پیش بین غلظت آلاینده $PM_{2.5}$ با استفاده از شبکه ترکیبی (ANN-GA) مطالعه موردی : شهر ارومیه

محمد طایفه طاهرلو [۱], امیر اسدی وایقان [۲*]

۱- دانشجوی کارشناسی ارشد، گروه عمران، دانشکده فنی و مهندسی، دانشگاه ارومیه

۲*- استادیار، گروه عمران، استادیار، گروه عمران، دانشکده فنی و مهندسی، دانشگاه ارومیه

ایمیل نویسنده مسئول: a.asadi@urmia.ac.ir

تاریخ دریافت: ۱۴۰۲/۰۶/۳۰ تاریخ پذیرش:۱۴۰۲/۰۸/۰۱

چکیده

به دلیل اهمیت مشکلات مربوط به محیط زیست و سلامتی که ناشی از آلودگی هوا است، روش های پیش بینی آلاینده ها به عنوان یک ابزار مهم در تحقیقات مربوط به آلودگی هوا مد نظر بوده اند. در میان آلاینده های مختلف اثرگذار بر کیفیت هوا، ذرات با قطر آیرودینامیکی کمتر از ۲/۵ میکرومتر ($PM_{2.5}$) یکی از مسائل اصلی در مدیریت کنترل آلودگی هوا هستند. در این مطالعه، شبکه های عصبی مصنوعی (ANN) در ترکیب با الگوریتم ژنتیک (GA)، برای پیش بینی ذرات $PM_{2.5}$ در یک دوره ی کوتاه مدت در شهر ارومیه، استفاده شده اند. از فیلتر Savitzky-Golay (SG) جهت پیش پردازش و هموار سازی داده های ایستگاه انداز ه گیری ذرات $PM_{2.5}$ استفاده گردید. دو روش پرکردن شکاف داده ها (روش های KNN و SPLINE) به منظور به حداقل رساندن انحراف آموزشی و بهبود دقت شبکه به کار گرفته شده اند. داده‌های PM_{10}، $PM_{2.5}$، دی اکسید نیتروژن، دی اکسید گوگرد ، مونوکسید کربن و داده های هواشناسی نیز برای این پیش بینی ها استفاده شده اند. طبق نتایج به دست آمده، روش ANN-GA (ترکیب روش های شبکه عصبی مصنوعی و الگوریتم ژنتیک)، یک بهبود ۴۰ درصدی در همبستگی نتایج پیش بینی نسبت به روش شبکه عصبی مصنوعی ارائه داد. خطای MSE ۰/۰۰۱ (در مقیاس ۱-۰) و ضریب همبستگی R، به مقدار ۰/۹۱ در پیش بینی مشاهده گردید.

کلمات کلیدی

" پیش بینی آلودگی هوا "، "شبکه عصبی مصنوعی"، "آلودگی هوا"، "الگوریتم ژنتیک"، "$PM_{2.5}$ "

۱- مقدمه

در ۵۰ سال گذشته، فعالیت هایی مانند شهرنشینی، صنعتی شدن و رشد جمعیت، هوا را به بخشی جدایی ناپذیر از زندگی ما تبدیل کرده اند (Harishkumar et al., ۲۰۲۰). آلودگی هوا را می توان به عنوان وجود مواد شیمیایی یا ترکیبات سمی در هوا تعریف کرد تا جایی که برای سلامتی خطرناک باشد. انتشار گاز اتومبیل ها، مواد شیمیایی گیاهی، گرد و غبار، گرده و هاگ های کپک به عنوان ذرات معلق (PM) معرفی می شوند. به گزارش سازمان بهداشت جهانی، آلودگی هوای محیط باعث مرگ ۴/۲ میلیون نفر در اثر سکته مغزی، بیماری های قلبی، سرطان ریه و بیماری های مزمن تنفسی می شود. از بین آلاینده های مختلف موثر بر کیفیت هوا، ذرات معلق کمتر از ۱۰ میکرون مشکل اصلی آلودگی هوا هستند (Scibor et al., ۲۰۲۰). همچنین، شواهد فزآینده‌ای از اثرات PM_{10} و $PM_{2.5}$ بر بیماری‌های قلبی عروقی (CVD) و بیماری‌های تنفسی (DR) وجود دارد (Samoli et al., ۲۰۰۵؛ Schwartz et al., ۲۰۱۲). با توجه به اهمیت آلاینده های معیار و رفاه عمومی، ایستگاه های پایش در نقاط مختلف شهر برای اندازه گیری لحظه ای آلاینده ها وجود دارد. داده های نظارت بلند مدت باید برای پیش بینی توسط مدل های مختلف جمع آوری شود. پیش بینی آلاینده های هوا فرصتی را برای تعیین شدت آلودگی هوا در مناطق مختلف و جلوگیری از اثرات غیرقابل بازگشت فراهم می کند. علاوه بر این، این مدل‌ها به تصمیم گیرندگان اجازه می‌دهند تا برای پیشگیری یا کنترل PM در آینده آماده شوند. (Zhang et al., ۲۰۱۹). پیش پردازش داده یک فرایند آماده سازی داده های خام و مناسب کردن آن برای یک مدل یادگیری ماشین است. این اولین و مهمترین گام در ایجاد یک مدل یادگیری ماشین است. در طول دهه‌های گذشته، بسیاری از روش‌های جدید برای بازسازی داده‌های سری زمانی پیوسته توسعه یافته‌اند. این روش‌ها را می‌توان به چهار نوع طبقه‌بندی کرد: (۱) روش‌های مبتنی بر زمانی، (۲) روش‌های مبتنی بر فرکانس، (۳) روش‌های ترکیبی، و (۴) روش‌های همجوشی چند منبع. در این مقاله به دلیل محبوبیت گسترده و روش‌های تثبیت شده در این زمینه، بر روی روش‌های مبتنی بر زمانی تمرکز دارد. روش‌های مبتنی بر زمانی را می‌توان بیشتر به چهار دسته تقسیم کرد: (الف) روش‌های درونیابی-جایگزینی زمانی، (ب) روش‌های فیلتر زمانی، (ج) روش‌های برازش عملکرد زمانی و (د) مدل های یادگیری عمیق زمانی (Liang et al., ۲۰۲۳). در میان روش های فیلتر زمانی، فیلتر Savitzky-Golay در سال ۱۹۶۴ معرفی و در بسیاری از زمینه های پیش پردازش و بازسازی داده ها بکار رفته است. فیلترهای SG معمولا برای نقاط داده ای هم فاصله اعمال می شوند و بر اساس برازش چند جمله ای درجه n در یک همسایگی (معمولا متقارن) k-m تا k+m از هر نقطه داده k (این محدود شامل ۲m+۱ نقطه داده است) استوار است (Luo et al., ۲۰۰۵ ؛ Schmid et al., ۲۰۲۲). SG همچنین می تواند روی مجموعه ای از نقاط داده اعمال شود تا آنها را روان کند (صاف کند) و دقت داده ها را بدون به خطر انداختن جهت گیری آنها افزایش دهد (Luo et al., ۲۰۰۵). از ویژگی های مهم این فیلتر که باعث شده است در زمینه های مختلف به کار برده شود می توان به ۱) سادگی و اجرای آسان آن ۲) تاثیر پذیری کمتر عملکرد فیلتر ناشی از نویز های باقیمانده (Residual Noise) و ۳) بهبود بازسازی مقادیر داده های

مفقود شده در سری زمانی اشاره کرد (Chen et al., ۲۰۲۱). Kandasamy و همکاران در سال ۲۰۱۳ به بررسی مقایسه ای روش های هموار سازی و پر کردن شکاف داده های ماهواره MODIS پرداختند. روش های مورد مقایسه به ترتیب شامل (iterative caterpillar singular spectrum analysis (ICSSA))، empirical mode decomposition (EMD)، low pass filtering (LPF) و Whittaker smoother (Whit) بودند. بر اساس نتایج تحقیق، فیلتر SG برای سری های زمانی دارای شکاف بیش از ۳۰ درصد عملکرد بهتری نسبت به سایر روش ها داشته است. همچنین این فیلتر در پیش بینی زمان بندی فنولوژیکی سری زمانی دارای شکاف بیش از ۶۰ درصد دقت بهتری داشته است. در سال ۲۰۲۱ Chen و همکاران از فیلتر (SG) جهت بازسازی داده های سری زمانی ماهواره لندست (NDVI) استفاده کردند. نتایج تحقیق نشان داد که این فیلتر در مقایسه با سایر روش های معمول بازسازی داده ها (IFSDAF، STAIR و Fill-and-Fit) عملکرد بهتری دارد.

