



تحلیل تطبیقی غلظت دی اکسید گوگرد (SO_2) در هوای آزاد با داده های ماهواره ای توسط الگوریتم های یادگیری ماشین - مطالعه موردی: شهر ارومیه

ثمین رحیم زاده^۱، مهدی قنبرزاده لک^{۲*}، آرزو سلیمانی^۳

دریافت: ۱۴۰۳/۱۰/۱۰ / پذیرش: ۱۴۰۴/۰۲/۱۵ / دسترسی اینترنتی: ۱۴۰۴/۱۰/۰۶

چکیده

آلودگی هوا به عنوان یکی از معضلات اصلی زیست محیطی، تأثیرات منفی جدی بر سلامت عمومی و کیفیت زندگی دارد. دی اکسید گوگرد، یکی از آلاینده های معیار کیفیت هوای آزاد، علاوه بر ایجاد اثرات زیان بار بر سلامت انسان، نقش مهمی در شکل گیری پدیده های جهانی همچون بارش های اسیدی دارد. پژوهش حاضر با هدف پایش و تحلیل تطبیقی غلظت پایش شده زمینی دی اکسید گوگرد در هوای شهر ارومیه با داده های برگرفته از سنجنده تروپومی از ماهواره ای سنتینل ۵-پی، از الگوریتم های یادگیری ماشین، بهره برده است. داده ها در بازه زمانی ۱۳۹۸ تا ۱۴۰۲ جمع آوری و پس از پیش پردازش، با استفاده از شاخص های

ضریب

ثمین رحیم زاده^۱، مهدی قنبرزاده لک^{۲*} (✉)، آرزو سلیمانی^۳

۱. دانشجوی کارشناسی ارشد مهندسی عمران-محیط زیست، دانشکده فنی و مهندسی، دانشگاه ارومیه، ایران

۲. دانشیار گروه مهندسی عمران، دانشکده فنی و مهندسی، دانشگاه ارومیه، ارومیه،

ایران

۳. دانش آموخته دکتری محیط زیست-آلودگی های محیط زیست، دانشکده منابع طبیعی و محیط زیست، دانشگاه ملایر، ملایر، ایران

Doi: 10.82447/girs.2025.1194923

پست الکترونیکی مسئول مکاتبات: ghanbarzadehlak@urmia.ac.ir

تعیین و ریشه میانگین مربعات خطا تحلیل شدند. نتایج نشان داد که ادغام داده های زمینی و ماهواره ای، همراه با پارامترهای هواشناسی شامل دمای هوا، رطوبت نسبی، میزان تبخیر و تعرق، دمای تر، میزان ابرناکی و فشار هوا در ارتفاع ایستگاه، عملکرد مدل ها را به طور قابل توجهی بهبود بخشیده است. الگوریتم $XGBoost$ با مقادیر R^2 برابر ۰/۸۹ و $RMSE$ معادل $0.57 ppb$ ، به عنوان بهترین مدل شناسایی شد. تحلیل های بدون در نظر گرفتن پارامترهای هواشناسی، کاهش چشمگیر دقت مدل ها را نشان داد، که اهمیت این پارامترها را تأیید می کند (عملکرد مدل $XGBoost$ بدون استفاده از پارامترهای هواشناسی با R^2 معادل ۰/۳۸ و $RMSE$ برابر با $0.75 ppb$ ، کاهش قابل توجهی را نشان می دهد). داده های ماهواره ای قابلیت ایفای نقش کلیدی در پوشش شکاف های اطلاعاتی ناشی از محدودیت های ایستگاه های زمینی را دارند. همچنین، ترکیب فناوری های نوین و داده های محیطی می تواند به شناسایی دقیق تر آلاینده ها و بهبود کیفیت هوا کمک کرده و به عنوان الگویی برای سایر شهرهای ایران و کشورهای در حال توسعه مطرح شود.

چکیده مبسوط

مقدمه: آلودگی هوا یکی از مهم ترین چالش های محیط زیستی در سراسر جهان محسوب شده و تأثیرات گسترده ای بر سلامت عمومی، اکوسیستم ها و اقتصاد جوامع دارد. این معضل به ویژه در مناطق

ارومیه دارند. این مدل می‌تواند به‌عنوان ابزاری کارآمد در مدیریت کیفیت هوا و کاهش اثرات منفی آلودگی در مناطق مختلف مورد استفاده قرار گیرد. همچنین، این تحقیق با بهره‌گیری از الگوریتم‌های پیشرفته یادگیری ماشین، تلاش کرده است تا ارتباطات پیچیده بین متغیرهای مختلف جوی و داده‌های مربوط به غلظت آلاینده‌ها را شناسایی کند. به‌طور خاص، در این مطالعه به بررسی تأثیر متغیرهای جوی نظیر دمای هوا، رطوبت نسبی، میزان تبخیر و تعرق، دمای تر، میزان ابرناکی و فشار هوا در ارتفاع ایستگاه، بر تغییرات غلظت SO_2 پرداخته شده است. هدف دیگر این پژوهش، ارائه راهکارهایی برای بهبود مدیریت کیفیت هوا با استفاده از تکنیک‌های پیشرفته تحلیل داده است.

روش تحقیق: برای انجام این پژوهش، داده‌های ماهواره‌ای مربوط به SO_2 از سنتینل-۵پی با استفاده از پلتفرم گوگل ارث انجین Google Earth Engine استخراج گردید. این داده‌ها با اطلاعات پایش زمینی که از ایستگاه‌های پایش کیفیت هوا (شامل غلظت SO_2 و سایر آلاینده‌ها) و داده‌های هواشناسی شامل دمای هوا، رطوبت نسبی، میزان تبخیر و تعرق، دمای تر، میزان ابرناکی و فشار هوا در ارتفاع ایستگاه به دست آمد، ترکیب شدند. پیش‌پردازش داده‌ها شامل حذف مقادیر پرت و پاک‌سازی داده‌های ناهنجار با هدف بهبود دقت تحلیل‌ها انجام شد. پس از نرمال‌سازی داده‌ها، برای شناسایی و حذف داده‌های پرت از روش Z-Score استفاده گردید؛ در این روش، داده‌هایی با مقدار Z بیشتر از ۳ یا کمتر از -۳ به‌عنوان داده‌های پرت در نظر گرفته شده و از مجموعه داده حذف شدند. علاوه بر این، داده‌های زمینی مورد استفاده در این پژوهش از ایستگاه‌های مستقر در ارومیه جمع‌آوری گردید که شامل داده‌های تاریخی مربوط به غلظت آلاینده‌ها و شرایط جوی بود. سپس از الگوریتم‌های یادگیری ماشین نظیر Random Decision Tree، XGBoost، Gradient Boosting، Extra Trees، Forest CatBoost و AdaBoost برای مدل‌سازی استفاده گردید. در این فرآیند، ۸۰ درصد از داده‌ها به‌عنوان داده‌های آموزش (train) و ۲۰ درصد به‌عنوان داده‌های آزمایش (test) تخصیص داده شدند. ارزیابی عملکرد مدل‌ها با استفاده از معیارهای R^2 و RMSE انجام شد. به‌منظور سنجش میزان تأثیرگذاری متغیرهای جوی بر نتایج مدل، روابط همبستگی بین هر متغیر ورودی و غلظت SO_2 مورد بررسی قرار گرفت. در نهایت، نتایج مدل‌های مختلف مقایسه شدند و پیشنهاداتی برای بهبود دقت تحلیل‌ها ارائه گردید.

شهری با تراکم جمعیتی بالا و فعالیت‌های صنعتی، برجسته‌تر است. آلودگی هوا منجر به بروز مشکلات تنفسی، بیماری‌های قلبی و عروقی و اختلالات مختلف در عملکرد سیستم‌های حیاتی بدن انسان می‌شود. در میان آلاینده‌های گوناگون، دی‌اکسید گوگرد (SO_2) به دلیل نقش برجسته آن در ایجاد بیماری‌های ریوی، تولید باران اسیدی و اثرات منفی بر کیفیت هوا، به‌عنوان یکی از اصلی‌ترین آلاینده‌ها شناخته می‌شود. همچنین، این آلاینده می‌تواند بر کیفیت خاک و منابع آبی نیز اثرگذار باشد و چرخه‌های طبیعی اکوسیستم را مختل کند. این پژوهش با هدف بررسی دقیق غلظت SO_2 در شهر ارومیه، با استفاده از داده‌های ماهواره‌ای سنتینل-۵پی (Sentinel-5P) و ترکیب آن‌ها با روش‌های پیشرفته یادگیری ماشین، انجام شده است.

طرح مسئله: در بسیاری از مناطق در حال توسعه مانند شهر ارومیه، محدودیت‌های جدی در تأمین داده‌های دقیق و پیوسته از کیفیت هوا وجود دارد. کمبود ایستگاه‌های پایش زمینی و نارسایی در ثبت داده‌های مورد نیاز، از جمله چالش‌های عمده در تحلیل روند تغییرات آلودگی و شناسایی عوامل مؤثر بر آن محسوب می‌شوند. این محدودیت‌ها باعث ایجاد شکاف‌های اطلاعاتی در مدیریت کیفیت هوا و تدوین برنامه‌های کاربردی برای کاهش آلودگی می‌شود. در چنین شرایطی، دستیابی به روش‌های نوین که بتوانند این کمبودها را جبران کنند، بسیار ضروری است. استفاده از داده‌های ماهواره‌ای، به‌ویژه داده‌های حاصل از سنجنده تروپومی (TROPOMI) با وضوح زمانی و مکانی بالا، می‌تواند به‌عنوان راهکاری کارآمد برای پر کردن این شکاف‌ها و ارائه اطلاعات مفید در زمینه کیفیت هوا مورد استفاده قرار گیرد. این داده‌ها با ارائه اطلاعات جامع از توزیع فضایی آلاینده‌ها، امکان تحلیل دقیق‌تر الگوهای انتشار و تغییرات زمانی آلاینده‌ها را فراهم می‌کنند. علاوه بر این، تحلیل همبستگی بین داده‌های ماهواره‌ای و داده‌های سطح زمین می‌تواند به درک بهتر الگوهای فضایی و زمانی انتشار آلاینده‌ها کمک کند. در این پژوهش، تلاش شده است تا با ترکیب این داده‌ها و استفاده از الگوریتم‌های یادگیری ماشین، راهکارهای عملی و علمی برای مدیریت بهتر کیفیت هوا ارائه شود.

هدف: این پژوهش با هدف استفاده از داده‌های ماهواره‌ای و ترکیب آن‌ها با داده‌های زمینی برای تحلیل دقیق‌تر غلظت SO_2 و شناسایی عوامل جوی مؤثر بر آن انجام شده است. یکی از اهداف اصلی این مطالعه، ارائه مدلی جامع و قابل‌اعتماد برای پیش‌بینی و مدیریت غلظت SO_2 در مناطق شهری است که محدودیت‌هایی مشابه شهر

می‌تواند الگویی مفید برای سایر شهرهای ایران و کشورهای در حال توسعه باشد. یافته‌های این تحقیق نه تنها به بهبود برنامه‌ریزی‌های زیست‌محیطی کمک می‌کند، بلکه می‌تواند به کاهش اثرات منفی آلودگی بر سلامت عمومی و ارتقای کیفیت زندگی در مناطق شهری کمک نماید. در مناطقی که دسترسی به داده‌های پایش زمینی محدود است، این روش امکان ارائه اطلاعات جامع و دقیق را فراهم کرده و راهکارهای مؤثری برای مقابله با چالش‌های مربوط به آلودگی هوا ارائه می‌دهد. همچنین، این تحقیق نشان داد که استفاده از داده‌های ماهواره‌ای به همراه داده‌های هواشناسی، می‌تواند ابعاد جدیدی از تحلیل کیفیت هوا را آشکار سازد و به توسعه دانش در این حوزه کمک نماید. به طور خاص، این روش می‌تواند در پیش‌بینی تغییرات بلندمدت آلودگی هوا و طراحی سیاست‌های کاهش آلودگی نیز کاربرد داشته باشد. علاوه بر این، تأکید بر اهمیت داده‌های با وضوح بالا و استفاده از تکنیک‌های پیشرفته، نشان‌دهنده ظرفیت بالای این روش‌ها در حل مسائل پیچیده محیط‌زیستی است.

واژگان کلیدی: آلودگی هوای آزاد، SO_2 ، Sentinel-۵P، الگوریتم یادگیری ماشین، ارومیه، پارامترهای هواشناسی

نتایج و بحث: نتایج این پژوهش نشان داد که ترکیب داده‌های ماهواره‌ای و زمینی، دقت مدل‌سازی غلظت SO_2 را به‌طور چشمگیری افزایش می‌دهد. استفاده از الگوریتم‌های پیشرفته یادگیری ماشین توانست روابط پیچیده‌ای که بین متغیرهای مختلف وجود دارد را شناسایی کرده و تحلیل نماید. همچنین، حذف داده‌های مربوط به پارامترهای جوی منجر به کاهش قابل‌توجهی در دقت مدل‌ها شد، که اهمیت بالای این داده‌ها در تحلیل‌های مربوط به آلودگی هوا را نشان می‌دهد. به‌طور کلی، این نتایج تأیید می‌کند که داده‌های ماهواره‌ای می‌توانند به‌عنوان جایگزینی مناسب یا مکملی برای داده‌های زمینی در مناطقی که فاقد ایستگاه‌های پایش هستند، مورد استفاده قرار گیرند. علاوه بر این، استفاده از داده‌های با وضوح بالا به‌ویژه در زمان‌ها و مکان‌های بحرانی، نقش بسزایی در بهبود دقت مدل‌های پیش‌بینی دارد. نتایج مدل‌سازی همچنین نشان داد که تأثیر متغیرهای جوی بر غلظت SO_2 در زمان‌ها و شرایط مختلف متغیر است و این موضوع نیاز به تحلیل دقیق‌تر در مطالعات آتی دارد.

