر را در غرب استان شناسایی کرد، اما در تفکیک دقیق مرزهای خطر به اندازه RF موفق نبود.

۴-۳- کاربرد عملی و محدودیت‌ها

چارچوب پیشنهادی این پژوهش، ابزاری دوگانه در اختیار تصمیم‌گیران قرار می‌دهد: نقشه‌های احتمال (با طیف رنگی پیوسته) برای درک پیوستگی فضایی خطر و اولویت‌بندی پژوهشی مناسب هستند، در حالی که نقشه‌های دسته‌بندی ریسک (با طبقات گسسته) برای اقدامات فوری مدیریتی و تخصیص منابع کاربرد دارند .با این حال، این مطالعه با محدودیت‌هایی نیز روبرو بود. نخست، عدم تعادل ذاتی داده‌های بیماری ممکن است بر عملکرد مدل‌های خطی پایه مانند LR تأثیر منفی گذاشته باشد. دوم، اگرچه مدل‌های پیچیده‌ای مانند RF دقت بالایی دارند، اما تفسیرپذیری آن‌ها نسبت به مدل‌های ساده‌تر دشوارتر است؛ هرچند تحلیل اهمیت ویژگی‌ها (شکل ۱۲) تا حد زیادی به رفع این چالش کمک کرد.

در مجموع، نتایج نشان داد که به‌کارگیری یادگیری ماشین در کنار تحلیل‌های مکانی، می‌تواند الگوهای پنهان انتشار بیماری را که با روش‌های سنتی قابل‌شناسایی نیستند، آشکار سازد و مبنایی علمی برای سیستم‌های هشدار زودهنگام فراهم کند.



شکل ۱۲- نمایش میزان تاثیر معیارها را در خروجی الگوریتم جنگل تصادفی

۵- نتیجه‌گیری

این پژوهش با هدف توسعه یک چارچوب پیش‌بینی مکانی دقیق برای بیماری لیشمانیوز جلدی در کانون اندمیک استان ایلام انجام شد. با تلفیق داده‌های بالینی دوره‌ی زمانی ۱۳۹۲ تا ۱۳۹۷ (معادل سال میلادی ۲۰۱۴ تا ۲۰۱۹) و متغیرهای کلیدی اقلیمی بر پایه GIS کارایی سه الگوریتم یادگیری ماشین RF، SVM و LR مورد ارزیابی قرار گرفت.

نتایج نشان داد که الگوریتم جنگل تصادفی با غلبه بر پیچیدگی‌های ذاتی و غیرخطی داده‌های اکولوژیک، بالاترین دقت و حساسیت را در شناسایی مناطق مستعد بیماری ارائه می‌دهد. تحلیل اهمیت ویژگی‌ها، نقش حیاتی متغیر بیشینه میانگین دما (TMax_m) را به‌عنوان اصلی‌ترین محرک محیطی در چرخه انتقال بیماری برجسته کرد؛ یافته‌ای که همسو با وابستگی شدید فیزیولوژی پشه‌های خاکی به آستانه‌های حرارتی است.

نوآوری اصلی این مطالعه، ارائه یک چارچوب نقشه‌دوگانه شامل نقشه‌های پیوسته (احتمال وقوع) و گسسته (دسته‌بندی ریسک) است. این خروجی‌ها ابزاری قدرتمند برای گذار از "پاسخ‌های واکنشی" به "مدیریت پیشگیرانه" فراهم می‌کنند و به مقامات بهداشتی اجازه می‌دهند تا منابع محدود خود را دقیقاً در کانون‌های پرخطر[1] متمرکز کنند.

۶-پیشنهادات

اگرچه این مطالعه بر مدل‌سازی اکولوژیک و اقلیمی ناقل بیماری متمرکز بود، اما باید توجه داشت که شیوع لیشمانیوز پدیده‌ای چندوجهی است. بنابراین، پیشنهاد می‌شود در تحقیقات آتی، این مدل‌های اقلیمی با متغیرهای اجتماعی-اقتصادی و رفتاری (مانند تراکم جمعیت، وضعیت مسکن و مهاجرت) تلفیق شوند تا تصویر جامع‌تری از "آسیب‌پذیری انسانی" در کنار "خطر محیطی" حاصل گردد.