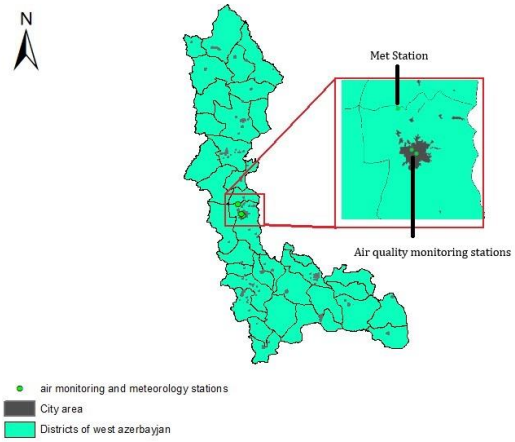
برخی از مدل‌های مورد استفاده در مطالعات پیش‌بینی آلودگی هوا عبارتند از میانگین متحرک یکپارچه رگرسیون خودکار (ARIMA)، شبکه عصبی مصنوعی (ANN)، مدل کیفیت هوای چند مقیاسی جامعه (CMAQ)، مدل تحقیق و پیش‌بینی آب و هوا (WRF) همراه با شیمی(WRF-CHEM)، مدل‌های فازی، مدل‌های خاکستر و/یا مدل‌های ترکیبی (Farhadi et al., ۲۰۲۰). روش ANN به طور گسترده توسط دانشمندان برای ارائه راه حل های سریع و مقرون به صرفه برای کاهش اثرات منفی آلودگی هوا در سراسر جهان استفاده شده است (Biancofiore et al., ۲۰۱۷؛ Cabaneros et al., ۲۰۱۷؛ Coman et al., ۲۰۰۸؛ Ibarra-Berastegi et al., ۲۰۰۸). شبکه‌های عصبی به عنوان یک جایگزین، با موفقیت در پیش‌بینی آلودگی هوا مورد استفاده قرار گرفته‌اند و نتایج دقیقی را در داده‌های سری زمانی تولید کرده‌اند. انواع مختلف نویز و ساختار غیرخطی در داده ها وجود داشت (Elangasinghe et al., ۲۰۱٤). در دو دهه گذشته، الگوریتم ژنتیک (GA)، که یک روش جستجو برای شبیه سازی مکانیزم تکامل ژنتیکی است، یک تکنیک بهینه‌سازی محبوب و پیشرفته بوده است (Dede et al., ۲۰۱۱). GA با تکامل مبتنی بر جمعیت، بقای بهترین ها، هدایت تصادفی و عدم اتکا به اطلاعات گرادیان مشخص می شود (Ding et al., ۲۰۱۱). GA یک فرآیند محاسباتی تکراری است که شامل رمزگذاری، مقداردهی اولیه جمعیت، انتخاب، عملیات ژنتیکی، ارزیابی و تصمیم توقف است (Salsedo-Sanz, ۲۰۰۶). رویکردهای مدل‌سازی ترکیبی کاربردهای گسترده‌ای دارند که در آن روش‌ها یا ویژگی‌های متعددی با هم ادغام می‌شوند تا مدلی پیچیده‌تر با عملکرد برتر در سناریوهای خاص ایجاد کنند (Dede et al., ۲۰۱۱). به گفته (Asghari & Nematzadeh, ۲۰۰٦)، عملکرد (سرعت دستیابی به راه حل های بهتر) و دقت نتایج را می توان توسط ادغام یک شبکه عصبی با الگوریتم ژنتیک (ANN-GA) بهبود بخشید. GA راه‌حل‌های مختلفی را برای یک موضوع معین تولید می‌کند و آنها را در چندین نسل توسعه می‌دهد. هر راه حل شامل تمام پارامترهایی است که می توانند نتایج را بهبود بخشند. در نتیجه، تمام وزن ها در ANN در یک راه حل GA قرار می گیرند (de Mattos Neto et al., ۲۰۱۳؛ Antanasijević et al., ۲۰۱٤؛ Zhao et al, ۲۰۱۰). برای پیش بینی $PM_{۲,۵}$، (Kow et al., ۲۰۲۰) یک مدل ترکیبی را ایجاد کردند که انتشار پس‌انداز را با شبکه‌های عصبی حلقه ای (دارای لوپ) ترکیب می‌کند. یافته ها نشان داد که مدل ترکیبی از نظر کیفیت و دقت از شبکه های منفرد بهتر عمل می کند. این مدل همچنین می تواند برای آزمایش مکانیسم های انتشار ذرات استفاده شود. (Delavar et al., ۲۰۱۹) از یک مدل ترکیبی متشکل از یک شبکه عصبی مصنوعی و یک مدل خود رگرسیون غیرخطی برای پیش‌بینی آلودگی هوای شهر تهران استفاده کردند. نتایج نشان داد که خروجی های الگوریتم ژنتیک بهینه شده به طور قابل توجهی دقیق هستند. سه مدل برای پیش‌بینی کیفیت هوا در لانژو چین در سال ۲۰۱۷ استفاده شد: BP، GA و BP-GA. با توجه به نتایج، BP-GA دارای سطح بالایی از تعمیم و قابلیت جستجوی جهانی و همچنین عملکرد پیش بینی بهتری بوده است (Qu & Kang, ۲۰۱۷). (Esfandani & Nematzadeh, ۲۰۱۵)، یک تکنیک ترکیبی BP-GA را برای پیش‌بینی $PM_{۱۰}$ در تهران معرفی کردند. بر خلاف ANN، BP-GA دارای ۵۴/۹ درصد عملکرد بهتر از شبکه عصبی مصنوعی بوده است. علاوه بر این، آنها کشف کردند که BP-GA نتایج دقیق تری را برای دوره های زمانی کوتاه تر به دلیل نوسانات داده های قابل توجه در بلندمدت ارائه می دهد که بر عملکرد شبکه تأثیر منفی می گذارد . در آذرماه سال ۱۳۹۶ مطالعه‌ای در مورد آلاینده‌های مسئول در بحران آلودگی هوای شهر ارومیه و همچنین منشأ پیدایش آنها توسط نوری و همکاران مورد بررسی قرار گرفت. نتایج حاصل از این مطالعه نشان داد که آلاینده مسئول بحران آلودگی هوای شهر ارومیه در زمان مورد بررسی، $PM_{۲,۵}$ می‌باشد. همچنین تردد خودروها و ترافیک شهری به عنوان سناریو اصلی در انتشار این آلاینده معرفی شد (نوری و همکاران، ۱۳۹۶).در این مطالعه، ANN-GA به همراه داده‌های گمشده برای پیش‌بینی $PM_{۲,۵}$ در ارومیه، استفاده شد تا نشان دهد چگونه پر کردن شکاف داده‌ها و روش‌های پیش پردازش می‌توانند عملکرد مدل‌های ترکیبی را بهبود بخشد.

۲- روش انجام تحقیق

• محدوده مورد مطالعه

شهر ارومیه مرکز استان آذربایجان غربی و واقع در شمال غربی ایران است. این شهر توسط کوه هایی از قبیل کوهسیر، قیز قلعه (قلعه دختر)، کوه جهودان، کوه چهل مر شهیدان احاطه شده است. ارومیه به دلیل ترافیک های پی در پی و پر ازدحام، افزایش سطوح $CO_۲$ و PM و کمبود دانش کافی در تنظیم و مکان یابی واحدهای تولیدی صنعتی، یکی از آلوده ترین شهرهای ایران به حساب می آید. گرد و غبار برخواسته از عراق که بر منطقه تأثیر منفی می گذارد و همچنین وارونگی که ۹۰ روز در سال رخ می دهد، نمونه هایی از آلودگی هوای مختص این منطقه است. علاوه بر این، خشک شدن دریاچه ارومیه که می تواند منجر به طوفان نمک شود، یکی از نگرانی های حیاتی است

که در آینده نزدیک به آلودگی قابل توجهی منجر خواهد شد (نوری و همکاران, ۱۳۹۶). در شکل ۱ موقعیت منطقه مطالعاتی به همراه ایستگاه های هواشناسی و پایش کیفیت هوا در شهر ارومیه نشان داده شده است.



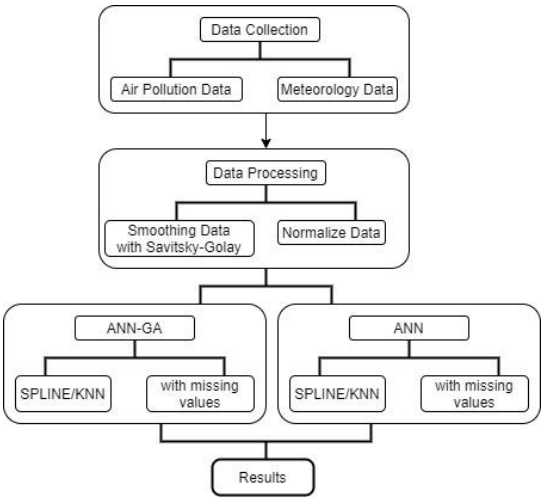
شکل ۱- شهر مورد مطالعه و موقعیت ایستگاه های هواشناسی و پایش کیفیت هوا

- **داده های هواشناسی و غلظت آلاینده ها**

غلظت آلاینده های هوا (مونوکسید کربن، دی اکسید نیتروژن و دی اکسید گوگرد) و همچنین داده های هواشناسی (دما، رطوبت نسبی و سرعت باد) به عنوان ورودی در این تحقیق برای پیش بینی $PM_{2.5}$ مورد استفاده قرار گرفتند. غلظت آلودگی هوا و داده های هواشناسی طی یک دوره دو ساله از ایستگاه پایش شماره ۳ شهرداری ارومیه و سایت هواشناسی ایران (Data.irimo.ir) اخذ شد.

- **مراحل پیش بینی غلظت ذرات $PM_{2.5}$**

در شکل ۲ مراحل انجام پیش بینی غلظت ذرات $PM_{2.5}$ آورده شده است. مرحله اول شامل جمع آوری داده های آلاینده ها و هواشناسی، مرحله دوم شامل پیش پردازش داده ها (نرمالیزه کردن و هموار سازی) است که توسط فیلتر SG انجام گرفت. مرحله سوم نیز مدل سازی و پیش بینی توسط شبکه ANN و ANN-GA در دو حالت بدون پر کردن شکاف داده و با پر کردن شکاف داده توسط روشهای SPLINE و KNN است.