نتیجه‌گیری: این پژوهش نشان داد که ترکیب داده‌های ماهواره‌ای و زمینی، به همراه الگوریتم‌های پیشرفته یادگیری ماشین، ابزاری قوی برای تحلیل و مدیریت کیفیت هوا فراهم می‌کند. این رویکرد

مقدمه

سنجش آلاینده‌های معیار کیفیت هوای آزاد، رصد تغییرات زمانی آن‌ها و چاره‌اندیشی درخصوص نحوه کنترل آلودگی محتمل هوا قبل از وقوع، ازجمله دغدغه‌های اصلی پیشروی متولیان حفظ سلامت و ارتقای بهداشت عمومی جوامع به شمار می‌رود (۲)، ۱۳، ۳۵، ۳۶ و ۴۰). قرار گرفتن طولانی‌مدت در معرض آلاینده‌های هوا می‌تواند علاوه بر افزایش خطرات بروز مشکلات روانی همچون استرس، اضطراب، افسردگی و زوال عقل، پیش‌زمینه ابتلا به بیماری‌های جسمانی نظیر آسم، مشکلات پوستی، سرطان ریه، بیماری‌های قلبی-عروقی و تنفسی را فراهم آورد (۲، ۱۳، ۳۵، ۳۶ و ۴۰). برآوردها نشان می‌دهد که آلودگی هوای آزاد، نه‌تنها سالانه جان حدود پنج میلیون نفر را در سراسر جهان می‌گیرد (۱۳)، بلکه عامل اصلی در مرگ زودرس ۸/۸ میلیون نفر در سال با کاهش میانگین امید به زندگی تا ۲/۹ سال، می‌باشد (۲۱ و ۳۷).

بر اساس یک دسته‌بندی جامع، آلاینده‌های اصلی هوای آزاد را می‌توان در چهار طبقه اصلی قرار داد: (۱) آلاینده‌های گازی (همچون دی‌اکسید گوگرد، مونواکسید کربن، دی‌اکسید نیتروژن، اُزن و ترکیبات آلی فرار)؛ (۲) آلاینده‌های آلی دیرپا (همچون دی‌اکسین‌ها)؛ (۳) ذرات ریز با منشأ فلزات سنگین (نظیر سرب و جیوه)؛ و (۴) ذرات جامد معلق که خود به دو دسته ذرات جامد معلق با قطر آئرو دینامیکی کمتر از ۲/۵ میکرومتر ($PM_{2.5}$) و ذرات جامد معلق با قطر آئرو دینامیکی کمتر از ۱۰ میکرومتر (PM_{10}) تقسیم می‌شوند (۱۷ و ۳۱). در این بین، دی‌اکسید گوگرد (SO_2) به‌عنوان یکی از آلاینده‌های معیار کیفیت هوای آزاد با طول عمر نسبتاً کوتاه، عمدتاً به انتشارهای محلی، نسبت داده می‌شود. بروز مشکلات زیست‌محیطی در مقیاس فرامنطقه‌ای مانند بارش‌های اسیدی (۱۴)، کاهش میدان دید افقی و آسیب به لایه اُزن، همچنین اثرات منفی قابل توجه بر سلامت انسان شامل تحریک دستگاه تنفسی، برونشیت، آسم، سرطان، خشکی چشم، تیرگی قرنیه و بیماری‌های قلبی-عروقی

(۲۴)، از جمله پیامدهای تولید و نشر SO_2 در هوای آزاد به‌شمار می‌روند.

براساس آمار منتشر شده، حداکثر غلظت روزانه SO_2 در برخی مناطق جنوب شرقی اروپا به ۲۱۲۵ میکروگرم بر متر مکعب رسیده است که بیش از ۱۰۰ برابر حد مجاز اعلام شده توسط سازمان بهداشت جهانی (WHO^1) است (۳۳). همچنین دی‌اکسید گوگرد جزء اصلی "مه‌دود لندن" تاریخی است و عامل اصلی مه شدید فعلی در هند و چین به‌شمار می‌رود که هر سال موجب مرگ زودرس بیش از یک میلیون نفر در این مناطق می‌شود (۲۲). کشورهایی با رشد صنعتی سریع مانند چین و هند، به عنوان بزرگ‌ترین تولیدکنندگان SO_2 در جهان شناخته می‌شوند، جایی که افزایش انتشار ناشی از سوزاندن زغال‌سنگ چشمگیر بوده است (۲۹ و ۳۹). نتایج غوزیکالی و همکاران^۲ (۲۰۱۶) در تبریز نشان داد که ۰/۵ درصد از موارد بررسی شده انسداد ریه (در سال ۱۳۹۳)، مربوط به مواجهه شخص با دی‌اکسید گوگرد در غلظت بیش از ۱۰ میکروگرم بر متر مکعب بوده است (۱۲). در سال ۱۳۹۰، غلظت بالای دی‌اکسید گوگرد در تبریز، موجب مرگ ۱۳ نفر شد (۱۶). این وضعیت ضرورت پایش و کنترل انتشار آلاینده دی‌اکسید گوگرد از طریق طرح‌ریزی سیاست‌ها و بهره‌برداری از فناوری‌های نوین به عنوان اولویت اصلی در بهبود کیفیت هوا و سلامت عمومی را مطرح می‌کند (۳۴).

با توجه به محدودیت‌های ایستگاه‌های سنجش زمینی نظیر عدم برداشت داده، کالیبره نبودن (عدم واسنجی) تجهیزات و پوشش محدود فضایی این ایستگاه‌ها، استفاده از داده‌های ماهواره‌ای به عنوان یک منبع داده مکمل، ضروری به نظر می‌رسد. اطلاعات اخیر می‌توانند خلأهای موجود در پوشش زمانی و مکانی ایستگاه‌های سنجش زمینی را پُر کرده و امکان نظارت دقیق‌تری بر کیفیت هوا فراهم آورند. استفاده ترکیبی از داده‌های ماهواره‌ای و اندازه‌گیری‌های زمینی به طور فزاینده‌ای برای بهبود دقت

² Ghozikali et al.

¹ World Health Organization

در پژوهش خود که به تحلیل فعالیت‌های آتشفشانی و انتشار دی‌اکسید گوگرد در منطقه استرومبولی، ایتالیا^۷ پرداخته‌اند، از داده‌های Sentinel-5P در بازه زمانی ۲۰۱۸ تا ۲۰۲۱ برای شناسایی دود آتشفشانی و انتشار دی‌اکسید گوگرد بهره بردند (۷). علاوه بر آن، قنادی و همکاران^۸ (۲۰۲۲)، از داده‌های ماهواره Sentinel-5P در پایش آلودگی SO₂ در نواحی صنعتی ایران استفاده کردند. این مطالعه در بازه زمانی ۲۰۱۸ تا ۲۰۲۰ انجام شده است (۱۱). با توجه به مطالعات پیشین، داده‌های برگرفته از ماهواره Sentinel-5P به دلیل تفکیک بالاتر و قدرت در پایش دقیق آلایندگی دی‌اکسید گوگرد در مقایسه با Aura، در بازه‌های زمانی کوتاه‌مدت تا بلندمدت، ترجیح داده می‌شود. این ماهواره به‌ویژه برای مطالعات در مقیاس شهری و صنعتی بسیار مناسب است.

انتخاب پارامترهای هواشناسی کمکی در تحلیل همبستگی بین داده‌های برگرفته از ایستگاه‌های زمینی پایش SO₂ و اطلاعات حاصل از ماهواره، نقشی کلیدی ایفا می‌کند. در مطالعه کوفانو و همکاران^۹ (۲۰۲۱)، از پارامترهای سرعت و جهت باد، فشار هوا و ارتفاع ستون گاز برای بررسی غلظت دی‌اکسید گوگرد استفاده شد (۷). رحمان و همکاران^{۱۰} (۲۰۲۱)، پارامترهایی همچون دما، رطوبت نسبی و سرعت باد را برای بررسی تأثیر قرنطینه‌های مرتبط با کووید-۱۹ بر غلظت آلایندگی در شهر داکا، بنگلادش، به کار گرفتند (۳۰). در پژوهش ویچزورک^۴ (۲۰۲۳) از متغیرهای کمکی دما، میزان تابش خورشیدی و جهت باد استفاده شد (۳۸). در ایران، کمره‌ای و همکاران^{۱۱} (۲۰۱۷)، با بهره‌گیری از داده‌های دما و فشار هوا، همراه با اطلاعات فعالیت‌های صنعتی، اثرات دی‌اکسید گوگرد و دی‌اکسید نیتروژن را بر سلامت عمومی کمی‌سازی کردند (۱۶). مشاهده چنین تنوعی در انتخاب پارامترهای کمکی، نشان می‌دهد که شرایط جغرافیایی و اقلیمی هر منطقه تأثیر بسزایی در تعیین عوامل مؤثر بر غلظت آلایندگی

مدل‌های کیفیت هوا و پر کردن شکاف‌های داده‌ای در مناطقی که ایستگاه‌های زمینی کمتر توسعه یافته‌اند، توصیه می‌شود (۹). ماهواره سنتینل ۵-پی (Sentinel-5P)، مجهز به ابزار سنجنده پیشرفته تروپومی (TROPOMI)، از سال ۲۰۱۷ به‌طور پیوسته داده‌هایی با وضوح فضایی بالا (۵/۵ × ۳/۵ کیلومتر) را ارائه می‌دهد. این داده‌ها به‌طور خاص برای پایش محلی آلایندگی‌هایی مانند دی‌اکسید گوگرد طراحی شده‌اند و امکان تحلیل دقیق‌تر در مناطق مختلف را فراهم می‌کنند (۳۲). علاوه بر آن، ماهواره آئورا (Aura) متعلق به ناسا، که از سال ۲۰۰۴ فعال است، مجهز به ابزار اُم آی (OMI) بوده و داده‌های جهانی و روزانه‌ای از آلایندگی‌های جوی ارائه می‌دهد. با این حال، وضوح فضایی کمتر اُم آی (OMI) (۲۶۰۰ کیلومتر عرض مشاهده) محدودیت‌هایی در تحلیل دقیق آلودگی‌های محلی ایجاد می‌کند (۲۰ و ۲۳).

در مطالعه جیون و همکاران^۳ (۲۰۲۳)، که به بررسی آلایندگی SO₂ در منطقه آسیا پرداخته‌اند، از داده‌های ماهواره‌ای Sentinel-5P در بازه زمانی ۲۰۱۷ تا ۲۰۲۱ برای شناسایی منابع آلودگی و نقشه‌برداری از مناطق بحرانی استفاده شده است. این تحقیق به‌ویژه از پارامترهای جوی مانند دما و رطوبت، برای تحلیل بهتر استفاده کرده است (۱۵). ویچزورک^۴ (۲۰۲۳)، در مطالعه خود، که به نقشه‌برداری الگوهای آلودگی دی‌اکسید گوگرد در شرق اروپا پرداخته است، از داده‌های سنجنده TROPOMI از ماهواره Sentinel-5P در بازه زمانی ۲۰۲۰ تا ۲۰۲۲ بهره برد (۳۸). سلیمانی و همکاران^۵ (۲۰۲۱)، با استفاده از داده‌های ماهواره‌های Sentinel-5P و Aura در بازه زمانی ۲۰۱۷ تا ۲۰۱۹، به بررسی آلایندگی دی‌اکسید گوگرد در کشور لیتوانی پرداختند. این مطالعه نشان داد که داده‌های برگرفته از Sentinel-5P به دلیل قدرت تفکیک بالاتر نسبت به داده‌های برگرفته از ماهواره Aura، عملکرد بهتری در پایش آلودگی‌های با منشأ دی‌اکسید گوگرد دارد (۳۴). کوفانو و همکاران^۶ (۲۰۲۱)،

⁷ Stromboli, Italy

⁸ Ghannadi et al.

⁹ Rahman et al.

¹⁰ Kamarehie et al.

³ Jion et al.

⁴ Wieczorek.

⁵ Soleimany et al.

⁶ Cofano et al.

(خصوصاً SO_2) خواهد داشت. در جدول ۱، پارامترهای کمکی مورد استفاده توسط دیگر محققان در تحلیل همبستگی بین داده‌های برگرفته از ایستگاه‌های زمینی پایش SO_2 و اطلاعات حاصل از ماهواره، بطور خلاصه آورده شده است.

با توجه به استفاده فراگیر از روش‌های مختلف یادگیری ماشین در تحلیل همبستگی بین داده‌های زمینی و ماهواره‌ای غلظت SO_2 در هوای آزاد، در مقاله حاضر با انتخاب یک مطالعه موردی (شهر ارومیه، مرکز استان آذربایجان غربی)، سعی شده است ضمن تعیین روش کارآمد یادگیری ماشین، پارامترهای کمکی دخیل در افزایش همبستگی، مورد مطالعه قرار گیرد. در جدول ۱، روش‌های کارآمد یادگیری ماشین در مطالعات پیشین قابل مشاهده است. در ارزیابی دقت پیش‌بینی‌ها، عمدتاً از شاخص‌های ضریب تعیین R^2 و ریشه میانگین مربعات خطا^{۱۲} (RMSE) استفاده شده است. حصول ضرایب تعیین بین ۰/۷۰ تا ۰/۹۰ در مطالعات پیشین، با توجه به پارامترهای هواشناسی مورد استفاده در تقویت همبستگی، لزوم بررسی موضوع در مناطق مختلف جغرافیایی را آشکار می‌سازد.

