

تخمین مکانی و زمانی غلظت $PM_{2.5}$ با استفاده از داده‌های عمق اپتیکی هواویز و هواشناسی بر مبنای الگوریتم‌های یادگیری ماشین و بهینه‌سازی (مطالعه موردی: شهر تهران)

امین غلامی¹، یاسر ابراهیمیان قاجاری^{2*}

¹ دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، دانشکدگان فنی، دانشگاه تهران، تهران، ایران
amingholami328@gmail.com

² دانشکده مهندسی عمران، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران
y.ebrahimian@nit.ac.ir

(تاریخ دریافت: مرداد 1401، تاریخ تصویب: آبان 1401)

چکیده

امروزه با قرارگیری ایستگاه‌های زمینی سنجش آلودگی هوای غلظت $PM_{2.5}$ در مکان‌های مختلفی از شهر، میزان غلظت با دقت بالایی اندازه‌گیری می‌شوند، اما ایراد اصلی که بر این ایستگاه‌های زمینی گرفته می‌شود، محدود بودن این ایستگاه‌ها و پراکندگی محدود آن‌هاست. در نتیجه، امکان برآورد غلظت در یک منطقه وسیع را در اختیار نمی‌گذارد؛ از این رو، در این تحقیق از داده‌های MODIS نظیر عمق اپتیکی هواویز اخذشده از تصاویر ماهواره‌ای NASA و همچنین، از داده‌های هواشناسی در سال‌های 1396 تا 1399 استفاده شده‌است. در مطالعه حاضر، ما ابتدا از الگوریتم‌های یادگیری ماشین نظیر درخت تصمیم‌گیری، رگرسیون خطی چندگانه و پرسپترون چند لایه با استفاده از آموزش داده‌های AOD و هواشناسی استفاده کردیم. نتایج نشان داد که برای این داده‌ها خطای روش‌ها به صورت $MLP < DC < MLR$ بوده؛ در نتیجه بهترین دقت، روش MLP بوده‌است ($RMSE = 11.46$ و $R^2 = 0.67$). همچنین، برای بهبود دقت از الگوریتم‌های بهینه‌سازی نظیر روش بهینه‌سازی ازدحام ذرات، الگوریتم ژنتیک و الگوریتم ژنتیک مبتنی بر مهاجرت استفاده شده‌است، الگوریتم‌های GA و الگوریتم MBGA دقت مطلوبی دادند که الگوریتم MBGA توانست خطا را بهینه کند، مدل برآورد $PM_{2.5}$ با استفاده از روش MBGA+MLP توانست غلظت $PM_{2.5}$ را با دقت بالا ($RMSE = 1.71$ و $R^2 = 0.99$) برآورد کند. این تحقیق یک الگوریتم پیشنهادی (MBGAMLP) برای تخمین غلظت ارائه می‌دهد که سازمان‌های دولتی می‌توانند در اجرای سیاست‌های حفاظت از محیط زیست استفاده کنند.

واژگان کلیدی: آلودگی هوا، تهران، الگوریتم ژنتیک مبتنی بر مهاجرت، AOD، $PM_{2.5}$

1- مقدمه

بسیاری از کلان شهرها با مشکل آلودگی هوا مواجه هستند که این آلودگی، بخشی بخاطر منابع متحرک مانند وسایل نقلیه و بخشی بعلا وجود منابع ثابت مانند صنایع، به وجود آمده است [1]. تأثیرات تدریجی و طولانی مدت آلودگی هوا سبب شده تا مسئولان و مردم کمتر به آن توجه کنند این در حالی است که روند صعودی آمارهای مرگ و میر، سرطان‌ها و سکت‌های قلبی ناشی از آلودگی هوا، نشانگر این است که آلودگی هوا سبب مرگ تدریجی انسان‌ها شده است [2]

در سال‌های اخیر محققان به این نتیجه رسیده‌اند که آلودگی هوا نه تنها باعث مشکلات قلبی نظیر حمله قلبی و سکت می‌شود، بلکه در دراز مدت باعث تکرار این مشکلات خواهد شد. بیماران قلبی ساکن در محیط‌هایی با آلودگی هوای بیشتر، 44 درصد بیشتر احتمال وقوع دومین حمله قلبی را دارند. بر اساس مطالعات سازمان بهداشت جهانی، ذرات معلق آلاینده هوا سیزدهمین علت منجر به مرگ در سرتاسر جهان و مسئول 800000 مرگ در سال می‌باشند [3].

قرار گرفتن افراد در معرض PM_{2.5} (ذرات با قطر آیرودینامیکی کمتر از 2.5 میکرومتر) باعث بروز مشکلات قلبی، عروقی و تنفسی می‌باشد [4]–[7]. منابع انتشار این ذرات، انسانی و طبیعی می‌باشد. اگرچه بخش کوچکی از ذرات موجود در جو زمین را تشکیل می‌دهد، این ذرات کیفیت هوا را کاهش می‌دهد و باعث آلودگی هوا می‌شوند و واکنش‌های فتوشیمیایی جوی را افزایش می‌دهد [8], [9]. به هر حال در رابطه با اندازه‌گیری میزان غلظت آلودگی هوا محدودیت‌هایی وجود دارند که از جمله محدودیت‌های مهم آن، تعداد و توزیع مکانی ایستگاه‌های سنجش آلودگی هوا است. به دلیل هزینه زیاد و کمبود امکانات لازم، نمی‌توان کل سطح شهر را با ایستگاه‌های سنجش آلودگی هوا پوشش داد؛ لذا از الگوریتم‌های درونیابی و پیش‌بینی جهت مدل‌سازی آلودگی هوا استفاده می‌شود. دقت این مدل‌ها از اهمیت ویژه‌ای برخوردار است و تحقیقات مختلفی جهت بهبود کیفیت مدل‌سازی آلودگی هوا انجام شده است [2], [10], [11].

در واقع، در تمامی کاربردهای مهندسی، مسئله دقت و زمان، از اهمیت بالایی برخوردار است. الگوریتمی که بتواند دقت بالاتری را ارائه دهد، ارجح می‌باشد. زمان اجرا نیز هر چقدر کمتر باشد، الگوریتم قابلیت تعمیم‌پذیری بالاتری دارد [12]؛ از این رو در چند سال اخیر، برای مدل‌سازی مکانی- زمانی آلودگی هوا از الگوریتم‌های یادگیری ماشین از جمله شبکه‌های عصبی مصنوعی و تکنیک‌های یادگیری عمیق استفاده شده است.

در راستای پیش‌بینی آلودگی هوا تحقیقات متعددی صورت گرفته است. یک برنامه کاربردی با هدف ساده‌سازی فرآیند مدل‌سازی رگرسیون کاربری اراضی¹ را برای مطالعات زیست‌محیطی پیشنهاد داد. با استفاده از قابلیت‌های آماری و GIS با زبان برنامه‌نویسی R ساخته شده است. در این نرم‌افزار از مجموعه داده‌های آزمایشی از 48 غلظت اندازه‌گیری شده NO₂ در شهر برادفورد کشور بریتانیا استفاده شده است. در این تحقیق از داده‌های تراکم جمعیت، فاصله از جاده‌ها و ترافیک نیز استفاده شده است و نتایج این تحقیق برای داده‌های آزمایشی، جذر میانگین مربعات خطای 5/8 – 7/6 (μg/m³) و ضریب تعیین 0/4 – 0/7 می‌باشد [13].

Delavar و همکاران به بررسی روش‌های مدل‌سازی زمانی آلودگی هوای شهر تهران بر اساس داده‌های هواشناسی و توپوگرافی و داده‌های زمینی ایستگاه‌های سنجش آلودگی هوا پرداختند. آن‌ها در تحقیق خود از روش‌های ماشین بردار پشتیبان رگرسیونی، رگرسیون وزن‌دار جغرافیایی، شبکه عصبی مصنوعی و شبکه عصبی غیرخطی اتورگرسیو با ورودی خارجی استفاده کردند. در نهایت، شبکه عصبی غیرخطی اتورگرسیو با ورودی خارجی را به عنوان روش بهینه با خطای 1/72 (μg/m³) برای مدل‌سازی PM_{2.5} انتخاب کرده‌اند [11].

با هدف پیش‌بینی غلظت دو آلاینده‌ی NO_x و CO در مکان‌های صنعتی با استفاده از مدل شبکه عصبی غیرخطی اتورگرسیو با ورودی خارجی انجام شده است. پایگاه داده مورد استفاده برای آموزش شبکه عصبی مطابق با سری زمانی تاریخی متغیرهای هواشناسی (سرعت باد، جهت باد، دما و رطوبت نسبی) و غلظت آلاینده‌ها در سایت پتروشیمی Skikda می‌باشد. در نتیجه عملکرد مدل پیشنهادی، مقادیر

1 Land Use Regression (LUR)

و استفاده از داده‌های ماهواره‌ای به عنوان یک ابزار تکمیلی در چنین مطالعاتی پیشنهاد می‌شود، زیرا نقشه‌ی حاصل از تصاویر ماهواره‌ای، به صورت پیوسته بوده و می‌تواند با نقشه‌ی حاصل از ایستگاه‌های زمینی ترکیب شود تا دقت بهتری حاصل شود از این رو، در چند سال اخیر، جهت برآورد غلظت ذرات معلق از عمق اپتیکی هواویزها (AOD) استفاده شده‌است که می‌توان اشاره کرد به:

Yang و همکاران به بررسی همبستگی AOD-PM_{2.5} در رزولیشن‌های مختلف تصاویر 1km، 3km و 10km پرداختند و از تصاویر سازمان زمین شناسی ایالات متحده⁹ و اصلاح چند زاویه‌ای تصحیح جوی¹⁰ استفاده کردند. نتایج آن‌ها نشان می‌دهد که می‌توان در سطح اطمینان 95% استفاده از داده‌های MODIS AOD برای ارزیابی غلظت PM_{2.5} ماهانه در هر دو سطح شهر و ایالت استفاده کرد از روش رگرسیون خطی برای بررسی کورلیشن استفاده کردند [14].