شکل ۲- مراحل پیش بینی آلودگی ذرات $PM_{2.5}$

- نرمال سازی داده ها

هنگامی که مقیاس یا دامنه اطلاعاتی داده های دریافتی مطابقت نداشت، از نرمال سازی داده ها برای یکسان سازی وزن داده ها بدون تغییر در ویژگی های اصلی استفاده می گردد. در نتیجه، شبکه هیچ حساسیتی نسبت به یک یا چند داده نشان نخواهد داد. از تابع "mapminmax" در نرم افزار متلب برای نرمال سازی در این تحقیق به دلیل سادگی استفاده گردید (رابطه ۱).

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (۱)$$

در رابطه (۱) X_n داده های نرمال شده، X_{min} کوچکترین مقدار داده های موجود، X_{max} بیشترین مقدار داده های موجود و X داده های اولیه هستند

- فیلتر ساویستکی -گولای (SG)

فیلتر SG در سال ۱۹۶۴ توسط یک شیمی دان به نام آبرام ساویتسکی و یک فیزیکدان به نام مارسل گولای ابداع گردید. این فیلتر به عنوان یک فیلتر پایین گذر عمل می کند، ساختاری شبیه به فیلترهای دارای پاسخ ضربه محدود (FIR) دارد و توسط دو پارامتر عملکرد آن کنترل می گردد. اولین پارامتر طول پنجره یا طول ناحیه برازش چند جمله ای و دومین پارامتر درجه چند جمله ای است. محبوبیت این فیلتر به دلیل ریاضیات و محاسبات ساده و کوتاه بوده و در زمینه های مختلف علوم کاربرد وسیعی دارد. از قبیل پردازش تصاویر سنجش از دور، پردازش سیگنال های دیجیتال، پردازش طیف اتمی و .... (خداقلی و باقری، ۱۳۹۸).

در این روش یک چند جمله ای به تعداد فرد از نمونه های داده برازش داده می شود. سپس مقدار این چند جمله ای به ازای مقدار میانی نمونه ها محاسبه و با مقدار نرم شده برای این نمونه مساوی قرار داده می شود. با جابجایی این زیر مجموعه از داد ها که به آن فاصله نرم کردن اطلاق می گردد، به اندازه یک نمونه و تکرار الگوریتم، مقدار نرم شده برای نمونه بعدی به دست می آید (خداقلی و باقری، ۱۳۹۸).

جهت نرم کردن و هموار سازی داده ها نمونه های داده ها که با $X[n]$ نمایش داده می شود در نظر گرفته می شود. ضرایب چند جمله ای $P(n)$ برازش داده شده از رابطه ۲ بدست می آید.

$$P(n) = \sum_{k=0}^{N} a_k n^k \quad (۲)$$

خطای برازش به روش کمترین مربعات از رابطه ۳ بدست می اید.

$$\varepsilon_N = \sum_{n=-M}^{M} (P(n) - x[n])^2 = \sum_{n=-M}^{M} (\sum_{k=0}^{N} a_k n^k - x[n])^2 \quad (۳)$$

در رابطه ۳، M نصف فاصله نرم کردن یا نصف فاصله برازش منحنی است. برای نمونه بعدی با جابجایی فاصله نرم کردن به اندازه یک نمونه و تکرار برازش چندجمله ای و محاسبه مقدار آن به ازای نقطه میانی بدست می آید. و این کار تا آخرین نمونه تکرار می شود.

- پر کردن شکاف داده ها

دو روش پر کردن داده ها، نزدیک ترین همسایه (KNN) و برازش منحنی غیرخطی (SPLINE)، به دلیل خرابی و خطاهای دستگاه ضبط داده استفاده شد. این تکنیک ها الگوریتم آموزشی را در پیش بینی پاسخ های دقیق تر قادر می سازد. هدف اصلی طبقه بندی، پیش‌بینی برچسب های نقاط داده آزمون با القای تمام نقاط داده آموزشی است (Lu et al., ۲۰۱٦). در روش استاندارد KNN، قانون اکثریت برای پیش‌بینی برچسب نقطه داده آزمایشی استفاده می‌شود. کلاس اصلی اکثر داده های آموزشی مشابه در فضای ویژگی برای پیش بینی برچسب استفاده می شود (Zhang et al., ۲۰۱۷). در مطالعات قبلی، KNN برای رسیدگی به مشکلات داده های گمشده به عنوان یک ابزار انتساب داده استفاده شده است (Zhang et al., ۲۰۰۷).

اسپلاین مکعبی یک تابع مکعبی تکه‌ای است که مجموعه‌ای از نقاط داده را درون‌یابی می‌کند و در عین حال نرمی را تضمین می‌کند. برازش چندجمله‌ای مکعبی به یک سری از داده های مشاهداتی، مبنایی برای انتساب است. برازش به گونه ای انجام می شود که تابع و دو مشتق اول آن در گره ها (جایی که مقاطع تکه ای به هم می پیوندند) پیوسته باشند (Junninen et al., ۲۰۰٤).

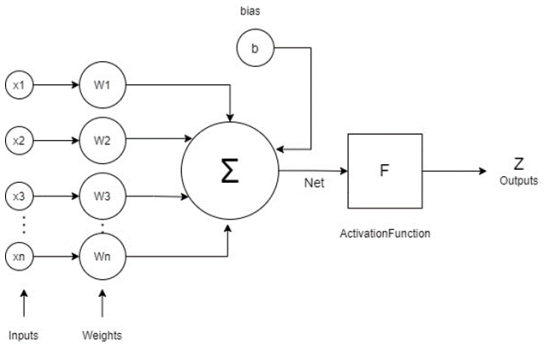
- **شبکه عصبی مصنوعی (ANN)**

ANN ها مدل های محاسباتی موازی عظیمی هستند که عملکرد مغز انسان را تقلید می کنند. یک ANN از پردازنده های ساده تشکیل شده است که توسط اتصالات وزنی به هم مرتبط شده اند (Dongare et al., ۲۰۱۲). مجموعه ای از نورون های مصنوعی مشتق شده از سلول های عصبی بیولوژیکی شبکه عصبی را تشکیل می دهند. هر نورون چندین ورودی و یک خروجی دارد که هر کدام وزنی دارند که در یک تابع خوب خطی به نام تابع انتقال جمع می‌شود (شکل۳). این نورون می تواند اطلاعات را از خروجی خود به نورون های دیگر ارسال کند. خروجی نهایی با آخرین لایه نورون یکسان است. یک شبکه پیشخور با اتصالات یک طرفه توسط نورون ها تشکیل می شود. با توپولوژی های شبکه پیشخور (FFN) حداقل سه لایه (ورودی، پنهان و خروجی) تشکیل می گردد که پرسپترون چند لایه (MLP) رایج ترین و موفق ترین نوع معماری شبکه عصبی است (Ordieres et al., ۲۰۰۵) شبکه یک تابع فعال سازی غیر خطی را در لایه های مخفی و خروجی به کار می گیرد. MLP اغلب تقریب هایی را برای مسائل بسیار پیچیده و پویا ارائه می دهد (Rosenblatt, ۱۹٦۱). رابطه (۴) فرآیندهای ریاضی هر نورون در شبکه را نشان می دهد.

$$Z = f(net)$$
$$net = b + W_1x_1 + W_2x_2 + W_3x_3 + \ldots + W_nx_n \quad (۴)$$
$$= b + \sum W_ix_i$$

که در X ورودی ... ال،

در این مطالعه، یک سیستم واحد شامل دو لایه پنهان و یک لایه خروجی می باشد. برای معرفی داده ها به شبکه از روش سری زمانی استفاده شد. از آنجا که شبکه عصبی پس از هر اجرا پاسخ های متفاوتی می دهد؛ مقدار نسبت داده‌های آموزشی، آزمایش و اعتبارسنجی، از طریق آزمون و خطا انجام شد و فرآیند معیار ارزیابی در صورت لزوم اصلاح شد. سناریوهای وارد کردن داده ها به دو صورت تعریف شدند. سناریوی اول از هیچ انتسابی استفاده نمی کرد، در حالی که سناریوی دوم از SPLINE و KNN برای پر کردن شکاف های داده استفاده می کرد. به عنوان یک تابع انتقال، یک لایه سیگموئید (logsig) برای لایه های پنهان و یک لایه خطی (Purelin) برای لایه خروجی استفاده شد. الگوریتم لونبرگ–مارکوارت بر اساس نوع مسئله و سرعت همگرایی به عنوان الگوریتم یادگیری انتخاب شد. برای بهبود نتایج، تعداد نورون‌ها، پارامترهای تکرار، تعداد ارزیابی‌های مجاز، پارامترهای الگوریتم لونبرگ و قابلیت اطمینان، همه از طریق فرآیند آزمون و خطا تنظیم شدند.



شکل ۳- ساختار شبکه عصبی مصنوعی (ANN) (Haykin, ۲۰۰۹)

- **الگوریتم ژنتیک (GA)**

الگوریتم‌های تکاملی روش‌های جستجوی تصادفی هستند که فرآیند تکامل طبیعی را برای حل مسائل بهینه‌سازی تقلید می‌کنند (Booker, ۱۹۸۹). این الگوریتم افراد را ارزیابی می کند تا کیفیت راه حل ها را با یک تابع هدف مانند هزینه و زمان منحصر به فرد بهبود دهند. افراد با عملکرد بهتر به عنوان والدین نسل بعدی افراد انتخاب می شوند. GA با استفاده از عملگرهایی که از نو ترکیبی جنسی (متقاطع) و جهش در موجودات طبیعی الهام گرفته اند، افراد جدیدی ایجاد می کند. راه حل های جدید ارزیابی می شوند و چرخه انتخاب و ایجاد افراد جدید تا زمانی که راه حل رضایت بخشی پیدا یا یک محدودیت زمانی از پیش تعیین شده سپری شود، تکرار می شود.