در مطالعات پیشین، استفاده از روش‌های مختلف یادگیری ماشین در مناطق جغرافیایی متفاوت و با بهره‌گیری از پارامترهای کمکی متنوع، نتایج متفاوتی را به همراه داشته است. به‌عنوان مثال، در مطالعه‌ای که توسط کاظمی گراجه و همکاران^{۱۱} (۲۰۲۳)، در شهر اراک (مرکز استان مرکزی) صورت گرفته است، از روش جنگل تصادفی (Random Forest) با پارامترهای کمکی شامل دما، سرعت و جهت باد و پوشش گیاهی استفاده شده و مقدار $R^2 = ۰/۹۱$ و $RMSE = ۰/۱۶۲$ ppb به دست آمد (۱۸). در مقابل، در مطالعه‌ای که توسط رحمان و همکاران^۹ (۲۰۲۱)، در شهر داکا پایتخت بنگلادش انجام شده است، از روش XGBoost با پارامترهای دما و رطوبت نسبی استفاده شد که مقادیر $R^2 = ۰/۷۲$ و $RMSE = ۰/۵۱۳$ ppb حاصل گردید (۳۰). همچنین، در مطالعه‌ای که توسط اوغوز و پکین^{۱۵} (۲۰۲۴)، در کشور ترکیه، با روش AdaBoost با استفاده از پارامترهای دما، رطوبت نسبی، سرعت باد و فشار اتمسفر به کار رفت، مقادیر $R^2 = ۰/۸۴$ و $RMSE = ۰/۲۱۵$ ppb به دست آمده است (۲۸).

جدول ۱. روش‌های یادگیری ماشین، شاخص‌های ارزیابی و پارامترهای کمکی هواشناسی مورد استفاده در مطالعات پیشین.

Table 1. Machine learning methods, evaluation metrics, and auxiliary parameters in previous studies.

منبع	روش یادگیری ماشین منتخب	RMSE	(R^2)	موقعیت مورد مطالعه	دوره زمانی	پارامترهای کمکی
کاظمی گراجه و همکاران ^{۱۱} ، (۲۰۲۳)	Random Forest, linear regression	۰/۱۶۲	۰/۹۱	شهر اراک مرکز استان مرکزی	۲۰۱۸-۲۰۲۰	دما، سرعت و جهت باد، پوشش گیاهی
ممتاز و همکاران ^{۱۴} ، (۲۰۲۳)	Random Forest, XGBoost	۰/۱۹۲	۰/۸۹	لاهور و اسلام‌آباد واقع در پاکستان	۲۰۲۰-۲۰۲۱	دما، رطوبت نسبی، سرعت و جهت باد، تغییرات فصلی
ویچزورک ^۴ ، ۲۰۲۳	Random Forest, Gradient Boosting	۰/۲۴۱	۰/۸۸	لهستان و اوکراین	۲۰۱۸-۲۰۲۰	دما، سرعت و جهت باد، بارندگی
بلانکو و همکاران ^{۱۵} ، (۲۰۲۴)	Gradient Boosting, Linear regression	۰/۲۳۳	۰/۸۷	شهر میلان واقع در ایتالیا	۲۰۲۰-۲۰۲۱	دما، رطوبت نسبی، سرعت و جهت باد، فشار هوا، نقطه شبنم، بارندگی، پوشش ابر، تابش خورشیدی

^{۱۴} Mumtaz et al.

^{۱۵} Blanco et al.

^{۱۱} Coefficient of Determination

^{۱۲} Root Mean Square Error

^{۱۳} Kazemi Garajeh et al.

منبع	روش یادگیری ماشین منتخب	RMSE	(R ²)	موقعیت مورد مطالعه	دوره زمانی	پارامترهای کمکی
آنگرانی و همکاران ^{۱۶} ، (۲۰۲۴)	Random Forest	۰/۲۸۰	۰/۸۵	جهانی	۲۰۲۰-۲۰۲۱	دما، رطوبت نسبی، سرعت باد، فشار اتمسفر، بارندگی، تابش خورشیدی
اوغوز و پیکین ^{۱۷} ، (۲۰۲۴)	AdaBoost	۰/۲۱۵	۰/۸۴	کشور ترکیه	۲۰۲۱	دما، رطوبت نسبی، سرعت باد، فشار اتمسفر
سلیمانی و همکاران ^{۱۸} ، (۲۰۲۱)	Decision Tree	۰/۴۵۶	۰/۸۳	شهر کاوناس واقع در لیتوانی	۲۰۱۸-۲۰۱۹	دما، رطوبت نسبی، سرعت باد، بارش
فانیا و همکاران ^{۱۸} ، (۲۰۲۴)	Gradient Boosting, Random Forest	۰/۳۹۸	۰/۷۹	کشور ایتالیا	۲۰۱۹-۲۰۲۱	دما، سرعت باد، فشار هوا، تابش خورشیدی
رحمان و همکاران ^{۱۹} ، (۲۰۲۱)	XGBoost	۰/۵۱۳	۰/۷۲	داکا پایتخت بنگلادش	۲۰۱۹-۲۰۲۰	دما، رطوبت نسبی

با توجه به اهمیت آلاینده دی‌اکسید گوگرد و اثرات آن بر سلامت انسان و محیط زیست، ضرورت پایش مستمر این آلاینده به‌ویژه در مناطق شهری بیش از پیش احساس می‌شود. با این حال، ایستگاه‌های زمینی در بسیاری از مناطق به دلایل مختلفی مانند محدودیت در پوشش مکانی، اختلالات فنی یا کاهش کیفیت داده‌ها در برخی بازه‌های زمانی، با چالش‌هایی مواجه‌اند. در چنین شرایطی، استفاده از داده‌های ماهواره‌ای می‌تواند به‌عنوان یک منبع مکمل، خلأ اطلاعاتی ناشی از نبود یا ضعف عملکرد ایستگاه‌های زمینی را تا حدی جبران کند. به‌ویژه در مناطقی که فاقد ایستگاه‌های فعال هستند، بهره‌گیری از این داده‌ها همراه با پارامترهای هواشناسی و تکنیک‌های نوین مانند یادگیری ماشین، امکان برآورد نسبی غلظت آلاینده‌ها را فراهم خواهد آورد. هدف اصلی این پژوهش، ارزیابی ارتباط میان داده‌های زمینی و داده‌های ماهواره‌ای غلظت دی‌اکسید گوگرد (SO₂) با استفاده از الگوریتم‌های مختلف یادگیری ماشین و تعیین مؤثرترین پارامترهای هواشناسی در افزایش دقت مدل است. بر این اساس،

در این پژوهش از ترکیب داده‌های ماهواره‌ای و زمینی برای مدل‌سازی غلظت SO₂ در هوای شهری بهره گرفته شد تا در شرایط کمبود یا نبود داده‌های زمینی، امکان تخمین غلظت آلاینده و ارائه هشدارهای لازم برای مدیریت کیفیت هوا فراهم گردد. این امر به‌ویژه در مواقع اضطراری اهمیت دارد، چرا که می‌تواند به اطلاع‌رسانی به‌موقع به شهروندان و تدوین راهکارهایی برای کاهش آلودگی منجر شود. پژوهش حاضر با استفاده از روش‌های یادگیری ماشین، امکان ارزیابی دقیق داده‌های ماهواره‌ای و تلفیق آن با منابع زمینی را برای پایش غلظت SO₂ در منطقه‌ای خاص مانند شهر ارومیه فراهم کرده است.

روش تحقیق

موقعیت منطقه مورد مطالعه

شهر ارومیه، مرکز استان آذربایجان غربی، در مختصات جغرافیایی ۳۷/۴۸ تا ۳۷/۶ درجه عرض شمالی و ۴۵ تا ۴۵/۱۲

¹⁶ Anggraini et al.

¹⁷ Oğuz & Pekin.

¹⁸ Fania et al.

درجه طول شرقی واقع شده است (شکل ۱). این شهر در مجاورت دریاچه ارومیه، که بزرگ‌ترین دریاچه داخلی ایران و دومین دریاچه آب شور در خاورمیانه محسوب می‌شود، واقع شده است. اقلیم منطقه مورد مطالعه تحت تأثیر جریان‌های هوایی اقیانوس اطلس و مدیترانه‌ای قرار داشته و در دسته مناطق نیمه خشک طبقه‌بندی می‌شود. میانگین سالانه بارندگی در این منطقه حدود ۳۵۰ میلی‌متر گزارش شده است (۱ و ۲۷).

روش استخراج داده‌ها

داده‌های غلظت آلاینده‌های PM_{10} , NO_x , NO_2 , NO , SO_2 ، از ایستگاه پایش آلودگی هوای ارومیه (ایستگاه شماره ۳ - شکل ۱) دریافت شدند. این ایستگاه در مختصات $37/5544$ درجه عرض جغرافیایی شمالی و $45/0484$ درجه طول جغرافیایی شرقی قرار دارد و اطلاعات مورد نیاز برای این مطالعه به صورت روزانه طی دوره زمانی پنج‌ساله، از سال ۱۳۹۸ تا ۱۴۰۲، جمع‌آوری و تحلیل شده است. ایستگاه سینوپتیک هواشناسی ارومیه (شکل ۱) که در محدوده فرودگاه این شهر واقع شده، به عنوان منبع داده‌های هواشناسی مورد استفاده قرار گرفت. این ایستگاه در مختصات جغرافیایی $37/5253$ درجه عرض شمالی و $39/3988$ درجه طول شرقی و در ارتفاع ۱۳۵۵ متری از سطح دریا قرار گرفته است. تأسیس این ایستگاه به سال ۱۳۳۰ بازمی‌گردد و داده‌های هواشناسی مورد نیاز (بازه زمانی ۱۳۹۸ تا ۱۴۰۲) شامل دمای حداقل، حداکثر و متوسط هوا، دمای تر، دمای نقطه شبنم، فشار بخار موجود و اشباع، رطوبت نسبی، فشار کاهیده به ارتفاع سطح دریا و فشار در ارتفاع ایستگاه، سرعت و جهت باد، مقادیر حداکثر سرعت و جهت باد، میزان تبخیر از تشت، مجموع تعداد ساعات آفتابی روزانه، میزان بارش ۶ و ۲۴ ساعته، درجه ابرناکی، میزان تابش ۲۴ ساعته خورشید و میزان دید افقی، از آن جمع‌آوری شده است (این داده‌ها با ارسال درخواست رسمی به سازمان هواشناسی کشور و از طریق پایگاه اینترنتی این سازمان دریافت شده‌اند: <https://data.irimo.ir/login/login.aspx>). علاوه بر

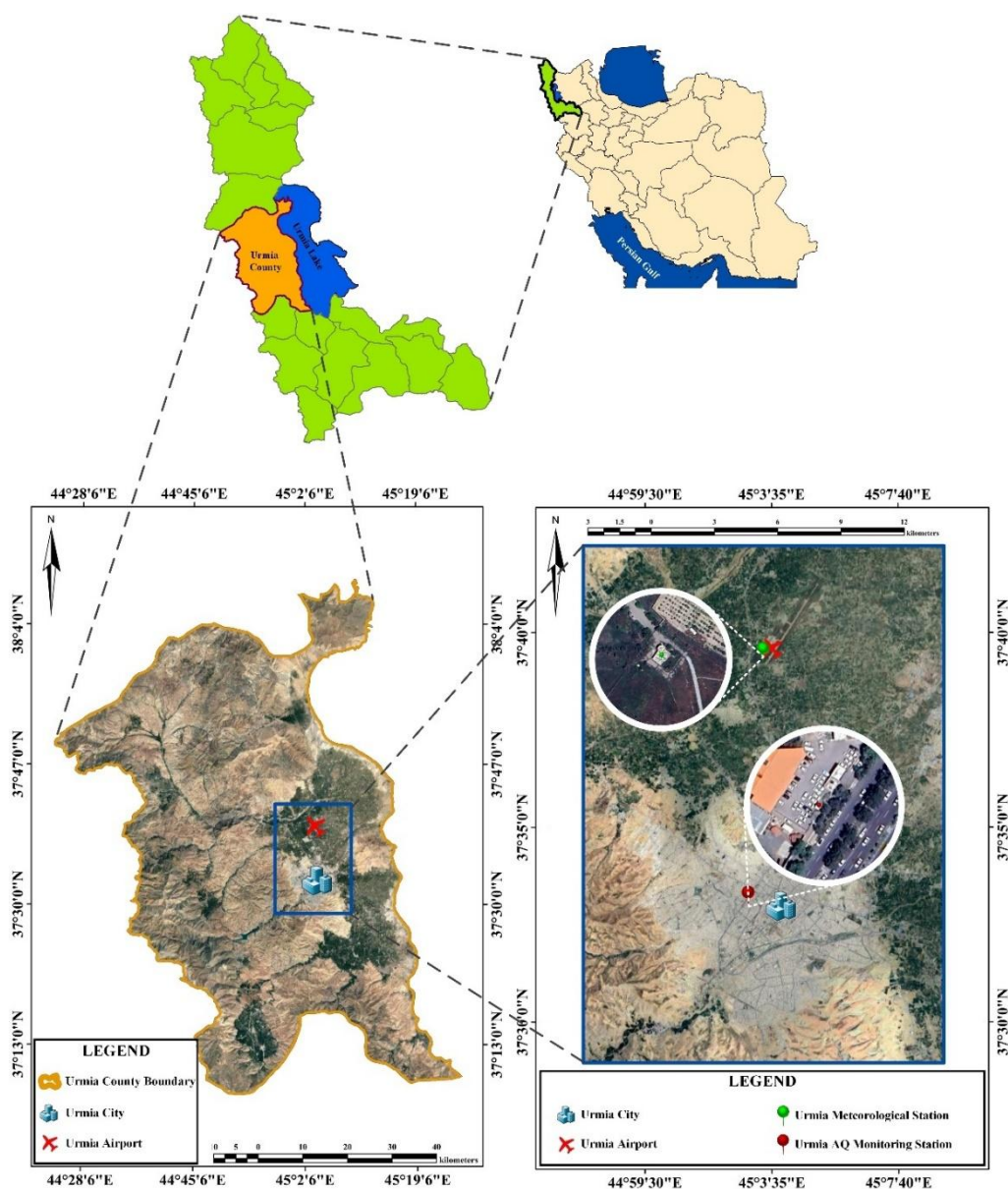
این، داده‌های ماهواره‌ای غلظت دی‌اکسیدگوگرد از ماهواره Sentinel-5P با استفاده از پلتفرم گوگل ارث انجین (Google Earth Engine - GEE) (۸) برای مختصات ایستگاه شماره ۳ سنجش زمینی آلاینده‌های هوای آزاد ارومیه طی دوره زمانی ۱۳۹۸ تا ۱۴۰۲ استخراج شدند (برای اطلاعات بیشتر به وبسایت <https://earthengine.google.com/> مراجعه شود). این داده‌ها پس از تأیید اعتبار برای استفاده در تحلیل‌های بعدی آماده شدند.