همچنین، بهره‌گیری از معماری‌های یادگیری عمیق می‌تواند به کشف الگوهای پیچیده‌تر زمانی-مکانی کمک نماید. در مجموع، این پژوهش گامی مؤثر در جهت استقرار سیستم‌های هشدار زودهنگام داده‌محور برای کنترل بیماری‌های گرمسیری نادیده‌گرفته‌شده در ایران محسوب می‌شود.

مراجع

[۱] WHO. Leishmaniasis. Available from: http:// wwww.hoint/ leishmaniasis/en/. [cited May 07,2025].

[۲] Neisani Samani, Z., Alesheikh, A. A., & Kalantari, M. (2025). A Geo-AI-Based Approach for Spatiotemporal Prediction of Infectious Diseases With Cluster Distribution Pattern in Urban Environments. *Transactions in GIS*, *29*(7), e70132.

[۳] Tabasi, M., Alesheikh, A. A., Sofizadeh, A., Saeidian, B., Pradhan, B., & AlAmri, A. (2020). A spatio-temporal agent-based approach for modeling the spread of zoonotic cutaneous leishmaniasis in northeast Iran. *Parasites & Vectors*, *13*(1), 572.

[۴] Babaie, E., Alesheikh, A. A., & Tabasi, M. (2022). Spatial modeling of zoonotic cutaneous leishmaniasis with regard to potential environmental factors using ANFIS and PCA-ANFIS methods. *Acta Tropica*, *228*, 106296.

[۵] Wolfe, C. M., Barry, A., Campos, A., Farham, B., Achu, D., Juma, E., ... & Impouma, B. (2024). Control, elimination, and eradication efforts for neglected tropical diseases in the World Health Organization African region over the last 30 years: A scoping review. *International Journal of Infectious Diseases*, *141*, 106943.

[۶] Niu, B., Qureshi, H., Khan, M. I., & Shah, A. (2025). Integrating AI for infectious disease prediction: A hybrid ANN-XGBoost model for leishmaniasis in Pakistan. *Acta Tropica*, 107628.

[۷] Mohammadi, A., Hamer, D. H., Pishagar, E., & Bergquist, R. (2025). Spatial modelling to identify high-risk zones for the transmission of cutaneous leishmaniasis in hyperendemic urban environments: A case study of Mashhad, Iran. *Health & Place*, *91*, 103394.

[۸] Donizette, A. C., Rocco, C. D., & de Queiroz, T. A. (2025). Predicting leishmaniasis outbreaks in Brazil using machine learning models based on disease surveillance and meteorological data. *Operations Research for Health Care*, *44*, 100453.

[۹] Portella, T. P., Sudbrack, V., Coutinho, R. M., Prado, P. I., & Kraenkel, R. A. (2024). Bayesian spatio-temporal modeling to assess the effect of land-use changes on the incidence of Cutaneous Leishmaniasis in the Brazilian Amazon. *Science of The Total Environment*, *953*, 176064.

[۱۰] da Silva Chagas, É. C., da Silva Ferreira, F. A., Mwangi, V. I., Terrazas, W. C. M., Becker, J. N., de Castro Simões, R., ... & de Oliveira, J. H. (2024). Spatio-temporal analysis of American Tegumentary Leishmaniasis incidences in the Brazilian state of Amazonas: 2011 to 2022. *Acta Tropica*, *256*, 107266.

[۱۱] Neto, A. L. S., de Oliveira Silva, L. E., Júnior, A. F. B. P., & Rocha, T. J. M. (2023). Geospatial analysis of american tegumentary leishmaniasis in Alagoas, 2007-2021. *Revista de Patologia Tropical/Journal of Tropical Pathology*, *52*(2), 107-116.

۱ Hotspots

[۱۲] shabanpour N, Kaffash Charandabi N, Shirzadi M R. Modeling and analysis of leishmaniasis distribution process using multilayer perceptron neural network and support vector regression (Case study: villages of Isfahan province). JGST 2023; 12 (2) : 1

[۱۳] Arunashantha, S., Jayarathne, M., Wijesekera, S., Sakalasooriya, N., & Kottage, C. (2023). GIS-Based Situational Analysis of Cutaneous Leishmaniasis Disease (CLD) in Sri Lanka. *Journal of Geoscience and Environment Protection*, *11*(3), 70-86.