به طور خاص، برای الگوریتم ژنتیک، مراحل زیر باید تایید شوند:

۱- تولید جمعیت اولیه و ارزیابی آن.

۲- بهترین والدین را از مرحله ۱ انتخاب کرده و ترکیب کنید تا (تقاطع) کودک را به دست آورید.

۳- چند نفر را برای جهش انتخاب کنید و افراد جهش یافته بسازید.

۴- بهترین افراد را برای مراحل ۱، ۲و ۳ انتخاب کنید.

۵- اگر راه حل ها مطابق با میل ما نیستند، مرحله ۲ را دوباره انجام دهید.

تقاطع روش اصلی برای ایجاد راه حل های جدید است (Cantu-Paz & Kamath, ۲۰۰۵). در این مقاله از GA با سه عملگر متقاطع تک نقطه ای، چند نقطه ای و یکنواخت استفاده شده است (جدول ۱).

جدول ۱- مشخصات شبکه ANN استفاده شده

نوع شبکه عصبی	پرسپترون چندلایه
تعداد لایه پنهان	۲
الگوریتم آموزش جهت تنظیم وزن و بایاس شبکه	پس انتشار لونبرگ-مارکوارت
تابع انتقال	در لایه پنهان logsig و در خروجی purelin
تعداد نورون هرلایه	۵و۱۰
تعداد تکرار	۲۰۰
درصد تقسیم داده	آموزش ۷۰٪
	ارزیابی ۱۵٪
	امتحان ۱۵٪
ارزیابی مدل	R و mse

- **شبکه ترکیبی (ANN-GA)**

دو راه برای راه اندازی شبکه های ترکیبی وجود دارد: استفاده از الگوریتم های ژنتیک برای آموزش شبکه ها یا استفاده از GA برای مشخص کردن شبکه های منتخب. هدف از آموزش شبکه عصبی یافتن مجموعه ای از وزن ها است که خطاهای اندازه گیری را به حداقل می رساند. وزن‌های شناسایی‌شده توسط GA بدون اصلاح بیشتر در روش آموزشی در شبکه استفاده می‌شوند (Caudell, Montana, ؛Fogel et al., ۱۹۹۰ ؛Dolan & ۱۹۸۹ ؛Davis & ۱۹۸۹ ؛Hanson & Whitley, ۱۹۸۹). در این مطالعه از GA به عنوان تابع آموزشی استفاده شد. پس از معرفی داده ها به عنوان سری زمانی و انتخاب میزان داده برای هر قسمت از یادگیری، ارزیابی و تست، ساختار و تعداد لایه های شبکه با تابع "newff" ایجاد شد. تفاوت اصلی این است که فرآیند یادگیری ژنتیکی به جای عملکرد "آموزش" استفاده می شود. شایان ذکر است که ویژگی های لایه شبکه در هر دو روش یکسان می باشد. برای یادگیری نحوه تکمیل فرآیند، تابع یادگیری جدید به چندین فرآیند جانبی از جمله ایجاد تابع هزینه، انتخاب، تقاطع و جهش نیاز دارد. در فرآیند انتخاب از دو روش انتخاب رولت و انتخاب مسابقات استفاده شده است. برای معرفی تابع هزینه، وزن‌ها توسط تابع "newff" ایجاد می‌شوند . مقادیر متفاوتی به متغیرهای جمعیت اولیه، حداکثر تعداد جهش و ضریب فشار انتخاب توسط روش آزمون و خطا اختصاص داده شده است. علاوه بر این دو سناریوی ورود داده (با و بدون انتساب) تعریف شدند. در جدول ۲ مشخصات شبکه ترکیبی ANN-GA استفاده شده در این مطالعه آورده شده است.

جدول ۱- مشخصات شبکه ترکیبی ANN-GA

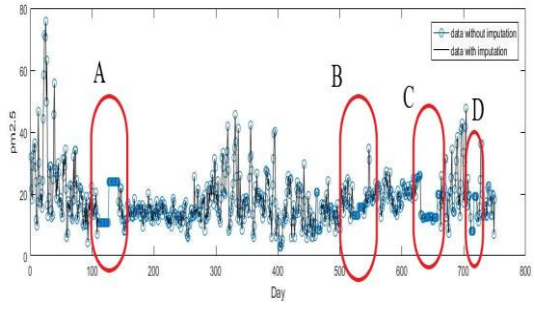
نوع شبکه عصبی	پرسپترون چندلایه
الگوریتم بهینه سازی	ژنتیک
تعداد لایه پنهان	۲
تکرار عملیات بهینه سازی	۳
نسبت تقاطع	۰/۸
نسبت جهش	۰/۴
فرآیند انتخاب	RWS،TS،Random
مقدار پارامتر جهش	۰/۱
مقدار فشار انتخاب(در صورت استفاده از RWS)	۱۰
تعداد جمعیت رقابت(در صورت استفاده از TS)	۳
تابع انتقال	در لایه پنهان logsig و در خروجی purelin
تعداد نورون هرلایه	۵و۱۰
تعداد تکرار	۲۰۰
درصد تقسیم داده	آموزش ۷۰٪
	ارزیابی ۱۵٪
	امتحان ۱۵٪
ارزیابی مدل	R و mse

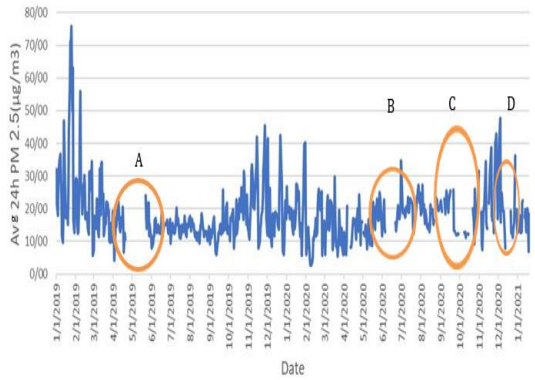
۳- نتایج

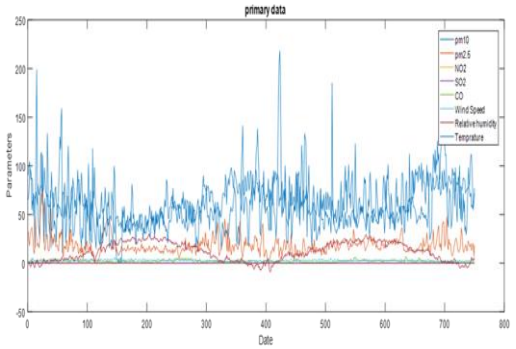
هموارسازی و پرکردن شکاف داده ها

شکل ۴ میانگین داده های روزانه $PM_{۲,۵}$ در یک دوره دو ساله را نشان می دهد. شکاف های داده در قسمت های A، B، C،D نشان داده شده است (به ترتیب A در بازه زمانی اوسط ماه چهارم تا ماه ششم ۲۰۱۹، B در بازه زمانی ماه ششم تا هفتم ۲۰۲۰، C در بازه زمانی ماه دهم تا یازدهم ۲۰۲۰ و D در بازه زمانی ماه دوازدهم ۲۰۲۰ تا ماه یکم ۲۰۲۱) . همچنین قبل از پر کردن شکاف داده ها از فیلتر SG برای هموارسازی تغییرات نامنظم و ناگهانی در داده های ورودی استفاده شد. نتایج استفاده از این فیلتر در شکل ۵ آورده شده است. پس از اعمال فیلتر SG بر روی داده های ورودی شامل غلظت آلاینده های هوا (مونوکسید کربن، دی اکسید نیتروژن و دی اکسید گوگرد) و همچنین داده های هواشناسی (دما، رطوبت نسبی و سرعت باد) نوسانات شدید در داده ها صاف تر گردیده است، اما ویژگی های آنها دست نخورده باقی مانده است. در نتیجه، قابلیت اطمینان پیش بینی شبکه و ضرایب خطای پیش‌بینی بهبود یافته است. در واقع با اعمال فیلتر SG بر روی داده های خام، یک الگوریتم آموزشی می تواند دقت قابل قبولی در پیش بینی ارائه دهد. شکل ۶ نشان می دهد که چگونه شکاف داده های روزانه $PM_{۲,۵}$

شکل ۴- مقادیر متوسط روزانه $PM_{۲,۵}$ در دوره دو ساله (۲۰۱۹-۲۰۲۱)

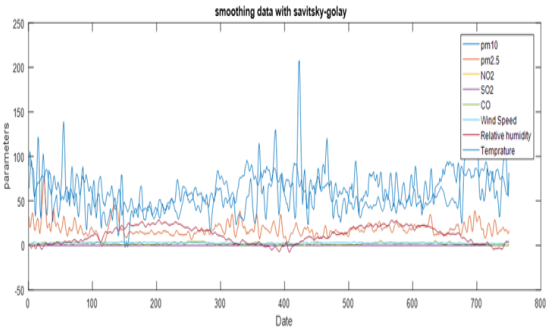






شکل ۶- نتایج پر کردن شکاف داده های $PM_{۲,۵}$ توسط دو روش به ترتیب KNN و SPLINE

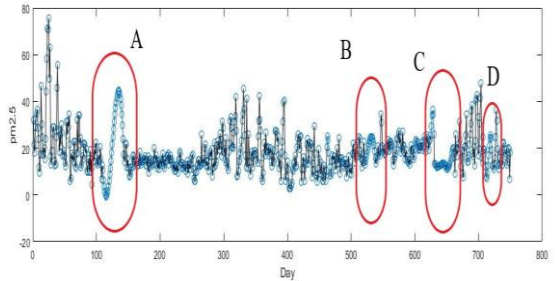
قسمت های A، B، C،D به درستی با ضرایب همبستگی بالا پر شده اند (۹۶ درصد برای SPLINE و ۸۹ درصد برای KNN). روش SPLINE در مقایسه با روش KNN عملکرد بهتری داشته است. دلیل عملکرد بهتر روش SPLINE استفاده از چند جمله ای های تکه ای درجه پایین است. در این روش این چند جمله ای ها تکه ای طوری انتخاب می گردند که باعث برازش هموارتر به داده ها گردد (Cho et al., ۲۰۲۰). اما در روش KNN از میانگین وزنی k داده مشابه استفاده می گردد که در صورت توزیع یکنواخت داده ها این برازش دقیق تر خواهد بود (Cho et al., ۲۰۲۰). در مورد داده های سری زمانی $PM_{۲,۵}$ به دلیل عدم توزیع یکنواخت داده ها، این روش پر شدگی ضریب همبستگی کمتری را نشان داده است (۹۶٪ روش SPLINE و ۸۹٪ روش KNN).