پیش پردازش داده‌ها

در مرحله اول، داده‌های گم‌شده یا دارای گپ (Missing Data) شناسایی و حذف شدند. برای این منظور، از روش‌های مناسب پُر کردن گپ‌های داده استفاده شد تا مجموعه داده‌ها از نظر کامل بودن، آماده باشند. سپس داده‌های پرت (Outliers) به دلیل تأثیر منفی آن‌ها بر دقت مدل‌های یادگیری ماشین، شناسایی و حذف گردیدند. شناسایی داده‌های پرت با استفاده از معیار Z-Score انجام شد. این داده‌ها ممکن است به دلایلی مانند خرابی دستگاه‌ها، خطای اندازه‌گیری، یا شرایط آبی و غیرمعمول که منجر به خطای نقطه‌ای می‌شود، ایجاد شده باشند.

در مرحله بعد، داده‌ها به‌طور جداگانه برای هر مجموعه (داده‌های زمینی و داده‌های ماهواره‌ای) به صورت روزانه میانگین‌گیری شدند تا مقادیر هم‌راستا و یکپارچه شوند. سپس، برای همسان‌سازی مقیاس داده‌ها و ایجاد قابلیت مقایسه بین آن‌ها، از فرآیند نرمال‌سازی با استفاده از روش StandardScaler استفاده شد. این عملیات موجب شد که تمامی متغیرها در یک دامنه استاندارد قرار گیرند. در نهایت، داده‌های نرمال‌شده و همگن‌شده بر اساس شاخص زمانی، به‌طور هم‌زمان ادغام شدند

تا برای استفاده در الگوریتم‌های یادگیری ماشین آماده گردند.



شکل ۱- موقعیت جغرافیایی مکانی: استان آذربایجان غربی؛ شهرستان ارومیه؛ شهر ارومیه؛ ایستگاه سنجش آلاینده‌های هوای شهر ارومیه و ایستگاه هواشناسی فرودگاه ارومیه.

Figure 1. Geographical location: West Azerbaijan Province; Urmia County; Urmia City; Urmia urban air quality monitoring station and Urmia Airport meteorological station

یادگیری ماشین

شده و داده‌هایی که برای یک روز کامل نبودند، حذف شدند. این پیش‌پردازش داده‌ها به منظور بهبود عملکرد مدل‌ها و افزایش دقت نتایج ضروری بوده است (۵). برای اجرای الگوریتم‌های یادگیری ماشین، از زبان برنامه‌نویسی پایتون نسخه ۳/۸ در محیط تعاملی جویپتر نوت‌بوک استفاده شد. جویپتر نوت‌بوک با قابلیت

در این پژوهش، از یادگیری ماشین به عنوان یکی از شاخه‌های هوش مصنوعی در تحلیل داده‌ها و ایجاد همبستگی بین داده‌های زمینی و ماهواره‌ای غلظت SO_2 استفاده شده است. داده‌های این مطالعه به صورت متوسط روزانه برای تمامی متغیرها جمع‌آوری

اجرای کدها به صورت تکه‌ای و نمایش همزمان نتایج، ابزاری مؤثر در تحلیل داده‌ها و توسعه مدل‌ها به شمار می‌رود (۱۸).

در ادامه، از کتابخانه‌های کلیدی زبان برنامه‌نویسی پایتون شامل نام پای (NumPy)، پنداس (Pandas)، مت پلات لیب (Matplotlib) و سای کیت-لرن (Scikit-learn) استفاده شد. کتابخانه NumPy به دلیل قابلیت‌های پیشرفته در پردازش داده‌ها و انجام محاسبات عددی، امکان مدیریت کارآمد آرایه‌ها و ماتریس‌های چندبعدی را فراهم می‌آورد و نقش مهمی در آماده‌سازی داده‌ها برای الگوریتم‌های یادگیری ماشین ایفا می‌کند. از سوی دیگر، کتابخانه Pandas با ساختارهای داده‌ای مانند چهارچوب داده (DataFrame)، به عنوان ابزاری قدرتمند برای خواندن، پردازش و تجزیه و تحلیل داده‌ها مورد استفاده قرار گرفت. این کتابخانه با ارائه قابلیت‌هایی نظیر پاک‌سازی و دستکاری داده‌ها، فرآیند آماده‌سازی داده‌ها را تسهیل می‌کند.

علاوه بر این، برای تجسم داده‌ها و نمایش نتایج مدل‌ها، از کتابخانه Matplotlib بهره گرفته شد. این کتابخانه امکان ترسیم نمودارهای متنوع دوبعدی و سه‌بعدی را فراهم آورده و به تحلیل و درک بهتر روابط بین متغیرها کمک می‌کند. در نهایت، الگوریتم‌های یادگیری ماشین با استفاده از کتابخانه Scikit-learn پیاده‌سازی شدند. این کتابخانه به دلیل پوشش گسترده الگوریتم‌های طبقه‌بندی، رگرسیون و خوشه‌بندی، انتخابی مناسب برای توسعه مدل‌های یادگیری ماشین محسوب می‌شود. Scikit-learn، همچنین، با بهینه‌سازی محاسبات و مستندسازی دقیق، کارایی و دقت مدل‌ها را در فرآیند تحلیل داده‌ها تضمین می‌کند. ترکیب این ابزارها و کتابخانه‌ها، بستری یکپارچه و کارآمد برای توسعه و ارزیابی مدل‌های یادگیری ماشین فراهم کرده و به دستیابی به نتایج دقیق و قابل اعتماد کمک شایانی خواهد کرد.

در این پژوهش، از (۱) الگوریتم‌های مبتنی بر درخت تصمیم مانند درخت تصمیم (Decision Tree)، جنگل تصادفی (Random Forest) و درخت‌های اضافی (Extra Trees)

به دلیل قدرت بالا در تحلیل داده‌ها و پیش‌بینی دقیق استفاده شده است. همچنین، از (۲) الگوریتم‌های مبتنی بر تقویت (Boosting) شامل گرادیان تقویتی (Gradient Boosting)، تقویت گرادیان پیشرفته (XGBoost)، کت‌بوست (CatBoost) و آدابوست (AdaBoost) بهره گرفته شد. این الگوریتم‌ها با ترکیب چندین مدل ضعیف و کاهش خطای پیش‌بینی، در تحلیل داده‌های پیچیده و برآورد غلظت آلاینده‌ها مانند SO_2 عملکرد مؤثری دارند.

نحوه تحلیل نتایج مدل‌سازی

جهت بررسی همبستگی بین غلظت SO_2 و پارامترهای هواشناسی و سایر آلاینده‌های معیار در سری زمانی سال‌های ۱۳۹۸ تا ۱۴۰۲، از شاخص ضریب همبستگی پیرسون (R) استفاده گردید. این شاخص شدت و جهت رابطه خطی بین دو متغیر پیوسته را ارزیابی می‌کند. ضریب پیرسون عددی بین ۱- و ۱+ است؛ که ۱+ نشان‌دهنده همبستگی کامل مثبت، ۱- نشان‌دهنده همبستگی کامل منفی و ۰ به معنای نبود رابطه خطی بین متغیرهاست. مقدار مثبت R به این معنی است که با افزایش یکی از متغیرها، دیگری نیز تمایل به افزایش دارد، و مقدار منفی آن نشان‌دهنده این است که با افزایش یکی از متغیرها، دیگری کاهش می‌یابد (۳۶). در این مطالعه، ضریب پیرسون به منظور بررسی شدت و جهت ارتباط بین غلظت دی‌اکسید گوگرد و پارامترهای هواشناسی و سایر آلاینده‌های هوا (برگرفته از سنجش زمینی) با استفاده از زبان برنامه‌نویسی Python محاسبه گردیده است (رابطه ۱).

برای ارزیابی دقت عملکرد مدل‌های یادگیری ماشین، ضریب تعیین (R^2) و ریشه میانگین مربعات خطا (RMSE) به عنوان معیارهای کلیدی، استفاده شدند. شاخص R^2 نشان‌دهنده درصد تغییرات متغیر وابسته است که توسط متغیرهای مستقل مدل توضیح داده می‌شود و مقادیر آن بین ۰ تا ۱ قرار دارد. مقدار نزدیک به ۱ به معنای قدرت بالای مدل در تطبیق با داده‌های واقعی و توضیح تغییرات آن‌هاست، در حالی که مقدار نزدیک

به صفر ضعف مدل را نشان می‌دهد (رابطه ۲). از سوی دیگر، شاخص RMSE به عنوان شاخصی معتبر برای سنجش میزان خطای مدل پیش‌بینی عمل می‌کند. این شاخص در صورتی که پیش‌بینی‌ها به داده‌های واقعی نزدیک باشند، به صفر میل کرده و در صورت وجود اختلافات قابل توجه، به طور چشمگیری افزایش می‌یابد (رابطه ۳). ترکیب مقدار پایین ریشه میانگین مربعات خطا (RMSE) و مقدار بالای ضریب تعیین (R²) نشان‌دهنده دقت مناسب مدل در تحلیل داده‌ها است (۴، ۲۸ و ۳۴).

$$R = \frac{\sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^n (b_i - \bar{b})^2}} \quad [1]$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad [2]$$

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \quad [3]$$

که در آن a_i و b_i زوج داده‌های i ام مدنظر جهت بررسی همبستگی و تعیین ضریب پیرسون، $\bar{a}$ و $\bar{b}$ میانگین مقادیر زوج داده‌های اخیر، y_i مقدار مشاهداتی i ام، $\hat{y}_i$ مقدار پیش‌بینی شده i ام توسط مدل، $\bar{y}$ میانگین مقادیر مشاهداتی و n تعداد نمونه‌ها است.

نتایج و بحث

در شکل ۲، غلظت‌های متوسط روزانه SO₂ بر اساس داده‌های اخذ شده از ایستگاه پایش آلودگی هوای شهر ارومیه آورده شده است. فقدان داده‌های زمینی در برخی از سال‌ها (به عنوان مثال، روزهای میانی سال ۱۳۹۸ و روزهای انتهایی سال ۱۴۰۲) و نوسانات غلظت در روزهای مختلف سال، از جمله دلایل اصلی در استفاده از داده‌های کمکی (در اینجا اطلاعات برگرفته از سنجنده TROPOMI از ماهواره Sentinel-5P)، به شمار می‌روند. براین اساس و با در دست داشتن اطلاعات ماهواره‌ای

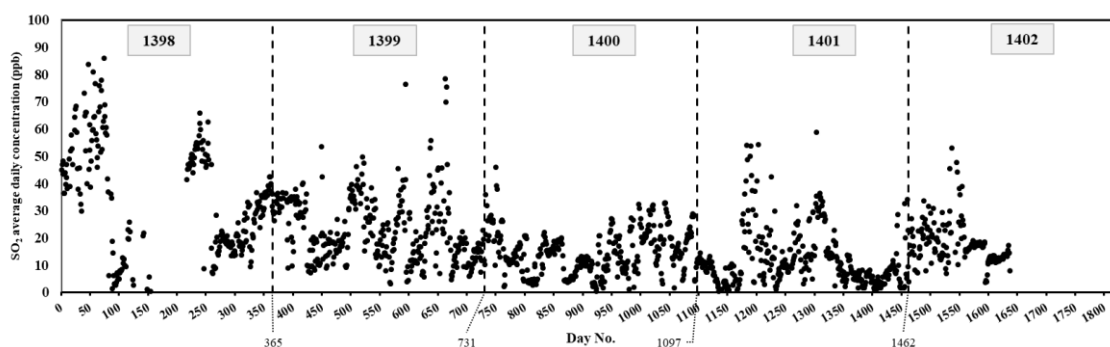
می‌توان روند تغییرات غلظت SO₂ در هوای آزاد را با دقت بیشتری مورد مطالعه قرار داد. ذکر این نکته ضروری است که در تمامی روزهای داده‌برداری شده طی سال‌های ۱۳۹۸ تا ۱۴۰۲، کیفیت هوای منطقه مورد مطالعه از حد استاندارد ملی (ppb) ۱۴۰ برای متوسط غلظت روزانه تجاوز ننموده و بیشترین غلظت گزارش شده طی این مدت مربوط به روز ۱۲ خرداد ۱۳۹۸ با مقدار ۸۵/۸۹ ppb می‌باشد. با این وجود، حداکثر غلظت ساعتی SO₂ در محدوده مورد مطالعه در ۶۸ روز از حد استاندارد ملی ۷۵ ppb تجاوز داشته است (با میانگین صدک ۹۹ ام معادل ۱۰۲/۶۲ ppb، ۱۰۰/۹۴ ppb و ۹۹/۱۴ ppb به ترتیب برای متوسط سه ساله ۱۳۹۸-۱۴۰۰، ۱۳۹۹-۱۴۰۱ و ۱۴۰۲-۱۴۰۰). این واقعیت، لزوم طرح‌ریزی مطالعات درخصوص انتشار آلاینده SO₂ در محدوده مطالعاتی را آشکار می‌سازد.