Li و همکاران در تحقیق خود به بررسی همبستگی AOD-PM_{2.5} با استفاده از مدل رگرسیون کاربری اراضی¹¹ پرداختند و از تصاویر MAIAC AOD و PM_{2.5} برای شهر Beijing استفاده کردند. نتایج نشان داد که AOD به تنهایی می‌تواند مستقیماً برای پیش‌بینی توزیع فضایی غلظت PM_{2.5} زمین در سطح فصلی مورد استفاده قرار گیرد، ($R^2 > 0.53$) که با قابلیت مدل‌های LUR قابل مقایسه است ($R^2 > 0.81$) [15].

Wang و همکاران به تخمین PM_{2.5} بر مبنای سیستم مدل‌سازی پیش‌بینی کیفیت هوای تو در تو¹² و AOD MODIS در شهرهای Beijing-Tianjin-Hebei برای داده‌های سال 2017 پرداختند و از مدل اثر مختلط خطی¹³ برای تخمین استفاده کردند. نتایجشان نشان داد که یک ارتباط قوی در تخمین غلظت PM_{2.5} با $R^2=0.78$ و $(\mu\text{g}/\text{m}^3)$ RMSE=26.44 وجود دارد [16].

Wang و همکاران پیش‌بینی و تجزیه و تحلیل PM_{2.5} در منطقه فولینگ چونگ کینگ¹⁴ توسط روش‌های انتشار

جذر میانگین خطا¹ برای آلاینده‌های NO_x و CO به ترتیب 0/8445 و 0/7618 و مقادیر میانگین خطای مطلق² برای آلاینده‌های NO_x و CO به ترتیب 0/1451 و 0/1598 حاصل شد. بررسی نتایج، اهمیت متغیر هواشناسی را جهت پیش‌بینی غلظت آلاینده‌ها و کارایی شبکه عصبی NARX نشان داد [10].

در زمینه دقت پیش‌بینی میانگین روزانه غلظت آلاینده PM_{2.5}، مدلی جدید که ترکیبی از تحلیل مسیر توده هوا و تحلیل موجک³ است، برای بهبود شبکه عصبی مصنوعی ارائه شده‌است [1]. این مدل از 13 ایستگاه مختلف پایش آلودگی هوا در شهرهای پکن⁴، تیانجین⁵ و هبی⁶ تهیه شده‌است. سرعت و جهت باد به عنوان پارامترهایی در محاسبه‌ی مقدار آلودگی هوا براساس مسیر توده‌ی هوا مورد توجه قرار گرفتند. علاوه بر این، سری زمانی اولیه غلظت PM_{2.5} با کمترین تغییرپذیری، به وسیله تبدیل موجک به چند زیرسری تجزیه شد. استراتژی پیش‌بینی روی هر یک از آن‌ها اعمال شد. متغیرهای پیش‌بینی روزانه هواشناسی و همچنین پیش‌بینی‌کننده‌های آلاینده‌های مربوطه به عنوان ورودی برای مدل پرسپترون چندگانه⁷ (MLP) از نوع شبکه عصبی پس انتشار⁸ استفاده شدند. تأیید آزمایشگاهی مدل پیشنهادی در طی یک دوره بیش از یک سال (سپتامبر 2013-اکتبر 2014) انجام شد و با بررسی عملکرد برآورد نشان داده شد که مدل جغرافیایی مبتنی بر مسیر و تبدیل موجک می‌تواند ابزار موثر برای بهبود دقت پیش‌بینی PM_{2.5} باشد. مقادیر جذر میانگین مربعات خطای مدل ترکیبی می‌تواند به طور متوسط تا 40% کاهش یابد.

در تحقیقات بالا فقط از داده‌های زمینی استفاده کردند، اما در برخی از تحقیقات نیز از ترکیب تصاویر ماهواره‌ای و داده‌های ایستگاه‌های زمینی به منظور مدل‌سازی مکانی-زمانی آلودگی هوا استفاده شده‌است، در واقع به علت پراکندگی ایستگاه‌های زمینی، برآورد آلودگی هوا در محل‌هایی که این ایستگاه‌ها موجود نباشند، دقت زیادی ندارد

8 Back-Propagation Neural Network

9 Unites States Geological Survey (USGS)

10 Multi-angle implementation of atmospheric correction

11 Land use regression model (LURM)

12 nested air-quality prediction modeling system (NAQPMS)

13 linear mixed effect (LME)

14 Fuling District of Chongqing

1 Mean Squared Error (MSE)

2 Mean Absolute Error (MAE)

3 Wavelet Analysis

4 Beijing

5 Tianjin

6 Hebei

7 Multi-Layer Perceptron

به عقب¹ و شبکه عصبی مصنوعی² برای داده‌های هواشناسی، PM_{2.5} و AOD بین سال‌های 2016 تا 2017 استفاده کردند. نتایج نشان می‌دهد که غلظت PM_{2.5} رابطه مستقیمی خطی با دما نسبت به نرخ انتشار و نرخ تبدیل پیش فرض دارد، با رطوبت نسبی همبستگی مثبت دارد و با سرعت باد همبستگی معکوس معنی‌داری دارد، اما با بارندگی رابطه خطی آشکاری ندارد و شباهت در مکانیسم تأثیر AOD و غلظت PM_{2.5} منجر به همبستگی مثبت خاصی بین آن‌ها می‌شود [17].

Zhang و همکاران برای تخمین PM_{2.5} با استفاده از AOD از مدل رگرسیون وزن‌دار جغرافیایی³ استفاده کردند. نتایج نشان داد این الگوریتم بهتر از روابط اقلیمی ساده عمل می‌کند. نتایج با استفاده از مجموعه رادیومتر تصویربرداری مادون قرمز مرئی⁴ AOD به صورت $R^2=0.59$ و $RMSE=5.66$ ($\mu g/m^3$) و با استفاده از تصویرگر خط‌پایه پیشرفته⁵ AOD به صورت $R^2=0.44$ و $RMSE=4.53$ ($\mu g/m^3$) می‌باشد [18].

قهرمانلو و همکاران برای پیش‌بینی PM_{2.5} با استفاده از MAIAC AOD از مدل‌های یادگیری ماشینی نظیر MLR، مدل جلوه‌های ترکیبی⁶ و جنگل تصادفی⁷ استفاده کردند که نتیجه گرفتند که دقت روش به ترتیب RF، MEM و MLR بهتر می‌باشد [19].

Ni و همکاران یک الگوی مکانی-زمانی برای پیش‌بینی غلظت PM_{2.5} با استفاده از داده‌های هواشناسی و AOD بر اساس روش یادگیری ماشینی نظیر شبکه عصبی انتشار به عقب ارائه داده‌است. برای داده‌های Beijing-Tianjin-Hebei نتایج به صورت $R^2 = 0.68$ ، $RMSE = 20.99$ برای اعتبار سنجی عمومی و $R^2 = 0.54$ ، $RMSE = 24.13$ برای اعتبارسنجی متقاطع می‌باشد [20].

در واقع AOD یک همبستگی مکانی بالایی با غلظت PM_{2.5} دارند. در اکثر تحقیقات ذکر شده، بر اهمیت استفاده از روش‌های یادگیری ماشینی جهت بررسی این همبستگی دارند. لذا این تحقیق در نظر دارد که به بررسی همبستگی بین PM_{2.5} و AOD با کمک داده‌های هواشناسی نظیر سرعت و جهت باد، دما، فشار و رطوبت هوا و همچنین داده‌های

MAIAC AOD بپردازد که یک مدل پیش‌بینی ماشین لرینینگ برای تخمین PM_{2.5} با استفاده از این داده‌ها ارائه شود. در این تحقیق از سه روش DC، MLP و MLR جهت پیش‌بینی غلظت PM_{2.5} استفاده شده‌است که داده‌های ورودی شامل داده‌های هواشناسی و همچنین داده‌های MAIAC AOD و داده خروجی، PM_{2.5} می‌باشد که جهت آموزش شبکه از این داده‌ها استفاده شده‌است و درنهایت در نظر داریم با استفاده از الگوریتم‌های بهینه‌سازی و توسعه آن به انتخاب ویژگی⁸ بپردازیم تا مشخص شود که کدام یک از داده‌ها می‌توانند کنار هم بیایند تا دقت پیش‌بینی بهبود یابد. نوآوری این مقاله استفاده از MBGA برای انتخاب پارامترهای تأثیرگذار در تخمین غلظت آلودگی هوا می‌باشد. این الگوریتم از روش‌های بهینه‌سازی مبتنی بر جغرافیای زیستی⁹ و GA برای بهینه‌سازی استفاده می‌کند. از طرف دیگر، با توجه به اینکه به علت پوشش ابری در بعضی از مناطق تصاویر ماهواره‌ای اخذ شده، داده‌های از دست‌رفته داریم، لذا در این تحقیق در نظر داریم که با استفاده از تغییر اندازه در محدوده بافر، این مشکل را برطرف کنیم.