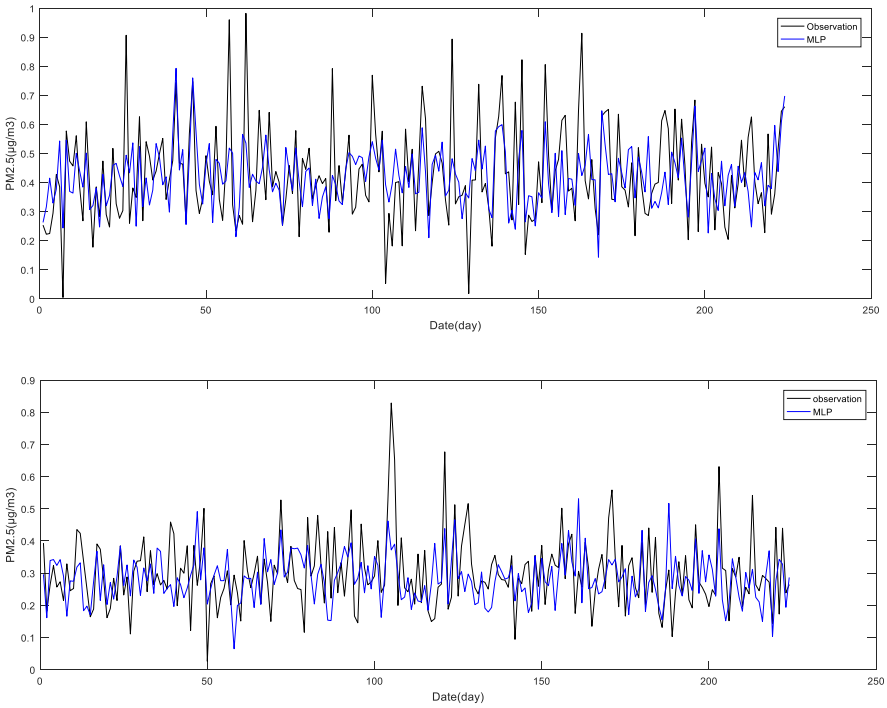


شکل ۵- داده های اولیه و هموار شده (نرم شده) توسط فیلتر SG

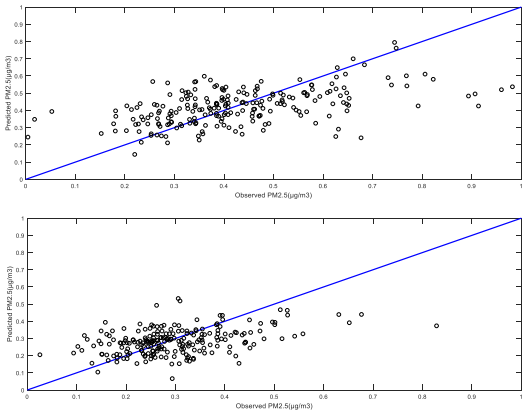
نتایج شبکه عصبی مصنوعی ANN

در این تحقیق در ابتدا غلظت $PM_{۲,۵}$ با روش شبکه عصبی مصنوعی (ANN) و با استفاده از دو روش پر کردن شکاف داده SPLINE و KNN پیش بینی گردید. جهت ارزیابی بهتر شبکه ANN با دو روش پر کردن شکاف داده ها ۵۰ تکرار انجام گرفت. بر اساس نتایج بدست آمده در روش SPLINE میانگین خطای MSE ۰/۰۱۴۹و ضریب همبستگی R، ۵۳/۶ ٪ می باشد. همچنین میانگین خطای MSE برای روش KNN برابر ۰/۰۲۳۰ و میانگین ضریب همبستگی R، ۵۴/۳ ٪ بدست آمد. اشکال ۷ و ۸ نتایج پیش بینی غلظت $PM_{۲,۵}$ را با استفاده از توابع SPLINE و KNN نشان می دهند. همانطور که مشاهده می شود، ضریب همبستگی پیش بینی هر دو روش دارای اختلاف حدود ۰/۷ ٪ بوده و این امر نشان دهنده پیش بینی نزدیک دو روش می باشد. اگر چه براساس میانگین خطای MSE، خطای روش SPLINE نسبت به روش KNN کمتر بوده است. همچنین با توجه به شکل ۸، پراکندگی پیش بینی در روش KNN بیشتر از روش SPLINE است. علت عملکرد بهتر روش SPLINE را می توان به نحوه پر کردن شکاف داده ها در این روش نسبت داد. با توجه به ضریب همبستگی های بدست آمده از این تحقیق می توان نتیجه گرفت که ANN حتی در صورت بهبود توسط روش های پر کردن (SPLINE و KNN) در پیش بینی غلظت $PM_{۲,۵}$ عملکرد خوبی نداشته است (ضرایب همبستگی ۵۳/۶ و ۵۴/۳). Asghari و Nematzadeh در سال ۲۰۱۶ به این نتیجه دست یافتند که شبکه ANN به تنهایی در پیش بینی غلظت $PM_{۱۰}$ دارای ضریب همبستگی ۵۷٪ بوده است و با ترکیب این شبکه با الگوریتم ژنتیک GA نتایج پیش بینی بهبود یافته است. همچنین نتایج تحقیق دیگر نشان داد که (BPNN) به دلیل اینکه نمی تواند به طور مؤثرتر و عمیق‌تری اطلاعات مفید را از مجموعه داده‌های با ابعاد بالا (الگوهای داده های ورودی-خروجی) یاد بگیرد و استخراج کند عملکرد ضعیف تری نسبت به شبکه ترکیبی CNN-BP داشته است (مقادیر $R^۲$ ۰/۵۶ برای BPNN در مقایسه با ۰/۸۰ برای (CNN-BP (Kow et al., ۲۰۲۰). Zaini و همکاران در سال ۲۰۲۲ نیز اذعان داشته اند که روش ANN به تنهایی به دلیل محدود بودن در حل مجموعه داده های سری زمانی غیر خطی بزرگ و عدم رهگیری موثر توزیع ویژگی های مجموعه داده های کیفیت هوا عملکرد ضعیف تری در پیش بینی $PM_{۲,۵}$ نسبت به شبکه های ترکیبی دارند.



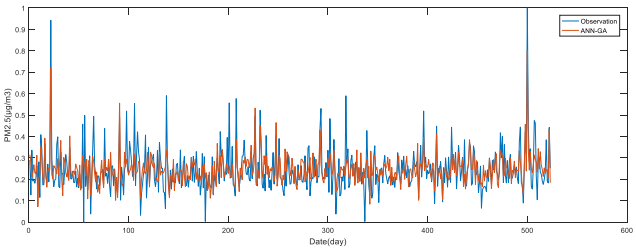


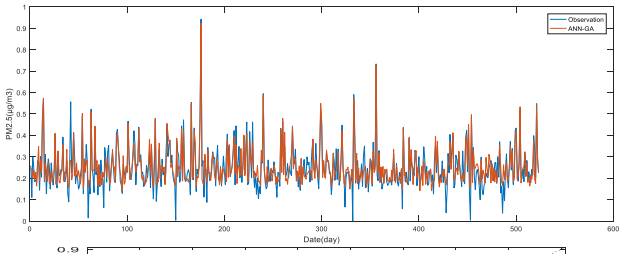
شکل ۸- همبستگی پیش بینی شبکه ANN و مقادیر واقعی مشاهده برای $PM_{2.5}$ توسط دو روش SPLINE و KNN

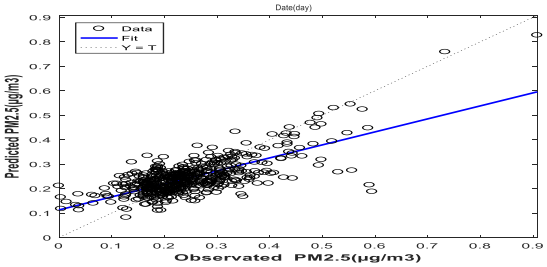


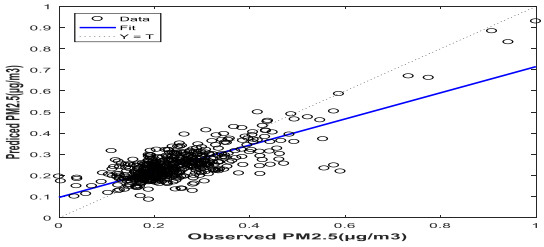
نتایج شبکه ترکیبی (ANN-GA)