به منظور تعیین پارامترهای هواشناسی تأثیرگذار در ارتقای شدت همبستگی بین داده‌های زمینی و ماهواره‌ای غلظت SO₂، در ابتدا لازم است وجود رابطه بین پارامترهای مختلف هواشناسی و غلظت زمینی آلاینده دی‌اکسید گوگرد در منطقه مورد مطالعه، مورد بررسی قرار گیرد. داده‌های هواشناسی از ایستگاه سینوپتیک فرودگاه ارومیه جمع‌آوری شده و شامل مجموعه‌ای از پارامترهای متنوع بود که برای تحلیل همبستگی و ارزیابی تأثیر آن‌ها مورد استفاده قرار گرفتند. این پارامترها عبارت‌اند از: دید افقی (vv)، میزان ابرناکی (n)، جهت باد (dd)، سرعت باد (ff)، دمای هوا (t)، دمای نقطه شبنم (td)، فشار کاهیده به ارتفاع سطح دریا (P)، فشار هوا در ارتفاع ایستگاه (P₀)، میزان بارش شش ساعته (rrr)، میزان بارش ۲۴ ساعته (rrr24)، رطوبت نسبی (u)، دمای حداقل هوا (t_{min})، دمای حداکثر هوا (t_{max})، میزان تابش ۲۴ ساعته خورشید (radglo24)، مجموع ساعات آفتابی روزانه (sshn)، میزان تبخیر و تعرق (evt)، دمای تر (twet)، حداکثر سرعت باد (ff_{max})، حداکثر جهت باد (dd_{max})، فشار بخار (ew) و فشار بخار اشباع (ews). به منظور بررسی تأثیر پارامترهای منتخب هواشناسی بر غلظت دی‌اکسید گوگرد، از تحلیل‌های آماری و نمودار حرارتی (Heatmap) استفاده شد

به دلایلی چون تغییرات سریع و مقطعی در شرایط جوی، تأثیرات کوتاه مدت یا ناپایدار پارامترها و یا حتی تفاوت های مقیاس زمانی بین اندازه گیری ها مربوط باشد. به عنوان مثال، سرعت باد و فشار هوا در ارتفاع ایستگاه ممکن است در مقیاس های زمانی طولانی تر یا در شرایط جوی خاص اثرات بیشتری داشته باشند. همچنین، شرایط خاص جغرافیایی و اقلیمی منطقه ارومیه نیز می تواند به عنوان عامل دیگری بر شدت یا ضعف همبستگی ها تأثیرگذار باشد.

(شکل ۳). نتایج تحلیل ها نشان داد که پارامترهای دمای هوا (با ضریب همبستگی پیرسون معادل $-0/23$)، رطوبت نسبی ($+0/20$)، میزان تبخیر و تعرق ($-0/20$)، دمای تر ($-0/19$)، میزان ابرناکی ($+0/17$) و فشار هوا در ارتفاع ایستگاه ($+0/13$)، بیشترین همبستگی با غلظت دی اکسید گوگرد زمینی در ارومیه داشته اند. میزان همبستگی این پارامترها با غلظت دی اکسید گوگرد در بازه $0/12$ تا $0/24$ متغیر بود.

در این مطالعه برخی پارامترهای هواشناسی همبستگی ضعیفی با غلظت دی اکسید گوگرد نشان دادند. این موضوع ممکن است



شکل ۲- غلظت متوسط روزانه SO_2 پایش زمینی طی بازه زمانی مورد مطالعه.

Fig 2- Daily average ground monitoring SO_2 concentrations during the studied period.

در ادامه و به منظور بررسی وجود ارتباط میان غلظت های زمینی دی اکسید گوگرد و سایر آلاینده های پایش شده زمینی شامل دی اکسید نیتروژن (NO_2)، اکسیدهای نیتروژن (NO_x)، مونوکسید نیتروژن (NO)، ذرات معلق با قطر کمتر از 10 میکرومتر (PM_{10})، ذرات معلق با قطر کمتر از $2/5$ میکرومتر ($PM_{2.5}$) و مونوکسید کربن (CO)، مجدداً از نمودار حرارتی (Heatmap) استفاده شد (شکل ۴). براساس این تحلیل، می توان با وارد کردن مقادیر غلظت سایر آلاینده ها در ایجاد همبستگی بین داده های برگرفته از ایستگاه های زمینی پایش SO_2 و اطلاعات حاصل از ماهواره، در راستای تقویت آن گام برداشت.

اگرچه دمای حداقل (t_{min}) و دمای حداکثر (t_{max}) نیز همبستگی بالایی با غلظت SO_2 نشان دادند (به ترتیب $-0/21$ و $-0/23$)، اما به دلیل همبستگی بسیار قوی آن ها با دمای هوا (t) که مقدار آن به ترتیب $0/93$ و $0/98$ است، در این پژوهش صرفاً دمای هوا (t) به عنوان نماینده این گروه انتخاب شده است. لازم به ذکر است که مقادیر ارائه شده در نمودار حرارتی (Heatmap) ضرایب همبستگی پیرسون هستند. برای تعیین پارامترهای تأثیرگذار، از قدر مطلق ضرایب پیرسون استفاده شده است. بنابراین، پارامترهایی که دارای بزرگترین مقادیر مطلق همبستگی بودند، به عنوان عوامل کلیدی در ارتقای همبستگی بین داده های زمینی و ماهواره ای SO_2 انتخاب شدند.

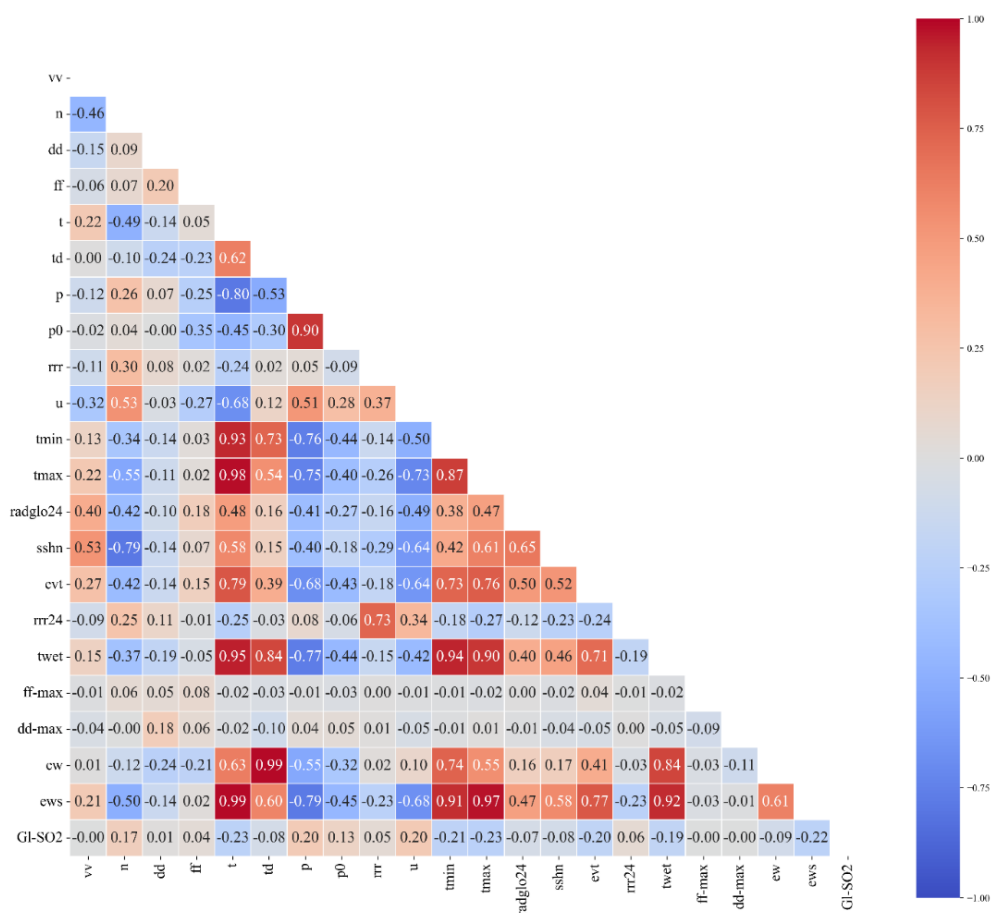
دامنه عددی متفاوتی دارند. به همین دلیل، برای ایجاد یکپارچگی و جلوگیری از تأثیرات منفی تفاوت مقیاس‌ها بر مدل‌های یادگیری ماشین، تمام داده‌ها پس از حذف مقادیر ناقص، نرمال‌سازی شدند. این فرایند باعث می‌شود که تمام پارامترها تأثیر مشابهی بر مدل داشته باشند و تحلیل‌ها دقیق‌تر و قابل اعتمادتر انجام شود.

مجموعه‌ای از الگوریتم‌های یادگیری ماشین برای تحلیل داده‌ها استفاده شد که به ترتیب از مدل‌های ساده به پیچیده‌تر انتخاب شدند. ابتدا الگوریتم درخت تصمیم (Decision Tree) برای مدل‌سازی اولیه روابط میان متغیرها به کار گرفته شد. سپس، الگوریتم جنگل تصادفی (Random Forest) به دلیل قابلیت کاهش بیش‌برازش (overfitting) و پردازش داده‌های پیچیده‌تر مورد بهره‌برداری قرار گرفت (۲۴). در نهایت، الگوریتم درخت‌های اضافی (Extra Trees) به دلیل ساخت درخت‌های تصادفی بیشتر انتخاب گردید. برای بهبود دقت، از الگوریتم‌های تقویت (Boosting) مانند الگوریتم تقویت گرادیان (Gradient Boosting) و نسخه بهینه‌سازی شده آن الگوریتم تقویت گرادیان پیشرفته (XGBoost¹⁹)، استفاده شد که دقت و سرعت بالاتری دارند. همچنین، الگوریتم کت بوست (CatBoost) برای پردازش داده‌های دسته‌ای به کار رفت. همچنین، الگوریتم آدابوست (AdaBoost²⁰) به عنوان الگوریتمی جهت تقویت مدل‌های ضعیف و بهبود پیش‌بینی‌ها در داده‌هایی با پراکندگی بالا، انتخاب شد. برای تقسیم‌بندی داده‌ها، ۸۰ درصد داده‌ها به عنوان داده‌های آموزش (train) و ۲۰ درصد به عنوان داده‌های آزمایش (test) انتخاب شدند. در جدول ۲، نتایج مربوط به ارزیابی دقت پیش‌بینی مدل‌های مختلف آورده شده است. مقادیر ضریب تعیین (R^2) و ریشه میانگین مربعات خطا (RMSE) برای مقایسه کارایی الگوریتم‌ها و شناسایی بهترین مدل، مورد استفاده قرار گرفت.

نتایج تحلیل همبستگی پیرسون نشان داد که ذرات معلق با قطر کمتر از ۱۰ و ۲/۵ میکرومتر همبستگی معناداری با غلظت دی‌اکسید گوگرد ندارند. این امر احتمالاً به دلیل تفاوت در منابع انتشار این آلاینده‌ها و تأثیرپذیری غلظت آن‌ها از شرایط جوی متفاوت است. از این رو، دو آلاینده مذکور در ادامه تحلیل‌ها حذف شدند. همچنین، مونوکسید کربن به دلیل محدودیت و ناقص بودن داده‌ها، از مجموعه داده‌ها کنار گذاشته شد تا اثرات منفی ناشی از کمبود اطلاعات کاهش یابد. آلاینده‌های دی‌اکسید نیتروژن، اکسیدهای نیتروژن و مونوکسید نیتروژن همبستگی قوی‌تری با غلظت دی‌اکسید گوگرد زمینی نشان دادند. این آلاینده‌ها معمولاً از منابع مشابه صنعتی و حمل‌ونقل تولید می‌شوند و تحت تأثیر عوامل جوی و فرآیندهای شیمیایی مشابه قرار می‌گیرند. بنابراین، این آلاینده‌ها به عنوان عوامل کلیدی در مدل‌سازی و تحلیل‌های بعدی انتخاب شدند.

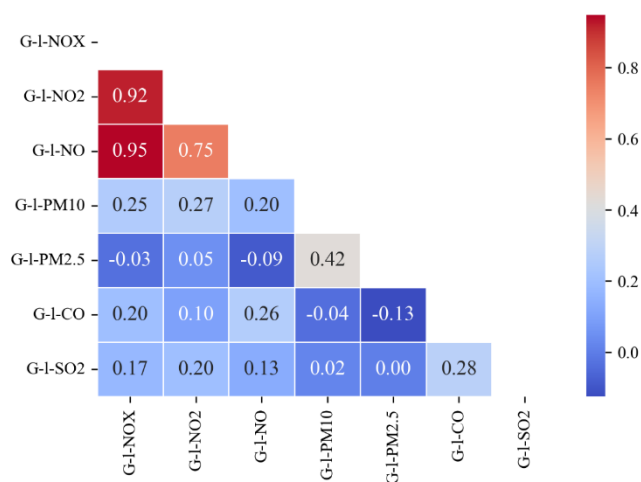
در ادامه، از آنجا که مدل‌های یادگیری ماشین به داده‌های ناقص حساس می‌باشند، برای جلوگیری از تأثیر منفی نقص داده بر دقت نهایی، تمامی روزهای با داده ناقص (روزهایی که یکی یا چند مورد از داده‌های دمای هوا، رطوبت نسبی، میزان تبخیر و تعرق، دمای تر، میزان ابرناکی، فشار هوا در ارتفاع ایستگاه، NO₂، NO_x، NO و SO₂ زمینی و ماهواره‌ای برای آن‌ها گزارش نشده است) شناسایی و از مجموعه بررسی‌ها حذف شدند. علاوه بر آن، با توجه به اینکه داده‌های ماهواره Sentinel-5P از سرویس گوگل ارث انجین تنها در ساعت ۱۳:۳۰ به وقت محلی ثبت می‌شوند، جهت هماهنگ‌سازی با داده‌های زمینی و هواشناسی، تمامی داده‌های اخیر به صورت میانگین روزانه محاسبه شدند تا امکان استفاده همزمان از داده‌های مختلف در مدل‌های یادگیری ماشین فراهم گردد. در این تحقیق از سه نوع داده شامل داده‌های هواشناسی، داده‌های زمینی و داده‌های ماهواره‌ای استفاده شد که هر کدام مقیاس و

²⁰ Adaptive Boosting¹⁹ Extreme Gradient Boosting



شکل ۳- نمودار حرارتی همبستگی بین غلظت زمینی SO₂ و پارامترهای هواشناسی.