هدف اصلی از این مقاله، مدل‌سازی و پیش‌بینی دقیق میزان غلظت PM_{2.5} با استفاده از الگوریتم‌های یادگیری ماشین می‌باشد اهداف تفصیلی تحقیق نیز به شرح زیر می‌باشند:

- 1- استفاده از AOD جهت پیش‌بینی دقیق‌تر آلودگی هوای غلظت PM_{2.5}
- 2- توسعه‌ی یک معماری روش‌های الگوریتم‌های یادگیری ماشینی مبتنی بر الگوریتم‌های بهینه‌سازی
- 3- مقایسه و ارزیابی الگوریتم‌های پیش‌بینی، جهت تعیین روش‌های بهینه
- 4- مشخص کردن پارامترهای تأثیرگذار در تخمین غلظت PM_{2.5}

سؤالات این تحقیق به شرح زیر می‌باشند:

- 1- چگونه می‌توان مسئله پیش‌بینی آلودگی هوا را در قالب الگوریتم‌های یادگیری ماشینی، طراحی و پیاده‌سازی نمود؟
- 2- چگونه می‌توان یک الگوریتم یادگیری ماشین را توسعه و بهبود داد؟

6 Mixed Effects Model (MEM)

7 Random Forest (RF)

8 Feature Selection

9 Biogeography-based Optimization

1 Back propagation (BP)

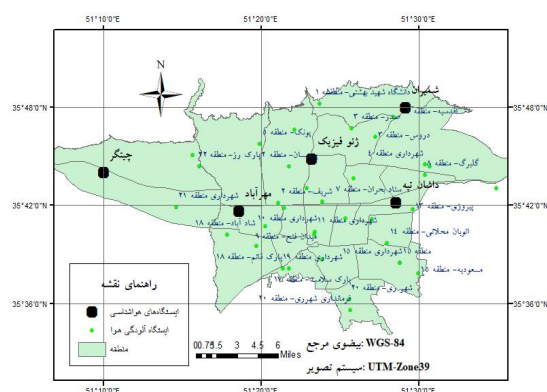
2 Artificial neural network (ANN)

3 Geographically weighted regression (GWR)

4 Visible Infrared Imaging Radiometer Suite (VIIRS)

5 Advanced Baseline Imager (ABI)

بیماری‌های ناشی از آلودگی هوا جلوگیری کرد. برای این منظور جهت پیش‌بینی آلودگی هوا، کل شهر تهران به عنوان منطقه مطالعاتی انتخاب شده‌است. در شکل (1) نقشه مناطق شهر تهران به همراه ایستگاه‌های هواشناسی، سنجش آلودگی هوا سازمان حفاظت محیط زیست و شرکت کنترل کیفیت شهرداری تهران نشان داده شده‌است.



شکل 1- مناطق شهر تهران و ایستگاه‌های هواشناسی، سنجش آلودگی هوای شرکت کنترل کیفیت شهرداری و سازمان حفاظت محیط زیست

2-2- جمع‌آوری داده

آلودگی هوا از جمله پدیده‌هایی است که عوامل مختلفی در آن تاثیر می‌گذارند. برای پیش‌بینی دقیق نیاز است این پارامترهای تاثیرگذار در آلودگی هوا به درستی شناسایی شوند. در جدول (1) جزئیات داده‌ها آمده‌است.

2-2-1- آلودگی هوای ایستگاه زمینی

در این تحقیق، داده‌های مربوط به غلظت آلاینده PM_{2.5} شرکت کنترل کیفیت هوای شهرداری تهران از سال 1396 تا 1399 دریافت گردیده‌است. غلظت این آلاینده‌ها در بازه‌ی روزانه جمع‌آوری شده و توسط این شرکت در اختیار قرار گرفته شده‌است. 1 شهر تهران دارای 41 ایستگاه سنجش آلودگی هوا است که توسط سازمان حفاظت محیط زیست و شرکت کنترل کیفیت هوای شهرداری تهران، کنترل و مدیریت می‌شوند. اما از آنجایی که در این تحقیق داده‌های روزانه مورد نیاز است و این داده‌ها برای تمامی ایستگاه‌ها در دسترس نمی‌باشد، تنها داده‌های مربوط به 22 ایستگاه مورد استفاده قرار گرفته‌اند.

3- آیا می‌توان دقت الگوریتم‌های مدلسازی و پیش‌بینی آلودگی هوا را بهبود داد؟

با توجه به موضوع پیشنهادی در این تحقیق، نویسندگان جنبه‌های زیر را به عنوان نوآوری‌های پیشنهادی در نظر دارند:

1- بهبود یک الگوریتم پیش‌بینی جهت برآورد دقیق‌تر میزان غلظت آلودگی هوا

2- توسعه‌ی یک الگوریتم یادگیری ماشینی با استفاده از الگوریتم‌های بهینه سازی و پالایش داده

2- روش‌ها و موادها

2-1- منطقه مورد مطالعه

شهر تهران به عنوان پایتخت کشور، مهمترین کلانشهر و مرکز سیاسی و تجاری کشور محسوب می‌شود که بالغ بر بیست درصد جمعیت کشور در آن ساکن هستند. تهران توسط رشته‌کوه‌های البرز از سمت شمال و دشت کویر از سمت جنوب احاطه شده‌است. آب و هوای شهر تهران تاثیر گرفته از موقعیت جغرافیایی آن است. جز مناطق شمالی تهران که تحت تاثیر کوهستان، تا اندازه‌ای معتدل و مرطوب هستند، آب و هوای دیگر مناطق شهر تقریباً گرم و خشک و در زمستان‌ها اندکی سرد است. رشته کوه البرز همچون سدی از نفوذ بسیاری از توده‌های هوا جلوگیری می‌کند، از همین روی سبب گردیده که تهران از آب و هوایی نسبتاً خشک برخوردار باشد. محصور بودن در بین کوه‌ها از سه طرف، که مانع خروج آلودگی‌ها از شهر می‌شود، از یک سو و افزایش بی‌رویه‌ی استفاده از وسایل نقلیه و گسترش صنایع از عوامل اصلی آلودگی هوا در شهر تهران می‌باشند.

آلودگی هوا در شهر تهران عمدتاً مصنوعی و ناشی از فعالیت وسایل نقلیه است که سهم بالایی در آلودگی هوای شهر دارند. میزان آلودگی هوای شهر گاهی از حد استاندارد بالاتر رفته و در برخی موارد موجب تعطیلی کامل شهر می‌گردد. بنابراین، پیش‌بینی و مدلسازی آلودگی هوا برای شهر تهران امری ضروری بوده تا اقدامات لازم جهت کنترل آلودگی انجام شده و مکان‌هایی که از نظر آلودگی در وضعیت خطرناکی قرار دارند، شناسائی گردند. همچنین، می‌توان با آگاه‌سازی عموم و بخصوص افراد بیمار، از گسترش

جدول 1- جزئیات داده‌های مورد استفاده

منبع	دقت مکانی	محدوده‌ی زمانی	داده	نوع داده
شرکت کنترل کیفیت شهرداری تهران	در دسترس نیست	1396-1399	PM _{2.5} (µg/m ³)	آلودگی هوای ایستگاه زمینی
NASA, LP DAAC	1km×1km	1396-1399	Optical_Depth_047(µm) Optical_Depth_055(µm)	تصاویر ماهواره‌ای AOD
سازمان هواشناسی ایران	در دسترس نیست	1396-1399	دما (°C) فشار (pa) رطوبت (%) سرعت باد (m/s) جهت باد (°)	داده‌های هواشناسی

2-2-2- داده‌های هواشناسی

داده‌های روزانه هواشناسی نظیر جهت باد (WDIR) و سرعت باد (WSPD)، دما (TEMP)، فشار هوا (APRS) و رطوبت نسبی (RHUM)، توسط سازمان هواشناسی ایران فراهم شده‌است. این داده‌ها در 5 ایستگاه: شمیران، داشان تپه، مهرآباد، چیتگر و ژئوفیزیک اندازه‌گیری شده‌است.

2-2-2-3- تصاویر ماهواره‌ای AOD

از طرف دیگر برای بهبود دقت پیش‌بینی و مدلسازی آلودگی هوا و همچنین بررسی همبستگی آن با از داده‌های AOD تصاویر ماهواره‌ای MAIAC استفاده شده‌است. در این مقاله از تصاویر ماهواره‌ای بین سال‌های 1396 تا 1399 از NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC) استفاده شده‌است. (<https://doi.org/10.5067/MODIS/MCD19A2.006>).

در این تحقیق از باندهای AOD47 و AOD55 این تصاویر استفاده شده‌است، بدین صورت که به ازای مکان هر یک از 22 ایستگاه سنجش ذرات معلق آلودگی هوای شهر

تهران، مقادیر پیکسل باندهای مذکور در بین سال‌های 1396 تا 1399 اخذ شده‌است.

2-3- روش تحقیق

ساختار اصلی روش تحقیق شامل دو بخش جمع‌آوری داده‌ها و پیش‌پردازش و توسعه مدل می‌باشد که در شکل (2) روش تحقیق نمایش داده شده‌است.