در این تحقیق از یک شبکه ترکیبی ANN-GA جهت پیش بینی $PM_{2.5}$ در ایستگاه سنجش آلودگی هوا شهر ارومیه استفاده گردید. از GA به عنوان تابع آموزشی استفاده شد. برای یادگیری نحوه تکمیل فرآیند آموزش، تابع یادگیری جدید به چندین فرآیند جانبی از جمله ایجاد تابع هزینه، انتخاب، تقاطع و جهش نیاز دارد. در فرآیند انتخاب از سه روش انتخاب رولت (Roulette Wheel)، مسابقات (Tournament) و تصادفی (Random) استفاده شده است. نتایج بدست آمده از این تحقیق بیانگر کارآیی مدل ترکیبی شبکه عصبی با الگوریتم ژنتیک می‌باشد. روش های پر کردن شکاف داده های KNN و SPLINE در ANN-GA باعث بهبود دقت پیش‌بینی $PM_{2.5}$ شده است. (جدول ۳). بر اساس نتایج جدول ۳، روش KNN بهترین عملکرد را در فرآیند انتخاب داشته است (مقدار ضریب همبستگی R ۰/۹۱ و MSE ۰/۰۰۱). در شکل های ۹ تا ۱۲ نیز عملکرد ANN-GA در پیش بینی توسط سه روش انتخابی و دو روش پر کردن شکاف داده ها نشان داده شده است. در مقایسه مقادیر ضریب همبستگی R و MSE در شبکه عصبی تنها (ANN) با KNN (۰/۵۳۴ و ۰/۰۲۳) و شبکه ترکیبی ANN-GA (۰/۹۱ و ۰/۰۰۱) می توان دریافت که ضریب R بهبود ۴۰ درصدی و MSE بهبود ۹۵ درصدی داشته است. نتایج تحقیقات پیشین نیز با این مطالعه تطابق دارند (Delavar et al., ۲۰۱۹؛ Ismail, ۲۰۱۲ Ghazali & ۲۰۱۴ ،de Mattos Neto et al.). Ghazali و Ismail در مطالعه خود برای پیش بینی وضعیت آلودگی هوا با استفاده از شبکه مصنوعی ساده به مقدار خطای MSE ۰/۰۶۷ و ضریب همبستگی ۰/۵۶ رسیدند. این مقادیر برای شبکه عصبی ساده با استفاده از روش های پر کردن شکاف داده ها به ترتیب ۰/۰۱ و ۰/۵۵ بود. در تحقیق دیگر با استفاده از فیلتر SG و کاربرد الگوریتم ژنتیک در انتخاب پارامتر ضریب همبستگی بهبود ۷۰ درصدی را داشته است (Delavar et al., ۲۰۱۹). همچنین در تحقیقی که توسط de Mattos Neto و همکاران در سال ۲۰۱۴ در مورد پیش بینی وضعیت آلودگی هوا با شبکه هیبریدی عصبی-ژنتیک انجام گردید، مقدار MSE ۰/۰۰۲۸ در پیش بینی $PM_{2.5}$ بود. دلیل عملکرد بهتر شبکه ترکیبی ANN-GA در پیش بینی $PM_{2.5}$ را می توان به افزایش دقت و سرعت تشخیص خطا با گنجاندن پارامترهای بیشتر در فرآیند بهینه سازی (مانند تعداد لایه های پنهان) توسط GA مرتبط دانست (Zaini et al., ؛Nematzadeh, ۲۰۱۶ Asghari & ۲۰۲۲).



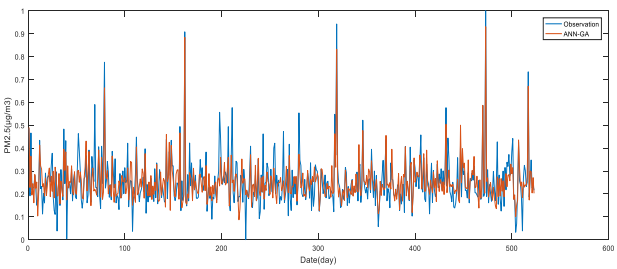


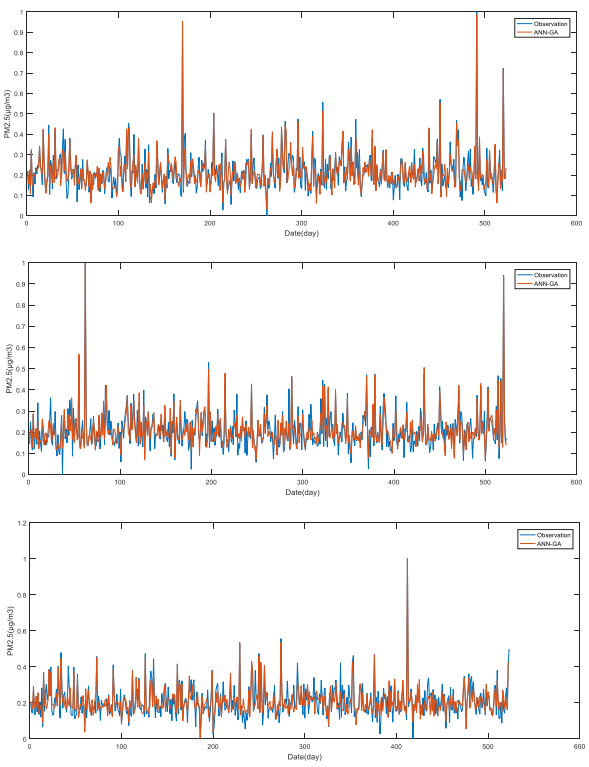




جدول ۳- نتایج شبکه ترکیبی ANN-GA در پیش بینی $PM_{2.5}$

Gap-Filling Method	Selection Process	MSE	R
SPLINE	Roulette Wheel	۰.۰۰۳۷	۰.۷۳
	Tournament	۰.۰۰۱۸	۰.۸۷
	Random	۰.۰۰۳۳	۰.۸۱
KNN	Roulette Wheel	۰.۰۰۱	۰.۹۱
	Tournament	۰.۰۰۱	۰.۸۹
	Random	۰.۰۰۰۱	۰.۸۵

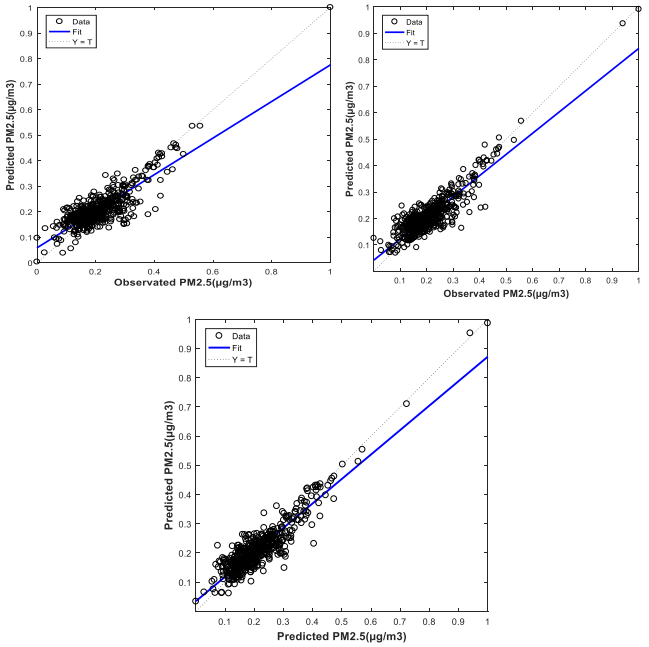




شکل ۱۱- مقایسه میزان پیش بینی شبکه ترکیبی ANN-GA (انتخاب به ترتیب با سه روش Roulette Wheel، Tournament و Random) و مقادیر واقعی مشاهده شده $PM_{2,5}$ به روش پر کردن KNN

۴- نتیجه‌گیری

روش های پیش بینی آلاینده ها به عنوان یک ابزار مهم در تحقیقات مربوط به آلودگی هوا مد نظر بوده اند. در میان آلاینده های مختلف اثرگذار بر کیفیت هوا، ذرات با قطر آیرودینامیکی کمتر از ۲/۵ میکرومتر ($PM_{2,5}$) یکی از مسائل اصلی در مدیریت کنترل آلودگی هوا هستند. در این مطالعه، مدلی برای پیش‌بینی غلظت آتی $PM_{2,5}$ توسط شبکه هیبریدی (ANN-GA) ایجاد شد. دو روش پر کردن داده های گمشده (روش های KNN و SPLINE) برای به حداقل رساندن سوگیری آموزشی و بهبود دقت شبکه استفاده شدند. PM_{10}، $PM_{2,5}$، دی اکسید نیتروژن، دی اکسید گوگرد، مونوکسید کربن و داده های هواشناسی برای پیش بینی ها استفاده شدند. نتایج نشان می‌دهد که شبکه عصبی چندلایه برای اهداف پیش‌بینی نسبتاً کارآمد است، اما دقت کافی برای پیش‌بینی را ندارد. شبکه ANN تنها با روش های پر کردن شکاف داده ها، خطای MSE ۰/۰۲۳ و ضریب همبستگی R ۰/۵۴۳را تولید کرد. به منظور بهبود مقادیر همبستگی و کاهش خطای شبکه، از الگوریتم ژنتیک در ترکیب با شبکه پرسپترون چند لایه (ANN-GA) استفاده شد. همانطور که نتایج نشان داد، MSE و ضریب همبستگی R برای شبکه هیبریدی (ANN-GA) به ترتیب ۰/۰۰۱ و ۰/۹۱ بودند. علاوه بر این، در مقایسه با شبکه منفرد، ضریب همبستگی ۴۰ درصد افزایش و MSE ۹۵ درصد بهبود یافته است. بنابراین می توان نتیجه گرفت که (ANN-GA) می تواند به عنوان ابزاری قدرتمند و قابل اعتماد برای پیش بینی آلودگی هوا مورد استفاده قرار گیرد.