Fig 3- Heatmap of the correlation between ground-level sulfur dioxide concentration and meteorological parameters.



شکل ۴- نمودار حرارتی همبستگی بین غلظت دی اکسید گوگرد (SO₂) و سایر آلاینده‌ها.

Fig 4- Heatmap of the correlation between sulfur dioxide (SO₂) concentration and other pollutants.

نتایج ارزیابی عملکرد مدل‌های مختلف یادگیری ماشین (جدول ۲) نشان می‌دهد که الگوریتم تقویت گرادیان پیشرفته (XGBoost) با ریشه میانگین مربعات خطا (RMSE) برابر با ۰/۵۷ ppb و ضریب تعیین (R^2) برابر با ۰/۸۹ بهترین عملکرد را در میان مدل‌های ارزیابی شده دارد. پس از آن، الگوریتم جنگل تصادفی (Random Forest) با ریشه میانگین مربعات خطا (RMSE) معادل ۰/۵۶ ppb و ضریب تعیین (R^2) برابر با ۰/۷۲ عملکرد خوبی از خود نشان داد. در حالی که الگوریتم آدا بوست (AdaBoost) با ریشه میانگین مربعات خطا (RMSE) برابر با ۰/۷۴ ppb و ضریب تعیین (R^2) معادل ۰/۵۰ کمترین دقت را در پیش‌بینی‌های خود نشان داد. این نتایج نشان می‌دهد که برای شرایط خاص این تحقیق، الگوریتم‌های تقویت گرادیان پیشرفته (XGBoost) و جنگل تصادفی (Random Forest) عملکرد مطلوبی داشتند. با این حال، ممکن است در مناطق و شرایط متفاوت، روش‌های دیگری بتوانند نتایج بهتری ارائه دهند.

با توجه به اینکه هرچه مقدار RMSE به صفر نزدیک‌تر باشد، مدل خطای کمتری دارد و همچنین هرچه مقدار R^2 به عدد یک نزدیک‌تر باشد، دقت پیش‌بینی مدل بالاتر است، الگوریتم XGBoost با کمترین میزان RMSE و بیشترین مقدار R^2 به عنوان دقیق‌ترین مدل در این پژوهش انتخاب گردید.

مجموعه‌ای از الگوریتم‌های یادگیری ماشین برای تحلیل داده‌ها استفاده شد که به ترتیب از مدل‌های ساده به پیچیده‌تر انتخاب شدند. ابتدا الگوریتم درخت تصمیم (Decision Tree) برای مدل‌سازی اولیه روابط میان متغیرها به کار گرفته شد. سپس، الگوریتم جنگل تصادفی (Random Forest) به دلیل قابلیت کاهش بیش‌برازش (overfitting) و پردازش داده‌های پیچیده‌تر مورد بهره‌برداری قرار گرفت (۲۴). در نهایت، الگوریتم درخت‌های اضافی (Extra Trees) به دلیل ساخت درخت‌های تصادفی بیشتر انتخاب گردید. برای بهبود دقت، از الگوریتم‌های تقویت (Boosting) مانند الگوریتم تقویت گرادیان (Gradient Boosting) و نسخه بهینه‌سازی شده آن الگوریتم تقویت گرادیان پیشرفته (XGBoost²¹)، استفاده شد که دقت و سرعت بالاتری دارند. همچنین، الگوریتم کت بوست (CatBoost) برای پردازش داده‌های دسته‌ای به کار رفت. همچنین، الگوریتم آدا بوست (AdaBoost²²) به عنوان الگوریتمی جهت تقویت مدل‌های ضعیف و بهبود پیش‌بینی‌ها در داده‌هایی با پراکندگی بالا، انتخاب شد. برای تقسیم‌بندی داده‌ها، ۸۰ درصد داده‌ها به عنوان داده‌های آموزش (train) و ۲۰ درصد به عنوان داده‌های آزمایش (test) انتخاب شدند. در جدول ۲، نتایج مربوط به ارزیابی دقت پیش‌بینی مدل‌های مختلف آورده شده است. مقادیر ضریب تعیین (R^2) و ریشه میانگین مربعات خطا (RMSE) برای مقایسه کارایی الگوریتم‌ها و شناسایی بهترین مدل، مورد استفاده قرار گرفت.

جدول ۲- ارزیابی عملکرد الگوریتم‌های یادگیری ماشین بکار رفته در تحقیق حاضر.

Table 2- Performance evaluation of machine learning algorithms used in current research.

الگوریتم یادگیری ماشین	ضریب تعیین (R^2)	ریشه میانگین مربعات خطا (RMSE) (ppb)
Decision Tree	۰/۵۸	۰/۶۸
Random Forest	۰/۷۲	۰/۵۶

²² Adaptive Boosting

²¹ Extreme Gradient Boosting

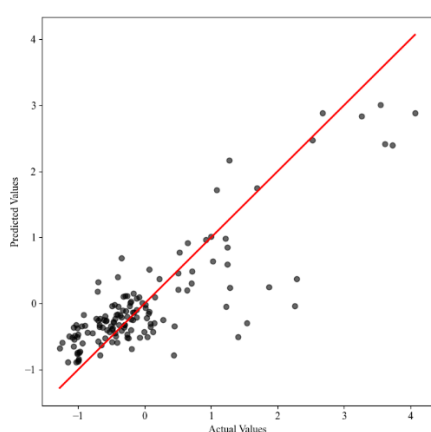
۰/۵۹	۰/۶۸	Extra Trees
۰/۶۱	۰/۶۶	Gradient Boosting
۰/۵۷	۰/۸۹	XGBoost
۰/۶۱	۰/۶۷	CatBoost
۰/۷۴	۰/۵۰	AdaBoost

و ماهواره‌ای، توانایی بالایی در تخمین دقیق غلظت SO_2 داشته و با روش‌های موفق سایر مطالعات قابل مقایسه هستند.

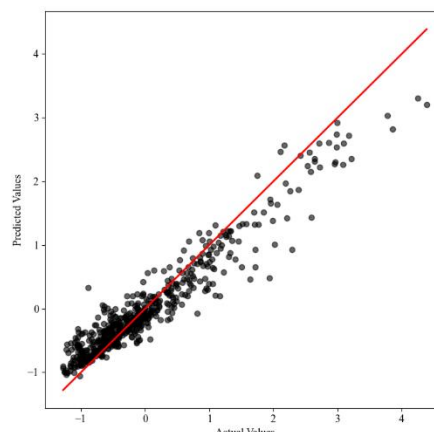
با توجه به نتایج حاصل از ارزیابی الگوریتم‌های مختلف یادگیری ماشین، الگوریتم **XGBoost** به‌عنوان بهترین روش برای پیش‌بینی غلظت SO_2 در این پژوهش انتخاب شد. به‌منظور بررسی دقت و عملکرد این مدل، نمودارهای پراکندگی مقادیر مشاهداتی و مدل‌شده SO_2 ، باقیمانده خطاها برای تحلیل نحوه توزیع و میزان خطا و هیستوگرام خطا برای مدل **XGBoost** ترسیم شده‌اند. این نمودارها به‌صورت ترکیبی، دقت و عملکرد مدل را در تحلیل داده‌ها و پیش‌بینی غلظت SO_2 به تصویر می‌کشند.

در شکل ۵-الف نمودار مربوط به داده‌های آموزش الگوریتم **XGBoost** و در شکل ۵-ب نمودار مربوط به داده‌های آزمایش همان مدل ترسیم شده است. شکل ۵ بیانگر این است که این مدل بهترین همبستگی را بین مقادیر واقعی و پیش‌بینی‌شده ارائه داده است (شکل ۵-ب). نقاط پیش‌بینی‌شده عمدتاً نزدیک به خط نیمساز قرار دارند، که این امر نشان‌دهنده عملکرد دقیق و قابل اعتماد مدل در پیش‌بینی داده‌ها است. این نتایج تأیید می‌کنند که روش **XGBoost** یک گزینه بسیار مناسب برای تحلیل و پیش‌بینی در این حوزه است.

مقایسه نتایج این مطالعه با سایر پژوهش‌ها نشان می‌دهد که روش‌های استفاده‌شده در این تحقیق از نظر دقت و ارزیابی عملکرد هم‌راستا با مطالعات پیشین هستند. برای مثال، در مطالعه (کاظمی گراجه و همکاران^{۱۱}، ۲۰۲۳)، استفاده از الگوریتم جنگل تصادفی (**Random Forest**) منجر به دستیابی به ضریب تعیین (R^2) برابر با ۰/۸۷ و ریشه میانگین مربعات خطا (**RMSE**) برابر با ۷/۲ ppb شد، که به نتایج مشابه در این مطالعه نزدیک است (۱۸). همچنین، در پژوهش (آنگرانی و همکاران^{۱۴}، ۲۰۲۴)، الگوریتم تقویت گرادیان (**Gradient Boosting**) با ضریب تعیین (R^2) برابر با ۰/۹۱ و ریشه میانگین مربعات خطا (**RMSE**) معادل ۶/۹ ppb عملکرد خوبی را نشان داد که تطابق آن با نتایج این تحقیق، کارایی بالای این مدل‌ها در تخمین غلظت دی‌اکسید گوگرد را تأیید می‌کند (۳). نتایج پژوهش (بلانکو و همکاران^{۱۳}، ۲۰۲۴) با استفاده از الگوریتم‌های تقویت گرادیان پیشرفته (**XGBoost**) و جنگل تصادفی (**Random Forest**) دقت مشابهی را با ضریب تعیین (R^2) حدود ۰/۸۸ و ریشه میانگین مربعات خطا (**RMSE**) برابر با ۷/۹ ppb ارائه کرده است (۴). این تطابق‌ها نشان می‌دهد که مدل‌های منتخب این مطالعه، در کنار داده‌های ترکیبی زمینی



(ب)



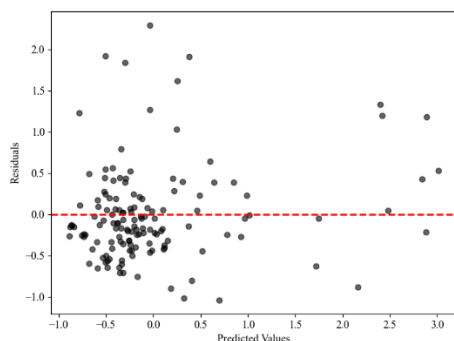
(الف)

شکل ۵- نمودار پراکندگی مقادیر مشاهده‌ای و مدل شده SO_2 با الگوریتم XGBoost: (الف) داده‌های آموزشی؛ و (ب) داده‌های آزمایش.

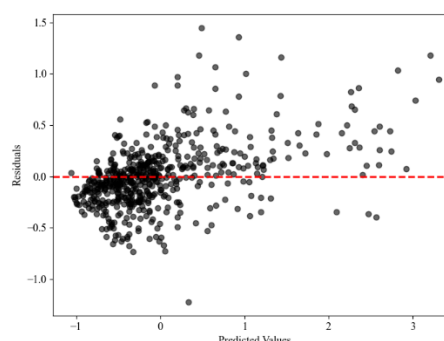
Fig 5- Scatter plot of observed and modeled SO_2 values using the XGBoost algorithm: (a) Train data; and (b) Test data.

در شکل ۷، هیستوگرام توزیع خطای مدل XGBoost برای داده‌های آموزشی (شکل ۷-الف) و آزمایشی (شکل ۷-ب) نشان داده شده است. توزیع خطاها در هر دو نمودار نزدیک به صفر و متقارن است، که بیانگر عملکرد مناسب مدل در پیش‌بینی مقادیر واقعی است. در داده‌های آموزشی (شکل ۷-الف)، مدل توانسته است الگوها را به خوبی یاد بگیرد و خطاهای بزرگی مشاهده نمی‌شود. همچنین در داده‌های آزمایشی (شکل ۷-ب) نیز توزیع مشابهی دیده می‌شود؛ که نشان می‌دهد مدل بر روی داده‌های جدید عملکرد قابل قبولی داشته است. بنابراین، مدل XGBoost به عنوان یک روش پیش‌بینی برای غلظت SO_2 ، عملکرد موفقیت‌آمیزی را ارائه داده است.

شکل ۶ نمودارهای باقی‌مانده خطا، (شکل ۶-الف نمودار مربوط به داده‌های آموزش و شکل ۶-ب نمودار مربوط به داده‌های آزمایش) برای مدل XGBoost را نشان می‌دهند. پراکندگی تصادفی نقاط اطراف محور صفر افقی در شکل ۶-ب، بیانگر این است که هیچ الگوی خاصی در باقی‌مانده‌ها وجود ندارد. این امر نشان‌دهنده این است که فرض همگنی واریانس رعایت شده است. همچنین، در نمودارها هیچ‌گونه الگو یا انحراف خاصی مانند مخروطی شدن یا انحنای ویژه مشاهده نمی‌شود. با وجود تعدادی نقطه پرت که فاصله بیشتری از خط صفر دارند، پراکندگی کلی نقاط به صورت تصادفی است و نشان‌دهنده عملکرد مناسب مدل در پیش‌بینی مقادیر واقعی است.