[۱۴] Shabanpour, N., Razavi-Termeh, S. V., Sadeghi-Niaraki, A., Choi, S. M., & Abuhmed, T. (2022). Integration of machine learning algorithms and GIS-based approaches to cutaneous leishmaniasis prevalence risk mapping. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102854.

[۱۵] Zeb, I., Qureshi, N. A., Shaheen, N., Zafar, M. I., Ali, A., Hamid, A., ... & Ashraf, A. (2021). Spatiotemporal patterns of cutaneous leishmaniasis in the district upper and lower Dir, Khyber Pakhtunkhwa, Pakistan: A GIS-based spatial approaches. *Acta Tropica*, *217*, 105861.

[۱۶] Tabasi M, Alesheikh A A. Development of an Agent-Based Model for Simulation of the Spatiotemporal Spread of Leishmaniasis in GIS (Case Study: Maraveh Tappeh). JGST 2019; 8 (3) :113-131

[۱۷] AhangarCani, M., & Farnaghi, M. (2019). Providing a model for Cutaneous Leishmaniasis risk mapping using GIS and neural network algorithm. *Scientific-Research Quarterly of Geographical Data (SEPEHR)*, *28*(109), 7-24.

[۱۸] Ghavibazou, L., Hosseini-Vasoukolaei, N., Akhavan, A. A., Jahanifard, E., Yazdani-Charati, J., & Fazeli-Dinan, M. (2018). Dispersal status of cutaneous leishmaniasis in Mazandaran province, 2009-2017. *Journal of Mazandaran University of Medical Sciences*, *28*(167), 58-70.

[۱۹] Salimi, M., Jesri, N., Javanbakht, M., Farahani, L. Z., Shirzadi, M. R., & Saghafipour, A. (2018). Spatio-temporal distribution analysis of zoonotic cutaneous leishmaniasis in Qom Province, Iran. *Journal of parasitic diseases*, *42*(4), 570-576.

[۲۰] Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731-1742.

[۲۱] Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological monographs*, *92*(1), e01486.

[۲۲] Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., ... & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*, *152*, 267-275.

[۲۳] Nele, L., Mattera, G., Yap, E. W., Vozza, M., & Vespoli, S. (2024). Towards the application of machine learning in digital twin technology: a multi-scale review. *Discover Applied Sciences*, *6*(10), 502.

[۲۴] Yeşilkanat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons & Fractals*, *140*, 110210.

[۲۵] Dutta, P., Paul, S., & Kumar, A. (2021). Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19. In *Electronic devices, circuits, and systems for biomedical applications* (pp. 521-540). Academic Press.

[۲۶] Virtriana, R., Ihsan, K. T. N., Anggraini, T. S., Deliar, A., Harto, A. B., Riqqi, A., & Sakti, A. D. (2025). Development of location suitability prediction for health facilities using random forest machine learning in 2030 integrating remote sensing and GIS in West Java, Indonesia. *Environmental Advances*, *19*, 100604.

[۲۷] Muna, U. M., Hafiz, F., Biswas, S., & Azim, R. (2025). GBDTSVM: Combined Support Vector Machine and Gradient Boosting Decision Tree Framework for efficient snoRNA-disease association prediction. *Computers in Biology and Medicine*, *192*, 110219.

[۲۸] Cherifi, M., El Korso, M. N., Fortunati, S., Mesloub, A., & Ferro-Famil, L. (2025). Robust inference with incompleteness for logistic regression model. *Signal Processing*, 110027.