2-4- جمع‌آوری داده‌ها و پیش‌پردازش

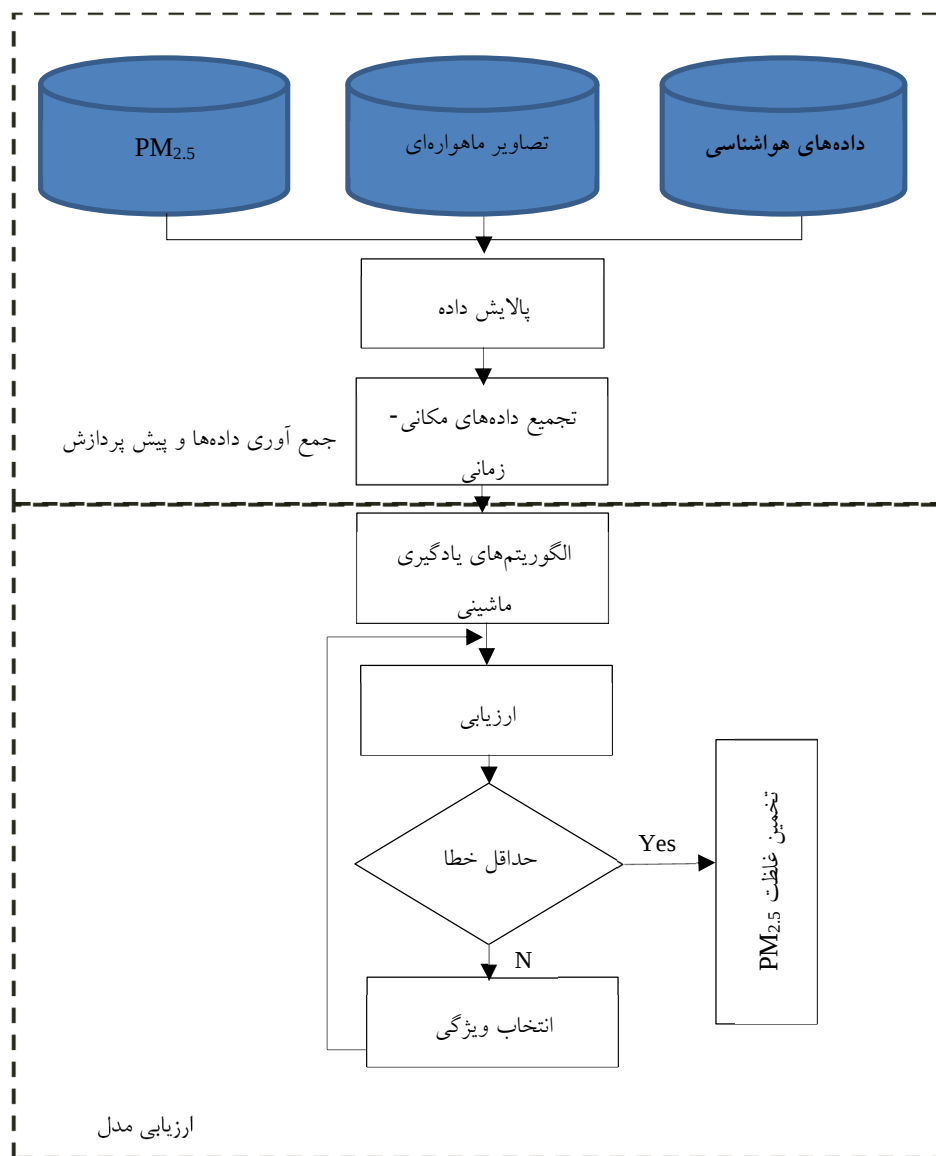
در این تحقیق از سیستم تصویر WGS 841 استفاده شده‌است. جهت درونیابی داده‌های هواشناسی و انتقال مکانی داده‌ها به 22 ایستگاه سنجش آلودگی هوای شهر تهران از مدل درونیابی Ordinary Kriging استفاده شده‌است.

2-5- پالایش داده‌ها

داده‌های ورودی هواشناسی ایستگاه‌های زمینی سنجش سازمان هواشناسی هم از لحاظ مکانی و هم به لحاظ زمانی به علت خطاهای دستگاهی، دچار missing data شده‌است. لذا طبق [11] از روش‌های Fourier Series و Spline جهت برطرف کردن این مشکل استفاده شده‌است. این دو روش برای یک سری داده زمانی آلودگی هوای شهر تهران به مدت 10 سال استفاده شده‌اند و برای چندین داده به صورت روزانه، هفتگی و ماهانه استفاده شدند و مدلی برازش داده شدند که طبق نتایج [11] روش سری فوریه برای داده‌های از دست رفته ماهانه و روش اسپیلاین برای داده‌های از دست رفته روزانه و هفتگی مناسب بودند.

جهت اخذ داده‌های AOD در 22 موقعیت مکانی ایستگاه‌های زمینی آلودگی هوا، از تحلیل بافر² استفاده شده‌است که در شکل (3) نمایش داده شده‌است.

طبق شکل (3) ابتدا برای هر نقطه، یک آنالیز بافر شکل می‌گیرد، از آنجایی که سایز پیکسل‌ها 1km×1km می‌باشد مقادیر distance برای این آنالیز به ترتیب 500، 1000 و 2000 متر در نظر گرفته شده‌است تا حد الامکان مقادیر پیکسل برای هر نقطه مدنظر پیدا شود. این کار تا زمانی انجام می‌شود که برای تمامی 22 ایستگاه در روزهای 2014 تا 2017، مقادیر AOD استخراج شود.



شکل 2- روش تحقیق

2-6- تجمیع داده‌های مکانی-زمانی

داده‌ها در این مرحله به دو دسته داده ورودی و خروجی برای ورود به الگوریتم‌های یادگیری ماشینی کار گرفته می‌شوند. داده‌های ورودی شامل پارامترهای هواشناسی و AOD می‌باشد، همچنین داده‌های خروجی فقط $PM_{2.5}$ می‌باشد.

2-7- الگوریتم‌های یادگیری ماشینی

جهت آموزش شبکه توسط داده‌های AOD و هواشناسی و تخمین غلظت $PM_{2.5}$ روش‌های نظیر DC، MLP و

MLR مقایسه خواهد شد. بهترین روش به عنوان الگوریتم انتخابی در انتخاب ویژگی¹ استفاده می‌شود.

2-7-1- مدل پرسپترون پیشخور چندلایه (MLP)

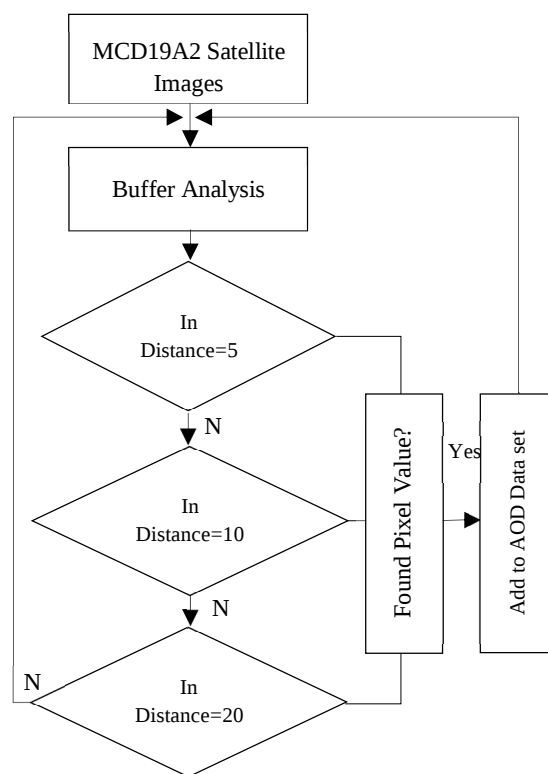
در ابتدایی ترین شکل خود، این مدلی است که از تعداد محدودی لایه‌های متوالی تشکیل شده‌است. هر لایه از تعداد محدودی واحد (اغلب به نام نورون) تشکیل شده‌است. هر واحد از هر لایه به هر واحد لایه بعدی (و بنابراین قبلی) متصل است. این اتصالات را عموماً لینک یا سیناپس می‌نامند. اطلاعات از یک لایه به لایه بعدی جریان می‌یابد (بنابراین اصطلاح پیشخور نامیده می‌شود). لایه اول که لایه

¹ Feature Selection

ورودی نام دارد از ورودی تشکیل شده است. سپس لایه های میانی وجود دارد که به آن ها لایه های پنهان می گویند. خروجی حاصل در آخرین لایه به دست می آید که لایه خروجی نامیده می شود [21].

2-7-2- درخت تصمیم گیری (DC)

یادگیری درخت تصمیم گیری یک رویکرد یادگیری تحت نظارت است که در آمار، داده کاوی و یادگیری ماشین استفاده می شود. درخت تصمیم طبقه بندی یا رگرسیون به عنوان یک مدل پیش بینی برای نتیجه گیری در مورد مجموعه ای از مشاهدات استفاده می شود.



شکل 3- پالایش داده ها

مدل های درختی که در آن متغیر هدف می تواند مجموعه ای از مقادیر مجزا بگیرد، درختان طبقه بندی نامیده می شوند. در این ساختارهای درختی، برگ ها نشان دهنده برجسب های کلاس و شاخه ها نشان دهنده ترکیبی از ویژگی هایی هستند که به آن برجسب های کلاس منتهی می شوند. درخت های تصمیم که در آن متغیر هدف می تواند

مقادیر پیوسته (معمولاً اعداد واقعی) بگیرد، درخت های رگرسیون نامیده می شوند [22].

2-7-3- رگرسیون خطی چندگانه (MLR)

یک روش رگرسیون-تحلیلی و یک مورد خاص از رگرسیون خطی است. رگرسیون خطی چندگانه یک تکنیک آماری است که سعی دارد یک متغیر وابسته مشاهده شده را بر حسب چندین متغیر مستقل توضیح دهد. مدل مورد استفاده در پارامترها خطی است و متغیر وابسته تابعی از متغیر مستقل است. این رابطه توسط یک تابع خطی تعریف می شود. رگرسیون خطی چندگانه، تعمیم رگرسیون خطی ساده را با توجه به تعداد رگرسیون ها نشان می دهد [23].