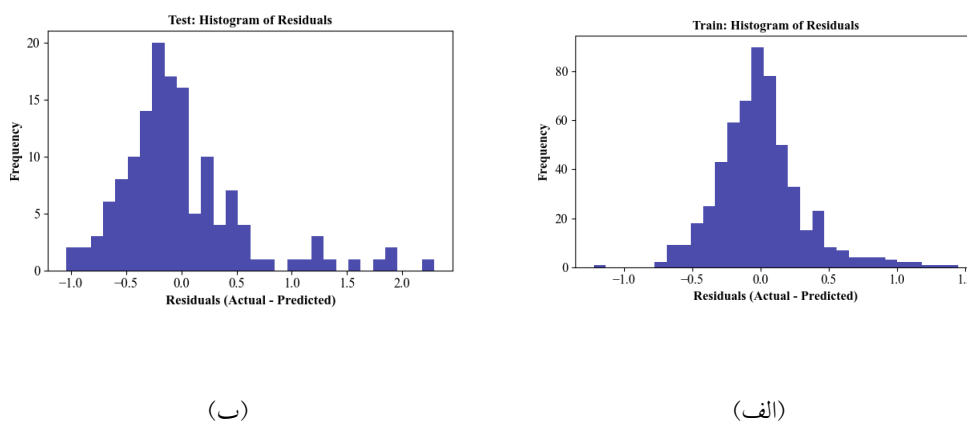
(ب)



(الف)

شکل ۶- نمودار باقیمانده خطاهای مدل XGBoost: (الف) داده‌های آموزشی؛ و (ب) داده‌های آزمایش.

Fig 6- Residual error plot of the XGBoost model: (a) Train data; and (b) Test data.



شکل ۷- نمودار هیستوگرام توزیع خطای مدل XGBoost: (الف) داده‌های آموزشی؛ و (ب) داده‌های آزمایش.

Fig 7- Histogram of the error distribution of the XGBoost model: (a) Train data; and (b) Test data.

نتیجه‌گیری

مقاله حاضر با تمرکز بر پایش دی‌اکسید گوگرد (SO_2) در هوای آزاد شهر ارومیه، ترکیبی از داده‌های برگرفته از ماهواره سنتینل ۵-پی (Sentinel-5P) و روش‌های یادگیری ماشین را برای تحلیل دقیق این آلاینده ارائه کرده است. نتایج نشان می‌دهد که داده‌های ماهواره‌ای، به‌ویژه با وضوح بالای سنجنده تروپومی (TROPOMI)، به‌عنوان منبعی ارزشمند و قابل‌اعتماد برای پایش کیفیت هوا در مناطقی که داده‌های زمینی محدود هستند، عمل می‌کند. این ترکیب داده‌ای، علاوه بر کاهش وابستگی به

به‌منظور بررسی نقش پارامترهای هواشناسی در بهبود عملکرد مدل‌های یادگیری ماشین، تحلیل‌ها یک بار دیگر و این بار بدون در نظر گرفتن پارامترهای کمکی، انجام شد (جدول ۳). همان‌گونه که در جدول ۳ مشاهده می‌شود، تمامی مدل‌ها کاهش عملکرد قابل‌توجهی را تجربه کرده‌اند. برای مثال، مدل XGBoost که در تحلیل‌های اولیه (با استفاده از پارامترهای هواشناسی) بهترین عملکرد را ارائه داده بود، در شرایط بدون این پارامترها به مقادیر $R^2 = 0.38$ و $RMSE = 0.75$ ppb رسید. این کاهش قابل‌توجه دقت، اهمیت استفاده از اطلاعات جوی در مدل‌سازی را تأیید می‌کند.

جدول ۳- ارزیابی عملکرد الگوریتم‌های یادگیری ماشین بدون اعمال پارامترهای هواشناسی.

Table 3- Performance evaluation of machine learning algorithms without applying meteorological parameters.

خطا	میانگین مربعات	ضریب تعیین (R^2)	الگوریتم یادگیری ماشین
	(ppb) (RMSE)		
	۰/۸۱	۰/۲۷	Decision Tree
	۰/۷۷	۰/۳۵	Random Forest

۰/۷۹	۰/۳۲	Extra Trees
۰/۷۷	۰/۳۵	Gradient Boosting
۰/۷۵	۰/۳۸	XGBoost
۰/۷۹	۰/۳۳	CatBoost
۰/۸۶	۰/۱۸	AdaBoost

هستند، بلکه می‌توانند به کاهش اثرات منفی آلودگی بر سلامت عمومی و ارتقای کیفیت زندگی در مناطق شهری منجر شوند.

در تحقیق‌های آتی، بررسی تأثیر خودروهای گازوئیلی و دیزلی و همچنین مصرف گازوئیل در پمپ‌بنزین‌ها می‌تواند به شناسایی منابع آلاینده مهم در مناطق شهری کمک کند. همچنین، توجه به شرایط ترافیکی و الگوهای حمل‌ونقل می‌تواند دقت مدل‌های پیش‌بینی آلودگی را افزایش دهد. استفاده از داده‌های بیشتر، از جمله اطلاعات ترافیکی و مصرف انرژی، می‌تواند استراتژی‌های مؤثرتری برای کاهش آلودگی هوا و بهبود کیفیت زندگی شهری ارائه دهد.

تقدیر و تشکر

نویسندگان مقاله مراتب تقدیر و تشکر خود را از معاونت محترم نظارت و پایش اداره کل حفاظت محیط‌زیست استان آذربایجان غربی، به دلیل در اختیار گذاشتن داده‌های پایش زمینی ایستگاه سنجش آلاینده‌های هوای شماره ۳ ارومیه، ابراز می‌دارند.

ایستگاه‌های زمینی، شکاف‌های اطلاعاتی موجود را پوشش داده و امکان تحلیل جامع‌تری از روند تغییرات آلودگی هوا را فراهم می‌سازد. استفاده از الگوریتم‌های یادگیری ماشین در این تحقیق، ارتباطات پیچیده میان داده‌های سطح زمین و ماهواره‌ای را آشکار ساخته و اثر عوامل جوی بر پیش‌بینی غلظت SO₂ را با دقت بیشتری بررسی کرده است. این رویکرد نشان می‌دهد که ترکیب فناوری‌های نوین با داده‌های محیطی می‌تواند به شناسایی دقیق‌تر آلاینده‌ها و مدیریت بهتر کیفیت هوا کمک کند. یافته‌های این پژوهش نشان داد که استفاده از پارامترهای هواشناسی تأثیر قابل توجهی بر بهبود عملکرد مدل‌های یادگیری ماشین در تحلیل و برآورد SO₂ دارد. حذف این پارامترها منجر به افت شدید دقت مدل‌ها، به‌ویژه در شاخص‌های R² و RMSE شد که اهمیت داده‌های جوی را در مدل‌سازی تأیید می‌کند. این نتایج تأکید می‌کند که در کنار استفاده از داده‌های زمینی و ماهواره‌ای، ادغام متغیرهای جوی می‌تواند به ارتقای دقت و قابلیت اطمینان مدل‌ها کمک کند. پژوهش حاضر می‌تواند به‌عنوان الگویی برای سایر شهرهای ایران و کشورهای در حال توسعه مطرح شود. نتایج به‌دست آمده نه تنها در بهبود برنامه‌ریزی‌های زیست‌محیطی مؤثر

منابع مورد استفاده

1. Ahmady-Birgani, H., Ravan, P., Yao, Z., & Afrasinei, G. M. (2022, May). Newly-Born Sand Dunes of Lake Urmia: Assessing Migration Rate and Morphodynamic Changes Using Remote Sensing

Techniques and Field Studies. In EGU General Assembly Conference Abstracts (pp. EGU22-12041). <https://doi.org/10.5194/egusphere-egu22-12041>

2. Aneja, V. P., Agarwal, A., Roelle, P. A., Phillips, S. B., Tong, Q., Watkins, N., & Yablonsky, R. (2001). Measurements and analysis of criteria pollutants in New Delhi, India. *Environment International*, 27(1), 35-42. [https://doi.org/10.1016/S0160-4120\(01\)00051-4](https://doi.org/10.1016/S0160-4120(01)00051-4)
3. Anggraini, T. S., Irie, H., Sakti, A. D., & Wikantika, K. (2024). Machine learning-based global air quality index development using remote sensing and ground-based stations. *Environmental Advances*, 15, 100456. <https://doi.org/10.1016/j.envadv.2023.100456>
4. Blanco, G., Barco, L., Innocenti, L., & Rossi, C. (2024). Urban Air Pollution Forecasting: a Machine Learning Approach leveraging Satellite Observations and Meteorological Forecasts. *arXiv preprint arXiv:2405.19901*. <https://doi.org/10.48550/arXiv.2405.19901>
5. Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Harmouch, H., & Naumann, F. (2022). The effects of data quality on machine learning performance. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2207.14529>
6. Chen, X., Zhang, L. W., Huang, J. J., Song, F. J., Zhang, L. P., Qian, Z. M., ... & Tang, N. J. (2016). Long-term exposure to urban air pollution and lung cancer mortality: A 12-year cohort study in Northern China. *Science of the Total Environment*, 571, 855-861. [10.1016/j.scitotenv.2016.07.064](https://doi.org/10.1016/j.scitotenv.2016.07.064)
7. Cofano, A., Cigna, F., Santamaria Amato, L., Siciliani de Cumis, M., & Tapete, D. (2021). Exploiting Sentinel-5P TROPOMI and ground sensor data for the detection of volcanic SO₂ plumes and activity in 2018–2021 at Stromboli, Italy. *Sensors*, 21(21), 6991. <https://doi.org/10.3390/s21216991>
8. Dastaran M, Jafari Sh, Moslemi H, Attarchi, S, Alavipanah, S. K. 2022. Monitoring Bakhtegan wetland using a time series of satellite data on the Google Earth Engine platform and predicting parameters with Facebook's Prophet model. *Journal of RS and GIS for Natural Resources*, 13(4): 1-20. <https://doi.org/10.30495/GIRS.2022.685454>. (In Persian).
9. Dubovik, O., Schuster, G. L., Xu, F., Hu, Y., Bösch, H., Landgraf, J., & Li, Z. (2021). Grand challenges in satellite remote sensing *Frontiers in Remote Sensing*, 2, 619818. doi: 10.3389/frsen.2021.619818
10. Fania, A., Monaco, A., Pantaleo, E., Maggipinto, T., Bellantuono, L., Cilli, R., ... & Bellotti, R. (2024). Estimation of Daily Ground Level Air Pollution in Italian Municipalities with Machine Learning Models Using Sentinel-5P and ERA5 Data. *Remote Sensing*, 16(7), 1206. <https://doi.org/10.3390/rs16071206>
11. Ghannadi, M. A., Shahri, M., & Moradi, A. (2022). Air pollution monitoring using Sentinel-5 (Case study: big industrial cities of Iran). *Environmental Sciences*, 20(2), 81-98. <http://dx.doi.org/10.52547/envs.2022.1026>. (In Persian).
12. Ghozikali, M. G., Heibati, B., Naddafi, K., Kloog, I., Conti, G. O., Polosa, R., & Ferrante, M. (2016). Evaluation of chronic obstructive pulmonary disease (COPD) attributed to atmospheric O₃, NO₂, and SO₂ using Air Q Model (2011–2012 year). *Environmental research*, 144, 99-105. <https://doi.org/10.1016/j.envres.2015.10.030>
13. Halder, B., Ahmadianfar, I., Heddarn, S., Mussa, Z. H., Goliatt, L., Tan, M. L., ... & Yaseen, Z. M. (2023). Machine learning-based country-level annual air pollutants exploration using Sentinel-5P and Google Earth Engine. *Scientific Reports*, 13(1), 7968. <https://doi.org/10.1038/s41598-023-34774-9>
14. Hassanzadeh, N., Hassanzadeh, R., Hosseinjanizadeh, M., & Honarmand, M. (2023). Evaluation of SO₂ and NO₂ concentration in the atmosphere of industrial areas using RS and GIS: Case study of Khatoonabad Copper Smelting Factory, Kerman, Iran. *Scientific-Research Quarterly of Geographical Data (SEPEHR)*, 32(125), 7-27. <https://doi.org/10.22131/SEPEHR.2023.559597.2896>. (In Persian).
15. Jion, M. M. M. F., Jannat, J. N., Mia, M. Y., Ali, M. A., Islam, M. S., Ibrahim, S. M., ... & Islam, A. R. M. T. (2023). A critical review and prospect of NO₂ and SO₂ pollution over Asia: Hotspots, trends, and