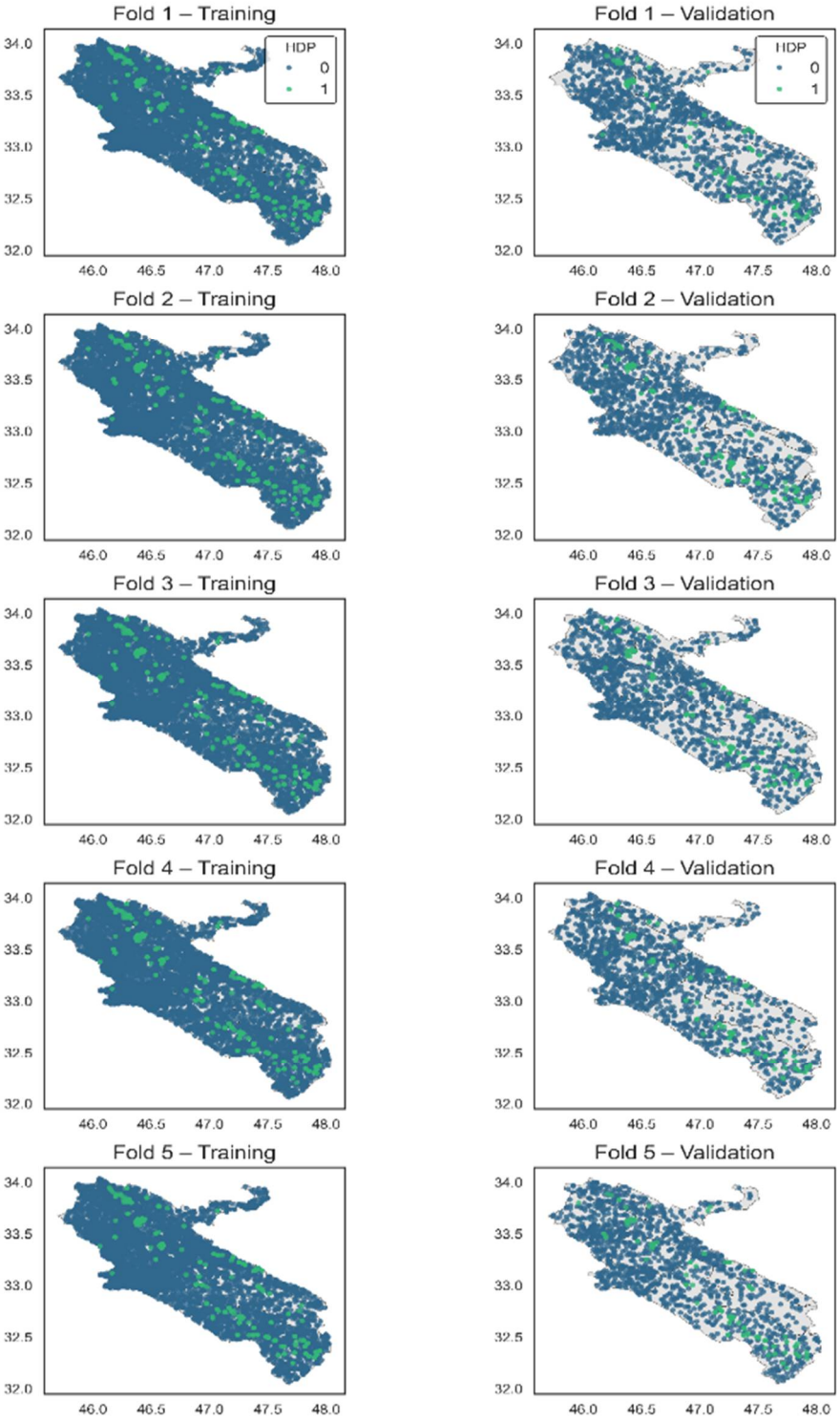
[۲۹] AbdiLynne, H., & Williams, J. (2010). Principal component analysis wiley interdisciplinary reviews. *Computational Statistics*, *2*(4), 433-459.