2-8- ارزیابی مدل

در این مرحله هر یک از الگوریتم های یادگیری ماشینی اعتبارسنجی خواهند شد که در این تحقیق از روش های متداول اعتبارسنجی نظیر ضریب تعیین¹، خطای میانگین نرمال شده²، جذر میانگین مربعات خطا³ و میانگین درصد خطای مطلق⁴ استفاده شده است [24]، [25]. فرمول های آن در روابط (1-4) نوشته شده است.

$$R^2 = 1 - \frac{\sum_{i=1}^N [(P_i - O_i)^2]}{\sum_{i=1}^N [(P_i - \bar{P})^2]} \quad (1)$$

$$NMB = \frac{\sum_{i=1}^N P_i - O_i}{\sum_{i=1}^N O_i} \quad (2)$$

$$MAPE = \frac{\sum_{i=1}^N \left| \frac{P_i - O_i}{O_i} \right|}{n} \times 100 \quad (3)$$

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^N [P_i - O_i]^2 \right)^{1/2} \quad (4)$$

جایی که N تعداد مشاهدات، O_i پارامتر مشاهده شده، P_i پارامتر محاسبه شده، $\bar{O}$ میانگین پارامتر مشاهدات، σ_p میانگین پارامتر محاسبات، σ_o انحراف معیار مشاهدات و σ_p انحراف معیار محاسبات است.

3 Root mean squared error (RMSE)

4 Mean absolute percent error (MAPE)

1 Coefficient of determination (R²)

2 Normalized mean bias (NMB)

2-9- انتخاب ویژگی

از آنجایی که داده‌های هواشناسی شامل 5 پارامتر و AOD شامل 3 پارامتر می‌باشد، بنابراین داده‌های ورودی الگوریتم‌های پیش‌بینی شامل 8 پارامتر می‌باشد. در این تحقیق فرض شده‌است که این 8 پارامتر، پارامترهای تاثیرگذار در پیش‌بینی آلودگی هوا می‌باشد؛ برای سرعت بخشیدن به اجرای الگوریتم‌ها و حذف پارامترهای با تاثیر کمتر (ضریب همبستگی پایین) در پیش‌بینی PM2.5 در این تحقیق از الگوریتم‌های بهینه‌سازی نظیر PSO و GA استفاده شده‌است. از طرف دیگر در این تحقیق، الگوریتم ژنتیک با استفاده از تغییر نرخ مهاجرت بهبود داده شده است. تابع فیتنس برای تمامی الگوریتم‌های بهینه‌سازی، همان RMSE الگوریتم‌های پیش‌بینی می‌باشد که هدف از الگوریتم‌های بهینه‌سازی مینیمم کردن RMSE می‌باشد.

2-9-1- الگوریتم ژنتیک¹

الگوریتم ژنتیک (GA) از جمله روش‌های فراابتکاری در مسائل بهینه‌سازی است که اولین بار در سال 1975 به‌وسیله Holland مطرح شده‌است و در واقع یک شبیه‌سازی مجازی از نظریه تکامل تدریجی داروین² می‌باشد [26]. این الگوریتم یک روش مبتنی بر جمعیت می‌باشد که در هر تکرار محاسباتی (نسل) روی جمعیتی از کروموزوم‌ها عمل کرده و تغییرات تصادفی بر روی آن‌ها از طریق اعمال عملگرهای ادغام و جهش ایجاد می‌کند. سپس جواب‌های مختلف به‌دست‌آمده از نظر عملکرد بر اساس تابع هدف ارزیابی می‌شوند. الگوریتم ژنتیک دارای سه عملگر اصلی انتخاب³، ادغام⁴ و جهش⁵ می‌باشد [27].

2-9-2- الگوریتم ژنتیک مبتنی بر مهاجرت (MBGA)⁶

قدرت exploitation و exploration کم، یکی از ضعف‌های الگوریتم ژنتیک استاندارد می‌باشد. لذا بهبود عملگرهای انتخاب، تقاطع و جهش می‌تواند سبب بهبود

عملکرد الگوریتم شود. در الگوریتم ژنتیک استاندارد زمانی که از چرخ گردان برای انتخاب والدین استفاده می‌شود، به والدین با تابع بهینگی خوب بها داده می‌شود. در واقع فرایند crossover بین دو والد با تابع بهینگی خوب اتفاق می‌افتد. در نتیجه، فقط به کروموزوم‌ها با تابع بهینگی خوب فرصت داده می‌شود. اما در بسیاری از مواقع، شاید از ترکیب یک کروموزوم با تابع بهینگی خوب و یک کروموزوم با تابع بهینگی بد، فرزندی خوب بوجود آید. یعنی جواب بهینه مسئله از ترکیب چنین والد‌هایی حاصل می‌شود. لذا در این مقاله، از نرخ‌های مهاجرت⁷ الگوریتم بهینه‌سازی مبتنی بر جغرافیای زیستی (BBO)⁸، برای انتخاب والدین در الگوریتم ژنتیک استفاده می‌شود. به این صورت که هر کروموزوم نرخ مهاجرت به بیرون⁹ و نرخ مهاجرت به داخل¹⁰ خاص خودش را دارد. مهاجرت به بیرون $(\mu_j(k))$ و مهاجرت به داخل $(\lambda_j(k))$ ، به‌عنوان توابعی از تعداد ژن‌های‌شان به‌صورت روابط (5) و (6) تعریف می‌گردند [28], [29].

$$\mu_{k(j)} = \frac{E}{2} \times (-\cos(\frac{k(j)\pi}{N}) + 1) \quad (5)$$

$$\lambda_{k(j)} = \frac{I}{2} \times (\cos(\frac{k(j)\pi}{N}) + 1) \quad (6)$$

جاییکه $k(j)$ نشان‌دهنده رتبه گونه‌ها (بر اساس عملکرد تناسب اندام) است که در زیستگاه j^{th} زندگی می‌کنند. E و I به ترتیب بالاترین نرخ emigration و بالاترین نرخ immigration را نمایش دهید.

این نرخ‌ها که به صورت سینوسی می‌باشند، نحوه انتخاب والدین را به صورت هوشمند تعیین می‌کنند. به این صورت که ابتدا یک والد با تابع بهینگی ضعیف (با نرخ مهاجرت به داخل بالا) و والد دیگر با تابع بهینگی خوب (با نرخ مهاجرت به بیرون بالا) برای عمل تقاطع انتخاب می‌شوند که نتیجه آن ایجاد دو فرزند می‌باشد. به عبارتی دیگر، نحوه انتخاب والدین بر اساس مقدار مستقیم تابع بهینگی نمی‌باشد. بلکه از طریق نرخ‌های مهاجرت می‌باشد. این نرخ‌های مهاجرت نیز از مقدار تابع بهینگی کروموزوم‌ها حاصل می‌شوند (با توجه به روابط 5 و 6). نحوه انتخاب دو

6 Migration-based genetic algorithm (MBGA)

7 Migration

8 Biogeography-based Optimization (BBO)

9 Emigration rate

10 Immigration rate

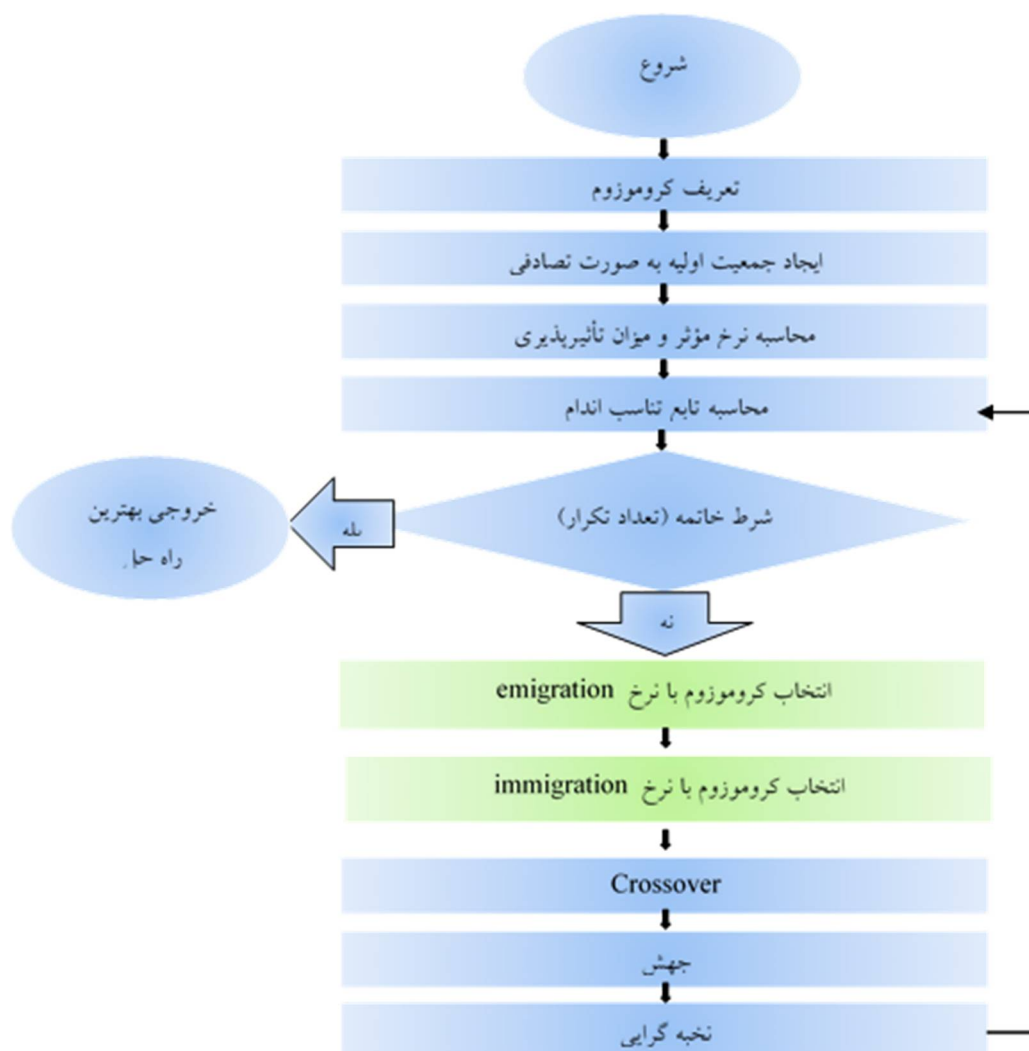
1 Genetic Algorithm (GA)