- sources. *Science of The Total Environment*, 876, 162851.
<https://doi.org/10.1016/j.scitotenv.2023.162851>
16. Kamarehie, B., Ghaderpoori, M., Jafari, A., Karami, M., Mohammadi, A., Azarshab, K., ... & Noorizadeh, N. (2017). Quantification of health effects related to SO₂ and NO₂ pollutants by using air quality model. *Journal of Advances in Environmental Health Research*, 5(1), 44-50.
<https://doi.org/10.22102/jaehr.2017.47757>
17. Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental pollution*, 151(2), 362-367. doi:10.1016/j.envpol.2007.06.012
18. Kazemi Garajeh, M., Laneve, G., Rezaei, H., Sadeghnejad, M., Mohamadzadeh, N., & Salmani, B. (2023). Monitoring trends of CO, NO₂, SO₂, and O₃ pollutants using time-series sentinel-5 images based on google earth engine. *Pollutants*, 3(2), 255-279.
<https://doi.org/10.3390/pollutants3020019>
19. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., & Bussonnier, M. (2016) Jupyter Notebooks as a Publishing Format for Reproducible Research. In *Proceedings of the 20th International Conference on Electronic Publishing* (pp. 87-90). doi:10.3233/978-1-61499-649-1-87
20. Krotkov, N. A., McLinden, C. A., Li, C., Lamsal, L. N., Celarier, E. A., Marchenko, S. V., ... & Streets, D. G. (2016). Aura OMI observations of regional SO₂ and NO₂ pollution changes from 2005 to 2015. *Atmospheric Chemistry and Physics*, 16(7), 4605-4629. doi:10.5194/acp-16-4605-2016
21. Lelieveld, J., Pozzer, A., Pöschl, U., Fnais, M., Haines, A., & Münzel, T. (2020). Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovascular research*, 116(11), 1910-1917.
<https://doi.org/10.1093/cvr/cvaa073>
22. Li, C., McLinden, C., Fioletov, V., Krotkov, N., Carn, S., Joiner, J., ... & Dickerson, R. R. (2017). India is overtaking China as the world's largest emitter of anthropogenic sulfur dioxide. *Scientific reports*, 7(1), 14304. <https://doi.org/10.1038/s41598-017-14639-8>
23. Lu, Z., Streets, D. G., De Foy, B., & Krotkov, N. A. (2013). Ozone Monitoring Instrument observations of interannual increases in SO₂ emissions from Indian coal-fired power plants during 2005–2012. *Environmental science & technology*, 47(24), 13993-14000. [dx.doi.org/10.1021/es4039648](https://doi.org/10.1021/es4039648)
24. Lu, Z., Streets, D. G., Zhang, Q., Wang, S., Carmichael, G. R., Cheng, Y. F., ... & Tan, Q. (2010). Sulfur dioxide emissions in China and sulfur trends in East Asia since 2000. *Atmospheric chemistry and physics*, 10(13), 6311-6331. doi:10.5194/acp-10-6311-2010
25. Mohammadian M, Morovati M, Omidipour R. Comparison of the effectiveness of machine learning methods in modeling fire-prone areas (Ilam Province, Darehshahr City). *Journal of RS and GIS for Natural Resources*, 15(1): 48-73. DOI: 10.30495/girs.2022.697864. (In Persian).
26. Mumtaz, R., Amin, A., Khan, M. A., Asif, M. D. A., Anwar, Z., & Bashir, M. J. (2023). Impact of Green Energy Transportation Systems on Urban Air Quality: A Predictive Analysis Using Spatiotemporal Deep Learning Techniques. *Energies*, 16(16), 6087. <https://doi.org/10.3390/en16166087>
27. Nouri, A., Ghanbarzadeh Lak, M., & Valizadeh, M. (2021). Prediction of PM_{2.5} concentrations using principal component analysis and artificial neural network techniques: A case study: Urmia, Iran. *Environmental Engineering Science*, 38(2), 89-98. <https://doi.org/10.1089/ees.2020.0089>
28. Oğuz, K., & Pekin, M. A. (2024). Prediction of Air Pollution with Machine Learning Algorithms. *Turkish Journal of Science and Technology*, 19(1), 1-12. <https://doi.org/10.55525/tjst.1224661>
29. Qu, Y., An, J., He, Y., & Zheng, J. (2016). An overview of emissions of SO₂ and NO_x and the long-range transport of oxidized sulfur and nitrogen pollutants in East Asia. *Journal of Environmental Sciences*, 44, 13-25. <https://doi.org/10.1016/j.jes.2015.08.028>
30. Rahman, M. S., Azad, M. A. K., Hasanuzzaman, M., Salam, R., Islam, A. R. M. T., Rahman, M. M., & Hoque, M. M. M. (2021). How air quality and COVID-19 transmission change under different

- lockdown scenarios? A case from Dhaka city, Bangladesh. *Science of The Total Environment*, 762, 143161.
<https://doi.org/10.1016/j.scitotenv.2020.143161>
31. Saieedifar Z, Khosroshahi M, Gohardust A, Ebrahimikhusfi Z, Lotfinasabasl S, Dargahian F. 2020. Investigation of the origin and spatial distribution of high dust concentrations and its synoptical analysis in Gavkhooni basin. *Journal of RS and GIS for Natural Resources*, 11(4): 47-64. <http://dorl.net/dor/20.1001.1.26767082.1399.11.4.3.3>. (In Persian).
32. Sameh, S., Zarzoura, F., & El-Mewafi, M. (2024). Spatio-temporal Analysis Mapping of Air Quality Monitoring in Cairo using Sentinel-5 satellite data and Google earth engine. *Mansoura Engineering Journal*, 49(1), 3. <https://doi.org/10.58491/2735-4202.3122>
33. Serbula, S. M., Milosavljevic, J. S., Kalinovic, J. V., Kalinovic, T. S., Radojevic, A. A., Trujic, T. L. A., & Tasic, V. M. (2021). Arsenic and SO₂ hotspot in South-Eastern Europe: An overview of the air quality after the implementation of the flash smelting technology for copper production. *Science of the Total Environment*, 777, 14598. <https://doi.org/10.1016/j.scitotenv.2021.145981>
34. Soleimany, A., Grubliauskas, R., & Šerevičienė, V. (2021). Application of satellite data and GIS services for studying air pollutants in Lithuania (case study: Kaunas city). *Air Qual Atmos Health* 14: 411–429. <https://doi.org/10.1007/s11869-020-00946-z>
35. Talebifard, R., Shafiee, A., & Yeganeh, A. (2021). Spatial evaluation of the effect of land use parameters on air quality in urban areas, a case study: Tehran. *Journal of Environmental Science Studies*, 6(3), 3890-3899. (In Persian).
36. Tavella, R. A., das Neves, D. F., Silveira, G. D. O., Vieira de Azevedo, G. M. G., Brum, R. D. L., Bonifácio, A. D. S., ... & da Silva Júnior, F. M. R. (2025). The Relationship Between Surface Meteorological Variables and Air Pollutants in Simulated Temperature Increase Scenarios in a Medium-Sized Industrial City. *Atmosphere*, 16(4), 363. <https://doi.org/10.3390/atmos16040363>
37. World Health Organization. (2021). WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide, 1302. <https://www.who.int/publications/i/item/9789240034228>
38. Wieczorek, B. (2023). Air Pollution Patterns Mapping of SO₂, NO₂, and CO Derived from TROPOMI over Central-East Europe. *Remote Sensing*, 15(6), 1565. <https://doi.org/10.3390/rs15061565>
39. Zhao, Y., Wang, S., Duan, L., Lei, Y., Cao, P., & Hao, J. (2008). Primary air pollutant emissions of coal-fired power plants in China: Current status and future prediction. *Atmospheric Environment*, 42(36), 8442-8452. <https://doi.org/10.1016/j.atmosenv.2008.08.021>
40. Zhang, Z., Wang, J., Chen, L., Chen, X., Sun, G., Zhong, N., ... & Lu, W. (2014). Impact of haze and air pollution-related hazards on hospital admissions in Guangzhou, China. *Environmental Science and Pollution Research*, 21, 4236-4244. <https://doi.org/10.1007/s11356-013-2374-6>



Comparative Analysis of Sulfur Dioxide (SO₂) Concentration with Satellite Data Using Machine Learning Algorithms - Case Study: Urmia City

Samin Rahimzadeh¹, Mehdi Ghanbarzadeh Lak^{2*}, Arezoo Soleimany³

Received: 2024-12-30 / Accepted: 2025-05-05 / Published: 2025-12-27

Abstract

Air pollution, as one of the main environmental challenges, has significant adverse effects on public health and quality of life. Sulfur dioxide (SO₂), one of the criteria pollutants of ambient air quality, not only causes harmful impacts on human health but also contributes to global phenomena such as acid rain. This study aims to monitor and analyze SO₂ concentrations in Urmia city using data from the TROPOMI sensor onboard the Sentinel-5P satellite, combined with ground-level data and machine learning algorithms. The research employed advanced machine learning techniques to develop models. Data were collected during the period 2019 to 2023, preprocessed, and analyzed using performance indicators such as the coefficient of

Samin Rahimzadeh¹, Mehdi Ghanbarzadeh Lak² (✉), Arezoo Soleimany³

1. MSc. Student of Civil and Environmental Engineering, School of Engineering, Civil Engineering Department, Urmia University, Urmia, Iran

2. Associate Professor, School of Engineering, Civil Engineering Department, Urmia University, Urmia, Iran

3. PhD Graduate in Environmental Sciences - Environmental Pollution, Faculty of Natural Resources and Environmental Sciences, Malayer University, Malayer, Iran

DOI: 10.82447/girs.2025.1194923

e-mail: m.ghanbarzadehlak@urmia.ac.ir

determination (R²) and root mean square error (RMSE). The results revealed that integrating ground-based and satellite data, alongside meteorological parameters such as air temperature, relative humidity, evapotranspiration, wet-bulb temperature, cloud cover, and atmospheric pressure at station height, significantly improved the performance of the models. The XGBoost algorithm, with R² of 0.89 and RMSE of 0.57 ppb, was identified as the best-performing model. Analyses excluding meteorological parameters showed a substantial decrease in model accuracy (i.e., the XGBoost model performance without meteorological parameters demonstrated a significant degradation, with R² dropping to 0.38 and RMSE increasing to 0.75 ppb), highlighting the critical importance of these variables. Satellite data demonstrated their potential to play a key role in bridging information gaps caused by the limitations of ground-based stations. Moreover, combining modern technologies with environmental data can contribute to more accurate pollutant identification, improve air quality, and serve as a model for other cities in Iran and developing countries.

Extended Abstract

Introduction: Air pollution is considered one of the most significant environmental challenges

worldwide, with far-reaching impacts on public health, ecosystems, and the economy. This issue is particularly pronounced in urban areas with high population density or industrial activities. Air pollution leads to respiratory problems, cardiovascular diseases, and various disorders in the functioning of vital human body systems. Among the various pollutants, sulfur dioxide (SO₂) is recognized as one of the primary pollutants due to its prominent role in causing respiratory diseases, producing acid rain, and negatively affecting air quality. Additionally, this pollutant can impact soil quality and water resources, disrupting natural ecosystem cycles. This study aims to accurately investigate SO₂ concentrations in Urmia city using data from the Sentinel-5P satellite, integrated with advanced machine learning methods.

Problem Statement: In many developing regions, such as Urmia, there are significant limitations in providing accurate and continuous air quality data. The lack of ground-based monitoring stations and deficiencies in recording essential data are major challenges in analyzing pollution trends and identifying influencing factors. These limitations create information gaps in air quality management and in developing practical pollution reduction programs. In such circumstances, adopting innovative methods to address these deficiencies is essential. Satellite data, particularly TROPOMI data with high temporal and spatial resolution, can serve as an effective solution. These data offer comprehensive information on the spatial distribution of pollutants, enabling more robust analysis of emission patterns and temporal variations. Additionally, correlation analysis between satellite and ground-based data enhances understanding of spatial and temporal pollutant distributions. This study combines these data sources with machine learning algorithms to propose practical and scientifically sound solutions for improved air quality management.

Objectives: This study aims to utilize satellite data in combination with ground-based data to achieve a

more accurate analysis of SO₂ concentrations and identify the meteorological factors influencing them. One primary objective is to provide a comprehensive and reliable model for managing SO₂ concentrations in urban areas with similar limitations as Urmia. Additionally, using advanced machine learning algorithms, this research seeks to uncover complex relationships between meteorological variables and pollutant concentrations. Specifically, it examines the impact of air temperature, relative humidity, evapotranspiration, wet-bulb temperature, cloud cover, and atmospheric pressure at station height. Another goal is to propose strategies for improving air quality management using advanced data analysis techniques.

Methodology: Satellite-based SO₂ data from Sentinel-5P were extracted using Google Earth Engine (GEE). These were integrated with ground-based monitoring information from air quality stations (SO₂ and other pollutants), and meteorological data as described above. Data preprocessing included outlier removal using Z-score (data with $Z > 3$ or $Z < -3$ were excluded). Several machine learning algorithms, including Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, CatBoost, and AdaBoost, were employed. 80% of data were allocated for training and 20% for testing. Model performance was evaluated using R^2 and RMSE. To assess meteorological influence, correlation coefficients between each input and SO₂ concentration were analyzed. Model performances were then compared, and recommendations provided.

Results and Discussion: Integrating satellite and ground-based data significantly enhanced the accuracy of SO₂ concentration modeling. Advanced machine learning algorithms identified complex variable relationships. Excluding meteorological parameters caused substantial accuracy reductions, demonstrating their essential role. Satellite data proved effective as a complementary or alternative

source in regions lacking monitoring stations. High-resolution data, particularly during critical times and locations, improved model accuracy. The modeling results revealed that the influence of meteorological variables on SO₂ concentration varies with time and conditions, indicating the need for further detailed analyses in future studies.

Conclusion: Integrating satellite and ground-based data with advanced machine learning algorithms provides a powerful tool for air quality analysis and management. This approach can serve as a model for other Iranian cities and developing countries. Findings contribute to environmental planning, mitigate pollution effects on public health, and improve urban quality of life. In areas with limited ground data access, this method delivers

comprehensive and accurate information. Using satellite data with meteorological variables uncovers new dimensions in air quality analysis. Specifically, this approach is suitable for long-term pollution trend analysis and designing pollution mitigation policies. Emphasizing high-resolution data and advanced techniques underscores the capacity of these methods to address complex environmental issues.

Keywords: Ambient air pollution, SO₂, Sentinel-5P, Machine Learning Algorithms, Urmia, Meteorological Parameters.

Please cite this article as: Rahimzadeh· Samin, Ghanbarzadeh Lak· Mehdi, Soleimany, Arezoo. (2025) **Comparative Analysis of Sulfur Dioxide (SO₂) Concentration with Satellite Data Using Machine Learning Algorithms Case Study: Urmia City.** *Journal of RS and GIS for Natural Resources*, 16(4): 43-67.