[۳۰] Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

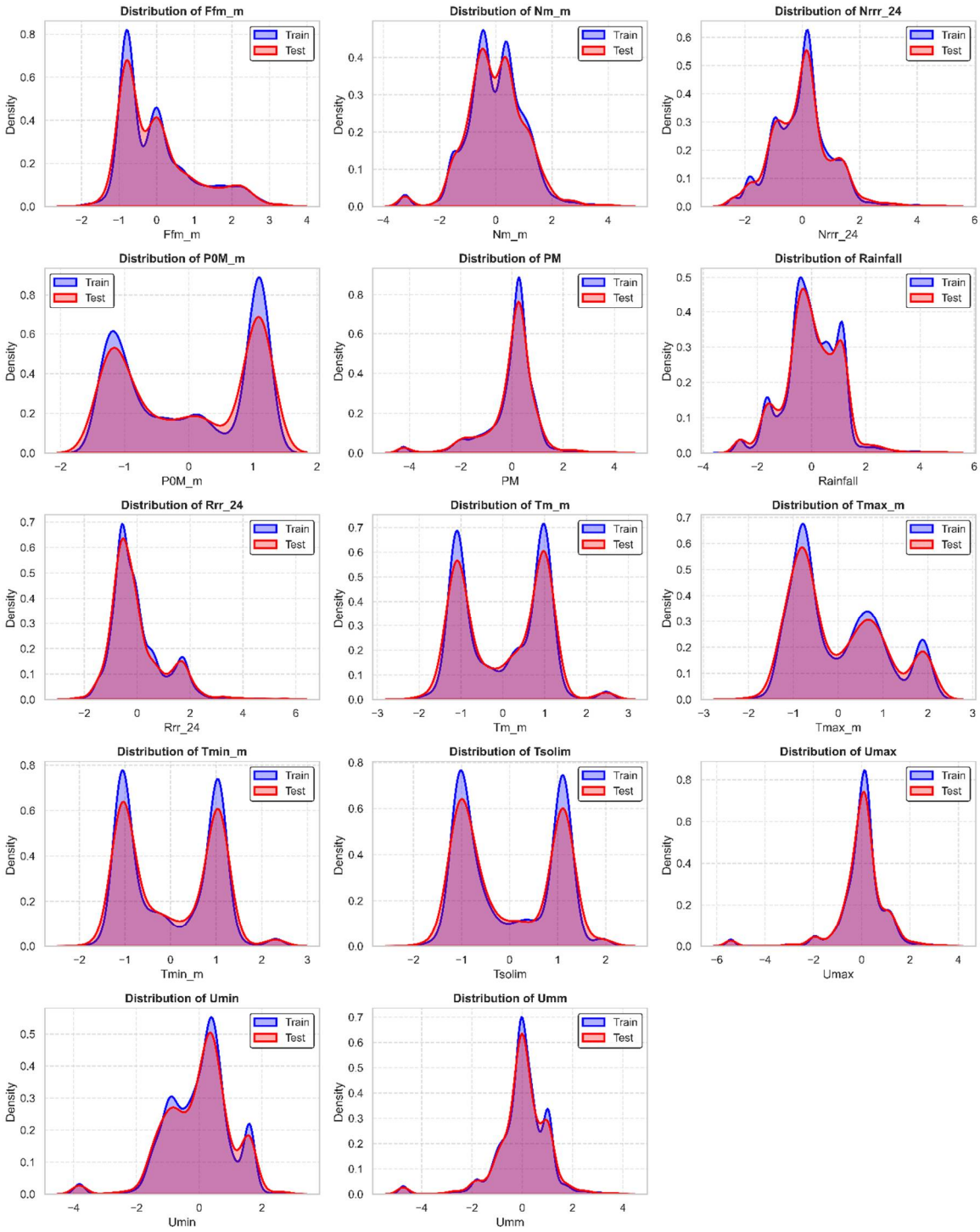
[۳۱] Friedman, S. K. (2011). Janet Franklin (with a contribution by Jennifer A. Miller): Mapping species distributions: Spatial inference and prediction: Cambridge University Press, 2009, 320 pp, Illus, maps; Hardback, ISBN–978-0-521-87635-3 (122.00), Paperback ISBN 978-0-521-70002-3, US $53.00.

[۳۲] Hefley, T. J., & Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, *1*(2), 87-97.

پیوست



شکل ۵- پارتیشن‌بندی و مصورسازی داده‌ها با استفاده از اعتبارسنجی متقابل k-fold برای آموزش و ارزیابی مدل.



شکل ۶ - مقایسه گرافیکی توزیع ویژگی‌ها بین مجموعه داده‌های آموزشی و آزمایشی