2 Darwin

3 Selection

4 Crossover

5 Mutation



شکل 4- الگوریتم MBGA

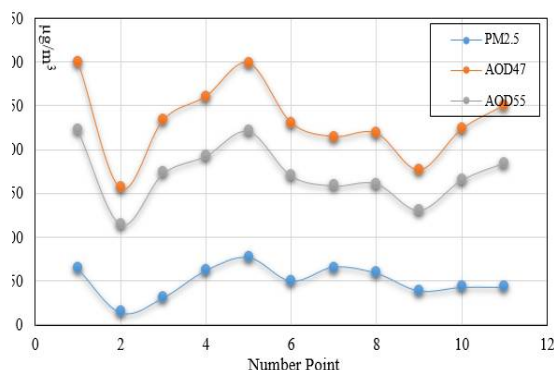
مهم‌ترین مزیت الگوریتم MBGA نسبت به GA این است که علاوه بر بهبود exploitation الگوریتم (از طریق نرخ‌های مهاجرت در عملگر ترکیب)، سبب بهبود exploration نیز می‌شود. زیرا ترکیب کروموزوم‌های ضعیف و کروموزوم‌های قوی سبب اکتشاف بیشتر می‌شود و کل فضای مسئله جستجو می‌شود. بخش‌های دیگر این الگوریتم از جمله تعریف کروموزوم، جمعیت اولیه، جهش و نحوه نخبه‌گرایی همانند الگوریتم ژنتیک می‌باشد.

در این تحقیق از 7 ژن، شامل پارامترهای AOD 55 , WSPD , WDIR, TEMP, RHUM, APRS and AOD 47 می‌باشد و کروموزوم‌ها شامل ترکیب از این 7 ژن می‌باشد که ممکن است در یک کروموزوم یک یا دو یا سه و... 7 ژن وجود داشته باشد که نشان‌دهنده پارامترهای تأثیر گذار می‌باشد، این کروموزوم‌ها باینری می‌باشند که عدد یک در هر ژن، نشان‌دهنده انتخاب آن پارامتر به عنوان پارامتر

والد بر اساس چرخ گردان مبتنی بر نرخ‌های مهاجرت می‌باشد. ترکیب در MBGA به این شکل می‌باشد که کروموزوم با نرخ مهاجرت به بیرون بالا، ژن‌های خود را به سمت کروموزوم با نرخ مهاجرت به داخل بالا می‌فرستد و تمامی جواب‌ها رو به اصلاح خواهند رفت که باعث بهبود عملکرد الگوریتم می‌شود. در شکل (4) روند کلی الگوریتم پیشنهادی آمده‌است. همانطور که مشاهده می‌شود، از نرخ‌های مهاجرت برای انتخاب و ترکیب کروموزوم‌ها استفاده می‌شود.

تفاوت اصلی الگوریتم توسعه داده‌شده و الگوریتم ژنتیک استاندارد در این است که در MBGA به جای آنکه کروموزوم‌های ضعیف حذف شوند، بلکه با استفاده از کروموزوم‌هایی با تابع بهینگی بالا بهبود داده می‌شوند. این باعث می‌شود اولاً سرعت همگرایی الگوریتم افزایش پیدا کند، ثانیاً الگوریتم کمتر در بهینه محلی گرفتار شود.

بین مقادیر $PM_{2.5}$ و AOD داده‌ها، همبستگی وجود دارد که برای داده‌های این تحقیق این مقدار بین 0.17 و 0.83 می‌باشد، در شکل (5) برای داده‌های یک روز، این همبستگی نشان داده شده‌است که مقدار همبستگی این داده‌ها برابر 0.7417 می‌باشد.



شکل 5- همبستگی داده‌های زمینی و هوایی

تاثیرگذار می‌باشد و عدد صفر نشان دهنده عدم انتخاب آن پارامتر به عنوان پارامتر تاثیرگذار می‌باشد.

تابع فیتنس برای هر کروموزوم نشان‌دهنده مقدار خطای RMSE آموزش شبکه با استفاده از پارامترهای تاثیرگذار منتخب در کروموزم جهت تخمین غلظت $PM_{2.5}$ می‌باشد که هدف این است که این تابع فیتنس با انتخاب بهترین کروموزم در 100 نسل، کمینه شود.

3- نتایج و بحث

جدول (2) جزئیات آماری داده‌های $PM_{2.5}$ ، AOD و هواشناسی نشان می‌دهد. همانطور که نشان داده شده‌است، داده $PM_{2.5}$ غیر نرمال و دارای نویزهای زیادی است، زیرا مقادیر کشیدگی (Kurtosis) و چولگی (Skewness) خیلی بالایی دارد، لذا طبق [11] از فیلتر Savitzky-Golay جهت حذف نویز استفاده خواهیم کرد.

جدول 2- شاخص‌های آماری داده‌های ورودی

پارامتر	کمینه	بیشینه	میانگین	انحراف معیار	چولگی	کشیدگی	چارک اول	میانه	چارک سوم
AOD47	19	859	05.228	86.113	66.0	41.3	143	213	299
AOD55	14	821	88.168	42.85	84.0	83.4	5.105	5.157	221
WDIR	23.53	6.262	58.190	216.19	54.1-	94.8	94.180	190	201
WSPD	08.2	11	06.4	77.0	63.1	60.7	45.3	99.3	6.4
TEMP	65.0	90.31	14.19	87.8	2.0-	57.1	56.10	7.20	71.27
RHUM	461.11	04.63	3.33	73.11	56.0	04.2	25.23	25.29	44
APRS	21.868	5.884	28.874	08.3	3.0	24.2	8.871	9.873	7.876
$PM_{2.5}$	2	426	4.31	47.17	3.3	68.40	20	28	38

همبستگی $AOD-PM_{2.5}$ از روش‌های MLR، DC و MLP استفاده شده‌است. داده‌های پالایش‌شده در هر یک از شبکه‌ها آموزش داده شدند و در نهایت پارامترهای ارزیابی هر یک از الگوریتم‌ها محاسبه شد و در جدول (3) نتایج آن نشان داده شده‌است.

جدول 3- نتایج الگوریتم‌های یادگیری ماشین

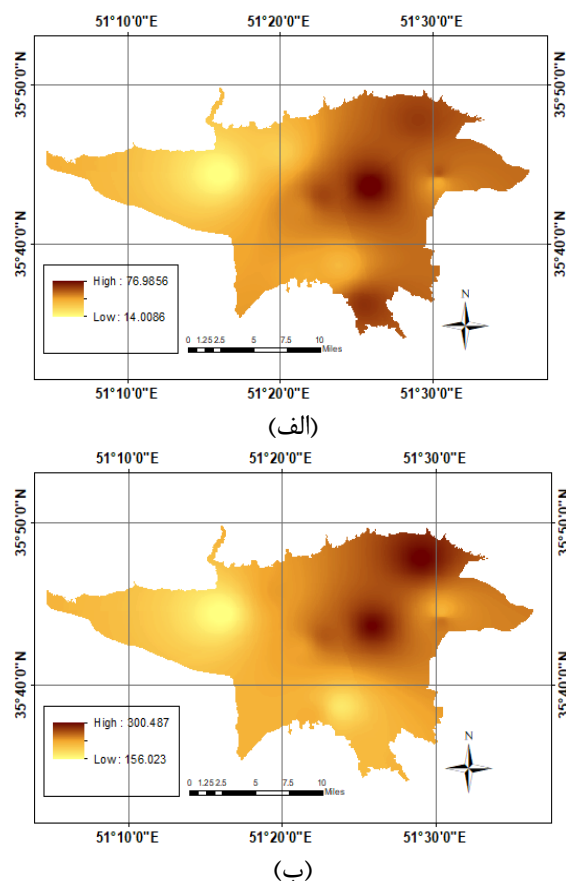
الگوریتم	RMSE	R^2	NMB	MAPE
DC	14.02	0.55	-0.0044	0.0018
MLR	15.14	0.46	0.014	0.002
MLP	11.46	0.67	-0.004	0.0015

نقشه آلودگی هوا بر اساس داده‌های $PM_{2.5}$ و AOD مطابق شکل (6) می‌باشد، همانطور که از شکل‌های (6) پیداست، داده‌های AOD می‌تواند به عنوان یک پارامتر تاثیرگذار در پیش‌بینی غلظت $PM_{2.5}$ استفاده کرد.