شکل ۱۲- مقایسه همبستگی پیش بینی شبکه ترکیبی ANN-GA (انتخاب به ترتیب با سه روش Roulette Wheel، Tournament و Random)و مقادیر واقعی مشاهده شده $PM_{2,5}$ به روش پر کردن KNN

منابع

- خداقلی، ب.، باقری، م.، ۱۳۹۸. استفاده از فیلتر ساویتزکی-گولای برای ارتقا کیفیت تصاویر لرزه ای، اولین همایش ملی پردازش سیگنال و تصویر در ژئوفیزیک، شاهرود، ایران.
- نوری, ا., قنبرزاده لک, م., موسوی مغانجوقی, س.، ۱۳۹۶. بررسی منشا بحران آلودگی هوای شهر ارومیه، اولین کنفرانس ملی مهندسی زیرساخت ها، ارومیه، ایران.
- Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M.Đ. and Perić-Grujić, A.A. ۲۰۱۳. PM_{10} emission forecasting using artificial neural networks and gen algorithm input variable optimization. Science of the Total Environment, ٤٤٣, pp.٥١١-٥١٩.
- Asghari, M., & Nematzadeh, H. ٢٠١٦. Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network. Journal of AI and Data Mining, ٤(١), ٤٩-٥٤.
- Azami, H., Mohammadi, K. and Hassanpour, H. ٢٠١١. An improved signal segmentation method using genetic algorithm. International Journal of Computer Applications, ٢٩(٨), pp.٥-٩

- Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G. and Di Carlo, P. ۲۰۱۷. Recursive neural network model for analysis and forecast of PM۱۰ and PM۲. ۵. Atmospheric Pollution Research, ۸(٤), pp.٦٥٢-٦٥٩.
- Booker, L.B., Goldberg, D.E. and Holland, J.H. ۱۹۸۹. Classifier systems and genetic algorithms. Artificial intelligence, ٤٠(۱-۳), pp.۲۳۵-۲۸۲.
- Cabaneros, S.M.S., Calautit, J.K.S. and Hughes, B.R. ۲۰۱۷ Hybrid artificial neural network models for effective prediction and mitigation of urban roadside NO۲ pollution. Energy Procedia, ۱٤۲, pp.۳٥۲٤-۳٥۳۰.
- Coman, A., Ionescu, A. and Candau, Y. ۲۰۰۸. Hourly ozone prediction for a ۲٤-h horizon using neural networks. Environmental Modelling & Software, ۲۳(۱۲), pp.۱٤۰۷-۱٤۲۱.
- Caudell, T.P. and Dolan, C.P. ۱۹۸۹. Parametric connectivity: training to constrained networks using genetic algorithms. In Proceedings of the Third International Conference on Genetic Algorithms, pp. ۳۷۰-۳۷٤.
- Cantú-Paz, E. and Kamath, C. ۲۰۰٥. An empirical comparison of combinations of evolutionary algorithms and neural networks for classification problems. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), ۳٥(٥), pp.۹۱٥-۹۲۷
- Chen, Y., Shi, R., Shu, S., & Gao, W. ۲۰۱۳Ensemble and enhanced PM۱۰ concentration forecast model based on stepwise regression and wavelet analysis. Atmospheric Environment, ۷٤, pp.۳٤٦-۳٥۹.
- Chen, Y., Cao, R., Chen, J., Liu, L. and Matsushita, B. ۲۰۲۱. A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter. ISPRS Journal of Photogrammetry and Remote Sensing, ۱۸۰, pp.۱۷٤-۱۹۰.
- Cho, B., Dayrit, T., Gao, Y., Wang, Z., Hong, T., Sim, A. and Wu, K. ۲۰۲۰. Effective missing value imputation methods for building monitoring data. In IEEE International Conference on Big Data (Big Data), pp. ۲۸٦٦-۲۸۷٥. DOI: ۱۰,۱۱۰۹/BigData٥۰۰۲۲,۲۰۲۰,۹۳۷۸۲۳۰.
- De Mattos Neto, P. S., Madeiro, F., Ferreira, T. A., & Cavalcanti, G. D. ۲۰۱٤. Hybrid intelligent system for air quality forecasting using phase adjustment. Engineering Applications of Artificial Intelligence, ۳۲, ۱۸٥-۱۹۱.
- Dede, T., Bekiroğlu, S. and Ayvaz, Y. ۲۰۱۱. Weight minimization of trusses with genetic algorithm. Applied Soft Computing, ۱۱(۲), pp.۲٥٦٥-۲٥۷٥.
- Delavar, M.R., Gholami, A., Shiran, G.R., Rashidi, Y., Nakhaeizadeh, G.R., Fedra, K. and Hatefi Afshar, S. ۲۰۱۹. A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS International Journal of Geo-Information, ۸(۲), p.۹۹.
- Ding, S., Su, C. and Yu, J. ۲۰۱۱. An optimizing BP neural network algorithm based on genetic algorithm. Artificial intelligence review, ۳٦(۲), pp.۱٥۳-۱٦۲.
- Dongare, A.D., Kharde, R.R. and Kachare, A.D. ۲۰۱۲. Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT), ۲(۱), pp.۱۸۹-۱۹٤.
- Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. ۲۰۱٤. Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmospheric Pollution Research, ٥(٤), ٦۹٦-۷۰۸.
- Esfandani, M.A. and Nematzadeh, H. ۲۰۱٥. Prediction of air pollution in Tehran based on evolutionary models. Indian journal of science and technology, ۸(۳٥), pp.۱-٥.
- Farhadi, R., Hadavifar, M., Moeinaddini, M. and Amintoosi, M. ۲۰۲۰. Prediction of the Air Quality by Artificial Neural Network Using Instability Indices in the City of Tehran-Iran. AUT Journal of Civil Engineering, ٤(٤), pp.۹-۹.
- Fogel, D.B., Fogel, L.J. and Porto, V.W. ۱۹۹۰. Evolving neural networks. Biological cybernetics, ٦۳(٦), pp.٤۸۷-٤۹۳.
- Ghazali, S. and Ismail, L.H. ۲۰۱۲. Air quality prediction using artificial neural network. In Proceedings of the International Conference on Civil Environmental Engineering Sustainability, Johor Bahru, Malaysia, Vol. ۳٥, pp. ۱٥.
- Haykin, S. S. ۲۰۰۹. Neural networks and learning machines/Simon Haykin. In: New York: Prentice Hall.
- Harishkumar, K., Yogesh, K., & Gad, I. ۲۰۲۰. Forecasting air pollution particulate matter (PM۲.٥) using machine learning regression models. Procedia Computer Science, ۱۷۱, ۲۰٥۷-۲۰٦٦.

- Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A. and de Argandoña, J.D. ۲۰۰۸. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. Environmental Modelling & Software, ۲۳(۵), pp.۶۲۲-۶۳۷.
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. ۲۰۰٤. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, ۳۸(۱۸), pp.۲۸۹۵-۲۹۰۷.
- Kandasamy, S., Baret, F., Verger, A., Neveux, P. and Weiss, M. ۲۰۱۳. A comparison of methods for smoothing and gap filling time series of remote sensing observations–application to MODIS LAI products. Biogeosciences, ۱۰(٦), pp.٤۰۵۵-٤۰۷۱.
- Kang, Z. and Qu, Z. ۲۰۱۷. Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou. ۲nd IEEE international conference on computational intelligence and applications, (ICCIA) (pp. ۱۵۵-۱٦۰). IEEE.
- Kow, P.-Y., Wang, Y.-S., Zhou, Y., Kao, I. F., Issermann, M., Chang, L.-C., & Chang, F.-J. ۲۰۲۰. Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM۲٫۵ forecasting. Journal of Cleaner Production, ۲٦۱, ۱۲۱۲۸۵.
- Liang J., Ren C., Li Y., Yue W., Wei Z., Song X., Zhang X., Yin A., Lin X. ۲۰۲۳. Using Enhanced Gap-Filling and Whittaker Smoothing to Reconstruct High Spatiotemporal Resolution NDVI Time Series Based on Landsat ۸, Sentinel-۲, and MODIS Imagery. ISPRS International Journal of Geo-Information. ۱۲(٦):۲۱٤. https://doi.org/۱۰,۳۳۹۰/ijgi۱۲۰٦۰۲۱٤.
- Luo J., Ying K., Bai J. ۲۰۰۵. Savitzky–Golay smoothing and differentiation filter for even number data. Signal processing. ۸۵(۷), pp.۱٤۲۹-۳٤.
- Lu, H.-C., Hsieh, J.-C., & Chang, T.-S. ۲۰۰٦. Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network. Atmospheric Research, ۸۱(۲), pp.۱۲٤-۱۳۹.
- Lu, Y., Shi, K., Yong, J., Gu, H. and Song, H. ۲۰۱٦. A B-spline curve extension algorithm. Science China Information Sciences, ۵۹(۳), pp.۱-۹.
- Momeni, E., Nazir, R., Armaghani, D.J. and Maizir, H. ۲۰۱٤. Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN. Measurement, ۵۷, pp.۱۲۲-۱۳۱.
- Montana, D.J. and Davis, L. ۱۹۸۹. August. Training feedforward neural networks using genetic algorithms. In IJCAI, ۸۹, pp.۷٦۲-۷٦۷.
- Ordieres, J. B., Vergara, E. P., Capuz, R. S., & Salazar, R. E. ۲۰۰۵. Neural network prediction model for fine particulate matter (PM۲٫۵) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environmental Modelling & Software, ۲۰(۵), pp.۵٤۷-۵۵۹.
- Rosenblatt, F. ۱۹٦۱. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Cornell Aeronautical Lab Inc Buffalo, NY.
- Salcedo-Sanz, S., Xu, Y. and Yao, X. ۲۰۰٦. Hybrid meta-heuristics algorithms for task assignment in heterogeneous computing systems. Computers & operations research, ۳۳(۳), pp.۸۲۰-۸۳۵.
- Samoli, E., Analitis, A., Touloumi, G., Schwartz, J, Anderson, H.R., Sunyer, J., Bisanti, L., Zmirou, D., Vonk, J.M., Pekkanen, J. and Goodman, P. ۲۰۰۵. Estimating the exposure–response relationships between particulate matter and mortality within the APHEA multicity project. Environmental health perspectives, ۱۱۳(۱), pp.۸۸-۹۵.
- Schmid, M., Rath, D. and Diebold, U. ۲۰۲۲. Why and how Savitzky–Golay filters should be replaced. ACS Measurement Science, ۲(۲), pp.۱۸۵-۱۹٦.
- Schwartz, J. and Lepeule, J. ۲۰۱۲. Is ambient PM۲. ۵ sulfate harmful? Schwartz and Lepeule Respond. Environmental Health Perspectives, ۱۲۰(۱۲), pp. a٤۵٤-a٤۵۵.
- Ścibor, M., Bokwa, A. and Balcerzak, B. ۲۰۲۰. Impact of wind speed and apartment ventilation on indoor concentrations of PM۱۰ and PM۲. ۵ in Kraków, Poland. Air Quality, Atmosphere & Health, ۱۳(۵), pp.۵۵۳-۵٦۲.
- Whitley, L.D. and Hanson, T. June. ۱۹۸۹. Optimizing neural networks using FasterMore accurate genetic search. In Proceedings of the ۳rd international conference on genetic algorithms, pp. ۳۹۱-۳۹۷.
- Zaini, N.A., Ean, L.W., Ahmed, A.N., Abdul Malek, M. and Chow, M.F. ۲۰۲۲. PM۲٫۵ forecasting for an urban area based on deep learning and decomposition method. Scientific Reports, ۱۲(۱), p.۱۷۵٦۵.