از آنجایی که همبستگی $AOD-PM_{2.5}$ برای ایستگاه‌های مختلف متفاوت است لذا الگوی خاصی در این همبستگی پیدا نشد، بنابراین سعی بر آن شد که با استفاده از الگوریتم‌های ماشین لرنینگ و الگوریتم‌های بهینه‌سازی این همبستگی را به یک مقدار مناسبی برای کل داده‌ها برازش داده شود. در این تحقیق جهت برآورد تخمین غلظت $PM_{2.5}$ و افزایش

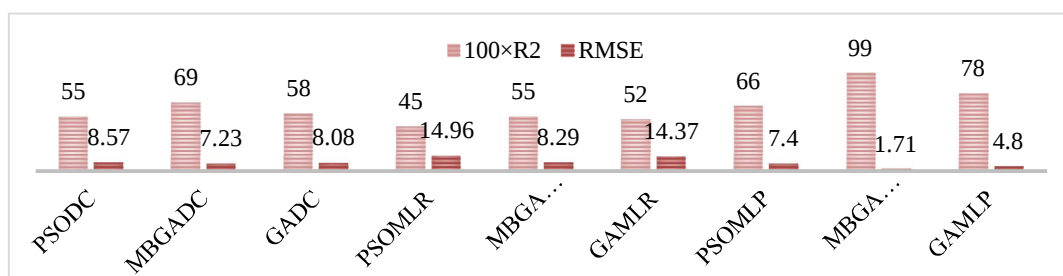
روش در شاخص‌های NMB و MAPE بهتر عمل کرده‌است. روش MLR بیشترین RMSE و کمترین R^2 را ایجاد کرده‌است و نشان می‌دهد که نمی‌توان برای برازش این داده‌ها از روش خطی رگرسیون استفاده کرد و باید روش غیرخطی دیگری انتخاب کرد، از طرف دیگر روش DC نتوانست دقت خوبی برای داده‌های پیچیده بدهد. در نهایت از آنجایی که روش MLP از توابع غیرخطی استفاده می‌کند این روش همانند جعبه سیاه عمل می‌کند و می‌تواند روابط پیچیده و همبستگی AOD-PM_{2.5} را همراه با داده‌های هواشناسی مدل‌سازی کند.

در این مرحله می‌خواهیم طبق شکل (روش تحقیق) پارامترهایی که تاثیر کمتری در تخمین غلظت PM_{2.5} دارند را حذف کنیم (Feature Selection)؛ بنابراین از روش‌های بهینه‌سازی نظیر GA، PSO و روش پیشنهادی MBGA استفاده شده‌است، ورودی‌های این روش‌های بهینه‌سازی، باینری می‌باشند که مقدار 1 نشان‌دهنده موثر بودن آن پارامتر می‌باشد که در ورودی الگوریتم‌های ماشین لرنینگ به عنوان یک ویژگی استفاده خواهد شد و مقدار 0 نشان‌دهنده ناموثر بودن آن پارامتر است که از لیست پارامترها حذف خواهد شد. در نهایت، هدف مینیمم کردن RMSE می‌باشد که در شکل (7) ترکیب هر یک از روش‌ها به همراه نتایج آن نشان داده شده‌است.



شکل 6- مقایسه (الف) نقشه PM_{2.5} و (ب) نقشه AOD

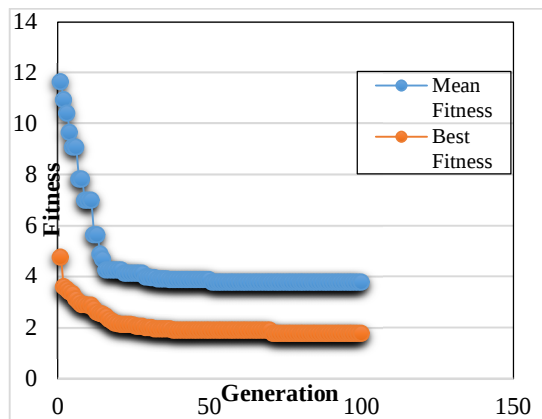
همانطور که از جدول (3) پیداست به ترتیب دقت MLR، DC و MLP بهتر می‌باشد که کمترین RMSE و بیشترین R^2 مربوط به روش MLP می‌باشد، همچنین این



شکل 7- مقایسه ترکیب روش‌های بهینه‌سازی با روش‌های یادگیری ماشین

(Output) و مقادیر واقعی (Target) ضریب تعیین برابر 0.99 ایجاد کند. R^2 نزدیک به یک نشان‌دهنده نزدیکی این دو مقدار به همدیگر می‌باشد، از این رو الگوریتم‌هایی که دارای R^2 نزدیکتری به یک باشند، می‌توانند تخمین صحیح‌تری را داشته باشند. در شکل (8) همبستگی داده‌های Output و Target برای دو روش MLP و MBGAMLP نشان داده شده‌است و همانطور که مشخص است الگوریتم MBGA

همانطور که مشخص است، الگوریتم بهینه‌سازی GA از PSO عملکرد بهتری داشته است و این دو الگوریتم بهینه‌سازی برای الگوریتم MLR، تغییرات محسوسی نداشته است و توانسته‌اند برای دو الگوریتم DC و MLP عملکرد بهتری داشته باشند و الگوریتم توسعه داده‌شده MBGA بهترین عملکرد را داشته و توانسته برای تمامی روش‌های ماشین لرنینگ بین مقادیر خروجی آموزش شبکه عصبی



شکل 9- میانگین و بهترین تغییرات تناسب اندام در MBGAMLP

4- نتیجه‌گیری

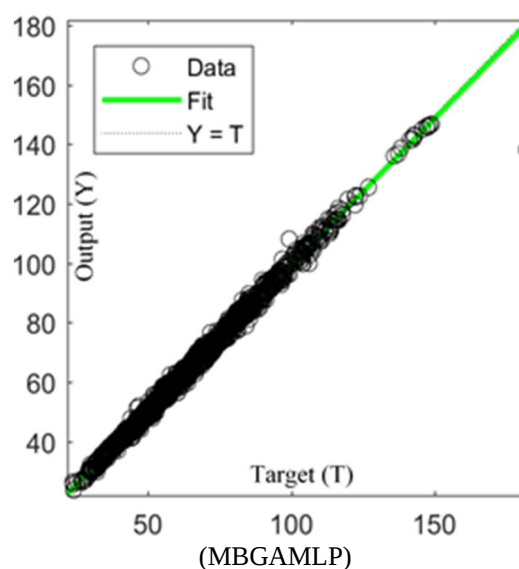
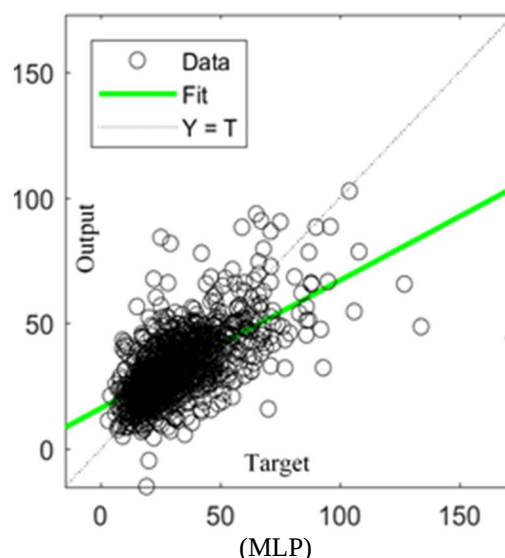
از آنجایی‌که با شیوع کرونا به عنوان بیماری همه‌گیر و در نتیجه تعطیلی اقتصاد جهانی، تغییراتی در مقدار AOD در سرتاسر جهان به همراه داشته است، اما مقدارهای کنونی نیز برای انسان خطرناک بوده‌است و این تحقیق در نظر دارد آلودگی هوا را با دقت مناسبی مدل‌سازی کند.

در این تحقیق، ما توزیع فضایی-زمانی PM_{2.5} را در شهر تهران برای سال‌های 1396 تا 1399 با استفاده از مدل‌های الگوریتم ماشین لرنینگ، با ورودی AOD و داده‌های هواشناسی برآورد کردیم. داده‌های ما شامل 22 ایستگاه برای داده‌های Ground-level PM_{2.5} و دو پارامتر AOD₄₇ WSPD, WDIR, TEMP, AOD₅₅ برای AOD و 5 پارامتر RHUM, APRS برای داده‌های هواشناسی می‌باشد.

در این راستا چند هدف کلی مورد بررسی قرار گرفت. هدف اول بررسی روش‌های یادگیری ماشین در تخمین غلظت آلاینده PM_{2.5} با استفاده از تصاویر سنجنش از دوری بوده است. از بین روش‌های DC, MLR و MLP دقت روش MLP با مشخصات ($RMSE=11.46$ and $R^2=0.67$) از همه بهتر بود. همچنین برای بهبود دقت، GA و الگوریتم پیشنهادی و توسعه داده‌شده MBGA دقت مطلوبی دادند که الگوریتم MBGA توانست خطا را بهینه کند.

مدل برآورد PM_{2.5} با استفاده از روش MBGAMLP توانست غلظت PM_{2.5} را با دقت بالا ($RMSE=1.71$ and $R^2=0.99$) برآورد کند، زیرا الگوریتم‌های GA و PSO با حذف کروموزوم‌های ضعیف و تقویت کروموزوم‌های قوی، باعث می‌شود اولاً سرعت همگرایی کاهش پیدا کند و همچنین تابع در یک بهینه محلی گرفتار شود؛ اما از

توانسته به خوبی این دو داده را به خط $Y=T$ فیت کند. در واقع هر دو شاخص ارزیابی نشان دادند که MBGAMLP در تخمین غلظت آلاینده PM_{2.5} خوب عمل کرده‌است.



شکل 8- همبستگی داده‌های Output و Target برای دو روش MLP و MBGAMLP

در الگوریتم MBGA، پارامترهای موثر، WSPD، TEMP، RHUM و AOD₄₇ و AOD₅₅ انتخاب شدند که توانستند با استفاده از الگوریتم MLP خطا را تا $1.71(\mu g/m^3)$ کاهش دهند. از آنجایی‌که داده‌های AOD₄₇ و AOD₅₅ همبستگی بالایی دارند، می‌توانیم این داده‌ها را یکسان در نظر بگیریم و در مدل یکی از آنها استفاده کنیم. در شکل (9) تغییرات فیتنس میانگین و بهترین فیتنس در 100 نسل نشان داده شده‌است. بهترین فیتنس در نسل 71ام با مقدار $1.71(\mu g/m^3)$ انتخاب شده‌است.

آنجایی که الگوریتم بهینه‌سازی MBGA توانسته با ادغام کروموزوم‌های ضعیف و قوی، سرعت همگرایی را افزایش دهد و دیگر تابع در یک بهینه محلی گرفتار نمی‌شود و تمامی فضای مسئله جستجو شود.

متغیرهای ورودی نظیر WDIR، WSPD، APRS، RHUM و AOD برای برآورد توزیع فضایی-زمانی PM_{2.5} استفاده شدند که از بین آن‌ها پارامترهای موثر در تخمین غلظت PM_{2.5} پارامترهای WSPD، TEMP، RHUM و AOD انتخاب شدند. نتایج تخمین خوبی از PM_{2.5} ارائه می‌دهد، اما ما هنوز محدودیت‌ها و عدم قطعیت‌هایی در برآورد داریم. برای تحقیقات آتی، پیشنهادات زیر جهت بهبود دقت تخمین غلظت PM_{2.5} ارائه می‌گردند:

- اندازه‌گیری‌های غلظت PM_{2.5} در سطح زمین و اندازه‌گیری AOD استخراج‌شده از تصاویر ماهواره‌ای در سطح اتمسفر می‌باشد. در نتیجه، می‌توان در آینده با در نظر گرفتن مشخصات عمودی AOD از LIDAR فضایی مانند Cloud-Aerosol Lidar با پلاریزاسیون متعامد (CALIOP)

و در نظر گرفتن مشخصات عمودی غلظت PM_{2.5}، دقت تخمین و همبستگی AOD-PM_{2.5} را بهبود داد.

- علاوه بر این، توزیع غلظت PM_{2.5} به دلیل منابع طبیعی و انسانی در مناطق مختلف بسیار متفاوت است و غلظت‌ها در فضا و زمان متفاوت است. به عنوان مثال، منابع طبیعی عمدتاً شامل نمک دریا، گرد و غبار، فوران آتشفشان، آتش‌سوزی جنگل‌ها و آتش‌سوزی در مراتع است و منابع انسانی عمدتاً شامل احتراق سوخت‌های فسیلی و فرآیندهای صنعتی است [20]، لذا با در نظر گرفتن این پارامترها و همچنین پارامترهای خصوصیات کاربری زمین، داده‌های رشد شهری و جمعیت شناختی و داده‌های انتشارات محلی ممکن است قابلیت پیش‌بینی غلظت PM_{2.5} را بهبود بخشد و می‌تواند در تحقیقات آینده گنجانده شود.

- داده‌های AOD تحت تاثیر پوشش ابری می‌باشد از آنجایی که در بعضی روزها و مکان‌ها این نویزها باعث از دست رفتن تصاویر ماهواره‌ای می‌شود، می‌توان با استفاده از الگوریتم‌های شبکه عصبی و درونیایی، پیکسل‌هایی که از دست رفته‌اند را با دقت مناسبی بازیابی کرد.

مراجع

- [1] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of PM_{2.5} pollution using air mass trajectory based geographic model and wavelet transformation," *Atmos. Environ.*, vol. 107, pp. 118–128, 2015.
- [2] X. Li and X. Zhang, "Predicting ground-level PM_{2.5} concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach," *Environ. Pollut.*, vol. 249, pp. 735–749, 2019.
- [3] F. Dominici, A. McDermott, M. Daniels, S. L. Zeger, and J. M. Samet, "Mortality among residents of 90 cities," *Spec. Rep. Revis. Anal. Time-Ser. Stud. Air Pollut. Health*, pp. 9–24, 2003.
- [4] C. A. Pope Iii et al., "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution," *Jama*, vol. 287, no. 9, pp. 1132–1141, 2002.
- [5] J. Lepeule, F. Laden, D. Dockery, and J. Schwartz, "Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009," *Environ. Health Perspect.*, vol. 120, no. 7, pp. 965–970, 2012.
- [6] W. R. W. Mahiyuddin, M. Sahani, R. Aripin, M. T. Latif, T.-Q. Thach, and C.-M. Wong, "Short-term effects of daily air pollution on mortality," *Atmos. Environ.*, vol. 65, pp. 69–79, 2013.
- [7] F. Dominici et al., "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases," *Jama*, vol. 295, no. 10, pp. 1127–1134, 2006.
- [8] F. M. Yang, Y. L. Ma, and K. B. He, "A brief introduction to PM_{2.5} and related research," *World Env.*, vol. 4, pp. 33–35, 2000.
- [9] Y. J. Kaufman, D. Tanré, and O. Boucher, "A satellite view of aerosols in the climate system," *Nature*, vol. 419, no. 6903, pp. 215–223, 2002.

- [10] N. Djebbari and M. Rouainia, "Artificial neural networks based air pollution monitoring in industrial sites," in *Engineering and Technology (ICET), 2017 International Conference on*, 2017, pp. 1–5.
- [11] M. R. Delavar *et al.*, "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 2, p. 99, 2019.
- [12] C. C. Aggarwal, "Neural networks and deep learning," *Springer*, vol. 10, pp. 978–3, 2018.
- [13] D. W. Morley and J. Gulliver, "A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment," *Environ. Model. Softw.*, vol. 105, pp. 17–23, 2018.
- [14] Z. Yang, C. Zdanski, D. Farkas, J. Bang, and Harris Williams, "Evaluation of Aerosol Optical Depth (AOD) and PM_{2.5} associations for air quality assessment," *Remote Sens. Appl. Soc. Environ.*, vol. 20, p. 100396, Nov. 2020, doi: 10.1016/j.rsase.2020.100396.
- [15] R. Li, T. Ma, Q. Xu, and X. Song, "Using MAIAC AOD to verify the PM_{2.5} spatial patterns of a land use regression model," *Environ. Pollut.*, vol. 243, pp. 501–509, Dec. 2018, doi: 10.1016/j.envpol.2018.09.026.
- [16] Q. Wang *et al.*, "Estimating PM_{2.5} concentrations based on MODIS AOD and NAQPMS data over Beijing–Tianjin–Hebei," *Sensors*, vol. 19, no. 5, p. 1207, 2019.
- [17] X. Wang, J. Yuan, and B. Wang, "Prediction and analysis of PM_{2.5} in Fuling District of Chongqing by artificial neural network," *Neural Comput. Appl.*, vol. 33, no. 2, pp. 517–524, 2021.
- [18] H. Zhang and S. Kondragunta, "Daily and hourly surface PM_{2.5} estimation from satellite AOD," *Earth Space Sci.*, vol. 8, no. 3, p. e2020EA001599, 2021.
- [19] M. Ghahremanloo, Y. Choi, A. Sayeed, A. K. Salman, S. Pan, and M. Amani, "Estimating daily high-resolution PM_{2.5} concentrations over Texas: Machine Learning approach," *Atmos. Environ.*, vol. 247, p. 118209, 2021.
- [20] X. Ni, C. Cao, Y. Zhou, X. Cui, and R. P. Singh, "Spatio-temporal pattern estimation of PM_{2.5} in Beijing-Tianjin-Hebei Region based on MODIS AOD and meteorological data using the back propagation neural network," *Atmosphere*, vol. 9, no. 3, p. 105, 2018.
- [21] "Approximation theory of the MLP model in neural networks | Acta Numerica | Cambridge Core." <https://www.cambridge.org/core/journals/acta-numerica/article/abs/approximation-theory-of-the-mlp-model-in-neural-networks/18072C558C8410C4F92A82BCC8FC8CF9> (accessed Oct. 25, 2022).
- [22] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *2016 International Conference on Computing Communication Control and automation (ICCUBE)*, Pune, Aug. 2016, pp. 1–7. doi: 10.1109/ICCUBE.2016.7860040.
- [23] S. Abdullah, M. Ismail, and S. Y. Fong, "Multiple linear regression (MLR) models for long term PM₁₀ concentration forecasting during different monsoon seasons," *J. Sustain. Sci. Manag.*, vol. 12, no. 1, pp. 60–69, 2017.
- [24] Y. Sathe, S. Kulkarni, P. Gupta, A. Kaginalkar, S. Islam, and P. Gargava, "Application of Moderate Resolution Imaging Spectroradiometer (MODIS) Aerosol Optical Depth (AOD) and Weather Research Forecasting (WRF) model meteorological data for assessment of fine particulate matter (PM_{2.5}) over India," *Atmospheric Pollut. Res.*, vol. 10, no. 2, pp. 418–434, 2019.
- [25] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," *Ecol. Inform.*, vol. 6, no. 3–4, pp. 228–241, Jul. 2011, doi: 10.1016/j.ecoinf.2010.12.003.
- [26] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA, USA: A Bradford Book, 1992.
- [27] L. Rarità, I. Stamova, and S. Tomasiello, "Numerical schemes and genetic algorithms for the optimal control of a continuous model of supply chains," *Appl. Math. Comput.*, vol. 388, p. 125464, 2021.

- [28] M. Kaveh and M. S. Mesgari, "Improved biogeography-based optimization using migration process adjustment: An approach for location-allocation of ambulances," *Comput. Ind. Eng.*, vol. 135, pp. 800–813, 2019.
- [29] O. Rostami and M. Kaveh, "Optimal feature selection for SAR image classification using biogeography-based optimization (BBO), artificial bee colony (ABC) and support vector machine (SVM): a combined approach of optimization and machine learning," *Comput. Geosci.*, vol. 25, no. 3, pp. 911–930, 2021.