- Zhao, H., Zhang, J., Wang, K., Bai, Z. and Liu, A. ۲۰۱۰. A GA-ANN model for air quality predicting. International Computer Symposium, pp. ٦٩٣-٦٩٩.
- Zhang, M.L. and Zhou, Z.H. ۲۰۰۷. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, ٤٠(٧), pp.٢٠٣٨-٢٠٤٨.
- Zhang, S., Li, X., Zong, M., Zhu, X. and Cheng, D. ۲۰۱۷. Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), ٨(٣), pp.١-١٩.
- Zhang, Y., Wang, J., Chen, L., Yang, H., Zhang, B., Wang, Q., Hu, L., Zhang, N., Vedal, S., Xue, F. and Bai, Z. ۲۰۱۹. Ambient PM۲. ٥ and clinically recognized early pregnancy loss: A case-control study with spatiotemporal exposure predictions. Environment International, ١٢٦, pp.٤٢٢-٤٢٩.
- Zuo, C., Chen, Q., Yu, Y. and Asundi, A. ۲۰۱۳. Transport-of-intensity phase imaging using Savitzky-Golay differentiation filter-theory and applications. Optics express, ٢١(٥), pp.٥٣٤٦-٥٣٦٢.

Prediction of $PM_{2.5}$ using a hybrid network (ANN-GA) Case study: Urmia city

Mohammad Teyefeh Taherloo[۱]; Amir Asadi Vaighan[۲*]

[۱] MSc., Department of Civil Engineering, Faculty of Engineering, Urmia University, Urmia, Iran

[۲*] Assistant Professor, Department of Civil Engineering, Faculty of Engineering, Urmia University, Urmia, Iran

Abstract

Introduction

For the last ۵۰ years, activities like urbanization, industrialization and population growth, make air as a significant inseparable part of our life. Air pollution can be defined as the presence of chemicals or toxic compounds in the air to extent that they pose a health risk. Emissions from cars, plant chemicals, dust, pollen and mold spores are introduced as particulate matter (PM). The World Health Organization reported that ambient air pollution causes ٤,٢ million deaths from strokes, heart disease, lung cancer and chronic respiratory diseases. Of the various pollutants affecting air quality, particulate matter smaller than ٢,٥ microns is the major air pollution problem (Ścibor et al., ٢٠٢٠). As well, there is growing evidence of the effects of PM١٠ and PM٢,٥ on cardiovascular disease (CVD) and respiratory disease (DR). Forecasting air pollutants provides an opportunity to determine the intensity of air pollution in different areas and prevent irreversible impacts. In addition, these models also allow decision-makers to make the right decisions and prepare for the prevention or control of the PMs in the future. Some of the models used in air pollution forecasting studies are auto-regressive Integrated Moving Average (ARIMA), artificial neural network (ANN), Community Multiscale Air Quality Model (CMAQ), the Weather Research and Forecasting (WRF) model coupled with Chemistry (WRF-CHEM), Fuzzy models, grey model and/or hybrid models. ANN has been used extensively by scientists to provide rapid and parsimonious solutions to mitigate the negative impacts of air pollution worldwide. Neural networks, as an alternative, have been successfully used in air pollution forecasting and have produced accurate results in time series data. Different types of noise and nonlinear structure were present in the data. Hybrid modeling approaches have a wide variety of applications in which numerous methods or attributes are merged to create a more sophisticated model with superior performance in certain scenarios. Urmia is one of Iran's most polluted cities, owing to continuous traffic and traffic congestion, growing $CO_{۲}$ and PM levels, and a lack of knowledge on regulating and locating industrial manufacturing units. Dust from Iraq affects the region, as well as inversion, which occurs ۹۰ days a year, are instances of region-specific air pollution. In addition, the drying of Urmia Lake, which can result in salt storms, is one of the critical concerns that will lead to significant pollution in the near future. In this study, ANN-GA with missing data imputation was used to predict $PM_{۲,۵}$ in Urmia, Iran, in the short-term to demonstrate how data-gap filling and preprocessing methods could improve hybrid models' performance.

Methodology

The concentrations of air pollutants (carbon monoxide, nitrogen dioxide, and sulfur dioxide) as well as meteorological data (temperature, relative humidity, and wind velocity) were used as inputs in this research to predict $PM_{۲,۵}$. Air pollution concentrations and meteorological data over a two-year period were obtained from Monitoring Station No. ۳, Urmia municipality, and Iran's meteorology website (Data.irimo.ir). The data was then preprocessed with the Savitsky-Golay filter before being fed into the ANN and ANN-GA networks. Data gaps and imputed data (KNN/SPLINE method) were used as input in each network, and the results were compared.In this study, a single system contains two hidden layers and one output layer. The time series method was used to introduce the data to the network. The data was divided into three parts. ۷۰٪ of the data is used for training, ۱۵٪ for validation, and ۱۵٪ for testing. Data import scenarios were defined in two ways. The first scenario used no imputation, while the second used SPLINE and KNN to fill in data gaps. As a transfer function, a sigmoid (logsig) layer was used for hidden layers,

and a linear layer (Purelin) was used for the output layer. The Levenberg-Marquardt algorithm was chosen as the learning algorithm based on the type of problem and the speed of convergence. To improve the results, the number of neurons, repetition parameters, number of permitted evaluations, Levenberg algorithm parameters, and reliability were all adjusted through a trial-and-error process.New ANN-GA network was used in this study and GA was used as a training function. After introducing the data as a time series and selecting the amount of data for each episode of learning, evaluation, and testing, the structure and number of network layers were created with the "newff" function. The main difference is that the genetic learning process was used instead of the "train" function. It's worth noting that the network layer characteristics in both methods were the same. To learn how to complete the process, the new learning function requires several side processes, including cost function creation, selection, intersection, and mutation. Three methods of roulette selection, tournament selection and random were used in the selection process. To introduce the cost function, weights were taken from those created by the "newff" function. Different values were assigned to the initial population variables, maximum mutation number, and selection pressure coefficient by trial-and-error method. Moreover, two data import scenarios were defined.

Conclusion

Forecasting methods have been considered an important tool in research on air pollution. Among the various pollutants that influence air quality, particles with an aerodynamic diameter of less than ۲٫۵ micrometers ($PM_{۲٫۵}$) are one of the key issues in air pollution control management. In this study, a model for predicting future concentrations of $PM_{۲٫۵}$ was developed by the Hybrid Network (ANN-GA). Two methods of data imputation (KNN and SPLINE) were used to minimize training issues and improve network accuracy. $PM_{۱۰}$, $PM_{۲٫۵}$, nitrogen dioxide, oxide, carbon monoxide, and weather data were used for predictions. The results show that multi-line neural networks are relatively efficient for predictive purposes but lack sufficient accuracy to predict. The ANN network produced MSE error of ۰٫۰۲۳ and coherence coefficient of R ۰٫۵۴۳ only with data gap filling methods. In order to improve R and reduce network errors, a genetic algorithm was used in combination with a multi-layer neural network (ANN-GA). As the results showed, MSE and R for hybrid networks (ANN-GA) were improved (R=۰٫۹۱ and MSE=۰٫۰۰۱). In addition, compared to ANN, the R increased by ۴۰ percent and the MSE improved by ۹۵ percent. Thus, it can be concluded that ANN-GA can be used as a powerful and reliable tool for predicting air pollution.

Keywords

Air pollution Prediction; Artificial Neural Network; Genetic Algorithm; $PM_{۲٫۵}$